From representation learning to thematic classification - Application to hierarchical analysis of hyperspectral images

Adrien Lagrange

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (Toulouse INP)

Discipline ou spécialité :

Signal, Image, Acoustique et Optimisation

Présentée et soutenue par :

M. ADRIEN LAGRANGE

le mercredi 6 novembre 2019

Titre :

From representation learning to thematic classification - Application to hierarchical analysis of hyperspectral images

---

Ecole doctorale :

Mathématiques, Informatique, Télécommunications de Toulouse (MITT)

Unité de recherche :

Institut de Recherche en Informatique de Toulouse (IRIT)

Directeur(s) de Thèse :

M. NICOLAS DOBIGEON

M. MATHIEU FAUVEL

Rapporteurs :

M. JEROME BOBIN, CEA SACLAY

M. PAUL SCHEUNDERS, UNIVERSITE INSTELLING ANTWERPEN

Membre(s) du jury :

M. CHARLES BOUVEYRON, CNRS COTE D'AZUR, Président

M. BERTRAND LE SAUX, ONERA - CENTRE DE PALAISEAU, Membre

M. MATHIEU FAUVEL, INRA TOULOUSE, Membre

M. MAURO DALLA MURA, GIPSA-LABO GRENOBLE CAMPUS, Membre

Mme EMILIE CHOUZENOUX, UNIVERSITE PARIS-EST MARNE LA VALLEE, Membre

M. NICOLAS DOBIGEON, TOULOUSE INP, Membre

M. STEPHANE MAY, CENTRE NATIONAL D'ETUDES SPATIALES CNES, Invité

Remerciements

Sans emphases et sans ambages, je tiens à remercier toutes les personnes qui m'ont aidées durant les trois années que j'ai consacré à ma thèse.

Tout d'abord, je remercie Mathieu Fauvel et Nicolas Dobigeon, mes deux directeurs de thèse, pour leurs disponibilités, leurs conseils et leurs idées. Ce fut un plaisir pour moi de travailler avec eux. Merci également à Stéphane May pour les fructueuse discussions que j'ai eu le plaisir d'avoir avec lui.

Je remercie ensuite mes rapporteurs Pr. Paul Scheunders et Jérome Bobin d'avoir pris le temps d'évaluer mon travail. J'aimerais également remercier Pr. Charles Bouveyron pour avoir présidé le jury ainsi que M. Bertrand Le Saux, M. Mauro Dalla Mura et Mme Émilie Chouzenoux d'avoir siégé dans ce même jury. Ce fut un honneur et un plaisir d'avoir pu leur présenter mes travaux.

Je remercie également le Pr. José Bioucas-Dias de m'avoir accueilli à Lisbonne et de m'avoir conseillé dans mes travaux.

J'aimerais avoir les mots pour remercier individuellement chacun de mes collègues de l'équipe SC de l'IRIT. Je les prie de bien vouloir m'excuser d'exprimer mes remerciements aussi succinctement mais qu'ils sachent que la bonne humeur, la simplicité et la convivialité présentent au sein de l'équipe ont grandement adouci ces trois années intenses. Merci à Louis, Olivier, Étienne, Dylan, Pierre-Antoine, Vinicius, Yanna, Tatsumi, Claire, Vinicius, Maxime, Camille, Serdar, Pierre-Hugo, Baha, Mouna, Marie, Thomas, Cédric, Emmanuel et tout ceux que j'oublie...

J'ajoute également un remerciement à tout le personnel administratif de l'IRIT, et en particulier à Annabelle, pour leur aide, leur patience et leur gentillesse.

Enfin, je remercie toute ma famille pour leur soutien durant durant ces trois années. Si je l'exprime peu, leurs soutiens m'est précieux. Dernière évoquée même si première dans mon cœur, je remercie finalement Marion qui m'accompagne et me soutiens tous les jours.

Résumé

De nombreuses approches ont été développées pour analyser la quantité croissante de donnée image disponible. Parmi ces méthodes, la classification supervisée a fait l'objet d'une attention particulière, ce qui a conduit à la mise au point de méthodes de classification efficaces. Ces méthodes visent à déduire la classe de chaque observation en se basant sur une nomenclature de classes prédéfinie et en exploitant un ensemble d'observations étiquetées par des experts. Grâce aux importants efforts de recherche de la communauté, les méthodes de classification sont devenues très précises. Néanmoins, les résultats d'une classification restent une interprétation haut-niveau de la scène observée puisque toutes les informations contenues dans une observation sont résumées en une unique classe. Contrairement aux méthodes de classification, les méthodes d'apprentissage de représentation sont fondées sur une modélisation des données et conçues spécialement pour traiter des données de grande dimension afin d'en extraire des variables latentes pertinentes. En utilisant une modélisation basée sur la physique des observations, ces méthodes permettent à l'utilisateur d'extraire des variables très riches de sens et d'obtenir une interprétation très fine de l'image considérée.

L'objectif principal de cette thèse est de développer un cadre unifié pour l'apprentissage de représentation et la classification. Au vu de la complémentarité des deux méthodes, le problème est envisagé à travers une modélisation hiérarchique. L'approche par apprentissage de représentation est utilisée pour construire un modèle bas-niveau des données alors que la classification, qui peut être considérée comme une interprétation haut-niveau des données, est utilisée pour incorporer les informations supervisées. Deux paradigmes différents sont explorés pour mettre en place ce modèle hiérarchique, à savoir une modélisation bayésienne et la construction d'un problème d'optimisation. Les modèles proposés sont ensuite testés dans le contexte particulier de l'imagerie hyperspectrale où la tâche d'apprentissage de représentation est spécifiée sous la forme d'un problème de démélange spectral.

Mots clés : analyse d'image, classification, apprentissage de représentation, télédétection, imagerie hyperspectrale.

Abstract

Numerous frameworks have been developed in order to analyze the increasing amount of available image data. Among those methods, supervised classification has received considerable attention leading to the development of state-of-the-art classification methods. These methods aim at inferring the class of each observation given a specific class nomenclature by exploiting a set of labeled observations. Thanks to extensive research efforts of the community, classification methods have become very efficient. Nevertheless, the results of a classification remains a high-level interpretation of the scene since it only gives a single class to summarize all information in a given pixel. Contrary to classification methods, representation learning methods are model-based approaches designed especially to handle high-dimensional data and extract meaningful latent variables. By using physic-based models, these methods allow the user to extract very meaningful variables and get a very detailed interpretation of the considered image.

The main objective of this thesis is to develop a unified framework for classification and representation learning. These two methods provide complementary approaches allowing to address the problem using a hierarchical modeling approach. The representation learning approach is used to build a low-level model of the data whereas classification is used to incorporate supervised information and may be seen as a high-level interpretation of the data. Two different paradigms, namely Bayesian models and optimization approaches, are explored to set up this hierarchical model. The proposed models are then tested in the specific context of hyperspectral imaging where the representation learning task is specified as a spectral unmixing problem.

keywords: image analysis, classification, representation learning, remote sensing, hyperspectral imaging.

Contents

List of Figures

List of Tables

Introduction (in French)

Au cours des dernières décennies, d'importants progrès ont été accomplis dans le domaine connu actuellement sous le nom d'intelligence artificielle ou d'apprentissage automatique. L'un des moteurs de cette révolution a été le développement d'algorithmes pour l'interprétation automatique d'images. Il est par exemple possible de citer l'émergence dans les années 90 des machines à vecteurs de support (SVM), introduites d'abord pour la reconnaissance de chiffres manuscrits [BGV92]. Dans les années qui suivirent, les réseaux de neurones profonds convolutionnels ont également été conçus pour résoudre ce même problème [LeC+98] et sont maintenant l'une des méthodes d'apprentissage les plus populaires.

L'attention croissante dont ont bénéficié ces technologies de pointe a amené les chercheurs et les utilisateurs à appliquer ces méthodes d'interprétation automatique dans de nombreux domaines d'application. En imagerie, de nombreuses méthodes d'analyse d'images ont été développées depuis la reconnaissance de chiffres manuscrits pour de nombreux cas d'application, par exemple la génération de cartes thématiques [LKC15], la segmentation d'images médicales [Ban08], la reconnaissance faciale [JL11], etc. Les méthodes de classification très populaires, telles que les SVMs ou les réseaux de neurones profonds, fournissent dorénavant de très bons résultats pour bon nombre de ces tâches.

Cependant, même si ces méthodes se sont révélées très efficaces, elles sont encore confrontées à des problèmes délicats comme la grande dimension des données, le manque de données labellisées, leur mauvaise labellisation ou encore le caractère multi-modale des classes considérées. Il a également été avancé que les résultats fournis par un classifieur, qui sont généralement un unique label par élément (un pixel, une image, ...), sont quelque peu limités. En particulier, nombre de ces algorithmes restent très obscurs dans leur processus de décision. Les réseaux de neurones profonds sont par exemple souvent considérés comme des algorithmes "boîte noire", bien que leur décision soit très précise [Cas16 ; Moo+17]. De plus, les méthodes de classification les plus utilisées ne recourent généralement pas à une modélisation du signal observé. Pour cette raison, il est difficile pour un spécialiste de guider l'interprétation par des connaissances experts sur la donnée observée.

Pour surmonter ces limitations, une alternative consiste à recourir à des approches fondées sur une modélisation des données. Les méthodes de classification sont principalement empiriques, c'est-à-dire que la règle de décision est uniquement apprise à partir d'un ensemble d'exemples. Au contraire, les approches de type modélisation reposent sur une modélisation physique des données (signaux observés, images ou mesures). Par exemple, en imagerie médicale, les modalités d'image sont généralement associées à un modèle physique du signal mesuré, dérivé des modalités particulières d'acquisition et d'un bruit spécifique [Cav+18b]. Parmi les approches basées sur des modèles, les méthodes d'apprentissage de représentation ont fait l'objet d'une attention importante.

Ces méthodes sont fondées sur l'hypothèse que les observations ne couvrent pas tout l'espace d'observation, mais sont en réalité contenues dans un sous-espace [BN08]. L'apprentissage de représentation vise à identifier ce sous-espace et à estimer la représentation de chaque observation dans celui-ci afin d'obtenir une représentation plus compacte, c'est-à-dire de dimension plus faible. Cette représentation de faible dimension est vue comme un ensemble de facteurs latents. Lorsque le modèle est construit à l'aide de connaissances a priori sur le domaine d'application, ces facteurs latents ont généralement une signification physique. Du point de vue de l'utilisateur, la possibilité de guider la méthode d'analyse afin d'estimer des paramètres spécifiques permet une interprétation beaucoup plus riche des résultats. Les produits annexes de ces méthodes d'apprentissage de représentation peuvent en effet présenter un grand intérêt. Par exemple, dans le cas du démélange hyperspectral, chaque vecteur de la base du sous-espace latent est associé à un matériau présent dans la scène observée [Bio+12].

Bien que la classification et l'apprentissage de représentation sont deux méthodes couramment utilisées, elles n'ont que très rarement été envisagées conjointement. L'objectif de cette thèse est d'introduire le concept d'apprentissage de représentation et de classification conjoints. Les modèles unifiés développés sont ensuite testés sur un cas d'application particulier qu'est l'analyse d'images hyperspectrales.

Structure du manuscrit

La première approche envisagée vise à mettre en place un nouveau modèle bayésien permettant d'estimer simultanément les classes et les représentations latentes. Pour cela, l'algorithme d'apprentissage de représentation considéré intègre une segmentation spatiale selon l'homogénéité des vecteurs de représentation latente. Dans l'approche proposée, le modèle de segmentation est complété de sorte à dépendre également des classes. La classification

est donc intégrée au modèle et exploite à la fois la donnée supervisée et la segmentation, qui intègre l'information bas-niveau, obtenant ainsi un classifieur robuste aux erreurs sur les données externes. L'algorithme fournit alors une description hiérarchique de l'image en termes de vecteurs latents, de segmentation spatiale et de classification thématique.

La deuxième approche considérée s'appuie sur la même description hiérarchique mais l'inférence est formulée comme un problème d'optimisation. La fonction de coût comprend alors trois termes principaux correspondant aux trois tâches considérées : l'apprentissage de représentation, la segmentation et la classification. Le problème obtenu s'apparente à un problème de cofactorisation de matrices avec un terme de segmentation liant les activations des deux factorisations agissant respectivement comme modèle de représentation et de classification. Une solution de ce problème non-convexe et non-lisse est ensuite approchée à l'aide d'un algorithme de descente de gradient proximal alternée.

Le troisième travail réalisé vise à intégrer dans le processus de démélange hyperspectral une information spatiale complémentaire. L'originalité de la proposition réside dans le fait que l'information spatiale n'est pas introduite via un terme de régularisation mais comme un second terme d'attache aux données calculé à partir d'une image panchromatique de la scène. Ce modèle complète en particulier les deux approches précédentes en mettant en place une méthode de démélange permettant une bonne estimation des spectres élémentaires en capitalisant sur la méthode de cofactorisation développée précédemment.

Principales contributions

Chapitre 1. La principale contribution de ce chapitre réside dans l'introduction d'un cadre bayésien pour unifier les approches de modélisation physique bas-niveau et de classification. Le modèle propose une utilisation de champs de Markov aléatoires pour relier tous les niveaux de modélisation afin de réaliser une estimation conjointe. La deuxième contribution est la conception d'une méthode de classification permettant de tenir compte des erreurs de labellisation dans l'ensemble d'apprentissage et de les corriger. Enfin, la dernière contribution réside dans le potentiel d'interprétation du modèle, notamment grâce à des produits annexes intéressants. En particulier, une des matrices estimées décompose chacune des classes en un ensemble de clusters chacun caractérisé par son vecteur d'abondance moyen. L'utilisateur peut ainsi analyser clairement la structure des données considérées.

Chapitre 2. Un modèle de cofactorisation est utilisé pour développer un cadre unifié alternatif. Ce modèle diffère des autres modèles de cofactorisation principalement par le terme

de couplage proposé. Premièrement, il permet une interprétation riche des résultats avec à nouveau l'idée de décomposer les classes en un ensemble de clusters. Et deuxièmement, il permet de conserver une flexibilité entre les deux tâches à accomplir contrairement aux modèles précédemment proposés [ZL10] où le modèle introduit deux objectifs antagonistes au lieu d'objectifs coopératifs. La dernière contribution réside dans la proposition d'une méthode d'optimisation avancée pour minimiser la fonctionnelle proposée. En effet, un algorithme de minimisation proximale linéarisée alternée est utilisé pour résoudre le problème à la fois non convexe et non lisse, avec une garantie de convergence vers un point critique de la fonction objectif.

Chapitre 3. La principale contribution de ce chapitre est une nouvelle proposition pour enrichir directement le modèle de démélange spectral avec de l'information spatiale. Elle consiste à utiliser un terme supplémentaire d'attache aux données au lieu de recourir à des méthodes de régularisation. Ce nouveau modèle améliore les résultats du démélange. Mais plus important encore, le modèle produit une carte de clustering caractérisant différentes zones de l'image par leur signature spectrale et leur configuration spatiale. Cela permet d'obtenir une représentation compacte, complète et visuelle de la scène analysée. À notre connaissance, cette méthode introduit pour la première fois le concept de démélange spatial et spectral conjoint.

Introduction

Over the last decades major progresses have occurred in the field of artificial intelligence. Many man-made activities have been successfully replaced by algorithms that are able to learn a given task directly from data. In particular, advances in image interpretation algorithms have been one of the driving force in this revolution. In the nineties, kernel methods, such as support vector machines (SVMs), were introduced firstly to identify handwritten digits [BGV92] and consisted in a major breakthrough. Specifically, SVMs highlighted non-probabilistic methods by proposing to minimize both a convex loss function while exploiting a set of labeled examples, and additional regularization terms to ensure a better separability of the classes. Following this trend, convolutional deep neural networks (CNN), which can automatically learn spatial features from the data, have become the top ranked methods for image recognition [LeC+98] and are now at the foundation of the most popular family of methods. Contrary to SVMs, the decision function of CNNs is a non-convex function composed of a sequence of differentiable operations. The parameters of this function are then optimized by minimizing a loss function, generally by using stochastic gradient descent. Although there is usually no convergence guarantee, CNNs manage to benefit from the huge quantity of available data to get state-of-the-art results.

The always increasing attention brought by these breakthrough technologies has pushed researchers and end-users to consider automatic interpretation methods in many fields of application such as remote sensing imaging [LKC15], medical imaging [Ban08], face recognition [JL11]. In particular, classification methods have received considerable attention. These methods aim at attributing a class to each elements of the analyzed dataset. These elements can take many forms ranging from simple pixels [Pla+09] to objects [ALL17] or images [KSH12]. The first step in classification generally consists in extracting a representation of each element of the dataset either automatically as with CNNs [BCV13], or with handcrafted features [DT05; Low99; PB01]. Then, two main cases may come forth. The first case is unsupervised classification methods for which no additional information is available with the dataset. The concept behind these methods is generally to try to identify groups

of similar elements to which the same class is assigned. A typical case is a clustering task trying for example to separate organs in a medical image [Thi+14]. The second case occurs when a so-called training set is available with the data. This training set is a collection of observations that were classified manually by an expert. The set of examples is then used to train the classification model for the considered task. Many recent works have pointed out that the use of large training set is indeed very beneficial to the classification [KSH12; Mag+16].

However, even if supervised classification methods have proven to be very efficient, they still face challenging issues:

- The **dimension of the observation** is usually a major issue. The work of [Hug68] introduced the so-called *curse of dimensionality*. It showed in particular that statistical methods made for low or moderate dimensional spaces do not adapt well to high dimensional spaces. The rate of convergence of the statistical estimation decreases when the dimension grows while the number of parameters to estimate simultaneously increases, making the estimation of the model parameters very difficult [Don00]. Beyond a certain limit, the classification accuracy actually decreases as the number of features increases [Hug68]. These problems may arise when considering observations with redundant information such as hyperspectral images [Cam+14] or video stream [Kar+14].

- The **dependence to ground-truth data** is also a recurrent limiting factor. The production of labeled data by experts is a critical work which is usually costly and time consuming. It is therefore common to be confronted with a lack of labeled data. For this reason, it is necessary to develop methods that leverage their dependence to GT data and are robust to overfitting [FM04; CFB08]. Semi-supervised methods are for example an attempt to deal with the lack of labeled data by using unlabeled data [CK05].

  Another issue regarding the training data is the presence of incorrect labels [BF99; FV13]. This can be due to ambiguity regarding the set of classes or mistakes of the expert. In any case, the robustness to such labeling noise can be an interesting feature to characterize the performance of a classification algorithm [BG09].

- Handling **multi-modal and/or composite classes** with intrinsic intra-class variability is also a recurrent issue [HT96a]. For instance, for a generic classification task, a class referred to as *humans* gathers distinct genders, or physical attributes.

When using too basic classifier, *e.g.* linear classifiers, it may actually be impossible to regroup the different modes in a single class [MP17].

It has also been argued that the outputs provided by a classifier, which are generally a unique label per elements (a pixel, an image,...) of the dataset, are somehow limited. In particular, many of these algorithms remains very obscure in their decision process. First among them, CNN algorithms are nowadays often seen as black box algorithms although very accurate in their decision [Cas16; Moo+17]. Moreover, the most used classification methods are usually model-free, *i.e.*, they are not based on a modeling of the observed signal. For this reason, when considering a specific task, it is difficult for a specialist to guide the interpretation by some prior knowledge.

To overcome this limitations, one alternative consists in resorting to model-based approaches [Idi13]. Model-free classification methods are mostly empirical in the sense that the decision rule is only learned from a set of examples. On the contrary, model-based approaches rely on a modeling of the data (observed signals, images or measurements). For example, in medical imaging, image modalities are generally associated with a specific physics-based model of the measured signal, derived from the acquisition process and particular noise corruption [Cav+18b]. Among model-based approaches, representation learning methods have received a considerable attention. Depending on the research community, representation learning has been referred to as dictionary learning methods [RPE12], matrix factorization [LS99], source separation [Bob+07], factor analysis [Cav+18b] or subspace learning [Li+15b]. These names denote representation learning methods differing mainly by the specific set of considered constraints enforced to ensure the physical interpretation of the data.

Representation learning – Representation learning is generally considered for modeling high-dimensional data. The main assumption underlying these methods is that the observations do not span the whole observation space but are actually located in a subspace [BN08]. Representation learning aims at identifying this subspace and at estimating the representation of each observation in this subspace to get a more compact representation, *i.e.*, of lower dimensionality [Ess+12]. This compact low-dimensional representation is a collection of latent factors. When the model is built in accordance with knowledge about the application field, these latent factors usually carry some physical meaning [El +06]. From the end-user point-of-view, the possibility to guide the analysis method in order to estimate specific parameters offers a richer interpretation of the results. The byproducts provided by representation learning methods can indeed be of the highest interest. For example, in the

case of hyperspectral unmixing, each vector of the basis spanning the subspace is identified to a material present in the observed scene [Bio+12].

However, a major drawback of this family of methods is the complexity of the targeted results. First, generally, representation learning results in very challenging estimation problems. Indeed, physics-based models often introduce non-convex problems [RCP14; Bob+15]. When considering optimization frameworks, such problems remain difficult to tackle and it is generally impossible to ensure convergence to a global optimum of the objective function. Some advanced methods can at least guarantee convergence to some local optimum [BST14; WYZ19]. However, the quality of the results then highly depends on the possibility to propose an initialization point close enough to the solution. Additionally to the non-convexity issues, representation learning commonly includes non-smooth terms because of the constraints inherent to compact representations, such as sparsity, or the constraint imposed on the search space, such as non-negativity constraints. One possibility to deal with this second issue is to resort to advanced optimization tools such as proximal methods [CP11].

To avoid estimation problems related to non-convexity or non-smoothness, one possibility is to resort to Markov chain Monte Carlo (MCMC) methods [Per+12; Per+15; EDT11]. Contrary to optimization methods, these methods use a Bayesian modeling of the problem. Each estimated variable is assigned a prior distribution model and the main concept of MCMC algorithm is to generate samples according to the joint posterior distribution [RC04; Bro+11]. The Bayesian estimators of the parameters of interest can then be approximated using these samples. Besides, these samples can be used to provide a full description of the posterior distribution of interest, beyond a simple point estimation (*e.g.*, maximum a posteriori estimators). For example, it gives the possibility to provide confidence sets. Moreover, the convexity of the problem is not required to ensure convergence of the estimation. Nevertheless, one major drawback of these methods is that, even if the convergence is guaranteed, it is not possible to predict when convergence will be reached. MCMC methods thus allow users to deal with complex settings but fail in many cases to scale to real practical problems due to the extensive computational burden needed to get the results [Per+15].

Additionally to estimation problems, another recurrent issue is the difficulty to include exogenous data into a representation learning task [MBP12]. As discussed previously, supervised classification methods are nowadays considered the most efficient methods to extract information from data. It could be argued that this efficiency comes from the ability of these methods to incorporate the information coming from the examples provided by the user. Unfortunately it would be tedious to copy such a process to representation learning methods. The problem comes in particular from the difficulty to gather handmade exam-

ples. Most of the time it is impossible for experts to estimate the expected output from the image. In order to get a rich output and to benefit from exogenous data, a possibility is to consider the development of joint methods [Mai+09]. Such methods have the advantages to solve the problem of high-dimensional data for the classification by producing meaningful low-dimensional representations. Besides, some of the information contained in the classification training set is likely to be transfered to the representation learning problem and help solve it. The development of image analysis methods proposing a joint classification and representation learning approach is one of the key interest of this manuscript. For this reason it is interesting to get a closer look at the works which have already proposed in the literature to conduct classification and representation learning jointly.

Joint classification and representation learning – Many of the works on joint approaches have been published in the dictionary learning community, in which representation learning is actually referred to as dictionary learning [AEB06; RBE10]. In these approaches, it is usual to identify the subspace containing the observations by inferring a so-called dictionary. This dictionary is a collection of elementary vectors, referred to as atoms, spanning the representation subspace. The idea of supervised dictionary learning has been popularized in particular by the work of Mairal *et al.* [Mai+09; MBP12]. The core concept of supervised dictionary learning is to develop models in which the dictionary is built for a specific classification application. The dictionary should both demonstrate a reconstruction ability and a discriminative ability. The Discriminative K-SVD (DKSVD) described in [ZL10] proposes for instance to directly consider an optimization problem composed of a data fitting term and a linear classification term. Authors performed a face recognition task with a two-step algorithm including a training step to learn a relevant dictionary followed by an inference step to classify unknown samples using the learned representation. This work was implemented for the same task in [JLD11] with the difference that the learned dictionary promoted the use of different dictionary atoms for each class.

Going further, some works aimed at recovering class-specific dictionaries. These class-specific dictionaries are learned to ensure a good discrimination of the classes. To solve an object classification problem, the authors of [FRZ18] proposed for example to promote structural incoherence between the dictionaries of the various classes using an orthogonality penalization between the dictionary atoms. Further attempts were also made to exploit the training set of the classification task more thoroughly. For example in [CNT11], all the pixels of the training set were used as dictionary and a sparse representation of the unknown pixels was then inferred. Moreover, since dictionary atoms were associated with a

class label, the contribution of each class for the reconstruction of each pixel was computed using a reconstruction error metric. Finally, pixels were assigned to the class contributing the most to their reconstruction.

Broadly speaking, the idea of performing two complementary tasks simultaneously has already been investigated and has resulted into a family of models called cofactorization models [HDD13]. In particular, joint representation learning and classification can be cast as a cofactorization problem. Both tasks are interpreted as individual factorization problems and a coupling term and/or constraints between the dictionaries and coding matrices associated with the two problems are then introduced. These cofactorization-based models have proven to be highly efficient in many application fields, *e.g.*, for text mining [WB11], music source separation [Yoo+10], and image analysis [YYI12; AM18].

A common thread found in all the aforementioned works is the perspective chosen to tackle the problem of joint classification and representation learning. The main objective is generally to design a classification method and the representation learning process is only considered as a mean to this end [BCV13]. More specifically, representation learning is used to solve the statistical issues occurring with high-dimensional data. It operates as a dimensionality reduction method aiming at providing the best low-dimensional representation for the classification [LC09]. Such perspective is very likely to be detrimental to the representation learning process since it appears as secondary. Indeed, the discriminative and reconstruction abilities of the dictionary are often seen as adversarial in these models.

The work presented in this manuscript gathers new strategies to tackle the problem of joint approaches. The main objective is to provide truly cooperative joint representation learning and classification methods by considering a coherent hierarchical modeling using both methods. To illustrate the relevance of the methods proposed in this manuscript, an application to hyperspectral image analysis has been considered through the dual scope of spectral unmixing and classification.

Analysis of hyperspectral images – Hyperspectral images are particularly well-suited to be studied with representation learning methods due to the high dimension of the pixels of this specific modality of images. As a reminder, conventional color imaging has been designed in order to mimic human eye and, for this reason, these images are composed of three bands corresponding each to the reflectance measured for the blue, green and red wavelengths. However the spectral information contained in such an image is eventually very limited. Indeed, the reflectance, defined as the fraction of incident electromagnetic power that is reflected, varies for each wavelength depending on the electromagnetic prop-

erties of the observed scene [SM02; Lan02]. It is actually possible to measure this reflectance spectrum for hundreds of specific wavelength and thus obtain a very accurate electromagnetic characterization of the scene [Pla+09]. In the case of hyperspectral imaging, hundreds of measurements are performed in order to get a fine sampling of the reflectance spectrum of the area underlying each pixel. Moreover, measurements are not limited to the visible domain but usually include a larger part of the electromagnetic spectrum, *e.g.*, the infrared domain [Van+93].

The study of the electromagnetic properties of matter has shown that every material can actually be characterized by a specific reflectance spectrum [Hap93]. Unfortunately, due to the limited spatial resolution of the hyperspectral sensors, the area described by a given pixel usually includes a collection of materials. The result is the creation of mixels, *i.e.*, pixels representing a mixture of elementary reflectance spectra of pure material, usually referred to as endmembers, as shown in Figure .1. Spectral unmixing aims at identifying these endmembers and estimating the proportions of each pure material inside each pixel [Bio+12]. This method of interpretation actually fits in the family of representation learning methods where the learned representation subspace is the subspace spanned by the identified endmembers and the latent representation is the vector of proportions of pure materials, generally called abundance vector.

Spectral unmixing is widely-used to interpret hyperspectral images particularly because of the richness of the data that allows a physical interpretation of the results. Moreover, the high dimension and high redundancy of the hyperspectral pixels may make it difficult to perform a classification [Cam+14; Fau+13]. Therefore, it is often necessary to use dimensionality reduction methods prior to classification [ZD16; LFG17].

Bearing in mind the aim to propose joint representation learning and classification methods, reviewing the previous works that attempted to link both methods in the specific context of hyperspectral imaging appeared of great interest. The frameworks proposed in the literature for a joint use of spectral unmixing and classification are generally based on a sequential use of the two approaches. The most simple way to implement a sequential approach is to use spectral unmixing as a feature extraction method. The abundance vectors can be used as feature vectors for the classification which is then performed with the help of a conventional classifier, as done with SVM in [LC09; Vil11; Dóp+11; Dóp+12] or with a deep neural network in [Ala+17]. Spectral unmixing as feature extraction method presents the benefit of reducing drastically the dimension as well as proposing features with physical meaning. However, these features remain rather simple and do not maximize the separability of the classes. Additionally, spectral unmixing does not profit at all from any information

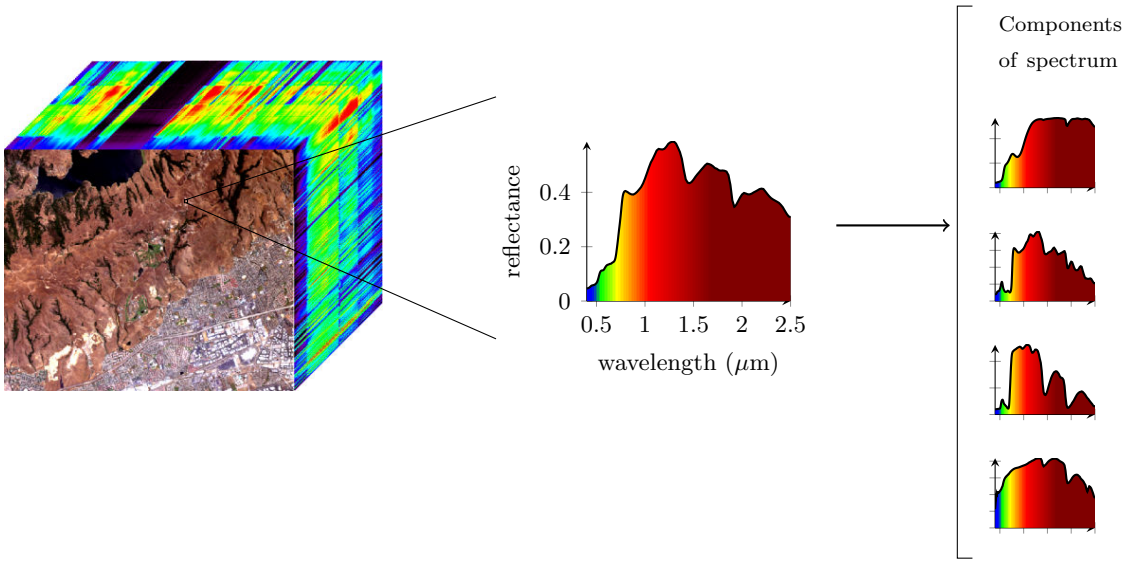


Figure .1.: Hyperspectral images are images with a fine spectral resolution. The measured reflectance spectrum of a pixel is explained as a mixture elementary components each representing a specific material.

coming from the classification. The classification map is actually the only considered result and spectral unmixing is only a tool to help the classification.

Spectral unmixing has also been used to improve classification results, more precisely to perform sub-pixel mapping [Vil+11b; Vil+11a]. The main idea is to identify mixed pixels, *i.e.*, pixels representing areas containing several classes, and then split these pixels to increase the spatial resolution. Spectral unmixing is used to assign classes to the newly created pixels. For example, if unmixing shows that a pixel contains 80% of vegetation and 20% of soil, 80% of the underlying new pixels are assigned to class *vegetation* and 20% to *soil*. A major limitation to these methods is that an endmember has to be equivalent to a class.

Nevertheless, several works used this assumption of equivalence between classes and endmembers. In the semi-supervised classification methods proposed in [Dóp+14; Li+15a], the spectral unmixing method was used directly as a classifier where the abundance vectors were directly interpreted as vectors collecting the probabilities to belong to each of the classes. Spectral unmixing was used side-by-side with a multinomial logistic regression (MLR). Besides, the two classifiers were used in an active learning method combining them to increase the size of the training set by generating labels to identified informative pixels.

Andrejchenko *et al.* [And+16] also exploited spectral unmixing as a classifier. This work,

focusing on classification problems with small training sets, introduced the idea of using all the pixels of the training set as endmembers. A sparse spectral unmixing method was then used to infer the abundance vectors. Finally, for each unlabeled pixel the predominant endmember was identified and its class was attributed the unlabeled pixel.

The same authors also proposed in [And+19] a classification method based on a decision fusion framework where the results of two classifiers were merged with the help of Markov or conditional random fields. The first classifier was a conventional MLR classifier and the second was similar to the one of [And+16] with the difference that fractional abundances were computed by summing all the abundances of endmembers of a same class yielding a probability vector to belong to each of the classes.

Another family of methods makes the link between classification and spectral unmixing by assuming that all the pixels of a given class live in a class-specific subspace. In particular, the early work [LBP12] proposed a segmentation method combining projection in class-specific subspaces with a MLR algorithm. From an unmixing point of view, this assumption also means that it is possible to use class-specific endmember matrices. Authors of [Sun+17] thus proposed to use a training set to estimate an endmember matrix for each class, then to concatenate all these endmember matrices to get a global endmember matrix and finally to use a sparse spectral unmixing method and classification method based on fractional abundances to get both unmixing results and classification results. The idea developed in these two latter works were combined in [Xu+19] in which the authors proposed to evaluate class-specific endmember matrices and the identified subspaces were then used to create a transformation function applied to the data, then used to feed a MLR algorithm.

This brief overview shows that very few attempts have been conducted to propose a joint spectral unmixing and classification method. Moreover, these methods generally tackle the problem by using the two approaches sequentially and, in most cases, with the final idea to get an improved classification method. Convinced that the representation learning results are as worthy of consideration as classification results for an end-user, the work presented in this manuscript is an attempt to propose truly joint representation learning and classification methods. The aim of these methods is to provide a hierarchical description of the considered data.

The work presented in this manuscript has been carried out within the Signal and Communications group of the Institut de Recherche en Informatique de Toulouse. This thesis was funded by the Centre National d'Études Spatiales (CNES) and Région Occitanie.

Structure of the manuscript

Chapter 1 introduces a hierarchical Bayesian model, inspired by [EDT11], to jointly perform low-level modeling and supervised classification. The low-level modeling intends to extract the latent structure of the data whereas classification is considered as a high-level modeling. These two stages of the hierarchical model are linked through a clustering stage which aims at identifying groups of pixels with similar latent representation. A Markov random field (MRF) is then used to ensure a spatial regularity of the cluster labels and a coherence with classification labels. The final stage used for classification exploits a set of possibly corrupted labeled data provided by the end-user. The parameters of the overall Bayesian model are estimated using a Markov chain Monte Carlo (MCMC) algorithm in the specific case where the image is an hyperspectral image and the low-level modeling is a spectral unmixing model.

Chapter 2 considers a different approach by using of a cofactorization model. The representation learning task and the classification task are both modeled as factorization matrix problems. A coupling term is then introduced to enable a joint estimation. Based on the same idea developed in model of Chapter 1, the coupling term is interpreted as a clustering task performed on the low-dimensional representation vectors. Finally, the cluster attribution vectors are used as features vectors for classification. The overall non-smooth, non-convex optimization problem is solved using a proximal alternating linearized minimization (PALM) algorithm ensuring convergence to a critical point of the objective function. The quality of the obtained results is finally assessed on synthetic and real data for the analysis of hyperspectral image using spectral unmixing and classification.

Chapter 3 intends to enrich the previous model by adding spatial information. In the previous models, the spatial information is only exploited through regularization terms such as Potts-MRF or total variation regularization. With this mechanism, spatial information is introduced at a late stage in an indirect manner. To introduce a more direct spatial information, a cofactorization model with two data fitting terms is considered. The first term is a spectral mixture model based on the hyperspectral image and thus accounts for the spectral information. The second term is a representation learning model based on an image aggregating the spatial information. This image can be computed from a panchromatic image, *e.g.*, by extracting spatial features or by concatenating the neighborhood of each pixel. The coupling term is again a clustering task identifying groups of pixels with similar spectral and spatial signatures. The resulting model performs an unsupervised unmixing

task and could be merged with the model of Chapter 2 to derive a richer supervised model.

Main contributions

Chapter 1. The main contribution of this chapter lies in the introduction of a Bayesian framework to unify representation learning and classification approaches. The model proposes a resourceful use of MRF to link all the levels of the model to conduct a joint estimation. The second contribution is the design of a classification method robust to labeling errors in the training set. The method additionally proposes a correction of erroneous labels. Finally, the last contribution is in the potential of interpretation of the results due to meaningful byproducts. In particular, a matrix decomposing the classes into a collection of clusters is estimated and each of these clusters are characterized by their mean abundance vector. These byproducts allow the user to clearly visualize the structure of the considered data.

Chapter 2. A cofactorization model is used to develop another unified framework for representation learning and classification. This model differs from other cofactorization model mainly by the proposed coupling term. Firstly, it allows a rich interpretation of the results with again the idea of decomposing classes in a collection of clusters. And secondly, it keeps flexibility between the two tasks at hand, contrary to previous models such as DKSVD [ZL10] where the model introduces two adversarial goals instead of cooperative ones. The final contribution lies in the proposition of a powerful optimization method dedicated to the criterion to be minimized. Indeed, a proximal alternating linearized minimization algorithm (PALM) is used to solve the non-convex, non-smooth problem at hand with guarantee of convergence to a critical point of the objective function.

Chapter 3. The main contribution of this chapter is a new proposition to enrich spectral mixture model with spatial information directly using an additional data fitting term instead of resorting to regularization methods. This new model tends to improve the results of the unmixing process. But more importantly, the model produces a segmentation map identifying several areas by their spectral signature and their spatial pattern. We actually obtain a very compact, complete and visual representation of the analyzed scene. Up to our knowledge, this method introduces the new concept of joint spatial-spectral unmixing.

List of publications

Submitted

[Lag+19c] A. Lagrange, M. Fauvel, S. May, J. Bioucas-Dias, and N. Dobigeon. "Matrix Cofactorization for Joint Representation Learning and Supervised Classification – Application to Hyperspectral Image Analysis". In: *arXiv:1902.02597 [cs, eess]* (Feb. 2019). arXiv: 1902.02597 [cs, eess] (cit. on p. 49).

[Lag+19e] A. Lagrange, M. Fauvel, S. May, and N. Dobigeon. "Matrix Cofactorization for Joint Spatial-Spectral Unmixing of Hyperspectral Images". In: *arXiv:1907.08511 [cs, eess]* (July 2019). arXiv: 1907.08511 [cs, eess] (cit. on p. 83).

International journals

[Lag+19d] A. Lagrange, M. Fauvel, S. May, and N. Dobigeon. "Hierarchical Bayesian Image Analysis: From Low-Level Modeling to Robust Supervised Learning". In: *Patt. Recognition* 85 (2019), pp. 26–36 (cit. on p. 19).

International conferences

[Lag+18] A. Lagrange, M. Fauvel, S. May, and N. Dobigeon. "A Bayesian Model for Joint Unmixing and Robust Classification of Hyperspectral Images". In: *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2018, pp. 3399–3403 (cit. on pp. 19, 55).

[Lag+19b] A. Lagrange, M. Fauvel, S. May, J. M. Bioucas-Dias, and N. Dobigeon. "Matrix Cofactorization for Joint Unmixing and Classification of Hyperspectral Images". In: *Proc. European Signal Process. Conf. (EUSIPCO)*. Sept. 2019 (cit. on p. 49).

National conferences

[Lag+17] A. Lagrange, M. Fauvel, S. May, and N. Dobigeon. “Un Modèle Bayésien Pour Le Démélange, La Segmentation et La Classification Robuste d’images Hyperspectrales”. In: *Actes du Colloque GRETSI*. 2017, pp. 1–4 (cit. on p. 19).

[Lag+19a] A. Lagrange, M. Fauvel, S. May, J. M. Bioucas-Dias, and N. Dobigeon. “Cofactorisation de Matrices Pour Le Démélange et La Classification Conjoints d’Images Hyperspectrales”. In: *Actes du Colloque GRETSI*. Aug. 2019 (cit. on p. 49).

Publications prior to the Ph.D. work

[Cam+16] M. Campos-Taberner, A. Romero-Soriano, C. Gatta, G. Camps-Valls, A. Lagrange, B. Le Saux, A. Beaupère, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, M. Ferecatu, M. Shimoni, G. Moser, and D. Tuia. “Processing of Extremely High-Resolution LiDAR and RGB Data: Outcome of the 2015 IEEE GRSS Data Fusion Contest - Part A: 2-D Contest”. In: *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.* 9.12 (Dec. 2016), pp. 5547–5559.

[Lag+15] A. Lagrange, B. Le Saux, A. Beaupere, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, and M. Ferecatu. “Benchmarking Classification of Earth-Observation Data: From Learning Explicit Features to Convolutional Networks”. In: *Proc. IEEE Int. Conf. Geosci. Remote Sens. (IGARSS)*. IEEE, 2015, pp. 4173–4176.

[LFG17] A. Lagrange, M. Fauvel, and M. Grizonnet. “Large-scale feature selection with Gaussian mixture models for the classification of high dimensional remote sens. images”. In: *IEEE Trans. Comput. Imag.* 3.2 (2017), pp. 230–242 (cit. on pp. 11, 74).

Chapter 1.

Hierarchical Bayesian model for joint classification and spectral unmixing

This chapter has been adapted from the journal paper [Lag+19d]. This work has also been discussed in the conference papers [Lag+18; Lag+17].

Contents

1.1. Introduction (in French)

Dans le contexte de l'interprétation d'images, de nombreuses méthodes ont été développées pour extraire l'information utile. Parmi ces méthodes, les modèles génératifs ont reçu une attention particulière du fait de leurs solides bases théoriques, mais aussi de la facilité d'interprétation des modèles estimés en comparaison des modèle discriminatifs, comme les réseaux de neurones profonds. Ces méthodes sont basées sur une modélisation statistique explicite des données. Ils permettent la construction de modèles dédiés pour chaque application [WG13], ou bien la construction de modèles plus génériques comme les modèles de mélange de gaussiennes pour la classification [Ker14]. L'utilisation de modèles spécialisés ou génériques représente deux approches différentes pour obtenir une description interprétable des données. Par exemple, lorsqu'on analyse des images, les modèles spécialisés visent à reconstituer la structure latente (potentiellement basée sur un modèle physique) de chacune des mesures pixeliques [DTC08] tandis que la classification produit une information haut-niveau réduisant la caractérisation des pixels à un unique label [FCB12].

La principale contribution de ce chapitre réside dans la définition d'un nouveau modèle bayésien développant un cadre unifié pour réaliser classification et modélisation des structures latentes de manière jointe. Ce modèle a l'avantage d'estimer des descriptions bas-niveau et haut-niveau cohérentes de l'image en réalisant une analyse hiérarchique de l'image. De plus, il est possible d'espérer une amélioration des résultats de chacune des méthodes grâce à la complémentarité des approches. En particulier, l'utilisation de données labellisées n'est plus limitée à l'analyse haut-niveau, *i.e.*, la classification. Il est également possible d'informer l'analyse bas-niveau, c'est-à-dire, la modélisation des structures latentes, qui profite en général mal de telles informations a priori. D'autre part, les variables latentes de la modélisation bas-niveau peuvent être utilisées comme descripteurs pour la classification. Un effet collatéral direct est la réduction de dimension explicite réalisée sur les données avant la classification [JL98]. Enfin, le modèle hiérarchique introduit permet de rendre la classification robuste à la corruption des labels d'entraînement. En effet, les performances d'une méthode de classification supervisée peuvent se dégrader si ces derniers ne sont pas entièrement fiables comme c'est souvent le cas puisque ces labels sont estimés par des experts humains pouvant commettre des erreurs. Pour cette raison, le problème de développer des méthodes de classification robustes aux erreurs de labellisation a été largement considéré dans la communauté [BG09 ; Pel+17]. S'inscrivant dans ce cadre, le modèle proposé tient

explicitement compte de la présence de labels corrompus.

L'interaction entre les modèles bas-niveau et haut-niveau est géré par l'utilisation de champs de Markov aléatoires (MRF) non-homogènes [Li09]. Les MRFs sont des modèles probabilistes largement utilisés pour décrire des interactions spatiales. C'est pourquoi, lorsqu'ils sont utilisés comme a priori dans une modélisation bayésienne, ils sont tout à fait adaptés pour capturer les dépendances spatiales entre les structures latentes des images [ZBS01 ; Tar+10 ; And+19 ; Che+17]. Le modèle proposé inclut lui deux instances de MRFs assurant (i) la cohérence entre les modélisations bas-niveau et haut-niveau, (ii) la cohérence avec les labels fournis par les experts comme donnée d'entraînement et (iii) une régularité spatiale.

La suite de ce chapitre est organisée de la manière suivante. La Section 1.3 présente le modèle bayésien hiérarchique proposé comme cadre unifié pour l'interprétation bas-niveau et haut-niveau d'images. Une méthode de Monte Carlo par chaîne de Markov est explicitée dans la Section 1.4 pour permettre l'échantillonnage selon la loi postérieur jointe des paramètres du modèle. Ensuite, une instance particulière du modèle est considérée dans la Section 1.5 où, en se recentrant sur le cas d'étude de ce manuscrit, des images hyperspectrales sont analysées à la fois du point de vue du démélange et de la classification. La Section 1.6 présente les résultats obtenus avec la méthode proposée et les compare à ceux obtenus avec des méthodes établies en utilisant des données synthétiques puis réelles. Finalement, la Section 1.7 conclut ce chapitre et ouvre quelques perspectives de recherche dans la suite de ce travail.

1.2. 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 [WG13], or either more general models to be implemented to solve generic tasks, such as Gaussian mixture models for classification [Ker14]. Task-specific and classification-like models are two different ways to reach an interpretable description of the data with respect to a particular applicative issue. For instance, when analyzing images, task-specific models aim at recovering the latent (possibly physics-based) structures underlying each pixel-wise measurement [DTC08] while classification provides a high-level information, reducing the pixel characterization to a unique

label [FCB12].

The contribution of this chapter lies in the derivation of a unified Bayesian framework able to perform classification and latent structure modeling jointly. 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. In particular, 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 [JL98]. 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 [BG09; Pel+17]. 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) [Li09]. 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 [ZBS01; Tar+10; And+19]. For example, Chen *et al.* [Che+17] proposed to use MRFs to perform clustering. The proposed framework incorporates two instances of MRF, ensuring (i) consistency between the low and high level modeling, (ii) consistency with external data available as prior knowledge and (iii) a more classical spatial regularization.

The remaining of the chapter is organized as follows. Section 1.3 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 1.4 to sample according to the joint posterior distribution of the resulting model parameters. Then, focusing on the problem at hand in this manuscript, a particular and illustrative instance of the proposed framework is presented in Section 1.5 where hyperspectral images are analyzed under the dual scope of unmixing and classification. Section 1.6 presents the results obtained

with the proposed method and compares them to the results of well-established methods using synthetic and real data. Finally, Section 1.7 concludes the chapter and opens some research perspectives to this work.

1.3. Hierarchical 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 1.3.1, 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 1.3.2. Finally, the last stage of the model, explained in Section 1.3.3, 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 Figure 1.1.

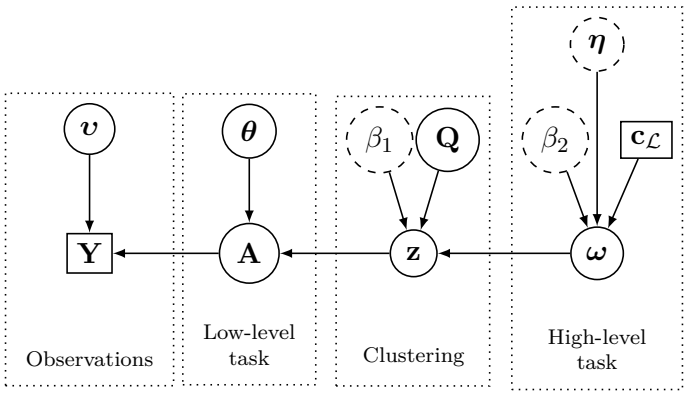


Figure 1.1.: Directed acyclic graph of the proposed hierarchical Bayesian model. (User-defined parameters appear in dotted circles and external data in squares).

1.3.1. 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.*, $R \leq d$. 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

$$\mathbf{Y}|\mathbf{A}, \boldsymbol{\upsilon} \sim \Psi\left(\mathbf{Y}; f_{\text{lat}}\left(\mathbf{A}\right), \boldsymbol{\upsilon}\right), \tag{1.1}$$

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

$$\mathbf{Y}|\mathbf{A}, \boldsymbol{\upsilon} \sim \prod_{p=1}^{P} \Psi\left(\mathbf{y}_p; f_{\text{lat}}\left(\mathbf{a}_p\right), \boldsymbol{\upsilon}\right). \tag{1.2}$$

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 $f_{\text{lat}}(\cdot)$ 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) [FCB06]. In such case, the model (1.1) should be interpreted as a probabilistic counterpart of PCA [TB99] 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 [CNJ09]. As a typical illustration, by assuming an additive white and centered Gaussian statistical model $\Psi(\cdot)$ and a linear latent function $f_{\text{lat}}(\cdot)$, the generic model (1.2)

can be particularly instanced as

$$\mathbf{Y}|\mathbf{A}, s^2 \sim \prod_{p=1}^{P} \mathcal{N}\left(\mathbf{y}_p; \mathbf{M}\mathbf{a}_p, s^2\mathbf{I}_d\right) \tag{1.3}$$

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 [CJ11], provided that

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

From a different perspective, the generic formulation of the statistical latent structure (1.2) can also result from a thorough analysis of more complex physical processes underlying observed measurements, resulting in specific yet richer physics-based latent models [Per+12; Alb+14]. For sake of generality, this latent structure will not be specified in the rest of this manuscript, except in Section 1.5 where the linear Gaussian model (1.3) will be more deeply investigated as an illustration in a particular applicative context.

1.3.2. 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

$$\mathbf{a}_p|z_p = k, \boldsymbol{\theta}_k \sim \Phi(\mathbf{a}_p; \boldsymbol{\theta}_k), \tag{1.4}$$

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 $\boldsymbol{\theta}_k = \{\boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k\}$, would lead

to a conventional Gaussian mixture model (GMM) for the latent structure, as in [EDT11] (see Section 1.5).

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 parametrized 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 1.3.3. Pursuing the objective of analyzing images, the second potential is associated with a Potts-MRF [Wu82] 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

$$\mathbb{P}[\mathbf{z}|\boldsymbol{\omega}, \mathbf{Q}] = \frac{1}{C(\boldsymbol{\omega}, \mathbf{Q})} \exp\left(\sum_{p\in\mathcal{P}} V_1(z_p, \omega_p, q_{z_p,\omega_p}) + \sum_{p\in\mathcal{P}} \sum_{p'\in\mathcal{V}(p)} V_2(z_p, z_{p'})\right) \tag{1.5}$$

where $\mathcal{V}(p)$ stands for the neighborhood of p, $q_{k,j}$ is the k-th element of the j-th 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

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

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 [Li09]

$$\begin{aligned} C(\boldsymbol{\omega}, \mathbf{Q}) &= \sum_{\mathbf{z}\in\mathcal{K}^P} \exp\left(\sum_{p\in\mathcal{P}} V_1(z_p, \omega_p, q_{z_p,\omega_p}) + \sum_{p\in\mathcal{P}} \sum_{p'\in\mathcal{V}(p)} V_2(z_p, z_{p'})\right) \\ &= \sum_{\mathbf{z}\in\mathcal{K}^P} \prod_{p\in\mathcal{P}} q_{z_p,\omega_p} \exp\left(\beta_1 \sum_{p'\in\mathcal{V}(p)} \delta(z_p, z_{p'})\right) \end{aligned} \tag{1.6}$$

The equivalence between Gibbs random fields and MRF stated by the Hammersley-Clifford theorem [Li09] provides the prior probability of a particular cluster label condi-

tionally upon its neighbors

$$\mathrm{P}[z_p = k|\mathbf{z}_{\mathcal{V}(p)}, \omega_p = j, q_{k,\omega_p}] \propto \exp\left(V_1(k, j, q_{k,j}) + \sum_{p' \in \mathcal{V}(p)} V_2(k, z_{p'})\right) \tag{1.7}$$

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 inter-dependency, 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.*,

$$\mathbf{q}_j \sim \mathrm{Dir}(\mathbf{q}_j; \zeta_1, \ldots, \zeta_K). \tag{1.8}$$

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 1.4. 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.

1.3.3. 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 chapter, 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 1.3.2), 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 intends 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

$$\mathbb{P}[\omega_p = j | \boldsymbol{\omega}_{\mathcal{V}(p)}, c_p, \eta_p] \propto \exp\left(W_1(j, c_p, \eta_p) + \sum_{p' \in \mathcal{V}(p)} W_2(j, \omega_{p'}) \right). \tag{1.9}$$

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.

1.4. 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

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

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 $\mathbf{A}$, $\boldsymbol{\Theta}$ and $\mathbf{Q}$ can be approximated through empirical averages

$$\hat{\mathbf{x}}_{\text{MMSE}} = \mathbb{E}[\mathbf{x}|\mathbf{Y}] \approx \frac{1}{N_{\text{MC}}} \sum_{t=1}^{N_{\text{MC}}} \mathbf{x}^{(t)} \tag{1.10}$$

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

$$\hat{\mathbf{x}}_{\mathrm{MAP}} = \underset{\mathbf{x}}{\operatorname{argmax}}\, p(\mathbf{x}|\mathbf{Y}) \approx \underset{\mathbf{x}^{(t)}}{\operatorname{argmax}}\, p(\mathbf{x}^{(t)}|\mathbf{Y}) \tag{1.11}$$

which basically amounts at retaining the most frequently generated label for these specific discrete parameters [Kai+12].

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.

1.4.1. Latent parameters

Given the likelihood function resulting from the statistical model (1.2) and the prior distribution in (1.4), the conditional posterior distribution of a latent vectors can be expressed as follows

$$\begin{aligned} p(\mathbf{a}_p|\mathbf{y}_p, \boldsymbol{\upsilon}, z_p = k, \boldsymbol{\theta}_k) &\propto p(y_p|\mathbf{a}_p, \boldsymbol{\upsilon}) p(\mathbf{a}_p|z_p = k, \boldsymbol{\theta}_k) \\ &\propto \Psi\left(\mathbf{y}_p; f_{\mathrm{lat}}\left(\mathbf{a}_p\right), \boldsymbol{\upsilon}\right) \Phi(\mathbf{a}_p; \boldsymbol{\theta}_k). \end{aligned} \tag{1.12}$$

1.4.2. 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 $\mathcal{K}$

$$\begin{aligned} \mathbb{P}(z_p = k|\boldsymbol{\theta}_k, \omega_p = j, q_{k,j}) &\propto p(\mathbf{a}_p|z_p = k, \boldsymbol{\theta}_k) \mathbb{P}(z_p = k|\mathbf{z}_{\mathcal{V}(p)}, \omega_p = j, q_{k,j}) \\ &\propto \Phi(\mathbf{a}_p; \boldsymbol{\theta}_k) q_{k,j} \exp\left(\beta_1 \sum_{p' \in \mathcal{V}(p)} \delta(k, z'_p)\right). \end{aligned} \tag{1.13}$$

1.4.3. 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

$$\begin{aligned} p(\mathbf{q}_j|\mathbf{z}, \mathbf{Q}_{\setminus j}, \boldsymbol{\omega}) &\propto p(\mathbf{q}_j)\mathbb{P}(\mathbf{z}|\mathbf{Q}, \boldsymbol{\omega}) \\ &\propto \frac{\prod_{k=1}^{K} q_{k,j}^{n_{k,j}}}{C(\boldsymbol{\omega}, \mathbf{Q})} \mathbb{1}_{\mathbb{S}_K}(\mathbf{q}_j). \end{aligned} \tag{1.14}$$

where $\mathbf{Q}_{\setminus j}$ 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 [Ris+10]. As alternatives, one could use to likelihood-free Metropolis Hastings algorithm [Per+13], auxiliary variables [Mol+06] or pseudo-likelihood estimators [Bes75]. 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

$$\mathbf{q}_j|\mathbf{z}, \boldsymbol{\omega} \sim \mathrm{Dir}(\mathbf{q}_j; n_{1,j} + 1, \ldots, n_{K,j} + 1), \tag{1.15}$$

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

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

which is a biased empirical estimator of $\mathrm{P}\left[z_p = k|\omega_p = j\right]$. 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 (1.13) with $\beta_1 > 0$ while the columns of the interaction matrix are sampled according to (1.15). 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 (1.15) reduces to (1.14) and the iteraction matrix is properly sampled according to its exact conditional posterior distribution.

1.4.4. 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

$$\begin{aligned}\mathbb{P}[\omega_p = j|\mathbf{z}, \boldsymbol{\omega}_{\setminus p}, \mathbf{q}_j, c_p, \eta_p] &\propto \mathbb{P}[z_p|\omega_p = j, \mathbf{q}_j, \mathbf{z}_{\nu(p)}]\mathbb{P}[\omega_p = j|\boldsymbol{\omega}_{\mathcal{V}(p)}, c_p, \eta_p] \\ &\propto \frac{q_{z_p,j}\pi_j \exp\left(\beta_2 \sum_{p' \in \nu(p)} \delta(j, \omega_{p'})\right)}{\sum_{k'=1}^{K} q_{k',j} \exp\left(\beta_1 \sum_{p' \in \nu(p)} \delta(k', z_{p'})\right)},\end{aligned} \tag{1.16}$$

where $\boldsymbol{\omega}_{\setminus p}$ 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

[1] This strategy can also be interpreted as choosing $C(\boldsymbol{\omega}, \mathbf{Q}) \times \mathrm{Dir}(\mathbf{1})$ instead of the Dirichlet distribution (1.8) as prior for $\mathbf{q}_j$.

conditional posterior probability reads

$$\mathbb{P}[\omega_p = j|\mathbf{z}, \boldsymbol{\omega}_{\backslash p}, \mathbf{q}_j, c_p, \eta_p] \propto \mathbb{P}[z_p|\omega_p = j, \mathbf{q}_j, \mathbf{z}_{\nu(p)}]\mathbb{P}[\omega_p = j|\boldsymbol{\omega}_{\mathcal{V}(p)}, c_p, \eta_p]$$
$$\propto \begin{cases} \dfrac{q_{z_p,j}\eta_p \exp\left(\beta_2 \sum\limits_{p'\in\nu(p)} \delta(j,\omega_{p'})\right)}{\sum\limits_{k'=1}^{K} q_{k',j} \exp\left(\beta_1 \sum\limits_{p'\in\nu(p)} \delta(k',z_{p'})\right)} & \text{when } \omega_p = c_p \\ \dfrac{(1-\eta_p)q_{z_p,j} \exp\left(\beta_2 \sum\limits_{p'\in\nu(p)} \delta(j,\omega_{p'})\right)}{(C-1)\sum\limits_{k'=1}^{K} q_{k',j} \exp\left(\beta_1 \sum\limits_{p'\in\nu(p)} \delta(k',z_{p'})\right)} & \text{otherwise} \end{cases} \tag{1.17}$$

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 $\sum\limits_{k'=1}^{K} q_{k',j} \exp\left(\beta_1 \sum\limits_{p'\in\nu(p)} \delta(k', z_{p'})\right) = 1$ in this specific case.

1.5. 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 [Cam+14; Man+14; Ma+14]. 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 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 [Bio+12]. Per se, spectral unmixing can be cast as a blind source separation

or a nonnegative matrix factorization (NMF) task [Ma+13]. The particularity of spectral unmixing, also known as spectral mixture analysis in the microscopy literature [DB12], 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 [Dóp+14; Vil+11b]. 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 1.3. 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 1.3.1 and 1.3.2, are specified in what follows, while the classification task is directly implemented as in Section 1.3.3.

1.5.1. Low-level model

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.*,

$$\mathbf{y}_p = \sum_{r=1}^{R} a_{r,p}\mathbf{m}_r + \mathbf{e}_p \tag{1.18}$$

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

$$\mathbf{Y} = \mathbf{M}\mathbf{A} + \mathbf{E} \tag{1.19}$$

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 [Bio+12]. Under this assumption, the LMM matrix formulation defined by (1.19) can be straightforwardly interpreted as a partic-

ular instance of the low-level interpretation (1.1) 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

$$s^2|\delta \sim \mathcal{IG}(s^2; 1, \delta), \quad \delta \propto \frac{1}{\delta}\mathbb{1}_{\mathbb{R}^+}(\delta). \tag{1.20}$$

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

$$s^2|\mathbf{Y}, \mathbf{A} \sim \mathcal{IG}\left(s^2; 1 + \frac{Pd}{2}, \frac{1}{2}\sum_{p=1}^{P} \|\mathbf{y}_p - \mathbf{M}\mathbf{a}_p\|^2\right) \tag{1.21}$$

which is easy to sample from, as an additional step within the Gibbs sampling scheme described in Section 1.4.

1.5.2. Clustering

In the current problem, the latent modeling $\Phi(\cdot; \cdot)$ in (1.4) is chosen as Gaussian distributions elected for the latent vectors $\mathbf{a}_p$ ($p \in \mathcal{P}$),

$$\mathbf{a}_p|z_p = k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k \sim \mathcal{N}(\mathbf{a}_p; \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) \tag{1.22}$$

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

$$p(\mathbf{a}_p|z_p = k, \boldsymbol{y}_p, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) \propto |\boldsymbol{\Lambda}_k|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{a}_p - \boldsymbol{\mu}_{k,p})^t \boldsymbol{\Lambda}_k^{-1}(\mathbf{a}_p - \boldsymbol{\mu}_{k,p})\right) \tag{1.23}$$

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

$$\sigma_{k,r}^2 \sim \mathcal{IG}(\sigma_{k,r}^2; \xi, \gamma) \tag{1.24}$$

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

$$\sigma_{r,k}^2|\mathbf{A}, \mathbf{z}, \psi_{r,k} \sim \mathcal{IG}\left(\sigma_{k,r}^2; \frac{n_k}{2} + \xi, \gamma + \sum_{p\in\mathcal{I}_k} \frac{(a_{r,p} - \psi_{r,k})^2}{2}\right) \tag{1.25}$$

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 $\mathrm{Dir}(\mathbf{1})$. 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 $\mathrm{E}[\mathbf{a}_p|z_p = k] = \boldsymbol{\psi}_k$. The resulting conditional posterior distribution of the mean vector $\boldsymbol{\psi}_k$ is the following multivariate Gaussian distribution

$$\boldsymbol{\psi}_k|\mathbf{A}, \mathbf{z}, \boldsymbol{\Sigma}_k \sim \mathcal{N}_{\mathbb{S}_R}\left(\boldsymbol{\psi}_k; \frac{1}{n_k}\sum_{p\in\mathcal{I}_k} \mathbf{a}_p, \frac{1}{n_k}\boldsymbol{\Sigma}_k\right) \tag{1.26}$$

truncated on the probability simplex

$$\mathbb{S}_R = \left\{\mathbf{x} = [x_1, \ldots, x_R]^T | \forall r,\ x_r \geq 0 \text{ and } \sum_{r=1}^{R} x_r = 1\right\}. \tag{1.27}$$

Sampling according to this truncated Gaussian distribution can be achieved following the strategies described in [AMD14].

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.

Algorithm 1: Inference using Gibbs sampling

Initialize all variables;
for $N_{\text{MC}} + N_{\text{burn}}$ *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 (1.13);
  foreach $j \in \mathcal{J}$ **do** sample $\mathbf{q}_j$ from $\text{Dir}(n_{1,j}+1, \ldots, n_{K,j}+1)$;
  foreach $p \in \mathcal{P}$ **do** sample ω_p from (1.16) and (1.17);
  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 $r \in \{1, \ldots, R\}$ **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 δ from $\mathcal{IG}(1, s^2)$;
  if *iteration* $> N_{\text{burn}}$ **then**
    update MMSE and MAP estimators
  end
end

1.6. Experiments

1.6.1. 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 Figure 1.2, 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 Figure 1.2. 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 Figure 1.2. 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= 30dB. The real interaction matrix $\mathbf{Q}$ presented in Figure 1.2 (e) and (f) summarized the data structure by providing the probability to be in a given cluster when belonging to a given class.

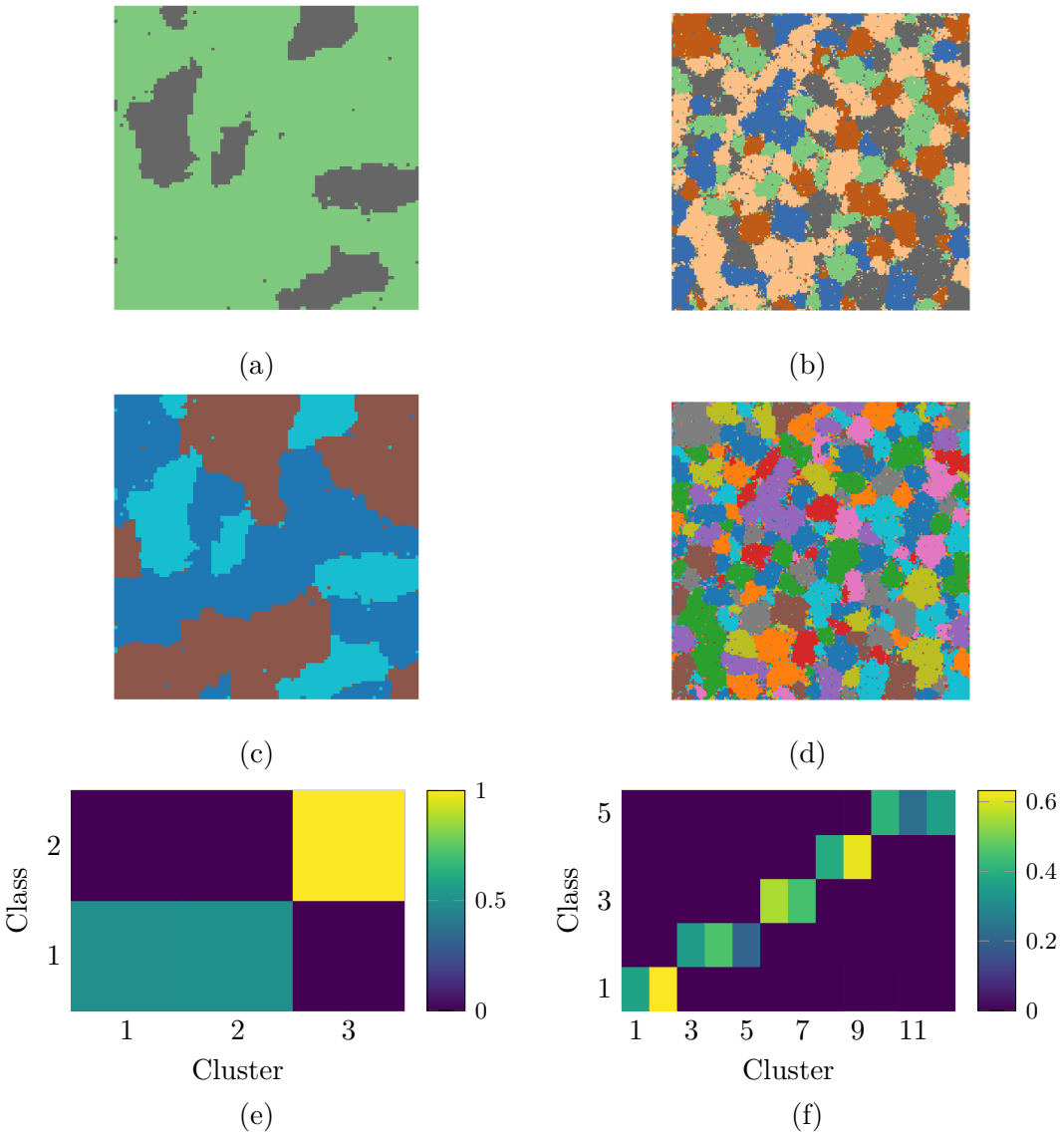


Figure 1.2.: Synthetic data. Classification maps of Image 1 (a) and Image 2 (b), corresponding clustering maps of Image 1 (c) and Image 2 (d), corresponding interaction matrix $\mathbf{Q}$ of Image 1 (e) and Image 2 (f).

Figure 1.3 represents 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.

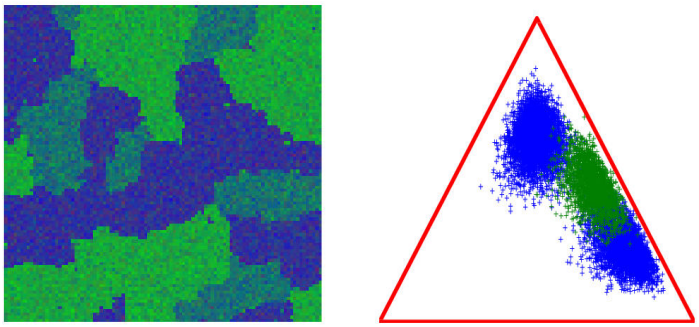


Figure 1.3.: Image 1. Left: colored composition of abundance map. Right: abundance vectors in the probabilistic simplex (red triangle) with Class 1 (blue) and Class 2 (green).

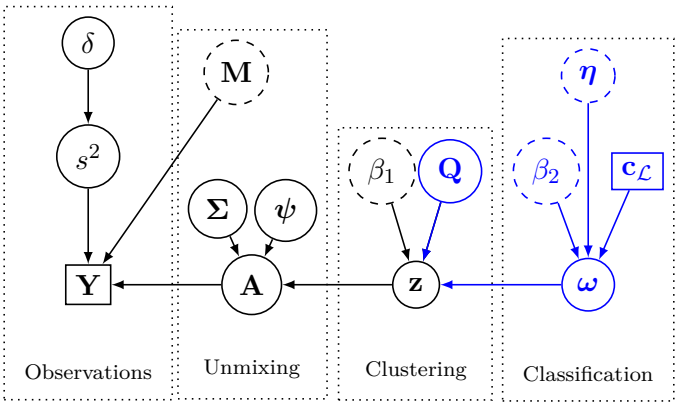


Figure 1.4.: Directed acyclic graph of the proposed model in the described hyperspectral framework. Part in blue is the extension made to the Eches model.

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 [EDT11] (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. Figure 1.4 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

Table 1.1.: Unmixing and classification results for all datasets.

		RMSE($\mathbf{A}$)	Kappa	Time (s)
Image 1	Proposed model	3.23×10^{-3} ($\pm 1.6 \times 10^{-5}$)	0.932 ($\pm$0.018)	171 ($\pm$5.4)
	Eches model	3.24×10^{-3} ($\pm 1.4 \times 10^{-5}$)	0.909 ($\pm$0.012)	146 ($\pm$0.7)
Image 2	Proposed model	1.62×10^{-2} ($\pm 1.62 \times 10^{-4}$)	0.961 ($\pm$0.04)	950 ($\pm$11)
	Eches model	1.61×10^{-2} ($\pm 2.71 \times 10^{-5}$)	0.995 ($\pm$0.0004)	676 ($\pm$2.1)
MUESLI image	Proposed model	N\ A	0.837 ($\pm 5 \times 10^{-3}$)	7175 ($\pm$102)
	Random Forest	N\ A	0.879 ($\pm 5 \times 10^{-4}$)	34 ($\pm$1.3)
	Gaussian model	N\ A	0.818 ($\pm 8.7 \times 10^{-5}$)	4 ($\pm$0.01)

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 [BG09]. The performance of the spectral unmixing task has been evaluated using the root global mean square error (RMSE) associated with the abundance estimation

$$\text{RMSE}(\mathbf{A}) = \sqrt{\frac{1}{PR} \left\| \hat{\mathbf{A}} - \mathbf{A} \right\|_F^2} \tag{1.28}$$

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. Details about evaluation metrics are available in Appendix A. Results reported in Table 1.1 show that the obtained RMSE 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) [HT96b] 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 Figure 1.5 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).

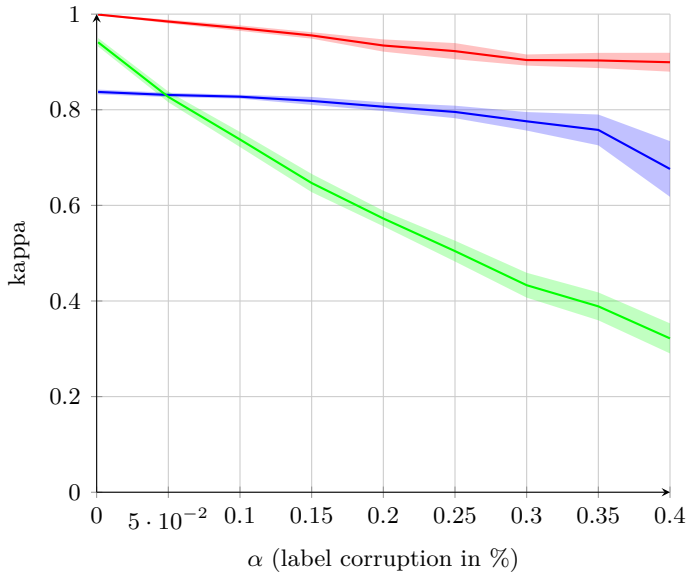


Figure 1.5.: Classification accuracy measured with Cohen's kappa as a function of the percentage of label corruption α: proposed model (red), MDA with abundance vectors (blue) and MDA with measured reflectance (green). Shaded areas denote the intervals corresponding to the standard deviation computed over 20 trials.

Moreover, as already explained, another advantage of the proposed model is the interesting by-products provided by the method. As an illustration, Figure 1.6 presents the interactions matrices $\mathbf{Q}$ 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 Figure 1.2 (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 [BF10]. 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 Figure 1.7 has been chosen in order to be highly correlated (4 vegetation spectra and 2 soils spectra). Moreover the endmembers matrix $\mathbf{M}$ has been augmented by 9 endmembers not present in the image. The obtained data is indeed both realistic and challenging in

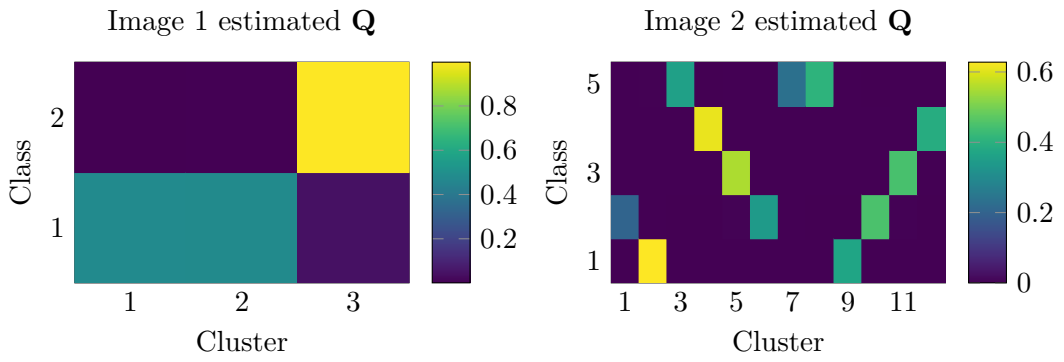


Figure 1.6.: Estimated interaction matrix $\mathbf{Q}$ for Image 1 (left) and Image 2 (right).

term of unmixing. A panchromatic view of the resulting image, made by summing all spectral bands, is presented in Figure 1.8 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= 10dB.

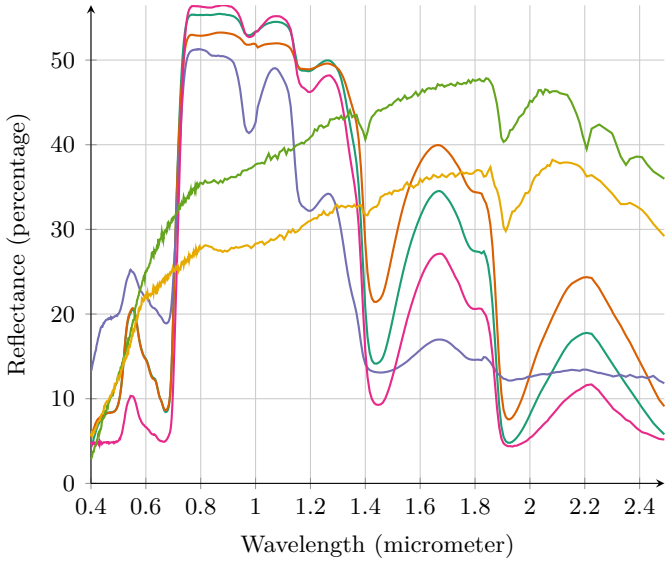
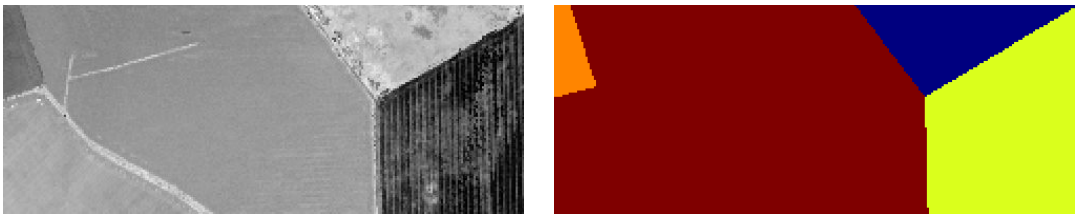


Figure 1.7.: Spectra used to generate the semi-synthetic image. 4 spectra are vegetation spectra and 2 are soil spectra.

Figure 1.9 shows the evolution of RMSE computed at each iteration for 250 iterations using the sampled $\hat{\mathbf{A}}^{(t)}$ matrix and the known $\mathbf{A}$ 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



(a) (b)

Figure 1.8.: Semi-synthetic image. Panchromatic view of the hyperspectral image (a), ground-truth (b).

and $\beta_1 = 1.2$ for Eches model. The evolution of the RMSE 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 Figure 1.10, 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 the Eches model is induced by the tendency to over-segment the image in more clusters than necessary.

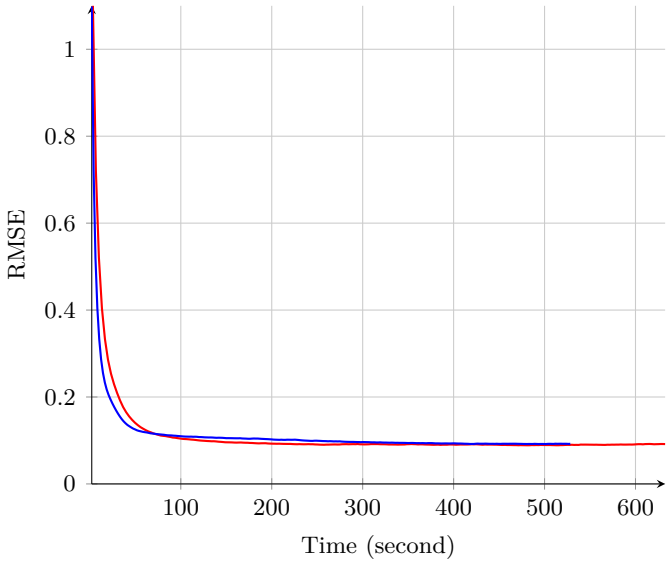
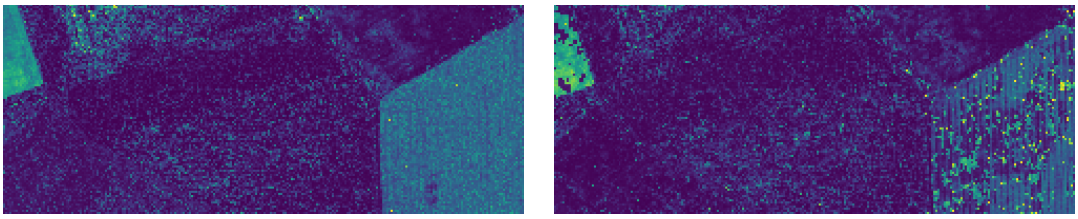


Figure 1.9.: Evolution of RMSE of the sampled $\hat{\mathbf{A}}^{(t)}$ matrix in function of the time for the proposed model (red) and Eches model (blue). Results are averaged in time and score over 10 trials.



(a) (b)

Figure 1.10.: Semi-synthetic image. Example of error map ($\|\hat{\mathbf{a}}_p - \mathbf{a}_p\|_2$) for proposed model (a), example of error map for the Eches model (b).

1.6.2. 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 [ND05] to obtain matrix $\mathbf{M}$. 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. Figure 1.11 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 1.1. Note that no performance measure of the unmixing step is provided since no abundance groundtruh is available for this real dataset. For comparison purpose, classification has been conducted with two conventional classifier namely random forest (RF) and a Bayesian Gaussian model (GM) using the 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 fitting 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

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

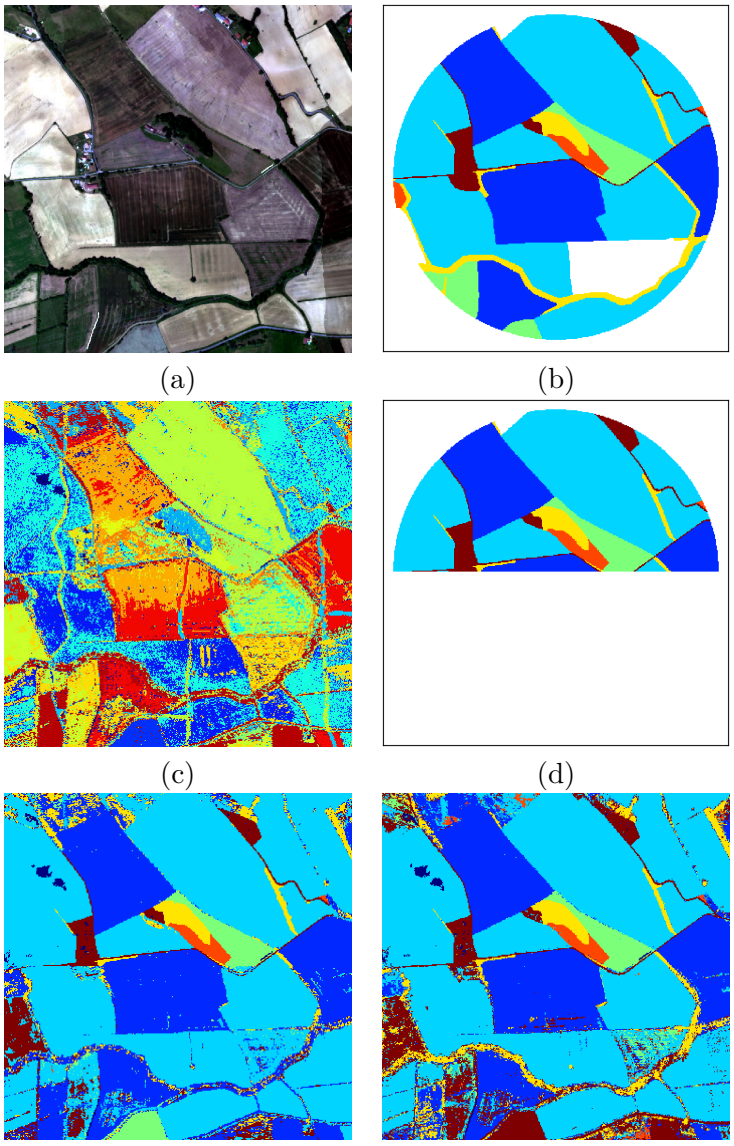


Figure 1.11.: Real MUESLI image. (a) colored composition of the hyperspectral image, (b) expert ground-truth, (c) estimated clustering, (d) training data, (e) estimated classification with proposed model and (f) estimated classification with random forest.

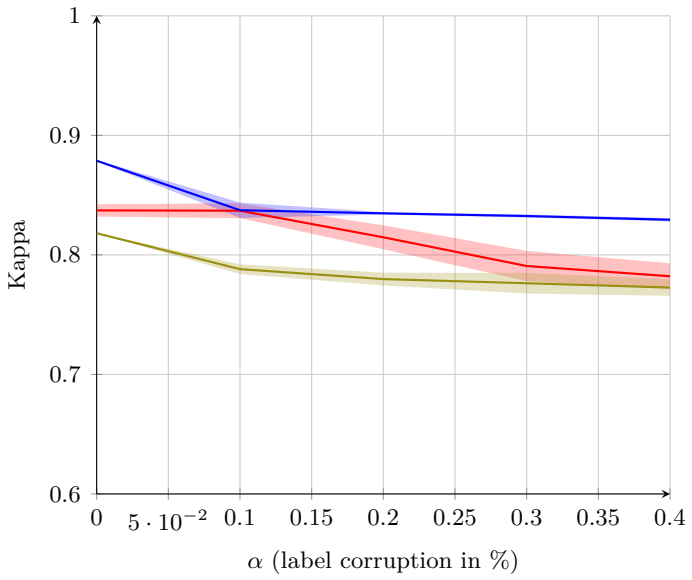


Figure 1.12.: Real MUESLI image. Classification accuracy measured with Cohen's kappa as a function of the percentage of label corruption α: proposed model (red), random forest (blue), PCA + Gaussian model (green). Shaded areas denote the intervals corresponding to the standard deviation computed over 10 trials.

additionally a spectral unmixing and estimates by-products of high interest for the user, for example matrix $\mathbf{Q}$.

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 1.6.1). 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 Figure 1.12 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 methods to classify remote sensing data and that the robustness to noise in labeled data is a well-documented property of this classification technique [Pel+17].

1.7. Conclusion and perspectives

This chapter 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.

A specific application of this model has been considered in the context of hyperspectral images to conduct 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. In the context of hyperspectral unmixing, it means including the estimating of the endmember matrix in the model. Instantiations of the proposed model in other applicative contexts will be also considered.

1.8. Conclusion (in French)

Ce chapitre introduit un modèle bayésien permettant d'effectuer conjointement une modélisation bas-niveau et une classification robuste d'une image. Ce modèle hiérarchique repose sur la mise en place de deux champs de Markov aléatoires promouvant une cohérence entre les différents niveaux de modélisation, à savoir, i) entre le clustering effectué sur les variables latentes de la modélisation bas-niveau et les labels de classes estimés, et ii) entre les labels de classes estimés et la donnée labellisée fournie par les experts utilisée dans la cadre de la classification supervisée.

Le modèle proposé a été construit de sorte à obtenir une classification robuste aux erreurs de labellisation potentiellement présentes de les données d'apprentissage. De plus, la méthode introduite va plus loin en proposant une correction de ces erreurs de labellisation.

Une instance particulière du modèle a été considérée dans le contexte de l'imagerie hyperspectral pour effectuer conjointement le démélange spectral et la classification d'une image. Une évaluation quantitative et qualitative a ensuite été réalisée sur des images synthétiques

puis réelles. Les résultats montrent la pertinence et les bonnes performances de la méthode introduite. La possibilité d'une interprétation très riche des résultats a également été mise en lumière. Une piste de travail envisagée pour la suite de ce travail est la généralisation du modèle pour gérer un cas entièrement non-supervisé pour la modélisation bas-niveau. Dans le contexte du démélange hyperspectral, cela équivaudrait à inclure l'estimation de la matrice des endmembers dans le modèle. Des instanciations du modèle dans d'autres contextes applicatifs seront également envisagées.

Chapter 2.

Matrix cofactorization approach for joint classification and spectral unmixing

This chapter has been adapted from the journal paper [Lag+19c]. This work was carried out in cooperation with Pr. José M. Bioucas-Dias, partly during a one month stay in Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa. This work has also been discussed in the conference papers [Lag+19b; Lag+19a].

Contents

2.1. Introduction (in French)

Continuant à explorer les idées introduites dans le chapitre précédent, ce chapitre met en jeu des concepts similaires mais sous une perspective différente. En effet, la modélisation bas-niveau des données peut être vue comme un problème d'apprentissage de représentation. Ce problème a été traité sous différentes perspectives et sous différentes dénominations telles que l'apprentissage de dictionnaires [AEB06], la séparation de sources [ZP01], l'analyse de facteurs [Cav+18a], la factorisation de matrices [KBV09] ou l'apprentissage de sous-espaces [EV13]. Nombre de ces méthodes visent à identifier un dictionnaire et un mélange en minimisant, à l'aide d'un méthode d'optimisation, une erreur de reconstruction mesurant une divergence entre le modèle et les données. Un des avantages de recourir à une méthode d'optimisation est le possibilité de s'appuyer sur des schémas d'optimisation rapides, bien documentés et bien établis comme les méthodes de *splitting* de variables [Boy+11], les méthodes proximales [BST14], etc. L'intérêt pratique de ces méthodes est généralement préféré à une estimation plus précise mais coûteuse comme celle réalisée avec une méthode MCMC.

De plus, comme expliqué dans l'introduction, l'idée de combiner l'apprentissage de représentation et la classification a déjà été considérée dans ce contexte [MBP12]. Certains travaux ont même introduit l'idée de réaliser ces tâches de manière simultanées [Zha+18b ; ZL10 ; JLD11]. En particulier, l'apprentissage de représentation et la classification conjoints peuvent être exprimés comme un problème de cofactorisation. Les deux tâches s'écrivent individuellement comme des problèmes de factorisation puis des contraintes entre les dictionnaires et les matrices de codage des deux problèmes sont imposées. Ces modèles de cofactorisation ont prouvé leur efficacité dans de nombreux champs d'application, tels que la fouille de texte [WB11], la séparation de sources audio [Yoo+10], ou l'analyse d'images [YYI12 ; AM18].

Cependant, le plupart de ces méthodes se focalisent sur les résultats de la classification et opposent les capacités de reconstruction et de discrimination des modèles au lieu de construire une structure cohérente qui permettrait de concilier ces deux capacités. Capitalisant sur le modèle bayésien développé dans le chapitre 1, ce chapitre propose une méthode de cofactorisation pour l'analyse d'images. L'apprentissage de représentation et la classification sont liés par les matrices de codage des deux problèmes de factorisation. Un clustering des représentations de faible dimension est réalisé et les vecteurs d'attribution aux clusters sont utilisés comme vecteurs de codage du problème de classification, c'est-à-dire comme de descripteurs. Cette méthode de couplage novatrice engendre un modèle hiérarchique co-

hérent et entièrement interprétable. Pour résoudre le problème d'optimisation non-convexe et non-lisse résultant, un algorithme de minimisation linéarisée alternée proximale est mis en place de sorte à fournir la garantie de convergence vers un point critique de la fonction objectif [BST14].

Ce chapitre s'organise de la façon suivante. La Section 2.3 pose les deux modèles de factorisation utilisés pour effectuer respectivement l'apprentissage de représentation et la classification puis expose le problème de cofactorisation. Le schéma d'optimisation utilisé pour trouver une solution au problème non-convexe résultant est également détaillé. Revenant ensuite au cas d'étude de ce manuscrit, une application au cas de l'analyse d'images hyperspectrales est considérée dans la Section 2.4 en considérant la classification et le démélange spectrale conjoints. Les performances sont illustrées à l'aide d'expérimentations sur données synthétiques puis réelles dans le Section 2.5. Enfin, la Section 2.6 conclut se chapitre et présente quelques perspectives de recherche.

2.2. Introduction

Following the work presented in the previous chapter, this chapter introduces similar concepts but proposes a different perspective. Indeed, representation learning has been considered from different perspectives, in particular known as dictionary learning [AEB06], source separation [ZP01], factor analysis [Cav+18a], matrix factorization [KBV09] or subspace learning [EV13]. Many of these methods attempt to identify a dictionary and a mixture by minimizing a reconstruction error measuring the discrepancy between the chosen model and the dataset with the help of an optimization method. One of the advantage to rely on an optimization approach is the possibility to rely on fast, well-established and well-documented optimization schemes such as variable splitting methods [Boy+11], proximal methods [BST14], etc. The practical interest of these methods is generally preferred to the exhaustive estimation produced by Bayesian model with MCMC estimation.

Moreover, as explained in the introduction section, the idea of combining the representation learning and classification tasks has already been considered in this context [MBP12]. Some works introduce the idea of performing the two tasks simultaneously [Zha+18b; ZL10; JLD11]. In particular, joint representation learning and classification can be cast as a cofactorization problem. Both tasks are interpreted as individual factorization problems and constraints between the dictionaries and coding matrices associated with the two problems can then be imposed. These cofactorization-based models have proven to be highly efficient in many application fields, *e.g.* for text mining [WB11], music source separation [Yoo+10],

and image analysis [YYI12; AM18].

However, most of the available methods tends to focus on classification results and oppose reconstruction and discriminative ability of the models instead of building a coherent hierarchical structure allowing to conciliate both abilities. Capitalizing on the Bayesian setting proposed in Chapter 1, this chapter proposes a particular cofactorization method, with a dedicated application to multivariate image analysis. The representation learning and classification tasks are related through the coding matrices of the two factorization problems. A clustering is performed on the low-dimensional representations and the clustering attribution vectors are used as coding vectors for the classification. This novel coupling approach produces a coherent and fully-interpretable hierarchical model. To solve the resulting non-convex non-smooth optimization problem, a proximal alternating linearized minimization (PALM) algorithm is derived, yielding guarantees of convergence to a critical point of the objective function [BST14].

This chapter is organized as follows. Section 2.3 defines the two factorization problems used to perform representation learning and classification and further discusses the joint cofactorization problem. It also details the optimization scheme developed to solve the resulting non-convex minimization problem. Focusing on the use case in this manuscript, an application of the introduced generic framework to hyperspectral image analysis is conducted in Section 2.4 through the dual scope of spectral unmixing and classification. Performance of the proposed framework is illustrated thanks to experiments conducted on synthetic and real data in Section 2.5. Finally, Section 2.6 concludes the chapter and presents some research perspectives to this work.

2.3. Proposed generic framework

The representation learning and classification tasks are generically defined as factorization matrix problems in Sections 2.3.1 and 2.3.2. To derive a unified cofactorization formulation, a third step consists in drawing the link between these two independent problems. In this work, this coupling is ensured by imposing a consistent structure between the two coding matrices corresponding to the low-dimensional representation and the feature matrices, respectively. As detailed in Section 2.3.3, it is expressed as a clustering task where the parameters describing the attribution to the clusters are the feature vectors, *i.e.*, the coding matrix resulting from the classification task. Particular instances of these three tasks will be detailed in Section 2.4 for an application to multiband image analysis.

2.3.1. Representation learning

The fundamental assumption in representation learning is that the P d-dimensional samples, gathered in matrix $\mathbf{Y} \in \mathbb{R}^{d\times P}$, belong to a R-dimensional subspace such that $R \ll d$. The aim is then to recover this manifold, where samples can be expressed as combinations of elementary vectors, herein the column of the matrix $\mathbf{M} \in \mathbb{R}^{d\times R}$ sometimes referred to as dictionary. These samples can be subsequently represented thanks to the so-called coding matrix $\mathbf{A} \in \mathbb{R}^{R\times P}$. Formally, identifying the dictionary and the coding matrices can be generally expressed as a minimization problem

$$\min_{\mathbf{M},\mathbf{A}} \mathcal{J}_{\mathrm{r}}(\mathbf{Y}|\psi(\mathbf{M},\mathbf{A})) + \lambda_m \mathcal{R}_m(\mathbf{M}) + \imath_{\mathbb{M}}(\mathbf{M}) + \lambda_a \mathcal{R}_a(\mathbf{A}) + \imath_{\mathbb{A}}(\mathbf{A}) \tag{2.1}$$

where $\psi(\cdot)$ is a mixture function (e.g., linear or bilinear operator), $\mathcal{J}_{\mathrm{r}}(\cdot)$ is an appropriate cost function, for example derived from a β-divergence [CJ11], $\mathcal{R}_{\cdot}(\cdot)$ denote penalizations weighted by the parameter $\lambda_{\cdot}$ and $\imath_{\cdot}(\cdot)$ is the indicator functions defined here on the respective sets $\mathbb{M} \subset \mathbb{R}^{d\times R}$ and $\mathbb{A} \subset \mathbb{R}^{R\times P}$ imposing some constraints on the dictionary and coding matrices.

In the case of a linear embedding adopted in this work, the mixture function writes

$$\psi(\mathbf{M},\mathbf{A}) = \mathbf{M}\mathbf{A}. \tag{2.2}$$

In this context, the problem (2.1) can be cast as a factor analysis driven by the cost function $\mathcal{J}_{\mathrm{r}}(\cdot)$. Depending on the applicative field, typical data-fitting measures include the Itakura-Saito, the Euclidean and the Kullback-Leibler divergences [CJ11]. Assuming a low-rank model (i.e., $R \leq d$), specific choices for the sets $\mathbb{A}$ and $\mathbb{M}$ lead to various standard factor models. For instance, when $\mathbb{M}$ is chosen as the Stiefel manifold, the solution of (2.1) is given by a principal component analysis (PCA) [Jol86]. When $\mathbb{M}$ and $\mathbb{A}$ impose nonnegativity of the dictionary and coding matrix elements, the problem is known as nonnegative matrix factorization [LS99; PT94].

Within a supervised context, the dictionary $\mathbf{M}$ can be chosen thanks to a end-user expertise or estimated beforehand. Without loss of generality but for the sake of conciseness, the framework described in this chapter assumes that this dictionary is known, possibly overcomplete as proposed in the experimental illustration described in Section 2.5. In this case, as in many applications, it makes sense to look for a sparse representation of the signal of interest to retrieve its most achievable compact representation [MBP12; BEZ08]. Following this strategy, we propose to consider an ℓ_1-norm sparsity penalization on the coding vectors,

leading to representation learning task defined by

$$\min_{\mathbf{A}} \mathcal{J}_{\mathrm{r}}(\mathbf{Y}|\mathbf{M}\mathbf{A}) + \lambda_a \|\mathbf{A}\|_1 + \imath_{\mathbb{A}}(\mathbf{A}) \tag{2.3}$$

where $\|\mathbf{A}\|_1 = \sum_{p=1}^{P} \|\mathbf{a}_p\|_1$ with $\mathbf{a}_p$ denoting the pth column of $\mathbf{A}$.

2.3.2. Supervised classification

To clearly define the classification task, let first introduce some key notations. The index subset of samples with an available groundtruth is denoted as $\mathcal{L}$ while the index subset of unlabeled samples is $\mathcal{U}$ such that $\mathcal{L} \cap \mathcal{U} = \emptyset$ and $\mathcal{L} \cup \mathcal{U} = \mathcal{P}$ with $\mathcal{P} \triangleq \{1, \ldots, P\}$. Classifying the unlabeled samples consists in assigning each of them to one of the C classes. This can be reformulated as the estimation of a $C \times P$ matrix $\mathbf{C}$ whose columns correspond to unknown C-dimensional attribution vectors $\mathbf{c}_p = [c_{1,p}, \ldots, c_{C,p}]^T$. Each vector is made of 0 except for $c_{i,p} = 1$ when the pth sample is assigned the ith class.

Numerous classification rules have been proposed in the literature [HTF09]. Most of them rely on a $K \times P$ matrix $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_P]$ of features $\mathbf{z}_p$ ($p \in \mathcal{P}$) associated with each sample and derived from the raw data. Within a supervised framework, the attribution matrix $\mathbf{C}_{\mathcal{L}}$ and feature matrix $\mathbf{Z}_{\mathcal{L}}$ of the labeled data are exploited during the learning step, where $\cdot_{\mathcal{L}}$ denotes the corresponding submatrix whose columns are indexed by $\mathcal{L}$. For a wide range of classifiers, deriving a classification rule can be achieved by solving the optimization problem

$$\min_{\mathbf{Q}} \mathcal{J}_{\mathrm{c}}(\mathbf{C}_{\mathcal{L}}|\phi(\mathbf{Q}, \mathbf{Z}_{\mathcal{L}})) + \lambda_q \mathcal{R}_q(\mathbf{Q}) + \imath_{\mathbb{Q}}(\mathbf{Q}) \tag{2.4}$$

where $\mathbf{Q} \in \mathbb{R}^{C \times K}$ is the set of classifier parameters to be inferred, $\mathcal{R}_q(\cdot)$ and $\imath_{\mathbb{Q}}(\cdot)$ refer respectively to regularizations and constraints imposed on $\mathbf{Q}$ and $\mathcal{J}_{\mathrm{c}}$ is a cost function measuring the quality of the classification such as the quadratic loss [ZL10] or cross-entropy [KB05]. Moreover, in (2.4), $\phi(\mathbf{Q}, \cdot)$ defines a element-wise nonlinear mapping between the features and the class attribution vectors parametrized by $\mathbf{Q}$, *e.g.*, derived from a sigmoid or a softmax operators. In this work, the classifier is assumed to be linear, which leads to a vector-wise post-nonlinear mapping

$$\phi(\mathbf{Q}, \mathbf{Z}_{\mathcal{L}}) = \phi(\mathbf{Q}\mathbf{Z}_{\mathcal{L}}) \tag{2.5}$$

with

$$\phi(\mathbf{X}) = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_p)]\,. \tag{2.6}$$

Once the classifier parameters have been estimated by solving (2.4), the unknown attribution vectors $\mathbf{C}_{\mathcal{U}}$ can be subsequently inferred during the testing step by applying the nonlinear transformation to the corresponding predicted features $\hat{\mathbf{Z}}_{\mathcal{U}}$ associated with the unlabeled samples. The obtained outputs are relaxed attribution vectors $\hat{\mathbf{c}}_p = \phi(\mathbf{Q}\hat{\mathbf{z}}_p)$ $(p \in \mathcal{U})$ and the most probable predicted sample class can be computed as $\operatorname{argmax}_i c_{i,p}$.

Under the proposed formulation of the classification task, the learning and testing steps can be conducted simultaneously, a framework usually referred to as semi-supervised, with the beneficial opportunity to introduce additional regularizations and/or constraints on the submatrix of unknown attribution vectors $\mathbf{C}_{\mathcal{U}}$. The initial problem (2.4) is thus extended to the following one

$$\min_{\mathbf{Q},\mathbf{C}_{\mathcal{U}}} \mathcal{J}_c(\mathbf{C}|\phi(\mathbf{Q}\mathbf{Z})) + \lambda_q \mathcal{R}_q(\mathbf{Q}) + \lambda_c \mathcal{R}_c(\mathbf{C}) + \imath_{\mathbb{Q}}(\mathbf{Q}) + \imath_{\mathbb{C}}(\mathbf{C}_{\mathcal{U}}) \tag{2.7}$$

where $\mathbf{C} = [\mathbf{C}_{\mathcal{L}}\ \mathbf{C}_{\mathcal{U}}]$ and $\mathbb{C} \subset \mathbb{R}^{C\times|\mathcal{U}|}$ denotes a feasible set for the attribution matrix $\mathbf{C}_{\mathcal{U}}$. In particular, nonnegativity and sum-to-one constraints can be introduced such that each attribution vector $\mathbf{c}_p$ $(p \in \mathcal{U})$ can then be interpreted as a probability vector of belonging to each class. In such a case, the feasible set is chosen as $\mathbb{C} = \mathbb{S}_C^{|\mathcal{U}|}$ where

$$\mathbb{S}_C \triangleq \left\{ \mathbf{u} \in \mathbb{R}^C | \forall k,\ u_k \geq 0 \text{ and } \sum_{k=1}^{C} u_k = 1 \right\}. \tag{2.8}$$

2.3.3. Coupling representation learning and classification

Up to this point, the representation learning and supervised classification tasks have been formulated as two independent matrix factorization problems given by (2.2) and (2.5), respectively. This work proposes to join them by drawing an implicit relation between two factors involved in these two problems. Inspired by hierarchical Bayesian models such as the one proposed in [Lag+18], both problems are coupled through the activation matrices $\mathbf{A}$ and $\mathbf{Z}$, as illustrated in Figure 2.1. More precisely, the coding vectors in $\mathbf{A}$ are clustered such that the feature vectors in $\mathbf{Z}$ are defined as the attribution vectors to the K clusters. Ideally, clustering attribution vectors $\mathbf{z}_p$ are filled with zeros except for $z_{k,p} = 1$ when $\mathbf{a}_p$ is associated with the kth cluster. Thus, the vectors $\mathbf{z}_p$ $(p \in \mathcal{P})$ are assumed to be defined on the K-dimensional probability simplex $\mathbb{S}_K$ similarly defined as (2.8) and ensuring nonnegativity and sum-to-one constraints. Many clustering algorithms can be expressed as optimization problem such as the well-known k-means algorithm and many of its variants [Con17; Pom+14]. Adopting this formulation, and denoting $\boldsymbol{\theta}$ the set of parameters of the

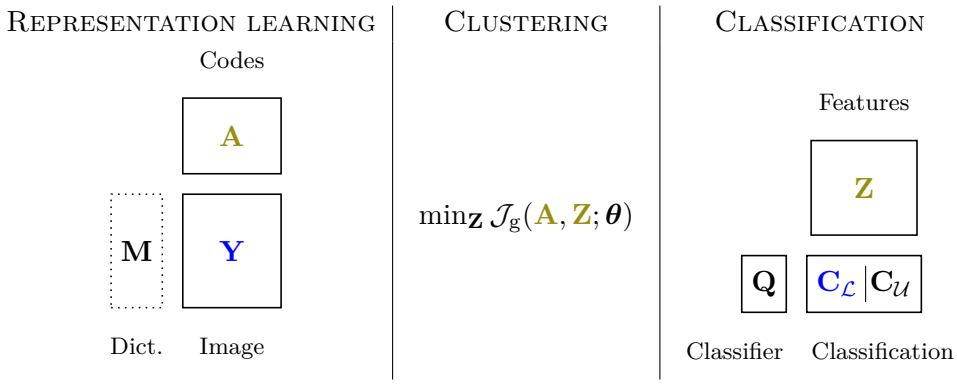


Figure 2.1.: Structure of the cofactorization model. Variables in *blue* stand for observations or available external data. Variables in *olive green* are linked through the clustering task here formulated as an optimization problem. The variable in a dotted box is assumed to be known or estimated beforehand in this work.

clustering algorithm, the clustering task can be defined as the minimization problem

$$\min_{\mathbf{Z},\boldsymbol{\theta}} \mathcal{J}_{\mathrm{g}}(\mathbf{A}, \mathbf{Z}; \boldsymbol{\theta}) + \lambda_z \mathcal{R}_z(\mathbf{Z}) + \lambda_\theta \mathcal{R}_\theta(\boldsymbol{\theta}) + \imath_{\mathbb{S}_K^P}(\mathbf{Z}) + \imath_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \tag{2.9}$$

where $\boldsymbol{\Theta}$ defines a feasible set for the parameters $\boldsymbol{\theta}$.

It is worth noting that introducing this coupling term is one of the major novelty of the proposed approach. When considering task-driven dictionary learning methods, it is usual to intertwine the representation learning and the classification tasks by directly imposing $\mathbf{A} = \mathbf{Z}$ [ZL10; SNT15]. Since these methods generally rely on a linear classifier, one major drawback of such approaches is their unability to deal with non-separable classes in the low-dimensional representation space. In such cases, the underlying model cannot be discriminative and descriptive simultaneously and the resulting tasks become adversarial. When considering the proposed coupling term, the cluster attribution vectors $\mathbf{z}_p$ offer the possibility of linearly separating any group of clusters from the others. As a consequence, the model benefits from more flexibility, with both discriminative and descriptive abilities in a more general sense.

2.3.4. Global cofactorization problem

Unifying the representation learning task (2.3) and the classification task (2.7) through the clustering task (2.9) leads to the following joint cofactorization problem

$$\begin{aligned}
&\min_{\substack{\mathbf{A},\mathbf{Q},\mathbf{C}_{\mathcal{U}},\\ \mathbf{Z},\boldsymbol{\theta}}} \lambda_0 \mathcal{J}_{\mathrm{r}}(\mathbf{Y}|\mathbf{M}\mathbf{A}) + \lambda_a \|\mathbf{A}\|_1 \\
&+ \lambda_1 \mathcal{J}_{\mathrm{c}}(\mathbf{C}|\phi(\mathbf{Q}\mathbf{Z})) + \lambda_q \mathcal{R}_q(\mathbf{Q}) + \lambda_c \mathcal{R}_c(\mathbf{C}) \\
&+ \lambda_2 \mathcal{J}_{\mathrm{g}}(\mathbf{A},\mathbf{Z};\boldsymbol{\theta}) + \lambda_z \mathcal{R}_z(\mathbf{Z}) + \lambda_\theta \mathcal{R}_\theta(\boldsymbol{\theta}) \\
&+ \imath_{\mathbb{A}}(\mathbf{A}) + \imath_{\mathbb{Q}}(\mathbf{Q}) + \imath_{\mathbb{S}_K^{|\mathcal{U}|}}(\mathbf{C}_{\mathcal{U}}) + \imath_{\mathbb{S}_K^{P}}(\mathbf{Z}) + \imath_{\boldsymbol{\Theta}}(\boldsymbol{\theta})
\end{aligned} \tag{2.10}$$

where λ_0, λ_1 and λ_2 control the respective contribution of each task data-fitting term. All notations and parameter dimensions are summarized in Table 2.1. A generic algorithmic scheme solving the problem (2.10) is proposed in the next section.

Table 2.1.: Overview of notations.

	parameter
$P \in \mathbb{R}$	number of observations
$d \in \mathbb{R}$	dimension of observations
$C \in \mathbb{R}$	number of classes
$K \in \mathbb{R}$	number of features/clusters
$\mathcal{P} = \{1, \ldots, P\}$	index set of observations
$\mathcal{L} \subset \mathcal{P}$	index set of labeled samples
$\mathcal{L}_i \subset \mathcal{L}$	index set of labeled samples in the ith class
$\mathcal{U} = \mathcal{P} \backslash \mathcal{L}$	index set of unlabeled samples
$\mathbf{Y} \in \mathbb{R}^{d \times P}$	observations
$\mathbf{M} \in \mathbb{R}^{d \times R}$	dictionary
$\mathbf{A} \in \mathbb{R}^{R \times P}$	coding matrix
$\mathbf{Q} \in \mathbb{C}^{C \times P}$	classifier parameters
$\mathbf{C}_{\mathcal{L}} \in \mathbb{R}^{C \times \lvert\mathcal{L}\rvert}$	attribution matrix of labeled data
$\mathbf{C}_{\mathcal{U}} \in \mathbb{R}^{C \times \lvert\mathcal{U}\rvert}$	attribution matrix of unlabeled data
$\mathbf{C} = [\mathbf{C}_{\mathcal{L}}\ \mathbf{C}_{\mathcal{U}}]$	class attribution matrix
$\mathbf{Z} \in \mathbb{R}^{K \times P}$	cluster attribution matrix
$\boldsymbol{\theta} \in \boldsymbol{\Theta}$	clustering parameters

2.3.5. Optimization scheme

The minimization problem defined by (2.10) is multi-convex, *i.e.*, convex according to each variable independently, but not globally convex. To reach a local minimizer, we propose

to resort to the proximal alternating linearized minimization (PALM) algorithm introduced in [BST14]. This algorithm is guaranteed to converge to a critical point of the objective function even in the case of non-convex problem. This means that, if the initialization is good enough, it is expected to likely converge to a solution close to the global optimum. To implement PALM, the problem (2.10) is rewritten in the form of an unconstrained problem expressed as a sum of a smooth coupling term $g(\cdot)$ and separable non-smooth terms $f_j(\cdot)$ ($j \in \{0, \ldots, 4\}$) as follows

$$\min_{\substack{\mathbf{A},\boldsymbol{\theta},\mathbf{Z},\\ \mathbf{Q},\mathbf{C}_\mathcal{U}}} f_0(\mathbf{A}) + f_1(\boldsymbol{\theta}) + f_2(\mathbf{Z}) + f_3(\mathbf{C}_\mathcal{U}) + f_4(\mathbf{Q}) + g(\mathbf{A}, \boldsymbol{\theta}, \mathbf{Z}, \mathbf{C}_\mathcal{U}, \mathbf{Q}) \tag{2.11}$$

where

$$\begin{aligned} f_0(\mathbf{A}) &= \imath_{\mathbb{A}}(\mathbf{A}) + \lambda_a \|\mathbf{A}\|_1 \\ f_1(\boldsymbol{\theta}) &= \imath_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \end{aligned} \qquad \begin{aligned} f_2(\mathbf{Z}) &= \imath_{\mathbb{S}_K^P}(\mathbf{Z}) \\ f_3(\mathbf{C}_\mathcal{U}) &= \imath_{\mathbb{S}_K^{|\mathcal{U}|}}(\mathbf{C}_\mathcal{U}) \\ f_4(\mathbf{Q}) &= \imath_{\mathbb{Q}}(\mathbf{Q}) \end{aligned}$$

and the coupling function is

$$\begin{aligned} g(\mathbf{A}, \boldsymbol{\theta}, \mathbf{Z}, \mathbf{C}_\mathcal{U}, \mathbf{Q}) = \lambda_0 \mathcal{J}_\mathrm{r}(\mathbf{Y}|\mathbf{M}\mathbf{A}) \\ + \lambda_1 \mathcal{J}_\mathrm{c}(\mathbf{C}|\phi(\mathbf{Q}\mathbf{Z})) + \lambda_q \mathcal{R}_q(\mathbf{Q}) + \lambda_c \mathcal{R}_c(\mathbf{C}) \\ + \lambda_2 \mathcal{J}_\mathrm{g}(\mathbf{M}, \mathbf{Z}; \boldsymbol{\theta}) + \lambda_z \mathcal{R}_z(\mathbf{Z}) + \lambda_\theta \mathcal{R}_\theta(\boldsymbol{\theta}). \end{aligned} \tag{2.12}$$

To ensure the stated guarantees of PALM, each of the independent non-smooth term has to be a proper, lower semi-continuous function $f_j : \mathbb{R}^{n_j} \to (-\infty, +\infty]$, which ensures in particular that the associated proximal operator is well-defined. Additionally, sufficient conditions on the coupling function are that $g(\cdot)$ is a $\mathcal{C}^2$ function (i.e., with continuous first and second derivatives) and that its partial gradients are globally Lipschitz. For example, partial gradient $\nabla_\mathbf{A} g(\mathbf{A}, \boldsymbol{\theta}, \mathbf{Z}, \mathbf{C}_\mathcal{U}, \mathbf{Q})$ should be globally Lipschitz for any fixed $\boldsymbol{\theta}$, $\mathbf{Z}$, $\mathbf{C}_\mathcal{U}$, $\mathbf{Q}$, that is

$$\begin{aligned} \|\nabla_\mathbf{A} g(\mathbf{A}_1, \boldsymbol{\theta}, \mathbf{Z}, \mathbf{C}_\mathcal{U}, \mathbf{Q}) - \nabla_\mathbf{A} g(\mathbf{A}_2, \boldsymbol{\theta}, \mathbf{Z}, \mathbf{C}_\mathcal{U}, \mathbf{Q})\| \leq \\ L_\mathbf{A}(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{C}_\mathcal{U}, \mathbf{Q}) \|\mathbf{A}_1 - \mathbf{A}_2\|, \quad \forall \mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{R \times P} \end{aligned} \tag{2.13}$$

where $L_\mathbf{A}(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{C}_\mathcal{U}, \mathbf{Q})$, simply denoted $L_\mathbf{A}$ hereafter, is the Lipschitz constant. For the sake of conciseness, we refer to [BST14] to get further details.

The main idea of the algorithm is then to update each variable of the problem alternatively using a proximal gradient descent. The overall scheme is summarized in Algorithm 2. For a

practical implementation, one needs to compute the partial gradients of $g(\cdot)$ explicitly and their Lipschitz constants to perform a gradient descent step, followed by a proximal mapping associated with the non-smooth terms $f_j(\cdot)$. The objective function is then monitored at each iteration and the algorithm is stopped when convergence is reached. Note that, when a specific penalization $\mathcal{R}.(\cdot)$ is non-smooth or non-gradient-Lipschitz, it is possible to move it into the corresponding independent term $f_j(\cdot)$ to ensure the required property of the coupling function $g(\cdot)$. This is for instance the case for the sparse penalization used over $\mathbf{A}$ which has been moved into $f_0(\cdot)$. Nonetheless, as mentioned above, the proximal operator associated with each $f_j(\cdot)$ is needed. Thus, even when the function consists of several terms, a closed-form expression of this operator should be known. Alternatively, one should be able to compose the proximal operators associated with each term of $f_j(\cdot)$ [Yu13].

Algorithm 2: PALM

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

2.4. Application to hyperspectral images analysis

A general framework has been introduced in the previous section. To illustrate, a particular instance of this generic framework is now considered, where explicit representation learning, classification and clustering are introduced. The specific case of hyperspectral images analysis is considered for this use case example.

Contrary to conventional color imaging which only captures the reflectance measure for three wavelengths (red, blue, green), hyperspectral imaging makes it possible to measure reflectance of the observed scene for several hundreds of wavelengths from visible to invisible domain. Each pixel of the image can thus be represented as a vector of reflectance, called spectrum, which characterizes the observed material.

One drawback of hyperspectral images is usually a weaker spatial resolution due to sensor limitations. The direct consequence of this poor spatial resolution is the presence of mixed

pixels, *i.e.*, pixels corresponding to areas containing several materials. Observed spectra are in this case the result of a specific mixture of the elementary spectra, called endmembers, associated with individual materials present in the pixel. The problem of retrieving the proportions of each material in each pixel is referred to as spectral unmixing [Bio+12]. This problem can be seen as a specific case of representation learning where the dictionary is composed of the set of endmembers standing for the endmember spectra and the coding matrix is the so-called abundance matrix containing the proportion of each material in each pixel.

Spectral unmixing is introduced as a representation learning task in Section 2.4.1. The specific classifier used for this application is then explained in Section 2.4.2 and finally Section 2.4.3 presents the clustering adopted to relate the abundance matrix and the classification feature matrix.

2.4.1. Spectral unmixing

As explained, each pixel of an hyperspectral image is characterized by a reflectance spectrum that physics theory approximates as a combination of endmembers, each corresponding to a specific material, as illustrated in Figure 2.2. Formally, in this applicative scenario, the d-dimensional sample $\mathbf{y}_p$ denotes the L-dimensional spectrum of the pth pixel of the hyperspectral image ($p \in \mathcal{P}$). Each observation vectors $\mathbf{y}_p$ can be expressed as a function of the endmember matrix $\mathbf{M}$ (containing the R elementary spectra) and the abundance vector $\mathbf{a}_p \in \mathbb{R}^R$ with $R \ll d$.

In the case of the most commonly adopted linear mixture model, each observation $\mathbf{y}_p$ is assumed to be a linear combination of the endmember spectra $\mathbf{m}_r$ ($r = 1, \ldots, R$) corrupted by some noise, underlying the linear embedding (2.2). Assuming a quadratic data-fitting term, the cost function associated with the representation learning task in (2.1) is written

$$\mathcal{J}_{\mathrm{r}}(\mathbf{Y}|\mathbf{MA}) = \frac{1}{2} \|\mathbf{Y} - \mathbf{MA}\|_{\mathrm{F}}^2 . \tag{2.14}$$

The abundance vector $\mathbf{a}_p$ is usually interpreted as a vector of proportions describing the proportion of each elementary component in the pixel. Thus, to derive an additive composition of the observed pixels, a nonnegative constraint is considered for each element of the abundance matrix $\mathbf{A}$, *i.e.*, $\mathbb{A} = \mathbb{R}_+^{R \times P}$. In this work, no sum-to-one constraint is considered since it has been argued that leaving this constraint offers a better adaptation to possible changes of illumination in the scene [Dru+16]. Additionally, as the endmember matrix $\mathbf{M}$ is the collection of reflectance spectra of the endmembers, it is also expected to

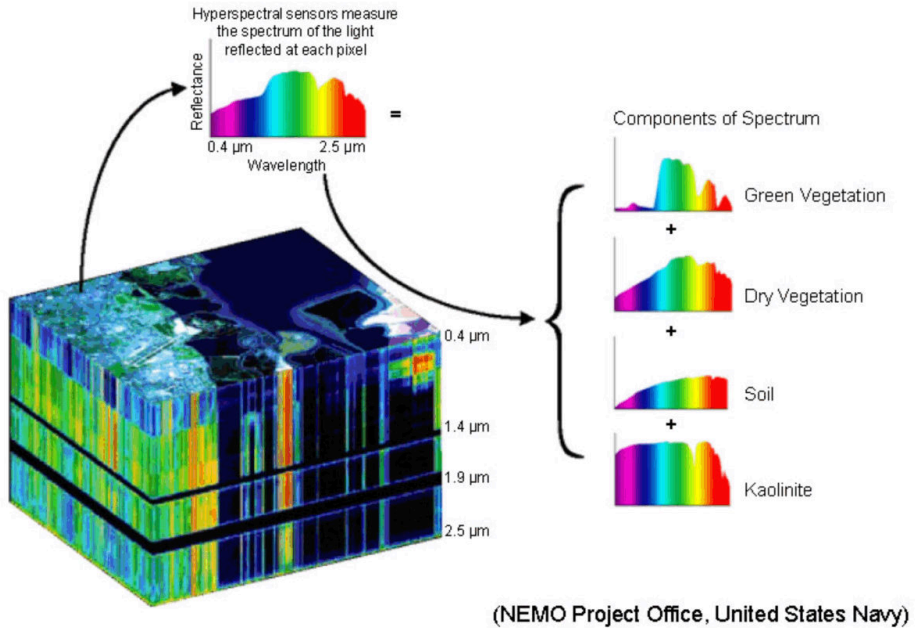


Figure 2.2.: Spectral unmixing concept (source US Navy NEMO).

be non-negative. When this dictionary needs to be estimated, the resulting problem is a sparse non-negative matrix factorization (NMF) task. When the dictionary is known or estimated beforehand, the resulting optimization problem is the nonnegative sparse coding problem

$$\min_{\mathbf{A}} \frac{1}{2} \|\mathbf{Y} - \mathbf{M}\mathbf{A}\|_{\mathrm{F}}^2 + \lambda_a \|\mathbf{A}\|_1 + \imath_{\mathbb{R}_+^{R\times P}}(\mathbf{A}) \tag{2.15}$$

where the sparsity penalization actually supports the assumption that only a few materials are present in a given pixel.

2.4.2. Classification

In the considered application, two loss functions associated with the classification problem have been investigated, namely quadratic loss and cross-entropy loss. One advantage of these two loss functions is that they can be used in a multi-class classification (i.e., with more than two classes). Moreover, this choice may fulfill the required conditions stated in Section 2.3.5 to apply PALM since, coupled with an appropriate $\phi(\cdot)$ function, both loss costs are smooth and gradient-Lipschitz according to each estimated variables.

Quadratic loss

The quadratic loss is the most simple way to perform a classification task and have been extensively used [JLD11; ZBS01; Yan+11]. It is defined as

$$\mathcal{J}_{\mathrm{c}}(\mathbf{C}|\hat{\mathbf{C}}) = \frac{1}{2}\left\|\mathbf{C}\mathbf{D} - \hat{\mathbf{C}}\mathbf{D}\right\|_{\mathrm{F}}^{2} \tag{2.16}$$

where $\hat{\mathbf{C}}$ denotes the estimated attribution matrix. In (2.16), the $P \times P$ matrix $\mathbf{D}$ is introduced to weight the contribution of the labeled data with respect to the unlabeled one and to deal with the case of unbalanced classes in the training set. Weights are chosen to be inversely proportional to class frequencies in the input data. The weight matrix is defined as the diagonal matrix $\mathbf{D} = \mathrm{diag}[d_1, \ldots, d_P]$ with

$$d_p = \begin{cases} \sqrt{\frac{1}{|\mathcal{L}_i|}}, & \text{if } p \in \mathcal{L}_i; \\ \sqrt{\frac{1}{|\mathcal{U}|}}, & \text{if } p \in \mathcal{U}; \end{cases} \tag{2.17}$$

where $\mathcal{L}_i$ denotes the set of indexes of labeled pixels of the ith class ($i = 1, \ldots, C$). Thus, considering a linear classifier, the generic classification problem in (2.7) can be specified for the quadratic loss

$$\min_{\mathbf{Q},\mathbf{C}_{\mathcal{U}}} \frac{1}{2}\left\|\mathbf{C}\mathbf{D} - \mathbf{Q}\mathbf{Z}\mathbf{D}\right\|_{\mathrm{F}}^{2} + \lambda_c \mathcal{R}_c(\mathbf{C}) + \imath_{\mathbb{S}_C^{|\mathcal{U}|}}(\mathbf{C}_{\mathcal{U}}) \tag{2.18}$$

where no additional constraints nor penalization is applied to the classifier parameters $\mathbf{Q}$. Besides, when samples obey a spatially coherent structure, as it is the case when analyzing hyperspectral images, it is often desirable to transfer this structure to the classification map. Such a characteristics can be achieved by considering a spatial regularization $\mathcal{R}_c(\mathbf{C})$ applied to the attributions vectors. Following this assumption, this work considers a regularized counterpart of the weighted vectorial total variation (vTV), promoting a spatially piecewise constant behavior of the classification map [Liu+18]

$$\|\mathbf{C}\|_{\mathrm{vTV}} = \sum_{m,n} \beta_{m,n} \sqrt{\left\|[\nabla_{\mathrm{h}}\mathbf{C}]_{m,n}\right\|_2^2 + \left\|[\nabla_{\mathrm{v}}\mathbf{C}]_{m,n}\right\|_2^2 + \epsilon} \tag{2.19}$$

where (m, n) are the spatial position pixel indexes and $[\nabla_{\mathrm{h}}(\cdot)]_{m,n}$ and $[\nabla_{\mathrm{v}}(\cdot)]_{m,n}$ stand for horizontal and vertical discrete gradient operators evaluated at a given pixel[1], respectively,

[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).

i.e.,

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

The weights $\beta_{m,n}$ can be computed beforehand to adjust the penalizations with respect to expected spatial variations of the scene. They can be estimated directly from the image to be analyzed or extracted from a complementary dataset as in [UFD18]. They will be specified during the experiments reported in Section 2.5. Moreover, the smoothing parameter $\epsilon > 0$ ensures the gradient-Lipschitz property of the coupling term $g(\cdot)$, as required in Section 2.3.5.

Cross-entropy loss

The quadratic loss has the advantage to be expressed simply and the associated Lipschitz constant of the partial gradients are trivially obtained. However, this loss function is known to be highly influenced by outliers which can result in a degraded predictive accuracy [Hub64]. A more sophisticated way to conduct the classification task is to consider a cross-entropy loss

$$\mathcal{J}_{\mathrm{c}}(\mathbf{C}|\hat{\mathbf{C}}) = -\sum_{p\in\mathcal{P}} d_p^2 \sum_{i\in\mathcal{C}} c_{i,p} \log\left(\hat{c}_{i,p}\right) \tag{2.20}$$

combined with a logistic regression, *i.e.*, where the nonlinear mapping (2.5) is element-wise defined as

$$[\phi\left(\mathbf{X}\right)]_{i,j} = \frac{1}{1+\exp(-x_{i,j})} \tag{2.21}$$

with $i \in \{1,\ldots,C\}$ and $p \in \mathcal{P}$. This classifier can actually be interpreted as a one-layer neural network with a sigmoid non-linearity. Cross-entropy loss is indeed a very conventional loss function in the neural network/deep learning community [Goo+16]. In the present case, the corresponding optimization problem can be written

$$\min_{\mathbf{Q},\mathbf{C}_{\mathcal{U}}} -\sum_{p\in\mathcal{P}} d_p^2 \sum_{i\in\mathcal{C}} c_{i,p} \log\left(\frac{1}{1+\exp(-\mathbf{q}_{i:}\mathbf{z}_p)}\right) + \lambda_q \mathcal{R}_q(\mathbf{Q}) + \lambda_c \left\|\mathbf{C}\right\|_{\mathrm{vTV}} + \imath_{\mathbb{S}_C^{|\mathcal{U}|}}(\mathbf{C}_{\mathcal{U}}) \tag{2.22}$$

where $\mathbf{q}_{i:} \in \mathbb{R}^{1\times K}$ denotes the ith line of the matrix $\mathbf{Q}$. The penalization $\mathcal{R}_q(\mathbf{Q})$ is here chosen as $\mathcal{R}_q(\mathbf{Q}) = \frac{1}{2}\left\|\mathbf{Q}\right\|_{\mathrm{F}}^2$ to prevent the loss function to artificially decrease when $\|\mathbf{q}_{i:}\|^2$ is increasing. This regularization has been extensively studied in the neural network literature where it is referred to as *weight decay* [Goo+16]. In (2.22), the regularization $\mathcal{R}_c(\mathbf{C}_{\mathcal{U}})$

applied to the attribution matrix is chosen again as a vTV-like penalization (see (2.19)).

2.4.3. Clustering

For the considered application, the conventional k-means algorithm has been chosen because of its straightforward formulation as an optimization problem. By denoting $\boldsymbol{\theta} = \{\mathbf{B}\}$ a $R \times K$ matrix collecting K centroids, the clustering task (2.9) can be rewritten as the following NMF problem [Pom+14]

$$\min_{\mathbf{Z},\mathbf{B}} \frac{1}{2} \|\mathbf{A} - \mathbf{B}\mathbf{Z}\|_{\mathrm{F}}^2 + \lambda_z \mathcal{R}_z(\mathbf{Z}) + \imath_{\mathbb{S}_K^P}(\mathbf{Z}) + \imath_{\mathbb{R}_+^{R\times K}}(\mathbf{B}) \tag{2.23}$$

where $\mathcal{R}_z(\mathbf{Z})$ should promote $\mathbf{Z}$ to be composed of orthogonal lines. Combined with the nonnegativity and sum-to-one constraints, it would ensure that $\mathbf{z}_p$ is a vector of zeros except for its kth component equal to 1, *i.e.*, meaning that the pth pixel belongs to the kth cluster. However, handling this orthogonality property within the PALM optimization scheme detailed in Section 2.3.5 is not straightforward, in particular because the proximal operator associated to this penalization cannot be explicitly computed. In this work, we propose to remove this orthogonality constraint since relaxed attribution vectors may be richer feature vectors for the classification task.

2.4.4. Multi-objective problem

Based on the quadratic and cross-entropy loss functions considered in the classification task, two distinct global optimization problems are obtained. When considering the quadratic loss of Section 2.4.2, the multi-objective problem (2.10) writes

$$\begin{aligned}
&\min_{\substack{\mathbf{A},\mathbf{Q},\mathbf{Z}\\ \mathbf{C}_{\mathcal{U}},\mathbf{B}}} \frac{\lambda_0}{2} \|\mathbf{Y} - \mathbf{M}\mathbf{A}\|_{\mathrm{F}}^2 + \lambda_a \|\mathbf{A}\|_1 + \imath_{\mathbb{R}_+^{R\times P}}(\mathbf{A}) \\
&+ \frac{\lambda_1}{2} \|\mathbf{C}\mathbf{D} - \mathbf{Q}\mathbf{Z}\mathbf{D}\|_{\mathrm{F}}^2 + \lambda_c \|\mathbf{C}\|_{\mathrm{vTV}} + \imath_{\mathbb{S}_C^{|\mathcal{U}|}}(\mathbf{C}_{\mathcal{U}}) \\
&+ \frac{\lambda_2}{2} \|\mathbf{A} - \mathbf{B}\mathbf{Z}\|_F^2 + \imath_{\mathbb{S}_K^P}(\mathbf{Z}) + \imath_{\mathbb{R}_+^{R\times K}}(\mathbf{B}).
\end{aligned} \tag{2.24}$$

Instead, when considering the cross-entropy loss function proposed in Section 2.4.2, the optimization problem (2.10) is defined as

$$\begin{aligned}
&\min_{\substack{\mathbf{A},\mathbf{Q},\mathbf{Z}\\ \mathbf{C}_{\mathcal{U}},\mathbf{B}}} \frac{\lambda_0}{2}\left\|\mathbf{Y}-\mathbf{M}\mathbf{A}\right\|_{\mathrm{F}}^2+\lambda_a\left\|\mathbf{A}\right\|_1+\imath_{\mathbb{R}_+^{R\times P}}(\mathbf{A})\\
&-\frac{\lambda_1}{2}\sum_{p\in\mathcal{P}} d_p^2\sum_{i\in\mathcal{C}} c_{i,p}\log\left(\frac{1}{1+\exp(-\mathbf{q}_{i:}\mathbf{z}_p)}\right)\\
&+\frac{\lambda_q}{2}\left\|\mathbf{Q}\right\|_{\mathrm{F}}^2+\lambda_c\left\|\mathbf{C}\right\|_{\mathrm{vTV}}+\imath_{\mathbb{S}_C^{|\mathcal{U}|}}(\mathbf{C}_{\mathcal{U}})\\
&+\frac{\lambda_2}{2}\left\|\mathbf{A}-\mathbf{B}\mathbf{Z}\right\|_F^2+\imath_{\mathbb{S}_K^P}(\mathbf{Z})+\imath_{\mathbb{R}_+^{R\times K}}(\mathbf{B}).
\end{aligned} \tag{2.25}$$

Both problems are particular instances of nonnegative matrix co-factorization [YYI12; Yoo+10]. To summarize, the hyperspectral pixel is first described as a combination of elementary spectra through the learning representation step, aka spectral unmixing. Then, assuming that there exist groups of pixels resulting from the same mixture of materials, a clustering is performed among the abundance vectors. And finally, attribution vectors to the clusters are used as feature vectors for the classification supporting the idea that classes are made of a mixture of clusters. For both multi-objective problems (2.24) and (2.25), all conditions required to the use of PALM algorithm described in Section 2.3.5 are met. Details regarding the two optimization schemes dedicated to these two problems are reported in the Appendix.

2.4.5. Complexity analysis

Regarding the computational complexity of the proposed Algorithm 2, deriving the gradients shows that it is dominated by matrix product operations. It yields that the algorithm has an overall computational cost in $\mathcal{O}(NK^2P)$ where N is the number of iterations.

2.5. Experiments

2.5.1. Implementation details

Before presenting the experimental results, it is worth clarifying the choices which have been made regarding the practical implementation of the proposed algorithms for the considered application. Important aspects are discussed below.

Convergence diagnosis and stopping rule – In all experiments conducted hereafter, the value of the objective function is monitored at each iteration to determine if convergence has been reached. The normalized difference between the last two consecutive values of the objective function is compared to a threshold and the algorithm stops when the criterion is smaller than this threshold (set as 10^{-4} for the conducted experiments). Figure 2.3 shows one example of the behavior of the objective function along the iterations as well as the behavior of several terms composing this overall objective function. As it can be observed from the figure, the global objective function is decreasing over the iteration, which is theoretically ensured by the PALM algorithm.

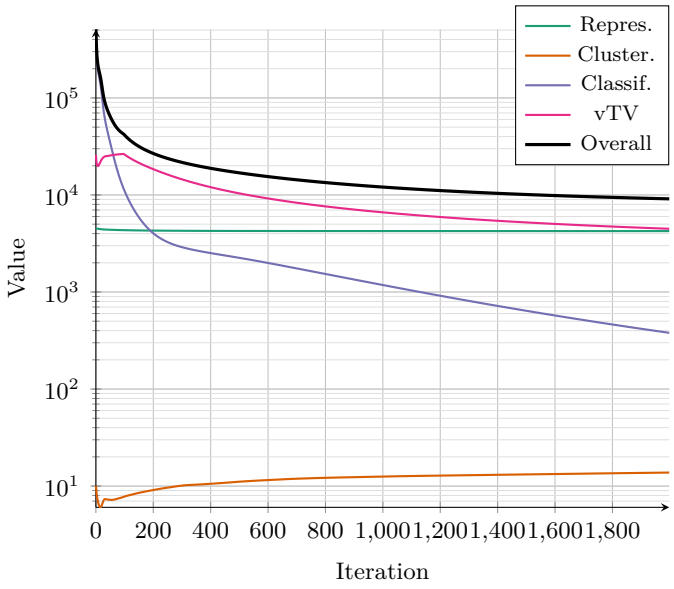


Figure 2.3.: Convergence of the various terms of objective function (representation learning, clustering, classification, vTV, total).

Initialization – As PALM algorithm only ensures convergence to a critical point and not a global optimum, it remains sensitive to initialization, which needs to be carefully chosen to reach relevant solutions. The initialization of the parameters associated with the learning representation and clustering steps relies on the self-dictionary learning method proposed in [GL18]. This method proposes to use observed pixels of the image as dictionary elements. The underlying assumption is that the image contains pure pixels, *i.e.*, composed of only a

single material. Formally, the initial estimate $\mathbf{A}^0$ of $\mathbf{A}$ is chosen as

$$\mathbf{A}^0 = \underset{\mathbf{A}}{\operatorname{argmin}} \frac{1}{2} \left\| \mathbf{Y} - \tilde{\mathbf{Y}}\mathbf{A} \right\|_{\mathrm{F}}^2 + \alpha \left\| \mathbf{A} \right\|_{1,2} \tag{2.26}$$

where $\|\mathbf{A}\|_{1,2} = \sum_{r=1}^{R} \|\mathbf{a}_{r,:}\|_2$ promotes the use of a reduced number of pixels as dictionary elements and $\tilde{\mathbf{Y}}$ is a submatrix of $\mathbf{Y}$ containing the pixel candidates to be used as dictionary elements. Following the strategy similarly proposed in [GL18], this subset $\tilde{\mathbf{Y}}$ is built as follows: *i*) for each class of the training set, a k-means is applied to the labeled samples to identify J clusters, *ii*) within a given class, one candidate is retained from each cluster as the pixel the farthest away from the centers of the other clusters (in term of spectral angle distance). This procedure provides a subset $\tilde{\mathbf{Y}}$ composed of $J \times C$ spectrally diverse candidates extracted from the labeled samples.

Then, regarding the representation learning step, only active elements in $\tilde{\mathbf{Y}}$, *i.e.*, those associated with non-zero rows in $\mathbf{A}^0$, are kept to define the dictionary $\mathbf{M}$. Finally, to initialize the variables involved in the clustering step, a k-means is conducted on $\mathbf{A}^0$ and the identified centroids are chosen as $\mathbf{B}^0$ while the corresponding attribution vectors define $\mathbf{Z}^0$. Finally, the classification parameters $\mathbf{Q}^0$ and attribution vectors $\mathbf{C}_{\mathcal{U}}^0$ are randomly initialized.

Weighting the vTV – As explained in Section 2.3.2, the classification is regularized by a weighted smooth vTV regularization. When all not fixed to the same value, the weights offer the possibility to account for natural boundaries in the observed scene, *i.e.*, variations in the classification map are expected to be localized at the edges in the image. As in [UFD18], an auxiliary dataset informing about the spatial structure of the image can be used to adjust these weights. Instead, in this work, we assume that no such external information is available. Thus these weights are directly computed from the hyperspectral image. More precisely, a virtually observed panchromatic image $\mathbf{y}_{\mathrm{PAN}} \in \mathbb{R}^P$, *i.e.*, a single band image, is first synthetized by averaging the bands of the hyperspectral image $\mathbf{Y}$. Then, the weights are chosen as

$$\beta_{m,n} = \frac{\tilde{\beta}_{m,n}}{\sum_{p,q} \tilde{\beta}_{p,q}} \text{ with } \tilde{\beta}_{m,n} = \frac{1}{\left\| [\nabla \mathbf{y}_{\mathrm{PAN}}]_{m,n} \right\|_2 + \sigma} \tag{2.27}$$

where $\nabla(\cdot) = [\nabla_{\mathrm{h}}(\cdot)\ \nabla_{\mathrm{v}}(\cdot)]^T$ is the gradient operator and σ is an hyperparameter chosen as $\sigma = 0.01$ to avoid numerical problems and to control the adaptive weighting (the larger σ, the less variation in the weighting) [SBC97].

Hyperparameter scaling – To balance the size and the dynamics of the matrices involved in the cofactorization problem, the hyperparameters λ_0 and λ_q in (2.24) and (2.25) have been set as

$$\lambda_0 = \frac{1}{d\,\|\mathbf{Y}\|_\infty^2}\tilde{\lambda}_0, \quad \lambda_q = \frac{P}{C}\tilde{\lambda}_q. \tag{2.28}$$

Then, for each experiment presented hereafter, the parameters $\tilde{\lambda}_\cdot$ have been empirically adjusted to obtain consistent results.

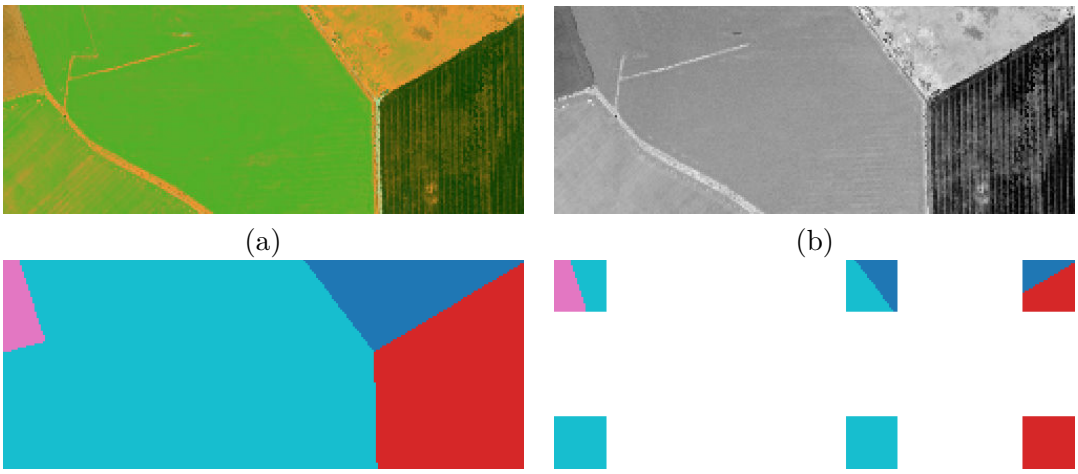


Figure 2.4.: Synthetic image: (a) colored composition of the hyperspectral image $\mathbf{Y}$, (b) panchromatic image $\mathbf{y}_{\text{PAN}}$, (c) classification ground-truth, (d) training set.

2.5.2. Synthetic hyperspectral image

Data generation – First, to assess the relevance of the proposed model, experiments have been conducted on synthetic images. These synthetic images have been generated using a real hyperspectral image which has been unmixed using the well-established unmixing method SUNSAL [BF10]. The extracted abundance maps and a set of 6 pure spectra from the hyperspectral library ASTER have been used to build a synthetic hyperspectral images with a realistic spatial organization. The resulting 100-by-250 pixel image presented in Figure 2.4 is composed of $d = 385$ spectral bands. The image is associated with a classification groundtruth ($C = 4$) based on the groundtruth of the original real image and a subpart of this groundtruth is assumed known and therefore used as training dataset for the supervised classification step.

Moreover, in this experiment, the endmember matrix $\mathbf{M}$ comprises the 6 spectra actually

Table 2.2.: Synthetic data: unmixing and classification results.

Model	F1-mean	Kappa	RMSE($\hat{\mathbf{A}}$)	RE	Time (s)
RF	0.913 ($\pm 1.4 \times 10^{-3}$)	0.907 ($\pm 1.3 \times 10^{-4}$)	N\A	N\A	0.9 (± 0.08)
FC-SUNSAL	0.893 ($\pm 6.4 \times 10^{-4}$)	0.912 ($\pm 3.7 \times 10^{-4}$)	0.120 ($\pm 3.1 \times 10^{-6}$)	0.37 ($\pm 5.1 \times 10^{-5}$)	6 (± 0.3)
CSR-SUNSAL	0.888 ($\pm 1.0 \times 10^{-3}$)	0.911 ($\pm 5.0 \times 10^{-4}$)	0.125 ($\pm 3.0 \times 10^{-6}$)	0.36 ($\pm 4.2 \times 10^{-5}$)	9 (± 0.5)
D-KSVD	0.520 ($\pm 3.1 \times 10^{-3}$)	0.653 ($\pm 3.4 \times 10^{-2}$)	N\A	0.23 ($\pm 4.1 \times 10^{-2}$)	382 (± 9)
LC-KSVD	0.879 ($\pm 3.7 \times 10^{-4}$)	0.904 ($\pm 1.0 \times 10^{-4}$)	N\A	30.4 ($\pm 1.0 \times 10^{-4}$)	96 (± 1)
Cofact-Q	0.911 ($\pm 3.5 \times 10^{-3}$)	0.893 ($\pm 3.5 \times 10^{-3}$)	0.0528 ($\pm 1.1 \times 10^{-4}$)	0.32 ($\pm 8.9 \times 10^{-4}$)	80 (± 6)
Cofact-CE	0.899 ($\pm 5.4 \times 10^{-2}$)	0.880 ($\pm 6.2 \times 10^{-2}$)	0.0524 ($\pm 1.3 \times 10^{-4}$)	0.27 ($\pm 2.2 \times 10^{-3}$)	61 (± 4)

used to generate the image. To evaluate the robustness of the method in a challenging scenario, these 6 initial endmember spectra are complemented with 9 endmembers not present in the image but very correlated with the 6 actually used ones. The endmember matrix is thus composed of $R = 15$ spectra depicted in Figure 2.5.

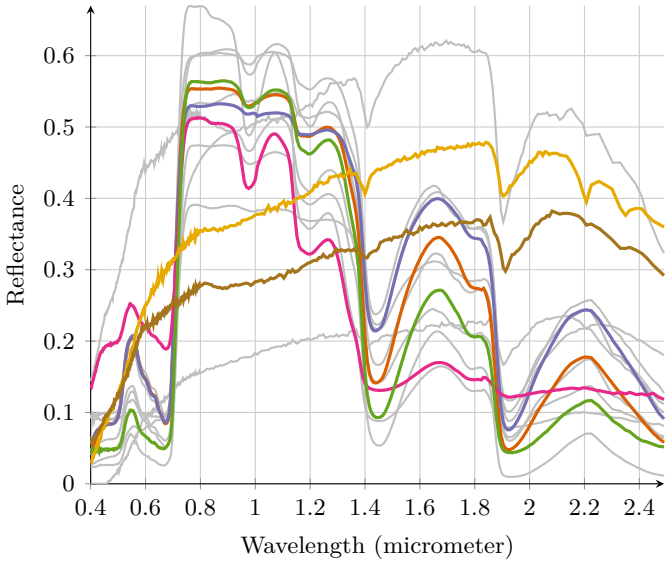


Figure 2.5.: Spectra used as dictionary $\mathbf{M}$ to generate the synthetic image. The 6 color spectra have been used to generate the semi-synthetic image (4 vegetation spectra and 2 soil spectra).

Compared methods – The proposed methods with quadratic (Q) and cross-entropy (CE) classification losses, denoted respectively by Cofact-Q and Cofact-CE, have been compared with state-of-the-art classification and unmixing methods. First, one considered competing method is the random forest (RF) classifier, which has been extensively used for the hy-

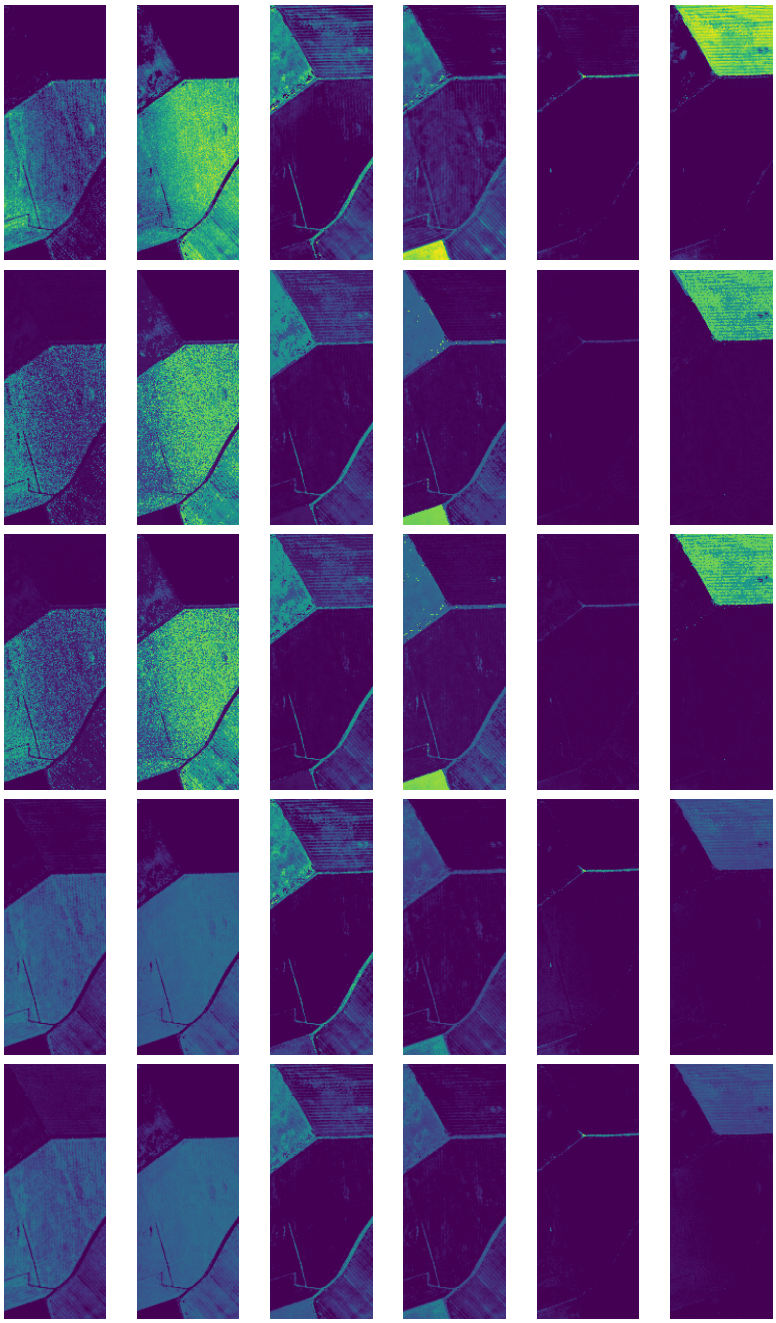


Figure 2.6.: Synthetic data: abundance maps of the 6 actual endmembers (from left to right): (1st row) ground-truth, (2nd row) Cofact-Q, (3rd row) Cofact-CE, (4rd row) FC-SUNSAL and (5th row) CSR-SUNSAL.

perspectral image classification. Parameters of the random forest (depth, number of trees) have been tuned using cross-validation with a grid-search strategy and the implementation provided in the *scikit-learn* Python library has been used [Ped+11]. Then, two unmixing methods proposed in [BF10] has been tested, namely the fully constrained least squares (FC-SUNSAL) and the constrained sparse regression (CSR-SUNSAL). FC-SUNSAL basically relies on the same data fitting term (2.14) considered in the proposed cofactorization method, under non-negativity and sum-to-one constraints applied to the abundance vectors. Conversely, the CSR-SUNSAL problem removes the sum-to-one constraint and introduces a ℓ_1-norm penalization on the abundance vectors. It thus solves (2.15) where the associated regularization parameter λ_a is tuned using a grid-search strategy. These two methods use an augmented Lagrangian splitting algorithm to recover the abundance vectors. Additionally, these abundance vectors are subsequently used as input features of a multinomial logistic regression classifier. This classifier is linear and its combination with the SUNSAL-based unmixing algorithms yields a sequential counterpart of the proposed Cofact-CE method.

Besides, the proposed method has been also compared with the discriminative K-SVD (D-KSVD) method proposed in [ZL10]. The D-KSVD problem has strong similarities with the proposed cofactorization problem. Indeed, it corresponds to a ℓ_0-penalized representation learning and a classification with a quadratic loss. It aims at learning a dictionary suitable for the classification problem and performs a linear classification on the coding vectors. For this reason, the dictionary $\mathbf{M}$ is only used as an initialization for D-KSVD, while it remains fixed for the unmixing and proposed cofactorization methods. Similarly, the label consistent K-SVD (LC-KSVD) is also considered [JLD11]. This model has been proposed as an improvement of D-KSVD where an additional term ensures that the dictionary elements are class-specific. Hyperparameters of D-KSVD and LC-KSVD have been manually adjusted in order to get the best results. When implementing the PALM algorithm proposed in Section 2.3.5, the normalized regularization parameters in (2.28) have been fixed as $\tilde{\lambda}_0 = 100$, $\lambda_1 = \lambda_2 = 1$, $\lambda_a = \lambda_q = 0.1$ and $\tilde{\lambda}_c = 10^{-3}$. Finally, the number of clusters has been set to $K = 10$.

Figure-of-merits – Several metrics are computed to quantify the quality of the classification and unmixing tasks (see Appendix A for details). For classification, two widely-used metrics are used, namely Cohen's kappa and the averaged F1-score over all classes [CG08]. For unmixing, reconstruction error (RE) and root global mean squared error (RMSE) are

computed as follows

$$\begin{aligned} \mathrm{RE} &= \sqrt{\frac{1}{Pd}\left\|\mathbf{Y} - \mathbf{M}\hat{\mathbf{A}}\right\|_{\mathrm{F}}^2}, \\ \mathrm{RMSE}(\hat{\mathbf{A}}) &= \sqrt{\frac{1}{PR}\left\|\mathbf{A}_{\mathrm{true}} - \hat{\mathbf{A}}\right\|_{\mathrm{F}}^2} \end{aligned} \tag{2.29}$$

where $\mathbf{A}_{\mathrm{true}}$ and $\hat{\mathbf{A}}$ are the actual and estimated abundance matrices.

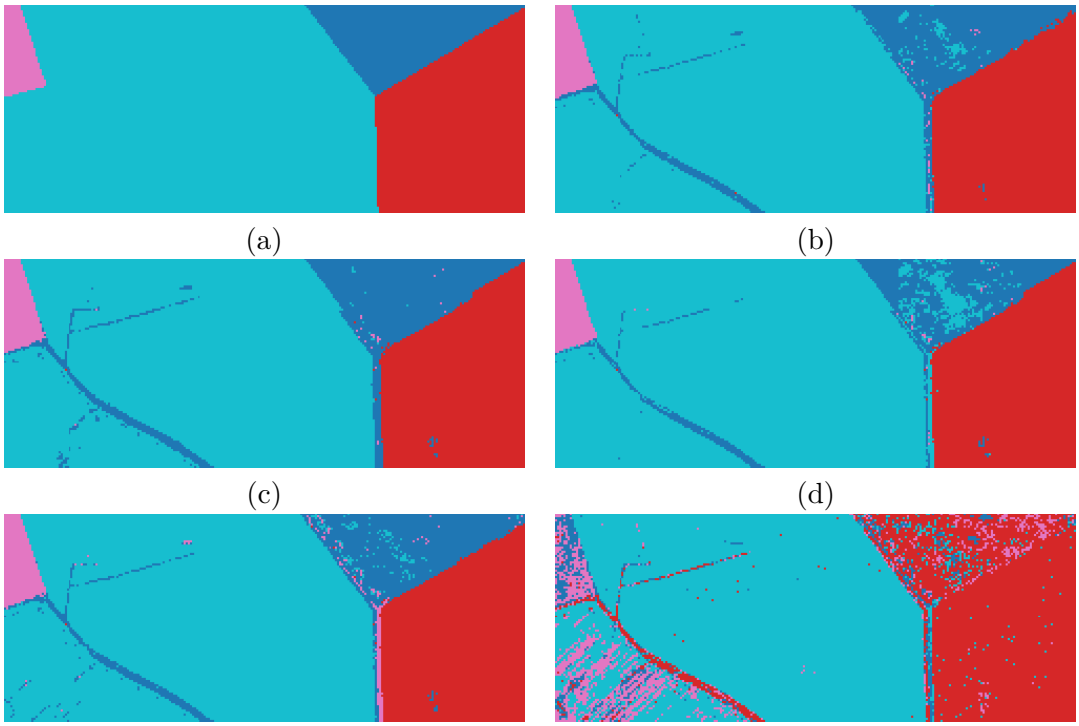


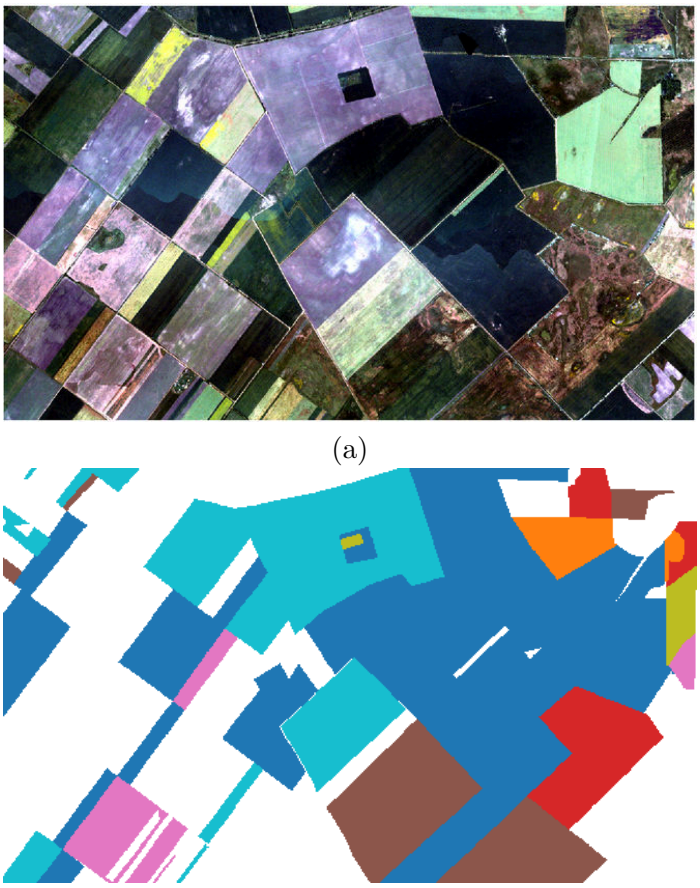
Figure 2.7.: Synthetic data, classification maps: (a) groundtruth, (b) RF, (c) Cofact-Q, (d) Cofact-CE, (e) LC-KSVD, (f) D-KSVD.

Performance evaluation – Quantitative results obtained on the synthetic dataset are reported in Table 2.2 and are visually depicted in Figures 2.7 and 2.6 for the classification and abundance maps, respectively. Metrics and their standard deviation have been computed over 20 trials. For each trial, a Gaussian white noise is added the observed image such that the SNR = 30 db. From these results, the proposed method appears to be competitive with the compared state-of-the-art methods. In term of classification results, even though

the spatial regularization is very weak in this setting, the cofactorization methods are as good as the RF classifier, which is very satisfying since this latter classifier is one of the most prominent one to deal with HS images. However, classification results of FC-SUNSAL and CSR-SUNSAL show that a classifier using abundance vectors can already perform well on this toy example where classes are linearly separable. As for LC-KSVD, it slightly performs worse regarding the F1-mean score whereas results of D-KSVD are clearly the worst. In term of unmixing performance, FC-SUNSAL, CSR-SUNSAL, Cofact-Q and Cofact-CE obtain very similar REs. Note however this metrics only evaluates the quality of the reconstructed data. However, the RMSE is lower with the cofactorization methods and the abundance estimations provided by FC-SUNSAL and CSR-SUNSAL significantly degrade. Even if it is not possible to produce a quantitative evaluation of the representation learnt by D-KSVD and LC-KSVD, REs tends to show that D-KSVD successfully estimated a representation of the data (without being easily interpretable) whereas LC-KSVD seems to focus mostly on the discriminative power of the representation at the price of an inaccurate representation. Moreover, the results produced by LC-KSVD have been obtained by increasing the dimension of the representation R to 40 while the results obtained by the other methods have been obtained for $R = 15$ to get good classification performances. The rather poor performance obtained by these two dictionary learning methods, when compared to the proposed cofactorization model, can be explained by the lack of flexibility of the corresponding models which try to recover a descriptive and discriminative representation simultaneously. On the contrary, some flexibility is offered by the clustering step included in the proposed method. Finally, comparison in term of processing times shows that D-KSVD, LC-KSVD and the proposed cofactorization methods are significantly slower, which is expected since these methods conducts representation learning and classification jointly. Nonetheless, the cofactorization methods appears faster than D-KSVD and LC-KSVD. It should be also noted that it is necessary to tune manually the number of iterations when using the two latter methods. Conversely, standard convergence criterion can be implemented for the proposed optimization-based methods.

2.5.3. Real hyperspectral image

Description of the dataset – The Aisa dataset was acquired by the AISA Eagle sensor during a flight campaign over Heves, Hungary. It contains $d = 252$ bands ranging from 395 to 975nm. A set of $C = 7$ classes have been defined for a total of $358,534$ referenced pixels, according to the class-wise repartition given in Table 2.3. To split the full dataset into two test and train subsets, special care has been taken to ensure that training samples are picked



(a)

(b)

Figure 2.8.: AISA dataset: (a) colored composition of the hyperspectral image $\mathbf{Y}$, (b) ground-truth [arable land: dark blue, forest: orange, grassland: red, fallowland: brown, leguminosae: pink, reed: green, row crops: light blue].

out from distinct areas than test samples. The polygons of the reference map are split in smaller polygons on a regular grid pattern and then 50% of the polygons are taken randomly for training and the remaining 50% for testing (see [LFG17] for a similar procedure). Figure 2.8 shows a colored composition of the image and the classification ground-truth. Several reasons justify the choice of this particular dataset. First, it is very challenging both in term of classification and unmixing mostly because the spectral signatures of the classes are very similar, leading in particular to very correlated endmember spectra in $\mathbf{M}$. Secondly, the ground-truth associated to this image is composed of two levels of classification. Thus, an additional ground-truth is available where the 7 considered classes have been subdivided into 14 classes also detailed in Table 2.3. These subclasses could be compared

Table 2.3.: AISA data: information about classes.

Class	Nb. of samples	Subclasses
Arable land	177,350	millet, rape, winter barley, winter wheat, oat
Forest	9,274	forest
Grassland	25,399	meadow, pasture
Green fallowland	44,370	fallow treated last year, fallow with shrubs
Leguminosae	17,628	leguminosae
Reed	4,776	reed
Row crops	79,737	maize, sunflowers

Table 2.4.: AISA data: unmixing and classification results.

Model	F1-mean	Kappa	RE	Time (s)
RF	0.711 ($\pm 1.4 \times 10^{-2}$)	0.835 ($\pm 1.2 \times 10^{-2}$)	N\A	41 ($\pm$1)
FC-SUNSAL	0.339 ($\pm 2.7 \times 10^{-2}$)	0.433 ($\pm 3.8 \times 10^{-2}$)	0.298 ($\pm 1.9 \times 10^{-3}$)	512 ($\pm$96)
CSR-SUNSAL	0.535 ($\pm 5.0 \times 10^{-2}$)	0.618 ($\pm 8.0 \times 10^{-2}$)	0.304 ($\pm 2.0 \times 10^{-5}$)	529 ($\pm$61)
D-KSVD	0.224 ($\pm 2.1 \times 10^{-2}$)	0.406 ($\pm 9.9 \times 10^{-2}$)	0.303 ($\pm 7.6 \times 10^{-6}$)	10475 ($\pm$129)
LC-KSVD	0.350 ($\pm 3.1 \times 10^{-2}$)	0.594 ($\pm 3.0 \times 10^{-2}$)	0.303 ($\pm 4.0 \times 10^{-6}$)	3780 ($\pm$320)
Cofact-Q	0.503 ($\pm 4.7 \times 10^{-2}$)	0.652 ($\pm 2.5 \times 10^{-2}$)	0.310 ($\pm 1.6 \times 10^{-4}$)	7303 ($\pm$139)
Cofact-CE	0.697 ($\pm 4.5 \times 10^{-2}$)	0.759 ($\pm 3.5 \times 10^{-2}$)	0.310 ($\pm 1.4 \times 10^{-4}$)	4382 ($\pm$257)

to the clustering outputs obtained by the proposed cofactorization method, *e.g.*, to verify either the clusters are consistent with the underlying subclasses.

Compared methods – The proposed algorithm is compared to the same methods introduced above. However, note that the D-KSVD method has experienced some difficulties to scale with the size of this new dataset, which is significantly bigger. Thus to obtain results in a decent amount of time, the algorithm has been interrupted prematurely; *i.e.*, before convergence. For the proposed cofactorization method, regularization parameters have been set to $\tilde{\lambda}_0 = \tilde{\lambda}_1 = \tilde{\lambda}_2 = \tilde{\lambda}_c = 1.$ and $\tilde{\lambda}_a = \tilde{\lambda}_q = 0.01$ and the number of clusters to $K = 30$. The initialization step described in Section 2.5.1 has been performed and the resulting dictionary $\mathbf{M}$ is depicted in Figure 2.9 ($R = 13$). The same dictionary has been used for the compared unmixing methods.

Performance evaluation – All quantitative results are presented in Table 2.4. Metrics

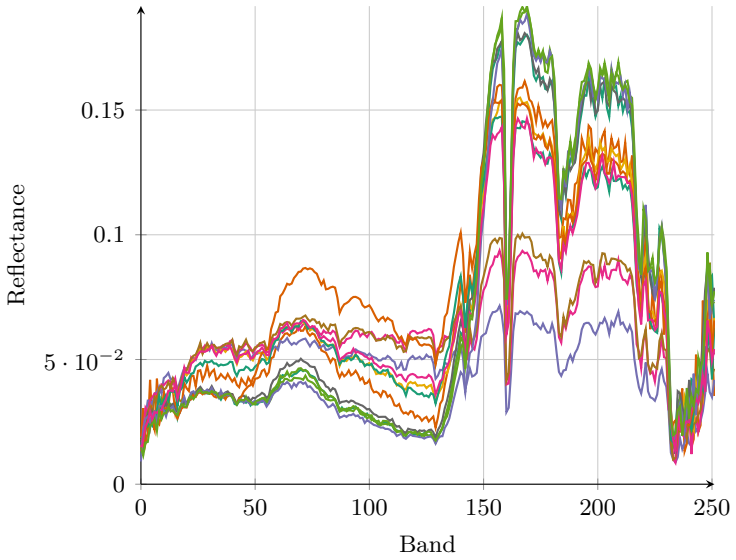


Figure 2.9.: AISA data: spectra used as the dictionary **M** identified by the self-dictionary method.

and their standard deviation have been computed over 5 trials. RMSE metrics have been removed since no groundtruth is available to assess the quality of the estimated abundance maps. RE is thus the only used figure-of-merit to assess the quality of the representation learning. Note however, as previously explained, RE does not directly evaluate the correctness of the abundance maps. In the present case, REs appear to be very similar for all algorithms. Contrary to the previous dataset, this is also the case for LC-KSVD, which can be explained by the fact that spectra are similar in the whole image and it is thus quite easy to get a very low RE with any estimated dictionary. This is the reason why qualitative evaluation remains interesting. Figure 2.11 shows a subset of the estimated abundance maps. It is difficult to draw any incontestable conclusion but it is clear that, despite similar REs, significantly different result are obtained for each method. This behavior is strengthened by the very high correlation between the endmembers in this dataset, which may lead to probable mismatch between endmember spectra. Nevertheless the Cofact methods seems to give slightly more consistent results. Indeed, edges in the abundance maps appear to be more consistent with boundaries observed in the hyperspectral image. Additionally, for the compared methods, some abundance maps seem to be influenced by the presence of two flight lines in the image. This phenomenon clearly appears in the abundance maps recovered by FC-SUNSAL (3rd row).

Concerning classification results, the results reported in Table 2.4 show that the classifica-

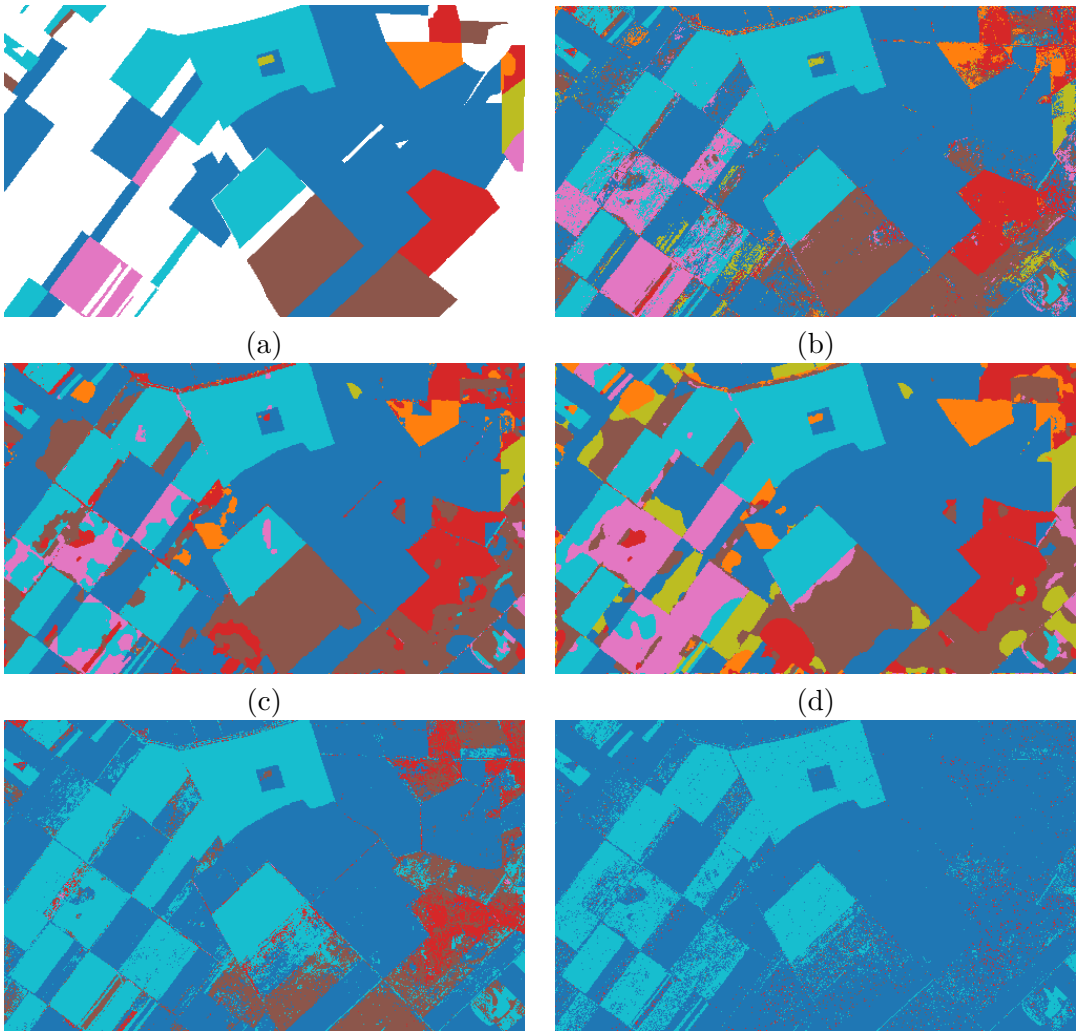


Figure 2.10.: AISA image, classification maps:(a) groundtruth, (b) RF, (c) Cofact-Q, (d) Cofact-CE, (e) LC-KSVD, (f) D-KSVD.

tion maps recovered by the Cofact-CE is very closed to the one obtained by RF. Figure 2.10 shows in particular that the cofactorization methods encounter some trouble distinguishing very similar classes, for example *grassland* (red) from *fallowland* (brown). Nevertheless, the obtained classification appears to be consistent and it seems reasonable to expect a lesser degradation of the classification results when considering less correlated spectral signatures. This confusion explains the less convicing results of the proposed method with quadratic loss. The results also show that the proposed method is beneficial to the classification since

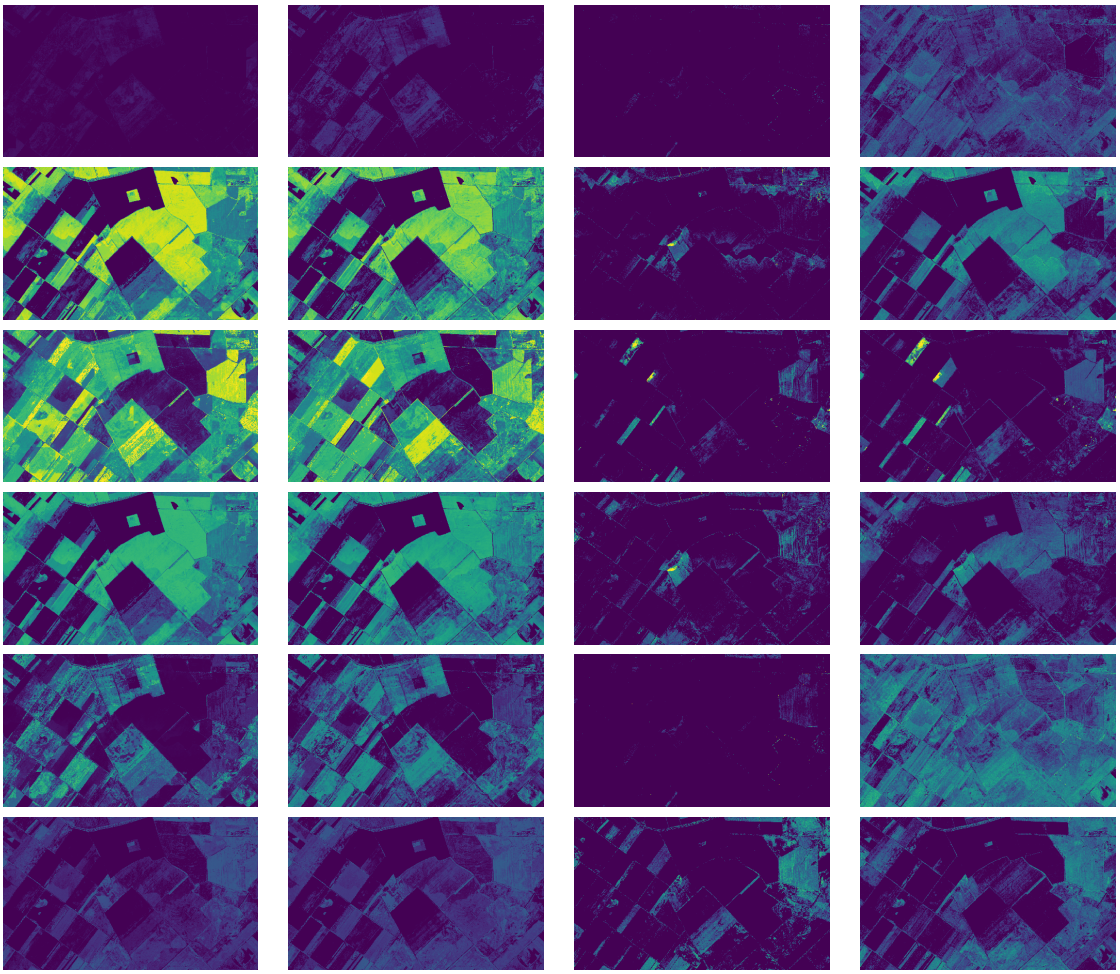


Figure 2.11.: AISA dataset, abundance maps of the 6 components: (1st column) Cofact-Q, (2nd column) Cofact-CE, (3rd column) FC-SUNSAL and (4th column) CSR-SUNSAL.

FC-SUNSAL, CSR-SUNSAL and Cofact-CE use the same classifier and the latter performs clearly better. The comparison between the representation learning-based algorithms is clear and the both Cofact methods perform better than LC-KSVD and D-KSVD.

In term of processing time, LC-KSVD, D-KSVD and the Cofact methods are clearly more time consuming. Nevertheless, all those methods provide more outputs than the other methods. The comparison between these methods seems to give an advantage for LC-KSVD. However, it should be noted that it is very difficult to monitor the convergence of LC-KSVD and D-KSVD since the value of the objective function over the iteration is not monotonic. The proposed algorithms and their implementations thus give a practical advantage since they do not need to be applied with different numbers of iterations to ensure good results.

One of very interesting feature of the Cofact method is the possibility of examining the clusters obtained as a byproduct. Given the formulation (2.23), the centroids $\mathbf{B}$ estimated by the Cofact method can be interpreted as average behaviors of abundance vectors. Corresponding virtual spectral signatures can be obtained by right-multiplying the dictionary $\mathbf{M}$ by this estimated abundance-like matrix $\mathbf{B}$. The first line in Figure 2.12 shows these spectral centroids for each cluster. Accessing this kind of information is precious in term of image interpretation since it offers the possibility of visualizing any class multi-modality. To illustrate, the second line of Figure 2.12 shows the mean spectra associated with the subclass groundtruth. Clearly, both lines exhibit strong similarities, with spectral diversity (hence multi-modality) for the 1st, 3rd and 4rd classes. This illustrates the relevance of the clusters recovered by the proposed cofactorization method.

2.6. Conclusion and perspectives

This chapter proposed a cofactorization model to unify a representation learning task and a classification task. The coding matrices associated with the two factorization problems, which respectively are the low-dimensional representations and the feature vectors, were related thanks to a clustering step. The low-dimensional representation vectors were clustered and the resulting attribution vectors were used as features vectors. These three tasks were jointly formulated as a non-convex non-smooth minimization problem, whose solution was approximated thanks to a PALM algorithm which ensured some convergence guarantees.

This model was instanced for a specific applicative scenario, namely hyperspectral image analysis, to jointly conduct unmixing and classification. It provided convincing results on synthetic and real data both quantitatively and qualitatively. Moreover, byproducts of the model appeared to be a relevant added value to interpret the obtained results.

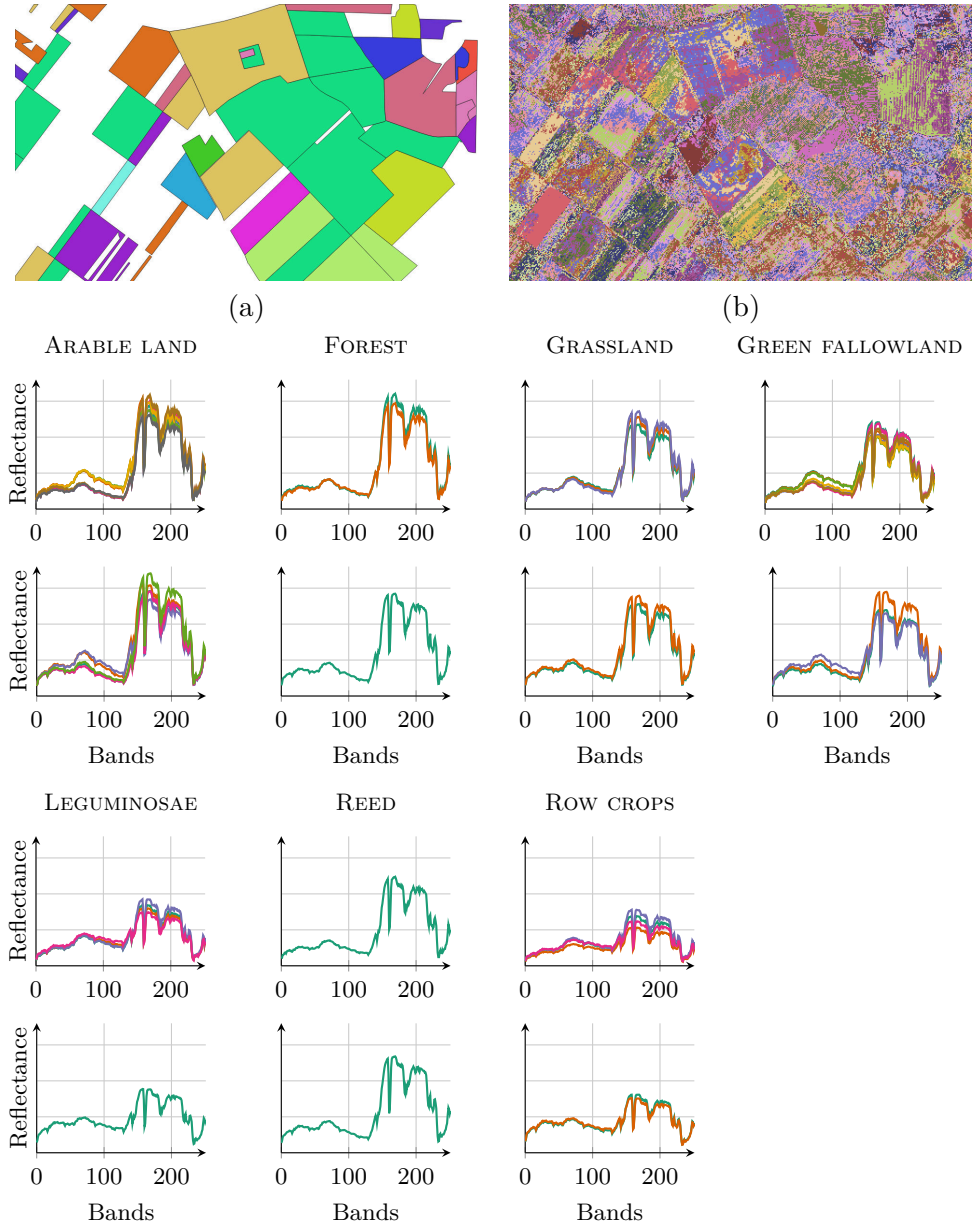


Figure 2.12.: AISA data. (a) Groundtruth map of subclasses. (b) Clustering recovered by Cofact-CE. For each class: (top) spectral centroids recovered by Cofact-CE, (bottom) mean spectra of all subclasses of the corresponding class evaluated using groundtruth.

To further improve the developed model, it would be particularly interesting to investigate the best way to learn an appropriate dictionary. Firstly, it would be relevant to directly exploit the supervised information to get a better dictionary initialization. Secondly, updating the dictionary when solving the cofactorization problem would be also of interest.

2.7. Conclusion (in French)

Ce chapitre introduit un modèle de cofactorisation permettant de réaliser conjointement un apprentissage de représentation et une classification. Les matrices de codage associées aux deux problèmes de factorisation, contenant respectivement les représentations en faible dimension et les vecteurs de descripteurs, sont reliées grâce à une étape de clustering. Ce clustering est effectué sur les représentations en faible dimension et les vecteurs d'attribution aux clusters sont ensuite utilisés comme vecteurs de descripteurs. Ces trois tâches ont été formulées conjointement sous la forme d'un problème de minimisation non-convexe et non-lisse dont la solution est approximée à l'aide d'un algorithme PALM assurant certaines garanties de convergence.

Ce modèle a été particularisé pour un scénario applicatif, à savoir l'analyse d'images hyperspectrales, permettant ainsi de réaliser conjointement démélange spectral et classification. Des résultats quantitatifs et qualitatifs convaincants ont été obtenus sur données synthétiques puis réelles. De plus, les produits annexes du modèle permettent une interprétation détaillée des résultats.

Pour améliorer le modèle développé, il apparaît particulièrement intéressant d'envisager une meilleure manière de réaliser l'apprentissage du dictionnaire. Premièrement, il serait certainement bénéfique d'exploiter directement l'information supervisée pour obtenir une meilleure initialisation du dictionnaire. Puis dans un second temps, la mise à jour du dictionnaire au cours de la résolution du problème d'optimisation semble la voie à suivre.

Chapter 3.

Matrix cofactorization for spatial and spectral unmixing

This chapter has been adapted from paper [Lag+19e].

Contents

3.1. Introduction (in French)

Ce chapitre se concentre sur le problème d'apprentissage de représentation et en particulier sur la possibilité d'enrichir le modèle à l'aide de données externes. Dans les chapitres précédents, le problème de l'apprentissage du dictionnaire a été écarté pour se concentrer sur la mise en place d'un modèle permettant l'apprentissage de représentation et la classification conjoints. Afin de traiter le problème dans sa globalité, ce point crucial du problème d'apprentissage de représentation est en particulier abordé dans ce dernier chapitre. Afin de proposer une approche très concrète, les contributions de ce chapitre sont directement exprimées dans le cadre de notre cas d'étude, c'est-à-dire, l'imagerie hyperspectrale.

Sachant que les images hyperspectrales contiennent une information spectrale riche, de nombreuses méthodes de démélange se concentrent sur l'idée d'exploiter au mieux cette information et négligent souvent l'information spatiale disponible. Un grand nombre des méthodes les plus reconnues traitent les pixels sans tenir compte de l'idée de base selon laquelle les pixels voisins sont souvent très similaires. La seule information partagée entre pixels est alors la matrice de endmembers [BF10 ; TDT15]. Néanmoins, plusieurs méthodes ont déjà été proposées pour effectuer un démélange spatial-spectral [SW14]. L'approche la plus classique consiste à envisager une régularisation spatiale locale des cartes d'abondances. Plusieurs travaux, tels que SUNSAL-TV [IBP12] ou S2WSU [Zha+18a], ont proposé d'utiliser une régularisation en norme TV comme régularisation spatiale. L'identification de groupes de pixels spectralement similaires, dispersés en petits clusters, a également été utilisée pour imposer un lissage spatial des abondances, par exemple dans [Wan+17 ; EDT11 ; Ech+13]. Avec une approche différente, d'autres travaux ont utilisé un voisinage local pour identifier le sous-ensemble de endmembers présents dans le voisinage. Cette approche a un intérêt en particulier dans le cas où on considère un grand nombre de endmembers [Can+11 ; DW13]. Enfin, dans une moindre mesure, l'information spatiale a également été utilisée pour faciliter l'extraction des endmembers. En effet, l'extraction des endmembers est souvent effectuée avant d'estimer les vecteurs d'abondance. Certains prétraitements ont été proposés pour faciliter cette extraction ou l'identification de pixels purs, tels que prendre la moyenne des spectres sur des superpixels [Tho+10] ou calculer des indicateurs d'homogénéité spatiale [ZP09].

Dans l'ensemble, il est intéressant de noter que toutes ces approches visent à exploiter l'idée très simple selon laquelle des pixels voisins sont similaires et ont des variables latentes similaires. Cependant, l'information spatiale est plus riche que cette simple idée. Par exemple, deux pixels très similaires du point de vue spectral peuvent être discriminés en

utilisant leur contexte, *e.g.*, une prairie naturelle et une culture sont très proches d'un point de vue spectral, mais la prairie est spatialement homogène alors que la culture est organisée en rangées. L'exploitation de modèles spatiaux et de descripteurs de textures devrait donc aider le processus de démélange. Pour exploiter cette idée, ce chapitre propose un modèle basé sur une approche par apprentissage de dictionnaires couplés permettant d'inférer conjointement des signatures spatiales et spectrales caractéristiques.

Des méthodes de cofactorisation, parfois appelées apprentissage de dictionnaires couplés, ont été utilisées avec succès dans de nombreux domaines, tels que la fouille de texte [WB11], la séparation de source en musique [Yoo+10] ou encore l'analyse d'image [YYI12 ; AM18]. L'idée principale est de définir un problème d'optimisation reposant sur deux modèles de factorisation, complétés par un terme de couplage imposant une dépendance entre les deux modèles. La méthode proposée dans ce chapitre, appelée SP2U pour *spatial-spectral unmixing*, considère conjointement un modèle de démélange spectral et une décomposition de descripteurs contextuelles calculées à partir de l'image panchromatique de la scène. Le terme de couplage s'interprète comme un clustering identifiant des groupes de pixels partageant des signatures spectrales et des contextes spatiaux similaires. Cette méthode présente deux avantages majeurs : *i)* elle fournit des résultats très compétitifs bien qu'elle soit non supervisée (c'est-à-dire qu'elle estime les endmembers et les cartes d'abondance) et *ii)* elle fournit des résultats très complets et pertinents car la scène se retrouve divisée en zones caractérisées par leurs signatures spectrales et spatiales.

Le reste du chapitre s'organise de la manière suivante. La section 3.3 définit les modèles spectral et spatial puis introduit le problème de cofactorisation. La section 3.4 détaille ensuite le schéma d'optimisation développé pour résoudre le problème de minimisation non-convexe et non-lisse qui en résulte. Une évaluation du modèle proposé est ensuite effectuée sur des données synthétiques dans la section 3.5, puis sur des données réelles dans la section 3.6. Enfin, la section 3.7 conclut ce chapitre et présente quelques perspectives de recherche pour ce travail.

3.2. Introduction

In this chapter, the focus is on the problem of representation learning problem and in particular on the possibility to enrich the model using exogenous data. In the previous chapters, the problem of learning a relevant dictionary has been left aside and, now that a method for joint representation learning and classification has been proposed, this key issue needs to be addressed. In order to be very concrete, the developments are directly presented

in the context of hyperspectral images.

As hyperspectral images contain a rich spectral information, many unmixing methods focus on exploiting it and often neglect spatial information. Many well-established methods process pixels without taking in consideration the basic idea that neighboring pixels are often very similar. The only shared information between pixels is a common endmember matrix [BF10; TDT15]. Nevertheless, advanced methods have been proposed to perform spatial-spectral unmixing [SW14]. The most direct approach is to consider local spatial regularization of the abundance maps. Several works, such as SUnSAL-TV [IBP12] or S2WSU [Zha+18a], proposed to use TV-norm regularization to achieve this goal. Identification of clusters of spectrally similar pixels, scattered in small groups, was also used to impose spatial smoothing of the abundances, e.g., in [Wan+17; EDT11; Ech+13]. In a different way, other works used the local neighborhood to identify the subset of endmembers present in the neighborhood. It is especially useful when dealing with a large number of endmembers [Can+11; DW13]. Finally, at a lesser extent, the spatial information has also been used to help the extraction of endmembers. Indeed, endmembers extraction is often performed before estimating the abundance vectors. Some preprocessing were proposed to ease the extraction and identification of pure pixels as the averaging of spectra over superpixels [Tho+10] or the use of spatial homogeneity scalar factors [ZP09].

Overall it is noticeable that all these approaches tend to exploit the very simple idea that neighboring pixels should be similar. However, spatial information is richer than this simple statement. For example, two spectrally very similar pixels can be discriminated using their context, *e.g.* a natural grassland and a crop field are spectrally very closed but the first is spatially homogeneous when the second in organized in rows. Exploiting spatial patterns and textures descriptors is thus expected to be helpful to the unmixing process. To exploit this assumption, this chapter proposes a model based on a cofactorization task to jointly infer common spatial and spectral signatures from the image.

Cofactorization methods, sometimes referred to as coupled dictionary learning, have been implemented with success in many application fields, *e.g.*, for text mining [WB11], music source separation [Yoo+10] and image analysis [YYI12; AM18], among others. The main idea is to define an optimization problem relying on two factorizing models supplemented by a coupling term enforcing a dependence between the two models. The method proposed in this chapter, called SP2U for spatial-spectral unmixing, jointly considers a spectral unmixing model and a decomposition of contextual features computed from the panchromatic image of the same scene. The coupling term is interpreted as a clustering identifying groups of pixels sharing similar spectral signatures and spatial contexts. This method exhibits two major

advantages: *i)* it provides very competitive results even though the method is unsupervised (i.e., it estimates both endmember signatures and abundance maps) and *ii)* it provides very insightful results since the scene is partitioned into areas characterized by spectral and spatial signatures.

The remaining of the chapter is organized as follows. Section 3.3 defines the spectral and the spatial models and further discusses the joint cofactorization problem. Section 3.4 then details the optimization scheme developed to solve the resulting non-convex non-smooth minimization problem. An evaluation of the proposed joint model is then conducted first on synthetic data in Section 3.5 and then on real data in Section 3.6. Finally, Section 3.7 concludes the chapter and presents some research perspectives to this work.

3.3. Towards spatial-spectral unmixing

The main goal of this section is to introduce a model capable of spectrally and spatially characterizing an hyperspectral image. In particular, instead of incorporating prior spatial information as a regularization [IBP12], the concept of spatial unmixing, detailed in Section 3.3.2, is introduced alongside a conventional spectral unmixing model in order to propose a new joint framework of spatial-spectral unmixing.

3.3.1. Spectral mixture model

Spectral unmixing aims at identifying the elementary spectra and the proportion of each material in a given pixel [Bio+12]. Each of the P pixels $\mathbf{y}_p$ is a d_1-dimensional measurement of a reflectance spectrum and is assumed to be a combination of R_1 elementary spectra $\mathbf{m}_r$, called endmembers, with $R_1 \ll d_1$. The so-called abundance vector $\mathbf{a}_p \in \mathbb{R}^{R_1}$ refers to the corresponding mixing coefficients in this pixel. In a general case, where no particular assumption is made on the observed scene, the conventional linear mixture model (LMM) is widely adopted to describe the mixing process. It assumes that the observed mixtures are linear combinations of the endmembers. Within an unsupervised framework, *i.e.*, when both endmember signatures and abundances should be recovered, linear spectral unmixing can be formulated as the following minimization problem

$$\min_{\mathbf{M},\mathbf{A}} \|\mathbf{Y} - \mathbf{M}\mathbf{A}\|_{\mathrm{F}}^2 + \imath_{\mathbb{R}_+^{d_1 \times R_1}}(\mathbf{M}) + \imath_{\mathbb{S}_{R_1}^P}(\mathbf{A}) \tag{3.1}$$

where the matrices $\mathbf{Y} \in \mathbb{R}^{d_1 \times P}$ gathers all the observed pixels, $\mathbf{M} \in \mathbb{R}^{d_1 \times R_1}$ the endmembers, $\mathbf{A} \in \mathbb{R}^{R_1 \times P}$ the abundance vectors and $\imath_{\mathbb{R}_+^{d_1 \times R_1}}(\cdot)$ and $\imath_{\mathbb{S}_{R_1}^P}(\cdot)$ are respectively indi-

cator functions on the non-negative quadrant and the column-wise indicator function on the R_1-dimensional probability simplex denoted by $\mathbb{S}_{R_1}$. The non-negative constraint over $\mathbf{M}$ is justified by the fact that endmember signatures are reflectance spectra and thus non-negative. The second indicator function enforces non-negative and sum-to-one constraints on the abundance vectors $\mathbf{a}_p$ $(p = 1, \ldots, P)$ in order to interpret them as proportion vectors. It is worth noting that the sum-to-one constraint is sometimes disregarded since it has been argued that relaxing this constraint out offers a better adaptation to possible changes of illumination in the scene [Dru+16]. Due to the general ill-conditioning of the endmember matrix $\mathbf{M}$, the objective function underlying (3.1) is often granted with additional regularizations promoting expected properties of the solution. In particular, numerous works exploited the expected spatial behavior of the mixing coefficients to introduce spatial regularizations enforcing piecewise-constant [EDT11; IBP12] or smoothly varying [TDT15; MIC12] abundance maps, possibly driven by external knowledge [UFD18]. Conversely, this work does not consider spatial information as a prior knowledge but rather proposes a decomposition model dedicated to the image spatial content, paving the way towards the concept of *spatial unmixing*. This contribution is detailed in what follows.

3.3.2. Spatial mixing model

As previously mentioned, this chapter proposes to complement the conventional linear unmixing problem (3.1) with an additional data-fitting term accounting for spatial information already contained in the hyperspectral image. To do so, for sake of generality, we assume that the scene of interest is characterized by vectors of spatial features $\mathbf{s}_p \in \mathbb{R}^{d_2}$ describing the context around the corresponding hyperspectral pixel indexed by p. The features can be extracted from the hyperspectral image directly or from any other available image of any modality of the same scene, with possibly better spatial resolution. For instance, one possibility consists in generating a virtual panchromatic image associated with the scene by averaging the hyperspectral bands and defining the features as the panchromatic pseudo-observations in a prescribed neighborhood. As a proof-of-concept but without limitation, this is the approach followed in Sections 3.5 and 3.6 dedicated to numerical experiments.

To capture common spatial patterns, akin to a so-called *spatial unmixing*, these P d_2-dimensional spatial features vectors $\mathbf{s}_p$ gathered in a matrix $\mathbf{S} \in \mathbb{R}^{d_2 \times P}$ are assumed to be linearly decomposed according to the following optimization problem

$$\min_{\mathbf{D},\mathbf{U}} \|\mathbf{S} - \mathbf{D}\mathbf{U}\|_{\mathrm{F}}^2 + \imath_{\mathbb{R}_+^{d_2 \times R_2}}(\mathbf{D}) + \imath_{\mathbb{S}_{R_2}^P}(\mathbf{U}) \tag{3.2}$$

where $\mathbf{D} \in \mathbb{R}^{d_2 \times R_2}$ is a dictionary matrix and $\mathbf{U} \in \mathbb{R}^{R_2 \times P}$ the corresponding coding matrix.

This model can be interpreted as a dictionary-based representation learning task [AEB06]. It means that the image in the considered feature space can be decomposed as a sum of elementary patterns collected in the matrix $\mathbf{D}$ of spatial signatures. The corresponding coding coefficients are gathered in $\mathbf{U}$. The non-negativity constraints are imposed to ensure an additive decomposition similarly to what is done in the context of non-negative matrix factorization [LS99]. Finally, without any constraint on the norms of $\mathbf{U}$ and $\mathbf{D}$, the problem would suffer from a scaling ambiguity between $\mathbf{U}$ and $\mathbf{D}$. Additional sum-to-one constraints are thus imposed on the columns of $\mathbf{U}$. It is worth noting that a similar model was implicitly assumed in [Vas+15; Vas+16; Vas+18] where a single-band image acquired by scanning transmission electron microscopy is linearly unmixed by principal component analysis [Jol86], independent component analysis [AJE01], N-FINDR [Win99] or thanks to a deep convolutional neural networks. However, in these works, the spatial feature space is defined by the magnitude of a sliding 2D-discrete Fourier transform, which unlikely ensures the additivity, or at least linear separability, assumptions underlying the mixtures.

3.3.3. Coupling spatial and spectral mixing models

After defining the spatial and spectral mixing models, we propose to relate both models by a coupling term, ensuring a joint spatial-spectral unmixing of the hyperspectral image. In this work, the coupling term is chosen such that it links the two coding matrices $\mathbf{A}$ and $\mathbf{U}$, corresponding to the *spectral* and *spatial* abundances, respectively. More precisely, the coupling is formulated as the following penalized least-square problem

$$\min_{\mathbf{B},\mathbf{Z}} \left\| \begin{pmatrix} \mathbf{A} \\ \mathbf{U} \end{pmatrix} - \mathbf{B}\mathbf{Z} \right\|_{\mathrm{F}}^2 + \frac{\lambda_z}{2} \mathrm{Tr}(\mathbf{Z}^T \mathbf{V} \mathbf{Z}) + \imath_{\mathbb{S}_K^P}(\mathbf{Z}) \tag{3.3}$$

with $\mathbf{V} = \mathbf{1}_K \mathbf{1}_K^T - \mathbf{I}_K$ where $\mathbf{I}_K$ is the $K \times K$ identity matrix, $\mathbf{1}_K$ is the $K \times 1$ vector of ones and $\mathrm{Tr}(\cdot)$ is the trace operator. This coupling term can be interpreted as a clustering task. The two coding matrices are concatenated and the clustering is then conducted on the columns of the resulting whole coding matrix. Centroids of the K clusters define the columns of the matrix $\mathbf{B} \in \mathbb{R}^{(R_1+R_2)\times K}$. Interestingly, each centroid is then the concatenation of a spatial signature and a spectral signature. In particular, it means that the pixels of a given cluster share the same spectral properties and a similar spatial context. Finally, the matrix $\mathbf{Z} \in \mathbb{R}^{K \times P}$ describes the assignments to the clusters, where $\mathbf{z}_p$ gathers the probabilities of belonging to each of the clusters, hence the non-negativity and sum-to-one constraint

enforced on it. It is accompanied with a specific regularization (see 2nd term of the right-hand side of (3.3)). This penalty promotes orthogonality over the lines of $\mathbf{Z}$ since it can be rewritten as $\mathrm{Tr}(\mathbf{Z}^T\mathbf{V}\mathbf{Z}) = \sum_{k_1 \neq k_2} \langle \mathbf{z}_{k_1,:} | \mathbf{z}_{k_2,:} \rangle$. This term becomes minimum when the assignments to clusters obey a hard decision, *i.e.*, when one component of $\mathbf{z}_p$ is equal to 1 and the others are set to 0. A strict orthogonality constraint would make the clustering problem equivalent to a k-means problem [Pom+14].

3.3.4. Joint spatial-spectral unmixing problem

Given the spectral mixing model recalled in Section 3.3.1, the spatial mixing model introduced in Section 3.3.2 and their coupling term proposed in Section 3.3.3, we propose to conduct spatial-spectral unmixing jointly by considering the overall minimization problem

$$\begin{aligned}\min_{\mathbf{M},\mathbf{A},\mathbf{D},\mathbf{U},\mathbf{B},\mathbf{Z}} & \frac{\lambda_0}{2}\left\|\mathbf{Y}-\mathbf{M}\mathbf{A}\right\|_{\mathrm{F}}^2 + \imath_{\mathbb{R}_+^{d_1\times R_1}}(\mathbf{M}) + \imath_{\mathbb{S}_{R_1}^P}(\mathbf{A}) \\ & + \frac{\lambda_1}{2}\left\|\mathbf{S}-\mathbf{D}\mathbf{U}\right\|_{\mathrm{F}}^2 + \imath_{\mathbb{R}_+^{d_2\times R_2}}(\mathbf{D}) + \imath_{\mathbb{S}_{R_2}^P}(\mathbf{U}) \\ & + \frac{\lambda_2}{2}\left\|\begin{pmatrix}\mathbf{A}\\ \mathbf{U}\end{pmatrix} - \mathbf{B}\mathbf{Z}\right\|_{\mathrm{F}}^2 + \frac{\lambda_z}{2}\mathrm{Tr}(\mathbf{Z}^T\mathbf{V}\mathbf{Z}) \\ & + \imath_{\mathbb{R}_+^{(R_1+R_2)\times K}}(\mathbf{B}) + \imath_{\mathbb{S}_K^P}(\mathbf{Z}) \end{aligned} \tag{3.4}$$

where λ_0, λ_1 and λ_2 adjust the respective contribution of the various fitting terms. It is worth noting that, thanks to the sum-to-one constraints enforced on the spectral abundance vectors $\mathbf{a}_p$ and spatial abundance vectors $\mathbf{u}_p$, all these coding vectors have the same unitary ℓ_1-norm. It has the great advantage of avoiding a reweighing of the $\mathbf{A}$ and $\mathbf{U}$ in the coupling term regardless of the number of endmembers and dictionary atoms. However, it is still necessary to adjust the three parameters $\lambda_\cdot$ to weigh the various contribution terms. The strategy used in the experimental sections is to simply ensure that all terms as a similar weight by taking into account the size and dynamic of the involved matrices. The next section describes the optimization scheme adopted to solve the joint spatial-spectral unmixing problem (3.4),

3.4. Optimization scheme

3.4.1. PALM algorithm

The cofactorization problem (3.4) is a non-convex, non-smooth optimization problem. For these reasons, the problem remains very challenging to solve and requires the use of advanced optimization tools. The choice has been made to resort to the proximal alternating linearized minimization (PALM) algorithm [BST14]. The core concept of PALM is to update each block of variables alternatively according to a proximal gradient descent step. This algorithm has the advantage to ensure the converge to a critical point of the objective function even in the case of a non-convex, non-smooth problem.

In order to obtain these convergence results, the objective function has to ensure a specific set of properties. Firstly, the various terms of the objective function have to be separable in a sum of one smooth term $g(\cdot)$ and a set of independent non-smooth terms. Then, each of the independent non-smooth term has to be a proper, lower semi-continuous function $f_i : \mathbb{R}^{n_i} \to (-\infty, +\infty]$. Finally, a sufficient condition is that the smooth term is a $\mathcal{C}^2$-continuous function and that its partial gradients are globally Lipschitz with respect to the derivative variable. Further details are available in the original paper [BST14].

In problem (3.4), the smooth term $g(\cdot)$ is composed of the three quadratic terms and the orthogonality-promoting regularization. All these terms obviously verify the gradient Lipschitz and $\mathcal{C}^2$-continuous properties. Moreover, the non-smooth terms f_i are separable into independent terms. Moreover, since they are all indicators functions on convex sets, their proximal operators are well-defined and, more specifically, are defined as the projection on the corresponding convex set. The projection on the non-negative quadrant is a simple thresholding of the negative values and the projection on the probability simplex can be achieved by a simple sort followed by a thresholding as described in [Con16].

A summary of the overall optimization scheme is given in Algo. 3 where $L_{\mathbf{X}}$ stands for the Lipschitz constant of the gradient of $g(\cdot)$ considered as a function of $\mathbf{X}$. Partial gradients and Lipschitz moduli are all provided in Appendix C.1. Additional details regarding the implementation are discussed in what follows.

3.4.2. Implementation details

Initialization and convergence – As explained, the PALM algorithm only ensures convergence to a critical point, *i.e.*, a local minimum, of the objective function. Hence, it is important to have a good initialization of the variables to be estimated. In the following experiments, the initial endmember matrix $\mathbf{M}^0$ has been chosen as the output of the ver-

Algorithm 3: PALM

Initialize variables $\mathbf{M}^0$, $\mathbf{A}^0$, $\mathbf{D}^0$, $\mathbf{U}^0$, $\mathbf{B}^0$ and $\mathbf{Z}^0$;
Set $\alpha > 1$;
while *stopping criterion not reached* **do**

$$\mathbf{M}^{k+1} \in \operatorname{prox}^{\alpha L_{\mathbf{M}}}_{\iota_{\mathbb{R}^{d_1 \times R_1}_+}}\left(\mathbf{M}^k - \tfrac{1}{\alpha L_{\mathbf{M}}}\nabla_{\mathbf{M}} g(\mathbf{M}^k, \mathbf{A}^k, \mathbf{D}^k, \mathbf{U}^k, \mathbf{B}^k, \mathbf{Z}^k)\right);$$

$$\mathbf{A}^{k+1} \in \operatorname{prox}^{\alpha L_{\mathbf{A}}}_{\iota_{\mathbb{S}^P_{R_1}}}\left(\mathbf{A}^k - \tfrac{1}{\alpha L_{\mathbf{A}}}\nabla_{\mathbf{A}} g(\mathbf{M}^{k+1}, \mathbf{A}^k, \mathbf{D}^k, \mathbf{U}^k, \mathbf{B}^k, \mathbf{Z}^k)\right);$$

$$\mathbf{D}^{k+1} \in \operatorname{prox}^{\alpha L_{\mathbf{D}}}_{\iota_{\mathbb{R}^{d_2 \times R_2}_+}}\left(\mathbf{D}^k - \tfrac{1}{\alpha L_{\mathbf{D}}}\nabla_{\mathbf{D}} g(\mathbf{M}^{k+1}, \mathbf{A}^{k+1}, \mathbf{D}^k, \mathbf{U}^k, \mathbf{B}^k, \mathbf{Z}^k)\right);$$

$$\mathbf{U}^{k+1} \in \operatorname{prox}^{\alpha L_{\mathbf{U}}}_{\iota_{\mathbb{S}^P_{R_2}}}\left(\mathbf{U}^k - \tfrac{1}{\alpha L_{\mathbf{U}}}\nabla_{\mathbf{U}} g(\mathbf{M}^{k+1}, \mathbf{A}^{k+1}, \mathbf{D}^{k+1}, \mathbf{U}^k, \mathbf{B}^k, \mathbf{Z}^k)\right);$$

$$\mathbf{B}^{k+1} \in \operatorname{prox}^{\alpha L_{\mathbf{B}}}_{\iota_{\mathbb{R}^{(R_1+R_2)\times K}_+}}\left(\mathbf{B}^k - \tfrac{1}{\alpha L_{\mathbf{B}}}\nabla_{\mathbf{B}} g(\mathbf{M}^{k+1}, \mathbf{A}^{k+1}, \mathbf{D}^{k+1}, \mathbf{U}^{k+1}, \mathbf{B}^k, \mathbf{Z}^k)\right);$$

$$\mathbf{Z}^{k+1} \in \operatorname{prox}^{\alpha L_{\mathbf{Z}}}_{\iota_{\mathbb{S}^P_K}}\left(\mathbf{Z}^k - \tfrac{1}{\alpha L_{\mathbf{Z}}}\nabla_{\mathbf{Z}} g(\mathbf{M}^{k+1}, \mathbf{A}^{k+1}, \mathbf{D}^{k+1}, \mathbf{U}^{k+1}, \mathbf{B}^{k+1}, \mathbf{Z}^k)\right);$$

end
return $\mathbf{M}^{end}, \mathbf{A}^{end}, \mathbf{D}^{end}, \mathbf{U}^{end}, \mathbf{B}^{end}, \mathbf{Z}^{end}$

tex component analysis (VCA) [ND05]. Abundance matrix is then initialized by solving the fully constrained least square problem $\min_{\mathbf{A}\in\mathbb{S}^P_{R_1}} \|\mathbf{Y} - \mathbf{M}\mathbf{A}\|_{\mathrm{F}}^2$. Finally, $\mathbf{D}^0$ and $\mathbf{U}^0$ are initialized by performing a k-means algorithm on columns of $\mathbf{S}$. Similarly $\mathbf{B}^0$ and $\mathbf{Z}^0$ are initialized by a k-means on the concatenation of $\mathbf{U}^0$ and $\mathbf{A}^0$.

As stated in Algo. 3, a criterion is needed to monitor the convergence of the optimization algorithm. In the following experiments, the residual error of the objective function is computed at each iteration and, when the relative gap between the two last iterations is below a given threshold (10^{-4} for these experiments), the algorithm is stopped.

Hyperparameters – Several weighting coefficient $\lambda_\cdot$ have been introduced in problem (3.4) to adjust the respective contribution of each term. In the following experiments, some of these coefficients have been renormalized to take in consideration the respective dimensions and dynamics of the matrices defining each term, yielding

$$\begin{cases} \lambda_0 &= \frac{1}{d_1\|\mathbf{Y}\|_\infty^2}\tilde{\lambda}_0 \\ \lambda_1 &= \frac{1}{d_2\|\mathbf{S}\|_\infty^2}\tilde{\lambda}_1 \end{cases}. \tag{3.5}$$

3.5. Experiments using simulated data

Performance of the proposed spatial-spectral unmixing method has been assessed thanks to experiments conducted on both synthetic and real data. The use of synthetic data makes quantitative validation possible whereas it is not possible with real data since there is no

reference data.

3.5.1. Data generation

In order to properly evaluate the relevance of the proposed model, two synthetic images referred to as Image 1 and Image 2 have been generated such that they incorporate consistent spatial and spectral information. For this reason, the first step of the image synthesis consists in generating two so-called segmentation maps which separate the images into J regions. In this work, for each image, the segmentation maps has been randomly generated according to a Potts-Markov random field [Li09].

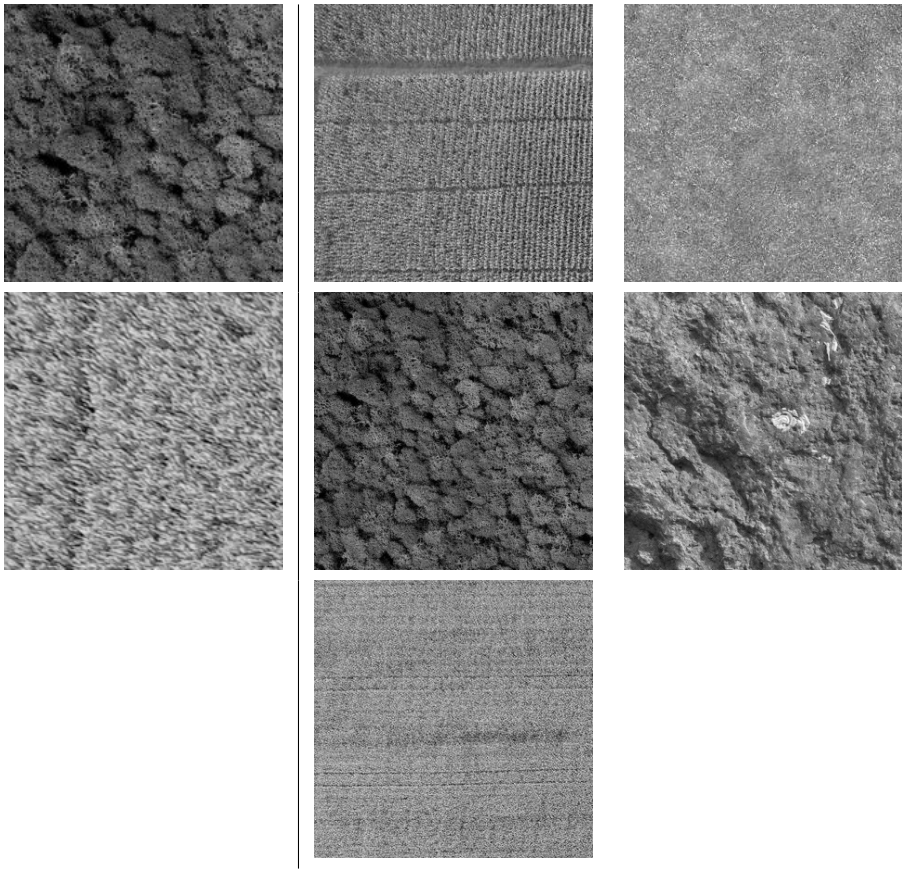


Figure 3.1.: Synthetic dataset: textures (forest, wheat) for Image 1 (left) and textures (corn, grass, forest, rock, wheat) for Image 2 (right).

The second step is to assign specific spatial and spectral signatures to each area of the segmentation map. In order to get realistic images, grayscale textures are extracted from real remote sensing images and a distinct texture is assigned to each cluster of the segmentation.

The textures are depicted in Fig. 3.1 for Image 1 and Image 2. Then, when the pth pixel belongs to the jth region ($j = 1, \ldots, J$), its spectral abundance vector has been generated as the convex combination of two predefined extremal spectral behaviors $\boldsymbol{\psi}_{i,1}$ and $\boldsymbol{\psi}_{i,2}$ characterizing the jth region, *i.e.*,

$$\mathbf{a}_p = t_p^{(j)} \boldsymbol{\psi}_{j,1} + (1 - t_p^{(j)}) \boldsymbol{\psi}_{j,2} \tag{3.6}$$

where $t_p^{(j)}$ is the intensity of the pth pixel of the jth grayscale texture. In other words, the texture intensity spatially modulates the spectral content differently in each region. The generated abundance maps are shown in Fig. 3.2.

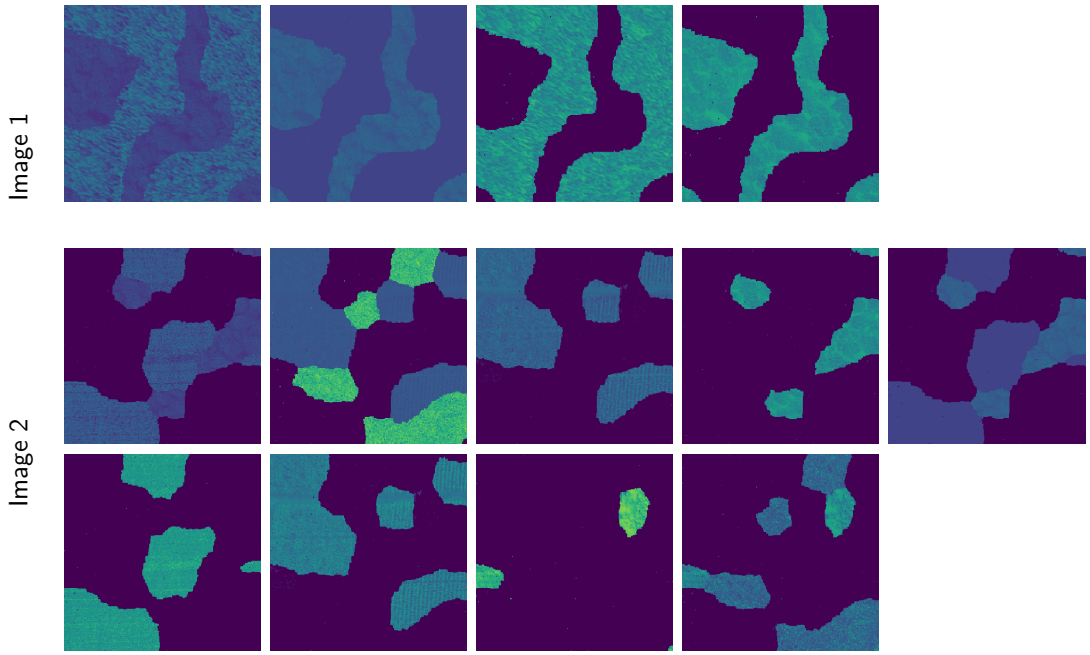


Figure 3.2.: Synthetic dataset: abundance maps.

The final step boils down to generating the hyperspectral image according to a linear mixing model. The endmember signatures have been extracted from the ASTER library. Two images have been generated according to this process. Image 1 is a 200×200-pixel image composed of $R_1 = 4$ endmembers and $J = 2$ regions. Image 2 is a 300×300-pixel image with $R_1 = 9$ endmembers and $J = 5$ regions. Additionally, corresponding panchromatic images are generated by normalizing and summing all spectral bands. The generated hyperspectral and panchromatic images are shown in Fig. 3.3.

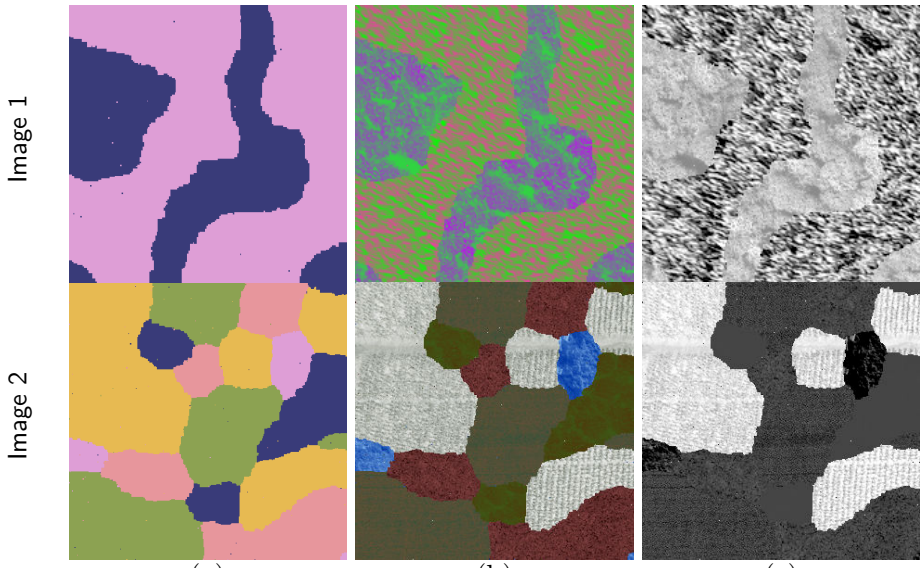


Figure 3.3.: Synthetic dataset: (a) segmentation map, (b) color composition of the hyperspectral image, (c) panchromatic image.

3.5.2. Compared methods

In order to assess the performance of the proposed spatial-spectral unmixing model, referred to as SP2U, the unmixing results have been compared to several well-established methods. First, the result of the initialization method has been used as baseline. This method is conventional [BF10] and consists in extracting endmembers using VCA method [ND05] and then solving a fully constrained least square (FCLS) problem. This first method is referred to as by VCA+FCLS hereafter.

The second compared method uses again a FCLS method to estimate the abundance vectors but uses an alternative endmember extraction algorithm. This method, called SISAL [Bio09], tries to estimate the minimum volume simplex containing the observed hyperspectral data by solving a non-convex problem using a splitting augmented Lagrangian technique.

The third compared method relies on a similar linear mixing model assumed by VCA+FCLS and SISAL+FCLS. However, instead of estimating the endmember signatures and abundances sequentially, it performs a joint estimation, yielding a non-negative matrix factorization (NMF) task with an additional sum-to-one constraint. This method referred to as NMF in the sequel, is a depreciated version of the SP2U problem (3.4) where $\lambda_1 = \lambda_2 = \lambda_z = 0$ and has been solved and initialized similarly.

The fourth method SUnSAL-TV was introduced in [IBP12] and proposes to solve a conventional linear unmixing problem with an additional spatial regularization term to incorporate spatial information. The regularization term is chosen as a total variation applied to the abundance maps $\mathbf{A}$. It promotes in particular similarity of abundance vectors of neighboring pixels. In this case, the local information is used whereas SP2U method relates pixels sharing the same spatial context, akin to a non-local framework. It is important to note that this method does not estimate the endmember matrix which is estimated beforehand using VCA or SISAL.

The fifth method, denoted n-SP2U, is a naive counterpart of the proposed SP2U method. Instead of using the coupling term introduced in Section 3.3.3, the abundance matrix $\mathbf{A}$ and the coding coefficients $\mathbf{U}$ are directly considered equal yielding the following problem

$$\begin{aligned}\min_{\mathbf{M},\mathbf{A},\mathbf{D}} \frac{\lambda_0}{2} \|\mathbf{Y} - \mathbf{M}\mathbf{A}\|_{\mathrm{F}}^2 + \imath_{\mathbb{R}_+^{d_1 \times R_1}}(\mathbf{M}) \\ + \frac{\lambda_1}{2} \|\mathbf{S} - \mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2 + \imath_{\mathbb{R}_+^{d_2 \times R_2}}(\mathbf{D}) + \imath_{\mathbb{S}_{R_1}^P}(\mathbf{A}).\end{aligned} \tag{3.7}$$

This method is considered for comparison since it may come naturally to mind when willing to couple factorizations associated with spatial and spectral unmixing. However, it actually appears very unlikely to perform well in real scenarios. It would mean that the mixture proportions are always similar in the spatial and spectral domains. However a given spectral signal is obviously expected to appear in various spatial contexts. To account for distinct spatial patterns of a given spectral content, some endmembers would need to appear several times in the $\mathbf{M}$ matrix, which is generally not a desired property.

3.5.3. Performance criteria

Performance of all methods has been assessed in term of endmember estimation using the average spectral angle mapper (aSAM)

$$\mathrm{aSAM}(\mathbf{M}) = \frac{1}{R_1} \sum_{r=1}^{R_1} \arccos\left(\frac{\langle \mathbf{m}_r^{(\mathrm{ref})} | \mathbf{m}_r \rangle}{\|\mathbf{m}_r^{(\mathrm{ref})}\|_2 \|\mathbf{m}_r\|_2} \right), \tag{3.8}$$

and also in term of abundance estimation using the root mean square error (RMSE)

$$\mathrm{RMSE}(\mathbf{A}) = \sqrt{\frac{1}{PR_1} \|\mathbf{A}^{(\mathrm{ref})} - \mathbf{A}\|_{\mathrm{F}}^2}, \tag{3.9}$$

Table 3.1.: Image 1: quantitative results of unmixing (averaged over 10 trials).

Model	aSAM($\mathbf{M}$)	RE	RMSE($\mathbf{A}$)	Time (s)
VCA+FCLS	$0.180\ (\pm 1.1 \times 10^{-2})$	$6.86 \times 10^{-3}\ (\pm 6.3 \times 10^{-3})$	$0.150\ (\pm 1.9 \times 10^{-2})$	**19** (± 11)
SISAL+FCLS	$0.151\ (\pm 3.4 \times 10^{-3})$	$\mathbf{2.81 \times 10^{-3}}\ (\pm 3.5 \times 10^{-6})$	$0.114\ (\pm 3.9 \times 10^{-3})$	23 (± 0.1)
NMF	$0.175\ (\pm 5.6 \times 10^{-3})$	$3.86 \times 10^{-3}\ (\pm 9.8 \times 10^{-4})$	$0.151\ (\pm 2.1 \times 10^{-2})$	27 (± 29)
VCA+SUnSAL-TV	$0.180\ (\pm 1.1 \times 10^{-2})$	$7.61 \times 10^{-3}\ (\pm 4.5 \times 10^{-3})$	$0.132\ (\pm 3.2 \times 10^{-2})$	27 (± 0.1)
SISAL+SUnSAL-TV	$0.151\ (\pm 2.9 \times 10^{-3})$	$4.6 \times 10^{-3}\ (\pm 1.1 \times 10^{-4})$	**0.0989** $(\pm 4.1 \times 10^{-3})$	28 (± 0.3)
n-SP2U	$0.188\ (\pm 1.5 \times 10^{-2})$	$28.1 \times 10^{-3}\ (\pm 1.2 \times 10^{-3})$	$0.192\ (\pm 9.6 \times 10^{-3})$	93 (± 14)
SP2U	**0.108** $(\pm 2.2 \times 10^{-2})$	$6.88 \times 10^{-3}\ (\pm 3.5 \times 10^{-4})$	$0.166\ (\pm 7.2 \times 10^{-2})$	409 (± 38)

Table 3.2.: Image 2: quantitative results of unmixing (averaged over 10 trials).

Model	aSAM($\mathbf{M}$)	RE	RMSE($\mathbf{A}$)	Time (s)
VCA+FCLS	$0.176\ (\pm 5.8 \times 10^{-3})$	$8.80 \times 10^{-3}\ (\pm 2.2 \times 10^{-3})$	$0.246\ (\pm 4.2 \times 10^{-3})$	100 (± 27)
SISAL+FCLS	$0.187\ (\pm 1.7 \times 10^{-2})$	$\mathbf{4.61 \times 10^{-3}}\ (\pm 5.0 \times 10^{-6})$	$0.145\ (\pm 2.3 \times 10^{-2})$	**57** (± 0.5)
NMF	$0.178\ (\pm 5.9 \times 10^{-3})$	$4.87 \times 10^{-3}\ (\pm 6.3 \times 10^{-3})$	$0.246\ (\pm 4.2 \times 10^{-3})$	109 (± 26)
VCA+SUnSAL-TV	$0.176\ (\pm 5.8 \times 10^{-3})$	$9.48 \times 10^{-3}\ (\pm 6.4 \times 10^{-4})$	$0.229\ (\pm 3.6 \times 10^{-3})$	81 (± 0.7)
SISAL+SUnSAL-TV	$0.189\ (\pm 9.6 \times 10^{-3})$	$4.74 \times 10^{-3}\ (\pm 5.4 \times 10^{-5})$	$0.131\ (\pm 1.2 \times 10^{-2})$	81 (± 2)
n-SP2U	$0.190\ (\pm 1.8 \times 10^{-2})$	$35.3 \times 10^{-3}\ (\pm 4.1 \times 10^{-3})$	$0.212\ (\pm 3.0 \times 10^{-2})$	518 (± 77)
SP2U	**0.155** $(\pm 1.4 \times 10^{-2})$	$9.74 \times 10^{-3}\ (\pm 4.3 \times 10^{-4})$	**0.125** $(\pm 3.9 \times 10^{-2})$	1174 (± 62)

where $\mathbf{m}_r^{(\text{ref})}$ and $\mathbf{A}$ are the rth actual endmember signature and the actual abundance matrix, respectively.

Two additional information have also been included in the results. The processing time includes the initialization, the endmembers extraction and the abundances estimation. Moreover we also consider the reconstruction error which measure how the model fits to the observed data

$$\text{RE} = \sqrt{\frac{1}{Pd_1} \|\mathbf{Y} - \mathbf{MA}\|_{\text{F}}^2}. \tag{3.10}$$

3.5.4. Results

As stated in Section 3.3.2, the spatial feature matrix $\mathbf{S}$ has been generated using the panchromatic image. For each pixel, the spatial feature vector is directly obtained by concatenating the values of the pixels in a 11×11-pixel neighborhood around the considered pixel. This choice is very basic but designing the best spatial feature is out of the scope of this chapter. Moreover, this choice has the advantage of offering a direct interpretation of the spatial content and cluster centroids as small 11-by-11 pixels images. For these experiments, the actual number of endmembers has been assumed known and thus $R_1 = 4$ for Image 1 and

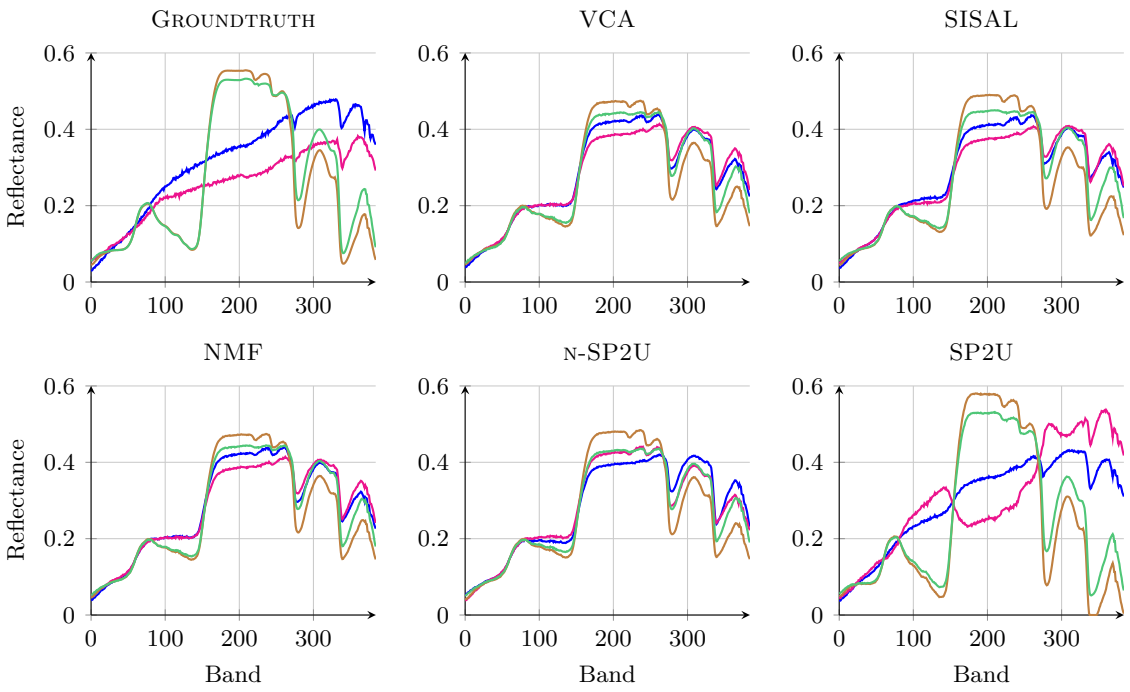


Figure 3.4.: Image 1: estimated endmembers.

$R_1 = 9$ for Image 2. The number of dictionary atoms and clusters have been empirically adjusted and set such that $R_2 = 20$ and $K = 30$ for Image 1 and $R_2 = 30$ and $K = 40$ for Image 2. It is worth noting that increasing these two parameters tends to improve the performance up to a certain point where a slow decreasing can be observed. Hence, the choice of these values is not critical as long as they are high enough. It can be explained by the fact that a sufficient number of atoms and centroids is needed to explain the data. However, beyond a certain value, increasing these parameters reduces the regularization induced by the clustering. In a more general case, using features more robust to rotation and translation deformation would likely allow to reduce the number of needed clusters and dictionary atoms. Moreover, the weighting terms of the various methods have been adjusted manually using a gridsearch algorithm in order to obtain consistent results. In particular, weighting coefficients of SP2U method have been set to $\tilde{\lambda}_0 = \tilde{\lambda}_1 = \lambda_2 = 1.0$ and $\lambda_z = 0.1$.

As the solution of the considered problem suffers from a permutation ambiguity inherent to factor models, a reordering of the endmembers is thus necessary before any evaluation. In this experiment, this relabeling is performed such that the aSAM is minimum. The quantitative results, averaged over 10 trials, has then been computed for Image 1 and Image

2 and are presented respectively in Tables 3.1 and 3.2.

The first conclusion of these results is that SP2U method gives the best estimation of the endmember matrix. All other endmember extraction algorithms are clearly behind. In particular, from Fig. 3.4, we can see that SP2U is the only method identifying that there are two spectra very different from the others which corresponds to the two soil spectra. Another interesting remark is that the NMF model barely improves the initializing point given by VCA+FCLS. It appears to converge in a few iteration to a local minimum close to initialization. Overall, it seems that including the spatial information allows to identify more clearly the endmembers in particular in the considered case where the pure pixel assumption does not hold.

Then, regarding the estimation of abundances, the evaluation is less straightforward since it depends on the estimation of the endmembers. RMSE is computed after the reordering of the endmembers and, for Image 1, the best abundance maps are obtained with SISAL+FCLS but they are not associated with the best estimated set of endmembers. The case of Image 2 is easier to discuss since the best abundance maps, obtained by SP2U, are associated with the best set of endmembers. It is also interesting to consider a qualitative evaluation of the obtained abundance maps depicted in Fig. 3.5. Even if the quantitative results seem to support the quality of the abundance maps retrieved by SUnSAL-TV, the results visually appear overly smooth. On the other hand, abundance maps estimated by SP2U seem visually relevant but the corresponding RMSE suffers from an overestimation of abundances corresponding to soil spectra. Additionally, we can see that the RE is of the same order for every model except for n-SP2U. This means that all models are equally good at finding a mixture explaining the observed data excepted n-SP2U, which was expected as explained in Section 3.5.2. Some methods such as SISAL+FCLS get a slightly lower RE but it is mostly because the method is simply a direct minimization of the RE and it does not translate necessarily in a better RMSE. Finally, it is interesting to have a look at the computational times. SP2U appears as the slowest method since it inherits from a much richer model. However, the reported computational times should be taken cautiously. Indeed, SUnSAL-TV and SISAL+FCLS were implemented with a fixed number of iterations and are based on Lagrangian augmented splitting methods. Conversely, other methods use a PALM algorithm with a different stopping criterion (see Section 3.4.2).

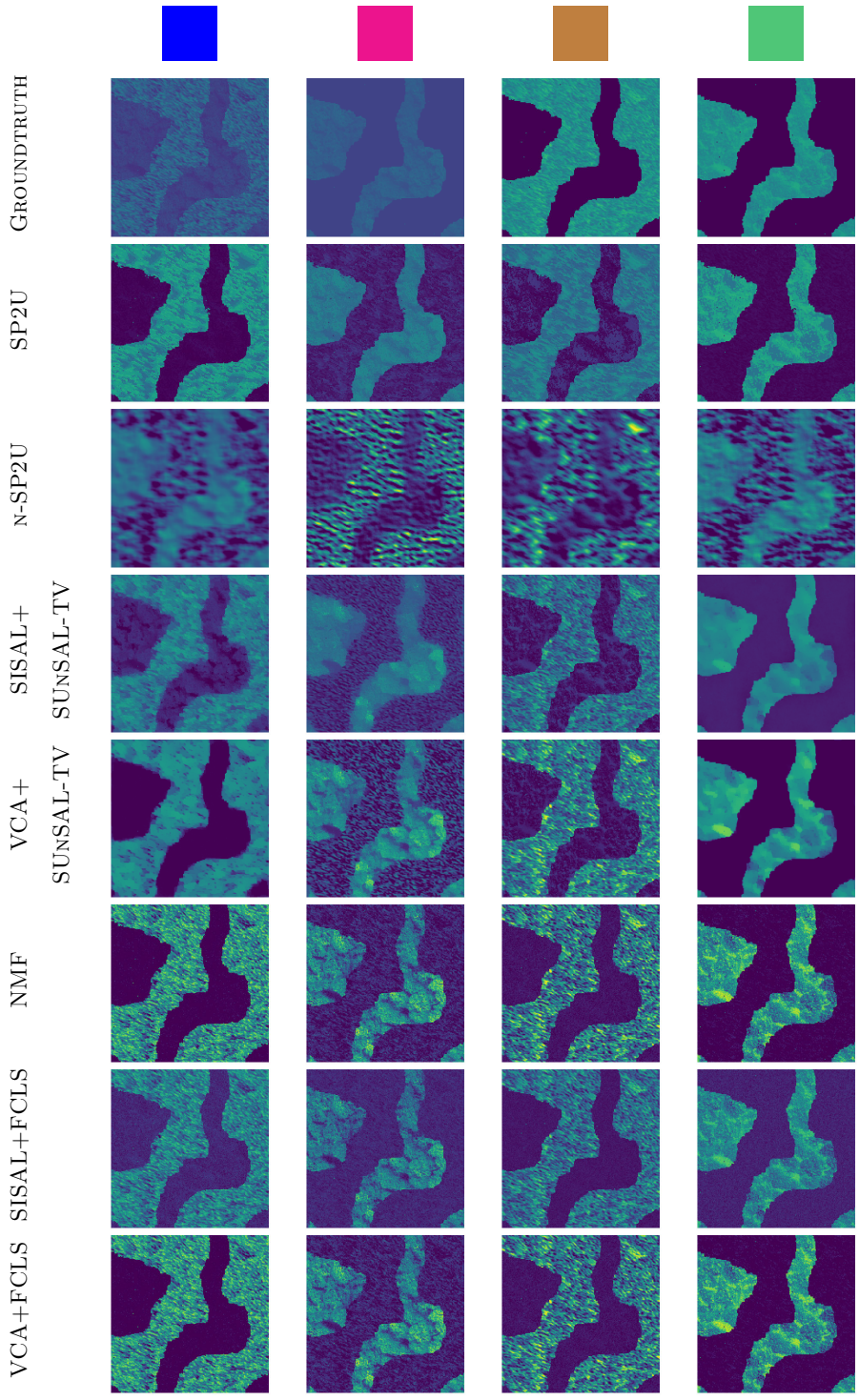


Figure 3.5.: Image 1: abundance maps (the colored squares refer to the colors used to plot endmembers in Fig. 3.4).

3.6. Experiments using real data

3.6.1. Real dataset

The real aerial hyperspectral image used to conduct the following experiment was acquired by AVIRIS in 2013 on a site called Citrus Belt 3, California. The image is composed of 224 spectral bands from 400 to 2500 nanometers with a spatial resolution of 3m per pixel. After removing bands corresponding to water absorption, a 751×651-pixel image with $d_1 = 175$ spectral bands has been finally obtained. A panchromatic image of the scene is computed by normalizing then summing all spectral bands. The resulting image and a color composition of the scene are presented in Fig. 3.6. It is possible to state that the scene includes a desert area and several vegetation areas. Thus several soil and vegetation spectra are expected to be identified.

Figure 3.6.: AVIRIS image: color composition of hyperspectral image (left) and corresponding panchromatic image (right).

3.6.2. Compared methods

As explained in Section 3.3, it is common to consider a sum-to-one constraint for abundance vectors to interpret them as proportion vectors. However, this assumption is not always fulfilled in practical scenarios. In the specific case of the considered AVIRIS image, we

decide to drop this constraint due to important illumination variation in the image. For example, the desert area on the upper part of the image is a hill and the spectrum energy is almost doubled on its sunny side. In order to get a well-defined problem after dropping the sum-to-one constraint, it is necessary to introduce a new constraint such that there is no scaling ambiguity between $\mathbf{M}$ and $\mathbf{A}$. The choice has been made to enforce a unit norm of the endmember spectra. Thus, the initial sum-to-one constraint was moved from columns of $\mathbf{A}$ to columns of $\mathbf{M}$. Then, to get abundance maps summing to one, it is possible to normalize the obtained solution a posteriori. Similarly the sum-to-one was removed for SUnSAL-TV, n-SP2U and NMF. Moreover, similarly to the synthetic case, parameters of the problem have been adjusted manually and set to $\tilde{\lambda}_0 = \tilde{\lambda}_1 = \lambda_2 = 1$ and $\lambda_z = 0.1$, $R_1 = 6$, $R_2 = 20$ and $K = 30$.

3.6.3. Results

Since no groundtruth is available for this dataset only qualitative evaluations of the various methods are performed. First, Fig. 3.7 shows the endmembers estimated by all compared methods. As explained in the previous paragraph, endmembers have been normalized except for SISAL and VCA. Regarding SISAL results, it is possible to note that the method estimates endmember signatures taking negative values. Negative endmembers can not be interpreted as real reflectance spectra and SISAL thus appears the worst compared methods. This method tries to identify a minimum volume simplex containing the observations under the assumption that the observations belong to a (R_1-1)-dimensional affine set. Thus, these poor results could be explained by a high noise level or non-linear mixtures. It is difficult to objectively compare the results of the other methods. However, the result obtained with SP2U method seems consistent with the visual content of the image since we can clearly identify *i)* two vegetation spectra (plotted in pink and orange) with strong absorbance in the visible domain and strong reflectance in the near-infrared domain [Myn+95] *ii)* two soil spectra (plotted in blue and brown) with an increase of the reflectance from 0.4μm to 1μm [Bau+86].

Regarding the abundance maps presented in Fig. 3.8, it seems again that the maps produced by SP2U are consistent with the actual content of the scene. They are in particular spatially consistent with natural edges in the image. Additionally, SP2U results seem to be sparse in the sense that only a few endmembers are used for a given pixel while other methods recover very similar abundance maps with all endmembers, see, *e.g.*, VCA+SUnSAL-TV. From Table 3.3, it seems that ensuring the sum-to-one constraint makes more difficult to fit to the observations since VCA+FCLS has the highest RE. And, again as expected, SP2U

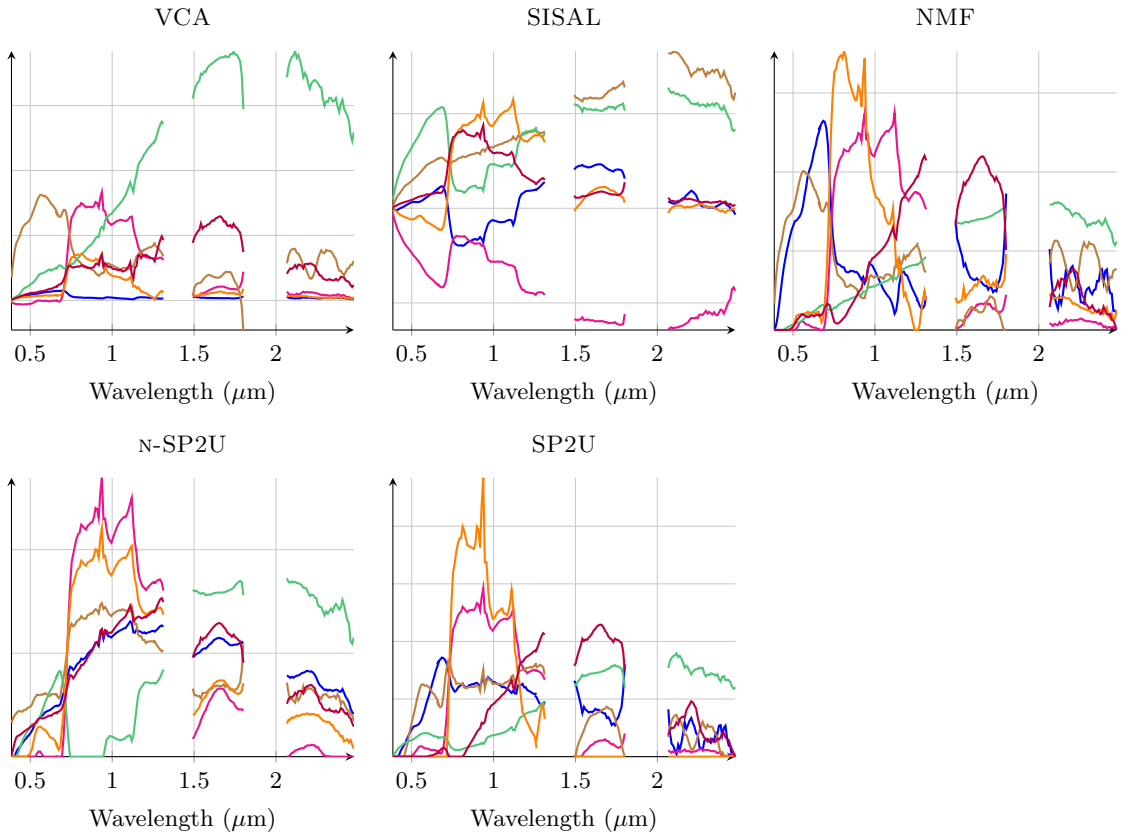


Figure 3.7.: AVIRIS image: estimated endmembers. Note that endmembers estimated by NMF, n-SP2U and SP2U have been normalized to avoid scaling ambiguity intrinsic of the estimation method.

method remains the slowest due to the overload of data to manipulate.

Table 3.3.: AVIRIS image: quantitative results of unmixing.

Model	RE	Time (s)
VCA+FCLS	2.8×10^{-3}	12
SISAL+FCLS	0.14×10^{-3}	214
NMF	0.13×10^{-3}	2054
VCA+SUnSAL-TV	0.88×10^{-3}	471
SISAL+SUnSAL-TV	0.15×10^{-3}	455
n-SP2U	1.1×10^{-3}	1347
SP2U	1.4×10^{-3}	7162

Besides, SP2U is not uniquely a spectral unmixing method and provides much richer interpretation. In Fig. 3.9, the results of the clustering performed by the coupling term are displayed. In particular, this figure shows the spatial position of the clusters, the spatial pattern characterizing the clusters and the mean spectra of the clusters. In this example, the first three clusters correspond to soil areas whereas the last two are vegetation, more precisely trees. For instance, the recovered spatial patterns associated with soil are smoother when the wooded areas are characterized by variations of higher frequencies.

3.7. Conclusion and perspectives

This chapter proposed a new model to interpret hyperspectral images. This method enriched the traditional spectral unmixing modeling by incorporating a spatial analysis of the data. Two data fitting terms, bringing respectively spectral and spatial information, were considered jointly, yielding a spatial-spectral unmixing. This coupled learning process was made possible by the introduction of a clustering-driven coupling term linking the two coding matrices. This clustering process identified groups of pixels with similar spectral and spatial behaviors.

The experiments conducted on synthetic and real data showed that the proposed method performed very well both at identifying endmembers and estimating abundances. Moreover the relevance of this method was not limited to the unmixing results since the outputs of the clustering task were also of high interest. The identified clusters were characterized by their average spectral signature and spatial context.

To further explore the potential of the proposed model, it would be particularly interesting to investigate the use of more complex spatial features instead of using directly observations

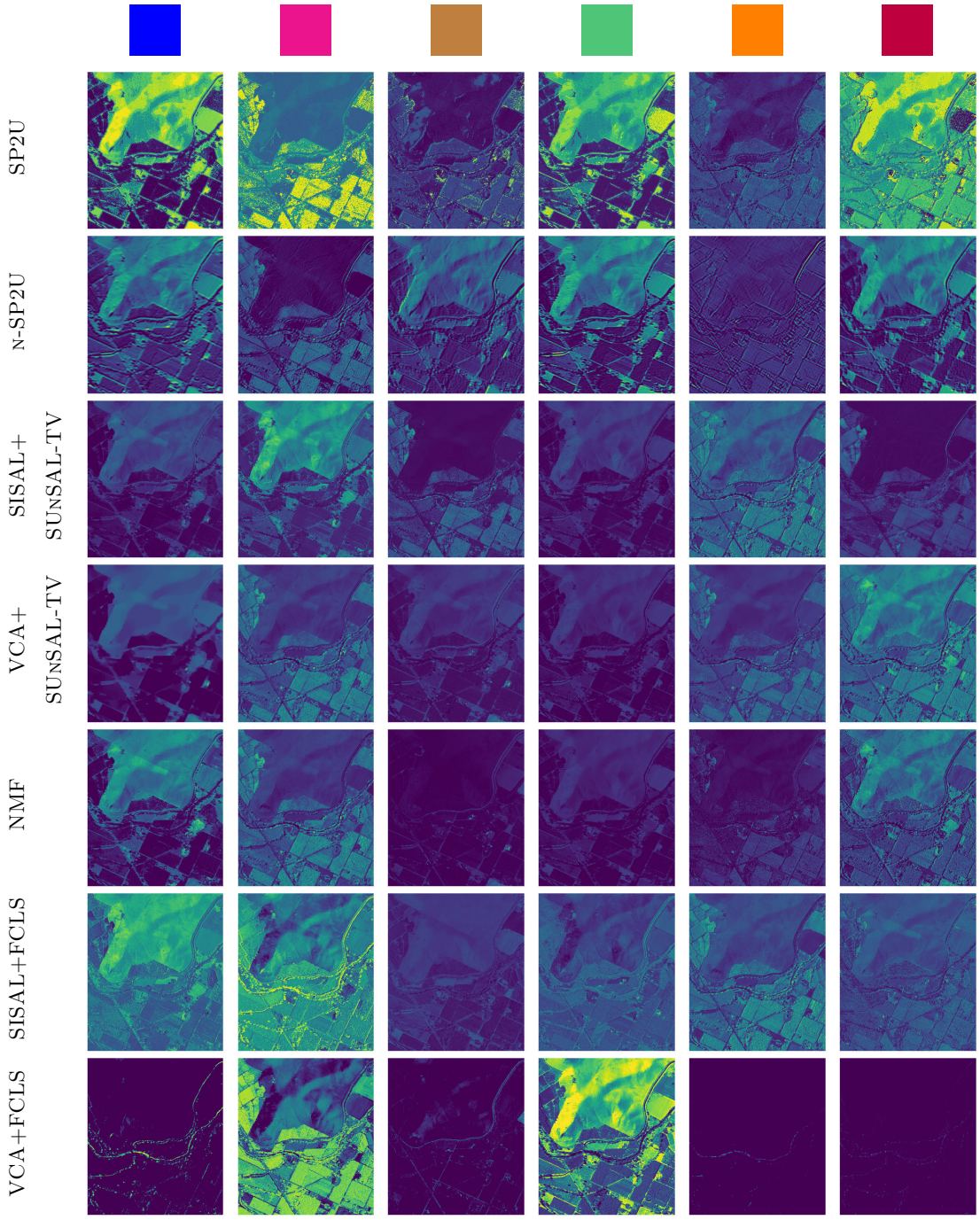


Figure 3.8.: AVIRIS image: estimated abundance maps. The colored squares refer to the colors used to plot endmembers in Fig. 3.7. However, no reordering has been performed, *i.e.*, endmembers have no particular relationship between methods.

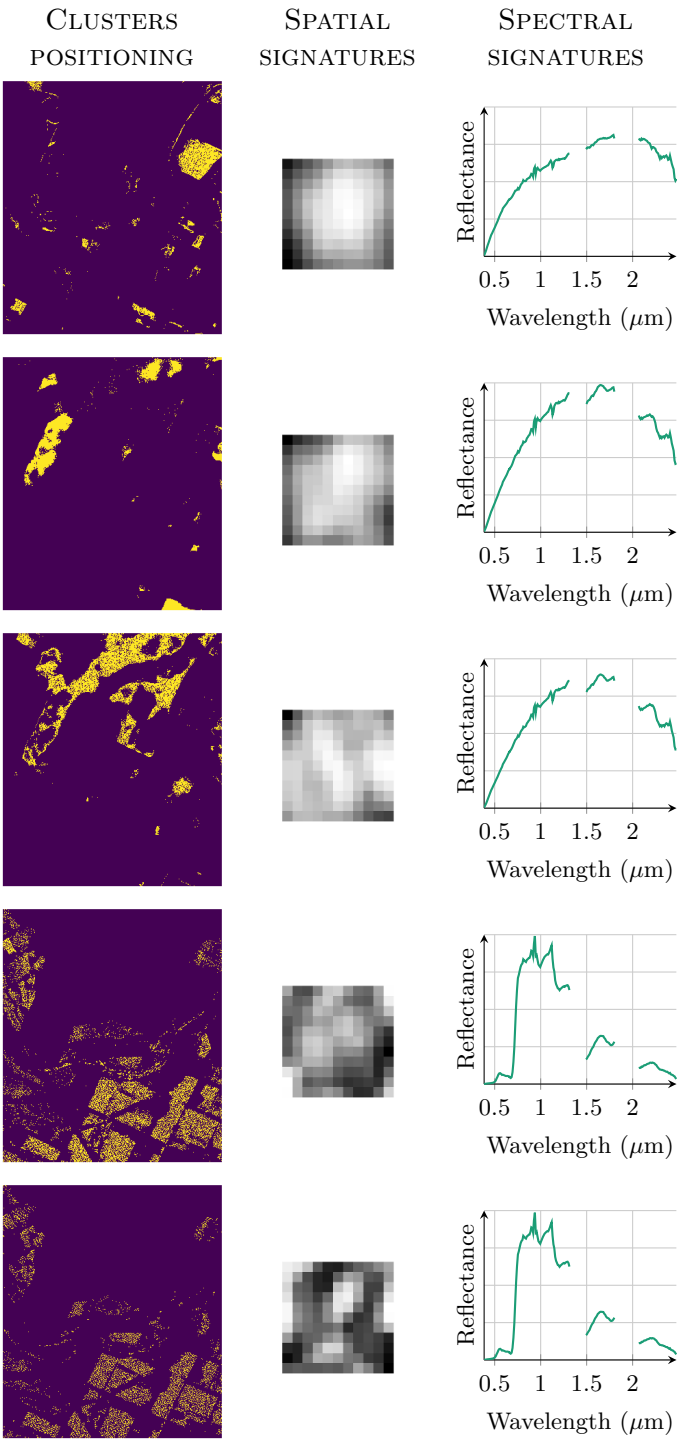


Figure 3.9.: AVIRIS image: 5 particular clusters described by their spatial positioning (left), spatial signature (middle) and spectral signature (right).

in a given neighborhood. For example, it would be relevant to use features more robust to rotation and translation in order to identify a texture instead of a fixed spatial pattern.

3.8. Conclusion (in French)

Dans ce chapitre, une nouvelle méthode pour l'interprétation d'images hyperspectrales a été introduite. Cette méthode enrichi la modèle de mélange spectrale classique en le complétant par une modélisation spatiale de l'image. Deux termes d'attache aux données, apportant respectivement des informations spectrales et spatiales, ont été utilisés conjointement, proposant ainsi un démélange spatial et spectral. Le processus d'apprentissage couplé a été rendu possible par l'introduction d'un terme de couplage, interprété comme un clustering, qui relie les deux matrices de codage. Ce clustering permet plus précisément d'identifier des groupes de pixels ayant des comportements spectraux et spatiaux similaires.

Les expériences menées sur données synthétiques puis réelles ont montré que la méthode proposée permet une bonne identification des endmembers et une bonne estimation des abondances. De plus, l'intérêt de cette méthode ne se limite pas aux résultats de démélange, puisque les résultats de la tâche de clustering présentent également un grand intérêt. Les clusters identifiés sont caractérisés par leur signature spectrale moyenne et leur contexte spatial.

Pour explorer d'avantage le potentiel du modèle proposé, il serait particulièrement intéressant d'étudier l'utilisation de descripteurs spatiaux plus complexes au lieu d'utiliser directement le voisinage du pixel comme descripteurs. Par exemple, l'utilisation de descripteurs robustes au rotation et translation permettrait de caractériser une texture plutôt qu'un motif spatial déterminé.

Conclusions

The problem of building a coherent model for joint representation learning and classification was addressed in this manuscript. The main objective was to explore the complementarity of the two image analysis methods. This complementarity was emphasized by the dependence structure developed in the proposed models.

This hierarchical modeling actually follows an intuitive line of thought. The analysis provided by a classification algorithm is based on a set of predefined classes. These classes are built around semantic concepts, such as *man-made surfaces* or *vegetation*, usually gathering heterogeneous observations. It results in multi-modal classes. Each of the modes corresponds to a set of observations, the features of which result from the same distribution. As expected by a representation learning task, this yields to the estimation of a specific low-dimensional space where observations are located and can be represented by a few latent variables, which allows the modes to be to easily identified. The representation learning process thus appears as a low-level model of the observation, potentially based on physical concepts, and the classification appears as a high-level semantic interpretation.

These models proved to offer many possibilities and advantages:

1. The multi-modality of the classes was easily handled since it is at the core of the developed models. As stated in the introduction, multi-modality of classes may be an issue for classification tasks. In particular, it makes the separation of the classes in the feature space very difficult, especially when considering linear classifiers as in the proposed methods. But with the proposed hierarchical model, the classification problem was decomposed into two steps. The modes/clusters were first identified in the low-dimensional space where they are the most easy to separate. Then the classification itself was performed using the cluster attribution vectors. In the case of a hard clustering, the attribution vectors to different clusters are orthogonal and it is thus possible to separate any union of clusters from the remaining. It means that the classification problem is more likely linearly separable in this feature space.

2. The robustness to labeling noise in the training set was studied, especially in chap-

ter 1. This robustness was mostly due to the clustering step and the semi-supervised approach. Indeed, clusters are identified using both labeled and unlabeled data and the large amount of data tends to reduce the influence of outliers or wrongly labeled data. Then, since it was expected to get a single class per cluster, wrongly labeled pixels were easily identified and eventually corrected.

3. The possibility to complement the original data with additional information was also explored in chapter 3. This chapter showed in particular that it is possible to integrate in the model a set of chosen spatial features within the hierarchical structure. It may help to get rid of ambiguities appearing when observations are considered individually and also brought new possibilities in term of interpretation of the results since the clusters were characterized from a different perspective.

Along with these model considerations, different paradigms of estimation were explored in this manuscript. Even if the purpose of this work was not to introduce new theoretical contributions regarding estimation methods, interesting remarks can be made about the practical implementation of the considered advanced estimation methods.

About MCMC estimation – The first advantage of MCMC estimation is that it allowed to handle very easily the hyperparameters of the model. Their estimation is included in the overall estimation and there is no need to rely on time-consuming selection methods such as gridsearch or cross-validation. The second advantage is the guarantee of convergence to the actual distribution even in non-convex case. Unfortunately, this nice property is somehow balanced by the difficulty to monitor the convergence of the Markov chain. It is difficult to know how many samples are necessary for the burn-in period and also how many samples are necessary for a correct estimation of the posterior distribution. The consequence is usually a long processing time. Moreover, the processing time is also strongly impacted by the distributions used as priors. When possible, it is usually clever to rely on conventional distributions yielding posterior distributions easy to sample from. When dealing with complex case, it is then necessary to call upon more complex sampling strategies, such as Metropolis-Hastings methods or Hamiltonian Monte Carlo methods, which tends to increase significantly the computational burden and the convergence time.

About optimization estimation – One of the major advantages of optimization methods is that they are easier to set up. This strength comes mostly from the extensive literature and numerous freely-available softwares. As shown by chapters 2 and 3, they also make it possible to handle larger datasets because of their shorter processing time. It is also practically easier to monitor the convergence of the algorithms since it is possible to com-

pute the value of the objective function and use it to check the convergence. However, if monitoring the convergence is common, it does not guarantee the convergence to a global optimum in most cases. Difficulties especially arise in the case of non-convex problems, where convergence to a local optimum is generally the best to hope for. For this reason, optimization methods remain very sensitive to initialization. The possibility to get an initial guess close enough to the global optimum is an indispensable prerequisite to rely on these methods. Finally, contrary to MCMC methods, the selection of relevant hyperparameters is a long-lasting problem and often remains a time-consuming and empirical process.

Perspectives and future works

There are many possibilities to continue the work detailed in this manuscript. Some of these perspectives are summarized in the following sections.

Estimation aspects

Regarding the two paradigms of estimation, namely MCMC and optimization methods, some conclusions can be drawn from this work. For the considered problem, it seems that MCMC was not the most optimal solution, in particular because of the very high computational burden. Even though the images used to test the Bayesian model were smaller, the processing time remained very long. Acceleration of MCMC methods remains a challenging problem. The generalization of splitting methods to MCMC inference has been recently explored [VDC19] and could result in the creation of distributed MCMC methods. In any case, a complete estimation of the posterior distribution is not of interest for the problem considered in this manuscript. Providing confidence sets is not as much a priority as being able to process large dataset. For these reasons, the optimization framework appears more efficient, with no loss in term of result accuracy or richness of the model.

The optimization framework is also easier to improve. It is for example possible to use improved version of the PALM algorithm with very little effort and no loss in term of convergence proof and no increase of the computational complexity. The iPALM algorithm [PS16], standing for inertial PALM, is for instance an interesting option to consider. This algorithm leverages the same acceleration idea developed by Nesterov [Nes83] in the context of convex optimization methods where an additional term depending on the previous iteration is introduced in the gradient descent expression.

A second path would be to explore the use of distributed methods [TDT18] in addition with stochastic optimization approaches which has proved to be very efficient to minimize

non convex functions in the context of deep neural networks. Stochastic methods have the advantage of better exploring the search space and are less likely to be trapped in a local minimum of the objective function. Considering a practical case, the type of problems we focus on is the following

$$\begin{aligned}\min_{\substack{\mathbf{M},\mathbf{A},\mathbf{Q},\\ \mathbf{Z},\mathbf{C}_{\mathcal{U}},\mathbf{B}}} & \frac{\lambda_0}{2}\left\|\mathbf{Y}-\mathbf{M}\mathbf{A}\right\|_{\mathrm{F}}^2+\lambda_a\left\|\mathbf{A}\right\|_1+\imath_{\mathbb{R}_+^{R\times P}}(\mathbf{A})\\ &+\frac{\lambda_1}{2}\left\|\mathbf{C}\mathbf{D}-\mathbf{Q}\mathbf{Z}\mathbf{D}\right\|_{\mathrm{F}}^2+\lambda_c\left\|\mathbf{C}\right\|_{\mathrm{vTV}}+\imath_{\mathbb{S}_C^{|\mathcal{U}|}}(\mathbf{C}_{\mathcal{U}})\\ &+\frac{\lambda_2}{2}\left\|\mathbf{A}-\mathbf{B}\mathbf{Z}\right\|_F^2+\imath_{\mathbb{S}_K^P}(\mathbf{Z})+\imath_{\mathbb{R}_+^{R\times K}}(\mathbf{B}),\end{aligned} \tag{3.1}$$

where we recognize the problem of chapter 2 with the additional optimization of matrix $\mathbf{M}$. It is possible to separate the optimized variables into two groups. The first group includes $\mathbf{A}$, $\mathbf{Z}$ and $\mathbf{C}_{\mathcal{U}}$ which are defined pixelwise, *i.e.*, each column of these matrices is related to a specific pixel. The second group is $\mathbf{M}$, $\mathbf{B}$ and $\mathbf{Q}$ which are global variables of significantly lower sizes. For the first group of variables, all the columns of these variables, indexed by pixel, are independent if the TV norm is not considered. It would thus be possible to distribute the pixels on a collection of independent processing units and possibly apply the spatial regularization on different areas independently. For the second group of variables, they could be optimized on a master node using stochastic gradient descent and then communicated to the slave nodes handling the pixelwise variables. The communication cost would be fairly limited since only a reduced number of pixels would be used to update the global variables and these smaller global variables are easy to transmit to slave nodes.

Hyperspectral images analysis

The analysis of hyperspectral images has proven to be a very challenging problem. The main limitation of the proposed approaches is the endmember matrix estimation. The estimation of an accurate endmember matrix is both crucial and very difficult due to high correlation of the endmembers. The most promising path to solve this issue is to try to further exploit additional data. Two paths emerged from the experience of this manuscript:

1. **Exploit the labeled data** – One possibility is to extract from the training set a reduced set of candidate endmembers from each of the classes. Then, imposing group sparsity penalization could ensure that only a few of the candidate endmembers are used in the overall image. Such method would be an adaptation of self-dictionary methods [GL18; GL14] with a specific selection of candidate endmembers. The ad-

vantages of this method would be the use of the supervised information and the definition of a convex problem for unmixing. However, the increase of the dimension of the problem would be a negative side-effect since the dimension of $\mathbf{M}$ containing the candidate endmembers would be increased.

2. **Exploit exogenous data** – It corresponds to the idea developed in chapter 3 where additional data of a different modality of image is included in the model. The use of spatial information has proven to be relevant to improve the estimation of the endmember matrix. Nonetheless, it is necessary to further explore this contribution. The next step is to consider more robust spatial features. One possibility, that has been explored but not included in this manuscript due to too early results, is the use of scattering transform [Mal12] to extract the features. However, this choice has the disadvantage of reducing interpretation possibilities since the spatial signature of the cluster is expressed in the feature space and the scattering coefficients are not invertible. Another possibility to focus on is the use of real panchromatic images with spatial resolution finer than the hyperspectral images. It would overcome the lack of texture of hyperspectral images due to their poor resolution.

Model developments

Before proposing new developments, it would be wise to consider a better evaluation of the proposed models. Hyperspectral unmixing is very difficult to objectively evaluate and a new application case may be a good way to assess the accuracy and the generality of the models introduced in this manuscript. As explained in the introduction, representation learning results are in general very difficult to evaluate due to the lack of groundtruth data. Nevertheless, the context of medical imaging seems a promising field to get feedback from the results of the methods. The medical experts are used to evaluate complex medical problems. If it is difficult for them to produce groundtruth for the data, it is easier to evaluate the coherence of the obtained results since they usually have expectations regarding the possible results. For example, the analysis of PET [Cav+18b] or fMRI images [Cha+12] could be interesting cases to investigate.

However, regarding the models, there are still possibilities of improvements. First of all, the suggestions made to enhance the hyperspectral image analysis, *i.e.*, the further exploitation of external data, also stands for the general models. Then, another opportunity is the improvement of the coupling term. The clustering methods that were used were the very conventional k-means algorithm and a Gaussian clustering with simplified covariance ma-

trices. These two clustering methods are easy to integrate in the joint model but remain very basic. One method to alleviate this limitation without increasing the complexity would consist in performing a clustering using a more advanced/complex method directly on the hyperspectral or/and the panchromatic images and then use the result to build a regularization term over the cluster attribution matrix $\mathbf{Z}$. A side-effect of such a regularization would also be a faster convergence of the estimation and thus a reduced processing time.

Conclusions (in French)

L'objectif principal de ces travaux a été le développement d'un modèle cohérent pour l'apprentissage de représentation et la classification conjoints. Pour cela, la complémentarité de ces deux méthodes d'analyse d'images a été étudiée ce qui a permis de mettre en place une structure de dépendance entre les deux approches.

La modélisation hiérarchique proposée correspond à une conception intuitive du problème. L'analyse fournie par un algorithme de classification est basée sur un ensemble de classes prédéterminées. Ces classes sont construites autour de concepts sémantiques larges, tels que *surfaces artificielles* ou *végétation*, regroupant généralement des observations hétérogènes ce qui a tendance à créer des classes multi-modales. Chacun des modes d'une classe correspond à un ensemble d'observations dont les caractéristiques sont générées par la même distribution. L'objectif de l'apprentissage de représentation est ensuite d'estimer l'espace de faible dimension dans lequel se trouvent ces observations et où elles peuvent être exprimées à l'aide de quelques variables latentes qui permettent de mieux distinguées ces modes. L'apprentissage de représentation apparaît ainsi comme une modélisation bas-niveau des observations, qui se basent éventuellement sur des concepts physiques, et la classification apparaît comme une interprétation sémantique haut-niveau.

Ces modèles ont offert de nombreuses possibilités et avantages :

1. Le caractère multimodal des classes est facilement pris en compte puisqu'il est au centre des modèles développés. Comme indiqué dans l'introduction, la multimodalité des classes peut constituer un problème pour les tâches de classification. En particulier, il peut devenir difficile de séparer les classes dans l'espace des descripteurs, notamment lorsqu'un classifieur linéaire est utilisé comme c'est le cas dans les méthodes proposées. La modélisation hiérarchique utilisée permet de décomposer le problème de classification en deux étapes. D'abord, les modes/clusters sont identifiés dans l'espace de faible dimension où ils sont les plus faciles à séparer. Et ensuite, la classification est effectuée à l'aide des vecteurs d'attribution à ces clusters. Dans le cas d'un clustering dur, les vecteurs d'attribution aux différents clusters sont en fait orthogonaux et il est

donc possible de séparer n'importe quelle union de clusters du reste. Cela signifie que le problème de classification est toujours séparable linéairement dans cet espace de descripteurs.

2. La robustesse au bruit dans les labels de l'ensemble d'apprentissage a également été étudiée, notamment dans le chapitre 1. Cette robustesse est principalement obtenue grâce à l'étape de clustering et à l'approche semi-supervisée considérée. En effet, les clusters sont identifiés à l'aide de données labellisées et non labellisées et la grande quantité de données considérées permet de réduire l'influence des données corrompues. De plus, puisque les éléments d'un cluster sont sensés appartenir à la même classe, il devient facile d'identifier les pixels mal étiquetés et éventuellement de les corriger.

3. La possibilité de compléter les données par de l'information complémentaire a également été explorée dans le chapitre 3. Ce chapitre montre en particulier qu'il est possible d'intégrer dans le modélisation hiérarchique un ensemble de descripteurs spatiaux. Ces descripteurs peuvent aider à lever les ambiguïtés qui apparaissent lorsque les observations sont considérées individuellement et indépendamment de leur contexte. Ils apportent également de nouvelles possibilités d'interprétation des résultats puisque les clusters sont caractérisés alors de manière beaucoup plus complète.

Parallèlement à ces développement en terme de modèle, différents paradigmes d'estimation ont été considérés dans ce manuscrit. Même si le but de ce travail n'était pas d'introduire de nouvelles contributions théoriques concernant les méthodes d'estimation, ce travail permet tout de même de tirer des enseignements concrets qu'en à la mise en œuvre des méthodes d'estimation considérées.

À propos de l'estimation par MCMC – Le premier avantage de l'estimation par MCMC est qu'elle permet de maîtriser très facilement les hyperparamètres du modèle en incluant leur estimation dans l'estimation globale. Il n'est donc pas nécessaire de recourir à des méthodes de sélection coûteuse en temps telles que la méthode de validation croisée. Le deuxième avantage est la garantie de convergence vers la distribution réelle, même avec des modèles non convexes. Malheureusement, cette propriété intéressante est contrebalancée par la difficulté à contrôler la convergence de la chaîne de Markov. Il est en effet très difficile de savoir combien d'échantillons sont nécessaires pour la période de *burn-in* tout comme il est difficile de savoir combien d'échantillons sont nécessaires pour une estimation correcte de la distribution a posteriori. Les méthodes MCMC s'avèrent donc finalement très coûteuses en temps de calcul. De plus, le temps de traitement est également fortement influencé par les distributions utilisées comme prior. Lorsque que c'est possible, il est généralement

judicieux d'avoir recours à des distributions conventionnelles pour obtenir des distributions a posteriori faciles à échantillonner. Dans des cas plus complexes, il est nécessaire de recourir à des stratégies d'échantillonnage plus avancées, telles que les méthodes de Metropolis-Hastings ou de Hamiltonian Monte Carlo, qui augmentent très significativement la charge de calcul et le temps de convergence.

À propos de l'estimation par optimisation – L'un des avantages principaux des méthodes d'optimisation est qu'elles sont plus faciles à utiliser. Cette force vient notamment de la littérature prolifique et des nombreux logiciels disponibles gratuitement. Comme on peut le voir dans les chapitres 2 et 3, ces méthodes permettent également de gérer des ensembles de données plus volumineux en raison de leur temps de traitement plus court. Il est également plus facile de surveiller la convergence de ces algorithmes dans la mesure où il est possible de calculer la valeur de la fonction objectif et de l'utiliser pour vérifier la convergence. Cependant, si contrôler la convergence de cette manière est classique, la convergence vers un optimum global n'est dans la plupart des cas pas assurée. Des problèmes se posent surtout dans le cas des problèmes non convexes, où la convergence vers un optimum local est généralement la meilleure garantie qu'on puisse attendre. Pour cette raison, les méthodes d'optimisation restent très sensibles à l'initialisation choisie. La possibilité d'obtenir une initialisation suffisamment proche de l'optimum global est une condition indispensable pour pouvoir compter sur ces méthodes. Enfin, contrairement aux méthodes MCMC, la sélection d'hyperparamètres pertinents est un problème récurrent et reste souvent un processus long et imprécis.

Perspectives

Il existe de nombreuses possibilités pour continuer le travail exposé dans ce manuscrit. Certaines de ces perspectives ont été résumées dans les sections suivantes.

Estimation

En ce qui concerne les deux paradigmes d'estimation testés, à savoir les méthodes MCMC et les méthodes d'optimisation, plusieurs conclusions peuvent être tirées de ce travail. Pour le problème considéré, les méthodes MCMC n'apparaissent pas comme le choix le plus approprié, en particulier en raison de leur très conséquente charge de calcul. En effet, même si les images utilisées pour tester le modèle bayésien étaient plus petites, le temps de traitement s'est tout de même avéré très long. De plus, une estimation complète de la distribution a posteriori n'est pas nécessaire dans les problèmes considérées. La capacité à pouvoir traiter

de grands ensembles de données est prioritaire devant la possibilité de fournir des intervalles de confiance. Pour ces raisons, le cadre d'optimisation semble plus approprié et efficace, d'autant qu'il ne dégrade ni les résultats ni la richesse des sorties du modèle.

Il est également plus simple de suggérer des améliorations de la méthode d'optimisation. Il est par exemple possible de mettre en place avec très peu d'effort une version améliorée de l'algorithme PALM, tout cela sans limiter la preuve de convergence ni augmenter la complexité de calcul. L'algorithme iPALM [PS16], pour *inertial PALM*, propose en particulier des améliorations intéressantes. Cet algorithme tire parti de la même idée d'accélération développée par Nesterov [Nes83] dans le contexte de méthodes d'optimisation convexe, où un terme supplémentaire dépendant de l'itération précédente est introduit dans l'expression de descente de gradient.

Une seconde voie pourrait être de considérer l'utilisation d'une méthode distribuée [TDT18] couplée avec des méthodes d'optimisation stochastique. Ces dernières se sont révélées très efficaces pour optimiser des fonctions non convexes dans le contexte de réseaux de neurones profonds. Les méthodes stochastiques ont l'avantage de mieux explorer l'espace de recherche et sont moins susceptibles d'être piégées dans un minimum local de la fonction objectif. Pour revenir sur un cas pratique, un cas type des problèmes abordés dans ce manuscrit est le suivant

$$\begin{aligned}\min_{\substack{\mathbf{M},\mathbf{A},\mathbf{Q},\\ \mathbf{Z},\mathbf{C}_{\mathcal{U}},\mathbf{B}}} & \frac{\lambda_0}{2}\left\|\mathbf{Y}-\mathbf{M}\mathbf{A}\right\|_{\mathrm{F}}^2+\lambda_a\left\|\mathbf{A}\right\|_1+\imath_{\mathbb{R}_+^{R\times P}}(\mathbf{A})\\ &+\frac{\lambda_1}{2}\left\|\mathbf{C}\mathbf{D}-\mathbf{Q}\mathbf{Z}\mathbf{D}\right\|_{\mathrm{F}}^2+\lambda_c\left\|\mathbf{C}\right\|_{\mathrm{vTV}}+\imath_{\mathbb{S}_C^{|\mathcal{U}|}}(\mathbf{C}_{\mathcal{U}})\\ &+\frac{\lambda_2}{2}\left\|\mathbf{A}-\mathbf{B}\mathbf{Z}\right\|_F^2+\imath_{\mathbb{S}_K^P}(\mathbf{Z})+\imath_{\mathbb{R}_+^{R\times K}}(\mathbf{B}),\end{aligned} \tag{3.1}$$

où on peut reconnaître le problème du chapitre 2 avec l'optimisation supplémentaire de la matrice de endmembers $\mathbf{M}$. Il est possible dans ce cas de séparer les variables à optimiser en deux groupes. Le premier groupe comprend $\mathbf{A}$, $\mathbf{Z}$ et $\mathbf{C}_{\mathcal{U}}$, qui sont les variables où chaque colonne de la matrice correspond à un pixel de l'image. Le deuxième groupe est constitué de $\mathbf{M}$, $\mathbf{B}$ et $\mathbf{Q}$, qui sont des variables globales de tailles nettement inférieures. Pour le premier groupe de variables, toutes les colonnes de ces variables, indexées par pixel, sont indépendantes si la norme TV n'est pas considérée. Il serait donc possible de répartir les pixels sur un ensemble d'unités de calcul indépendantes et éventuellement d'appliquer la régularisation spatiale sur différentes zones de manière indépendante. Pour le deuxième groupe de variables, elles pourraient être optimisées sur un unité de calcul centrale en utilisant une descente de gradient stochastique, puis communiquées aux nœuds esclaves qui gèrent les va-

riables au niveau du pixel. Les coûts de communication seraient alors assez limité, puisque seul un nombre réduit de pixels serait utilisé pour mettre à jour les variables globales et les variables globales plus petites seraient faciles à transmettre aux nœuds esclaves.

Analyse d'images hyperspectrales

L'analyse d'images hyperspectrales s'est avérée être un problème très complexe. La principale limite des approches proposées réside dans l'estimation de la matrice de endmember. L'estimation d'une bonne matrice de endmember est à la fois cruciale et très difficile en raison de la forte corrélation des endmembers. La voie la plus prometteuse pour améliorer son estimation consiste à essayer d'exploiter davantage de données externes. Deux pistes se dégagent en particulier :

1. **L'exploitation des données labellisées** – Une possibilité serait d'extraire de l'ensemble d'apprentissage un ensemble réduit de candidats endmembers dans chacune des classes. L'utilisation d'une pénalisation de parcimonie groupée pourrait ensuite garantir le fait que seuls quelques candidats endmembers seraient utilisés dans l'ensemble de l'image. Une telle méthode serait une adaptation des méthodes de *self-dictionary* [GL18 ; GL14] avec une sélection spécifique des endmembers candidats. Cette méthode aurait pour avantage d'utiliser l'information supervisée et de rendre le problème de démélange convexe. Cependant, un effet secondaire négatif serait l'augmentation de la dimension du problème puisque la dimension de la matrice de endmembers $\mathbf{M}$ contenant les candidats serait augmentée.

2. **L'exploitation de données exogènes** – Cette piste correspond au travail débuté dans le chapitre 3 où des données supplémentaires d'une modalité d'image différente sont incluses dans le modèle. L'utilisation d'informations spatiales s'est révélée pertinente pour améliorer l'estimation de la matrice de endmembers. Néanmoins, il est nécessaire d'explorer d'avantage cette contribution. L'étape suivante consiste à s'intéresser à l'utilisation des descripteurs spatiaux plus robustes. Une possibilité, qui a été explorée mais qui n'a pas été incluse dans ce manuscrit en raison de résultats trop précoces, est l'utilisation de la transformée en *scattering* [Mal12] pour extraire les descripteurs. Cependant, ce choix a pour inconvénient de réduire les possibilités d'interprétation, car la signature spatiale d'un cluster est exprimée dans l'espace des descripteurs et, étant donné que la transformée en *scattering* n'est pas inversible, il ne serait pas possible de visualiser la signature spatiale comme une image.

Une autre possibilité serait d'étudier l'avantage de recourir à une image panchromatique réelle avec une résolution spatiale plus fine que l'image hyperspectrale. Cela permettrait notamment de compenser le manque de textures des images hyperspectrales en raison de leur faible résolution spatiale.

Modélisation

Avant de proposer de nouveaux développements, une évaluation plus poussée des modèles proposés semble judicieuse afin d'en identifier plus clairement les limitations. Le démélange hyperspectral, qui a été utilisé comme cas particulier d'apprentissage de représentation, est en effet très difficile à évaluer objectivement. Considérer un autre cas d'application pourrait être un bon moyen d'évaluer plus précisément le modèle et le cadre général dans lequel il a été introduit. Comme expliqué dans l'introduction, les résultats d'apprentissage de représentation sont généralement très difficiles à évaluer en raison du manque de vérité terrain. Néanmoins, le contexte de l'imagerie médicale semble être un domaine prometteur pour une étude plus poussée des résultats. Les experts médicaux sont en effet habitués à évaluer/interpréter des résultats médicaux complexes. S'il leur est difficile de produire une vérité terrain pour les données, il leur est plus facile d'évaluer la cohérence des résultats obtenus. Ils possèdent généralement une idée précise des résultats potentiels et peuvent donc en évaluer la cohérence. Par exemple, l'analyse d'images TEP [Cav+18b] ou d'IRM fonctionnel [Cha+12] pourrait être un cas d'application intéressant à considérer.

Concernant le modèle lui-même, il existe tout de même des possibilités d'amélioration. Tout d'abord, les suggestions faites pour améliorer l'analyse des images hyperspectrales, c'est-à-dire l'utilisation plus poussée de données externes, sont également valables pour le modèle général. Une autre possibilité est l'amélioration du terme de couplage. Des méthodes de clustering très conventionnelles ont été utilisées comme couplage, plus précisément k-means et un clustering gaussien avec des matrices de covariance simplifiées. Ces deux méthodes de clustering sont faciles à intégrer dans le modèle global mais restent très basiques. Un moyen d'atténuer cette limitation sans augmenter la complexité serait de réaliser un clustering en utilisant une méthode plus avancée/complexe directement sur les images hyperspectrales ou/et panchromatiques, puis d'utiliser le résultat pour construire un terme de régularisation sur la matrice d'attribution aux clusters $\mathbf{Z}$. Une telle régularisation permettrait d'incorporer les résultats d'une méthode plus complexe et également d'accélérer la convergence de l'estimation des termes de clustering et donc de réduire le temps de calcul.

Appendices

Appendix A.

Assessing the accuracy

This appendix provides some details regarding the metrics used to assess the quality of the results of classification methods and classification methods.

A.1. Assessing performance: spectral unmixing

Measuring the quality of a spectral unmixing is a particularly difficult task. It is actually almost impossible to obtained a reliable groundtruth since it is very difficult for an expert to evaluate the subpixel information using the image and also very difficult to quantify the proportion of components on the field. Additionally, the definition of an elementary component is not straightforward. Depending of the user interest, endmembers can have a general meaning, e.g. *vegetation* or *rock*, or a very precise meaning, e.g. *maize* and *wheat* or *granite* and *limestone*. For this reason, the groundtruth is not uniquely defined.

Nevertheless, authors need to evaluate quantitatively their unmixing methods. In order to do it, they generally resort to synthetic data which are synthetic hyperspectral images generated using predefined elementary components which are mixed using some specific model. In this case, groundtruth endmember matrix and abundance matrix, denoted respectively by $\mathbf{M}_{\text{ref}}$ and $\mathbf{A}_{\text{ref}}$ are known and it is possible to define the following metrics,

- **Average spectral angle mapper (aSAM)** – This metric evaluates the quality of the estimated endmembers using the spectral angle. It means that only the shape of the endmember are compared to the groundtruth without taking into account the scale which is usually linked to an illumination factor and thus irrelevant.

$$\text{aSAM}(\mathbf{M}) = \frac{1}{R}\sum_{r=1}^{R} \arccos\left(\frac{{\mathbf{m}_r^{(\text{ref})}}^t.\mathbf{m}_r}{\left\|\mathbf{m}_r^{(\text{ref})}\right\|_2 \|\mathbf{m}_r\|_2}\right) \tag{A.1}$$

- **Root mean square error (RMSE)** – The mean square error measures the overall quality the abundance vectors by comparing to the groundtruth

$$\mathrm{RMSE}(\mathbf{A}) = \sqrt{\frac{1}{PR}\left\|\mathbf{A}_{\mathrm{ref}} - \mathbf{A}\right\|_{\mathrm{F}}^2} \tag{A.2}$$

- **Reconstruction error (RE)** – The reconstruction error is an indirect measure since it compares the original data to its inferred modeling. This metric is interesting because it does not require to have a groundtruth data. However, if RE is low when endmembers and abundance vectors are well-estimated, a low RE does not imply necessarily a good estimation. RE thus needs to be considered as a complementary measure which gives information about the convergence of the estimation and the risk of overfitting

$$\mathrm{RE} = \sqrt{\frac{1}{Pd}\left\|\mathbf{Y} - \mathbf{M}\mathbf{A}\right\|_{\mathrm{F}}^2} \tag{A.3}$$

As explained, it is not always possible to get quantitative measurement of the quality of the obtained unmixing. It is then very important to use qualitative evaluation. We expect in particular to obtain abundance map with spatial coherence which respect the natural boundaries of the observed scene.

A.2. Assessing performance: classification

There are many ways to measure the quality of a classification map [CG08]. As a major principle, one should always separate the available groundtruth into a training set and a validation set such that the performances of the algorithm are tested on the validation set which has not been used as training set. Additional, training and validation pixels should not be taken randomly in the available labeled data. The two sets should be spatially decorrelated meaning that some area of the image will only be used for training and some only for validation. Failing to do so would result in an overrated estimation of the performance.

Apart from that, the choice of metrics is mainly a choice of convenience and habit. We chose in this manuscript the two following metrics, largely used in the remote sensing community:

- **Cohen's kappa (κ)** – is a metric measuring agreement between two sets of labels typically the reference labels and the predicted labels [Coh60]. This metric takes into account the probability of agreement occurring by chance. Contrary to a basic percentage of agreement, Cohen's kappa gives equal importance to all classes even in

the case of unbalanced classes. Kappa is defined as follows

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{A.4}$$

where p_o is the probability of agreement between the two sets of labels estimated by computing the percentage of identical labels and $p_e = \frac{1}{P^2} \sum_{c=1}^{C} n_c^{(ref)} n_c^{(pred)}$ is the probability of random agreement with $n_c^{(ref)}$ and $n_c^{(pred)}$ are the number of pixels belonging to class c respectively in the reference and the prediction, C is the number of classes and P the number of pixels. Cohen's kappa is always inferior to 1 with $\kappa = 1$ being a perfect classification and $\kappa = 0$ a totally random classification.

- **Averaged F1-score over all classes (F1-mean)** – is an aggregation of the F1-scores computed for each class [CG08]. The F1-score of a class c is the harmonic mean between precision and recall, which are respectively the percentage of pixels classified c that actually belong to class c and the percentage of pixels of class c in the reference correctly classified as c.

$$\mathrm{F1-mean} = \frac{1}{C} \sum_{c=1}^{C} 2 \frac{\mathrm{precision}_c \mathrm{recall}_c}{\mathrm{precision}_c + \mathrm{recall}_c}. \tag{A.5}$$

The averaged F1-score is thus between 0 and 1 with $\mathrm{F1-mean} = 1$ being a perfect classification and it is also clear that a good accuracy is necessary for all classes to obtain a good score even in the case of unbalanced classes.

Appendix B.

Appendix to chapter 2

This appendix provides some details regarding the optimization schemes instanced for the cofactorization model proposed in Chapter 2 with the classification quadratic and cross-entropy losses.

B.1. Cofactorization model with quadratic loss function

Using notations consistent with (2.11), the smooth coupling term of the quadratic (Q) loss model can be expressed as

$$g(\mathbf{H},\mathbf{B},\mathbf{Z},\mathbf{C}_{\mathcal{U}},\mathbf{Q}) = \frac{\lambda_0}{2}\left\|\mathbf{Y}-\mathbf{W}\mathbf{H}\right\|_{\mathrm{F}}^2 + \frac{\lambda_1}{2}\left\|\mathbf{C}\mathbf{D}-\mathbf{Q}\mathbf{Z}\mathbf{D}\right\|_{\mathrm{F}}^2 + \lambda_c\left\|\mathbf{C}\right\|_{\mathrm{vTV}} + \frac{\lambda_2}{2}\left\|\mathbf{H}-\mathbf{B}\mathbf{Z}\right\|_{\mathrm{F}}^2. \tag{B.1}$$

For a practical implementation, one needs to compute the partial gradients of $g(\cdot)$ explicitly and their Lipschitz moduli to perform the gradient descent. Regarding the $\mathbf{H}$ and $\mathbf{B}$ variables, these computations are the same for the two models (quadratic and cross-entropy losses) and lead to

$$\nabla_{\mathbf{H}} g(\mathbf{H},\mathbf{B},\mathbf{Z},\mathbf{C}_{\mathcal{U}},\mathbf{Q}) = \lambda_0(\mathbf{W}^t\mathbf{W}\mathbf{H}-\mathbf{W}^t\mathbf{Y}) + \lambda_2(\mathbf{H}-\mathbf{B}\mathbf{Z}), \tag{B.2}$$

$$\nabla_{\mathbf{B}} g(\mathbf{H},\mathbf{B},\mathbf{Z},\mathbf{C}_{\mathcal{U}},\mathbf{Q}) = \lambda_2(\mathbf{B}\mathbf{Z}\mathbf{Z}^t-\mathbf{H}\mathbf{Z}^t), \tag{B.3}$$

Regarding the variables $\mathbf{Z}$, $\mathbf{Q}$ and $\mathbf{C}_{\mathcal{U}}$ involved in the classification step with quadratic loss, they writes

$$\begin{aligned}
\nabla_{\mathbf{Z}} g(\mathbf{H},\mathbf{B},\mathbf{Z},\mathbf{C}_{\mathcal{U}},\mathbf{Q}) &= \lambda_2(\mathbf{B}^T\mathbf{B}\mathbf{Z} - \lambda_1\mathbf{B}^T\mathbf{H}) \\
&\quad + \lambda_1(\mathbf{Q}^T\mathbf{Q}\mathbf{Z}\mathbf{D}^2 - \mathbf{Q}^T\mathbf{C}\mathbf{D}^2), \\
\nabla_{\mathbf{Q}} g(\mathbf{H},\mathbf{B},\mathbf{Z},\mathbf{C}_{\mathcal{U}},\mathbf{Q}) &= \lambda_1(\mathbf{Q}\mathbf{Z}\mathbf{D}^2\mathbf{Z}^T - \mathbf{C}\mathbf{D}^2\mathbf{Z}^T), \\
\nabla_{\mathbf{C}_{\mathcal{U}}} g(\mathbf{H},\mathbf{B},\mathbf{Z},\mathbf{C}_{\mathcal{U}},\mathbf{Q}) &= \lambda_c \nabla_{\mathbf{C}_{\mathcal{U}}} \|\mathbf{C}\|_{\mathrm{vTV}} \\
&\quad + \lambda_1(\mathbf{C}_{\mathcal{U}}\mathbf{D}_{\mathcal{U}}^2 - \mathbf{Q}\mathbf{Z}_{\mathcal{U}}\mathbf{D}_{\mathcal{U}}^2).
\end{aligned} \tag{B.4}$$

For sake of brevity, the gradient $\nabla.\, \|\cdot\|_{\mathrm{vTV}}$ of the vectorial TV regularization is not explicitly given. Readers are referred to [Get12] for further details.

All partial gradients are globally Lipschitz as functions of the corresponding partial variables. The following Lipschitz moduli can be derived as

$$\begin{aligned}
L_{\mathbf{H}} &= \left\|\lambda_0\mathbf{W}^T\mathbf{W} + \lambda_2\mathbf{I}_R\right\|, \\
L_{\mathbf{B}}(\mathbf{Z}) &= \left\|\lambda_2\mathbf{Z}\mathbf{Z}^T\right\|, \\
L_{\mathbf{Z}}(\mathbf{B},\mathbf{Q}) &= \max_p \left\|\lambda_2\mathbf{B}^T\mathbf{B} + \lambda_1 d_p\mathbf{Q}^T\mathbf{Q}\right\|, \\
L_{\mathbf{Q}}(\mathbf{Z}) &= \left\|\lambda_1\mathbf{Z}\mathbf{D}^2\mathbf{Z}^T\right\|, \\
L_{\mathbf{C}_{\mathcal{U}}} &= \lambda_1 \max_p d_p^2 + \lambda_c \frac{\sqrt{8}\max_p \beta_p}{\epsilon}.
\end{aligned} \tag{B.5}$$

B.2. Cofactorization model with cross-entropy loss function

When using cross-entropy as the classification loss function, the coupling term writes

$$\begin{aligned}
g(\mathbf{H},\mathbf{B},\mathbf{Z},\mathbf{C}_{\mathcal{U}},\mathbf{Q}) &= \frac{\lambda_0}{2}\|\mathbf{Y} - \mathbf{W}\mathbf{H}\|_{\mathrm{F}}^2 \\
&- \frac{\lambda_1}{2}\sum_{p\in\mathcal{P}} d_p^2 \sum_{i\in\mathcal{C}} c_{i,p}\log\left(\frac{1}{1+\exp(-\mathbf{q}_{i:}\mathbf{z}_p)}\right) \\
&+ \frac{\lambda_q}{2}\|\mathbf{Q}\|_2^2 + \lambda_c\|\mathbf{C}\|_{\mathrm{vTV}} + \frac{\lambda_2}{2}\|\mathbf{H} - \mathbf{B}\mathbf{Z}\|_F^2
\end{aligned} \tag{B.6}$$

and the specific partial gradients are

$$\begin{aligned}
\nabla_{\mathbf{Z}} g(\mathbf{H}, \mathbf{B}, \mathbf{Z}, \mathbf{C}_{\mathcal{U}}, \mathbf{Q}) &= -\frac{\lambda_1}{2}\mathbf{Q}^T\mathbf{G} \\
\nabla_{\mathbf{Q}} g(\mathbf{H}, \mathbf{B}, \mathbf{Z}, \mathbf{C}_{\mathcal{U}}, \mathbf{Q}) &= -\frac{\lambda_1}{2}\mathbf{G}\mathbf{Z}^T + \lambda_q\mathbf{Q}, \\
\nabla_{\mathbf{C}_{\mathcal{U}}} g(\mathbf{H}, \mathbf{B}, \mathbf{Z}, \mathbf{C}_{\mathcal{U}}, \mathbf{Q}) &= \lambda_c \nabla_{\mathbf{C}_{\mathcal{U}}} \|\mathbf{C}_{\mathcal{U}}\|_{\mathrm{vTV}} \\
&\quad - \frac{\lambda_1}{2}\sum_{p\in\mathcal{P}} d_p^2 \sum_{i\in\mathcal{C}} \log\left(\frac{1}{1+\exp(-\mathbf{q}_{i:}\mathbf{z}_p)}\right)
\end{aligned} \tag{B.7}$$

where $\mathbf{G}$ is a $C \times P$ matrix with elements given by

$$g_{i,p} = \frac{d_p^2 c_{i,p}}{1+\exp(\mathbf{q}_{i:}\mathbf{z}_p)}. \tag{B.8}$$

It should be noticed that $\mathbf{G}$ depends on $\mathbf{Z}$, $\mathbf{Q}$ and $\mathbf{C}$ and is only introduced here to get compact notations. The following Lipschitz moduli can be derived

$$\begin{aligned}
L_{\mathbf{Z}}(\mathbf{B}, \mathbf{Q}) &= \lambda_1 \sum_{p\in\mathcal{P}} d_p^2 \sum_{i\in\mathcal{C}} c_{i,p} \|\mathbf{q}_{j:}\|_2^2 + \left\|\lambda_2 \mathbf{B}\mathbf{B}^T\right\|, \\
L_{\mathbf{Q}} &= \lambda_1 \sum_{p\in\mathcal{P}} d_p^2 + \lambda_{\mathrm{q}}, \\
L_{\mathbf{C}_{\mathcal{U}}} &= \lambda_c \frac{\sqrt{8}\max_p \beta_p}{\epsilon}.
\end{aligned} \tag{B.9}$$

B.3. Computing the proximal operators

For a practical implementation of the PALM algorithm, the proximal operators associated with each $f_j(\cdot)$ in (2.12) need to be computed. It is clear that all these functions are proper lower semi-continuous functions for both models instanced in Section 2.4.4. The involved indicator functions are defined on convex sets. Thus, their proximal operators can be expressed as projections. The projection on the non-negative quadrant is a simple thresholding of negative values. The projection on the simplices $\mathcal{S}_\cdot$ can be conducted as detailed in [Con16]. The case of $f_0(\cdot)$ defined by a nonnegativity constraint complemented by a ℓ_1-norm sparsity promoting regularization is slightly more complex. It can be handled using a composition of proximal operators. As stated before, the proximal operator associated to the positivity constraint is the projection on the non-negative quadrant. The proximal operator associated with the ℓ_1-norm penalization is a soft-thresholding, i.e., $\mathrm{prox}^t_{\|\cdot\|_1}(x) = \mathrm{sign}(x)(|x| - \frac{1}{t})_+$ [Jen+11]. These two proximal operators satisfy the condi-

tions exhibited in [Yu13] required to be allowed to perform their compositions to get the proximal operator associated to $f_0(\cdot)$.

Appendix C.

Appendix to chapter 3

C.1. Computation details for optimization

This appendix provides some details regarding the optimization schemes instanced for the proposed cofactorization model.

Using notations adopted in Section 3.4, the smooth coupling term can be expressed as

$$g(\mathbf{M},\mathbf{A},\mathbf{D},\mathbf{U},\mathbf{B},\mathbf{Z}) = \frac{\lambda_0}{2}\left\|\mathbf{Y}-\mathbf{M}\mathbf{A}\right\|_{\mathrm{F}}^2 + \frac{\lambda_1}{2}\left\|\mathbf{S}-\mathbf{D}\mathbf{U}\right\|_{\mathrm{F}}^2 + \frac{\lambda_2}{2}\left\|\begin{pmatrix}\mathbf{A}\\ \mathbf{U}\end{pmatrix}-\mathbf{B}\mathbf{Z}\right\|_{\mathrm{F}}^2 + \frac{\lambda_z}{2}\mathrm{Tr}(\mathbf{Z}^T\mathbf{V}\mathbf{Z}).$$

For a practical implementation of PALM, the partial gradients of $g(\cdot)$ and their Lipschitz moduli need to be computed to perform the gradient descent. They are given by

$$\begin{aligned}
\nabla_{\mathbf{M}} g(\mathbf{M},\mathbf{A},\mathbf{D},\mathbf{U},\mathbf{B},\mathbf{Z}) &= \lambda_0(\mathbf{M}\mathbf{A}\mathbf{A}^T - \mathbf{Y}\mathbf{A}^T),\\
\nabla_{\mathbf{A}} g(\mathbf{M},\mathbf{A},\mathbf{D},\mathbf{U},\mathbf{B},\mathbf{Z}) &= \lambda_0(\mathbf{M}^T\mathbf{M}\mathbf{A} - \mathbf{M}^T\mathbf{Y}) + \lambda_2(\mathbf{A}-\mathbf{B}_1\mathbf{Z}),\\
\nabla_{\mathbf{D}} g(\mathbf{M},\mathbf{A},\mathbf{D},\mathbf{U},\mathbf{B},\mathbf{Z}) &= \lambda_1(\mathbf{D}\mathbf{U}\mathbf{U}^T - \mathbf{S}\mathbf{U}^T),\\
\nabla_{\mathbf{U}} g(\mathbf{M},\mathbf{A},\mathbf{D},\mathbf{U},\mathbf{B},\mathbf{Z}) &= \lambda_1(\mathbf{D}^T\mathbf{D}\mathbf{U} - \mathbf{D}^T\mathbf{S}) + \lambda_2(\mathbf{U}-\mathbf{B}_2\mathbf{Z}),\\
\nabla_{\mathbf{B}} g(\mathbf{M},\mathbf{A},\mathbf{D},\mathbf{U},\mathbf{B},\mathbf{Z}) &= \lambda_2(\mathbf{B}\mathbf{Z}\mathbf{Z}^T - \begin{pmatrix}\mathbf{A}\\ \mathbf{U}\end{pmatrix}\mathbf{Z}^T),\\
\nabla_{\mathbf{Z}} g(\mathbf{M},\mathbf{A},\mathbf{D},\mathbf{U},\mathbf{B},\mathbf{Z}) &= \lambda_2(\mathbf{B}^T\mathbf{B}\mathbf{Z} - \mathbf{B}^T\begin{pmatrix}\mathbf{A}\\ \mathbf{U}\end{pmatrix}) + \lambda_z\mathbf{V}\mathbf{Z},
\end{aligned}$$

where $\mathbf{B}_1$ and $\mathbf{B}_2$ correspond to the submatrices of $\mathbf{B}$ defined by the R_1 first rows and R_2

last rows, respectively, such that $\mathbf{B} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix}$.

All partial gradients are globally Lipschitz as functions of the corresponding partial variables. The following Lipschitz moduli can be explicitly derived as

$$\begin{aligned} L_{\mathbf{A}}(\mathbf{M}) &= \left\| \lambda_0 \mathbf{M}^T \mathbf{M} + \lambda_2 \mathbf{I}_{R_1} \right\|, \\ L_{\mathbf{M}}(\mathbf{A}) &= \left\| \lambda_0 \mathbf{A}\mathbf{A}^T \right\|, \\ L_{\mathbf{U}}(\mathbf{D}) &= \left\| \lambda_1 \mathbf{D}^T \mathbf{D} + \lambda_2 \mathbf{I}_{R_2} \right\|, \\ L_{\mathbf{D}}(\mathbf{U}) &= \left\| \lambda_1 \mathbf{U}\mathbf{U}^T \right\|, \\ L_{\mathbf{B}}(\mathbf{Z}) &= \left\| \lambda_2 \mathbf{Z}\mathbf{Z}^T \right\|, \\ L_{\mathbf{Z}}(\mathbf{B}) &= \left\| \lambda_2 \mathbf{B}^T \mathbf{B} + \lambda_z \mathbf{V} \right\|. \end{aligned} \tag{C.1}$$

Bibliography

[AEB06] M. Aharon, M. Elad, and A. Bruckstein. "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation". In: *IEEE Trans. Signal Process.* 54.11 (2006), p. 4311 (cit. on pp. 9, 50, 51, 89).

[AJE01] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. New York: John Wiley, 2001 (cit. on p. 89).

[Ala+17] F. I. Alam, J. Zhou, L. Tong, A. W. Liew, and Y. Gao. "Combining Unmixing and Deep Feature Learning for Hyperspectral Image Classification". In: *Proc. Int. Conf. on Digital Image Computing: Techniques and Applications (DICTA)*. Nov. 2017, pp. 1–8 (cit. on p. 11).

[Alb+14] M. Albughdadi, L. Chaari, F. Forbes, J. Y. Tourneret, and P. Ciuciu. "Model Selection for Hemodynamic Brain Parcellation in FMRI". In: *Proc. European Signal Process. Conf. (EUSIPCO)*. 2014, pp. 31–35 (cit. on p. 25).

[ALL17] N. Audebert, B. Le Saux, and S. Lefèvre. "Segment-before-Detect: Vehicle Detection and Classification through Semantic Segmentation of Aerial Images". en. In: *Remote Sens.* 9.4 (Apr. 2017), p. 368 (cit. on p. 5).

[AM18] N. Akhtar and A. Mian. "Nonparametric Coupled Bayesian Dictionary and Classifier Learning for Hyperspectral Classification". In: *IEEE Trans. Neural Netw. Learn. Syst.* 29.9 (2018), pp. 4038–4050 (cit. on pp. 10, 50, 52, 85, 86).

[AMD14] Y. Altmann, S. McLaughlin, and N. Dobigeon. "Sampling from a multivariate Gaussian distribution truncated on a simplex: a review". In: *Proc. IEEE-SP Workshop Stat. and Signal Process. (SSP)*. Invited paper. Gold Coast, Australia, July 2014, pp. 113–116 (cit. on p. 36).

[And+16] V. Andrejchenko, R. Heylen, P. Scheunders, W. Philips, and W. Liao. "Classification of Hyperspectral Images with Very Small Training Size Using Sparse Unmixing". In: *Proc. IEEE Int. Conf. Geosci. Remote Sens. (IGARSS)*. July 2016, pp. 5115–5117 (cit. on pp. 12, 13).

[And+19] V. Andrejchenko, W. Liao, W. Philips, and P. Scheunders. "Decision Fusion Framework for Hyperspectral Image Classification Based on Markov and Conditional Random Fields". In: *Remote Sens.* 11.6 (2019), p. 624 (cit. on pp. 13, 21, 22).

[Ban08] I. Bankman. *Handbook of Medical Image Processing and Analysis*. Elsevier, 2008 (cit. on pp. 1, 5).

[Bau+86] M. F. Baumgardner, L. F. Silva, L. L. Biehl, and E. R. Stoner. "Reflectance Properties of Soils". In: *Advances in Agronomy*. Vol. 38. Elsevier, 1986, pp. 1–44 (cit. on p. 102).

[BCV13] Y. Bengio, A. Courville, and P. Vincent. "Representation Learning: A Review and New Perspectives". In: *IEEE Trans. Patt. Anal. Mach. Intell.* 35.8 (2013), pp. 1798–1828 (cit. on pp. 5, 10).

[Bes75] J. Besag. "Statistical Analysis of Non-Lattice Data". In: *J. Roy. Stat. Soc. Ser. D* 24.3 (1975), pp. 179–195 (cit. on p. 31).

[BEZ08] A. M. Bruckstein, M. Elad, and M. Zibulevsky. "On the Uniqueness of Nonnegative Sparse Solutions to Underdetermined Systems of Equations". In: *IEEE Trans. Inf. Theory* 54.11 (Nov. 2008), pp. 4813–4820 (cit. on p. 53).

[BF10] J. M. Bioucas-Dias and M. A. Figueiredo. "Alternating Direction Algorithms for Constrained Sparse Regression: Application to Hyperspectral Unmixing". In: *Proc. IEEE GRSS Workshop Hyperspectral Image SIgnal Process.: Evolution in Remote Sens. (WHISPERS)*. IEEE, 2010, pp. 1–4 (cit. on pp. 41, 68, 71, 84, 86, 95).

[BF99] C. E. Brodley and M. A. Friedl. "Identifying Mislabeled Training Data". In: *J. Artif. Intell. Res.* 11 (1999), pp. 131–167 (cit. on p. 6).

[BG09] C. Bouveyron and S. Girard. "Robust Supervised Classification with Mixture Models: Learning from Data with Uncertain Labels". In: *Patt. Recognition* 42 (2009), pp. 2649–2658 (cit. on pp. 6, 20, 22, 40).

[BGV92] B. E. Boser, I. M. Guyon, and V. N. Vapnik. "A Training Algorithm for Optimal Margin Classifiers". In: *Proc. of the Fifth Annual Workshop on Computational Learning Theory*. ACM, 1992, pp. 144–152 (cit. on pp. 1, 5).

[Bio+12] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot. "Hyperspectral Unmixing Overview: Geometrical, Statistical, and Sparse Regression-Based Approaches". In: *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.* 5 (2012), pp. 354–379 (cit. on pp. 2, 8, 11, 33, 34, 60, 87).

[Bio09] J. M. Bioucas-Dias. "A Variable Splitting Augmented Lagrangian Approach to Linear Spectral Unmixing". In: *Proc. IEEE GRSS Workshop Hyperspectral Image SIgnal Process.: Evolution in Remote Sens. (WHISPERS)*. IEEE, 2009, pp. 1–4 (cit. on p. 95).

[BN08] J. M. Bioucas-Dias and J. M. Nascimento. "Hyperspectral Subspace Identification". In: *IEEE Trans. Geosci. Remote Sens.* 46.8 (2008), pp. 2435–2445 (cit. on pp. 2, 7).

[Bob+07] J. Bobin, J.-L. Starck, J. Fadili, and Y. Moudden. "Sparsity and Morphological Diversity in Blind Source Separation". In: *IEEE Trans. Image Process.* 16.11 (2007), pp. 2662–2674 (cit. on p. 7).

[Bob+15] J. Bobin, J. Rapin, A. Larue, and J.-L. Starck. "Sparsity and Adaptivity for the Blind Separation of Partially Correlated Sources". In: *IEEE Trans. Image Process.* 63.5 (2015), pp. 1199–1213 (cit. on p. 8).

[Boy+11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers". In: *Foundations and Trends® in Machine Learning* 3.1 (2011), pp. 1–122 (cit. on pp. 50, 51).

[Bro+11] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011 (cit. on p. 8).

[BST14] J. Bolte, S. Sabach, and M. Teboulle. "Proximal Alternating Linearized Minimization for Nonconvex and Nonsmooth Problems". en. In: *Math. Program.* 146.1-2 (Aug. 2014), pp. 459–494 (cit. on pp. 8, 50–52, 58, 91).

[Cam+14] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson. "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods". In: *IEEE Signal Process. Mag.* 31.1 (2014), pp. 45–54 (cit. on pp. 6, 11, 33).

[Cam+16] M. Campos-Taberner, A. Romero-Soriano, C. Gatta, G. Camps-Valls, A. Lagrange, B. Le Saux, A. Beaupère, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, M. Ferecatu, M. Shimoni, G. Moser, and D. Tuia. "Processing of Extremely High-Resolution LiDAR and RGB Data: Outcome of the 2015 IEEE GRSS Data Fusion Contest - Part A: 2-D Contest". In: *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.* 9.12 (Dec. 2016), pp. 5547–5559.

[Can+11] K. Canham, A. Schlamm, A. Ziemann, B. Basener, and D. Messinger. "Spatially Adaptive Hyperspectral Unmixing". In: *IEEE Trans. Geosci. Remote Sens.* 49.11 (2011), pp. 4248–4262 (cit. on pp. 84, 86).

[Cas16] D. Castelvecchi. "Can We Open the Black Box of AI?" In: *Nature News* 538.7623 (2016), p. 20 (cit. on pp. 1, 7).

[Cav+18a] Y. C. Cavalcanti, T. Oberlin, N. Dobigeon, S. Stute, M.-J. Ribeiro, and C. Tauber. "Factor Analysis of Dynamic PET Images: Beyond Gaussian Noise". In: *arXiv preprint arXiv:1807.11455* (2018) (cit. on pp. 50, 51).

[Cav+18b] Y. C. Cavalcanti, T. Oberlin, N. Dobigeon, S. Stute, M. Ribeiro, and C. Tauber. "Unmixing Dynamic PET Images with Variable Specific Binding Kinetics". In: *Med. Image Anal.* 49 (Oct. 2018), pp. 117–127 (cit. on pp. 2, 7, 113, 120).

[CFB08] M. Chi, R. Feng, and L. Bruzzone. "Classification of Hyperspectral Remote-Sensing Data with Primal SVM for Small-Sized Training Dataset Problem". In: *Adv. Space Res.* 41.11 (Jan. 2008), pp. 1793–1799 (cit. on p. 6).

[CG08] R. G. Congalton and K. Green. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. CRC press, 2008 (cit. on pp. 71, 124, 125).

[Cha+12] L. Chaari, T. Vincent, F. Forbes, M. Dojat, and P. Ciuciu. "Fast Joint Detection-Estimation of Evoked Brain Activity in Event-Related fMRI Using a Variational Approach". In: *IEEE Trans. Med. Imaging* 32.5 (2012), pp. 821–837 (cit. on pp. 113, 120).

[Che+17] F. Chen, K. Wang, T. Van de Voorde, and 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". In: *Remote Sens. Environment* 196 (2017), pp. 324–342 (cit. on pp. 21, 22).

[CJ11] C. Févotte and J. Idier. "Algorithms for nonnegative matrix factorization with the beta-divergence". In: *Neural Comput.* 23.9 (Sept. 2011), pp. 2421–2456 (cit. on pp. 25, 53).

[CK05] N. V. Chawla and G. Karakoulas. "Learning from Labeled and Unlabeled Data: An Empirical Study across Techniques and Domains". In: *J. Artif. Intell. Res.* 23 (2005), pp. 331–366 (cit. on p. 6).

[CNJ09] C. Févotte, N. Bertin, and J.-L. Durrieu. "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis". In: *Neural Comput.* 21.3 (Mar. 2009), pp. 793–830 (cit. on p. 24).

[CNT11] Y. Chen, N. M. Nasrabadi, and T. D. Tran. "Hyperspectral Image Classification Using Dictionary-Based Sparse Representation". In: *IEEE Trans. Geosci. Remote Sens.* 49.10 (Oct. 2011), pp. 3973–3985 (cit. on p. 9).

[Coh60] J. Cohen. "A Coefficient of Agreement for Nominal Scales". In: *Educ. Psychol. Meas.* 20.1 (1960), pp. 37–46 (cit. on p. 124).

[Con16] L. Condat. "Fast Projection onto the Simplex and the l1 Ball". In: *Math. Program.* 158.1-2 (2016), pp. 575–585 (cit. on pp. 91, 129).

[Con17] L. Condat. "A Convex Approach to K-Means Clustering and Image Segmentation". In: *Proc. Int. Workshop on Energy Minimization Methods Comput. Vis. Pattern Recognit. (EMMCVPR)*. Springer, 2017, pp. 220–234 (cit. on p. 55).

[CP11] P. L. Combettes and J.-C. Pesquet. "Proximal Splitting Methods in Signal Processing". en. In: *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Ed. by H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz. Springer Optimization and Its Applications. New York, NY: Springer New York, 2011, pp. 185–212. ISBN: 978-1-4419-9569-8 (cit. on p. 8).

[DB12] N. Dobigeon and N. Brun. "Spectral mixture analysis of EELS spectrum-images". In: *Ultramicroscopy* 120 (Sept. 2012), pp. 25–34 (cit. on p. 34).

[Don00] D. L. Donoho. "High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality". In: *AMS math challenges lecture* 1.32 (2000), p. 375 (cit. on p. 6).

[Dóp+11] I. Dópido, A. Villa, A. Plaza, and P. Gamba. "A Comparative Assessment of Several Processing Chains for Hyperspectral Image Classification: What Features to Use?" In: *Proc. IEEE GRSS Workshop Hyperspectral Image SIgnal Process.: Evolution in Remote Sens. (WHISPERS)*. June 2011, pp. 1–4 (cit. on p. 11).

[Dóp+12] I. Dópido, A. Villa, A. Plaza, and P. Gamba. "A Quantitative and Comparative Assessment of Unmixing-Based Feature Extraction Techniques for Hyperspectral Image Classification". In: *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.* 5.2 (2012), pp. 421–435 (cit. on p. 11).

[Dóp+14] I. Dópido, J. Li, P. Gamba, and A. Plaza. "A New Hybrid Strategy Combining Semisupervised Classification and Unmixing of Hyperspectral Data". In: *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.* 7 (2014), pp. 3619–3629 (cit. on pp. 12, 34).

[Dru+16] L. Drumetz, M.-A. Veganzones, S. Henrot, R. Phlypo, J. Chanussot, and C. Jutten. "Blind Hyperspectral Unmixing Using an Extended Linear Mixing Model to Address Spectral Variability". In: *IEEE Trans. Image Process.* 25.8 (2016), pp. 3890–3905 (cit. on pp. 60, 88).

[DT05] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection". In: *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. June 2005, 886–893 vol. 1 (cit. on p. 5).

[DTC08] N. Dobigeon, J.-Y. Tourneret, and C.-I Chang. "Semi-Supervised Linear Spectral Unmixing Using a Hierarchical Bayesian Model for Hyperspectral Imagery". In: *IEEE Trans. Signal Process.* 56.7 (July 2008), pp. 2684–2695 (cit. on pp. 20, 21).

[DW13] C. Deng and C. Wu. "A Spatially Adaptive Spectral Mixture Analysis for Mapping Subpixel Urban Impervious Surface Distribution". In: *Remote Sens. Environment* 133 (June 2013), pp. 62–70 (cit. on pp. 84, 86).

[Ech+13] O. Eches, J. A. Benediktsson, N. Dobigeon, and J.-Y. Tourneret. "Adaptive Markov random fields for joint unmixing and segmentation of hyperspectral image". In: *IEEE Trans. Image Process.* 22.1 (Jan. 2013), pp. 5–16 (cit. on pp. 84, 86).

[EDT11] O. Eches, N. Dobigeon, and J.-Y. Tourneret. "Enhancing Hyperspectral Image Unmixing with Spatial Correlations". In: *IEEE Trans. Geosci. Remote Sens.* 49 (2011), pp. 4239–4247 (cit. on pp. 8, 14, 26, 39, 84, 86, 88).

[El +06] G. El Fakhri, A. Sitek, R. E. Zimmerman, and J. Ouyang. "Generalized Five-Dimensional Dynamic and Spectral Factor Analysis". In: *Med. Phys.* 33.4 (2006), pp. 1016–1024 (cit. on p. 7).

[Ess+12] E. Esser, M. Moller, S. Osher, G. Sapiro, and J. Xin. "A Convex Model for Nonnegative Matrix Factorization and Dimensionality Reduction on Physical Space". In: *IEEE Trans. Image Process.* 21.7 (2012), pp. 3239–3252 (cit. on p. 7).

[EV13] E. Elhamifar and R. Vidal. “Sparse Subspace Clustering: Algorithm, Theory, and Applications”. In: *IEEE Trans. Patt. Anal. Mach. Intell.* 35.11 (2013), pp. 2765–2781 (cit. on pp. 50, 51).

[Fau+13] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton. “Advances in Spectral-Spatial Classification of Hyperspectral Images”. In: *Proc. IEEE* 101.3 (2013), pp. 652–675 (cit. on p. 11).

[FCB06] M. Fauvel, J. Chanussot, and J. A. Benediktsson. “Kernel Principal Component Analysis for Feature Reduction in Hyperspectrale Images Analysis”. In: *Proc. Nordic Signal Process. Symp. (NORSIG)*. 2006, pp. 238–241 (cit. on p. 24).

[FCB12] M. Fauvel, J. Chanussot, and J. A. Benediktsson. “A Spatial–Spectral Kernel-Based Approach for the Classification of Remote-Sensing Images”. In: *Patt. Recognition* 45.1 (Jan. 2012), pp. 381–392 (cit. on pp. 20, 22).

[FM04] G. M. Foody and A. Mathur. “Toward Intelligent Training of Supervised Image Classifications: Directing Training Data Acquisition for SVM Classification”. In: *Remote Sens. Environment* 93.1 (Oct. 30, 2004), pp. 107–117 (cit. on p. 6).

[FRZ18] H. Foroughi, N. Ray, and H. Zhang. “Object Classification with Joint Projection and Low-Rank Dictionary Learning”. In: *IEEE Trans. Image Process.* 27.2 (2018), pp. 806–821 (cit. on p. 9).

[FV13] B. Frénay and M. Verleysen. “Classification in the Presence of Label Noise: A Survey”. In: *IEEE Trans. Neural Netw. Learn. Syst.* 25.5 (2013), pp. 845–869 (cit. on p. 6).

[Get12] P. Getreuer. “Rudin-Osher-Fatemi Total Variation Denoising Using Split Bregman”. In: *Image Processing On Line* 2 (2012), pp. 74–95 (cit. on p. 128).

[GL14] N. Gillis and R. Luce. “Robust Near-Separable Nonnegative Matrix Factorization Using Linear Optimization”. In: *J. Mach. Learning Research* 15.1 (2014), pp. 1249–1280 (cit. on pp. 112, 119).

[GL18] N. Gillis and R. Luce. “A Fast Gradient Method for Nonnegative Sparse Regression with Self Dictionary”. In: *IEEE Trans. Image Process.* 27.1 (Jan. 2018), pp. 24–37. arXiv: 1610.01349 (cit. on pp. 66, 67, 112, 119).

[Goo+16] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep Learning*. Vol. 1. MIT press Cambridge, 2016 (cit. on p. 63).

[Hap93] B. Hapke. *Theory of Reflectance and Emittance Spectroscopy*. Cambridge university press, 1993 (cit. on p. 11).

[HDD13] L. Hong, A. S. Doumith, and B. D. Davison. “Co-Factorization Machines: Modeling User Interests and Predicting Individual Decisions in Twitter”. In: *Proc. of the Sixth ACM International Conference on Web Search and Data Mining*. ACM, 2013, pp. 557–566 (cit. on p. 10).

[HT96a] T. Hastie and R. Tibshirani. “Discriminant Analysis by Gaussian Mixtures”. In: *J. Roy. Stat. Soc. Ser. B* 58.1 (1996), pp. 155–176 (cit. on p. 6).

[HT96b] T. Hastie and R. Tibshirani. “Discriminant Analysis by Gaussian Mixtures”. In: *J. Roy. Stat. Soc. Ser. B* 58 (1996), pp. 155–176 (cit. on p. 40).

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer New York, 2009 (cit. on p. 54).

[Hub64] P. J. Huber. “Robust Estimation of a Location Parameter”. In: *Annals of Math. Stat.* 35.1 (1964), pp. 73–101 (cit. on p. 63).

[Hug68] G. Hughes. “On the Mean Accuracy of Statistical Pattern Recognizers”. In: *IEEE Trans. Inf. Theory* 14.1 (1968), pp. 55–63 (cit. on p. 6).

[IBP12] M.-D. Iordache, J. M. Bioucas-Dias, and A. Plaza. “Total Variation Spatial Regularization for Sparse Hyperspectral Unmixing”. In: *IEEE Trans. Geosci. Remote Sens.* 50.11 (Nov. 2012), pp. 4484–4502 (cit. on pp. 84, 86–88, 96).

[Idi13] J. Idier. *Bayesian Approach to Inverse Problems*. John Wiley & Sons, 2013 (cit. on p. 7).

[Jen+11] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. “Proximal Methods for Hierarchical Sparse Coding”. In: *J. Mach. Learning Research* 12.Jul (2011), pp. 2297–2334 (cit. on p. 129).

[JL11] A. K. Jain and S. Z. Li. *Handbook of Face Recognition*. Springer, 2011 (cit. on pp. 1, 5).

[JL98] L. O. Jimenez and D. A. Landgrebe. “Supervised Classification in High-Dimensional Space: Geometrical, Statistical, and Asymptotical Properties of Multivariate Data”. In: *IEEE Trans. Systems, Man, Cybernet. - Part C* 28.1 (Feb. 1998), pp. 39–54 (cit. on pp. 20, 22).

[JLD11] Z. Jiang, Z. Lin, and L. S. Davis. “Learning a Discriminative Dictionary for Sparse Coding via Label Consistent K-SVD”. In: *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*. June 2011, pp. 1697–1704 (cit. on pp. 9, 50, 51, 62, 71).

[Jol86] I. T. Jolliffe. *Principal Component Analysis*. en. 2nd ed. Springer Series in Statistics. New York: Springer-Verlag, 1986. ISBN: 978-0-387-95442-4 (cit. on pp. 53, 89).

[Kai+12] G. Kail, J.-Y. Tourneret, F. Hlawatsch, and N. Dobigeon. “Blind deconvolution of sparse pulse sequences under a minimum distance constraint: a partially collapsed Gibbs sampler method”. In: *IEEE Trans. Signal Process.* 60.6 (2012), pp. 2727–2743 (cit. on p. 30).

[Kar+14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. “Large-Scale Video Classification with Convolutional Neural Networks”. In: *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 1725–1732 (cit. on p. 6).

[KB05] D. M. Kline and V. L. Berardi. “Revisiting Squared-Error and Cross-Entropy Functions for Training Neural Network Classifiers”. In: *Neural Comput. Appl.* 14.4 (2005), pp. 310–318 (cit. on p. 54).

[KBV09] Y. Koren, R. Bell, and C. Volinsky. “Matrix Factorization Techniques for Recommender Systems”. In: *Computer* 8 (2009), pp. 30–37 (cit. on pp. 50, 51).

[Ker14] J. Kersten. “Simultaneous feature selection and Gaussian mixture model estimation for supervised classification problems”. In: *Patt. Recognition* 47.8 (2014), pp. 2582–2595 (cit. on pp. 20, 21).

[KSH12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagenet Classification with Deep Convolutional Neural Networks”. In: *Adv. in Neural Information Process. Systems.* 2012, pp. 1097–1105 (cit. on pp. 5, 6).

[Lag+15] A. Lagrange, B. Le Saux, A. Beaupere, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, and M. Ferecatu. “Benchmarking Classification of Earth-Observation Data: From Learning Explicit Features to Convolutional Networks”. In: *Proc. IEEE Int. Conf. Geosci. Remote Sens. (IGARSS)*. IEEE, 2015, pp. 4173–4176.

[Lag+17] A. Lagrange, M. Fauvel, S. May, and N. Dobigeon. “Un Modèle Bayésien Pour Le Démélange, La Segmentation et La Classification Robuste d’images Hyperspectrales”. In: *Actes du Colloque GRETSI*. 2017, pp. 1–4 (cit. on p. 19).

[Lag+18] A. Lagrange, M. Fauvel, S. May, and N. Dobigeon. “A Bayesian Model for Joint Unmixing and Robust Classification of Hyperspectral Images”. In: *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2018, pp. 3399–3403 (cit. on pp. 19, 55).

[Lag+19a] A. Lagrange, M. Fauvel, S. May, J. M. Bioucas-Dias, and N. Dobigeon. “Cofactorisation de Matrices Pour Le Démélange et La Classification Conjoints d’Images Hyperspectrales”. In: *Actes du Colloque GRETSI*. Aug. 2019 (cit. on p. 49).

[Lag+19b] A. Lagrange, M. Fauvel, S. May, J. M. Bioucas-Dias, and N. Dobigeon. “Matrix Cofactorization for Joint Unmixing and Classification of Hyperspectral Images”. In: *Proc. European Signal Process. Conf. (EUSIPCO)*. Sept. 2019 (cit. on p. 49).

[Lag+19c] A. Lagrange, M. Fauvel, S. May, J. Bioucas-Dias, and N. Dobigeon. “Matrix Cofactorization for Joint Representation Learning and Supervised Classification – Application to Hyperspectral Image Analysis”. In: *arXiv:1902.02597 [cs, eess]* (Feb. 2019). arXiv: 1902.02597 [cs, eess] (cit. on p. 49).

[Lag+19d] A. Lagrange, M. Fauvel, S. May, and N. Dobigeon. “Hierarchical Bayesian Image Analysis: From Low-Level Modeling to Robust Supervised Learning”. In: *Patt. Recognition* 85 (2019), pp. 26–36 (cit. on p. 19).

[Lag+19e] A. Lagrange, M. Fauvel, S. May, and N. Dobigeon. "Matrix Cofactorization for Joint Spatial-Spectral Unmixing of Hyperspectral Images". In: *arXiv:1907.08511 [cs, eess]* (July 2019). arXiv: 1907.08511 [cs, eess] (cit. on p. 83).

[Lan02] D. Landgrebe. "Hyperspectral Image Data Analysis". In: *IEEE Signal Process. Mag.* 19.1 (2002), pp. 17–28 (cit. on p. 11).

[LBP12] J. Li, J. M. Bioucas-Dias, and A. Plaza. "Spectral–Spatial Hyperspectral Image Segmentation Using Subspace Multinomial Logistic Regression and Markov Random Fields". In: *IEEE Trans. Geosci. Remote Sens.* 50.3 (2012), pp. 809–823 (cit. on p. 13).

[LC09] B. Luo and J. Chanussot. "Hyperspectral Image Classification Based on Spectral and Geometrical Features". In: *Proc. IEEE Workshop Mach. Learning for Signal Process. (MLSP)*. IEEE, 2009, pp. 1–6 (cit. on pp. 10, 11).

[LeC+98] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-Based Learning Applied to Document Recognition". In: *Proc. IEEE* 86.11 (1998), pp. 2278–2324 (cit. on pp. 1, 5).

[LFG17] A. Lagrange, M. Fauvel, and M. Grizonnet. "Large-scale feature selection with Gaussian mixture models for the classification of high dimensional remote sens. images". In: *IEEE Trans. Comput. Imag.* 3.2 (2017), pp. 230–242 (cit. on pp. 11, 74).

[Li+15a] J. Li, I. Dópido, P. Gamba, and A. Plaza. "Complementarity of Discriminative Classifiers and Spectral Unmixing Techniques for the Interpretation of Hyperspectral Images". In: *IEEE Trans. Geosci. Remote Sens.* 53.5 (2015), pp. 2899–2912 (cit. on p. 12).

[Li+15b] Z. Li, J. Liu, J. Tang, and H. Lu. "Robust Structured Subspace Learning for Data Representation". In: *IEEE Trans. Patt. Anal. Mach. Intell.* 37.10 (2015), pp. 2085–2098 (cit. on p. 7).

[Li09] S. Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer Science & Business Media, 2009 (cit. on pp. 21, 22, 26, 93).

[Liu+18] Y. Liu, F. Condessa, J. M. Bioucas-Dias, J. Li, P. Du, and A. Plaza. "Convex Formulation for Multiband Image Classification With Superpixel-Based Spatial Regularization". In: *IEEE Trans. Geosci. Remote Sens.* 56.5 (May 2018), pp. 2704–2721 (cit. on p. 62).

[LKC15] T. Lillesand, R. W. Kiefer, and J. Chipman. *Remote Sens. and Image Interpretation*. John Wiley & Sons, 2015 (cit. on pp. 1, 5).

[Low99] D. G. Lowe. "Object Recognition from Local Scale-Invariant Features." In: *Proc. IEEE Int. Conf. Computer Vision (ICCV)*. Vol. 99. 1999, pp. 1150–1157 (cit. on p. 5).

[LS99] D. D. Lee and H. S. Seung. "Learning the Parts of Objects by Non-Negative Matrix Factorization". In: *Nature* 401.6755 (1999), p. 788 (cit. on pp. 7, 53, 89).

[Ma+13] W.-K. Ma, J. M. Bioucas-Dias, P. Gader, T.-H. Chan, N. Gillis, A. Plaza, A. Ambikapathi, and C.-Y. Chi. "Signal Processing Perspective on Hyperspectral Unmixing: Insights from remote sens." In: *IEEE Signal Process. Mag.* (2013) (cit. on p. 34).

[Ma+14] W.-K. Ma, J. M. Bioucas-Dias, J. Chanussot, and P. Gader. "Signal and image processing in hyperspectral remote sens. [from the guest editors]". In: *IEEE Signal Process. Mag.* 31.1 (2014), pp. 22–23 (cit. on p. 33).

[Mag+16] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. "Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification". In: *IEEE Trans. Geosci. Remote Sens.* 55.2 (2016), pp. 645–657 (cit. on p. 6).

[Mai+09] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach. "Supervised Dictionary Learning". In: *Adv. in Neural Information Process. Systems*. 2009, pp. 1033–1040 (cit. on p. 9).

[Mal12] S. Mallat. "Group Invariant Scattering". In: *Comm. Pure Appl. Math.* 65.10 (2012), pp. 1331–1398 (cit. on pp. 113, 119).

[Man+14] D. Manolakis, E. Truslow, M. Pieper, T. Cooley, and M. Brueggeman. "Detection algorithms in hyperspectral imaging systems: An overview of practical algorithms". In: *IEEE Signal Process. Mag.* 31.1 (Jan. 2014), pp. 24–33 (cit. on p. 33).

[MBP12] J. Mairal, F. Bach, and J. Ponce. "Task-Driven Dictionary Learning". In: *IEEE Trans. Patt. Anal. Mach. Intell.* 34.4 (2012), pp. 791–804 (cit. on pp. 8, 9, 50, 51, 53).

[MIC12] S. Moussaoui, J. Idier, and E. Chouzenoux. "Alternating Direction Algorithms for Constrained Sparse Regression: Application to Hyperspectral Unmixing". In: *Proc. IEEE GRSS Workshop Hyperspectral Image SIgnal Process.: Evolution in Remote Sens. (WHISPERS)*. IEEE, June 2012 (cit. on p. 88).

[Mol+06] J. Moller, A. N. Pettitt, R. Reeves, and K. K. Berthelsen. "An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants". In: *Biometrika* 93.2 (June 2006), pp. 451–458 (cit. on p. 31).

[Moo+17] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. "Universal Adversarial Perturbations". In: *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1765–1773 (cit. on pp. 1, 7).

[MP17] M. Minsky and S. A. Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT press, 2017 (cit. on p. 7).

[Myn+95] R. B. Myneni, F. G. Hall, P. J. Sellers, and A. L. Marshak. "The Interpretation of Spectral Vegetation Indexes". In: *IEEE Trans. Geosci. Remote Sens.* 33.2 (1995), pp. 481–486 (cit. on p. 102).

[ND05] J. M. P. Nascimento and J. M. B. Dias. "Vertex Component Analysis: A Fast Algorithm to Unmix Hyperspectral Data". In: *IEEE Trans. Geosci. Remote Sens.* 43 (2005), pp. 898–910 (cit. on pp. 44, 92, 95).

[Nes83] Y. E. Nesterov. “A Method for Solving the Convex Programming Problem with Convergence Rate $\mathcal{O}(\frac{1}{K^2})$”. In: *Dokl. Akad. Nauk Sssr.* Vol. 269. 1983, pp. 543–547 (cit. on pp. 111, 118).

[PB01] M. Pesaresi and J. A. Benediktsson. “A New Approach for the Morphological Segmentation of High-Resolution Satellite Imagery”. In: *IEEE Trans. Geosci. Remote Sens.* 39.2 (2001), pp. 309–320 (cit. on p. 5).

[Ped+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. “Scikit-learn: Machine Learning in Python”. In: *J. Mach. Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 71).

[Pel+17] C. Pelletier, S. Valero, J. Inglada, N. Champion, C. Marais Sicre, and G. Dedieu. “Effect of Training Class Label Noise on Classification Performances for Land Cover Mapping with Satellite Image Time Series”. In: *Remote Sens.* 9 (2017), p. 173 (cit. on pp. 20, 22, 46).

[Per+12] M. Pereyra, N. Dobigeon, H. Batatia, and J. Y. Tourneret. “Segmentation of Skin Lesions in 2-D and 3-D Ultrasound Images Using a Spatially Coherent Generalized Rayleigh Mixture Model”. In: *IEEE Trans. Med. Imag.* 31 (2012), pp. 1509–1520 (cit. on pp. 8, 25).

[Per+13] M. Pereyra, N. Dobigeon, H. Batatia, and J.-Y. Tourneret. “Estimating the granularity coefficient of a Potts-Markov random field within an MCMC algorithm”. In: *IEEE Trans. Image Process.* 22.6 (June 2013), pp. 2385–2397 (cit. on p. 31).

[Per+15] M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. O. Hero, and S. McLaughlin. “A Survey of Stochastic Simulation and Optimization Methods in Signal Processing”. In: *IEEE J. Sel. Topics Signal Process.* 10.2 (2015), pp. 224–241 (cit. on p. 8).

[Pla+09] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, et al. “Recent Advances in Techniques for Hyperspectral Image Processing”. In: *Remote Sens. Environment* 113 (2009), S110–S122 (cit. on pp. 5, 11).

[Pom+14] F. Pompili, N. Gillis, P.-A. Absil, and F. Glineur. “Two Algorithms for Orthogonal Nonnegative Matrix Factorization with Application to Clustering”. In: *Neurocomputing* 141 (2014), pp. 15–25 (cit. on pp. 55, 64, 90).

[PS16] T. Pock and S. Sabach. “Inertial Proximal Alternating Linearized Minimization (iPALM) for Nonconvex and Nonsmooth Problems”. In: *SIAM J. Imag. Sci.* 9.4 (2016), pp. 1756–1787 (cit. on pp. 111, 118).

[PT94] P. Paatero and U. Tapper. "Positive Matrix Factorization: A Non-Negative Factor Model with Optimal Utilization of Error Estimates of Data Values". In: *Environmetrics* 5.2 (1994), pp. 111–126 (cit. on p. 53).

[RBE10] R. Rubinstein, A. M. Bruckstein, and M. Elad. "Dictionaries for Sparse Representation Modeling". In: *Proc. IEEE* 98.6 (2010), pp. 1045–1057 (cit. on p. 9).

[RC04] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. en. 2nd ed. Springer Texts in Statistics. New York: Springer-Verlag, 2004. ISBN: 978-0-387-21239-5 (cit. on p. 8).

[RCP14] A. Repetti, E. Chouzenoux, and J.-C. Pesquet. "A Preconditioned Forward-Backward Approach with Application to Large-Scale Nonconvex Spectral Unmixing Problems". In: *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2014, pp. 1498–1502 (cit. on p. 8).

[Ris+10] L. Risser, T. Vincent, J. Idier, F. Forbes, and 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". In: *J. Signal Process. Syst.* 60.1 (July 2010), pp. 1–14 (cit. on p. 31).

[RPE12] R. Rubinstein, T. Peleg, and M. Elad. "Analysis K-SVD: A Dictionary-Learning Algorithm for the Analysis Sparse Model". In: *IEEE Trans. Image Process.* 61.3 (2012), pp. 661–677 (cit. on p. 7).

[SBC97] D. M. Strong, P. Blomgren, and T. F. Chan. "Spatially Adaptive Local-Feature-Driven Total Variation Minimizing Image Restoration". In: *Statistical and Stochastic Methods in Image Processing II*. Vol. 3167. International Society for Optics and Photonics, 1997, pp. 222–234 (cit. on p. 67).

[SM02] G. Shaw and D. Manolakis. "Signal Processing for Hyperspectral Image Exploitation". In: *IEEE Signal Process. Mag.* 19.1 (2002), pp. 12–16 (cit. on p. 11).

[SNT15] X. Sun, N. M. Nasrabadi, and T. D. Tran. "Task-Driven Dictionary Learning for Hyperspectral Image Classification with Structured Sparsity Constraints". In: *IEEE Trans. Geosci. Remote Sens.* 53.8 (2015), pp. 4457–4471 (cit. on p. 56).

[Sun+17] Y. Sun, J. M. Bioucas-Dias, X. Zhang, Y. Liu, and A. Plaza. "A New Classification-Oriented Endmember Extraction and Sparse Unmixing Approach for Hyperspectral Data". In: *Proc. IEEE Int. Conf. Geosci. Remote Sens. (IGARSS)*. July 2017, pp. 3644–3647 (cit. on p. 13).

[SW14] C. Shi and L. Wang. "Incorporating Spatial Information in Spectral Unmixing: A Review". In: *Remote Sens. Environment* 149 (June 2014), pp. 70–87 (cit. on pp. 84, 86).

[Tar+10] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson. "SVM- and MRF-Based Method for Accurate Classification of Hyperspectral Images". In: *IEEE Geosci. Remote Sens. Lett.* 7 (2010), pp. 736–740 (cit. on pp. 21, 22).

[TB99] M. E. Tipping and C. M. Bishop. “Probabilistic Principal Component Analysis”. In: *J. Roy. Stat. Soc. Ser. B* 61.3 (Jan. 1999), pp. 611–622 (cit. on p. 24).

[TDT15] P.-A. Thouvenin, N. Dobigeon, and J.-Y. Tourneret. “Hyperspectral Unmixing with Spectral Variability Using a Perturbed Linear Mixing Model”. In: *IEEE Trans. Image Process.* 64.2 (2015), pp. 525–538 (cit. on pp. 84, 86, 88).

[TDT18] P.-A. Thouvenin, N. Dobigeon, and J.-Y. Tourneret. “Partially Asynchronous Distributed Unmixing of Hyperspectral Images”. In: *IEEE Trans. Geosci. Remote Sens.* (2018) (cit. on pp. 111, 118).

[Thi+14] B. Thirion, G. Varoquaux, E. Dohmatob, and J.-B. Poline. “Which fMRI Clustering Gives Good Brain Parcellations?” In: *Frontiers in neuroscience* 8 (2014), p. 167 (cit. on p. 6).

[Tho+10] D. R. Thompson, L. Mandrake, M. S. Gilmore, and R. Castano. “Superpixel Endmember Detection”. In: *IEEE Trans. Geosci. Remote Sens.* 48.11 (2010), pp. 4023–4033 (cit. on pp. 84, 86).

[UFD18] T. Uezato, M. Fauvel, and N. Dobigeon. “Hyperspectral Image Unmixing with LiDAR Data-Aided Spatial Regularization”. In: *IEEE Trans. Geosci. Remote Sens.* (2018) (cit. on pp. 63, 67, 88).

[Van+93] G. Vane, R. O. Green, T. G. Chrien, H. T. Enmark, E. G. Hansen, and W. M. Porter. “The Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)”. In: *Remote Sens. Environment* 44.2-3 (1993), pp. 127–143 (cit. on p. 11).

[Vas+15] R. K. Vasudevan, A. Belianinov, A. G. Gianfrancesco, A. P. Baddorf, A. Tselev, S. V. Kalinin, and S. Jesse. “Big data in reciprocal space: Sliding fast Fourier transforms for determining periodicity”. In: *Appl. Phys. Lett.* 106.9 (2015), p. 091601 (cit. on p. 89).

[Vas+16] R. K. Vasudevan, M. Ziatdinov, S. Jesse, and S. V. Kalinin. “Phases and Interfaces from Real Space Atomically Resolved Data: Physics-Based Deep Data Image Analysis”. In: *Nano Lett.* 16.9 (2016), pp. 5574–5581 (cit. on p. 89).

[Vas+18] R. K. Vasudevan, N. Laanait, E. M. Ferragut, K. Wang, D. B. Geohegan, K. Xiao, M. Ziatdinov, S. Jesse, O. Dyck, and S. V. Kalinin. “Mapping mesoscopic phase evolution during E-beam induced transformations via deep learning of atomically resolved images”. In: *Npj Comput. Mater.* 4 (2018) (cit. on p. 89).

[VDC19] M. Vono, N. Dobigeon, and P. Chainais. “Split-and-Augmented Gibbs Sampler —Application to Large-Scale Inference Problems”. In: *IEEE Trans. Image Process.* 67.6 (2019), pp. 1648–1661 (cit. on p. 111).

[Vil+11a] A. Villa, J. Chanussot, J. A. Benediktsson, and C. Jutten. “Unsupervised Classification and Spectral Unmixing for Sub-Pixel Labelling”. In: *Proc. IEEE Int. Conf. Geosci. Remote Sens. (IGARSS)*. July 2011, pp. 71–74 (cit. on p. 12).

[Vil+11b] A. Villa, J. Chanussot, J. A. Benediktsson, and C. Jutten. "Spectral Unmixing for the Classification of Hyperspectral Images at a Finer Spatial Resolution". In: *IEEE J. Sel. Topics Signal Process.* 5 (2011), pp. 521–533 (cit. on pp. 12, 34).

[Vil11] A. Villa. "Advanced spectral unmixing and classification methods for hyperspectral remote sens. data". PhD thesis. Université Grenoble Alpes, July 29, 2011 (cit. on p. 11).

[Wan+17] X. Wang, Y. Zhong, L. Zhang, and Y. Xu. "Spatial Group Sparsity Regularized Nonnegative Matrix Factorization for Hyperspectral Unmixing". In: *IEEE Trans. Geosci. Remote Sens.* 55.11 (Nov. 2017), pp. 6287–6304 (cit. on pp. 84, 86).

[WB11] C. Wang and D. M. Blei. "Collaborative Topic Modeling for Recommending Scientific Articles". In: *Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD)*. ACM, 2011, pp. 448–456 (cit. on pp. 10, 50, 51, 85, 86).

[WG13] C. S. Won and R. M. Gray. *Stochastic image processing*. Information Technology: Transmission, Processing, and Storage. New York: Springer Science & Business Media, 2013 (cit. on pp. 20, 21).

[Win99] M. E. Winter. "N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data". In: *Proc. SPIE Imaging Spectrometry V*. Ed. by M. R. Descour and S. S. Shen. Vol. 3753. 1. Denver, CO, USA: SPIE, 1999, pp. 266–275 (cit. on p. 89).

[Wu82] F.-Y. Wu. "The Potts Model". In: *Rev. Mod. Phys.* 54 (1982), p. 235 (cit. on p. 26).

[WYZ19] Y. Wang, W. Yin, and J. Zeng. "Global Convergence of ADMM in Nonconvex Nonsmooth Optimization". In: *SIAMSC* 78.1 (2019), pp. 29–63 (cit. on p. 8).

[Xu+19] S. Xu, J. Li, M. Khodadadzadeh, A. Marinoni, P. Gamba, and B. Li. "Abundance-Indicated Subspace for Hyperspectral Classification With Limited Training Samples". In: *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.* 12.4 (Apr. 2019), pp. 1265–1278 (cit. on p. 13).

[Yan+11] M. Yang, L. Zhang, X. Feng, and D. Zhang. "Fisher Discrimination Dictionary Learning for Sparse Representation". In: *Proc. IEEE Int. Conf. Computer Vision (ICCV)*. IEEE, 2011, pp. 543–550 (cit. on p. 62).

[Yoo+10] J. Yoo, M. Kim, K. Kang, and S. Choi. "Nonnegative Matrix Partial Co-Factorization for Drum Source Separation". In: *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2010, pp. 1942–1945 (cit. on pp. 10, 50, 51, 65, 85, 86).

[Yu13] Y.-L. Yu. "On Decomposing the Proximal Map". In: *Adv. in Neural Information Process. Systems*. 2013, pp. 91–99 (cit. on pp. 59, 130).

[YYI12] N. Yokoya, T. Yairi, and A. Iwasaki. "Coupled Nonnegative Matrix Factorization Unmixing for Hyperspectral and Multispectral Data Fusion". In: *IEEE Trans. Geosci. Remote Sens.* 50.2 (2012), pp. 528–537 (cit. on pp. 10, 50, 52, 65, 85, 86).

[ZBS01] Y. Zhang, M. Brady, and S. Smith. "Segmentation of Brain MR Images through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm". In: *IEEE Trans. Med. Imag.* 20 (2001), pp. 45–57 (cit. on pp. 21, 22, 62).

[ZD16] W. Zhao and S. Du. "Spectral–Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach". In: *IEEE Trans. Geosci. Remote Sens.* 54.8 (2016), pp. 4544–4554 (cit. on p. 11).

[Zha+18a] S. Zhang, J. Li, H. C. Li, C. Deng, and A. Plaza. "Spectral-Spatial Weighted Sparse Regression for Hyperspectral Image Unmixing". In: *IEEE Trans. Geosci. Remote Sens.* 56.6 (June 2018), pp. 3265–3276 (cit. on pp. 84, 86).

[Zha+18b] Z. Zhang, W. Jiang, J. Qin, L. Zhang, F. Li, M. Zhang, and S. Yan. "Jointly Learning Structured Analysis Discriminative Dictionary and Analysis Multiclass Classifier". In: *IEEE Trans. Neural Netw. Learn. Syst.* 29.8 (2018), pp. 3798–3814 (cit. on pp. 50, 51).

[ZL10] Q. Zhang and B. Li. "Discriminative K-SVD for Dictionary Learning in Face Recognition". In: *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 2691–2698 (cit. on pp. 4, 9, 15, 50, 51, 54, 56, 71).

[ZP01] M. Zibulevsky and B. A. Pearlmutter. "Blind Source Separation by Sparse Decomposition in a Signal Dictionary". In: *Neural Comput.* 13.4 (2001), pp. 863–882 (cit. on pp. 50, 51).

[ZP09] M. Zortea and A. Plaza. "Spatial Preprocessing for Endmember Extraction". In: *IEEE Trans. Geosci. Remote Sens.* 47.8 (2009), pp. 2679–2693 (cit. on pp. 84, 86).