Endmember Variability in hyperspectral image unmixing

Lucas Drumetz

UNIVERSITÉ GRENOBLE ALPES

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE ALPES

Spécialité : **Signal, Image, Parole, Télécoms**

Arrêté ministériel : 7 août 2006

Présentée par
Lucas DRUMETZ

Thèse dirigée par **Christian JUTTEN** et
codirigée par **Jocelyn CHANUSSOT**

préparée au sein du laboratoire **GIPSA-lab**
dans **l'école doctorale Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS)**

Endmember variability in hyperspectral image unmixing

Thèse soutenue publiquement le **25/10/2016**,
devant le jury composé de:

Yannick DEVILLE
Professeur, Université de Toulouse, Président
David BRIE
Professeur, Université de Lorraine, Rapporteur
Paolo GAMBA
Professeur, Université de Pavie, Rapporteur
José BIOUCAS DIAS
Professeur associé, Université de Lisbonne, Examinateur
Saïd MOUSSAOUI
Professeur, Ecole Centrale de Nantes, Examinateur
Devis TUIA
Professeur assistant, Université de Zurich, Examinateur
Christian JUTTEN
Professeur, Université Grenoble Alpes, GIPSA-lab, Directeur de thèse
Jocelyn CHANUSSOT
Professeur, Grenoble-INP, GIPSA-lab, Co-directeur de thèse

UNIVERSITÉ DE GRENOBLE ALPES

ÉCOLE DOCTORALE EEATS

Electronique, Electrotechnique, Automatique, Traitement du Signal

T H È S E

pour obtenir le titre de

docteur en sciences

de l'Université de Grenoble Alpes
Mention : Signal, Image, Parole, Télécoms

Présentée et soutenue par

Lucas DRUMETZ

Endmember variability in hyperspectral image unmixing

Thèse dirigée par Christian JUTTEN et Jocelyn CHANUSSOT

préparée au laboratoire Grenoble Image Parole Signal Automatique (GIPSA-lab)

soutenue le 25/10/2016

Jury :

Président :	Yannick DEVILLE	-	Université de Toulouse, France
Rapporteurs :	David BRIE	-	Université de Lorraine, France
	Paolo GAMBA	-	Université de Pavie, Italie
Examinateurs :	José BIOUCAS DIAS	-	Université de Lisbonne, Portugal
	Saïd MOUSSAOUI	-	Ecole Centrale de Nantes, France
	Devis TUIA	-	Université de Zurich, Suisse
Directeur :	Christian JUTTEN	-	Université Grenoble Alpes, GIPSA-lab, France
Co-directeur :	Jocelyn CHANUSSOT	-	Grenoble-INP, GIPSA-lab, France

Acknowledgments

Après trois années (et même un peu plus) passées au GIPSA-lab, dont je garderai un excellent souvenir, il faut se rendre à l'évidence: j'ai terminé mon doctorat, un moment qu'on s'imagine beaucoup en début de thèse, mais qui arrive finalement sans trop prévenir. Ce n'est finalement pas étonnant, dans un laboratoire qui est parfait pour y préparer une thèse, très dynamique, aussi bien scientifiquement que dans l'ambiance excellente qui y règne. Le moment est donc venu de mettre un point final à ce manuscrit à travers ces remerciements.

I would first like to thank all the members of my committee, for having kindly accepted to evaluate my work, and for coming to Grenoble for my defense. In particular, I would especially like to thank José Bioucas-Dias, whom I have met while I was starting in the lab, preparing my master thesis. Thanks for providing some advice and insight at that time, and also thanks for being one of the people who introduced me to spectral unmixing, when you were giving an overview talk on the subject.

Ensuite, je voudrais bien sûr remercier mes deux directeurs de thèse, Jocelyn et Christian. Merci de m'avoir fait confiance et de m'avoir proposé ce sujet de thèse. J'en profite pour remercier le Conseil Européen de la Recherche (European Research Coucil, ERC), pour avoir financé mes travaux dans le cadre du projet ERC CHESS[1] de Christian. Vous avez été les encadrants parfaits pour moi, en me guidant dans mes travaux et en m'apportant vos conseils avisés tout en me donnant une grande autonomie. J'ai pris beaucoup de plaisir à explorer mon sujet avec vous deux. Plus généralement, merci de m'avoir fait découvrir et apprécier le métier de chercheur. Merci de m'avoir soutenu dans tous mes choix. Jocelyn, je voudrais aussi te remercier pour m'avoir permis de passer ces quelques semaines à UCLA durant ma troisième année, et pour me permettre d'y retourner très prochainement.

I would also like to thank Miguel V., it has been a great pleasure to work with you. Thanks for "recruiting" my during my first year to work with you on Local Spectral Unmixing, which allowed me to interact with you throughout the thesis. I have learned so much from you, from all those countless discussions in your office or during meetings. This thesis and my knowledge about the topic would definitely not be the same without you. You were my unofficial advisor and I will always be grateful to you for this.

Merci également à Mauro pour avoir bien voulu me faire confiance pour mon stage de 3$^{\text{ème}}$ année à Phelma, qui m'a permis de découvrir GIPSA-lab et la recherche. Je ne serais probablement pas devenu docteur sans ce stage! C'est aussi l'occasion de remercier Guillaume, qui pouvait aussi être considéré comme mon encadrant non officiel pendant ce stage. C'était là aussi le début d'une collaboration qui allait perdurer tout au long de la thèse, et qui continue, pour longtemps encore j'espère.

I want to thank all the people I have interacted with during this thesis from a scientific

[1] Cette thèse a été entièrement financée par le Conseil Européen de la Recherche, via le septième programme-cadre de la Commission Européenne, par la bourse ERC AdG-2012-320684 du projet CHESS.

point of view: Manu and Rubén for working with me during my first year (both of you are one of the reasons there is a little bit of physics in my thesis), and Sylvain Douté, as well. Merci à Simon, j'ai aussi beaucoup appris grâce à toi (et quel dommage que tu sois resté si peu au labo!). Thanks to Ronald, for a few very fruitful discussions and insights you provided me, and to Easter as well, for all those discussions which helped me learn about convex optimization. Thanks to Mark Berman for these interactions when you were visiting us in Grenoble. I want to thank people I have met at UCLA: Wotao, from whom I learned a lot about optimization in a short period of time, Andrea Bertozzi for welcoming me to the Department of Mathematics at UCLA, Travis, for working with me on what started off as a small project, which has become very profitable. Merci à Alex pour avoir été mon compagnon français durant ce séjour à Los Angeles. A très bientôt!

Dans le même registre, merci aussi à Vincent pour cette visite de Tokyo en ta compagnie durant WHISPERS 2015 dont je garde un très bon souvenir, et qui m'a donné envie (comme toi j'imagine) d'y retourner!

Merci à Cyrille d'avoir été mon mentor pour ce qui était des mes enseignements en maths à l'IUT. C'était un plaisir de donner ces cours et d'échanger avec toi à ce sujet.

Vient le moment de remercier toute la Team GIPSA, ceux qui font de ce labo l'endroit parfait pour un doctorant. Merci à tous pour toutes ces coinches (et plus récemment ces parties de baby-foot) endiablées, ces discussions très sérieuses sur des sujets débiles, ou bien débiles sur des sujets très sérieux (les discussions dites "de vendredi" à la pause), ces soirées... Merci donc à (en espérant n'oublier personne): Victor (dont le seul tort aura été d'arriver trop tard au GIPSA), Miguel G.(G.) (idem), Taia l'embrouille, Quentin (que je tiens à remercier tout spécialement pour cette fameuse discussion sur l'introduction de mon manuscrit ;-)), Tim G. (pour toutes ces discussions et digressions sur tout et n'importe quoi -mais souvent sur mes problèmes informatiques en tout genre, désolé ;-) - dans ce fameux bureau D1189), Tim 2G., Raph, Guillaume (à quand le prochain ~~buffet~~ road trip aux US?) Alexis, Marielle, Paolo, Mélisande, Sophie, Céline, Cindy, Pascal, Arnaud, Aude, Robin, Manu, Pierre, Marc, Marion, Alexandre, Pedro et Pedro... Merci à tous, malheureusement je m'en vais bientôt en voyage (voyage)... D'ailleurs, j'y pense, vous avez perdu, comprenne qui pourra.

Un remerciement tout particulier à Anaïs, qui m'aura accompagné pendant la majeure partie de mes années grenobloises. Merci pour ton soutien et ta bonne humeur tout au long de ma thèse. Je suis content que tu aies enfin eu un aperçu pendant la soutenance de ce que je fais de mes journées. ;-)

Merci enfin à ma famille, qui m'a toujours soutenu. Merci à mes parents et grands parents, pour m'avoir très tôt appris la valeur du travail bien fait, et à Maxime, pour avoir été un modèle à suivre depuis tout petit. Merci à tous d'être venu me voir soutenir, même parfois avec l'assurance de n'y rien comprendre.

Contents

List of Acronyms

ADMM	- Alternating Direction Method of Multipliers
AEB	- Automated Endmember Bundles
AIC	- Akaike Information Criterion
AL	- Augmented Lagrangian
ALS	- Alternating Least Squares
ANC	- Abundance Nonnegativity Constraint
ASC	- Abundance Sum-to-one Constraint
BCM	- Beta Compositional Model
BIC	- Bayesian Information Criterion
BPT	- Binary Partition Tree
BSS	- Blind Source Separation
CD	- Coordinate Descent
CLSU	- (partially) Constrained Least Squares Unmixing
CP(D)	- Canonical Polyadic Decomposition
DECA	- DEpendent Component Analysis
DFC	- Data Fusion Contest
D-HIDENN	- Denoised Hyperspectral Intrinsic Dimensionality Estimation through Nearest Neighbor distance ratios
DTM	- Digital Terrain Model
ECG	- ElectroCardioGram
EEA	- Endmember Extraction Algorithm
EEG	- ElectroEncephaloGram
ELMM	- Extended Linear Mixing Model
FCLSU	- Fully Constrained Least Squares Unmixing
FDN	- Fisher Discriminant Nullspace
FISTA	- Fast Iterative Shrinkage Thresholding Algorithm
FOV	- Field of View
FFT	- Fast Fourier Transform
GNCM	- Generalized Normal Compositional Model
HSI	- HyperSpectral Image
HFC	- Harsanyi-Farrand-Chang algorithm

HIDENN - Hyperspectral Intrinsic Dimensionality Estimation through Nearest Neighbor distance ratios
HySIME - Hyperspectral Subspace Identification by Minimum Error
ICA - Independent Component Analysis
ICE - Iterative Constrained Endmembers algorithm
ID - Intrinsic Dimensionality
IQR - Inter Quartile Range
KKT - Karush-Kuhn-Tucker
LASSO - Least Absolute Shrinkage Selection Operator
LiDAR - Light Detection And Ranging
LMM - Linear Mixing Model
LSU - Local Spectral Unmixing
MAP - Maximum A Posteriori
MESMA - Multiple Endmember Spectral Mixture Analysis
MLE - Maximum Likelihood Estimator
MMOCA - Modified Maximum Orthogonal Complement Algorithm
MMSE - Minimum Mean Squared Error estimator
MNF - Maximum Noise Fraction
MOCA - Maximum Orthogonal Complement Algorithm
MVC-NMF - Minimum Volume Constraint-Nonnegative Matrix Factorization
MVSA - Minimum Volume Simplex Analysis
NABO - Negative ABundance Oriented unmixing algorithm
NCM - Normal Compositional Model
NMF - Nonnegative Matrix Factorization
NWHFC - Noise Whitened HFC algorithm
ODM - Outlier Detection Method
PCA - Principal Component Analysis
PDF - Probability Density Function
PLMM - Perturbed Linear Mixing Model
PPI - Pixel Purity Index
ProCoALS - Projected and Compressed Alternating Least Squares
RMSE - Root Mean Squared Error
RMT - Random Matrix Theory
RSSE - Robust SubSpace Estimator

SAM	- Spectral Angle Mapper
S-CLSU	- Scaled (partially) Constrained Least Squares Unmixing
SISAL	- Simplex Identification via Split Augmented Lagrangian
SML	- Second Moment Linear dimensionality
SNR	- Signal to Noise Ratio
SSA	- Single Scattering Albedo
SSE	- Signal Subspace Estimator
SU	- Spectral Unmixing
SUnSAL	- Sparse Unmixing by variable Splitting and Augmented Lagrangian
SV	- Spectral Variability
SVD	- Singular Value Decomposition
SVM	- Support Vector Machine
TV	- Total Variation
UAV	- Unmanned Aerial Vehicle
USGS	- United States Geological Survey
VCA	- Vertex Component Analysis

Notations

Vectors, matrices and tensors

s	- Scalars (normal case)
$\mathbf{s}$	- Vectors (normal case, bold font)
$\mathbf{S}$	- Matrices (upper case, bold font)
$\boldsymbol{\mathcal{S}}$	- Third-order tensors (calligraphy, upper case, bold font)

Sets and special matrices

$\mathbb{N}$	- Natural numbers
$\mathbb{R}^n$	- n-dimensional vectors with real coefficients
$\mathbb{R}^{m\times n}$	- m by n matrices with real coefficients
$\mathbb{R}^{m\times n\times p}$	- m by n by p tensors with real coefficients
$\mathbb{R}^n_+$	- Nonnegative orthant for $\mathbb{R}^n$ (extends to matrices and tensors)
$\mathcal{P}(\cdot)$	- Power set of a set
Δ_P	- $P-1$ unit simplex
$\mathbb{1}_.$	- Vector or matrix of ones whose size is given in index
$\mathbf{I}_.$	- Identity matrix whose size is given in index
$\mathbf{0}$	- Zero vector or matrix (the size can also be given in index)

Operations and Operators

$\odot$	- Schur-Hadamard product
$\oslash$	- Termwise quotient
$\times_i$	- Tensor-matrix product along the i^{th} mode
$\mathcal{H}$	- Linear operator (calligraphy letters)
$\mathcal{H}^*$	- Adjoint operator of $\mathcal{H}$
$\circ$	- Composition of operators
$\star$	- Circular convolution
$\text{vec}(\mathbf{A})$	- Vectorization of matrix $\mathbf{A}$ (stacking the columns)
$\text{diag}(\mathbf{d})$	- Diagonal matrix whose diagonal elements are the elements of vector $\mathbf{d}$
$\det(\mathbf{A})$	- Determinant of square matrix $\mathbf{A}$

$^\top$ - Matrix transposition

Norms

$||\cdot||_p$ - $\mathcal{L}_p$ norm of a vector, for $p \geq 1$ and $p = \infty$. For $0 < p < 1$, it is only a quasinorm. For $p = 0$, number of nonzero elements of a vector

$||\cdot||_{p,q}$ - $\mathcal{L}_{p,q}$ two-level mixed norm of a matrix

$||\cdot||_{p,q,r}$ - $\mathcal{L}_{p,q,r}$ three-level mixed norm of a third-order tensor

$||\cdot||_{G,p,q}$ - $\mathcal{L}_{G,p,q}$ two-level mixed norm of a vector (or matrix) endowed with the group structure G on its coefficients

$||\cdot||_F$ - Frobenius norm of a matrix or tensor: $||\mathbf{A}||_F = ||\text{vec}(\mathbf{A})||_2$

$|||\cdot|||$ - Operator norm of a linear operator

Convex Optimization

∇_f - Gradient field of differentiable function f

∂_f - Subdifferential of convex function f

f^* - Convex conjugate of function f

$\Gamma_0(\mathbb{R}^n)$ - Set of lower-semicontinuous extended real-valued convex functions on $\mathbb{R}^n$

$\mathbf{prox}_f$ - Proximal operator of function $f \in \Gamma_0(\mathbb{R}^n)$

β - Lipschitz constant of $f \in \Gamma_0(\mathbb{R}^n)$

$\mathcal{I}_C$ - Indicator function of a convex set C of $\mathbb{R}^n$

$\mathbf{proj}_C$ - Projection on convex set C of $\mathbb{R}^n$

$(\mathbf{x})_+$ - Projection of $\mathbf{x} \in \mathbb{R}^n$ on $\mathbb{R}^n_+$

$\lambda.$ - Regularization parameter

Spectral Unmixing

L - Number of spectral bands

d - Intrinsic Dimensionality of a dataset

P - Number of endmembers

m - Vertical spatial dimension of a HSI

n - Horizontal spatial dimension of a HSI

N - Number of pixels ($N = mn$)

$\mathbf{X}$ - Data matrix ($\mathbf{X} \in \mathbb{R}^{L \times N}$)

$\mathbf{E}$	- Noise matrix ($\mathbf{E} \in \mathbb{R}^{L \times N}$)
$\mathbf{Y}$	- Noiseless signal matrix ($\mathbf{Y} \in \mathbb{R}^{L \times N}$)
$\mathbf{A}$	- Abundance matrix ($\mathbf{A} \in \mathbb{R}^{P \times N}$)
$\mathbf{S}$	- (constant) Endmember matrix ($\mathbf{S} \in \mathbb{R}^{L \times P}$)
$\mathbf{S}_0$	- Reference endmember matrix
$\boldsymbol{\Psi}$	- Scaling factor matrix ($\boldsymbol{\Psi} \in \mathbb{R}^{P \times N}$)
$\boldsymbol{\psi}_k$	- Diagonal scaling factor matrix for pixel or subset k ($\boldsymbol{\psi}_k \in \mathbb{R}^{P \times P}$)
$\mathbf{B}$	- Spectral bundle or dictionary ($\mathbf{B} \in \mathbb{R}^{L \times Q}$), where Q is the number of atoms/candidate endmembers in the dictionary
ρ	- Bidirectional reflectance
ω	- Single scattering albedo
θ_0	- Incidence angle
θ	- Emergence angle
g	- Phase angle
ϕ	- Azimuth angle

Probability theory, Statistics

$\mathbf{K}_.$	- Empirical covariance matrix of a dataset in index ($\mathbf{K}_. \in \mathbb{R}^{L \times L}$)
$\mathbf{R}_.$	- Empirical correlation matrix of a dataset in index ($\mathbf{R}_. \in \mathbb{R}^{L \times L}$)
$p_.(\cdot)$	- Probability Density Function
$p_{.\|.}(\cdot\|\cdot)$	- Conditional PDF
$\sim$	- Distributed as
$\mathcal{N}(\cdot, \cdot)$	- Normal Distribution whose mean and covariance matrix are given as first and second argument
$\mathcal{W}(\cdot, \cdot)$	- Wishart Distribution whose mean and scale matrix are given as first and second argument
$\propto$	- Proportional to

Segmentation, Binary Partition Trees

$\|\cdot\|$	- Cardinality of a set
$\Pi(T)$	- Set of all partitions of the set T
π	- Partition, i.e. element of $\Pi(T)$
$\mathcal{R}$	- Region of a partition
$\mathcal{M}_{\mathcal{R}}$	- Region model for region $\mathcal{R}$ of a partition

$\mathcal{O}_{\cdot}$	- Merging criterion (name in index) for BPTs

Miscellaneous

$\hat{\cdot}$	- Estimated or extrapolated value
$\equiv$	- Identically equal to/associated to/equal up to an isomorphism
$\triangleq$	- Equal by definition
$\approx$	- Approximately equal to
$f = \mathcal{O}(g)$	- f is bounded above by g (up to a constant factor)
$f = o(g)$	- f is asymptotically dominated by g
$f \sim g$	- f is asymptotically equal to g

Introduction

The human eye is sensitive to three ranges of wavelengths of the electromagnetic spectrum, approximately corresponding to red, green, and blue, thanks to the three types of cone cells, provided illuminance is sufficient. The ability we have to distinguish many more colors, and in the end, to be sensitive to what we called the visible part of the electromagnetic spectrum (between 380 nm and 780 nm) actually results from the processing of these three types of information from the visual cortex of the brain. Then this color information, combined with the shape information given by the image formed on the retina are further processed by the brain in order to analyze and interpret the observed scene. This is possible only because of the source of light: what we visually perceive is only the result of the interaction of the incoming light with matter. Only from the "data" made of the color information and the spatial information of the image on the retina as well as the processing from the brain are we able to interpret what we see.

Somewhat similarly, digital cameras generally use color filters in order to separate the contributions of red, green and blue in the incoming light and direct them to different Charged Coupled Device arrays. The image is then recombined by a computer and can be displayed or processed, in order to be *automatically* interpreted.

Along with the development of computers and electronic photodetectors, this analogy spawned two challenges: to what extent is it possible to automatize the processing a human being can naturally and easily perform, and what would we gain to sense what is invisible to the human eye, that is the information contained in the remainder of the electromagnetic spectrum? In more technical terms, what benefits can we obtain from augmenting the spectral resolution of digital sensors? The former challenge gave birth to computer vision, and digital image analysis and processing, while the latter gave birth to spectroscopy. Combining the two by acquiring images whose individual *picture elements*, or pixels, are *spectra* of tens to hundreds of values, rather than a single value for gray-level images, or three values for RGB images is what defines multispectral and hyperspectral imaging, also known as *imaging spectroscopy*.

The spectral information contained in such images is considerably augmented with respect to conventional gray-level or color imaging. The chosen wavelength range can vary a lot depending on the application, but typically incorporates the visible and near infrared parts of the electromagnetic spectrum. Once a certain spectral resolution has been reached, the spectrum resulting from the acquisition of a pixel, corresponding to the observation of a certain material is usually considered as characterizing that material. This allows the identification and characterization of the materials present in the imaged scene, or sample, in a much finer way than what is possible with conventional images, or the naked human eye. However, in practice, there is a compromise to find between spectral and spatial resolutions, meaning that a hyperspectral image (HSI) of a scene will have a lower spatial resolution than a color or even a multispectral image acquired over the same scene.

The applications of hyperspectral imaging are very diverse, and include food processing, chemometrics, astrophysics, and so on. In this thesis, we will focus on remote sensing applications. Remote sensing is concerned with earth or planetary observation, in our case using hyperspectral sensors on board satellites, airplanes, or more recently Unmanned Aerial Vehicles (UAVs).

Hyperspectral imaging in remote sensing has been used in: planetary exploration (as an example, hyperspectral sensors mounted on satellites are frequently used to observe Mars) environmental measurements (monitoring of natural landscapes, damage evaluation in the event of a natural disaster...), agricultural (crop monitoring) or defense (surveillance, detection) applications. An example of HSI in a remote sensing context is presented in Fig. 1. Each wavelength is represented by a slice of the data cube. This provides a second way to apprehend a HSI, other than a collection of spatially arranged spectra. Indeed, each of these slices can be thought as gray level image, containing the radiance information for this particular wavelength. The HSI is then a collection of radiance (the physical quantity measured by the sensor, corresponding to the quantity of radiation reaching the sensor) images for different ordered wavelengths, also referred to as spectral bands.

In many cases, the effect of the atmosphere (even more so for satellite images) is unwanted, since we want to characterize the interaction of light with the materials on the ground, that is, we would like to have access to the optical properties of the materials. Physics-based algorithms are able to model and remove the contribution of atmospheric effects to the received radiance, providing an atmospherically corrected image in surface reflectance units. Reflectance is another physical quantity corresponding to the ratio between the light received at the sensor and the incident light on the area in the Field Of View (FOV) of the sensor during the acquisition of a pixel. Reflectance for a given wavelength range is then bounded between 0 (the material absorbs or scatters all the incident light away from the sensor) and 1 (the material reflects all the incident light towards the sensor).

We can immediately see that for the examples provided in the image of Fig. 1, the reflectance spectra of the different materials in the FOV of the sensor when it acquired a pixel are significantly different, which allows to distinguish between materials, and somewhat explains the use of the term “spectral signature” of a material.

The information contained in a hyperspectral image (HSI) also comes with an important quantity of data to handle, which calls for efficient signal and image processing algorithms to process such images, beginning with an efficient visualization of the information in HSIs.

Other applications of hyperspectral remote sensing include (the list is not exhaustive):

- Target detection, that is determining whether a material with a known signature is present or not in the observed scene [124], or the related problems, anomaly and change detection.
- Superresolution, a data fusion problem aiming at combining the information contained in two (or more) co-registered images of the same scene containing complementary in-

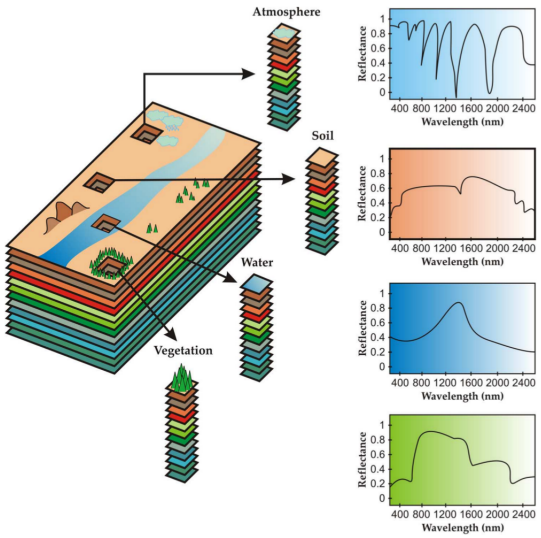


Figure 1: Representation of a HSI [20].

formation. For example, the images can have different resolutions [7]. Typically, we are interested in fusing a hyperspectral image (low spatial resolution and high spectral resolution) and a multispectral image (with converse resolution properties). The goal is to obtain an image with good resolutions for the two modalities. A related problem is pansharpening, where the goal is to fuse the information contained in a HSI and a panchromatic image, that is a gray level image incorporating information on all wavelengths in the visible light (minimal spectral resolution), but with a very high spatial resolution [106].

- HSI segmentation [24] is a problem whose goal is to partition a HSI into meaningful spatial (connected) regions in the sense of some predefined criterion (e.g. spectral homogeneity) in order to simplify its interpretation. A related problem is the supervised classification problem [62, 26], where a certain number of pixels have known labels, and the objective is to relate each pixel of the HSI to one of these predetermined classes.

- Spectral Unmixing (SU) [94]. Due to the limited spatial resolution of HSIs, many pixels cannot account for the spectral signature of only one material, because there were several distinct materials present in the FOV of the sensor during the acquisition. The acquired signature is then a mixture of the contributions of the different materials. In the example of Fig. 1, the pixels at the interface of the river and soil are a mixture of both reflectance signatures. Similarly, pixels whose field of view incorporate only a few isolated trees are a mixture of soil and vegetation. SU aims at identifying the signatures of the materials present in the image scene and to quantify their proportions inside each pixel of the image. SU will be the main focus of this thesis and this problem will be much more developed in in the next paragraph and in the remainder of this thesis.

Spectral unmixing can be seen as a blind source separation (BSS) problem whose goal is to be able to separate the contributions of the materials present in the FOV of the sensor

during each pixel's acquisition. The spectral signatures of the materials have to be extracted from the data (they are called *endmembers*) and their relative proportions in each pixel have to be estimated (these are called *fractional abundances*). The term "blind", coming from the signal processing community [43], means that we only make use of the observed data to solve the problem, without knowing either the endmembers or the mixing coefficients. We only use an equation describing how the observed pixel is related to the endmembers and abundances. This equation is called the *mixing model* and is of prime importance for spectral unmixing. Blind Source Separation (BSS) is an important topic in signal processing, and has been intensively studied in the signal processing community since its introduction in the 1980s [43]. Its applications are now very diverse. BSS was first formulated for a biological problem, but was quickly extended to many problems, such as the so-called "cocktail party" problem, when one tries to recover several audio sources from mixtures recorded by several microphones in the same room. Other applications include chemical analysis, geoscience, acoustics, and a wide range of biomedical problems: electroencephalography (EEG) signal analysis, Magnetic Resonance Imaging (MRI), electrocardiogram (ECG) analysis, and the list is far from exhaustive.

As we will see in the thesis, in SU, the mixture model is often considered linear as a first approximation [20], but it has particularities which make it a very specific BSS problem. We will see that usual BSS tools, such as the well-known Independent Component Analysis (ICA), cannot be straightforwardly applied to SU.

Objective and organization of the thesis

This thesis tries to address one of the usual limitations of the classical algorithms and models proposed for SU. One of these limitations is the questionable validity of the usual linear mixing model (LMM) in some practical applications, which has received a lot of interest in the last few years in the community [80]. The second limitation, which motivates this thesis is the so-called spectral variability (SV), or endmember variability problem. The idea behind this concept is that in most SU models and algorithms, a single material (or endmember) is implicitly assumed to be perfectly represented by a unique spectrum. This is in fact a strong assumption since the different materials always exhibit a certain intra-class variability, caused by different phenomena that we will detail in the next chapter. SV can considerably hamper the results of conventional SU techniques in practical scenarios. In other words, the term "spectral signature" is a bit misguiding, since a single spectrum cannot completely characterize a material. The community has long been aware of this issue, and some works exist to take it into account [143, 172], but in comparison to nonlinear mixtures, before this thesis started, it had received much less attention. However, the SV problem is currently in the spotlight and is becoming an important research topic in HSI image processing and SU. The main objective of this thesis is then to design models and algorithms specifically designed to tackle the SV problem. The manuscript is organized as follows:

Part I will review the state of the art methods for linear SU. It comprises two chapters. **Chapter 1** introduces more thoroughly the SU problem, and lays out a state of the art for

this topic. We focus on linear SU, which has been extensively studied in the past few decades, and for which many algorithms and methods have been developed. Then, in **Chapter 2** we make a review of most of the existing approaches to address the SV issue, in a linear SU framework, and we classify them depending on their lines of attack, and assumptions made to deal with the problem.

Part II and **Part III** gather the conributions of this thesis to the SU problem, accounting for endmember variability.

Part II is concerned with approaches dealing with sparsity to handle spectral variability. The methods we will introduce do not explicitly model spectral variability in the mixing model. In this sense, they can be considered "data driven". One of these approaches will perform SU in local, often small regions of the datasets (in which endmember variability effects are mitigated). In such cases, estimating the number of endmembers to use can become difficult, and overestimation is frequent. Then it makes sense to evaluate how this estimation is impacted by the small size of the regions. **Chapter 3** studies the problem of intrinsic dimensionality estimation for HSI in a local setting, by comparing and discussing several ID estimation algorithms of the literature. **Chapter 4** is divided into two parts, and deals with the use of sparsity to tackle spectral variability. The first part aims at finding a way to circumvent this overestimation problem in Local Spectral Unmixing (LSU) by eliminating irrelevant extracted endmembers using sparsity. We show how this improves LSU results on a simulated and a real dataset. The second part deals with the consideration of endmember "bundles", that is the representation of a material by a set of endmembers extracted from the data rather than a single spectrum. In this context, we show that introducing structured sparsity can also be beneficial to SU performance, both on synthetic and real datasets.

Part III introduces a new mixing model, called extended linear mixing model (ELMM), which is specifically designed to tackle the SV issue, starting from physical considerations, in particular to model the effect of changing illumination conditions on the spectrum of a given material. In **Chapter 5**, we derive the model from the physical radiative transfer model of Hapke, by resorting to simplifying assumptions in order to make the model tractable from a SU point of view. Further, we lay out two algorithms to estimate the parameters of this model, including several constraints and regularizations to provide better solutions, adapting optimization algorithms to our problem. We show results on synthetic and real datasets and compare our approaches to other algorithms of the literature. **Chapter 6** presents some extensions and applications of the ELMM to two other approaches for SU. The first application is a combination of the ideas of the ELMM to those of LSU, in order to be able to estimate SV in the LSU framework. The second application is the nonnegative Canonical Polyadic (CP) tensor decomposition of hyperspectral data, which can be shown to be connected to the ELMM.

Part I

A review on linear Spectral Unmixing methods and algorithms

Chapter 1

State of the art for the linear spectral unmixing problem

Contents

1.1 Introduction

The limited spatial resolution of hyperspectral sensors makes it very likely that the incoming light arriving to the sensor actually results from the interaction of photons with several distinct materials. The resulting reflectance (or radiance) is then a mixture of the contributions of the reflectances (or radiances) of the materials present in the Field Of View (FOV) of the sensor during the acquisition of a pixel. The objective of Spectral Unmixing (SU) is then to recover the signatures of the different materials present in the observed scene (called *endmembers*, for a reason which will become clear soon), and to quantify their proportions in each pixel of the scene (called *fractional abundances*). The problem is very simple to formulate, but solving it is not straightforward. We will first review the conventional linear spectral unmixing problem, and present the main algorithms developed to solve it. These methods have proven useful for SU, but they assume that the materials in the imaged scene are perfectly represented by a

single spectrum, which is a very strong implicit hypothesis. Then we briefly will discuss the main limitations of linear SU, including spectral variability, which will be the main focus of this manuscript.

This chapter's goal is to present the linear SU problem and to review the main approaches to tackle it. There are a large number of methods and algorithms for this problem in the literature, and we try to select some of the best representatives and we categorize them, following the outline of [20]. Let us denote each of the N hyperspectral pixels of the image by a vector $\mathbf{x}_k \in \mathbb{R}^L$, $k = 1, ..., N$, where L is the number of spectral bands. If we assume that there are P endmembers, with their signatures $\mathbf{s}_p \in \mathbb{R}^L$ to consider, we can store all of these signatures into the columns of a matrix $\mathbf{S} \in \mathbb{R}^{L \times P}$. There are then P abundance coefficients for each pixel, which we gather in vectors $\mathbf{a}_k$, $k = 1, ..., N$. Then for pixel k, we can relate the observation to the abundances and endmember matrix by:

$$\mathbf{x}_k = f(\mathbf{S}, \mathbf{a}_k) + \mathbf{e}_k, \tag{1.1}$$

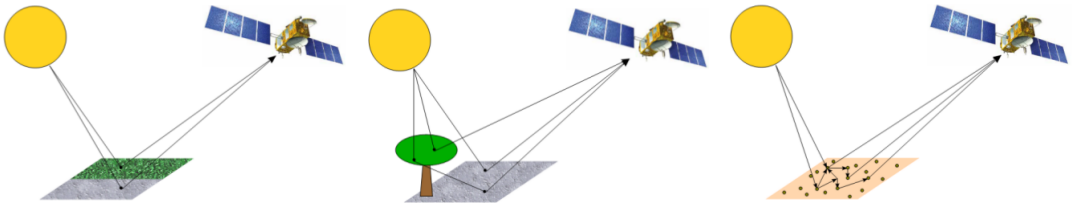
where f is a function modeling the mixing process, whose analytical form is the only information we have on the problem, other than the data. The vector $\mathbf{e}_k$ is an additive noise, often assumed to be Gaussian. If we store all pixels in a matrix $\mathbf{X} \in \mathbb{R}^{L \times N}$, all noise vectors in matrix $\mathbf{E} \in \mathbb{R}^{L \times N}$, and all the abundance vectors in a matrix $\mathbf{A} \in \mathbb{R}^{P \times N}$, and assuming the mixture model f is the same in the whole image, we can rewrite Eq. (1.1) as:

$$\mathbf{X} = f(\mathbf{S}, \mathbf{A}) + \mathbf{E}. \tag{1.2}$$

The scope of this thesis is on the classical case of the Linear Mixture Model. Therefore we will not address the case of nonlinear mixtures in detail. Reviews on the topic can be found in [80, 49].

1.2 Linear Spectral Unmixing

Considering that the endmembers are sources, and that the abundances are mixing coefficients, then estimating $\mathbf{S}$ and $\mathbf{A}$ is a Blind Source Separation (BSS) problem. For SU, a usual assumption is to consider a Linear Mixing Model (LMM), that is to consider that an observed pixel is a linear combination of the endmembers' spectra, weighted by the abundances. The LMM is a reasonable assumption when within the FOV of a pixel, each ray of light only interacts with a single material, before going to the sensor. This is typically the case in the so called "checkerboard" configuration, as shown Fig. 1.1 (a). Nonlinear mixing processes are typically linked to the limited validity of such a hypothesis in certain scenarios, for instance in tree canopies when the incoming light bounces off several times before reaching the sensor (Fig. 1.1 (b)), or when the mixture occurs at a microscopic level, e.g. in sand for instance (Fig. 1.1 (c)). The specificities of SU of HSIs with respect to other BSS problems are first due to the nonnegativity of the data: measurements are radiance or reflectance spectra, and should therefore be nonnegative. This means that the endmembers' spectra are nonnegative



(a) (b) (c)

Figure 1.1: Several mixing configurations. In a "checkerboard" scenario, the LMM is valid (a). Each material occupies a fraction of the (flat) surface in the FOV of the sensor. In (b) is displayed a case of multiple reflections of light before reaching the sensor. (c) shows a case of intimate mixing. The images were borrowed from [49].

as well. Besides, with the assumptions of Fig. 1.1, it is natural that the abundances should be interpreted as proportions. That is why the abundance coefficients are usually required to be positive and to sum to one in each pixel. The resulting constraints are usually called the abundance nonnegativity constraint (ANC), and the abundance sum-to-one constraint (ASC). The validity of the ASC can be questioned when nonlinearities or spectral variability (as we will see later in the manuscript) are not negligible, but it makes perfect sense in the classical LMM framework, provided there is no significant endmember missing. With the LMM, we can simply rewrite Eq. (1.1) as:

$$\mathbf{x}_k = \sum_{p=1}^{P} a_{pk}\mathbf{s}_p + \mathbf{e}_k, \tag{1.3}$$

and in the global case, Eq. (1.2) becomes:

$$\mathbf{X} = \mathbf{SA} + \mathbf{E}. \tag{1.4}$$

This problem is extremely ill-posed because the solution is not unique. Indeed, for instance, any invertible matrix $\mathbf{M} \in \mathbb{R}^{P\times P}$ will satisfy $\mathbf{SA} = (\mathbf{SM})(\mathbf{M}^{-1}\mathbf{A})$. In addition, constraining both the endmembers and abundances to be nonnegative, although physically sound, is not sufficient to alleviate this issue. However, this ill-posedness is typical in BSS problems, and additional assumptions have to be made to find suitable solutions. The best known of the assumptions classically used in BSS is probably the statistical independence of the sources to be separated, which led to the broad class of Independent Component Analysis (ICA) algorithms [43]. ICA techniques have proven to be extremely powerful and useful in many different situations. Unfortunately, in SU, its core assumption of statistical independence of the sources is not valid [121, 107].

There are two ways to consider Eq. (1.4) in an ICA framework. The most natural way to do so is to see the endmembers as the sources (in the rest of this thesis, we will often refer to the endmembers' spectra as "sources"), and the abundances as the mixing coefficients (with additional constraints). However, even though the endmembers are not usually modeled as random variables, empirical measures of statistical independence (such as Pearson correlation

coefficients between any two endmembers, at the second order) suggest that spectral signatures of the endmembers are far from independent [13]. They often share the same absorption bands, and some other spectral features. Another less intuitive way is to consider that the abundances are the sources, and that they are mixed by the coefficients of the endmembers' spectra. Nevertheless, one can immediately see that the ASC introduces a statistical dependence between the abundance coefficients [121]. The application of ICA techniques is then precluded in HSI data analysis for SU. However, it could still be useful separate artifacts [75] (e.g. the so-called "striping" and "smile" effects [139]) from the signal of interest.

This means that specific techniques have to be developed in order to reliably estimate the parameters of SU, even in a x linear case.

1.3 Convex geometry of the SU problem

Since ICA techniques usually fail in SU, techniques to recover endmembers and abundances have to rely on other approaches. In a LMM framework, and given the ANC and ASC, the geometric properties of the problem are going to be the cornerstones of most SU techniques. Both constraints mean that every abundance vector belongs to the unit (or probability) simplex of dimension $P-1$, denoted as Δ_P [1], and defined by $\Delta_P = \{\mathbf{a} \in \mathbb{R}^P, \forall p \in [\![1, ..., P]\!],\ a_p \geq 0,\ \text{and}\ \sum_{p=1}^{P} a_p = 1\}$. A simplex can be thought as a generalization of a triangle to higher dimensions: a n-simplex is a subset of an n dimensional affine subspace which is the simplest $n-1$ dimensional object in this subspace. For two points, the 1-simplex whose vertices are these two points is the line joining them. Three non collinear points can define the vertices of a 2-simplex, that is a triangle. A 3-simplex is a tetrahedron, and so on. In addition, since the mixture of the endmembers is linear, an observed pixel also lies in a $(P-1)$-simplex, whose vertices are precisely the P endmembers (hence their name). This means that although the data belongs to a high L-dimensional embedding space, they actually live in a much lower dimensional subspace, whose structure is relatively simple. This is because the data pixels are convex combinations of the endmembers. An example is given in Fig. 1.2. On this figure, the endmembers are represented as the red dots. A data point is shown in blue. The abundances of this pixel are the barycentric coordinates of the blue point in the simplex. This simplex is located in the affine subspace spanned by the endmembers. If the ANC is dropped and the ASC is kept, then the data can lie anywhere in this whole subspace, and not only in the simplex.

Conversely, if the ASC is dropped, the data still possess a strong geometric structure, since the ANC makes the data live in a convex cone spanned by the endmembers. This fact will be important when we will look into the effects of spectral variability later in the manuscript. The intersection of the cone and the affine subspace then also defines the data simplex when both constraints are enforced.

[1] We put P as a subscript to insist on the number of vertices, rather than on the dimensionality of the simplex.

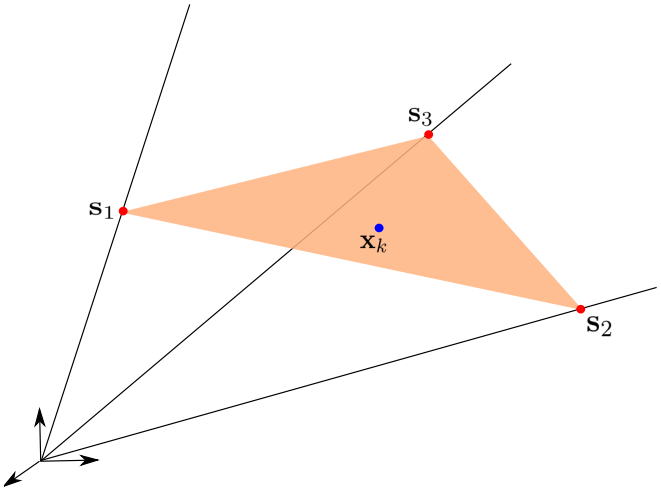


Figure 1.2: Geometric interpretation of the LMM in the case of three endmembers (red dots). The axes represent a basis of the linear subspace spanned by the endmembers.

This geometric interpretation of the LMM under the constraints suggests a two-step line of attack which is the most common way to perform SU: trying to extract the endmembers, knowing they are vertices of a simplex, and after that estimating the corresponding abundances. A third (implicit) step is to estimate the number of endmembers to extract, which is a difficult task and a research topic on its own, with much broader applications than SU.

1.3.1 Estimating the number of endmembers

Before extracting the sources, the number of endmembers to consider has to be estimated somehow. Recall that the definition of the endmembers is subjective and depends on the scale at which the endmember extraction has to be performed. This makes the estimation all the more difficult. However, an upper bound to this number can be obtained by estimating the Intrinsic Dimensionality (ID) of the dataset. For a HSI: $\mathbf{X} = \mathbf{Y} + \mathbf{E}$, decomposed as a signal part $\mathbf{Y}$ and a noise part $\mathbf{E}$, the ID of this dataset is defined (in the most common acceptation) as the dimension of the vector subspace spanned by the signals $\mathbf{y}_1, ..., \mathbf{y}_N$. A wide variety of algorithms exist in the literature (specific to hyperspectral remote sensing or more generally in the signal processing community) to perform this estimation task [132]. Chapter 3 will review some ID estimation algorithms of the literature and discuss their performance in particular settings for SU.

1.3.2 Endmember extraction with pixel purity

Once the number of endmembers to use has been determined, algorithms specifically designed for this task have to be designed. Since obtaining endmembers from geometric considerations is quite specific to HSI analysis, this has been an important topic in the literature, and many endmember extraction algorithms (EEA) have been designed. We review some of them below.

In most cases, another key assumption has to be made in order for algorithms to be efficient: the so-called *pure pixel* assumption [127]. As the name suggests, this consists in assuming that for each of the materials considered, at least one pixel in the image is composed of only this material (meaning that its abundance is 1 in this pixel). If this assumption holds, then (without considering noise), there exist data points corresponding to each vertex of the simplex. Depending on the endmembers considered, on the spatial resolution of the sensor, and on the characteristics of the imaged scene, the assumption does not always hold, and some algorithms take this into account [128]. A review of most of the popular EEAs assuming pixel purity can be found in [127]. In all cases, the idea is to find a way to extract the extreme points of the dataset, which in most cases are assumed to be the vertices of a simplex. Here, we briefly review a few popular algorithms of the literature assuming pixel purity holds for all materials in the image.

- Pixel Purity Index (PPI) [21]. This algorithm starts by whitening the data (so that the noise has zero mean and a unit covariance matrix) through a Maximum Noise Fraction (MNF) [69] transform. This is a dimension reduction step whose goal is to define image components sorted in decreasing order in terms of Signal to Noise Ratio (SNR), (unlike Principal Component Analysis (PCA), which sorts its image components in terms of explained variance of the data). A smaller number of components are retained in order to reduce the impact of the noise and the computational burden. Then, a large number of random vectors, called "skewers" are generated, and for each of those all data points are projected on the directions defined by the skewers, after what the data points maximizing these projections are kept. If a point is situated on the extreme parts of the dataset, they are likely to lead to important projection values for a large number of skewers. The number of times each pixel leads to large projection values of a skewer is counted. This defines a PPI score for each pixel. Finally, the pixels whose score is larger than a user defined threshold are candidate endmembers. A further selection can be made by an interactive visualization tool. Note that a faster and fully automated implementation of PPI was proposed in [39].

- N-FINDR [165] (for "N-finder", where N is the number of endmembers to extract). This algorithm aims at finding the pixels of the image forming the simplex with maximum volume among the possible simplices in the data. The idea is to grow a simplex inside the data until the one with maximum volume is found. In more details: an initial noise whitening step is also carried out using MNF to reduce the dimensionality of the data. Then P randomly chosen pixels make an initial guess of the endmember set. The volume of the simplex spanned by these initial endmembers is then computed. Then given a pixel, the volumes of P new simplices are computed by replacing the p^{th} endmember by the current pixel. Then the endmember which is absent in the simplex with largest volume is replaced by the current pixel. This operation is repeated for all pixels, after what the vectors spanning the largest volume simplex are retained as endmembers.

- Vertex Component Analysis (VCA) [122]. This algorithm also performs some dimension reduction using a PCA as its first step, in order to accommodate the noise, and projects the data onto a P-dimensional subspace. Then, a projective projection step (a projective

geometry concept, also known as perspective projection, or dark point fixed transform [47]) is carried out. It is a rescaling of the data in which every pixel $\mathbf{x}_k$ is transformed as: $\tilde{\mathbf{x}}_k = \frac{\mathbf{x}_k}{\mathbf{x}_k^\top \mathbf{u}}$, with $\mathbf{u} \in \mathbb{R}^P$ a vector chosen so that $\mathbf{x}_k^\top \mathbf{u} > 0,\ \forall k = [\![1, ..., N]\!]$. Here $\mathbf{u}$ is chosen as the mean of the data. The projected points lie on a hyperplane (recall that at this point, we are working in a P-dimensional subspace) defined by $\mathbf{v}^\top \mathbf{u} = 1,\ \forall \mathbf{v} \in \mathbb{R}^P$. This projection transforms a cone in $\mathbb{R}^P$ into a $(P-1)$-simplex, which makes the algorithm relatively insensitive to scaling variations of the data (for a high enough estimated SNR). This interesting property will be of interest later in the manuscript. The projective projection is only carried out when the noise power is not too important because it can amplify the noise. An iterative process is used in order to identify the endmembers. A random vector is generated, and the first endmember is the data point which maximizes the projection onto the direction of this vector. Then, a new random vector is generated, with the constraint that it has to be orthogonal to the subspace spanned by the already determined endmembers. At each step, the data point maximizing the projection on this random vector is identified as an endmember. The process stops when P endmembers have been identified.

These algorithms share the same core idea, which is to look for extreme points in the data distribution. However, their main drawbacks are their stochasticity, and their sensitivity to outliers, or to noise for low SNRs. For example, in the VCA, an outlier can easily be selected as a spurious endmember, and this will hamper the extraction of the following endmembers since the orthogonal projection step will be made in a suboptimal direction.

1.3.3 Abundance estimation

Once the endmembers have been identified, the last phase of the usual unmixing chain is abundance estimation. This is generally done by solving the following optimization problem (which is separable w.r.t. the pixels of the image):

$$\begin{aligned} \hat{\mathbf{A}} = \arg\min_{\mathbf{A}} \frac{1}{2}||\mathbf{X} - \mathbf{S}\mathbf{A}||_F^2 \quad &\Leftrightarrow \quad \hat{\mathbf{A}} = \arg\min_{\mathbf{A}} \frac{1}{2}\sum_{k=1}^{N} ||\mathbf{x}_k - \mathbf{S}\mathbf{a}_k||_2^2 \\ \text{s.t. } \mathbf{A} \in \Delta_P \quad &\text{s.t. } \mathbf{a}_k \in \Delta_P,\ \forall k = 1, ..., N, \end{aligned} \tag{1.5}$$

where $\mathbf{A} \in \Delta_P$ must be understood columnwise (each abundance vector for each pixel is constrained to belong to the probability simplex). $||.||_F$ denotes the Frobenius norm. This is the simplest way one can recover abundances subject to the ANC and ASC, once an endmember set is known, simply by looking for the abundances which reconstruct the data best in a least square sense, given $\mathbf{S}$. This way to obtain an estimation of the abundances is usually called Fully Constrained Least Squares Unmixing (FCLSU) in the literature. When the ASC is dropped, solving the optimization problem (1.5) is called (partially) Constrained Least Squares Unmixing (CLSU). The most widely used algorithm of the literature to solve these two problems is described in [76]. Without the ASC, the problem in pixel k simply

reduces to the nonnegative least squares problem:

$$\begin{aligned} \hat{\mathbf{a}}_k = \arg\min_{\mathbf{a}_k} \frac{1}{2}||\mathbf{x}_k - \mathbf{S}\mathbf{a}_k||_2^2 \\ \text{s.t. } \mathbf{a}_k \geq \mathbf{0}, \end{aligned} \tag{1.6}$$

which is solved in [76] and in the algorithm we will use throughout the thesis by an active set algorithm. The idea behind active set algorithms in nonnegative least squares (or more generally in quadratic programming) is that the optimal solution $\mathbf{a}^*$ (we drop for now the pixel index for brevity) will have a certain number of strictly positive entries, and the others will be zero (the nonnegativity constraint will be *active* for these entries). If the set of indices corresponding to zero entries is known (the *active set*), then the remaining entries (those of the *passive* set, where the constraints are *inactive*) of $\mathbf{a}^*$ are simply the solution of the unconstrained reduced problem:

$$\arg\min_{\mathbf{a}_{passive}} \frac{1}{2}||\mathbf{x} - \mathbf{S}_{passive}\mathbf{a}_{passive}||_2^2, \tag{1.7}$$

whose solution $(\mathbf{S}_{passive}\mathbf{S}_{passive}^\top)^{-1}\mathbf{S}_{passive}\mathbf{x}$ involves the Moore-Penrose pseudoinverse of the matrix $\mathbf{S}_{passive}$ (provided it has full rank, which is a reasonable assumption here). Then, the solution of problem (1.6) is $\mathbf{a}^* = [\mathbf{a}_{passive}, \mathbf{0}_{active}]$. Then the problem boils down to identifying the active and passive sets. To do that, we write the Lagrangian for problem (1.6):

$$\mathcal{L}(\mathbf{a}) = \frac{1}{2}||\mathbf{x} - \mathbf{S}\mathbf{a}||_2^2 + \boldsymbol{\mu}^\top \mathbf{a}, \tag{1.8}$$

where $\boldsymbol{\mu}$ is a vector of Lagrange multipliers for the inequality constraint $\mathbf{a} \geq \mathbf{0}$. The Karush-Kuhn-Tucker (KKT) conditions for this problem write:

$$\begin{aligned} &\mathbf{S}^\top(\mathbf{S}\mathbf{a}^* - \mathbf{x}) + \boldsymbol{\mu} = \mathbf{0} \\ &\boldsymbol{\mu}^\top \mathbf{a}^* = 0 \\ &\mathbf{a}^* \geq \mathbf{0} \\ &\boldsymbol{\mu} \geq 0. \end{aligned} \tag{1.9}$$

In particular, these conditions imply that the vector $\mathbf{w} = \mathbf{S}^\top(\mathbf{S}\mathbf{a}^* - \mathbf{x})$ only has negative entries. Suppose that we have initialized the indices of the active and passive sets, and solved the reduced problem (1.7) on this passive set. Let us further define $\hat{\mathbf{w}} = \mathbf{S}^\top(\mathbf{S}\mathbf{a}^\dagger - \mathbf{x})$, where $\mathbf{a}^\dagger = [\mathbf{a}^*_{passive}, \mathbf{0}_{active}]$ for the current active and passive sets. Then, if $\hat{\mathbf{w}} \leq \mathbf{0}$ on all the passive and active indices, then we have $\mathbf{a}^* = \mathbf{a}^\dagger$. Otherwise, it can be shown that if for some index p, $\hat{w}_p > 0$, then incorporating it to the passive set, and solving the updated problem (1.7) will decrease the objective function. This process of adding one of the indices (in practice the one corresponding to the largest entry in $\hat{\mathbf{w}}$) for which the condition is not satisfied to the passive set, is repeated until the true active set has been found (when all the entries of $\hat{\mathbf{w}}$ are negative). This first algorithm to take advantage of this idea was presented in [101], and is

proven to terminate and converge to the true solution of problem (1.6). There are many more ways to solve this problem, such as interior-point methods [22], or more modern proximal methods [42]. In order to solve the FCLSU problem, that is to incorporate the ASC, [76] augments the endmember matrix $\mathbf{S}$ with an additional vector of ones, and the pixel vector with an additional 1:

$$\tilde{\mathbf{S}} = \begin{bmatrix} \delta\mathbf{S} \\ \mathbb{1}_L^\top \end{bmatrix}, \text{ and } \tilde{\mathbf{x}} = \begin{bmatrix} \delta\mathbf{x} \\ 1 \end{bmatrix}, \tag{1.10}$$

where δ is a coefficient weighting the ASC, with respect to the data fit, and $\mathbb{1}_L$ is a vector of ones of length L. In order that the ASC should be almost perfectly enforced, δ should be very small. This algorithm solves the CLSU problem (1.6), simply replacing $\mathbf{S}$ by $\tilde{\mathbf{S}}$ and $\mathbf{x}$ by $\tilde{\mathbf{x}}$.

Note that the (F)CLSU problem could also be easily (and more efficiently, according to [130]) dealt with in a proximal framework, for instance by using splitting algorithms such as the Alternated Direction Method of Multipliers (ADMM), and adding Lagrange multipliers for the equality constraint in case the ASC is required, or by using a projected gradient scheme, using efficient algorithms to project on the unit simplex [45] if the ASC is considered. For more details about proximal methods, which will be at the core of the optimization schemes developed in this thesis, see [42], or Appendix A.

Other algorithms faster than the active set method of [76] (at least for a relatively low number of endmembers, usually less than 10 [130]) have been more recently developed in the community. We can cite for instance the work of [79], which does not rely on complex optimization algorithms, but rather on affine geometry, and linear algebra. It is based on the fact that data points outside the convex hull of the endmembers will necessarily have at least one zero abundance coefficient. The algorithm uses a method to identify a zero coefficient, and then finds the other zero coefficients one after the other by recursively projecting the data point on the affine subspace spanned by the endmember set, minus one which has been identified as having zero abundance. This operation is performed until the projected point is situated inside or on the facet of a lower dimensional simplex, spanned by the active endmembers for this pixel. Then the abundances can simply be computed using the unconstrained least squares solution (in a very low dimensional subspace, so this operation requires a much reduced computational load). Since the data point is inside the simplex, the ANC and ASC will naturally be satisfied. Another algorithm based similar concepts is slightly faster due to a more efficient computation of the barycentric coordinates of points inside the simplex [130].

1.3.4 Endmember extraction without pure pixels

Other algorithms have been designed in order to circumvent the pure pixel hypothesis [128]. Most of them are based on finding the simplex of minimum volume enclosing the data. This method is able to find the true endmembers of the dataset if there are enough data pixels on the facets of the simplex, in order that the scatterplot can be extrapolated into the true simplex, as can be seen on Fig. 1.3. The constraint of having all pixels belonging to the

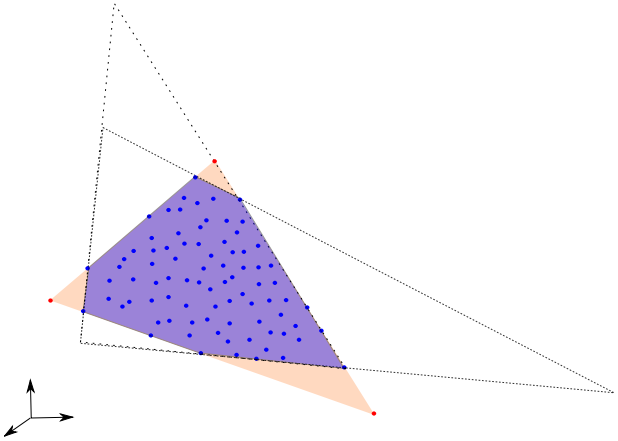


Figure 1.3: Finding the minimum volume simplex enclosing the data. The endmembers are in red, and the data points in blue. The convex hull of the data, whose extreme points would be identified as endmembers by pure pixel-based EEAs, is represented in another shade of blue. The true simplex is in orange, and two other simplices enclosing the data, but with a larger volume are represented in dashed lines.

simplex can be softened to avoid too much sensitivity w.r.t. outliers, which can highly affect the volume of the simplex. The minimum volume constraint is enforced through a hard, often nonconvex optimization problem, resulting in an increased complexity of the algorithms, w.r.t. pure pixel based EEAs. Also, these approaches are inherently able to jointly estimate the endmembers and the abundances, often in a Nonnegative Matrix Factorization (NMF) framework, which we will encounter again in section 1.4.2. We can cite as examples the Minimum Volume Simplex Analysis (MVSA) [103], and its extension, the so-called Simplex Identification via Split Augmented Lagrangian (SISAL) [17]. In both MVSA and SISAL, the first step is a dimension reduction in order to identify the P-dimensional signal subspace (using the algorithm of [19], which we will detail in Chapter 3). Since the volume of the simplex spanned by the columns of $\mathbf{S}$ is proportional to its determinant, MVSA solves the following optimization problem:

$$\begin{aligned} &\arg\min_{\mathbf{Q}} \; -\ln(|\det(\mathbf{Q})|) \\ &\text{s.t. } \mathbf{Q}\tilde{\mathbf{X}} \geq \mathbf{0},\; \mathbb{1}_P^\top \mathbf{Q}\tilde{\mathbf{X}} = \mathbb{1}_N^\top, \end{aligned} \tag{1.11}$$

where $\tilde{\cdot}$ is the projection of the matrix under the tilde onto the signal subspace and $\mathbf{Q} = \tilde{\mathbf{S}}^{-1}$. This matrix is used because the problem is better conditioned under this formulation than simply using $\tilde{\mathbf{S}}$. Such a hard problem is easier to solve in a lower dimensional subspace, where the noise is in addition reduced. The constraints on the second line of the problem are simply the ANC and ASC, since $\mathbf{Q}\tilde{\mathbf{X}}$ is the abundance matrix. This problem is nonconvex (in general) and very hard to solve, and in practice the MVSA only looks for a good enough local minimum. SISAL solves a similar problem, except that it replaces the hard ANC by a soft version in order to accommodate outliers, which can considerably hamper the results of minimum volume-based algorithms.

The Minimum Volume Constrained NMF (MVC-NMF) [118] or the Iterative Constrained Endmembers (ICE) algorithm [15] can also be mentioned. It replaces the nonconvex penalization of the volume by a simpler convex surrogate: the sum of the squared distances between pairs of endmembers in the feature space. This is much easier to handle than the volume of the simplex, which involves the absolute value of the determinant function. ICE considers the following the globally nonconvex NMF problem (under the ASC and ANC):

$$\underset{\mathbf{S}\geq\mathbf{0},\mathbf{A}\in\Delta_P}{\arg\min}\ \frac{1}{2}||\mathbf{X}-\mathbf{S}\mathbf{A}||_F^2+\lambda\sum_{i=1}^{P}\sum_{j=i+1}^{P}||\mathbf{s}_i-\mathbf{s}_j||_2^2, \tag{1.12}$$

where λ is a regularization parameter, weighting the importance of the regularization term w.r.t. the data fit. Note that NMF problems require the initialization of the variables involved. In practice the initialization of the endmembers is often performed using a pure pixel-based EEA, and the abundances can be initialized by FCLSU, or randomly initialized, although this may cause the algorithm to be stuck in a poor local minimum of the objective function.

1.4 Statistical approaches

The algorithms discussed here make assumptions on the probability density function (PDF) of the variables of the unmixing problem, either explicitly for Bayesian approaches, or implicitly for NMF-based approaches. In particular, these approaches are supposed to be more robust to heavy mixtures, outliers and noise than the usual geometric approaches because the pure pixel assumption can be relaxed to some extent.

1.4.1 Bayesian Approaches

This class of methods relies on the Bayesian estimation framework. Provided the endmember matrix $\mathbf{S}$ and the abundance matrix $\mathbf{A}$ are statistically independent, we can write the following relationship between their PDFs, using Bayes' rule:

$$p_{\mathbf{S},\mathbf{A}|\mathbf{X}}(\mathbf{S},\mathbf{A}|\mathbf{X})=\frac{p_{\mathbf{X}|\mathbf{S},\mathbf{A}}(\mathbf{X}|\mathbf{S},\mathbf{A})p_{\mathbf{S}}(\mathbf{S})p_{\mathbf{A}}(\mathbf{A})}{p_{\mathbf{X}}(\mathbf{X})}, \tag{1.13}$$

where $p_{\mathbf{X}|\mathbf{S},\mathbf{A}}(\mathbf{X}|\mathbf{S},\mathbf{A})$ is the likelihood function, which depends on the observation model, $p_{\mathbf{S}}(\mathbf{S})$ and $p_{\mathbf{A}}(\mathbf{A})$ are the prior densities on the sources and abundances, respectively. These priors are used to incorporate prior knowledge about the parameters. $p_{\mathbf{S},\mathbf{A}|\mathbf{X}}(\mathbf{S},\mathbf{A}|\mathbf{X})$ is the posterior density. The idea of Bayesian estimation is to estimate this posterior density, and from them to estimate the parameters of the model from this PDF. Classically used estimators are the Minimum Mean Squared Error estimator (MMSE), or the Maximum A Posteriori (MAP) estimator. No specific prior information about the sources and abundances results in having uniform priors, and in that case the posterior density is proportional to the likelihood, and the MAP estimator reduces to the Maximum Likelihood Estimator (MLE). However, the

resulting estimation problem is very ill-posed and regularizations are necessary for accurate estimation, which, as we will see, is equivalent to incorporating priors on the abundances and endmembers.

In most cases, the posterior density does not have a simple analytical expression, and the MAP estimators cannot be obtained easily from an optimization problem, nor can the MMSE estimator be computed in a straightforward way. To circumvent this, Markov Chain Monte Carlo (MCMC) algorithms are used to sample the posterior densities, giving access to an approximate PDF, which in turn allows to access the MAP or MMSE estimates.

For instance, the work of [50] lays out a hierarchical Bayesian model for the data, in which the noise is assumed to be white, Gaussian with a diagonal covariance matrix, with the same variances on each band. The endmembers are assumed to be Gaussian as well, and their means are determined by pure pixel EEAs. The variances of these endmember distributions are linked to the confidence in this assumption. The abundances are assumed to be uniform over the unit simplex. The model is said to be hierarchical because the different variances involved have to be estimated as well, so they have to be assigned (often non-informative) PDFs as well.

Another example of an algorithm designed for SU in a purely statistical context is the so called "Dependent Component Analysis" (DECA) [120]. In this algorithm, the abundances are modeled as a mixture of Dirichlet densities, which automatically enforces the ASC and ANC. The prior on the endmembers depends on a parameter λ, which influences the penalization of a large volume of the estimated simplex (as in the minimum volume based EEAs). An algorithm is then designed to obtain the MLE for this problem. However, out of simplicity, the model does not incorporate noise. Having for instance a Gaussian model for the noise would allow to access the Bayesian estimators of MAP and MMSE to be computed using MCMC methods.

Note that the last two approaches are not constrained by the pure pixel assumption, nor by the requirement to have enough pixels on the facets (as for the minimum volume based approaches). In this sense, these approaches are more suited for highly mixed scenarios.

1.4.2 Nonnegative Matrix Factorization Approaches

Let us assume for a moment that the noise follows an i.i.d. Gaussian distribution with zero mean and a diagonal covariance matrix with identical variances for each spectral band:

$$\mathbf{e}_k \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_L). \tag{1.14}$$

The observed image $\mathbf{X}$ can be decomposed as a sum of signal and noise $\mathbf{X} = \mathbf{Y} + \mathbf{E}$, where $\mathbf{Y} = \mathbf{SA}$ is the noiseless signal. Using an hypothesis on the independence of the observed pixels we can write the likelihood as:

$$p_{\mathbf{X}|\mathbf{S},\mathbf{A}}(\mathbf{X}|\mathbf{S},\mathbf{A}) \sim \prod_{k=1}^{N} \frac{1}{(2\pi)^{\frac{L}{2}}\sigma^L} \exp\left(-\frac{||\mathbf{x}_k - \mathbf{y}_k||_2^2}{2\sigma^2}\right). \tag{1.15}$$

If we assume uniform priors for the endmembers and abundances (over the convex sets incorporating the constraints, i.e. the positive orthant for the endmembers and the unit simplex for the abundances), then the posterior density is proportional to the likelihood (1.15), and the MAP estimator reduces to the MLE. The minimizer of the negative log-likelihood is the same as the maximizer of the likelihood, so in the end, after some straightforward computations, the MAP estimator is:

$$\underset{\mathbf{S}\geq\mathbf{0},\mathbf{A}\in\Delta_P}{\arg\min}\ \frac{1}{2}\sum_{k=1}^{N}||\mathbf{x}_k-\mathbf{S}\mathbf{a}_k||_2^2 \ \Leftrightarrow\ \underset{\mathbf{S}\geq 0,\mathbf{A}\in\Delta_P}{\arg\min}\ \frac{1}{2}||\mathbf{X}-\mathbf{S}\mathbf{A}||_F^2. \tag{1.16}$$

The right handside of Eq. (1.16) precisely defines the Nonnegative Matrix Factorization (NMF) problem, that is to decompose a data matrix into factor matrices (here the sources and abundances), with the common dimension of the two factors known beforehand (here, the number of endmembers P to consider). However, as such the problem is extremely ill posed since there are infinitely many ways to find factor matrices solving this problem. In addition, the problem is biconvex (convex w.r.t. each of the factor matrices), but not jointly convex, with possibly many local minima.

It is well known, however, that incorporating regularization terms to the problem helps making the problem better posed and result in better solutions [163]. The various possible regularizations can be interpreted in a probabilistic way in the Bayesian framework, using the MAP estimator. Indeed, in a more general case than the MLE estimation of Eq. (1.16), the MAP estimator can be written as:

$$\underset{\mathbf{S},\mathbf{A}}{\arg\min}\ -\ln(p_{\mathbf{X}|\mathbf{S},\mathbf{A}}(\mathbf{X}|\mathbf{S},\mathbf{A}))-\ln(p_{\mathbf{A}}(\mathbf{A}))-\ln(p_{\mathbf{S}}(\mathbf{S})). \tag{1.17}$$

This means that any regularized NMF problem: $\underset{\mathbf{S},\mathbf{A}}{\arg\min}\ \frac{1}{2}||\mathbf{X}-\mathbf{S}\mathbf{A}||_F^2+\mathcal{R}_{\mathbf{S}}(\mathbf{S})+\mathcal{R}_{\mathbf{A}}(\mathbf{A})$, where $\mathcal{R}_{\cdot}(\cdot)$ is a suitable regularization term on the variable in index, can be seen as a MAP estimation for a white i.i.d. Gaussian noise, with

$$p_{\mathbf{S}}(\mathbf{S})=\frac{\exp(-\mathcal{R}_{\mathbf{S}}(\mathbf{S}))}{\int_{\mathbb{R}_+^L}\exp(-\mathcal{R}_{\mathbf{S}}(\mathbf{U}))\mathrm{d}\mathbf{U}}\quad\text{and}\quad p_{\mathbf{A}}(\mathbf{A})=\frac{\exp(-\mathcal{R}_{\mathbf{A}}(\mathbf{A}))}{\int_{\Delta_P}\exp(-\mathcal{R}_{\mathbf{A}}(\mathbf{U}))\mathrm{d}\mathbf{U}}. \tag{1.18}$$

With this in mind, the minimum volume based EEA algorithms alluded to in section 1.3.2 are not only based on the geometry of the SU problem, but can be interpreted in a statistical framework as well. For example, in the case of the ICE algorithm [15], under the same i.i.d. white Gaussian noise assumption, choosing a uniform prior on the abundances, and $p_{\mathbf{S}}(\mathbf{S})\propto\exp\left(-\lambda\sum_{i=1}^{P}\sum_{j=i+1}^{P}||\mathbf{s}_i-\mathbf{s}_j||_2^2\right)$ yields Eq. (1.12) as a MAP estimator. Other examples of NMF in SU which are not solely based on the minimum volume constraint are numerous, and can be based on various hypotheses, including spatial (piecewise) smoothness of the abundances [176] and sparsity [170].

Note that in all cases, statistical methods for SU require the use of initial values for the parameters to estimate , either to define the prior densities, or for NMF problems to initialize

the nonconvex optimization problem in a smart way. In practice the initialization of the endmembers is generally performed using a pure pixel-based EEA.

1.5 Sparse Unmixing

This section is concerned with a different approach for SU, which takes advantage of all the recent developments in signal processing concerning the use of the sparsity hypothesis. A sparse matrix is a matrix for which many entries are zero. For a linear system of equations, a sparse matrix means that the equations are not very coupled, as would be the case for a dense matrix. Sparsity has been extensively used in the last decade for instance to find desirable solutions of underdetermined linear systems of equations, which is the core hypothesis of compressed sensing [27, 28], or simply to enforce sparsity in any linear system if it is a valid assumption in the problem at hand. The idea is to find a basis in which the desired latent variable has a sparse decomposition. For instance, in SU, the sparsity hypothesis can be used by considering that of all the endmember signatures available for a HSI, it is rare that more than 3 or 4 materials contribute to an observed pixel. This sparsity hypothesis makes all the more sense when the endmember matrix is a large dictionary, either built from the data, or available a priori.

1.5.1 A semi-blind approach

Sparse unmixing [87] aims at solving the SU in a semi-supervised way by using a large dictionary of endmembers, typically a spectral library such as the United States Geological Survey (USGS) spectral library for minerals. Whether the library is pruned beforehand or not, there is typically many more candidate endmembers than the actual number present in each pixel, meaning that for such a large source matrix $\mathbf{S}$, the abundance matrix $\mathbf{A}$ is extremely sparse. In this context, the sparsity hypothesis makes perfect sense, and allows to select the few endmembers of the library which are actually present in the image. This allows to circumvent the difficult endmember estimation problem. The sparsity hypothesis can also be justified in the blind case if there are a lot of materials to unmix (for a spatially large or complex image), in a more classical endmember extraction and abundance estimation SU framework.

1.5.2 Sparsity in spectral unmixing

For any vector $\mathbf{a} \in \mathbb{R}^P$, we can define its sparsity level as its number of nonzero entries. This number is called the $\mathcal{L}_0$ "norm" $||\cdot||_0$ (it is not a norm per se, because it does not satisfy the homogeneity property, i.e. $\exists(\mathbf{a}, \lambda)$, $||\lambda\mathbf{a}||_0 \neq |\lambda|||\mathbf{a}||_0$). With this definition, for a given pixel (whose index we will drop here for brevity), and in the context of semi-blind SU, the goal is to find the sparsest possible solution of the linear system of equations given by the LMM

(taking into account the noise):

$$\begin{aligned} &\arg\min_{\mathbf{a}} ||\mathbf{a}||_0 \\ &\text{s.t. } ||\mathbf{x} - \mathbf{S}\mathbf{a}||_2 \leq \delta,\ \mathbf{a} \geq \mathbf{0}. \end{aligned} \tag{1.19}$$

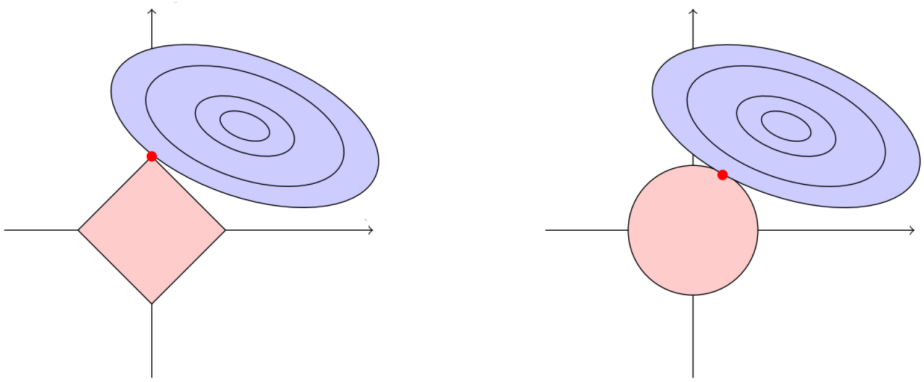
We leave aside the ASC for now for reasons that we will clarify a bit further. The uniqueness of the solution of the noiseless problem ($\delta = 0$) depends on the degree of coherence of the library $\mathbf{S}$, namely on a quantity called spark($\mathbf{S}$) (for sparsity rank), which is the smallest number of linearly dependent columns of the matrix $\mathbf{S}$. It should not be mistaken with the rank, which is the largest number of linearly independent columns of the matrix. What is more, we have the inequality spark($\mathbf{S}$) $\leq$ rank($\mathbf{A}$) + 1. For the noiseless problem, it can be shown that if there exists a solution $\mathbf{a}$ of the system $\mathbf{S}\mathbf{a} = \mathbf{x}$ with $||\mathbf{a}||_0 \leq \frac{\text{spark}(\mathbf{S})}{2}$, then this vector is the unique solution of problem (1.19), with $\delta = 0$ [51]. This condition is usually satisfied in SU applications, especially in the semi-supervised context. However, problem (1.19) is nonconvex, combinatorial and very hard to solve, which means workarounds have to be found to approximately solve it. One of them is the Orthogonal Matching Pursuit (OMP) [152], a greedy algorithm aimed at finding the most important nonzero components of the solution. It can easily be adapted to handle nonnegative solutions.

The most common way to deal with the problem is to replace the $\mathcal{L}_0$ norm by its convex relaxation, the $\mathcal{L}_1$ norm. The benefit of doing so is to turn the hard combinatorial problem (1.19) into a convex optimization problem. What is more, it is shown in [27] that under some assumptions on the linear system, the solution of

$$\begin{aligned} &\arg\min_{\mathbf{a}} ||\mathbf{a}||_1 \\ &\text{s.t. } ||\mathbf{x} - \mathbf{S}\mathbf{a}||_2 \leq \delta, \mathbf{a} \geq \mathbf{0} \end{aligned} \tag{1.20}$$

is a good approximation of a sparse solution of the linear system, and can sometimes give close to optimal solutions of problem (1.19) as well. The sparsity condition to obtain the uniqueness of the solution is, however, much more restrictive than with the $\mathcal{L}_0$ norm. In addition, there are also geometric arguments to explain the fact that the $\mathcal{L}_1$ norm does promote sparse solutions. An equivalent formulation of problem (1.20) is to minimize the data fit, under the constraint the the $\mathcal{L}_1$ norm is below some value. Fig. 1.4 shows why the $\mathcal{L}_1$ norm can enforce sparse solutions, compared to a classical Tikhonov regularization using a squared $\mathcal{L}_2$ norm. In this figure, the blue ellipses represent the level sets of a smooth function. The red domain is either the $\mathcal{L}_1$ or $\mathcal{L}_2$ balls of a given radius. It can be seen that the intersection between the $\mathcal{L}_1$ ball and the best possible level set of the smooth function is very likely to occur on the axes (i.e. on points with zero entries) with an $\mathcal{L}_1$ regularization because of the singularity of the $\mathcal{L}_1$ ball on the axes. On the contrary, this phenomenon has very little chance to happen with the $\mathcal{L}_2$ ball.

Problem (1.20) is also equivalent to the following Lagrange formulation, for an appropriate



(a) (b)

Figure 1.4: A geometric explanation (in two dimensions) of the fact that the $\mathcal{L}_1$ norm enforces sparse solutions (a), compared to an $\mathcal{L}_2$ norm regularization (b). These images were borrowed from [109].

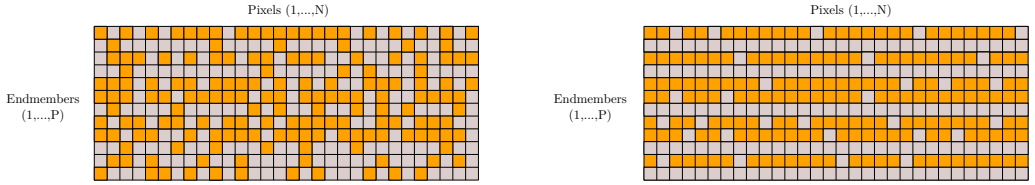
choice of the regularization parameter λ:

$$\begin{aligned} &\arg\min_{\mathbf{a}} \frac{1}{2}||\mathbf{x}-\mathbf{S}\mathbf{a}||_2^2 + \lambda||\mathbf{a}||_1 \\ &\text{s.t. } \mathbf{a} \geq \mathbf{0}. \end{aligned} \tag{1.21}$$

This problem is convex, but remains nondifferentiable and cannot be easily solved. In addition, the ANC has to be taken into account. Also, the problem with the $\mathcal{L}_1$ formulation is that the ASC can no longer be enforced if needed. The reason for this is that the ASC forces the nonnegative abundance coefficients to have a constant sum; hence the $\mathcal{L}_1$ norm of the abundance vector is constant as well and cannot be minimized without breaking the ASC. Without the ASC, problem (1.21) can be efficiently solved by proximal methods, such as the ADMM, as done in [87], with the Sparse Unmixing by variable Splitting and Augmented Lagrangian (SUnSAL) algorithm.

Other types of sparsity can be envisioned for SU: in particular, the so-called "collaborative" sparsity hypothesis is very sound in SU applications, and has been used in [85], in an adapted version of the SUnSAL algorithm. The rationale is that for SU with a spectral library, many irrelevant spectral signatures of the library will not be present in any pixel of the dataset, which amounts to say that the support of the nonzero coefficients is the same for all pixels. In terms of sparsity, this means that the number of nonzero *rows* of the abundance matrix $\mathbf{A} \in \mathbb{R}^{P\times N}$ has to be small. In other words, we force a certain number of rows of the abundance matrix to be entirely zero. The resulting optimization problem to solve has to be written for the whole image:

$$\begin{aligned} &\arg\min_{\mathbf{A}} \frac{1}{2}||\mathbf{X}-\mathbf{S}\mathbf{A}||_F^2 + \lambda\sum_{p=1}^{P}||\mathbf{a}_p||_2 \\ &\text{s.t. } \mathbf{A} \geq \mathbf{0}, \end{aligned} \tag{1.22}$$



(a) (b)

Figure 1.5: Difference between $\mathcal{L}_1$ sparsity (a) and collaborative sparsity (b) on the abundance matrix. The inactive pixels (where the abundance is zero) are shown in grey.

where $\mathbf{a}_p$ is a row (transposed in order to get a column vector) of the abundance matrix $\mathbf{A}$. The term $\sum_{p=1}^{P} ||\mathbf{a}_p||_2$ can be seen as the $\mathcal{L}_1$ norm of a P-dimensional ector containing in each entry the $\mathcal{L}_2$ norm of the abundance coefficients for a given pixel. In this sense, this term can be interpreted as a mixed norm for the matrix $\mathbf{A}$ [97]. The $\mathcal{L}_{p,q}$ mixed norm of a matrix $\mathbf{A} \in \mathbb{R}^{P\times N}$ is defined for any p and $q \geq 1$ (p and q can even take infinite values, in which case one has to take the limit for p or $q \to \infty$, which amounts to take a supremum over the corresponding dimension) as:

$$||\mathbf{A}||_{p,q} = \left(\sum_{i=1}^{P} \left(\sum_{j=1}^{N} |a_{ij}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}}. \tag{1.23}$$

With this definition, the penalty in Eq. (1.22) is a $\mathcal{L}_{2,1}$ mixed matrix norm.

The difference between the regular and collaborative sparsity penalties on the entries of the matrix on which they apply is shown in the diagram of Fig. 1.5. In our case, the abundance matrix becomes row-sparse. This type of norms will be encountered later in the manuscript and will prove useful to handle SV in a certain paradigm. In addition, we can remark that this collaborative penalization is no longer at odds with the ASC.

Finally, in SU in general, the abundance maps are known to exhibit piecewise smooth patterns. With a favorable enough SNR (say more than 20dB, which is reasonable for most airborne of satellite remote sensing HSIs), it is not always necessary to explicitly enforce this constraint to obtain visually meaningful abundance maps. However, in a semi-blind sparse unmixing context, the abundance maps can turn out to be a bit noisy due to the large size of the library, and it can be a good idea to explicitly enforce spatial smoothness of the abundance maps. This was first done in [88] using a Total Variation (TV) penalization, and was later refined using nonlocal means [174], in order to take advantage of similar spatial structures in the image rather than simply neighborhood information.

1.5.3 A blind variant

The sparse unmixing framework can be a way to avoid the estimation of the endmembers, but it requires an a priori known spectral library, and is therefore a semi-blind approach only.

Recently, using collaborative sparsity as introduced above, the sparse unmixing problem was reformulated in a completely blind way in [8] and [89]. These works are based on the idea that if the pure pixel assumption holds, then the endmembers are present among the pixels of the image, which means that some columns of the data matrix $\mathbf{X}$ are the endmembers. $\mathbf{X}$ is then used as a dictionary of N candidate endmembers (here we have $P = N$). This dictionary is used in the unmixing process, with the constraint that only a small of the pixels (the endmembers we want to recover) actually intervene to reconstruct all the image. Therefore, only a small number of rows of the abundance matrix should be nonzero. To enforce this, the following optimization problem is defined:

$$\begin{aligned} &\arg\min_{\mathbf{A}} \frac{1}{2}||\mathbf{X} - \mathbf{X}\mathbf{A}||_F^2 + \lambda||\mathbf{A}||_{2,1} \\ &\text{s.t. } \mathbf{A} \geq \mathbf{0}. \end{aligned} \tag{1.24}$$

The problem is formulated here without the ASC, but it could be included to the problem if required, contrary to $\mathcal{L}_1$ sparsity. The interest of this formulation is the replacement of the library $\mathbf{S}$ by the data matrix $\mathbf{X}$, used as a self-dictionary. With a large sparsity penalty, only the few most relevant spectral signatures of the data will be used as endmembers. This avoids an EEA to be used beforehand, and replaces the ID estimation step with the appropriate tuning of the regularization parameter, but requires pure pixels.

1.6 Main Limitations of the LMM

We have summarized here the different types of approaches for linear SU. Theses techniques have proved very useful, but are subject to two main limitations. Let us assume that we have extracted endmembers in a HSI, and have then a simplex we can work with. It can happen that a given pixel of the image falls outside of this simplex, as shown in Fig. 1.6.

In that case, solving the usual constrained least squares problem to obtain the abundances

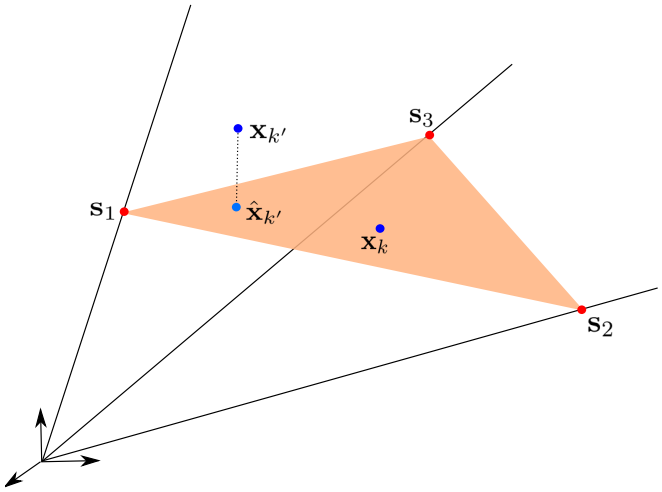


Figure 1.6: Example of a pixel which does not satisfy the usual LMM.

amounts to project the new pixel onto the simplex, and the obtained abundances are those of the projection. Then there is going to be an error on the abundance estimation for this pixel. In case there is no important missing endmember in this pixel, there can be two limitations of the LMM which can cause this:

- Nonlinearities: it may happen that in the considered pixel, the mixture between the contributions of the endmembers in this pixel is not linear. For example, when the FOV of the sensor for this pixel corresponds to a tree canopy, or in urban scenarios. In those cases, the light which reaches the sensor may have interacted with more than one material on the ground, and has bounced on objects multiple times. In those cases, it can be interesting to consider more complex mixing models, such as bilinear ones [123, 114], accounting for the interaction with up to two constituents on the ground, or "multilinear" ones, which can theoretically account for higher order interactions [81, 110]. An example of the geometric interpretation of a nonlinear mixture is shown in Fig. 1.7. Nonlinear mixing models and the corresponding unmixing algorithms have been a fertile research avenue in the community for the last decade, and comprehensive reviews can be found in [80, 49]. This subject will not be developed further in this thesis, which focuses on the second limitation of the LMM.

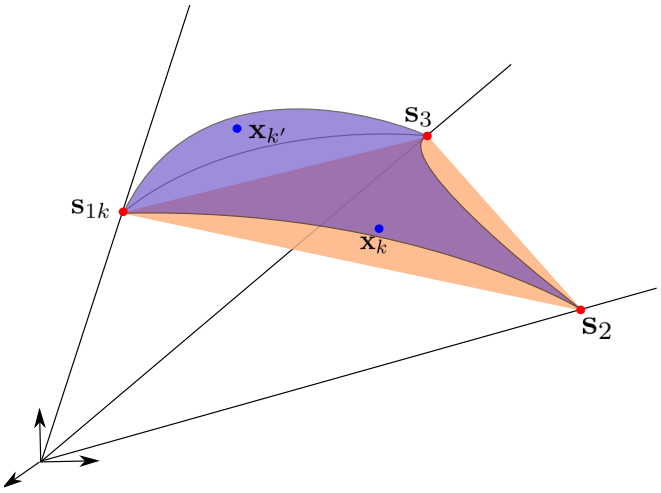


Figure 1.7: A nonlinear mixture of the three endmembers for pixel $\mathbf{x}_{k'}$.

- Spectral Variability: the simplex we are dealing with here may not be suited to a given pixel for another reason, with a fundamentally different physical cause. Indeed, if an endmember is usually considered to be a single point (spectrum) in the feature space, all materials present intra-class variability in practice, which can modify locally the spectrum of the pure materials, regardless of the mixing process. Different physical phenomena, which will be detailed in the next chapter, can cause this diversity [143, 172]. An example is presented in Fig. 1.8, where a pixel which was not in the initial simplex can be well explained by considering a local variant of the endmember $\mathbf{s}_1$. This allows to define a new local simplex, suited for this pixel. In this case, the mixture is still linear, but we consider that endmembers are not fixed in all pixels.

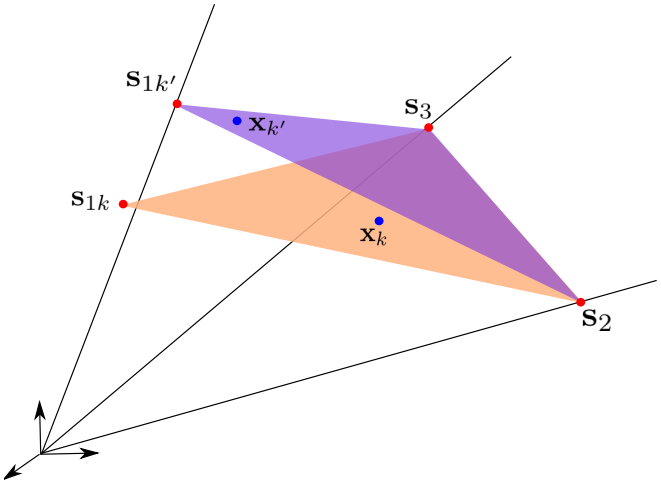


Figure 1.8: Spectral variability in a LMM framework.

Both limitations of the LMM are important, but have not received so far the same attention from the community. It can make sense to consider nonlinear mixtures when it is relevant, but the risk is to try to explain a bad fit of the LMM only because of nonlinear effects, while SV effects are equally important, if not predominant.

1.7 Partial Conclusion

This chapter introduced the spectral unmixing problem and reviewed the main lines of attack to address it, in a linear framework. We have separated the different approaches into three categories. The Spectral Unmixing (SU) problem is a very specific Blind Source Separation (BSS) problem for which convex geometry plays a great role due to the Abundance Nonnegativity (ANC) and Abundance Sum-to-one (ASC) constraints. The classical unmixing chain has three steps. The first task is to estimate the number of endmembers to consider. Then these endmembers are extracted from the data using geometric arguments. Finally, the abundance coefficients are estimated. Other types of more statistical approaches to SU exist and make explicit (or implicit in the case of Nonnegative Matrix Factorization (NMF) problems) assumptions on the statistical properties of the variables used. A final class of methods use a spectral library as a dictionary of candidate endmembers, and use the sparsity properties of the abundance vectors when decomposed in such large dictionaries of signatures in order to select the most relevant signatures and assess their proportions in the image.

However, all these approaches assume a constant spectral signature for each material. In other words, Spectral Variability (SV) is not taken into account at all. The next chapter will focus on linear SU techniques which take the intra-class variability of the materials into account.

Chapter 2

State of the art for spectral variability in spectral unmixing

Contents

2.1 Introduction

This Chapter addresses the spectral variability issue and how it has been handled so far in the literature. We present the different causes of spectral variability (SV), and review some existing approaches to tackle it. SV refers to endmember variability in a broad sense, that is to the fact that the spectral signature of a material can vary either in the spatial domain of the image, the temporal domain, and so on. We first introduce a general framework to express any kind of endmember variability. We then propose a classification of these methods in four categories, inside which they are described and compared. The different algorithms are described mostly considering variability in the spatial domain, since it is the most common in the literature, although some techniques specifically designed to address SV in the temporal or angular domains are also discussed.

Note that many developments on the topic have surfaced during the preparation of this thesis. That is why, at least in this Chapter, some recent developments which are directly connected to the work presented in this thesis will be not be discussed here, but will be introduced and compared in the appropriate chapters. In any case, note that a recent review of the latest published methods is proposed in [52].

SV is concerned with taking into account the fact that any endmember (whatever the

definition) always has a certain intra-class variability. This is a natural statement, and a phenomenon that the community has always been aware of. Notwithstanding, the problem has rarely been directly addressed in the works on SU. Indeed, the vast majority of SU techniques consider that the endmember matrix $\mathbf{S}$, once extracted, is fixed. In addition, SV, along with nonlinear effects, is one of the main causes of errors in a conventional SU framework. Since nonlinear mixtures have gathered much more interest in the community, errors due to SV can be easily mistaken for nonlinear effects, whereas the LMM does not always have to be questioned. The observed errors can also be due to the fact that the endmembers used for the whole image may locally not be good representatives of the different materials. Incorporating variability in SU amounts to allow this endmember matrix to vary somehow, in the spatial or temporal domain for instance.

The causes of variability can be very diverse. If we are interested in spatial variability within an image, SV effects can be related to:

- the changing illumination conditions during the acquisition process. It is well known that radiance and reflectance are physical quantities which vary depending on the incidence and viewing angles. On a flat surface, there is a priori no reason that these quantities should change along the image, because the sun and sensor (except maybe on UAVs) are far away from the scene. However, when the topography of the observed scene is not flat, then the acquisition angles change locally, since for a given spatial location they are defined w.r.t. the tangent plane to the surface. Fig. 2.1 illustrates these considerations. The physics of reflectance is a quite complex topic, which has been widely studied, in particular in planetary science. The observed reflectance is a function of the acquisition angles, but also of the optical parameters of each material. Even though the geometric parameters are the same for all materials, the effects of topography on the observed reflectance are material specific, but correlated along materials, because they share the same cause. Radiative transfer equations are used to model these phenomena, and some physical models exist to relate the geometric parameters during the acquisition, and the photometric parameters of the materials to the observed reflectance. Shadow effects can also be linked to the geometry of the scene, but this may involve nonlinear effects more than spectral variability since there is no direct illumination from the sun. It can still be interpreted as SV.
- the intrinsic variability of the materials. This type of variability is probably the most important in terms of impact, but also the hardest to model. Indeed, it is purely material dependent and usually corresponds to the variation of a hidden parameter, which is not taken into account in the mixing model. Examples of this include the variation of the concentration of chlorophyll in green vegetation, which will affect its color, or the effects of soil moisture on reflectance. More generally, intrinsic variability is due to physico-chemical variations within the materials. Hence, incorporating this in SU can be hard because different variability models have to be used for each constituent of the image.
- atmospheric effects. These effects can be present if the image is in radiance units, before any atmospheric correction has been performed on the data. In such cases, the

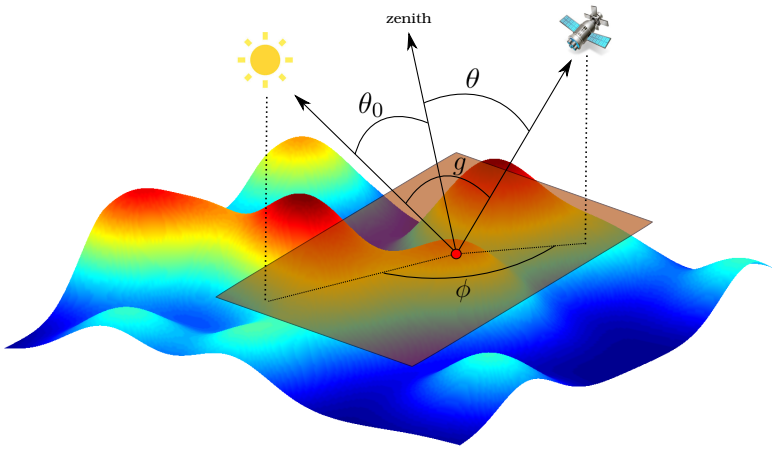


Figure 2.1: Acquisition angles for a given spatial location (red dot). The tangent plane at this point of the surface is in brown. The incidence angle is θ_0, the emergence angle is θ, and the angle between the projections of the sun and the sensor is the azimuthal angle, denoted as ϕ. g is the phase angle. θ_0 and θ are defined w.r.t. the zenith, which is defined locally (in each point of the observed surface) as the normal to the observed surface at this point.

properties of the atmosphere can vary locally in the image. Atmospheric correction cannot be perfect, and these effects can still happen when the data has been converted to reflectance units.

In the case of temporal variability, more effects can be found: appearance or disappearance of a material in a certain time frame, temporal variations, in particular seasonal variations for vegetation and snow... The atmosphere can also vary from one acquisition to the other. In the case of multiangular data, the acquisition angles vary from one acquisition to the other, in addition to the effects of topography.

2.2 A general framework

In this section, we formulate mathematically the SV problem in its most general form, for any type of variability in any domain (spatial, time, angular) whatsoever. We will use the LMM here, although everything in this paragraph can be easily generalized to a nonlinear mixture model.

As we have mentioned above, dealing with spectral variability can be seen as considering that the source matrix $\mathbf{S}$ is not constant in space, time or, more generally, between different datasets. Mathematically, let us consider a dataset, which is partitioned into K subsets indexed by k, and the corresponding endmember signatures:

$$\boldsymbol{\mathcal{X}} \equiv \{\mathbf{X}_k\} \quad \text{and} \quad \boldsymbol{\mathcal{S}} \equiv \{\mathbf{S}_k\} \quad \text{for} \quad k = 1, ..., K. \tag{2.1}$$

The index k can denote a partition in the spatial domain (mostly at the pixel scale, but also

possibly at the scale of larger spatial regions), or denote several datasets if we deal with a sequence of images acquired over the same scene at different time dates (temporal variability), or with different acquisition angles (angular variability), or, more generally, between distinct datasets (which are supposed to share at least one endmember). In each subset, the sources are still a linear mixture of the abundances:

$$\mathbf{X}_k = \mathbf{S}_k\mathbf{A}_k + \mathbf{E}_k. \tag{2.2}$$

Of course, the different subsets may not be independent: there are usually relationships and some sort of continuity between certain parameters. In the spatial domains, abundances as well as spectral variability effects can be spatially correlated (e.g. for a smooth enough topography). In the temporal domain, abundances are also very correlated from one date to another, although as in the spatial domain there can be discontinuities. Seasonal variations can exhibit continuous or even periodic effects. Angular variations can be a continuous function of the angles. These considerations advocate for a joint processing of the data rather than totally an separate processing of each subset, using appropriate continuity or correlation hypotheses.

2.3 Spectral variability in the spatial domain

This section reviews a number of techniques aimed at addressing SV in SU. Most of them are mentioned in one of the two review papers available on the subject [143, 172]. In [172], the methods are categorized in two classes, depending on whether the methods see an endmember class as a set of signatures, or as a probability distribution. Here, we categorize the techniques in four different classes: those which are based on the concept of endmember bundles, those based on local spectral unmixing, those based on computational models, and those based on parametric physics-based models.

2.3.1 Spectral Bundles

The most natural way to include spectral variability in SU is to replace the signature of each endmember by multiple candidate endmembers for the corresponding material. Ideally, if a library of spectral comprising several instances of each material under various variability conditions is available, then it should be used for SU as a dictionary, not unlike the approach described in section 1.5. Geometrically, the simplex of Fig. 1.2 in section 1.3 can now change in every pixel, and its vertices can be selected within the pool of candidates for each material, as shown in Fig. 2.2. However, the availability of such libraries is conditioned to controlled in situ measurements which are usually very specific and quite costly to carry out. In the light of this observation, we can wonder how this library could be directly build from the data, in a completely blind way. This question is exactly the motivation behind spectral bundles.

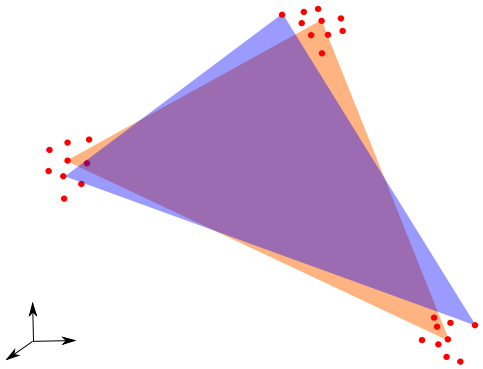


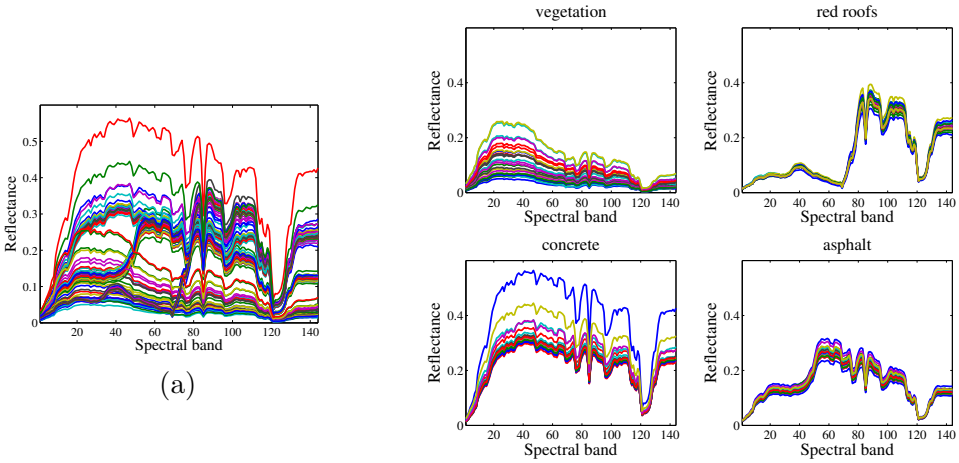
Figure 2.2: Concept of spectral bundles.

2.3.1.1 Extracting spectral bundles

The concept of spectral bundles was introduced in [145], under the name "Automated Endmember Bundles" (AEB). The underlying idea is very simple: let us assume that there are a certain number of pure pixels for each material in the image. Then each of these pure pixels explains a part of the SV present in the image for this material, and can be considered as a suitable candidate endmember. In order to extract them, several subsets of the image are randomly selected (possibly sampling without replacement to ensure that different endmember instances are selected every time). An endmember extraction algorithm (EEA) is run on each of these subsets, to extract as many signatures as the number of endmembers considered globally. If there is at least one pure pixel in each subset for each material, then different instances of each endmember are likely to be selected. All the candidate endmembers are then gathered in a dictionary of candidate endmembers. However, since most EEAs are stochastic, the extracted sources are not *aligned*, i.e. the order of the endmembers is not the same from one subset to the other, and there is a priori no grouping of the different signatures into classes containing different instances of the same endmembers. To solve this problem, a clustering step is required, in order to group the signatures into P bundles of candidate endmembers for the different materials. This can be done using for instance the k-means algorithm, with a suitable distance (see Fig. 2.3). The most two popular in HSI processing are probably the spectral angle and the Euclidean distance. The former has the advantage of being insensitive to scalings of the vectors, which will prove useful later to be robust to illumination changes in the image. However, it is not easy to tune the parameters of the bundle extraction (number and size of the subsets to use) in order to get optimal performance. Note that if there are no pure pixels in the image, the efficiency of this technique with non pure pixel based EEAs is unclear, since it is not guaranteed that the extrapolated endmembers can be explained from a SV point of view.

2.3.1.2 Abundance estimation for multiple endmember instances

Once the dictionary of bundles is built, the abundances still have to be estimated. Before describing how, let us denote the dictionary created from the bundles by $\mathbf{B}$. The clustering



(a) (b)

Figure 2.3: Example of an endmember pool extracted by the AEB approach (a), clustered into 4 meaningful classes (b).

step defines a group structure[1] on the abundance coefficients and on the dictionary. We denote this group structure by G, and each group G_i, $i = 1, ..., P$ contains m_{G_i} signatures, so that representative j of group G_i in the dictionary is denoted as $\mathbf{b}_{G_i,j}$. Then there are $Q = \sum_{i=1}^{P} m_{G_i}$ columns in $\mathbf{B}$. From the endmember bundles, there are multiple ways to extract abundances, several of which are detailed below.

The Fisher Discriminant Nullspace (FDN) approach [92] is a dimension reduction technique based on finding a linear transformation which maps the data to a subspace where the intra class variability is minimized, while the inter class variability is maximized. More precisely, the within class-scatter matrix $\mathbf{K}_w \in \mathbb{R}^{L\times L}$ and the between-class scatter matrix $\mathbf{K}_b \in \mathbb{R}^{L\times L}$ are defined as:

$$\begin{aligned}\mathbf{K}_w &= \frac{1}{N}\sum_{i=1}^{P}\sum_{j=1}^{m_{G_i}}(\mathbf{b}_{G_i,j} - \bar{\mathbf{b}}_{G_i})(\mathbf{b}_{G_i,j} - \bar{\mathbf{b}}_{G_i})^\top \\ \mathbf{K}_b &= \frac{1}{N}\sum_{i=1}^{P} m_{G_i}(\bar{\mathbf{b}}_{G_i} - \bar{\mathbf{b}})(\bar{\mathbf{b}}_{G_i} - \bar{\mathbf{b}})^\top,\end{aligned} \tag{2.3}$$

where $\bar{\mathbf{b}}$ is the mean vector of the whole dictionary $\mathbf{B}$ and $\bar{\mathbf{b}}_{G_i}$ is the mean vector for the instances of group G_i. Note that the covariance matrix of the whole dictionary $\mathbf{K}$ can be expressed as $\mathbf{K} = \mathbf{K}_w + \mathbf{K}_b$. With these definitions in mind, FDN looks for the linear transformation $\mathbf{W}$ solution of:

$$\arg\max_{\mathbf{W}} \frac{\det(\mathbf{W}^\top \mathbf{K}_b \mathbf{W})}{\det(\mathbf{W}^\top \mathbf{K}_w \mathbf{W})}. \tag{2.4}$$

[1] This denomination should not be understood in the sense of the algebraic structure. Instead, this simply means that the endmembers (and hence the abundance coefficients) are grouped into a certain number of clusters, which we call groups here.

It can be shown that if $\mathbf{K}_w$ is invertible, then the columns of $\mathbf{W}$ are the eigenvectors of $\mathbf{K}_w^{-1}\mathbf{K}_b$. Unfortunately, in most cases the scatter matrices are singular because of the limited number of samples (candidate endmembers) available compared to the dimension of the data. To solve this problem, a possibility is to first project the data onto the null space of $\mathbf{K}_w$ (where the denominator of Eq. (2.4) vanishes), and then to find vectors maximizing $\det(\mathbf{W}^\top \mathbf{K}_b \mathbf{W})$ (the eigenvectors of the between class scatter matrix in the projected domain associated to the largest eigenvalues). Once the projection $\mathbf{W}$ has been found, then the unmixing can be performed in the lower dimensional projected space, in a very conventional way using the classical Fully Constrained Least Squares Unmixing (FCLSU) algorithm (detailed in section 1.3.3), and the projected centroids of each endmember bundle as the endmembers.

Multiple Endmember Spectral Mixture Analysis (MESMA) [131] is a technique which aims at selecting in each pixel of the HSI the best endmember candidate for each material in terms of data fit. To do that, the FCLSU algorithm has to be run using all possible combinations of candidate endmembers. However, the problem thus becomes combinatorial and a brute force approach rapidly becomes untractable. To mitigate this, MESMA browses through combinations of 2 to a certain number of endmembers (fixed by the user) to limit the combinatorics of the problem. In addition, since we expect no more than 3 or 4 materials to be present simultaneously in each pixel, this technique makes sense. The method still remains computationally expensive. Interestingly, this approach was recently combined with sparse regression methods to prune the dictionary beforehand in order to alleviate its computational load [90].

Another technique to incorporate the bundle information to SU is to use machine learning approaches. For example, as proposed in [117], the availability of a bundle allows to generate training data by simulating mixtures of a selection of the candidate endmembers in various controlled proportions. Then classes are created by discretizing the unit abundance simplex into several areas corresponding to different mixing proportions. Then a multiclass Support Vector Machine (SVM) is trained on the simulated data, before being used to classify the actual data to be tested. A more recent technique uses Gaussian Processes to learn the function linking the training data to the abundances has been proposed [154]. It has the advantage of not requiring to discretize the solution space for the abundances.

The usual FCLSU algorithms can also be used, simply replacing the usual endmember matrix by the dictionary $\mathbf{B}$. Each instance of each endmember is then associated to an abundance map. Under the ANC and ASC, it seems natural to compute the global abundances of each material by summing the contributions of each instances within the corresponding bundle. It turns out that this has a very natural geometric interpretation. Indeed, we can write the LMM in the context of FCLSU with spectral bundles in one pixel in two different ways:

$$\mathbf{x} = \sum_{m=1}^{Q} a_m \mathbf{b}_m = \sum_{i=1}^{P} \left(\sum_{j=1}^{m_{G_i}} a_{G_i,j} \mathbf{b}_{G_i,j} \right), \tag{2.5}$$

where $\mathbf{b}_m$ is the m^{th} column of this dictionary, and $a_{G_i,j}$ is the abundance coefficient associated to $\mathbf{b}_{G_i,j}$. Now, if we want the global abundance of material i to be $\alpha_i = \sum_{j=1}^{m_{G_i}} a_{G_i,j}$, then we

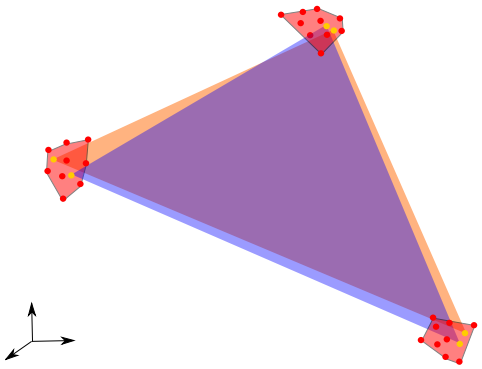


Figure 2.4: Geometric interpretation of using FCLSU on the whole extracted dictionary. The red polytopes are the convex hull of the different bundles. The yellow points are accessible endmembers when using FCLSU, whereas they were not extracted by the EEA.

have to rewrite Eq. (2.5) as:

$$\mathbf{x} = \sum_{i=1}^{P} \left(\sum_{j=1}^{m_{G_i}} a_{G_i,j} \right) \left(\frac{\sum_{j=1}^{m_{G_i}} a_{G_i,j} \mathbf{b}_{G_i,j}}{\sum_{j=1}^{m_{G_i}} a_{G_i,j}} \right) = \sum_{i=1}^{P} \alpha_i \mathbf{S}_i^*, \tag{2.6}$$

with $\mathbf{S}_i^* = \frac{\sum_{j=1}^{m_{G_i}} a_{G_i,j} \mathbf{b}_{G_i,j}}{\sum_{j=1}^{m_{G_i}} a_{G_i,j}}$. This matrix actually contains new "equivalent" endmembers for this pixel, associated with the global "intuitive" abundance coefficients. For a certain material, this new endmember is actually a weighted mean of all the available instances of this material, where the weights are the abundances extracted by FCLSU. Therefore, the normalized coefficients of this weighted mean are nonnegative and sum to one. This means that each element of $\mathbf{S}_i^*$ is a convex combination of the instances of the corresponding endmember. Geometrically, each equivalent endmember belongs to the convex hull of the elements of the bundle for this material. In this sense, finding abundances with FCLSU rather than MESMA allows more freedom in terms of SV: the latter constrains each pixel to come from a combination of the extracted sources, while the former theoretically allows any point inside the convex hull of the each bundle to be a local endmember. This geometric interpretation is shown in Fig. 2.4. The per-pixel equivalent endmember $\mathbf{S}_i$ is of course only defined if at least one instance group G_i is active in this pixel. Otherwise, it makes no sense trying to extract spectral variability in a pixel from a material which is not present. The main limitation of all the abundance estimation techniques using bundles is that the results are heavily dependent on the quality of the extracted bundle, which is not easy to assess, due to the unknown validity of the pure pixel hypothesis in each subset and the randomness, both in the subset generation and in the endmember extraction itself. Recent methods trying to extract bundles in a more refined way, in particular taking spatial information into account have also been developed [167, 135].

2.3.2 Local Spectral Unmixing

This section is concerned with Local Spectral Unmixing (LSU) approaches, i.e. techniques which perform SU on spatially coherent subsets of the data (sliding windows or spatial regions). This is related to bundle extraction since it is perfectly possible to use such subsets instead of random ones in the Automated Endmember Bundles (AEB) approach [143]. However, there is an additional important difference here: not only are the endmembers extracted locally, the abundances are also estimated in a local setting. The rationale behind LSU is that nonlinearities and especially SV are likely to be mitigated when working on local regions of the data. For example, when the observed scene as a slowly varying topography relatively to the spatial resolution, then illumination effects can be seen locally as approximately uniform. Even with a more uneven topography, defining two regions corresponding to two sides of a hill, for instance, makes sense in this context, since even if the materials involved are the same in both regions, two local endmembers will yield better unmixing results than a global one for both sides. This applies to any type of SV which could be spatially correlated. Also, pixels within local regions tend to be share the same active endmembers, in slowly varying proportions. This means that in local regions, multiple mixtures (when the data simplex can be partitioned into several sub-simplices, each accounting for a mixture of different materials [171]) are less likely to occur.

2.3.2.1 Sliding windows

The simplest way to define spatial regions is to use local sliding windows, as in the works of [66] and [29]. In [66], LSU is shown to be spatially adaptive compared to usual SU because it is able to perform the unmixing at different spatial scales. It is also able to deal with SV by allowing to split the vegetation class into two subclasses located at different locations of the image: rainforest and mangrove, which was not possible using conventional SU. However, an underlying problem related to LSU is that the final expected result for SU is usually a set of *global* abundance maps. Thus, it can be useful for interpretation to go from local to global SU results. The most straightforward way to do this is to group a posteriori all the local endmembers into a bundle, once again using a clustering algorithm, and to sum up the local contributions of the instances of a class in each pixel. We will come back later to this issue.

2.3.2.2 Binary Partition Tree-based LSU

Ideally, the local subsets used in LSU should be meaningful regions of a segmentation of the hyperspectral image, rather than just sliding windows. Here, we present an approach which goes even further [161]: instead of using an algorithm to perform HSI segmentation and defining the resulting regions as the subsets of LSU, it designs a segmentation which is optimal in terms of LSU performance. This strategy is based on an image processing tool [138] which has been recently adapted for HSI processing [155]: the Binary Partition Tree (BPT).

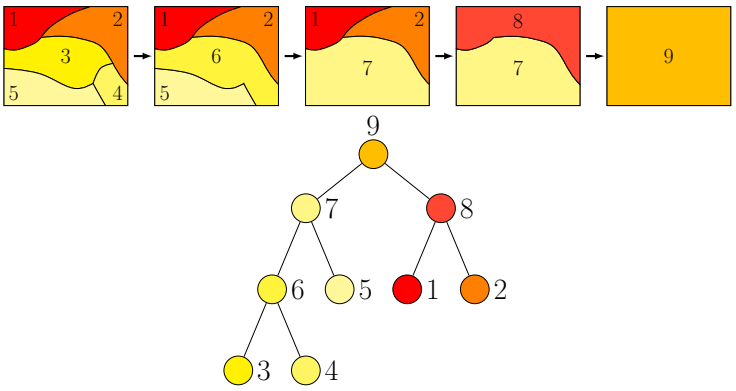


Figure 2.5: Example of the construction of a Binary Partition Tree. This image was taken from [150]. At each step of the merging process, the two most similar regions are merged.

A BPT is a hierarchical region-based representation of an image, useful for segmentation or object recognition, among others. Its interest is that the meaningful regions of the image are represented at different scales. The construction of a BPT is conceptually quite simple. It requires an initial partition of the image, fine enough not to undersegment meaningful spatial structures. It can be simply the pixel level of the HSI itself, or a preliminary segmentation of the image, for example using a watershed segmentation [146], superpixel generation techniques [1], or other segmentation algorithms such as a mean-shift clustering [41]. Then the BPT is built by iteratively merging the two most similar adjacent regions of the image, until there is only one region left, whose support is this of the whole image. By doing so, a tree structure is created, as shown in Fig. 2.5. Every region of the initial partition is called a leaf of the tree. The regions resulting from a merging are called nodes of the tree, and the last region is called the root of the tree. In order that such a merging process be possible, for any region $\mathcal{R}_i$ (where i is the index of the region), the definition of a *region model* $\mathcal{M}_\mathcal{R}$ is necessary and the notion of "similarity" of two regions should be precised, by defining a *merging criterion*, i.e. a similarity measure $\mathcal{O}(\mathcal{R}_i, \mathcal{R}_j)$ between any $\mathcal{R}_i$ and $\mathcal{R}_j$. For any region of a HSI, one of the simplest possible region models is the mean vector of the pixels composing it, that is:

$$\mathcal{M}_{\mathcal{R}_i} \equiv \bar{\mathbf{x}}_{\mathcal{R}_i} = \frac{1}{|\mathcal{R}_i|} \sum_{i \in \mathcal{R}_i} \mathbf{x}_i, \tag{2.7}$$

where $|\mathcal{R}_i|$ is the cardinality of region $\mathcal{R}_i$. As a simple merging criterion, one can choose the usual Euclidean distance, or the Spectral Angle Mapper (SAM):

$$\mathcal{O}_{SAM}(\mathcal{M}_{\mathcal{R}_i}, \mathcal{M}_{\mathcal{R}_j}) = SAM(\bar{\mathbf{x}}_{\mathcal{R}_i}, \bar{\mathbf{x}}_{\mathcal{R}_j}) \triangleq \arccos\left(\frac{\bar{\mathbf{x}}_{\mathcal{R}_i}^\top \bar{\mathbf{x}}_{\mathcal{R}_j}}{||\bar{\mathbf{x}}_{\mathcal{R}_i}||_2 ||\bar{\mathbf{x}}_{\mathcal{R}_j}||_2}\right). \tag{2.8}$$

This region model and those merging criteria, though quite simple, can already help to segment a HSI into meaningful regions [155]. However, for a SU application, it makes sense to design unmixing based region models and merging criteria. In [161], this is done by running an EEA on each region, thus extracting a certain number of local endmembers. This number is determined using an intrinsic dimensionality (ID) estimation algorithm on the pixels of this

region. We will come back in detail on the ID estimation issue in Chapter 3. The region model is then the set of endmembers extracted in the region:

$$\mathcal{M}_{\mathcal{R}_i} \equiv \mathbf{S}_{\mathcal{R}_i} = [\mathbf{s}_1^{\mathcal{R}_i}, \mathbf{s}_2^{\mathcal{R}_i}, ..., \mathbf{s}_{d_{\mathcal{R}_i}}^{\mathcal{R}_i}] \in \mathbb{R}^{L \times d_{\mathcal{R}_i}}, \tag{2.9}$$

where $d_{\mathcal{R}_i}$ is the estimated ID in region $\mathcal{R}_i$. Defining a merging criterion between two regions under this region model amounts to finding a similarity measure between two endmember matrices, of possibly (and very likely) different sizes. To do that, [161] defines a similarity matrix $\boldsymbol{\Upsilon}_{i,j} \in \mathbb{R}^{d_{\mathcal{R}_i} \times d_{\mathcal{R}_j}}$ for two regions $\mathcal{R}_i$ and $\mathcal{R}_j$, with:

$$\upsilon_{i,j}(k,l) = SAM(\mathbf{s}_k^{\mathcal{R}_i}, \mathbf{s}_l^{\mathcal{R}_i}). \tag{2.10}$$

The similarity measure is actually a distance between any two regions of hyperspectral pixels (regardless of their shapes) characterized by their sets of endmembers, with the same number of spectral bands [68]. It is defined as:

$$O_{\text{spectral}}(\mathcal{M}_{\mathcal{R}_i}, \mathcal{M}_{\mathcal{R}_j}) \triangleq ||\boldsymbol{\zeta}_{i,j}^r||_2 + ||\boldsymbol{\zeta}_{i,j}^c||_2, \tag{2.11}$$

with $\boldsymbol{\zeta}_{i,j}^r$ (resp. $\boldsymbol{\zeta}_{i,j}^c$) a vector containing the minimum value of each row (resp. column) of $\boldsymbol{\Upsilon}_{i,j}$ in its entries. The first term of Eq. (2.11) is a measure of the overall closeness of the endmembers of region $\mathcal{R}_j$ to the endmembers of region $\mathcal{R}_i$. The second term has converse properties and is used to make the expression symmetric. Note that an extension of this region model and this merging criterion has also been proposed in [161], using in addition the similarities of the abundance maps of each region.

With this region model and this merging criterion, a hierarchical representation of the image driven by the local endmembers' similarity can be built. From the BPT, a large number of different segmentations of the image can be defined, by retaining some of its nodes to define a partition of the support of the image. The question is: how can we select the optimal one in terms of SU performance?

From any BPT, there are many ways to recover a segmentation of the image. This process is called *pruning* the tree (an example is provided in Fig. 2.6). The most simple pruning strategies consist in cutting the tree at a given height, or in selecting the partition obtained after a certain number of merging operations during the construction of the tree. Another approach is to define a cost function over the set of possible partitions given by the structure of the tree T, and to retain the one providing the lowest value of the cost function:

$$\pi^* = \underset{\pi \in \Pi(T)}{\arg\min}\ \varepsilon(\pi), \tag{2.12}$$

where π^* is the optimal partition on the tree given the cost function ε, defined on the set of all possible partitions in the tree $\Pi(T)$. The cost function value for a given partition should somehow be a combination of the values of some energy function defined for each region of the partition. Conditions on the separability properties of the cost function to ensure that the optimal partition can be efficiently obtained by a dynamic program have been studied and can be found in [95, 70, 150]. The general flowchart of the construction and pruning of a BPT is shown in Fig. 2.7 For a LSU application, a cost function which depends on the

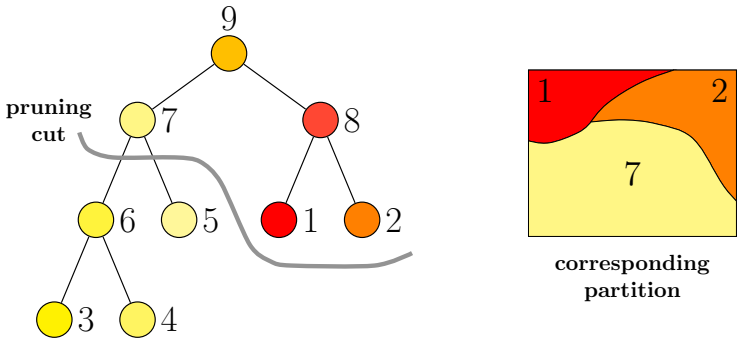


Figure 2.6: Example of the pruning of the BPT of Fig. 2.5. This image was taken from [150].

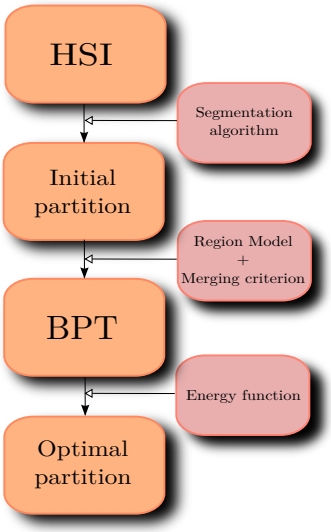


Figure 2.7: Flowchart of the construction and pruning of a BPT.

unmixing performance, but also on the cardinality (i.e. the number of regions $|\pi|$) of the partition is desirable, so as to obtain a tradeoff between reconstruction and complexity of the segmentation. This means that the abundance estimation has to be performed in each region using the local endmembers extracted. In the absence of ground truth on the materials and abundances present in the image to study, a widely used criterion for unmixing performance is the Root Mean Squared Error (RMSE) between the original pixel $\mathbf{x}$ and the reconstructed one $\hat{\mathbf{x}}$ using the endmembers, abundance vector and mixing model:

$$RMSE(\mathbf{x}, \hat{\mathbf{x}}) \triangleq \sqrt{\frac{1}{L}\sum_{l=1}^{L}(x_l - \hat{x}_l)^2}. \tag{2.13}$$

A low value of this measure means that the pixel is well reconstructed by a LMM, that is by a linear combination of the extracted endmembers, with the estimated abundances as weights. However, this measure is far from perfect since it is possible to reconstruct the data very well from poorly estimated endmembers (e.g. when the pixel is actually inside the simplex spanned by these irrelevant endmembers) and/or abundances. We denote by $\epsilon_{\mathcal{R}}(\mathbf{x}, \hat{\mathbf{x}})$ the

RMSE for a pixel $\mathbf{x}$ of region $\mathcal{R}$.

From the RMSE in each pixel, two region-wise energies can be defined as the average or the maximum value of RMSE over all the pixels of this region. Using these, the energy of a partition of the HSI can be defined as one of these two expressions:

$$\begin{aligned}\varepsilon_{\text{average}}(\pi) &= \frac{1}{N}\sum_{\mathcal{R}\in\pi}\sum_{\mathbf{x}\in\mathcal{R}}\epsilon_{\mathcal{R}}(\mathbf{x},\hat{\mathbf{x}}) + \lambda_{BPT}|\pi|,\\ \varepsilon_{\max}(\pi) &= \frac{1}{N}\sum_{\mathcal{R}\in\pi}|\mathcal{R}|\max_{\mathbf{x}\in\mathcal{R}}\epsilon_{\mathcal{R}}(\mathbf{x},\hat{\mathbf{x}}) + \lambda_{BPT}|\pi|.\end{aligned} \tag{2.14}$$

$\varepsilon_{\text{average}}$ looks for low values of the average RMSE in the regions of the partitions, but allows some large RMSE values within the regions, while $\varepsilon_{\max}$ is less sensitive to outliers, but can produce higher RMSE values on average.

With either of these two energies, the optimal partitions for any value of λ_{BPT} can be obtained through dynamic programming. When λ_{BPT} sweeps through the real line, the obtained partitions go from the initial partition to a partition with a single region containing only the root of the tree. In between are partitions with an decreasing number of regions when λ_{BPT} gets higher. In practice, especially for large values of this regularization parameter, a potentially large range of values of λ_{BPT} can lead to the same partition. The intervals defining the same partition are called persistent intervals and can be easily computed from the BPT [70]. In the end, through this approach, a segmentation of the image which is optimal in terms of LSU performance (in terms of RMSE, weighted by the number of regions in the partition) is obtained. This approach allows to perform the unmixing locally in a smarter way than simply defining sliding windows, which helps mitigating nonlinear and SV effects, and avoids the propagations of model errors the whole image, but it raises some questions. The first is how should we deal with the possible ID estimation problems in small regions, and limit their repercussions on the rest of the unmixing chain? The second comes from the fact that this LSU process generates regionwise SU results: the endmembers and abundances are only

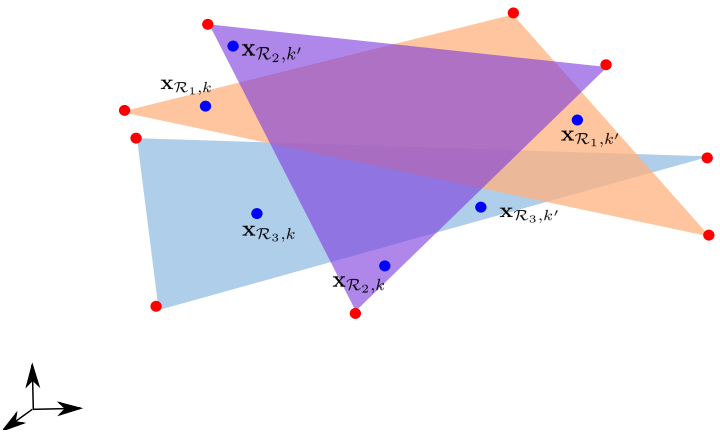


Figure 2.8: Geometric interpretation of LSU.

defined within each region. Indeed, as we show in Fig. 2.8, the geometric interpretation of LSU is clear for each region: we work inside a regionwise simplex. However, at the scale of the whole image, there is no global coherence, which prevents an immediate global interpretation of the results. How would it be possible to use these results and to reinterpret them at the global image scale, so that they could be compared to other SU approaches? These two points will be addressed later in the manuscript.

2.3.3 Computational Models

In this section, we present a different class of methods to address SV. This time, SV is explicitly taken into account in the mixing model used, contrary to bundles or LSU approaches which are more data-driven. Most of the time, the computational models we review below allow the sources to vary locally around some reference, as shown in Fig. 2.9. Theoretically, these techniques are quite powerful since they allow to capture any kind of SV. However, this flexibility can also be a drawback since no SV cause is explicitly modeled, making it sometimes hard to give a physical interpretation to the results.

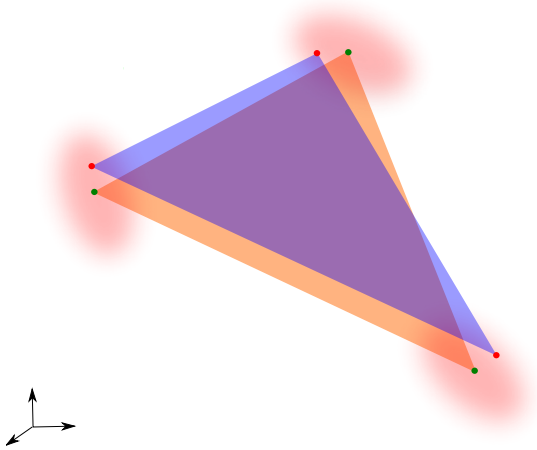


Figure 2.9: The fluctuations of local endmembers around references (in green) are at the core of most computational models to address SV.

Since the idea is to estimate parameters related to SV with little information on the physical processes causing it, it comes as no surprise that statistical techniques have been developed in this area, as was the case for highly mixed data in section 1.4.

For instance, several Bayesian approaches have been developed in the community to model the sources as statistical distributions. The Normal Compositional Model (NCM) [60] uses the usual LMM (discarding the noise term), but assumes in addition that the endmembers are normally distributed:

$$\mathbf{s}_p \sim \mathcal{N}(\mathbf{s}_{0p}, \sigma^2 \mathbf{I}_L), \tag{2.15}$$

where $\mathbf{s}_{0p}$ is a reference endmember spectrum for material p. The covariance matrix is diagonal

with the same variances in each band. This allows to write the likelihood as:

$$\mathbf{x}_k \sim \mathcal{N}\left(\sum_{p=1}^{P} a_{kp}\mathbf{s}_{0p}, \sum_{p=1}^{P} a_{kp}^2\sigma^2\mathbf{I}_L\right). \tag{2.16}$$

In addition, a uniform prior on the simplex is assumed for the abundances. The variance σ^2 is assigned a prior distribution as well, with an additional hyperparameter, which is itself assigned an uninformative prior. From this, the a posteriori distribution of the parameters is derived and sampled using MCMC methods, to access the Bayesian estimators. The fact that the variance of the endmembers has to be estimated allows to assess a posteriori the degree of SV in the processed image. However, the mean value of the endmembers is fixed by the prior endmember extraction step. Note that a similar Beta Compositional Model (BCM) was also proposed [59], replacing the Gaussian prior on the endmembers by a Beta distribution (which has the advantage of taking its values between 0 and 1, which is sound for reflectance endmembers).

More recently, the work proposed in [71] goes further by incorporating the noise term $\mathbf{e}_k$ in the model (with pixel dependent variances), and being capable of estimating both band dependent variances, as well as the means of the endmember distributions. The distribution of the noise and the prior on the endmembers are:

$$\begin{aligned} \mathbf{e}_k &\sim \mathcal{N}(\mathbf{0}, \nu_k^2\mathbf{I}_L), \\ \mathbf{s}_{kp} &\sim \mathcal{N}(\boldsymbol{\mu}_p, \mathrm{diag}(\boldsymbol{\sigma}_p)), \end{aligned} \tag{2.17}$$

where $\mathbf{s}_{kp}$ is the local endmember for material p in pixel k, ν_k^2 is the noise variance for pixel k, and $\boldsymbol{\mu}_p$ is the mean of the endmember distribution for endmember p. $\boldsymbol{\sigma}_p$ is a vector of variances for each band for endmember p, and $\mathrm{diag}(\boldsymbol{\sigma}_p) \in \mathbb{R}^{P\times P}$ is a diagonal matrix whose is diagonal is $\boldsymbol{\sigma}_p$. The mean of each endmember distribution is associated to a (truncated, in order to avoid negative values) Gaussian distribution, centered on the reference endmembers, extracted by the EEA:

$$\boldsymbol{\mu}_p \sim \mathcal{N}(\mathbf{s}_{0p}, \epsilon^2\mathbf{I}_L), \tag{2.18}$$

where ϵ^2 represents the confidence in the extracted reference. This allows to slightly correct the mean value of the endmember distributions, in case the reference endmembers are not optimal. The abundance prior is made of several Dirichlet distributed classes, so as to allow the simplex to be clustered into several distinct regions where the abundance vectors can live, and to encourage spatially close pixels to have correlated abundances. The parameters of this prior are also part of the hierarchical Bayesian model. This model is called the "Generalized Normal Compositional Model" (GNCM).

All these models are "computational" in the sense that they explicitly model SV (unlike bundles and LSU approaches), but without caring about the physical processes involved in generating the intra class variability. This is both their strength and weaknesses: they are theoretically able to model any kind of SV, but are not able to link it to physically interpretable parameters.

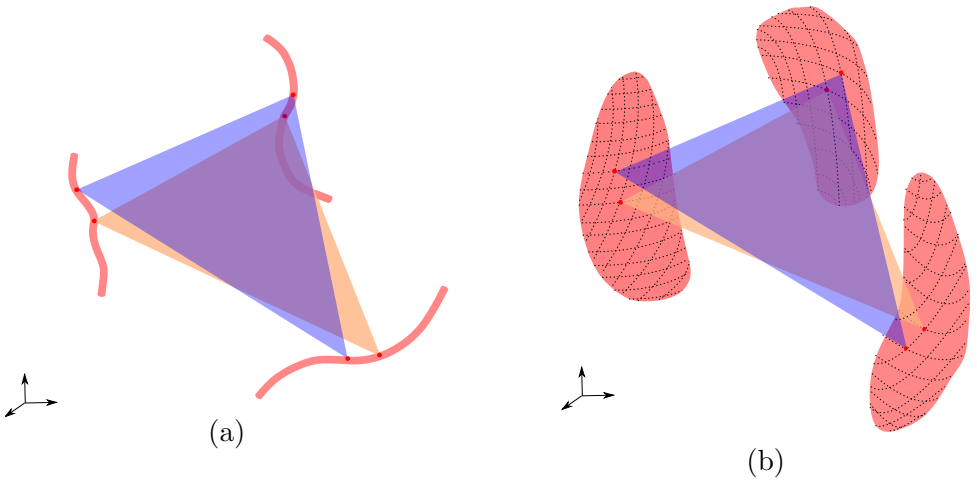


Figure 2.10: A Simple parametric model to deal with SV (1 free parameter) (a). A more complex model (2 free parameters) (b).

2.3.4 Parametric physics-based models

2.3.4.1 Recovering spectral variability from parametric models

In this last category of methods, the sources are also allowed to vary according to a specific model, but in a more constrained way than in section 2.3.3. Indeed, the idea is to use an explicit parametric model $f_p : \mathbb{R}^{n_p} \mapsto \mathbb{R}^L$ to define the achievable spectra for a material p:

$$\mathbf{s}_{kp} = f_p(\boldsymbol{\theta}_k^p), \tag{2.19}$$

where $\mathbf{s}_{kp}$ is the spectrum of material p in pixel k, and $\boldsymbol{\theta}_k^p \in \mathbb{R}^{n_p}$ is a vector of n_p physically interpretable parameters for model f_p. A limited number of n_p free parameters are included in the model describing material p (typically much less than the number of spectral bands), based on the modeling of some physical phenomenon. Fig. 2.10 shows the geometric interpretation of such models. These parametric models and the possible ranges of their parameters actually define an n_p-dimensional manifold for each material (which is the possible locus of the material's spectra), in the L-dimensional ambient space. In Fig. 2.10 are shown two examples, in the case of models with a single parameter (a), where the sources describe curves (i.e. 1-dimensional manifolds) in the ambient space. A second, more complex case is shown on Fig. 2.10 (b), where the model has two parameters. The models for the different materials are not bound to be completely independent. In some cases, e.g. when they share the same analytical expression (for a cause of variability affecting all materials in the same, or at least in a correlated way), and/or some physical parameters (for instance acquisition angles). However, in most cases, material specific models have to be used, since the causes of intrinsic variability are, by definition, material dependent.

Once the model is known, in a blind SU framework, the objective is to estimate the parameters of the models for each material and pixel (and hence the local sources), in addition

to the abundances (still with the ANC and ASC). The most straightforward way to do this is to resort to a least squares fit:

$$\left(\hat{\mathbf{a}}_k, \hat{\boldsymbol{\theta}}_k^1, ..., \hat{\boldsymbol{\theta}}_k^P\right) = \underset{\mathbf{a}_k, \boldsymbol{\theta}_k^1, ..., \boldsymbol{\theta}_k^P}{\arg\min} ||\mathbf{x}_k - \mathbf{S}_k \mathbf{a}_k||_2^2 \quad \text{s.t. } \mathbf{S}_k = [f_1(\boldsymbol{\theta}_k^1), ..., f_P(\boldsymbol{\theta}_k^P)]. \tag{2.20}$$

Refining the model further is theoretically possible by replacing the matrix product $\mathbf{S}_k\mathbf{a}_k$ by a nonlinear function of $\mathbf{S}_k$ and $\mathbf{a}_k$, but here we will limit ourselves to a LMM framework. In addition, it can make sense to add various constraints (e.g. physically plausible range of the parameters) and hypotheses using regularizations (e.g. continuity of some parameters in the spatial domain).

However, in all generality, the estimation process of Eq. (2.20) is very ambitious. The efficiency and tractability of this method in practice is heavily dependent on the number of parameters to estimate, and on the analytical expressions of the model. The problem is very difficult in a completely blind setting since the information in the data about the shape of the manifolds is directly conditioned by the number of pure or close to pure pixels. Regardless of the complexity of the physical models involved, obtaining good variability estimation results in a heavily mixed scenario seems extremely difficult, even a linear case. For blind SU, such a parameter estimation technique is expected to be efficient when the pixel is not too heavily mixed (and only for the predominant material), for a limited number of parameters to estimate, and with convenient enough analytical expressions for the models used (e.g. when the functions f_p are injective, to avoid identifiability issues). Using appropriate regularization terms (such as one using spatial information) can also help to make the problem better-posed.

2.3.4.2 Examples

There are several models in the literature which describe the reflectance of materials, such as vegetation and soil types, using various physical parameters. However, these models are usually cumbersome to use in a blind SU context, due to their complexity. Some of their parameters are rarely available beforehand in practice, and their efficient use in SU is likely to be conditioned to simplifying hypotheses, in order to make the models tractable. We mention here some models of the literature for the sake of illustration. Some examples include tree leaves reflectance models, soil moisture content reflectance models, or models describing the variations of the spectra due depending on geometric and photometric parameters. Most of these models are based on radiative transfer equations, which are adapted for specific materials, and their parameters are usually estimated using in situ measurements.

For instance, the work of [91] introduces a radiative transfer model describing the reflectance of tree leaves in the visible and near-infrared domains, depending on parameters linked to the mesophyll structure of leaves, water content and pigment concentration, among other optical parameters.

In [144], two parametric models (one is affine, the other exponential) are proposed to ex-

plain the variations in the spectrum of soil under different moisture conditions. Wavelength-dependent parameters of these two models are estimated by linear regression on in-situ measurements for two wavelengths. By assuming these parameters have been estimated for every wavelength (or at least for different wavelength ranges), the relationship between moisture and reflectance defines in each case a one-dimensional manifold (parametrized in each pixel by a real value accounting for the moisture level), which enables the local estimation of this parameter in the image. In this case, the equations are not directly related to physical modeling, but obtained in a more pragmatic and empirical way by regression of actual data acquired by varying the parameter of interest (here soil moisture content). This results in a more artificial model, which has the advantage of being mathematically tractable and can still be related to a physical parameter. Another example of a relatively simple physics based model for the same phenomenon can be found in [137].

Another well known radiative transfer model in the remote sensing and planetary science communities is the semi-analytical model designed by Hapke [73], to link reflectance of a material to its single scattering albedo (SSA), and to geometric parameters (namely the acquisition angles alluded to in section 2.1), as well as material specific photometric parameters. More details on this model will be found in Chapter 5. SSA is defined as the ratio between the scattered fraction of the incoming light to the total radiation coefficient (sum of the scattered and absorbed fractions of incoming light), in each wavelength. For instance, a perfect black body has an albedo of zero, fresh snow has a high albedo, and a perfect mirror has an albedo of one, in any wavelength range. Unlike reflectance, the albedo spectrum of a material does not depend on the acquisition angles and can thus be considered to truly characterize a material. The SSA of a material is very hard to access, as are the photometric parameters of the materials. In addition, the model is analytically complex and hence impossible to use as such for SU. However, in cases where the photometry, as well as the geometry are known, the method [112] is able to invert the model to access the SSA of the materials, and then performs the SU in the albedo domain. It has been suggested in the several works that topographic effects on the reflectance signatures can be empirically approximated by scaling variations [121, 122], hypothesis that we will validate further in this manuscript.

2.4 Temporal and Angular variabilities

In this section we focus briefly on methods to tackle SV in time sequences or in multiangular HSIs. The work of [66] performed the analysis of multitemporal HSIs using sliding windows-based LSU, but each frame was processed independently of the others. In order to take advantage of the correlations between adjacent time frames, and to better capture temporal variations, a joint process of the whole sequence is preferable.

The work in [77] exploits the additional modality of the data by using a multitemporal model for the endmembers, since they are seen as functions of the wavelength, but also of the time, thus defining a surface as an endmember rather than a single signature. These 3D signatures are then used to extract features in order to produce a classification map for the

whole time series.

More recently, a new approach to SU has been developed using the arsenal of tensor decomposition techniques [158]. Hyperspectral time series or multiangular series are seen as three-way arrays (or tensors, although this denomination is a bit abusive): $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{L \times N \times K} \equiv \{\mathbf{X}_k\}, k = 1, ..., K$, where K is the number of time frames or of angular acquisitions (one spatial modality, one spectral modality, and one temporal or angular modality). A multilinear unmixing is proposed using a nonnegative Canonical Polyadic (CP) Decomposition. The idea of the of the CP decomposition is to approximate a noisy tensor (ideally a low rank one) as a tensor of rank R, decomposed as sum of R rank 1 tensors (the tensor rank is the minimum number of terms in the decomposition required for the equality to hold exactly). More precisely, if we denote by $\mathbf{S} \in \mathbb{R}^{L \times R}$ a matrix of "spectral" factors, by $\mathbf{A}^\top \in \mathbb{R}^{N \times R}$ a matrix of "spatial" factors (the matrix $\mathbf{A}$ is transposed to comply with the definition of the abundance matrix used throughout the thesis), and by $\mathbf{T} \in \mathbb{R}^{K \times R}$ a matrix of temporal factors, the underlying multilinear model can be written as:

$$x_{lkm} = \sum_{r=1}^{R} s_{lr} a_{rk} t_{mr} \lambda_r, \tag{2.21}$$

where λ_r are scaling indeterminacies for each of the R terms. Rewriting the global model in a compact form, the nonnegative CP decomposition consists in solving the following optimization problem:

$$\begin{aligned} &\arg\min_{\mathbf{S},\mathbf{A},\mathbf{T}} ||\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{L}} \times_1 \mathbf{S} \times_2 \mathbf{A}^\top \times_3 \mathbf{T}||_F^2 \\ &\text{s.t. } \mathbf{S} \geq \mathbf{0}, \mathbf{A} \geq \mathbf{0}, \mathbf{T} \geq \mathbf{0}, \end{aligned} \tag{2.22}$$

where $\mathcal{L} \in \mathbb{R}^{R \times R \times R}$ is a diagonal tensor containing the scaling indeterminacies λ_r on the diagonal. The operator $\times_k$ is the tensor matrix product along the k^{th} mode [40], which is defined for a n^{th}-order tensor $\mathcal{Y} \in \mathbb{R}^{N_1 \times N_2 \times ... \times N_n}$ and a matrix $\mathbf{U}_k \in \mathbf{R}^{m \times N_k}$ as the tensor in $\mathbb{R}^{N_1 \times ... \times N_{k-1} \times m \times N_{k+1} \times ... \times N_n}$, whose entries are:

$$(\boldsymbol{\mathcal{Y}} \times_k \mathbf{U}_k)_{i_1,...,i_{k-1},i_k,i_{k+1},...,i_n} \triangleq \sum_{j=1}^{N_k} y_{i_1,...,i_{k-1},j,i_{k+1},...,i_n} u_{i_k,j}. \tag{2.23}$$

Note that the scaling indeterminacies can be absorbed into either of the three matrix terms, if needed. This problem is highly non-convex, yet many algorithms provide rather precise but costly computation. One of the most popular of these algorithms is an Alternated (Nonnegative) Least Squares (ALS) approach, in which the variables are optimized alternatively and iteratively (in a way similar to NMF).

We denote the spectral and spatial factor matrices with the same notations as the ones used for the endmembers and abundances throughout the thesis, although this is slightly abusive, since the CP model is a priori not motivated by physical considerations. However, in the context of CP decomposition of multitemporal or multiangular HSIs, they can be interpreted as such, and experimental evidence for this is provided in [158]. In addition, the rank R is linked to the number of endmembers to consider. The temporal factors are able to capture

seasonal variations of the materials, namely for snow cover monitoring [158]. In this work, the CP decomposition is performed in a compressed domain, making the algorithm (ProCoALS, for Projected and Compressed Alternating Least Squares) applicable to large datasets where usual tensor decomposition algorithms are limited by memory or computational load issues. Also, note that the ASC can be included if needed, depending on the SU context. The method was also adapted for multiangle images [157], where the angular factors are also physically interpretable. We will come back later to tensor decomposition approaches and show that they can be used to model variability not only in the temporal angular domains, but also in the spatial domain, using an appropriate tensor representation of HSIs.

2.5 Partial Conclusion

In this chapter, we have introduced the spectral variability (SV) issue and have presented a general framework to encompass all variability types, that is in the spatial, temporal, angular, or even multidataset contexts. We have presented the main algorithms and techniques developed to handle it in SU, mostly focusing on SV in the spatial domain, which has been far more studied than for other modalities. The methods developed for this domain have been categorized into four classes: spectral bundle methods, where several instances of each endmember are extracted from the data, Local Spectral Unmixing techniques, which perform SU locally in spatial regions of the image, computational models whose rationale is to allow local endmembers to vary around reference signatures, and finally discussed physics-based models, whose applicability is not straightforward and probably more limited than the other classes, but have the advantage of modeling variability causes using physically interpretable parameters.

We also have presented some of the few methods used to jointly process multimodal datasets in the context of temporal and angular variability, especially the recently proposed tensor decomposition based approaches.

The contributions of the thesis will be connected to several of the ramifications of the SV problem presented in this chapter:

- Part II will make the connection between sparsity and the bundles and local spectral unmixing (LSU) approaches, where the subsets on which SV occurs are spatial regions of the image.
  - Chapter 3 addresses one of the drawbacks of LSU, which is the need to estimate the intrinsic dimensionality (ID) of sometimes small regions of the HSI. This can lead to erroneous estimation of the number of endmembers to use in local regions. We review several algorithms of the literature to estimate the ID of HSIs and study their behavior experimentally in local datasets, before providing guidelines for their appropriate use in local settings.
  - In Chapter 4, we first apply sparse regression tools to LSU in order to eliminate the

contributions of the wrongly extracted local sources due to the possible overestimation of ID in small ill-conditioned regions of the segmentations. Then, we study the influence of social norms in SU using spectral bundles, by taking advantage of the group structure of the spectral bundles, after the clustering step.

- In part III, we explore the less studied direction of physics based models to model SV.
  - In Chapter 5, starting from the Hapke model, we make several simplifying assumptions to derive a simple tractable model, termed Extended Linear Mixing Model (ELMM) to address changing illumination conditions during SU, in particular due to uneven topography. In this case, the subsets where SV is handled are the individual pixels. We then design an optimization problem and two algorithms to estimate the parameters of the model, while adding regularizations to enforce desirable properties on the solutions.
  - In Chapter 6, we discuss two applications of the proposed ELMM. In the first application, we show that combining the ELMM with the LSU approach can help to interpret the LSU results at a global scale, and at the same time to extract SV related information. The second application shows the connection of tensor decomposition approaches to the ELMM, be it using time series, multiangular series of HSIs (the different frames are then the subsets on which variability is addressed here), or even a new tensor representation of a regular HSI, which is proven useful to extract spatially related SV content through nonnegative CP decomposition.

Part II

Sparsity and Spectral Variability

Chapter 3

Local Intrinsic Dimensionality

Contents

3.1 Introduction

In the previous part, we have seen that performing the unmixing in local regions of the image can mitigate nonlinear and spectral variability (SV) effects. However, doing so requires the estimation of the number of endmembers to use in each local region, which can become problematic, and lead to overestimations of this number, in which case we will see sparsity can play a role to eliminate unwanted spectra from the unmixing. Before that, it is necessary to evaluate and quantify the possible estimation problems we are faced with in small local regions. This chapter is then concerned with the problem of estimating the intrinsic dimensionality (ID) of HSIs, and especially in small datasets, such as spatial regions of a larger image.

Usually, the dimensionality of hyperspectral vectors, L, is large, with hundreds or thousands of spectral bands. Assuming the measurement may be decomposed into signal, $\mathbf{y}$, and noise, $\mathbf{e}$, that is, $\mathbf{x} = \mathbf{y} + \mathbf{e}$, authors in [32] introduce the following definition:

Definition 1. *The ID of a dataset, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, is the dimension, d, of the vector subspace spanned by the signals, $\mathbf{y}_1, \ldots, \mathbf{y}_n$.*

Different authors have given alternative definitions of the intrinsic dimension or of similar terms. Chang and Du [36] define the "*virtual dimensionality*" as the the number of endmembers necessary to give accurate unmixing. Bajorski [10] defines the "*effective dimensionality*"

as the dimensionality of the affine subspace giving an acceptable approximation to all pixels. Def. 1 is equivalent to the ones provided in [18, 140]. Besides conceptual aspects, all of them are used in spectral unmixing to estimate the actual number of endmembers or the dimensionality of the subspace spanned by these endmembers. Hereafter, for sake of clarity, we will make use of the ID term only.

We have reviewed in Chapter 2 several local approaches for SU [29, 105, 145, 66, 56], designed to overcome some of the issues of global approaches, *i.e.* spectral variability [143, 172]. Furthermore the local spectral unmixing (LSU) approach has proved to be a useful framework to propose new unmixing-based segmentation techniques [161] or to improve unmixing-based hyperspectral super-resolution techniques using the local low rank property of hyperspectral data [160, 104]. In addition to the latter works, we envision to incorporate the local spectral unmixing to other hyperspectral applications such as unmixing-based anomaly/target detection, spectral-spatial classification or visualization, among others. Thus, there is an increasing need to better understand the role of ID in local neighborhoods of hyperspectral data, *i.e.* in patches or segmentation regions. These results were originally published in [54].

3.2 Related work

There exist many ID methods for hyperspectral data in the literature, as well as more general techniques in the signal processing community [25]. Nevertheless, the specificity of hyperspectral data reside in two aspects: the 2D spatial arrangement of the signal and the high dimensionality induced by the numerous spectral bands. Most of the existing ID estimation algorithms are based on the eigen-decomposition of some data dependent statistical matrix, often second order statistics. The basic idea is that if some noiseless signals $\mathbf{y}$ span a d-dimensional vector space, then their covariance matrix $\mathbf{K_y}$ should have a rank which is equal to d. Then this covariance matrix should only have d nonzero eigenvalues. The main issue with this strategy is that noisy signals have more nonzero eigenvalues than their ID value, and the problem boils down to being able to sort the eigenvalues related to signal and the ones related to noise in the following eigenvalue decomposition:

$$\mathbf{K_x} = \mathbf{P}^\top \mathbf{D} \mathbf{P}, \tag{3.1}$$

where $\mathbf{P}$ is a change of basis matrix, and $\mathbf{D}$ is a diagonal matrix containing the eigenvalues of $\mathbf{K_x}$ on its diagonal. A simple baseline approach is to define the ID as the number of the largest eigenvalues that must be retained to represent a percentage of the total variance of the data [64], *i.e.*, 95% or 99%. Chang and Du [36] proposed the widely used *Harsanyi-Farrand-Chang (HFC)* method, based on the comparison of the eigenvalues obtained from the covariance and the correlation matrices. The validity of the HFC method has been questioned in [10, 11], and Bajorski proposed an alternative algorithm, called *Second Moment Linear dimensionality (SML),* based on similar concepts. Another popular algorithm to perform hyperspectral ID estimation is the *Hyperspectral Subspace Identification by Minimum Error (HySIME)* [19], which is an evolution of the *Signal Subspace Estimation (SSE)* algorithm presented in [18]. The HySIME algorithm works by identifying the signal subspace achieving

a residual error comparable to the estimated noise power. A different approach has been proposed in [32], where new results in *Random Matrix Theory (RMT)* are used to determine which eigenvalues are due to noise and which are due to signal have been adapted for the identification of the hyperspectral ID. The *Outlier Detection Method (ODM)* [9] is another eigen-based algorithm, although ODM focuses on modeling the noise and treats the signal as outliers to the noise distribution.

Three non eigen-based hyperspectral ID estimators have recently been proposed. The first one, introduced in [111] as part of a *Negative ABundance-Oriented (NABO)* unmixing algorithm, borrows the main idea from the HySIME algorithm. Basically, it decomposes the residual error from the unconstrained unmixing into two components, a first due to noise and a second due to ID. The algorithm works by starting from an underestimate of the ID, and then, iteratively increments the ID value until the unmixing error can be solely explained by the noise term. The second non eigen-based method, called *Hyperspectral Image Dimension Estimation through Nearest Neighbor distance ratios (HIDENN)* [82] is based on local geometrical properties of the data manifold. The technique is aimed at computing the correlation dimension of the dataset, which is itself closely related to the concept of fractal dimension. The basic idea is to count (in the neigborhood of one data point) the total number of pairs of points $g(\epsilon)$ which have a distance between them that is less than ϵ. Then it can be shown that if $n \to \infty$ and $\epsilon \to 0$, the so-called correlation integral $C(\epsilon)$ has the following asymptotic behavior:

$$C(\epsilon) \triangleq \frac{g(\epsilon)}{n^2} \underset{\epsilon \to 0}{\sim} \epsilon^{d-1}, \tag{3.2}$$

where $d-1$ is here the dimension of the manifold (and d is the ID of the data). This behavior can be intuitively understood by the fact that in higher dimensions, there are more possible ways for one point to reach neighboring points. One can then recover the ID by computing:

$$d - 1 = \lim_{\epsilon \to 0} \frac{\ln(C(\epsilon))}{\ln(\epsilon)}. \tag{3.3}$$

Note that since the ID is here estimated in each point of the data cloud, in the signal processing literature this category of ID estimation technique can be referred to as local ID estimation [31, 25]. However, the concept differs from the one we are interested in since we consider spatially local ID estimation.

In [99], Kyubeda *et al.* proposed the *Maximum Orthogonal Complement Algorithm (MOCA)*, which solves an optimiziation problem exploiting the sensitivity of the $\mathcal{L}_{2,\infty}$ norm to rare materials, so the signal subspace preserves them. In [2], Acito *et al.* proposed a version of MOCA, called *Robust Signal Subspace Estimator (RSSE)*, that improves the latter in terms of computational speed and lighter parametrization. The same authors summarized in [3] both approaches, MOCA and RSSE, using a common theoretical framework, and also proposed a more computationally efficient version of the MOCA algorithm named *Modified MOCA (MMOCA)*. They also derived from the RSSE algorithm a method to account for signal dependent noise [4]. Chang *et al.* [37] proposed a Neyman-Pearson detector version of MOCA linking the ideas behind MOCA with those of the HFC algorithm. Recently, Chang *et al.* [38] have proposed an extension of the latter work based on high-order statistics.

3.3 Contributions

In [33], Cawse-Nicholson *et al.* studied the effect of correlated noise on ID estimation, and Hasanlou and Samadzadegan performed in [74] a comparative study of some ID estimation algorithms for classification. A recent survey of ID estimation algorithms compares five methods, three of which are also considered in this study, mostly in terms of ID estimation performance on the whole image, and in terms of the impact of the noise correlation and estimation [132]. Here, we are interested in the performance of hyperspectral ID estimation algorithms when going from global to local studies, that is, the capacity of the algorithms to correctly estimate the ID on small regions or subsamples of a hyperspectral image. In addition, the present study includes several algorithms not considered in [132].

Hyperspectral ID estimation algorithms can be grouped according to two main characteristics: i) whether they are based on eigen-decomposition or not, and ii) the requirement of a denoising step or of a noise power estimation. When trying to identify the ID of local (often small) regions in hyperspectral images, eigenvalue-based methods can be severely affected by the so-called curse of dimensionality [100] and the high between-band correlation. The curse of dimensionality refers to: i) the empty space phenomenon in high dimensions, which makes it necessary to use more and more data samples for estimation purposes when the dimension becomes higher, and to ii) the fact that high-dimensional data often show multicollinearity, which can hamper noise estimation regression. The effects of the local denoising and the local estimation of the noise power can also influence ID estimation. Usually, small regions present a relatively high spectral homogeneity, in the sense that the materials in the different pixels of small regions are likely to be the same, with slowly varying abundance coefficients. Then, noise can be sometimes misinterpreted as a signal, compromising the local denoising and noise power estimation.

We describe and compare nine ID estimation algorithms when going from global to local studies of hyperspectral data. We catalog the ID algorithms according to their base methodologies and we highlight their main drawbacks when working on local, often small, subsets of data. We also provide some guidelines for a better use of these algorithms in local studies which can be summarized as: (i) perform a global denoising or estimation of the noise power, that is, avoid the use of local denoising or local noise power estimation; (ii) subsets below a size threshold produce unreliable estimations, usually presenting an overestimation peak and an increase in the error variance.

3.4 State of the art of hyperspectral ID estimation

In this section, some methods for the estimation of the ID of a hyperspectral image are listed and presented. These methods are the ones used for the experiments in Sections 3.5.1 and 3.5.2. Several algorithms in the following require a noise estimation step before computing the ID. The algorithm used in this paper to perform this noise estimation (originally suggested in [133]) is presented before the ID estimation algorithms themselves. In [65], sev-

eral algorithms for noise estimation for hyperspectral images, based on linear regression are compared. The noise estimation method suggested in [133] was shown to be relatively robust in the simulations of that study. It is also the most widely used in the community. Anyway, by running similar experiments as the ones described below with known noise values (or equivalently a perfect noise estimation), we obtained comparable results to those obtained by estimating the noise globally on the whole image. This shows that the noise estimation provided by this method seems suited for local ID estimation. Next, we describe all the algorithms compared in this study. Some of the properties of those are listed in Table 3.1.

3.4.1 Noise estimation

The noise estimation algorithm used in the experiments is based on the use of the high correlation between adjacent bands and was first brought to the hyperspectral imaging community in [133]. The idea behind this strategy is to perform a linear regression of each spectral band on all the other bands, that is to say to express all the pixels from one spectral band (stacked into a $n \times 1$ vector) as a linear combination of the pixel vectors of all the other bands. If we denote by $\mathbf{X}_{\neq k}$ the data matrix $\mathbf{X}$ with the k^{th} row $\mathbf{x}_k$ (one entire band) removed, we can estimate the optimal regression parameters $\mathbf{b}_k \in \mathbb{R}^{L-1}$ of $\mathbf{x}_k$ on $\mathbf{X}_{\neq k}$ in a least square sense by:

$$\mathbf{b}_k = \mathbf{x}_k \mathbf{X}_{\neq k}^\top (\mathbf{X}_{\neq k} \mathbf{X}_{\neq k}^\top)^{-1}, \tag{3.4}$$

and we can finally estimate the noise vector $\mathbf{e}_k \in \mathbb{R}^n$ in band k by:

$$\mathbf{e}_k = \mathbf{x}_k - \mathbf{b}_k \mathbf{X}_{\neq k}. \tag{3.5}$$

This difference between the observations in the considered spectral band and the result of the regression is assumed to be due to noise, providing the estimated noise values and allowing the estimation of the noise sample correlation matrix, which is assumed to be diagonal (and hence does not consider spectrally correlated noise) with difference variances in each spectral band.

Other methods exist to perform hyperspectral noise estimation, such as the so-called shift difference method [69] for instance, which assumes that the differences between adjacent pixels are mainly due to different realizations of i.i.d. noise, the signal component being practically the same. Two other noise estimation strategies [113, 134] which have been used in hyperspectral data analysis, have been evaluated and discussed (especially for their behavior in case of correlated noise) in [33].

3.4.2 Review of some ID estimation algorithms

3.4.2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an extremely popular technique for data analysis [93], which has been used extensively for dimension reduction, among other applications. The idea

is to perform a Singular Value Decomposition (SVD) on the sample covariance matrix of a given dataset. The resulting eigenvectors are then sorted by decreasing order of eigenvalues. The subspace spanned by the first d eigenvectors is the d-dimensional space whose explained variance percentage is the highest. This means that when the data cloud is projected onto this d-dimensional subspace, the relative difference between the variance of the data cloud and its projection is the lowest possible. To estimate the dimensionality of a dataset, one has to select a threshold on the percentage of the explained variance. However, for some applications, including hyperspectral imaging, the manual choice of a threshold is not an easy task, since explained variance is not directly linked to the number of sources, and also because variance can be very well explained in a very low-dimensional subspace while the intrinsic dimension of the data manifold might be higher. In the experiments, we selected a threshold of 95% of the explained variance to determine the ID values.

3.4.2.2 Harsanyi, Farrand, and Chang (HFC)

This dimensionality estimation method, termed HFC (for Harsanyi, Farrand, and Chang) is another rather simple and widely adopted technique to compute the ID of a hyperspectral dataset. The sample correlation and covariance matrices ($\mathbf{R_x}$ and $\mathbf{K_x}$, respectively) of the observations are both computed, and their eigenvalues are sorted in decreasing order. HFC assumes that the sources are deterministic and nonnegative, and that the noise is spectrally white (i.e. uncorrelated with constant variance) with zero mean. In this case, if the ID is d, then the d largest eigenvalues of $\mathbf{R_x}$ are supposed to be larger than those of $\mathbf{K_x}$ because in the corresponding components (coming from the transformation by the eigenvalue decomposition) an endmember contributes to the correlation eigenvalues in addition to the noise. Based on this, the algorithm performs a hypothesis test on each eigenvalue set to determine if the eigenvalues of the covariance and correlation matrices are statistically significantly different or not. Note that the algorithm's results depend on a user-tuned false alarm probability, set to $\alpha = 10^{-5}$ in the experiments. Every time the test fails in a component, the ID value is incremented. The ID finally corresponds to the number of times this test has failed. An alternative version of the algorithm, called Noise Whitened HFC (NWHFC), assumes that the noise is uncorrelated but with possibly non-constant variance. It includes a noise-whitening step before using the same methodology as HFC.

Bajorski has argued in [10, 11] that the HFC method can only measure the dependence of the difference between consecutive eigenvalues of the covariance to the average values of the bands, which is unrelated to the ID value. Therefore, the HFC method may be conceptually wrong. However, the method provides consistent results because the differences between consecutive covariance eigenvalues is in itself a useful indicator of the ID of the dataset, while relating this difference to eigenvalues of the correlation matrix is not relevant [11].

3.4.2.3 Hyperspectral Subspace Identification by Minimum Error (HySIME)

Another popular algorithm to perform hyperspectral intrinsic dimensionality estimation is Hyperspectral Subspace Identification by Minimum Error (HySIME) [19]. For this algorithm, the noise $\mathbf{e}$ is assumed to be zero-mean Gaussian distributed; the noise value $\mathbf{e}$ and the noise correlation matrix $\mathbf{R_n}$ are estimated using the band correlation method described in Section 3.4.1. The sample observation correlation matrix $\mathbf{R_x}$ is computed, as well as the signal sample correlation matrix $\mathbf{R}_{\hat{\mathbf{y}}}$, taking the signal values $\hat{\mathbf{y}}$ by subtracting the estimated noise values $\hat{\mathbf{e}}$ from the observations $\mathbf{x}$. The eigenvectors of the latter matrix are computed and sorted in descending order according to the corresponding eigenvalues. The subspace spanned by the first d eigenvectors corresponding to the d largest eigenvalues is the signal subspace, whereas the orthogonal complement is associated with the noise subspace. The separation between the two is found by looking for the value of d which minimizes the Root Mean Squared Error (RMSE) between the signal and the projection of the observations on the subspace spanned by the first d eigenvectors, taking into account the projection error power (decreasing function of d) as well as the noise power (increasing with d). Note that in this case, using the correlation matrix is meaningful because its d first eigenvectors define the subspace minimizing the RMSE between the projected data and the original data.

3.4.2.4 Random Matrix Theory (RMT)

This technique was recently introduced in [32] and makes use of the tools of Random Matrix Theory (RMT) to estimate the ID of a hyperspectral dataset. It requires a noise estimation step which, in [32], is performed by the method presented in [113]. The method extends an existing RMT-based method for dimensionality estimation [108] to the case of spectrally correlated Gaussian noise. The underlying mixing model is also assumed to be linear. The general idea is that, under the assumption that each column of the $L \times N$ noise image is distributed according to $\mathbf{e} \sim \mathcal{N}(0, \mathbf{\Phi})$, the random cross product matrix $\mathbf{e}\mathbf{e}^\top$ follows a Wishart distribution (which can be seen as a multivariate generalization of the χ^2 distribution) $\mathcal{W}_L(\mathbf{\Phi}, N)$, with L representing the degrees of freedom, and $\mathbf{\Phi}$ the $L \times L$ scale matrix. The probability density function of the largest eigenvalue of such matrices has been extensively studied in RMT. In the context of dimensionality estimation, a criterion has been found to test which is the largest sample covariance eigenvalue which is statistically consistent with the distribution of the largest eigenvalue of a Wishart matrix. In other words, this means that the eigenvalue of $\mathbf{K_x}$ found by this process is the largest noise eigenvalue, and that all the larger sample covariance eigenvalues are associated to a signal component. This criterion, originally derived for a number of samples $N \to \infty$ and a number of variables (bands in this application) $L \to \infty$, with their ratio constant: $\frac{L}{N} = c$ (usual conditions in RMT), has also shown to be reliable for large but finite N and L values (see [32] and references therein). The computation of the eigenvalues of interest to be tested against those of a Wishart matrix, as well as the testing criterion, differ in the general case if the uncorrelated noise assumption has been dropped, but the basic principle remains the same.

3.4.2.5 Outlier Detection Method (ODM)

The algorithm introduced in [9] estimates the ID of a hyperspectral image by focusing on the noise and treating the signal data points as outliers to the noise distribution. It comprises three steps: the first is a whitening step performed by a Maximum Noise Fraction (MNF) transform [69], in which the noise estimation is performed using once again the band-regression method. The noise is then whitened by an eigenvalue decomposition of the noise covariance matrix $\mathbf{K_e}$ and scaled so as to get equal variances in each band, thus defining a noise hypersphere in the spectral space, and a principal component analysis is performed on the transformed data to obtain the final transformed components. The final step is the ID estimation through outlier detection, using Inter-Quartile Range (IQR) to define a boundary between the noise and the "outliers". The Euclidean distances between the standard deviation of each transformed band and the standard deviation of the previous one are computed, and the ID is incremented every time the value is above the IQR threshold. It is a nonparametric technique, which does not make any assumption on the noise distribution (even though the band-regression based noise estimation algorithm used will provide optimal performance when the noise is Gaussian, because of the least squares step), and hence the final step is supposed to be robust to a small number of samples used for the estimation.

3.4.2.6 Vertex Component Analysis/Negative ABundance Oriented algorithm (VCA/NABO)

This technique [111] performs spectral unmixing and dimensionality estimation at the same time. It is noteworthy that this method is not eigenvalue-based. The idea is to start from an underestimation of the dimensionality of the dataset, and an estimation of the noise. Then an endmbember extraction (using any Endmember Extraction Algorithm (EEA)) is performed, and the abundances are computed through linear unconstrained least squares unmixing, dropping both the usual Abundance Sum-to-one Constraint (ASC) and the Abundance Nonnegativity Constraint (ANC). Then, the power of the Root Mean Square Error (RMSE) is compared to the estimated noise power. If the former is higher than the latter, the dimensionality is incremented until the error power becomes smaller than the estimated noise power. At this step, it should not be necessary to increase the dimensionality further since the potential gain in RMSE will not be meaningful, and so the number of endmembers has been found. It should be noted that the abundances are computed without using any constraints so that RMSE (in other words, the projection error) is not due to the projection of the data onto the feasible set of solutions but mainly to the fact that the subspace on which the data are projected has a too small dimension. In the experiments described below, the chosen EEA is Vertex Component Analysis (VCA) [122]. As this widely used EEA is stochastic by nature, the VCA/NABO algorithm is performed 20 times, and the final ID value is the (rounded) mean of the results of each iteration.

3.4.2.7 Hyperspectral Intrinsic Dimensionality Estimator through Nearest Neighbor distance ratios (HIDENN)

The dimensionality estimation method described here was presented in [82] and is called Hyperspectral Image Dimension Estimation through Nearest Neighbor distance ratios (HIDENN). As VCA/NABO (though the methods are completely different in nature), it differs from most of the other methods mentioned in this paper in the sense that it is not based on any eigenvalue decomposition whatsoever. The data is assumed to come from samples of a manifold (it does not require any particular mixing model, so long as the abundances are subject to the ANC and the ASC), whose dimension is equal to the number of endmembers in the image minus one. A particular case of this is the $(d-1)$-simplex defined by a linear mixture of d materials. The algorithm estimates the dimension of the manifold (locally isomorphic to $\mathbb{R}^{d-1}$) using geometrical properties and then provides the number of endmembers. In that case, the distance between each data sample and its l-nearest neighbor is computed for two well chosen values of l, and using a variant of Eq. (3.3), an estimator of the correlation dimension is built to estimate the ID at this location in the spectral space. The choice of these values is critical since they need to be small enough to reduce the influence of the noise, but also large enough to be statistically robust. The individual pixel values are then averaged to give the global ID of the dataset, requiring a sufficient number of samples for the estimation to be meaningful. As the estimation of the dimension of such a manifold in the spectral space is highly sensitive to noise, a denoising may be performed beforehand in order to allow a more robust estimation of the ID. The algorithm becomes D-HIDENN (for Denoised-HIDENN) and makes use once again of the band correlation noise estimation technique described in [133].

3.4.2.8 Modified Maximum Orthogonal Complement Algorithm (MMOCA)

This non eigenvalue-based ID estimation technique [3], MMOCA (for Modified Maximum Orthogonal Complement Analysis), is actually a combination of the NWHFC algorithm described above and the MOCA algorithm [99]. The former is used to provide an underestimation of the ID of the dataset, so that the latter can iterate on the ID values from this starting point. More precisely, for a given candidate ID value d, MOCA aims at finding a suboptimal solution to the following optimization problem:

$$\hat{\mathbf{M}} = \arg\min_{\tilde{\mathbf{M}}} ||\mathbf{P}^{\perp}_{\tilde{\mathbf{M}}}\tilde{\mathbf{X}}||_{2,\infty}, \tag{3.6}$$

where $\tilde{\mathbf{X}}$ is the whitened data matrix, and $\tilde{\mathbf{M}}$ is taken from the set of all possible bases of a d-dimensional subspace of $\mathbb{R}^L$. $\mathbf{P}^{\perp}_{\tilde{\mathbf{M}}}$ is the projection matrix on the orthogonal complement of the subspace spanned by $\mathbf{M}$, such that $\mathbf{P}^{\perp}_{\tilde{\mathbf{M}}}\tilde{\mathbf{X}}$ is the error of the projection of the whitened data on the signal subspace. The $\mathcal{L}_{2,\infty}$ norm is used for its sensitivity to rare materials, since a rare material not accounted for by the $\tilde{\mathbf{M}}$ matrix will result in a high error on the concerned pixels, even if they are very few. The stopping criterion for this iterative process is based on a hypothesis test using a Maximum A Posteriori (MAP) criterion. The idea is to determine whether $||\mathbf{P}^{\perp}_{\tilde{\mathbf{M}}}\tilde{\mathbf{X}}||_{2,\infty}$ depends only on the noise distribution or also on the residual signal.

Property \ Algorithm	HySIME	RMT	ODM	VCA/NABO	HIDENN	D-HIDENN	PCA	HFC	MMOCA
Eigenvalue based	✓	✓	✓				✓	✓	
Nearest neighbor distance ratios					✓	✓			
Subspace estimation	✓			✓					✓
Noise estimation step	✓	✓	✓	✓		✓			
Underlying Mixing Model	Free	LMM	Free	LMM	Free	Free	Free	Free	Free

Table 3.1: Properties of the algorithms used.

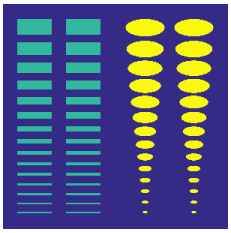


Figure 3.1: The spatial pattern used for the creation of the synthetic datasets.

3.5 Local Performance of the algorithms

3.5.1 Experiments on synthetic datasets

3.5.1.1 Datasets

The synthetic datasets built for this study were designed to evaluate how the previous algorithms behave from a local to a global scale, and to assess the effects of the SNR as well as the number of bands of the hyperspectral data in the ID estimation. A spatial pattern of 300×300 pixels comprising two kinds of aligned geometrical shapes (rectangles and ellipses) of various sizes was synthesized (see Fig. 3.1). Different variants of the dataset were created with different numbers of bands (480, 240, 120, 60 and 30 bands) and a spectrally and spatially white Gaussian noise was added so as to reach different values of SNR (20, 25, 30, 35, 40 dB), yielding a total of 25 synthetic images.

From the spatial pattern of Fig 3.1, three distinct mixtures were created: two mixtures of five endmembers and one of three endmembers. A mixture of three endmembers was employed to define the background, while two other mixtures of five endmembers were situated in the rectangles and the ellipses, respectively. The endmembers were randomly chosen from a mineral sublibrary of the United States Geological Survey (USGS) spectral library[1], with the constraint that the Spectral Angle Mapper between two signatures should not be less than 10 degrees or more than 30 degrees. This library contains the spectral signatures of various minerals acquired on the ground with a field spectrometer. The original endmembers were downsampled by a factor of 2, 4, 8 and 16 to provide datasets with the selected range of spectral bands. Note that some of the endmembers can be common to the different mix-

[1] http://speclab.cr.usgs.gov/spectral-lib.html

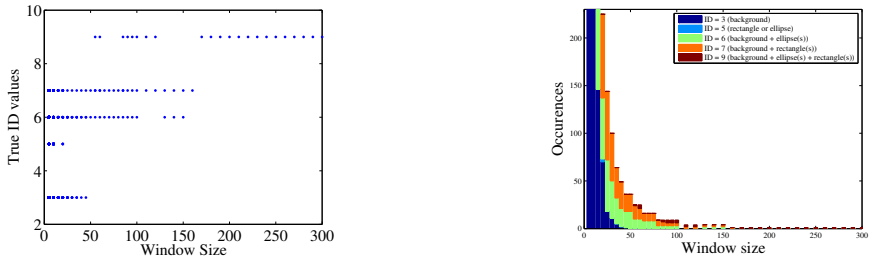
tures. In the end, there are 9 distinct endmembers in the image: 2 endmembers are common between the background and the ellipses, another is common between the background and the rectangles, and a last one is common between the ellipses and the rectangles. Thus we can deduce that there are 4 endmembers which are repeated among the different patterns in the image, leading to a total of 9 distinct endmembers. The abundances of each pixel are sampled from a uniform distribution over the probability simplex of the corresponding dimension (depending on the considered mixture), so that the ASC and ANC are enforced. The mixed pixels are finally generated using the LMM.

Furthermore, since we want to focus on the capability of the different algorithms for local ID estimation, we only consider Roger's method [133] for noise estimation. We tested the impact of this choice by comparing the results of the local ID estimation of section 3.5.1.3 using this noise estimation strategy on the whole image, to the use of the actual noise values. The results are similar in both cases, which shows that Roger's noise estimation strategy has little impact on the results, at least when the noise is estimated globally. Here, we have considered a spectrally and spatially white noise. However as shown in [132], coloration of the noise (different variances in each bands, but still a diagonal noise covariance matrix) and correlation between bands for the noise can be significant in real scenarios. We have performed experiments on synthetic datasets accounting for these two properties of the noise, in order to see the impact of non white noise on local ID estimation. However, the conclusions are very similar to those of the experiments with white noise. Hence, these results are not shown here but gathered Appendix C.

3.5.1.2 Experimental setup

Here we present the experimental methodology we followed to assess how the different algorithms behave in local ID estimation. Each of the 25 synthetic datasets was divided into non-overlapping square tiles of various sizes, ranging from 5×5 to 100×100 pixels with steps of 5×5 pixels, and from 100×100 to 300×300 pixels with steps of 10×10 pixels. Therefore, we can study the performance of ID estimation algorithms from a very small local subset (25 pixels) to a global scenario (90000 pixels). The actual ID of each tile depends on which region of the image it falls into (see Fig. 3.1). The possible actual ID values plotted against the tile length size are shown in Fig. 3.2 (a): 5 if the tile falls into a rectangle or an ellipse only, 3 if the tile falls into the background only, 6 if the tile falls into the background and one or multiple ellipses, 7 if the tile falls into the background and one or multiple rectangles, and 9 if the tile falls into the background, one or multiple ellipses and one or multiple rectangles. A summary of these considerations is presented in Fig. 3.2 (b), in which a stacked histogram of the tiles is shown. The first two tile sizes (the bars corresponding to 5×5 and 10×10 pixels are truncated for the sake of visibility, since 3600 and 900 tiles of this size can be fitted into the image, respectively).

For all the 25 configurations of SNR and number of bands, and for all tile sizes, each ID estimation algorithm is independently run on each tile. Since the noise is here spectrally white, we used the HFC algorithm rather than its noise whitened counterpart, which has minimal



(a) (b)

Figure 3.2: True ID values plotted against size of the subset (a). Stacked Histogram of the tiles for each size, depending on the ID value (b).

impact on the results. The ID estimation is performed in two different cases, depending on the way the noise is estimated: locally or globally. The *local noise estimation* makes use of the pixel values of the local subset only, while the *global noise estimation* makes use of the whole image. In both cases, we employed a fast implementation of Roger's method [133], due to [19], and presented in section 3.4.1.

Next, we describe the quality metrics defined to evaluate the performance of each algorithm. Given the set $\mathcal{S} = \{5, 10, 15, \cdots, 100, 110, 120, \cdots, 300\}$ of window sizes, let $s = |\mathcal{S}|$ be the number of possible lengths. N_i denotes the number of windows of size $\mathcal{S}_i$, $1 \leq i \leq s$. Let d_{ij} and $\hat{d}_{ij}$ respectively denote the actual and estimated ID values of the j^{th} window of size $\mathcal{S}_i$. We define μ_i as the average of the relative absolute errors committed on all windows of size $\mathcal{S}_i$:

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{|d_{ij} - \hat{d}_{ij}|}{d_{ij}}. \tag{3.7}$$

We also define μ as the average of all the μ_i values for all possible window lengths. This provides a single number to assess the overall performance of the algorithms from the most local (*i.e.* smallest window size) to global ID estimation:

$$\mu = \frac{1}{s} \sum_{i=1}^{s} \mu_i. \tag{3.8}$$

Finally, σ_i^2 is an estimator of the variance of the absolute relative error committed on all tiles of size $\mathcal{S}_i$:

$$\sigma_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} \left(\frac{|d_{ij} - \hat{d}_{ij}|}{d_{ij}} - \mu_i \right)^2. \tag{3.9}$$

3.5.1.3 Results

The results of the ID estimations on the 25 synthetic datasets are presented for all algorithms in Figs. 3.3 to 3.8. In Fig. 3.3 (a), the value of μ (see Eq. (3.8)) is displayed as an image,

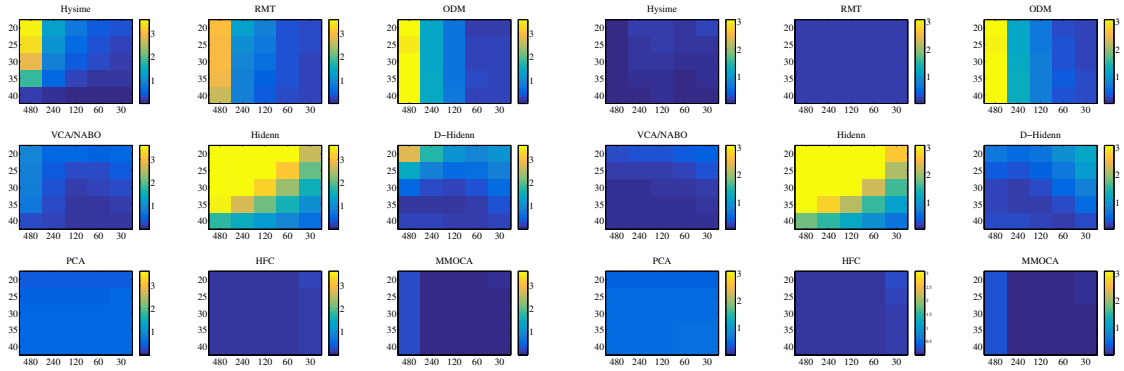


Figure 3.3: μ_{local} (a) and μ_{global} (b) and as a function of SNR in dB (y axis) and number of bands (x axis). The color scale ranges from blue (0.0002) to yellow (3.6 (a), 3.1 (b) or higher).

for all noise powers and numbers of bands, in the case of a local noise estimation. From this figure, we see that the algorithms of the bottom row (PCA, HFC and MMOCA) are nearly insensitive to the number of bands or the noise power. This is because these algorithms do not require any noise estimation. The results of PCA are highly dependent on the chosen threshold for the explained variance, which is not directly related to the ID value. MMOCA and HFC seem to perform relatively well in all cases. The case of HIDENN is different since estimating the dimension of a manifold is an operation which is highly sensitive to noise, and also dependent of the dimension of the ambient space. We can see that if any of the two tested parameters here are tuned to a more favorable value (higher SNR or lower number of bands), the overall results get better, while in unfavourable configurations, outliers in the estimated values severely decrease the performance. The denoised version of the algorithm, D-HIDENN, helps to reduce the impact of this phenomenon, although it is still present. This algorithm is still sensitive to the noise power, because the noise is not only estimated through its covariance matrix, but also subtracted from the observations. The last four algorithms, Hysime, RMT, ODM, and VCA/NABO present a more similar behavior. They all require a noise estimation, whose performance greatly impacts the ID estimation. We can notice immediately that the ID estimation for these algorithms is much more sensitive to the number of bands than to the noise power, which can be explained by the fact that the algorithm used for the noise estimation is based on a regression of each band on the others, an operation becoming less precise when the number of bands increases. This is due to the multicollinearity effect: when there are more bands, they are more correlated since adjacent wavelengths become closer and closer, and there are multiple good candidates for the regression coefficients. Hence a small change in the data can induce a large change in the regression coefficients (see section 3.4.1). Fig. 3.3 (b) shows the same metric μ in the case of a global noise estimation. For MMOCA, HFC, PCA and HIDENN, the results are very similar to the ones obtained for the local noise estimation since these algorithms do not estimate the noise (they are not exactly equivalent since for both experiments a different noise realization was used). However, for the other algorithms, notable differences are visible: the

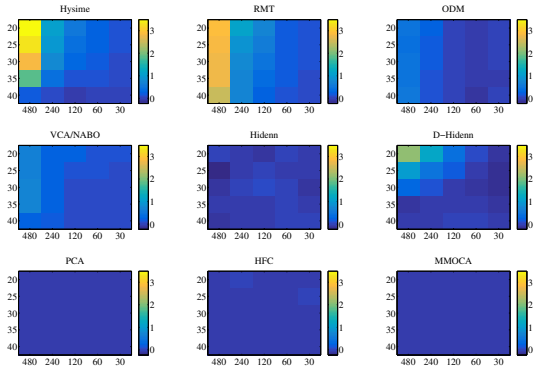


Figure 3.4: $\mu_{\text{local}} - \mu_{\text{global}}$ as a function of SNR in dB (y axis) and number of bands (x axis). The color scale ranges from blue (-0.48) to yellow (3.49 or higher).

algorithms perform much better in the least favorable cases. As we will see in the following, this is due to the fact that global noise estimation allows a much better ID estimation in small windows (provided the noise distribution is the same everywhere), where a precise local noise estimation is impossible because of the too low number of samples. The ODM algorithm does not seem very affected by the change in the noise estimation. This probably comes from the paradigm used in this algorithm: the objective of ODM is to identify the signal as an outlier in a noise distribution.

Fig. 3.4 sums up these considerations by showing the difference between the μ values estimated using local and global noise estimations, $\mu_{\text{local}} - \mu_{\text{global}}$. Thus, a positive value means that local noise estimation performed worse than global noise estimation, and vice versa. From the figure, it is clear than in almost all cases, global noise estimation performs better for algorithms sensitive to the way the noise is estimated. We see that when the configuration becomes more favorable, the results of local noise estimation become closer to the ones with global noise estimation. It happens in some cases that estimating the noise locally performs slightly better than doing it globally, but in most cases the results show that global noise estimation is much more robust.

In Fig. 3.5 we show in detail the results of the ID estimation for all algorithms and all window lengths in one representative noise and band number configuration, respectively 30 dB and 120 bands, corresponding to the central pixels of the images of Fig. 3.3. This configuration was chosen because it is representative of many real scenarios. These figures are to be compared to the actual ID values in Fig. 3.2. Two patterns in the ID estimations can be found for most algorithms: (i) a window size range where the ID estimate has a peak, which is too large, and (ii) a set of window sizes for which there is a slow stabilization of the results, until the support of the global image is reached. Fig. 3.5 shows, as expected, that for local noise estimation and for most algorithms, the ID estimation provides erroneous values for the smallest windows. For HySIME, RMT and VCA/NABO we can observe an important peak in the estimated ID values for a certain window size. This peak means that nearly all the values between zero and the maximum of the peaks were attained for the different windows of this size, confirming the instability of the algorithms, and more specifically of

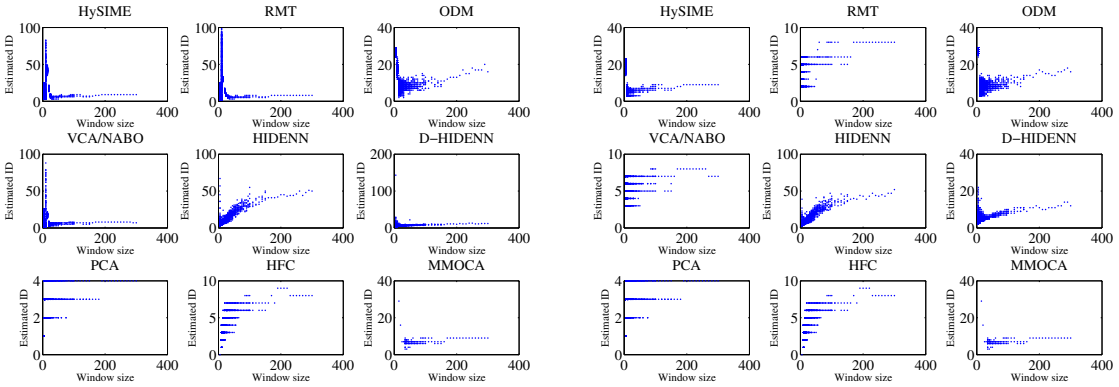


Figure 3.5: Estimated ID in the case of local (a) and global (b) noise estimation plotted against window size for all algorithms, for SNR = 30dB and 120 bands.

the noise estimation for small windows. The height and position of the peak depends on the noise and band configuration, as we will discuss in the following. The peak is also present, to a lesser extent for HIDENN/D-HIDDEN algorithms because they estimate the dimension of a manifold with too few samples, in which case noise is mistaken for signal, especially for small regions which are likely to have a low rank. ODM also shows this peak because of the noise estimation, although its importance is mitigated by the outlier in noise paradigm. For MMOCA, the peak has another origin since the low dimensional subspace is estimated by resorting to an optimization problem. In this case, for too small windows, this problem is very ill-conditioned, which entails erroneous estimations. Below a certain size, singular matrices appear during the estimation and the algorithm fails to produce an estimated value. Finally, PCA seems affected inasmuch as the (overall small) variance seems harder to capture with only a few dimensions. Finally, HFC seems to be less affected by the number of pixels in the local regions, since the estimation does not show a peak in the ID values but more a linear increase with the window size. Fig. 3.5 (b), the same plots are presented, but in this case for global noise estimation. As before, HIDDEN, PCA, HFC and MMOCA are not affected since they do not require a noise estimation step. ODM does not seem very affected either, probably because of its particular signal and noise model. For HySIME the peak is also present, because while the noise correlation matrix estimation is much more precise, the signal correlation matrix still has to be estimated in a small dataset. However, the peak decreases faster and is less important in amplitude than in the local case. However, RMT and VCA/NABO, seem very affected by the change in noise estimation. The corresponding plots are now quite similar in shape to the actual ID values in Fig. 3.2 (VCA/NABO does not require the estimation of the signal covariance matrix). Finally, for D-HIDENN, global noise estimation allows the suppression of the most aberrant outliers from the estimated ID values. Overall, it seems that global noise estimation is very beneficial to ID estimation, but it relies on the assumption that the noise is spatially i.i.d. in all the image.

Another aspect of local ID estimation shown in Figs. 3.6 and 3.7, is the transition between erroneous ID estimations for small windows to correct ID estimations when the window sizes

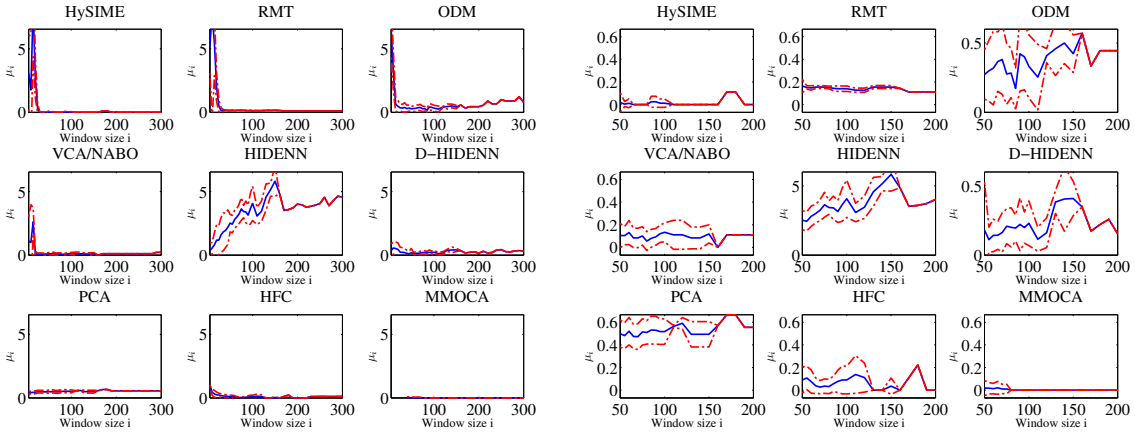


Figure 3.6: μ_i as a function of the window size $i \in S$ in the case of local noise estimation, for SNR = 30dB and 120 bands (a). The standard deviation of the estimated values is represented by the red curves. (b) is simply a zoomed version of the (a).

get sufficiently high, until the size of the whole image is reached. The quality metric μ_i (see Eq. (3.7)) is plotted in blue against the window length $\mathcal{S}_i$, while the dashed red curves correspond to the quality metric plus and minus one standard deviation (the standard deviation is defined as the square root of Eq. (3.9)) are shown in dashed red. Fig. 3.6 shows the value of μ_i for the local noise estimation. The righmost figure is a zoomed version to show what happens after the peak in the estimations. From this figure, we clearly see that for the algorithms concerned, the peak is accompanied by a large variance in the estimations, which quickly decreases as the number of samples get higher. Note that for large window sizes, this phenomenon is also due to the fact that there are fewer windows of this size that we can fit into the image. For global noise estimation (Figs. 3.7), we see that apart from HySIME and ODM, the estimations in small windows are less subject to a high variance, and the estimation for each window size in small windows is much more precise, which confirms the results of the previous figures. The observations drawn from these figures allow one to define empirically a size threshold above which the noise estimation will be reliable.

Finally, Figs. 3.8 depict a last but nonetheless important aspect of the noise estimation: for which window size does the peak appear? We discuss this particular point, very linked to the definition of a reliability threshold for the estimation, considering this time several band number configurations at fixed SNR, and vice versa, but only for the algorithms concerned (*i.e.* HySIME, RMT, ODM, VCA/NABO and MMOCA). For local noise estimation, we immediately see that the size at which the peak appears for all algorithms is much more related to the number of bands considered in the estimation than it is to the noise level, which more influences its height. The higher the number of bands, the later the peak appears, which means that larger windows will be necessary for a correct ID estimation. In the case of global noise estimation, many cases are favorable enough for the algorithms not to present a peak, since the noise is correctly estimated (except for MMOCA which does not require a

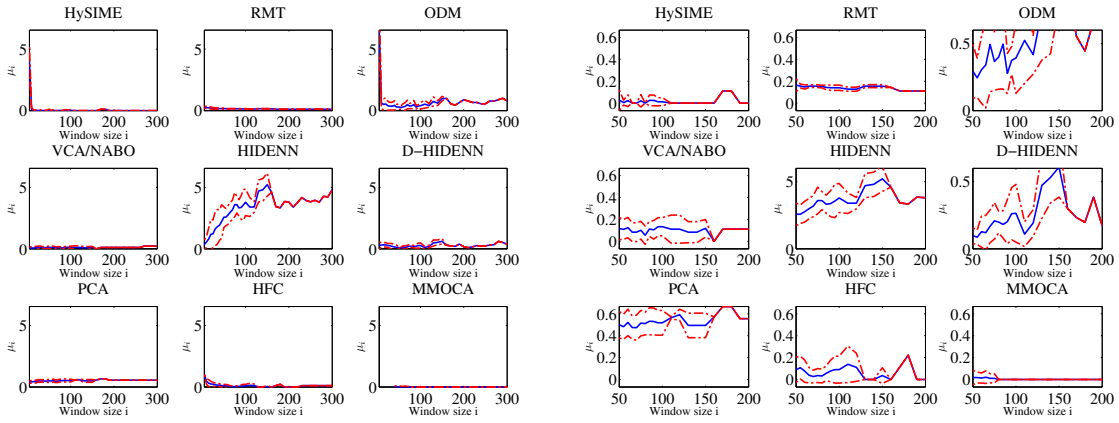


Figure 3.7: μ_i as a function of the window size $i \in S$ in the case of global noise estimation, for SNR = 30dB and 120 bands (a). The standard deviation of the estimated values is represented by the red curves. (b) is simply a zoomed version of (a).

noise estimation), and its position is less influenced by the number of bands.

3.5.2 Experiments on real datasets

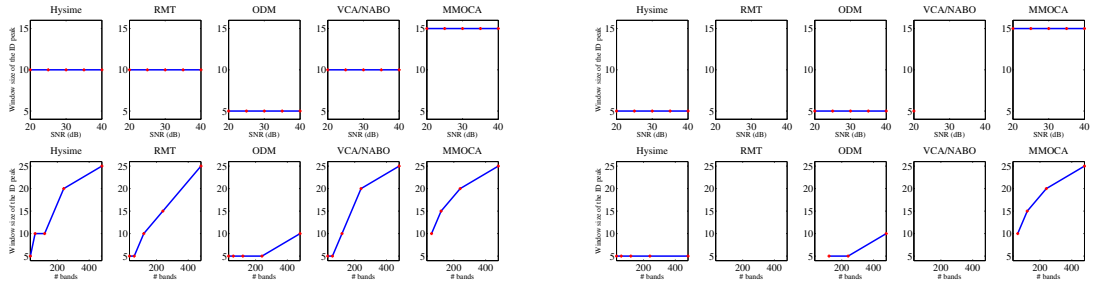
In this section we present the experiments we performed on two real datasets in order to validate the observations made on the synthetic datasets.

3.5.2.1 Datasets

The first dataset we used is an image acquired by NASA's AVIRIS sensor over the Cuprite mining district in Nevada, USA. It is a 350×350 image comprising 188 spectral bands, which has been often used to validate ID estimation algorithms. We estimated the SNR of each band of this image using the algorithm presented in section 3.4.1 and obtained an average SNR (over all bands) of 27dB. It is usually considered that there are at least 17 different materials (mostly minerals) in this image, based on ground observations and mineral maps of the site[2]. In addition, according to experiments performed in [33], the noise in this image in not very spectrally correlated. An RGB representation of this image is shown in Fig 3.9 (a), using bands 40, 30 and 20 of the image.

The second dataset was acquired by the CASI 1500 sensor over the region of Barrax, in the south of Spain, in 2005[3]. The 97×847 image comprises 144 bands in the VNIR region (370-1050 nm) and the estimated average SNR is 43dB. A RGB representation of this scene is also shown in Fig. 3.9 (b), using bands 52, 35 and 25.

[2] http://speclab.cr.usgs.gov/cuprite95.tgif.2.2um_map.gif.
[3] http://www.uv.es/~leo/sen2flex/



(a) (b)

Figure 3.8: Window size of the ID peak (corresponding to spurious estimation) plotted against SNR for 120 bands (top row) or against number of bands at fixed SNR = 30dB (bottom row), for local (a) and global (b) noise estimation. Blank values indicate that no peak is present in the corresponding configuration.

(a) (b)

Figure 3.9: RGB representation of the Cuprite (a) and Barrax (b) datasets.

3.5.2.2 Experimental setup

For both datasets, as for the synthetic data, we perform local ID estimation on non-overlapping square tiles of different sizes, from 5×5 to 100×100 pixels size with steps of 5×5 pixels, and from then on, from 100×100 pixels size to the maximum possible with steps of 10×10 pixels. For the Barrax dataset, we considered only the tiles in which no unobserved values were present. For both datasets, ID estimation was carried out for all algorithms, for local and global noise estimation. In the absence of ground truth, we cannot compute the metrics used for the synthetic datasets, but we can compare qualitatively the shapes of the local ID plots to the observations made for the synthetic data.

3.5.2.3 Results

First, we compared the results of the ID estimations for both datasets in the case of local noise estimation (see Figs. 3.10 (a) and 3.11 (a)). The results show that in both cases, the general behavior of the algorithms is similar to that of the synthetic datasets. We can see that HySIME, RMT, ODM and VCA/NABO show a clear peak in the ID estimations for small windows, which is clearer for the Cuprite dataset, probably because it is noisier than

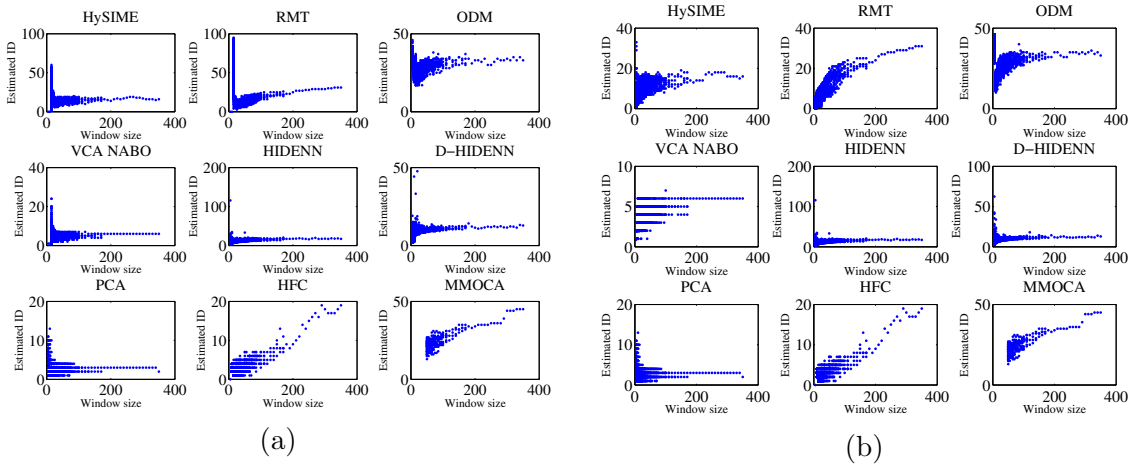


Figure 3.10: Estimated ID of the Cuprite dataset in the case of local (a) and global (b) noise estimation plotted against window size for all algorithms.

the Barrax data. The peaks appear roughly for the same window sizes as in the simulated data, which is logical since the number of bands is comparable in both datasets. Then the peaks quickly decrease and seem to stabilize around different values for each algorithm when we approach global ID estimation. Note that the zero values which can appear for very small windows and some algorithms are due to a very poor noise estimation. For example, in the case of HySIME, the objective function can be increasing with the dimensionality for poorly estimate noise values, hence the zero value for the ID. For HIDENN and D-HIDENN, the results are consistent with the synthetic data: large outliers appear for very small windows, and then, the algorithm quickly stabilizes with smaller outlier values if the data has been denoised beforehand. The performance of PCA still depends heavily on the arbitrary choice of the variance percentage (still 95% here). For a percentage lower than 95%, the estimated ID is rarely above 3 for the global images, showing that the ID is not linked to the variance of the data cloud. The performance improves for larger thresholds, but the tuning is empirical and data-dependent. HFC still obtains a more or less linear behavior with the increase in window size. Finally, the MMOCA algorithm fails to produce a value for a large range of window sizes because of the ill-conditioning of the subspace estimation problem (which explains why only windows bigger than 50×50 pixels appear for the Cuprite dataset and windows over 20×20 pixels for the Barrax dataset). For global noise estimation, and for the algorithms requiring noise estimation, the results are still consistent with the ones obtained on the synthetic data (on the Figs. 3.10 (b) and 3.11 (b)). The peak in the estimated values is still present for the HySIME and ODM algorithms with the Cuprite data, but very attenuated with respect to the case of local noise estimation. For RMT and VCA/NABO, as for the synthetic datasets in such noise and band configuration, the peak has vanished. We can see that when the windows get larger, both noise estimation strategies perform in an increasingly similar way, as expected. From the figures above, we can define an empirical threshold above which the ID estimation would be reliable: for instance, for the Cuprite dataset, we can set the window size threshold to 30×30 pixels for the case of local noise estimation, and a window size threshold of 15×15 pixels for global noise estimation, for all algorithms. Note that the algorithms can

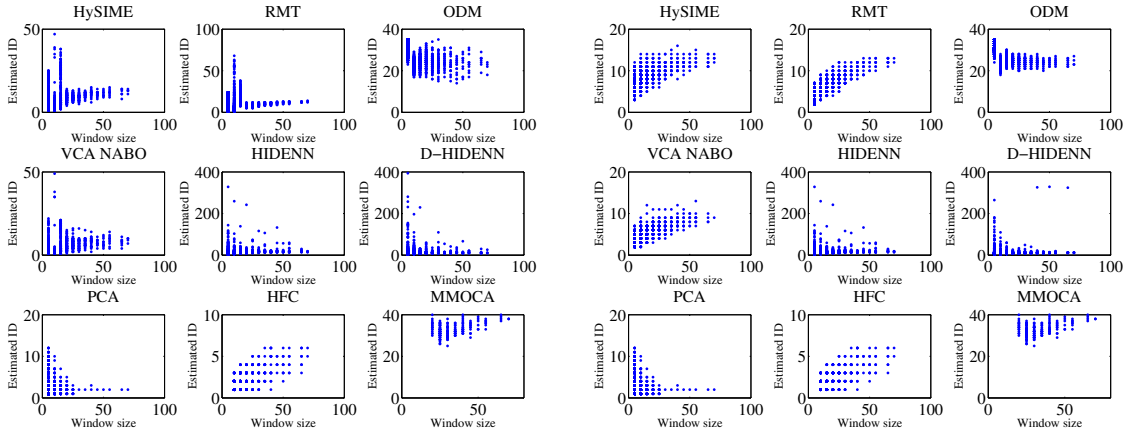


Figure 3.11: Estimated ID of the Barrax dataset in the case of local (a) and global (b) noise estimation plotted against window size for all algorithms. Some outliers for Hidenn and D-Hidenn are not displayed.

differ a lot in their estimated global ID values (in the Cuprite case, around 20 for HySIME, HFC and HIDENN, and less for VCA/NABO and PCA, and much higher for RMT, ODM and MMOCA). Notice that for both datasets, the estimated global noise ID values for large window sizes match the ones with the same windows, but obtained in a local noise estimation context. This tends to confirm that the spatial i.i.d. assumption for the noise holds in these datasets.

3.6 Discussion

In this section, we summarize the observations made for the synthetic and real datasets, and we provide some indications on how to use the ID estimation algorithms in a local setting. From the results, we observed that there are three main parameters influencing local ID estimation:

1. The number of pixels in the local region.
2. The number of spectral bands.
3. The noise level.

A majority of the tested algorithms require a noise estimation step. For these algorithms, a clear pattern can be seen when estimating ID in regions of the datasets at different scales, often comprising a peak in the overestimation of the ID in unfavorable cases. This pattern is especially present when the noise is estimated locally, in each tile of the image. When such a peak appears, its amplitude increases with the noise power, while its position is especially

determined by the number of bands in the image: the more bands, the larger the window size where the peak appears, which means that the ID estimation will be unreliable for larger windows than if there were fewer spectral bands. This phenomenon is linked to the curse of dimensionality: since higher dimensional spaces are sparser, more samples are required in order that estimation algorithms obtain reliable results. In addition, the multicollinearity phenomenon in high dimensions can also hamper noise estimation strategies exploiting the between band correlations.

These considerations raise the question of how to choose a minimum value for the window size, below which the ID estimation is unreliable. One has to take into account the position of the peak, but also the speed of the decrease after it. A threshold can be roughly defined visually from plots similar to those in Fig. 3.5. The most favorable configuration for local ID estimation is then a low number of bands and a good SNR. In any case, for those algorithms, a global noise estimation is preferable, since it largely reduces the uncertainty due to the noise estimation. The only case when a local noise estimation is preferable is in the case of a spatially non-i.i.d. noise. MMOCA does not require a noise estimation, but fails to produce a result when the underlying optimization problem is too ill-conditioned. For the other algorithms, HIDDEN and D-HIDDEN have a tendency to produce large outliers when the number of samples is too few. HFC seems to behave more naturally for small windows, since small estimated ID values come out in this case.

Next, we need to determine which algorithm to choose to estimate the ID locally. To guide the reader in his choice, we summarize below and in Table 3.2 the strengths and weaknesses of each tested algorithm:

- HySIME: relatively robust for local ID estimation, provided the noise is estimated globally, but still subject to overestimation when the window size is too small because it requires the estimation of the signal correlation matrix. It is also relatively fast and produced good results on synthetic datasets.
- RMT: comparable to HySIME, with good performance on the synthetic datasets. It does not show a peak in the ID values when the noise is estimated globally (at least for reasonable band number and noise configurations). Relatively fast.
- ODM: relatively fast, but less precise and more sensitive to the number of bands than the previous two algorithms. Less sensitive to local/global noise estimation.
- VCA/NABO: same advantages as the previous ones, which fall in the same category (although NABO is not eigenvalue-based), but quite computationally intensive since it requires a spectral unmixing step. Slightly more sensitive to noise than most algorithms.
- HIDENN / D-HIDENN: not eigenvalue based, but very sensitive to noise, even though its effect can be mitigated but not suppressed when a de-noising step is performed. Poor precision in low SNR cases. Relatively slow.
- HFC: Practically insensitive to noise and band number. Provides underestimated ID values independently of the scale, although they are overall relatively accurate. Depends

on a user-defined threshold. It can be argued that it is theoretically wrong and that the results depend on the average values of the bands and not directly on the ID of the data. Fast.

- PCA: definitely not a good candidate: the performance is conditioned to the arbitrary choice of the threshold.
- MMOCA: does not require a noise estimation, good performance. Computationally rather intensive, especially for small windows. Does not work for too small windows because of ill conditioning.

Property \ Algorithm	HySIME	RMT	ODM	HIDENN	D-HIDENN	VCA/NABO	PCA	HFC	MMOCA
Noise level sensitivity	+	+	+	++	++	+	-	- -	- -
Sample size sensitivity	++	+	+	++	++	+	+	-	+
Band number sensitivity	+	+	++	-	-	+	- -	- -	-
Computational burden	-	- -	- -	+	+	++	- -	- -	+
Overall estimation error on synthetic datasets	-	- -	+	++	+	-	+	-	- -

Table 3.2: Strengths and weaknesses of the tested algorithms for local ID estimation. One or two + signs means high or very high, and one or two - signs means low or very low.

3.7 Partial Conclusion

In this chapter, we presented a study of several Intrinsic Dimensionality estimation algorithms for hyperspectral imaging in the context of local ID estimation. The results on both synthetic and real data show that in general, when trying to use these algorithms on local subsets of a large image, one has to be careful with: (i) the number of samples in the subsets, which have to be sufficiently numerous for estimation processes to be reliable; and, (ii) when noise estimation or denoising is required, a local approach will yield a decrease in performance, although this problem can be highly mitigated by estimating the noise on the whole image. Two other important factors also have consequences on the results: the noise level and the number of spectral bands. A low SNR and a high number of bands will increase the chance of mistaking noise for signal and make the estimation more prone to fail in higher dimensional settings, respectively. We summed up the properties of nine ID estimation algorithms and showed how they behaved in local areas of the image, and evidenced their respective strengths and weaknesses for local ID estimation. Future work will include considerations developed in this chapter in the pipelines of algorithms designed for other applications on hyperspectral imaging which resort to local subsets of the image, especially for local spectral unmixing (LSU).

Chapter 4

Spectral bundles and Local Spectral Unmixing

Contents

4.1 Introduction

This Chapter is concerned with incorporating sparsity inducing penalties in the spectral unmixing (SU) problem in order to deal with the spectral variability (SV) issue. Here, we present two ways of doing this in two different contexts: Local Spectral Unmixing (LSU) and spectral bundles.

In the previous Chapter, we have seen that local intrinsic dimensionality (ID) estimation is an important step of LSU, and has to be carefully carried out in order to limit potential overestimations of the number of endmembers to use in each subset. However, as we will see, ID estimation algorithms provide more of an upper bound of the number of endmembers to use, because in any case, for real data, the definition of the endmembers is a subjective and scale dependent task. Binary partition tree (BPT) based LSU aims at improving the problem of this scale dependency by performing the SU at different scales of the hierarchy defined by the tree. Due to overestimated ID values, or a mismatch between estimated ID and expected number of endmember in local regions, the results can be hard to interpret, because meaningless signatures are often extracted as local endmembers.

On the other hand, the variety of existing sparsity inducing norms can be useful in a

SU context when spectral bundles are used. Indeed, after the clustering step, the bundles form a dictionary which has a strong group structure [1], which should be incorporated into the SU problem for a finer abundance estimation than simply using conventional abundance estimation approaches on the whole dictionary, such as the Fully Constrained Least Squares Unmixing (FCLSU) algorithm.

4.2 Contributions

In this Chapter, our contributions are twofold. First, we propose a new BPT based LSU chain in which the noise estimation is carried out globally, as per the recommendations of Chapter 3, and also incorporates collaborative sparsity in the regional unmixing problem, in order to obtain more interpretable region-wise unmixing results. The objective is to limit or suppress the negative effect of an overestimated local ID. To avoid having to tune a regularization parameter for sparsity in each region, we obtain an algorithmic regularization path, providing the sequence of successively active endmembers when the regularization parameter increases. Then we select the best model using the Bayesian Information Criterion (BIC), to favor models which reconstruct the data well and penalize those with too many parameters. The use of sparsity along with LSU was first sketched in [56], and the method described here can be found in [58]. The results of the proposed approach on a synthetic and a real dataset show the interest of the proposed LSU strategy.

Second, we propose to refine the abundance estimation in the bundles approach to SU with endmember variability by including "social" sparsity inducing norms into the optimization problem, in order to take into account the natural group structure of the bundles into the unmixing. We test several penalties and compare them on synthetic and real datasets, and show that they outperform the results a simple FCLSU algorithm, when used to estimate the abundances with the automated endmember bundles (AEBs) of [143]. This approach was initially proposed in [116]. To the best of our knowledge, this is the first time that group sparsity is used for bundle-based SU accounting for endmember variability. In addition, this is the first time a fractional mixed norm is used to combine within and inter group sparsity effects in a single compact penalty, which is in addition compatible with the ASC, unlike $\mathcal{L}_1$ regularization.

4.3 Local Spectral Unmixing and Sparsity

4.3.1 Motivation

As we have already mentioned in section 2.3.2, local ID estimation and LSU are linked. Indeed, an LSU approach requires the estimation of the ID in each subset used in order to

[1] We recall that this "group structure" denotes the organization of the bundle matrix and of the abundance coefficients into several clusters, not the algebraic structure.

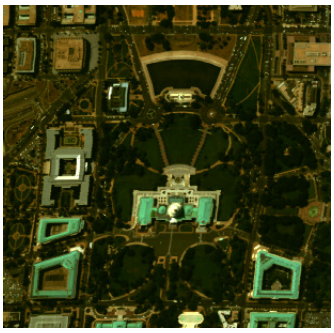


Figure 4.1: The Washington DC mall dataset

extract an appropriate number of local endmembers. This is all the more crucial with BPT based LSU since the extracted endmembers are part of the region model, which will be used to decide when and with which other region the current region will merge. This means that ID overestimation will propagate errors all along the construction and pruning of the BPT. We have already seen what could be done in order to use ID estimation algorithms in the best way possible in Chapter 3. However, even when the noise is estimated globally instead of locally, there can still be overestimations of the number of endmembers to use. In addition, the ID is not systematically linked to the number of endmembers to use, which has a hazy and empirical definition. On real datasets, the ID can be thought of an upper bound of the number of endmembers to use, depending on the desired scale and application. To evidence this phenomenon, we constructed a BPT on a $302 \times 307 \times 191$ subset of the Washington DC mall dataset (shown in Fig. 4.1), acquired by the HYDICE sensor, whose wavelengths range from 400nm to 2500nm, in 191 spectral bands, and with a spatial resolution of 2.8 m. This image was taken over Capitol Hill, where the United States Capitol can be found, in Washington DC, USA. We describe the parameters used for this BPT construction, most of which will be the same in all section 4.3. The initial segmentation was obtained using a mean shift clustering algorithm [41], giving an initial partition with 5760 regions. We used the spectral region model of Eq. (2.9), along with the mergion criterion of Eq. (2.11). We chose the Vertex Component Analysis (VCA) as the endmember extraction algorithm (EEA), but as many other related algorithms, it has a stochastic component. In order to deal with this, we ran VCA 20 times in each region, using the local ID estimate (computed with the Random Matrix Theory (RMT) algorithm of [32], using a global noise estimation) as the number of endmembers to extract, and kept the endmember set giving the largest simplex volume. The chosen energy function for any partition admissible for the BPT pruning is $\varepsilon_{\max}(\pi)$ of Eq. (2.14). This energy function minimizes the average of the maximum values of the RMSE in each region of the partition. This can produce higher region-wise RMSE values than simply minimizing the average RMSE over the regions of the partition, but it is less sensitive to outliers and has been shown to provide better looking segmentations [161]. From this BPT, we show in Fig. 4.2 a plot of the estimated ID as a function of the region size, similarly to what was done in Chapter 3, except that here the regions are not sliding windows anymore, but the regions of the BPT (before the pruning). Another difference is that since the regions are not square tiles, we use directly the number of pixels as the regions size instead

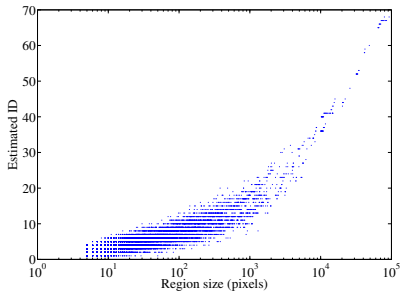


Figure 4.2: Local ID of the regions of the BPT on the Washington Mall dataset (on a logarithmic scale for the x-axis).

of the side of the tiles. Although there is no overestimation peak, thanks to the global noise estimation, the estimated ID seems relatively high, up to 70 for the largest regions of the BPT, and regularly over 10 for small regions, (even for regions of 100 pixels or less). In order to provide more evidence for this phenomenon, we compare visually in Fig. 4.3 two regions of the optimal segmentation obtained by keeping around 500 regions (keeping a certain number of regions is equivalent to an appropriate tuning of λ_{BPT} during the pruning step), as well as some of their properties. Although a visual inspection can be misleading, we do not expect more than three, perhaps four endmembers in the region on the left, and one or two on the region on the right. However, the estimated IDs were respectively of 20 and 9. The local endmembers extracted by VCA are shown in the bottom row. They clearly show that many signatures are very similar and are probably associated to the same materials. For instance, on the left there are at least 6 signatures with a very low spectrum, which are all associated to shadowed areas of the region. For the rooftop region, only 3 signatures or so are significantly different from one another in terms of spectral distance (that is, neglecting scaling effects of the signatures). Therefore, it hardly comes as a surprise that most of the abundance maps of these two regions are very sparse (sometimes only non negligible in extremely small regions or even isolated pixels), or only there to fit the noise. This means that most of them do not correspond to significant instances of the same materials, and are then not meaningful in terms of SV. However, their presence influences the root mean squared errors (RMSE) in the region, and especially if they are given the same weight as legitimate endmembers in the region model. In order to have better interpretable results in local regions, these dummy endmembers should be discarded in the BPT construction process.

4.3.2 Description of the approach

4.3.2.1 Using collaborative sparsity to discard irrelevant endmembers

To select the endmembers which should be discarded in the unmixing process, we would like to force the ones whose abundance maps are already very sparse or low in most pixels to be zero everywhere, for each region. To do that, we use collaborative sparsity. Indeed, the mixed $\mathcal{L}_{2,1}$ norm encourages row-wise sparsity in the abundance matrix, and will have the desired

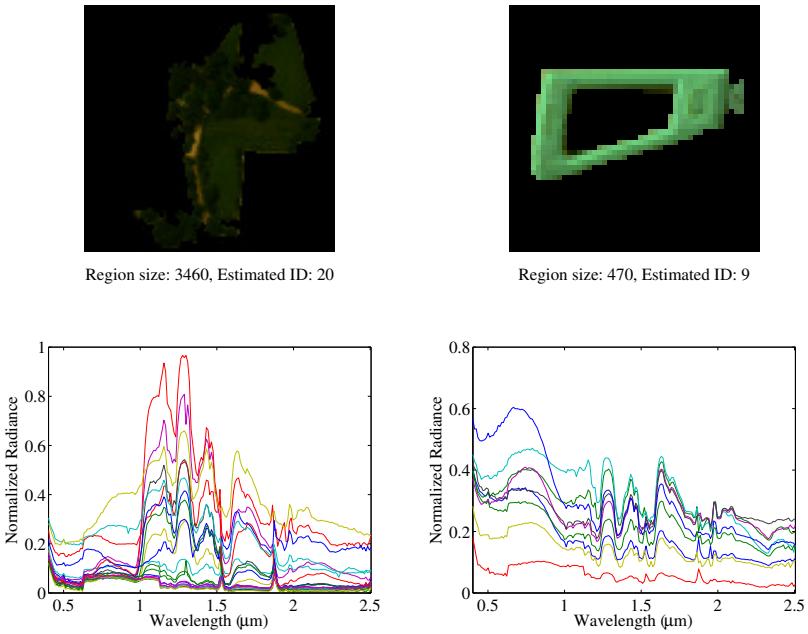


Figure 4.3: Two regions of a BPT built on the Washington dataset (top row), and the associated extracted endmembers (bottom row).

effect of nulling the abundance coefficients of irrelevant endmembers. Thus in each region $\mathcal{R}$, we are going to solve the following problem to estimate the abundances of each column of the local endmember matrix $\mathbf{S}_\mathcal{R} \in \mathbb{R}^{L \times d_\mathcal{R}}$, where $d_\mathcal{R}$ is the estimated ID in region $\mathcal{R}$:

$$\arg\min_{\mathbf{A}_\mathcal{R}} \frac{1}{2}||\mathbf{X}_\mathcal{R} - \mathbf{S}_\mathcal{R}\mathbf{A}_\mathcal{R}||_F^2 + \lambda_\mathcal{R}||\mathbf{A}_\mathcal{R}||_{2,1} + \mathcal{I}_{\Delta_{d_\mathcal{R}}}(\mathbf{A}_\mathcal{R}), \tag{4.1}$$

where $\mathcal{I}_{\Delta_{d_\mathcal{R}}}$ is the indicator function of the probability simplex, which has to be understood as being applied independently to each column of $\mathbf{A}_\mathcal{R}$. Depending on the value of $\lambda_\mathcal{R}$, some local endmembers will have zero abundance maps on the whole support of the region. To solve this (convex) problem, we introduce split variables such that the problem becomes:

$$\begin{aligned} &\arg\min_{\mathbf{A}_\mathcal{R}} \frac{1}{2}||\mathbf{X}_\mathcal{R} - \mathbf{S}_\mathcal{R}\mathbf{A}_\mathcal{R}||_F^2 + \lambda_\mathcal{R}||\mathbf{U}_\mathcal{R}||_{2,1} + \mathcal{I}_{\Delta_{d_\mathcal{R}}}(\mathbf{V}_\mathcal{R}) \\ &\text{s.t. } \mathbf{U}_\mathcal{R} = \mathbf{A}_\mathcal{R},\ \mathbf{V}_\mathcal{R} = \mathbf{A}_\mathcal{R}. \end{aligned} \tag{4.2}$$

With the problem in this form, we can then use the ADMM technique [22] (see Appendix A) to solve it. The ADMM consists in expressing the constrained problem defined in Eq. (4.2) in an unconstrained way using an Augmented Lagrangian (AL), and then minimizing it iteratively and alternatively for each of the variables introduced, including the Lagrange Multipliers appearing in the AL (the so-called *dual update*). ρ is the *barrier* parameter weighting the AL terms. Here, the Augmented Lagrangian writes :

$$\begin{aligned} \mathcal{L}(\mathbf{A}_\mathcal{R}, \mathbf{U}_\mathcal{R}, \mathbf{V}_\mathcal{R}) = &\ \frac{1}{2}||\mathbf{X}_\mathcal{R} - \mathbf{S}_\mathcal{R}\mathbf{A}_\mathcal{R}||_F^2 + \lambda_\mathcal{R}||\mathbf{U}_\mathcal{R}||_{2,1} + \mathcal{I}_{\Delta_{d_\mathcal{R}}}(\mathbf{V}_\mathcal{R}) \\ &+ \frac{\rho}{2}||\mathbf{A}_\mathcal{R} - \mathbf{U}_\mathcal{R} - \mathbf{C}_\mathcal{R}||_F^2 + \frac{\rho}{2}||\mathbf{A}_\mathcal{R} - \mathbf{V}_\mathcal{R} - \mathbf{D}_\mathcal{R}||_F^2 - \frac{\rho}{2}||\mathbf{C}_\mathcal{R}||_F^2 - \frac{\rho}{2}||\mathbf{D}_\mathcal{R}||_F^2, \end{aligned} \tag{4.3}$$

where $\mathbf{C}_\mathcal{R}$ and $\mathbf{D}_\mathcal{R}$ are the set of dual variables. The ADMM procedure to solve this problem in each region is summarized in Algorithm 1.

Data: $\mathbf{X}_\mathcal{R}$, $\mathbf{S}_\mathcal{R}$
Result: $\mathbf{A}_\mathcal{R}$
Initialize $\mathbf{A}_\mathcal{R}$ and choose $\lambda_\mathcal{R}$;
while *ADMM termination criterion is not satisfied* **do**
$\mathbf{A}_\mathcal{R} \leftarrow (\mathbf{S}_\mathcal{R}^\top \mathbf{S}_\mathcal{R} + 2\rho \mathbf{I}_{d_\mathcal{R}})^{-1}(\mathbf{S}_\mathcal{R}^\top \mathbf{X}_\mathcal{R} + \rho(\mathbf{U}_\mathcal{R} + \mathbf{V}_\mathcal{R} + \mathbf{C}_\mathcal{R} + \mathbf{D}_\mathcal{R}))$;
$\mathbf{U}_\mathcal{R} \leftarrow \mathbf{prox}_{(\lambda_\mathcal{R}/\rho)||\cdot||_{2,1}}(\mathbf{A}_\mathcal{R} - \mathbf{C}_\mathcal{R}) = \mathbf{soft}_{\lambda_\mathcal{R}/\rho}(\mathbf{A}_\mathcal{R} - \mathbf{C}_\mathcal{R})$;
$\mathbf{V}_\mathcal{R} \leftarrow \mathbf{prox}_{\mathcal{I}_{\Delta_{d_\mathcal{R}}}}(\mathbf{A}_\mathcal{R} - \mathbf{D}_\mathcal{R}) = \mathbf{proj}_{\Delta_{d_\mathcal{R}}}(\mathbf{A}_\mathcal{R} - \mathbf{D}_\mathcal{R})$;
$\mathbf{C}_\mathcal{R} \leftarrow \mathbf{C}_\mathcal{R} + \mathbf{U}_\mathcal{R} - \mathbf{A}_\mathcal{R}$;
$\mathbf{D}_\mathcal{R} \leftarrow \mathbf{D}_\mathcal{R} + \mathbf{V}_\mathcal{R} - \mathbf{A}_\mathcal{R}$;
end

Algorithm 1: ADMM to solve problem (4.2).

The update of $\mathbf{A}_\mathcal{R}$ has a closed form expression. The next two suproblems are separable w.r.t. the pixels. The updates of the two split variables $\mathbf{U}_\mathcal{R}$ and $\mathbf{V}_\mathcal{R}$ are proximal updates (which have to be understood columnwise for the simplex projection, and rowwise for the update of $\mathbf{U}_\mathcal{R}$). The update of $\mathbf{U}_\mathcal{R}$ requires the use of the proximal operator of the $\mathcal{L}_2$ norm, the so-called *block soft-thresholding* operator, denoted by **soft**, and of the proximal operator of the indicator function of the simplex (c.f. Appendix A), a projection on the simplex. This can be efficiently carried out using the algorithm of [45]. The last two updates are the dual updates of the Lagrange multipliers. We can stop the algorithm when the relative variation between two iterates of $\mathbf{A}_\mathcal{R}$ (measured in Frobenius norm) is below a certain tolerance, for example $\epsilon_A = 10^{-3}$.

However, there are two problems with this approach. The first is that since the linear constraints of the ADMM are only satisfied asymptotically, we have no guarantee that all the entries of the supposedly discarded rows of the abundance matrix will be exactly zero (and this actually happens in practice). Then an arbitrary thresholding step is required to eliminate endmembers with a small contribution [85, 8]. The second is that in order to obtain the appropriate sparsity level, the regularization parameter $\lambda_\mathcal{R}$ needs to be tuned, in every region. We could use a grid search over a set of parameters in each region, but this would be very computationally costly and would require a criterion to select the best run of the algorithm. We will see that we can find solutions for both issues.

4.3.2.2 Obtaining a an algorithmic regularization path

In order to tackle both the regularization parameter issue and the inexact sparsity of the collaborative sparse regression at once, we would like to obtain the regularization path of the solution, as a function of $\lambda_\mathcal{R}$. As we have already mentioned, computing this empirically would be computationally prohibitive. Regularization paths can sometimes be computed easily, for instance on the LASSO (for Least Absolute Shrinkage Selection Operator) problem,

which is nothing more than a least squares regression with a $\mathcal{L}_1$ regularization [61]. However, for more complex problems, such as ours, there is no way, to our knowledge, to obtain this regularization path easily, without an extensive grid search. However, we are going to find a convenient workaround for this by computing a so-called ADMM algorithmic regularization path, introduced in [83]. This approach is able to use the ADMM algorithm to quickly approximate the sequence of active supports of the variable of interest, when the regularization parameter increases, for certain sparsity regularized least squares problems. Even though there are as of today no theoretical guarantees on the efficiency of this algorithm, it was experimentally shown to be able to efficiently approximate the true sequence of active sets on several problems [83], including the LASSO. Here, we propose to extend this algorithm to collaborative sparsity.

Since exactly solving the optimization problem for a large number of regularization parameters would be too time consuming, we are more interested in finding the active set of endmembers when the weight of the sparsity increases w.r.t. this of the data fit term. The idea is, for each region involved in the construction of the BPT, to find a sequence of endmember matrices, whose number of endmembers are decreasing from the number of endmembers initially extracted to zero (when the model is fully sparse, for a very high penalty on dense solutions). Each new matrix contains the same endmembers as the previous one, except for one, which is the next endmember to be discarded when the weight of the sparsity term gets more important.

To do that, we modify the ADMM algorithm in order to quickly obtain the support of the regularization path, for each region. An iteration of Algorithm 1 is carried out for a very small value of the regularization parameter (which guarantees a fully dense solution). Then, the variables obtained at the end of the iteration are used as a warm start for another iteration with a new regularization parameter, slightly higher than the previous one. By repeating this for several iterations with higher and higher regularization parameters, the split variable $\mathbf{U}_\mathcal{R}$, which undergoes the soft thresholding becomes increasingly sparse. Since we are using warm starts, and because the new regularization parameter is only slightly different from the previous one, even if the ADMM is not fully converged at each iteration, the support of the active set is encoded in $\mathbf{U}_\mathcal{R}$, often in one iteration only, long before this active set is propagated to $\mathbf{A}_\mathcal{R}$ (this will be the case only at convergence, when the constraints of problem (4.2) are satisfied). With these modifications, we obtain Algorithm 2. The notation $||\mathbf{U}^i_\mathcal{R}||_{2,0}$ denotes the number of nonzero rows of the matrix $\mathbf{U}^i_\mathcal{R}$. In this algorithm, we have not used the same notation for the regularization parameter as in Algorithm 1 because since we are not solving exactly the ADMM for each value of $\gamma_\mathcal{R}$, the level of regularization at each iteration is not the same as before. Here, we are using a geometric progression for the values of $\gamma_\mathcal{R}$, whose common ratio is t. The value of t should be small to approximate the active sets of the regularization path well enough. We begin each iteration with the update of $\mathbf{U}_\mathcal{R}$ because it is the variable affected by a change in γ. The regularization space can be explored very quickly since the algorithm provides at most $d_\mathcal{R}$ endmember subsets of the full endmember set extracted by VCA, that need to be tested after this process. In practice we chose $\gamma^0_\mathcal{R} = 10^{-4}$ and $t = 1.04$, which allows $\gamma_\mathcal{R}$ to sweep from 10^{-4} to 5.10^4 in around 500 iterations (which is less than what is usually required in practice to reach a fully sparse model).

Data: $\mathbf{X}_{\mathcal{R}}$, $\mathbf{S}_{\mathcal{R}}$
Result: The sequence of $\mathbf{U}^i_{\mathcal{R}},\ i = 0, ..., i_{max}$
Initialize $\mathbf{A}^0_{\mathcal{R}}$ and choose $\gamma^0_{\mathcal{R}}$ and $t > 0$;
while $||\mathbf{U}^i_{\mathcal{R}}||_{2,0} \neq 0$ **do**
$\gamma^i_{\mathcal{R}} \leftarrow t\gamma^{i-1}_{\mathcal{R}}$;
$\mathbf{U}^i_{\mathcal{R}} \leftarrow \mathbf{prox}_{(\gamma^i_{\mathcal{R}}/\rho)||\cdot||_{2,1}}(\mathbf{A}^{i-1}_{\mathcal{R}} - \mathbf{C}^{i-1}_{\mathcal{R}}) = \mathbf{soft}_{\gamma^i_{\mathcal{R}}/\rho}(\mathbf{A}^{i-1}_{\mathcal{R}} - \mathbf{C}^{i-1}_{\mathcal{R}})$;
$\mathbf{A}^i_{\mathcal{R}} \leftarrow (\mathbf{S}^\top_{\mathcal{R}}\mathbf{S}_{\mathcal{R}} + 2\rho\mathbf{I}_{d_{\mathcal{R}}})^{-1}(\mathbf{S}^\top_{\mathcal{R}}\mathbf{X}_{\mathcal{R}} + \rho(\mathbf{U}^i_{\mathcal{R}} + \mathbf{V}^{i-1}_{\mathcal{R}} + \mathbf{C}^{i-1}_{\mathcal{R}} + \mathbf{D}^{i-1}_{\mathcal{R}}))$;
$\mathbf{V}^i_{\mathcal{R}} \leftarrow \mathbf{prox}_{\mathcal{I}_{\Delta_{d_{\mathcal{R}}}}}(\mathbf{A}^i_{\mathcal{R}} - \mathbf{D}^{i-1}_{\mathcal{R}}) = \mathbf{proj}_{\Delta_{d_{\mathcal{R}}}}(\mathbf{A}^i_{\mathcal{R}} - \mathbf{D}^{i-1}_{\mathcal{R}})$;
$\mathbf{C}^i_{\mathcal{R}} \leftarrow \mathbf{C}^{i-1}_{\mathcal{R}} + \mathbf{U}^i_{\mathcal{R}} - \mathbf{A}^i_{\mathcal{R}}$;
$\mathbf{D}^i_{\mathcal{R}} \leftarrow \mathbf{D}^{i-1}_{\mathcal{R}} + \mathbf{V}^i_{\mathcal{R}} - \mathbf{A}^i_{\mathcal{R}}$;
$i \leftarrow i + 1$
end

Algorithm 2: ADMM algorithmic regularization path for problem (4.2).

4.3.2.3 Selecting the best model

In order to obtain the sequence of active sets, one simply has to examine the successive active sets of the iterates of $\mathbf{U}_{\mathcal{R}}$, and store a sequence of at most $d_{\mathcal{R}}$ sparser and sparser candidate endmember matrices (denoted as $\mathbf{S}^i_{\mathcal{R}}$). The only operation left is to select the optimal active set in the sense of some criterion. In our case, we used the Bayesian Information Criterion (BIC) [141] (closely related to the Akaike Information Criterion (AIC) [6]), which helps selecting the best model in a set of candidate models, by favoring models with an important likelihood, while penalizing models with a high number of parameters.

More precisely, the purpose of the BIC is to select the model M_{opt} (out of a collection of models $M = \{M_i\}$) which maximizes the posterior density of the model, given the data:

$$M_{opt} = \arg\max_{M_i} P_{M_i|\mathbf{X}}(M_i|\mathbf{X}). \tag{4.4}$$

The posterior density for model M_i is, according to Bayes' Rule:

$$P_{M_i|\mathbf{X}}(M_i|\mathbf{X}) = \frac{P_{\mathbf{X}|M_i}(\mathbf{X}|M_i)P_{M_i}(M_i)}{P_{\mathbf{X}}(\mathbf{X})}. \tag{4.5}$$

This definition allows to put prior densities on each model. In our cases, since local ID estimation provides an upper bound of the number of endmembers to use, we chose to put the same constant prior on all candidate models. To compute $P_{\mathbf{X}|M_i}(\mathbf{X}|M_i)$, it is necessary to marginalize the joint distribution of the data and the parameters of the model, given the model M_i, over all the possible values of the parameters (stored in a vector $\boldsymbol{\theta}_i$):

$$P_{\mathbf{X}|M_i}(\mathbf{X}|M_i) = \int_{\boldsymbol{\Theta}_i} P_{\mathbf{X},\boldsymbol{\theta}_i|M_i}(\mathbf{X}, \boldsymbol{\theta}_i|M_i)\mathrm{d}\boldsymbol{\theta}_i. \tag{4.6}$$

This integral is very hard to compute exactly, and therefore we use an approximation of it, by resorting Laplace's integration method [102]. After some algebra, we obtain the general

formula for a quantity proportional to the posterior density, which defines the BIC:

$$BIC_i \triangleq \ln(n)k_i - 2\ln(\mathscr{L}_i), \tag{4.7}$$

where n stands for the number of samples, k_i is the number of parameters of model i, and $\mathscr{L}_i$ is the likelihood of model i. In our problem, with the assumption that the noise and modeling errors ($\mathbf{E}_\mathcal{R}$) are distributed according to a centered i.i.d. multivariate Gaussian whose covariance matrix is equal to $\sigma^2\mathbf{I}_L$, the formula (4.7) becomes [129]:

$$BIC_i = \ln(L)P_i + L\ln\left(\frac{||\mathbf{X}_\mathcal{R} - \mathbf{S}^i_\mathcal{R}\hat{\mathbf{A}}^i_\mathcal{R}||^2_F}{L}\right), \tag{4.8}$$

where P_i is the number of columns (i.e. the number of endmembers) in $\mathbf{S}^i_\mathcal{R} \in \mathbb{R}^{L\times P_i}$. We recall that L is the number of spectral bands. $\hat{\mathbf{A}}^i_\mathcal{R}$ is the abundance matrix estimated by FCLSU using the data and the endmember matrix $\mathbf{S}^i_\mathcal{R}$. The best model is simply the one minimizing the BIC value.

The BIC has interesting properties: it allows to add priors on the density of the set of models if needed (the formula has to be updated accordingly), and when the number of samples (here spectral bands) becomes very high, M_{opt} is the model minimizing the Kullback-Leibler divergence between the true density and that of the model, with the minimum number of parameters to avoid overfitting [102].

In our case, in order to alleviate the computational load, we perform these FCLSU steps from the smallest endmember matrices to the denser ones, and stop when the BIC value has increased for three consecutive iterations, to avoid performing numerous useless abundance matrix estimations in each region. The flowchart of this new way to build the region model is shown in Fig. 4.4.

4.3.3 Results

4.3.3.1 Results on synthetic data

In this section, we demonstrate the effect of the proposed spurious endmember elimination strategy for a small toy example synthetic image, which will mimic what happens in a local region of a larger hyperspectral image. First, we randomly picked 6 endmembers out of the mineral spectral library of the United States Geological Survey (USGS). We have resampled these spectra so that the number of spectral bands is 300. Then, we generated synthetic abundance maps for those endmembers using Gaussian Random Fields, which comply with the ASC and the ANC (shown in Fig. 4.5). We used the LMM to mix the data. Finally, we added an additive (spectrally and spatially) white Gaussian noise, such that the signal to noise ratio (SNR) is 25dB. This provides a $40 \times 40 \times 300$ simulated hyperspectral image. We know from Chapter 3 and [54] that with these settings, namely small spatial dimensions, large spectral dimension, and nonnegligible noise, ID estimation algorithms are likely to overestimate the actual number of endmembers. For instance, on this dataset, the Hyperspectral subspace

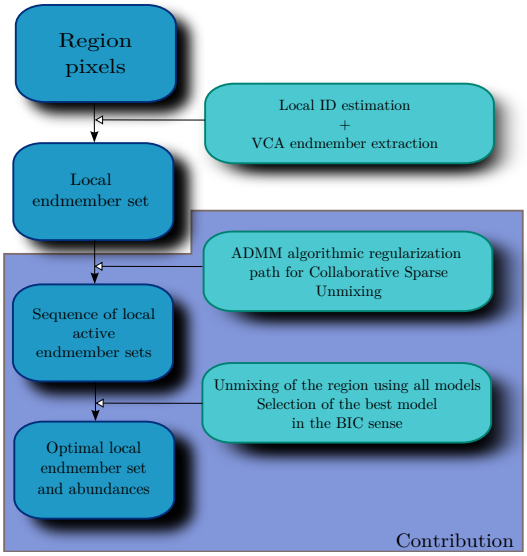


Figure 4.4: Flowchart of the proposed modifications to the region model.

identification by minimum error (HySIME) algorithm of [19] estimated the ID to be 20, and the Random Matrix Theory-based (RMT) algorithm of [32] returned an ID value of 26, whereas 6 endmembers were used to linearly mix the data.

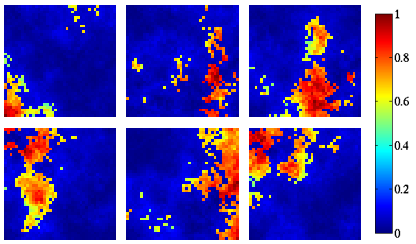
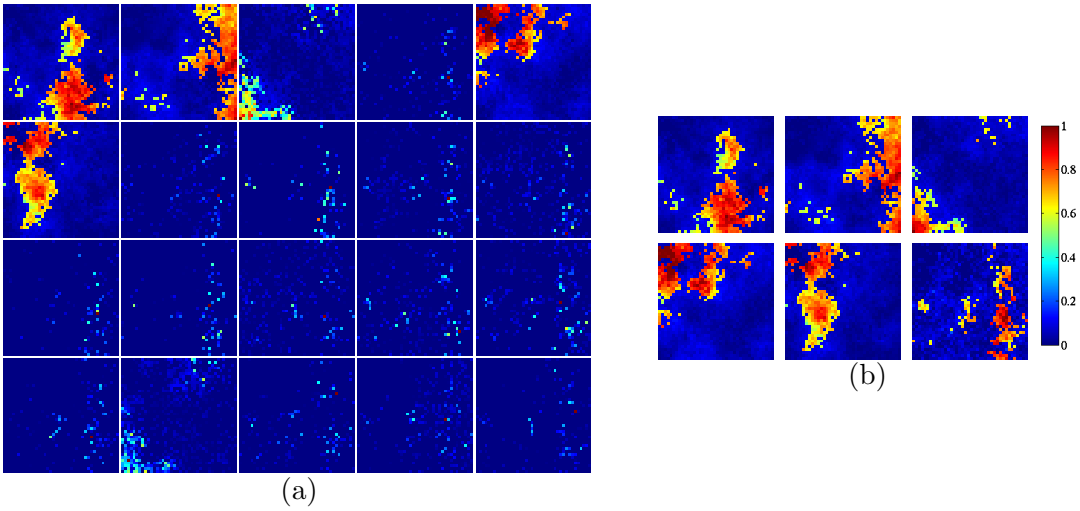


Figure 4.5: True abundance maps for the simulated data. The color scale goes from 0 (blue) to 1 (red).

Then, we ran a simple unmixing chain on the data, first extracting 20 endmembers (as recommended by HySIME) from the data using the VCA algorithm. We estimated the abundances using the FCLSU algorithm. They are shown in Fig. 4.6 (a). We can see that four of the extracted endmembers correspond to the actual ones which were used to generate the data. However the last two are not well recovered by this method. Indeed, for one of them (corresponding to the top left abundance map of Fig. 4.5), the abundances seem spread between two maps in Fig. 4.6 (a). The last abundance maps (middle of the bottom row of Fig. 4.5) is spread in the remaining maps of Fig. 4.6). This means that because of the ID overestimation VCA extracted redundant signatures, which make the abundance estimation results harder to interpret, even in such a simple case. The proposed endmember elimination scheme (with $t = 10^{-2}$ and $\lambda_0 = 10^{-3}$), however, is able to only keep the 6 most relevant endmembers, and to estimate their abundances correctly (see Fig. 4.6 (b)). We show the BIC

and the values of the two terms involved in its computation in Fig. 4.7. We can see that after the number of endmembers in the model goes over 6 (recall that the endmembers have been sorted so that each added endmember is less explicative of the dataset than the next), the data fit term (left) almost does not decrease anymore, while the number of parameters (middle) increases, which makes the BIC reach a minimum for 6 endmembers.

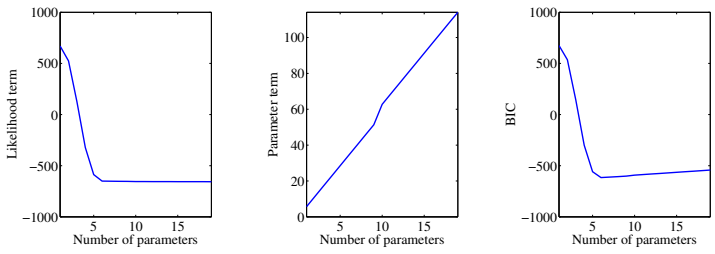


(a) (b)

Figure 4.6: Abundance maps estimated without sparsity by FCLSU on the whole set of endmembers extracted by VCA (a) and with the proposed model selection (b). The color scale goes from 0 (blue) to 1 (red).

4.3.3.2 Results on real data

In order to assess the impact of the proposed modifications to the region model, we built two BPTs on the Washington Mall dataset, one without sparsity, and one with the proposed region model. First we show the effect of collaborative sparsity on the local number of end-



(a) (b) (c)

Figure 4.7: BIC values (c) (decomposed into the likelihood term (a) and the parameter term (b)) for the sequence of endmember matrices obtained through the proposed method, for the simulated dataset.

members, in Fig. 4.8. We can see that using the BIC criterion on the sequence of models extracted by the modified ADMM significantly reduces the number of endmembers in each region, with respect to the estimated ID. We also computed, in each case, the optimal seg-

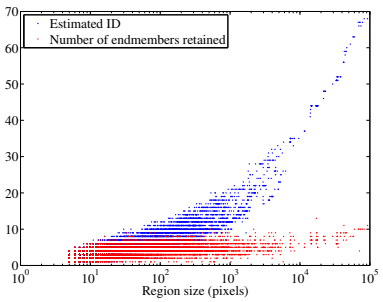
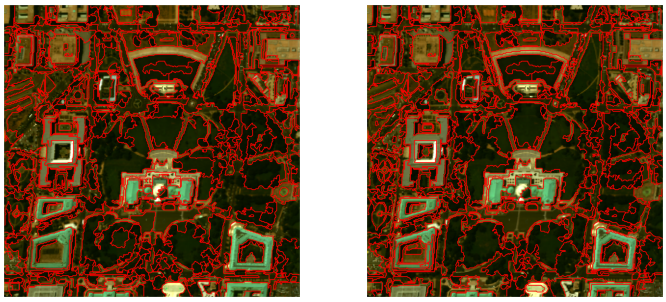


Figure 4.8: Local ID of the regions of the BPT on the Washington Mall dataset (in blue) and number of kept endmembers when the BPT is built with the proposed region model (in red) (on a logarithmic scale for the x-axis).

mentations whose number of regions is closest to 500 (finding the appropriate value of λ_{BPT} using the persistent intervals [70]). The regions of these segmentations can correspond to actual structures in the data, but it is not always the case, since we are looking for partitions minimizing the RMSE, not for homogeneous regions (in the sense of low variance ones). The segmentations are relatively similar, although some differences can be found. The similarities of the segmentations show that we have been able to discard the useless endmembers without significantly impacting the average RMSE. For the sake of illustration, we apply the proposed algorithm to discard the irrelevant endmembers of the region on the right of Fig. 4.3. The average RMSE of this region without sparsity is 0.0061, using 9 endmembers. The proposed approach (for a given run) only retains 3 endmembers, with a RMSE of 0.0064. This shows that we have been able to discard irrelevant endmembers by removing the redundant or meaningless information in the region.

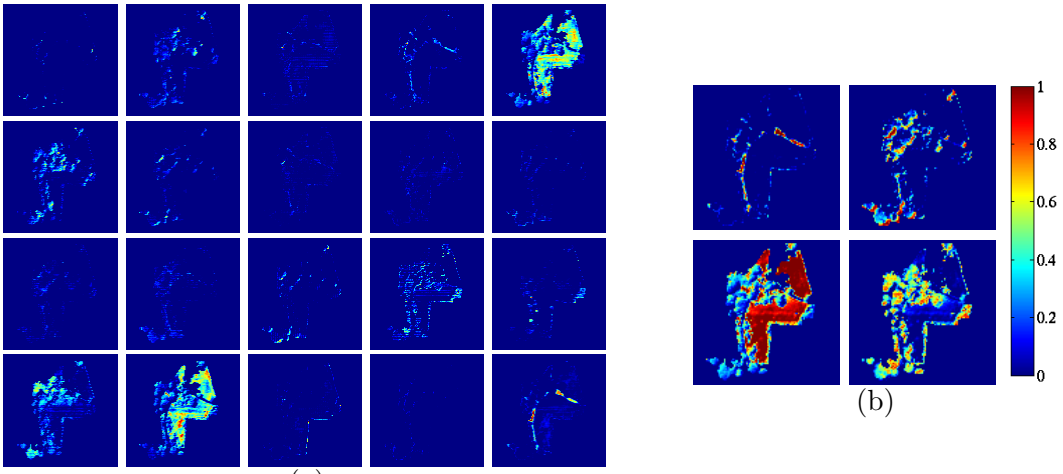
We are going to show that the sparsity imposed by the proposed region model also has a significant impact on the interpretability of the results. To do that, we take the region on the left of Fig. 4.3, taken from the partition of Fig. 4.9 (a), and show the difference in abundance maps with or without collaborative sparsity. These results are presented in Fig. 4.10. When no sparsity is applied, we can see that at least 8 abundance maps have negligible values on almost all the support of the region. Only around 5 abundance maps are really meaningful at the scale of the region. There seems to be 2 instances of grass, two instances of trees and one endmember associated to the gravel pathway. When we use the proposed scheme, only four endmembers are retained: one for grass, two for trees (including one for shadowed parts of the trees), and one for gravel. The different terms involved in the computation of the BIC are displayed in Fig. 4.11. These plots confirm that the likelihood term (which is very related to the mean RMSE in a region) does not decrease much when more than 4 endmembers are retained, while the parameter term increases linearly. The sparsity only kept the most relevant signatures, making the results more easily interpretable at the region scale.



(a) (b)

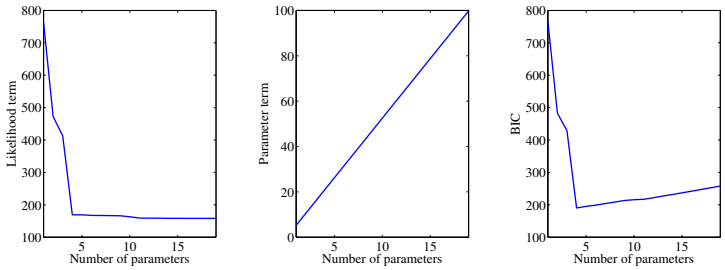
Figure 4.9: Optimal segmentations whose number of regions is the closest to 500, when no sparsity is considered (a) and using the proposed modifications to the region model (b).

It may also happen in some regions which seem relatively homogeneous visually that even with sparsity a significant number of endmembers are retained: this happens for instance in the water pond in the center of the upper part of the image: in the most central region, the estimated ID is 5, and no endmember was discarded after the collaborative unmixing. In this region, comprising a shallow water pond (around 50 cm deep), and some kind of concrete at the bottom, the mixing process is likely to be highly nonlinear. The BPT approach allowed to isolate this region from the rest of the image, by segmenting it, avoiding the propagation of the errors due to the endmembers of this region. Similarly, in regions which visually correspond to one macroscopic material (e.g. in the region on the right of Fig. 4.3), several endmembers (around 3 to 6 in this case, depending on the VCA runs) can be retained, because of spectral variability within the region. Since the used LMM does not account for SV, several endmembers are necessary to fit the data well. This shows that it would be interesting to have a mixing model incorporating SV explicitly in order to better unmix this type of regions.



(a) (b)

Figure 4.10: Abundance maps in the region on the left of Fig. 4.3, without sparsity (a) and with the proposed model selection (b). The color scale goes from 0 (blue) to 1 (red).



(a) (b) (c)

Figure 4.11: BIC values (c) (decomposed into the likelihood term (a) and the parameter term (b)) for the sequence of endmember matrices obtained through the proposed method, for the region on the left of Fig. 4.3.

4.4 Spectral Bundles and Social Norms

In this section, we leave aside LSU to focus on sparsity inducing penalties for SU in the context of endmember bundles. Indeed, as we have seen in section 2.3.2, after the clustering step, the dictionary of spectral bundles is organized into several groups, because it is divided in a certain number of bundles, each of them comprising several instances of the same endmember. For sufficiently large dictionaries, applying sparsity inducing penalties in the abundance estimation problem can be useful to discard irrelevant endmember candidates in each pixel. For instance, a simple $\mathcal{L}_1$ norm minimization will help, but it does not take into account the group structure of the bundles, and conflicts with the ASC. In the signal processing literature, several sparsity inducing penalties exist to incorporate the structure of the coefficients matrix. We detail some of them in the following sections, before testing their effects on the SU problem using AEB on synthetic and real datasets. These results were first presented in [116].

4.4.1 Social Sparsity

All the penalties we are going to detail in the following are based on applying a mixed norm on the abundance vector (in each pixel), which is endowed with a group structure G, which partitions the $Q = nP$ endmembers extracted by the n runs of VCA on random subsets into P groups (as many as the number of materials to unmix). We drop the pixel index for simplicity of the notation. In the most general form, the group two-level mixed $\mathcal{L}_{G,p,q}$ norm is defined, for any two positive real numbers p and q as [97]:

$$||\mathbf{a}||_{G,p,q} = \left(\sum_{i=1}^{P} \left(\sum_{j=1}^{m_{G_i}} |a_{G_i,j}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} = \left(\sum_{i=1}^{P} ||\mathbf{a}_{G_i}||_p^q \right)^{\frac{1}{q}}, \tag{4.9}$$

where m_{G_i} is the number of instances in group G_i, and $\mathbf{a}_{G_i}$ is a subvector of $\mathbf{a}$ comprising all the abundance coefficients associated to the endmembers of group G_i. This equation only

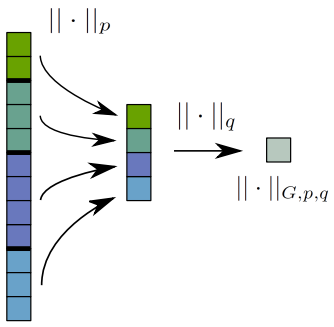


Figure 4.12: Illustration on how the $\mathcal{L}_{G,p,q}$ norm operates on a vector, given the group structure G.

defines a true norm for $p, q \geq 1$ (and also for p or $q = \infty$, by taking limits). As explained in Fig. 4.12, the idea is to take the p norm of each of the P subvectors of coefficient defined by the groups, to store the results in a P-dimensional vector, for which we are going to compute the q norm. We will see that with smart choices of p and q, different types of sparsity can be obtained when this mixed norm is used as a regularized in the unmixing with bundles. The definition of the $\mathcal{L}_{G,p,q}$ norm can easily be extended to a matrix $\mathbf{A} \in \mathbb{R}^{Q \times N}$, using the same expression, operating columnwise, and summing the results on all pixels:

$$||\mathbf{A}||_{G,p,q} = \sum_{k=1}^{N} ||\mathbf{a}_k||_{G,p,q}. \tag{4.10}$$

Note that this group matrix $\mathcal{L}_{G,p,q}$ norm is not the same as the mixed $\mathcal{L}_{p,q}$ matrix norm defined in section 1.5. However, there is a connection, because using a mixed matrix norm amounts to impose a particular group structure on the coefficients of the matrix, using its rows. Let us define a vector $\mathbf{a}^\dagger = \text{vec}(\mathbf{A}^\top)$ (vec being a vectorization operator which stacks the columns of a matrix). Then $\mathbf{a}^\dagger$ is a (column) vector with all the rows of $\mathbf{A}$ stacked. If we divide this vector into P groups, each comprising the coefficients corresponding to one row of $\mathbf{A}$, then we have $||\mathbf{a}^\dagger||_{G,p,q} = ||\mathbf{A}||_{p,q}$.

Here, we are interested in norms which can handle any group structure, while enforcing several types of sparsity. For instance, the use of sparsity in SU is based on the assumption that a few materials are active in each pixel. If the dictionary of endmembers has a group structure, it makes sense to enforce sparsity on the number of groups, rather than on the total number of signatures. This rationale is the basis of the so-called group LASSO [115], which is widely used in many signal processing applications. This method uses the $\mathcal{L}_{G,2,1}$ norm, (not to be confused with the collaborative case, which use a very particular group structure, as seen above), which enforces sparsity on the vector whose entries are the $||\mathbf{a}_{G_i}||_2$. This means that when one of these entries is zero, the whole group is discarded entirely since the vector $\mathbf{a}_{G_i}$ has a zero norm. Within each group, there is no sparsity and thus most or all signatures are likely to be active. The effect of this penalty on a matrix $\mathbf{A} \in \mathbb{R}^{P \times N}$ is shown in Fig. 4.13. In some cases, for example when we deal with a small number of groups, and we have reasons to believe that there is only one or few instances of each group which are active in a pixel,

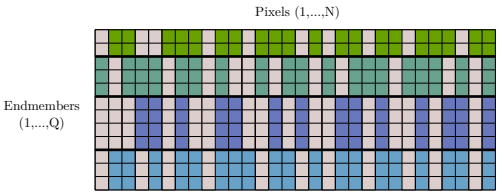


Figure 4.13: Effect of the group LASSO penalty on the abundance matrix. The group structure is shown in colors (the rows of the matrix have been sorted for more clarity). Inactive entries of the matrix are in gray. A small number of groups is selected in each pixel, but within each group the matrix is dense.

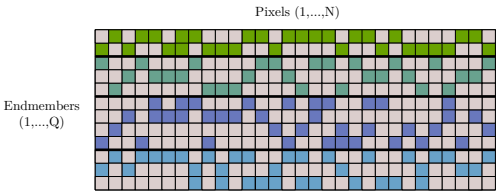


Figure 4.14: Effect of the elitist LASSO penalty on the abundance matrix. A small number of instance is each group is selected in each pixel, but all or almost all groups are active.

we may expect within-group sparsity, without group sparsity. In this case the elitist LASSO penalty [98] is suited to the problem, since it uses the $\mathcal{L}_{G,1,2}$ norm, which promotes a small value of the $\mathcal{L}_1$ norms of each $\mathbf{a}_{G_i}$. The effect of this penalty is shown in Fig. 4.14. Using a penalty which enforces both group sparsity and global sparsity (on the total number of active signatures) also seems appealing. The sparse group LASSO [142] uses a combination of the group lasso penalty and a classical $\mathcal{L}_1$ norm to benefit from both properties. It was recently used in a sparse SU context in [86]. However, in this case, benefiting from both penalties comes at the cost of having two regularization parameters to tune. In our case, this penalty is also at odds with the ASC due to the presence of the $\mathcal{L}_1$ norm in the objective, as we have already pointed out in section 1.5. The ASC can also be contradictory with sparsity in some other configurations: for instance, if each material has only one representative, the group LASSO reduces to the regular LASSO and the ASC conflicts with the objective. In order to avoid this issue, we are also using a fractional case, with the $\mathcal{L}_{G,1,q}$ "norm", with $q = \frac{a}{b}$ (a and $b \in \mathbb{N}$) and $0 < \frac{a}{b} < 1$. This penalty is no longer a norm, because we lose convexity due to the fact that $q \leq 1$, but it has the advantage of enforcing both group sparsity and within-group sparsity in a compact formulation, without conflicting with the ASC anymore. In addition, the $\mathcal{L}_q$ norm $q \leq 1$ is a better approximation of the $\mathcal{L}_0$ norm than the $\mathcal{L}_1$ norm. The effect of the $\mathcal{L}_{G,1,q}$ penalty on the abundance matrix is shown in Fig. 4.15. With either of those penalties, the optimization problem to solve is:

$$\arg\min_{\mathbf{A}} \frac{1}{2}||\mathbf{X} - \mathbf{B}\mathbf{A}||_F^2 + \lambda||\mathbf{A}||_{G,p,q} + \mathcal{I}_{\Delta_P}(\mathbf{A}). \tag{4.11}$$

Note that after solving this problem, in order to obtain the global abundances, one simply

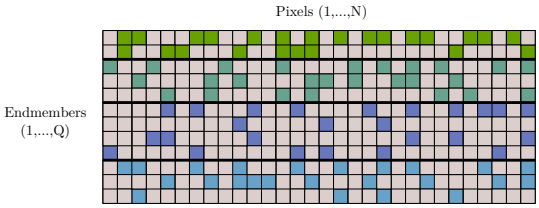


Figure 4.15: Effect of the fractional LASSO penalty on the abundance matrix. A small number of instance is each group is selected in each pixel, and the mixture is also sparse within each group.

has to sum the abundances within each instance of each group, as described in section 2.3.1. The optimization problem (4.11) is convex for both the group and elitist penalties, but not for the fractional one. In any case, we are going to use the ADMM once again to solve it. For both convex penalties (group and elitist), we will use the ADMM in a very similar way to Algorithm 1, with the only change being in the proximal update of the split variable $\mathbf{U}$ (and in using the whole image, not just the pixels of regions). Convergence to the global minimum is automatically guaranteed for the group and elitist penalties. For the nonconvex case, as we will see, the situation is more complex. The ADMM was designed to tackle convex problems, but it has been more and more (successfully) used for nonconvex problems as well, and recent works [164] show that if the nonconvex function satisfies some conditions, the ADMM is proven to converge to a stationary point in the nonconvex case. One of these cases includes the $\mathcal{L}_p$ quasinorm for $p < 1$. This results remains to be shown in the mixed $\mathcal{L}_{G,1,q}$ norm with $q < 1$ case, (but it is likely to satisfy the same conditions, being "less nonconvex" than the $\mathcal{L}_p$ quasinorm, because the unit ball of such a norm will have some nonconvex facets, but not all since some of them will be similar as the facets of the $\mathcal{L}_1$ ball). The next two sections introduce the proximal operators for the group and elitist penalties, while the last one shows how to handle the fractional case.

4.4.1.1 Group penalty case

The update of $\mathbf{u}$ (in each pixel) for the group penalty involves the following proximal operator, which is simply a group version of the block soft thresholding operator:

$$\mathbf{prox}_{\tau||\cdot||_{G,2,1}}(\mathbf{v}) = \begin{bmatrix} \mathbf{soft}_\tau(\mathbf{v}_{G_1}) \\ \vdots \\ \mathbf{soft}_\tau(\mathbf{v}_{G_P}) \end{bmatrix}. \tag{4.12}$$

4.4.1.2 Elitist penalty case

The proximal operator for the elitist norm is a bit more complex, but has a closed form (derived in [96]), which involves the regular ($\mathcal{L}_1$) soft thresholding operator:

$$\mathbf{prox}_{\tau||\cdot||_{G,1,2}}(\mathbf{v}) = \begin{bmatrix} \text{soft}_{\gamma_1}(\mathbf{v}_{G_1}) \\ \vdots \\ \text{soft}_{\gamma_P}(\mathbf{v}_{G_P}) \end{bmatrix}, \tag{4.13}$$

where the soft thresholding is applied entrywise, and $\gamma_i = \frac{\tau}{1+\tau}||\mathbf{v}_{G_i}||_1, \ \forall i \in [\![1, P]\!]$.

For both the group and elitist cases, the Augmented Lagrangian writes:

$$\mathcal{L}(\mathbf{A}, \mathbf{U}, \mathbf{V}) = \frac{1}{2}||\mathbf{X} - \mathbf{B}\mathbf{A}||_F^2 + \lambda||\cdot||_{G,p,q}(\mathbf{U}) + \mathcal{I}_{\Delta_Q}(\mathbf{V}) + \frac{\rho}{2}||\mathbf{A} - \mathbf{U} - \mathbf{C}||_F^2 + \frac{\rho}{2}||\mathbf{A} - \mathbf{V} - \mathbf{D}||_F^2, \tag{4.14}$$

where $(p, q) = (2, 1)$ in the group case and $(p, q) = (1, 2)$ in the elitist case. The ADMM algorithm to minimize the AL is summarized in Algorithm 3.

Data: $\mathbf{X}$, $\mathbf{B}$
Result: $\mathbf{A}$
Initialize $\mathbf{A}$ and choose λ ;
while *ADMM termination criterion is not satisfied* **do**
  $\mathbf{A} \leftarrow (\mathbf{B}^\top\mathbf{B} + 2\rho\mathbf{I}_P)^{-1}(\mathbf{B}^\top\mathbf{X} + \rho(\mathbf{U} + \mathbf{V} + \mathbf{C} + \mathbf{D}))$;
  $\mathbf{U} \leftarrow \mathbf{prox}_{(\lambda/\rho)||\cdot||_{G,p,q}}(\mathbf{A} - \mathbf{C})$;
  $\mathbf{V} \leftarrow \mathbf{prox}_{\mathcal{I}_{\Delta_P}}(\mathbf{A} - \mathbf{D}) = \mathbf{proj}_{\Delta_P}(\mathbf{A}_{\mathcal{R}} - \mathbf{D})$;
  $\mathbf{C} \leftarrow \mathbf{C} + \mathbf{U} - \mathbf{A}$;
  $\mathbf{D} \leftarrow \mathbf{D} + \mathbf{V} - \mathbf{A}$;
end

Algorithm 3: ADMM to solve problem (4.11) in the case of the group or elitist penalties.

4.4.1.3 Fractional penalty case

The problem is more complex for the fractional mixed norm. As we have pointed out above, there is no proof that the mixed $\mathcal{L}_{G,1,q}$ norm with $q < 1$ satisfies the required properties for the ADMM to converge. However, in our problem, with an appropriate variable splitting scheme, we can express the fractional case for problem (4.11) as a $\mathcal{L}_q$ regularized constrained least squares problem.

Let us suppose for simplicity (and without loss of generality), that the rows of $\mathbf{A}$ and the columns of $\mathbf{B}$ have been sorted such that, in each pixel, the abundance vector has the following form:

$$\mathbf{a} = [a_{1,1}, a_{1,2}, \cdots, a_{1,m_{G_1}}, a_{2,1}, a_{2,2}, \cdots, a_{2,m_{G_2}}, \cdots, a_{G_P,1}, a_{G_P,2}, \cdots, a_{G_P,m_{G_P}}]^\top. \tag{4.15}$$

Recall that m_{G_i} is the number of instances of one of the P groups, in this case G_i.

The problem we want to solve is:

$$\arg\min_{\mathbf{a}} \frac{1}{2}||\mathbf{X}-\mathbf{B}\mathbf{A}||_F^2 + \lambda||\mathbf{A}||_{G,1,q}^q + \mathcal{I}_{\Delta_Q}(\mathbf{A}), \tag{4.16}$$

with $Q = \sum_{i=1}^{P} m_{G_i}$ the total number of signatures in the bundle $\mathbf{B} \in \mathbb{R}^{L\times Q}$. With the following variable splitting scheme, Eq. (4.16) can be rewritten as:

$$\begin{aligned} &\arg\min_{\mathbf{a}} \frac{1}{2}||\mathbf{X}-\mathbf{B}\mathbf{A}||_F^2 + \lambda||\mathbf{U}||_q^q + \mathcal{I}_{\Delta_Q}(\mathbf{V}) \\ &\text{s.t. } \mathbf{\Gamma}\mathbf{A} + \mathbf{\Lambda}_1\mathbf{U} + \mathbf{\Lambda}_2\mathbf{V} = \mathbf{0}_{(G+Q)\times N}, \end{aligned} \tag{4.17}$$

with $\mathbf{0}$. the zero matrix whose size is given in index. $\mathbf{\Gamma} = \begin{bmatrix} \mathbf{M} \\ \mathbf{I}_Q \end{bmatrix}$, where

$$\mathbf{M} = \begin{bmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & & \vdots & \vdots & & & \vdots & \vdots & & \vdots & \vdots & & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{G\times Q}, \tag{4.18}$$

The i^{th} row having m_{G_i} consecutive ones. $\mathbf{I}$. denotes the identity matrix whose size is in index (we only provide one dimension for brevity since the matrix is square).

Finally, we have $\mathbf{\Lambda}_1 = \begin{bmatrix} -\mathbf{I}_G \\ \mathbf{0}_{Q\times G} \end{bmatrix} \in \mathbf{R}^{(G+Q)\times G}$, and $\mathbf{\Lambda}_2 = \begin{bmatrix} \mathbf{0}_{G\times Q} \\ -\mathbf{I}_Q \end{bmatrix} \in \mathbb{R}^{(G+Q)\times Q}$. All these variables are defined such that: $\mathbf{M}\mathbf{A} = \mathbf{U} \in \mathbb{R}^{G\times N}$ and $\mathbf{A} = \mathbf{V}$. This way, we have reduced the optimization problem to a $\mathcal{L}_q$ regularized least squares problem, where the variable on which the fractional norm is applied is a vector whose entries are the $\mathcal{L}_1$ norms of the abundance coefficients in each group. Note that the new problem is equivalent to the original one only thanks to the nonnegativity constraint, which allowed us to turn the $\mathcal{L}_1$ norm into linear constraints. This way, the convergence of the ADMM for our nonconvex problem is, in theory, guaranteed [164], should we be able to compute exact updates for all the subproblems of the ADMM.

However, even after this simplification of the problem, an issue remains: there is no closed form expression or known algorithm (to the best of our knowledge) to compute exactly the shrinkage operator of the $\mathcal{L}_q$ quasinorm (to the power q) when $q < 1$, except when $q = \frac{1}{2}$ or $q = \frac{2}{3}$ [30]. Here, we prefer the term "shrinkage operator" to the term "proximal operator" because the latter is usually defined for convex functions only. In addition, this operator is a discontinuous function, because of the nonconvexity of the quasinorm [169]. This limits the applicability of proximal methods to solve this type of problems, and other types of algorithms (or of nonconvex sparsity inducing penalties) have been investigated in remote sensing (see e.g. [153] and references therein).

In our case, in order to be able to apply ADMM nonetheless, we need an explicit shrinkage operator. We resort to an approximate q-shrinkage operator $S_{q,\lambda}$, as defined in [166]:

$$\forall \mathbf{x} \in \mathbb{R}^n,\ S_{q,\lambda}(\mathbf{x})_i = \text{sign}(x_i)(|x_i| - \lambda^{2-q}|x_i|^{q-1})_+. \tag{4.19}$$

This operator reduces to the soft thresholding operator when $q = 1$ and to the *hard thresholding* operator when $q = 0$. The hard thresholding operator is closely related to the shrinkage operator of the $\mathcal{L}_0$ norm [166]. In addition, it can be shown([166], theorem II.4) that the operator of Eq. (4.19) is actually the exact shrinkage operator of a nonconvex function (which we will denote as f_q) with desirable properties: it is separable w.r.t. each entry of $\mathbf{x}$ (with $f_q(\mathbf{x}) = \sum_{i=1}^{n} g_q(x_i)$), and the function g_q is even, continuous, strictly increasing and concave for $x_i > 0$, differentiable everywhere except in 0, and satisfies the triangle inequality. This function behaves in a way similar to the absolute value for small values of its argument, and more like the absolute value taken to the power q for larger arguments (see [166] for a graphical representation). For $q = 1$, this penalty function is the absolute value, but in general, unfortunately, there is no analytical expression for it. This result is interesting because we have an explicit shrinkage operator with nice properties to use with any proximal algorithm, to the cost of having a regularizer without an explicit expression. Nevertheless, the convergence of the ADMM in the nonconvex case remains to be proven, since we replaced the $\mathcal{L}_q$ norm with another nonconvex penalty, which should itself satisfy the required properties of [164] in order to guarantee convergence.

Finally, the optimization problem we solve is:

$$\begin{aligned} &\arg\min_{\mathbf{a}} \frac{1}{2}||\mathbf{X} - \mathbf{B}\mathbf{A}||_F^2 + \lambda f_q(\mathbf{U}) + \mathcal{I}_{\Delta_Q}(\mathbf{V}) \\ &\text{s.t. } \boldsymbol{\Gamma}\mathbf{A} + \boldsymbol{\Lambda}_1\mathbf{U} + \boldsymbol{\Lambda}_2\mathbf{V} = \mathbf{0}_{(G+Q)\times N}. \end{aligned} \tag{4.20}$$

The AL writes:

$$\mathcal{L}(\mathbf{A}, \mathbf{U}, \mathbf{V}) = \frac{1}{2}||\mathbf{X}-\mathbf{B}\mathbf{A}||_F^2+\lambda f_q(\mathbf{U})+\mathcal{I}_{\Delta_Q}(\mathbf{V})+\frac{\rho}{2}||\mathbf{M}\mathbf{A}-\mathbf{U}-\mathbf{C}||_F^2+\frac{\rho}{2}||\mathbf{A}-\mathbf{V}-\mathbf{D}||_F^2, \tag{4.21}$$

and we can ADMM algorithm to minimize it is shown in Algorithm 4 (the approximate q-shrinkage is performed coordinate-wise).

Data: $\mathbf{X}$, $\mathbf{B}$
Result: $\mathbf{A}$
Initialize $\mathbf{A}$ and choose λ ;
while *ADMM termination criterion is not satisfied* **do**
  $\mathbf{A} \leftarrow (\mathbf{B}^\top\mathbf{B} + \rho\mathbf{M}^\top\mathbf{M} + \rho\mathbf{I}_Q)^{-1}(\mathbf{B}^\top\mathbf{X} + \rho\mathbf{M}^\top(\mathbf{U} + \mathbf{V}) + \rho(\mathbf{C} + \mathbf{D}))$;
  $\mathbf{U} \leftarrow S_{q,\lambda/\rho}(\mathbf{M}\mathbf{A} - \mathbf{C})$;
  $\mathbf{V} \leftarrow \mathbf{prox}_{\mathcal{I}_{\Delta_P}}(\mathbf{A} - \mathbf{D}) = \mathbf{proj}_{\Delta_P}(\mathbf{A}_{\mathcal{R}} - \mathbf{D})$;
  $\mathbf{C} \leftarrow \mathbf{C} + \mathbf{U} - \mathbf{M}\mathbf{A}$;
  $\mathbf{D} \leftarrow \mathbf{D} + \mathbf{V} - \mathbf{A}$;
end

Algorithm 4: ADMM to solve problem (4.20).

4.4.2 Results

We want to test the performance of the three sparsity inducing penalties on a synthetic dataset and a real one. We compared them to a simple unmixing chain (based on VCA and FCLSU) which does not include spectral variability, and a bundle approach using FCLSU on the whole dictionary, which does not take into account the group structure of the dictionary in the optimization. For all cases, the bundle dictionary is the same, and is extracted using the VCA algorithm 5 times on random subsets of the data, comprising 80% of the image pixels. The groups were created using spectral clustering (which is more robust than k-means) [162]. Note that for the FCLSU algorithm, we used as endmembers the centroids of the bundles, which is already better than extracting only once the endmembers with VCA on the whole dataset. The fractional norm we use is a $\mathcal{L}_{G,1,\frac{9}{10}}$ norm, in order to limit the nonconvexity of the objective, while making the ASC compatible with it. We stop the ADMM algorithms after 300 iterations, or when the relative variation between the abundances (measured in Frobenius norm) goes below 10^{-4}.

4.4.2.1 Synthetic data

In order to test the different group penalties, we design a synthetic dataset with spectral variability. To do that, we consider 15 materials, whose spectra were randomly chosen from the USGS spectral library. We generated variability by computing scaled versions of these endmembers. As we have already mentioned earlier in this manuscript, scaling factors can be considered to be a good model for illumination variations along the observed scene [121, 122]. Therefore, we generated 15 spatial scaling factor maps using mixtures of Gaussians, so that in the end each material is associated to a different local endmember in each pixel. The abundance maps were defined using Gaussian Random Fields. They satisfy the ANC and ASC, and are sparse, in the sense that only 3 or 4 materials are active in each pixel, out of the 15 considered. In addition, there is only one pure pixel in each abundance map. The performance on the synthetic data were evaluated using the following metric (termed $aRMSE$ for abundance Root Mean Squared Error) on the estimated abundances:

$$aRMSE = \frac{1}{N}\sum_{k=1}^{N}\sqrt{\frac{1}{P}\sum_{i=1}^{P}\left(a_{G_i,k_{true}} - \sum_{j=1}^{m_{G_i}} a_{G_i,j,k}\right)^2}, \tag{4.22}$$

where $a_{G_i,k_{true}}$ is the global true abundance for material G_i in pixel k, and $a_{G_i,j,k}$ denotes the abundance coefficient in pixel k, for instance j of group G_i. This metric measures the quality of the estimation of the abundances of the materials (but not of the abundances of each instance within each group, which are not available here anyway). For this dataset, for each algorithm, the regularization parameter λ and the barrier parameter ρ of the ADMM were empirically set to 10^{-2} and 10^{-1}, respectively.

The visual results are gathered in Fig. 4.16, where we show the abundance maps for 5 materials (out of 15) obtained by all algorithms. The quantitative results, as well as the

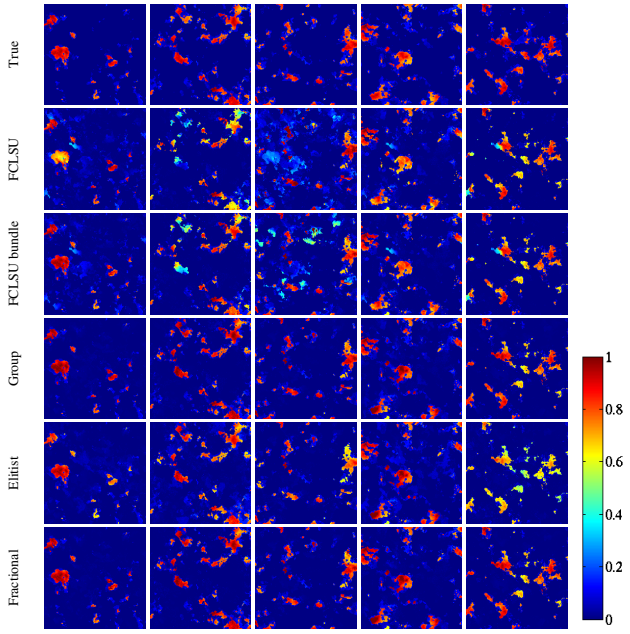


Figure 4.16: Abundance maps for all the compared algorithms for 5 of the 15 materials for the synthetic data. The color scale goes from 0 (blue) to 1 (red).

Algorithm	FCLSU	FCLSU Bundles	Group	Elitist	Fractional
aRMSE	0.0111	0.0067	0.0022	0.0043	0.0019
Running Time (s)	22	43	135	136	144

Table 4.1: Estimation error on the abundances, and running time for all the competing algorithms on the synthetic dataset. The best $aRMSE$ value is shown in red, and the second best is shown in blue.

running times of each algorithms are shown in Table 4.1. We also show in Fig. 4.17 the abundance matrices, displayed as images, to show the structure they induce on the abundance coefficients.

From the visual results, we can see that as expected, the standard linear SU chain, using VCA and FCLSU, performs rather poorly, since it is not able to take SV into account, which leads to significant estimation errors. All the other approaches make use of bundles, and consequently perform better both visually and quantitatively. When FCLSU is used to obtain the abundances of each instance of each group, the results are already much more satisfying, because this technique is already able to correct some of the wrongly estimated abundances of the batchless FCLSU approach. However, we can see that the results are far from perfect: the algorithm tends to estimate wrong mixtures of different materials. Indeed, it does not take the group structure into account, and it does not incorporate sparsity into the optimization problem. From both Fig. 4.16 and Table 4.1, we see that from the three tested algorithms here, the group and fractional penalty perform the best, while the elitist penalty is better than FCLSU, but worse than the other two. The elitist penalty assumes that few instances

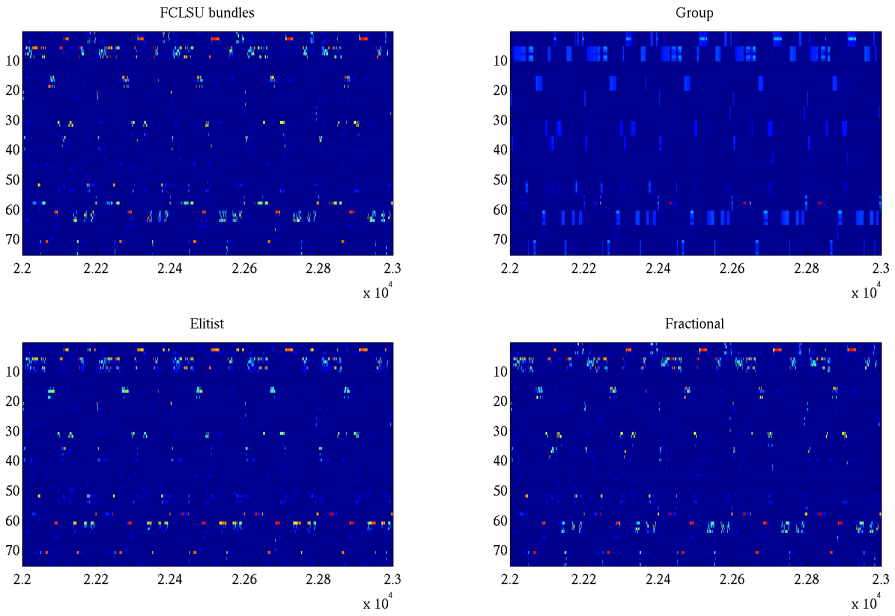
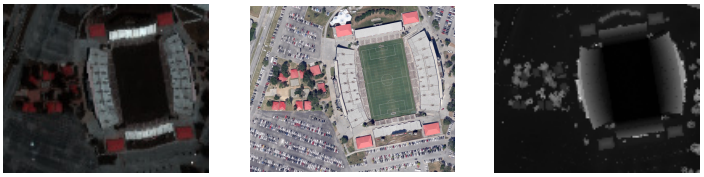


Figure 4.17: Abundance matrices for all the compared algorithms for 5 of the 15 materials for the synthetic data. A small part (1000) of the pixels are on the x-axis, and the groups, as well as their representatives are shown on the y-axis. The groups have been sorted for an easier visualization. Each group comprises 5 instances, which suggests that the endmember extraction and clustering were accurately performed.

are active within each group, but does not assume anything on the number of active groups. In practice, the mixture is dense over the groups, which explains the fact that this approach obtains less pure abundance maps than the group and fractional case. The group penalty, on the contrary, favors a sparse number of groups, but within each active group, it prefers a dense mixture. This phenomenon can clearly be seen in Fig. 4.17. If theoretically, with the geometrical interpretation of section 2.3.1 in mind, having a dense mixture within groups is an advantage, since it allows to explore more in detail the convex hull of the instances within a group, in practice we observe that the coefficients are often equally spread between the instances. The fractional penalty is able to slightly improve the results further, obtaining even sparser abundance maps than in the group case, for two reasons. The first is that it favors group sparsity as the group penalty does, but it also favors within group sparsity as the elitist penalty does. It is then able to combine the features of both other penalties at the same time. Furthermore, it makes use of a mixed $\mathcal{L}_{G,1,\frac{9}{10}}$ norm inducing sparser solutions than a classical $\mathcal{L}_1$ norm regularization, which disregards the group structure, and in addition conflicts with the ASC. Finally, the computational burden of all group sparsity strategies is heavier than the conventional SU chain, or even than the FCLSU algorithm used with the extracted bundles, but the running time is far from prohibitive. The group and elitist penalties have the same complexity, since only the proximal operator for the sparsity inducing norm changes, while the fractional case is slightly more expensive, since the constraints are more complex than those of the group and elitist cases.



(a) (b) (c)

Figure 4.18: A RGB representation of the Houston hyperspectral dataset (a). High spatial resolution color image acquired over the same area at a different time (b). Associated Lidar data (c), where black corresponds to 9.6m and white corresponds to 46.2m.

4.4.2.2 Real data

The real dataset we use in this Chapter is a subset of a hyperspectral image acquired over the University of Houston campus, Texas, USA, in June 2012. It was used in the 2013 IEEE GRSS Data Fusion Contest (DFC) [48]. The image comprises 144 spectral bands in the 380 nm to 1050 nm region, and comes with a LiDAR dataset acquired a day before over the same area, with the same spatial resolution (2.5 m). We are interested here in a $152 \times 108 \times 144$ subset of this image, acquired over Robertson stadium on the Houston Campus and its surroundings. Fig. 4.18 shows a RGB representation of the observed scene, as well as a high spatial resolution RGB image of the scene[2]. The estimated ID of the dataset using the Hysime algorithm [19] is 17, but we chose to consider 7 endmembers, to be able to reconstruct the data well while being able to keep easily visually interpretable results. The identified endmembers correspond to the following semantic classes: red metallic roofs, vertical structures surrounding the stadium, asphalt, healthy vegetation, an isolated red roof which was separated from the others, the concrete stands of the stadium and finally burnt vegetation. For this dataset, we keep the same setup as before, except that we (empirically) set the regularization parameter to λ to 0.5 for all algorithms. We will only evaluate the results visually in the absence of an available ground truth. The abundance maps estimated for all the algorithms are gathered in Fig. 4.19. As before, we also show the abundance matrices in Fig. 4.20.

First, we can see that the FCLSU results are not bad, except for the red metallic roofs surrounding the stadium, which should be pure, but are not here because of spectral variability induced by changing orientations of the facets of these roofs (as confirmed by the Lidar elevation image). Also, the concrete stands of the stadium are split into two abundance maps, whereas they are actually made of the same material (this much clearer from the high-resolution RGB image than from the RGB composition of the HSI). Here the different orientation of the stands with respect to the sun cause important spectral variability issues. The bundle approach, combined with FCLSU is already able to make the red roofs slightly purer, while gathering most of the stands in the same abundance map. In this case, the group penalty obtains comparable results, except that the global abundance of each material are much more spread across the various instances, contrary to FCLSU (as can be seen in

[2]Note that it was acquired at a different time with a few notable changes with respect to the dataset we are interested in (mainly parked cars).

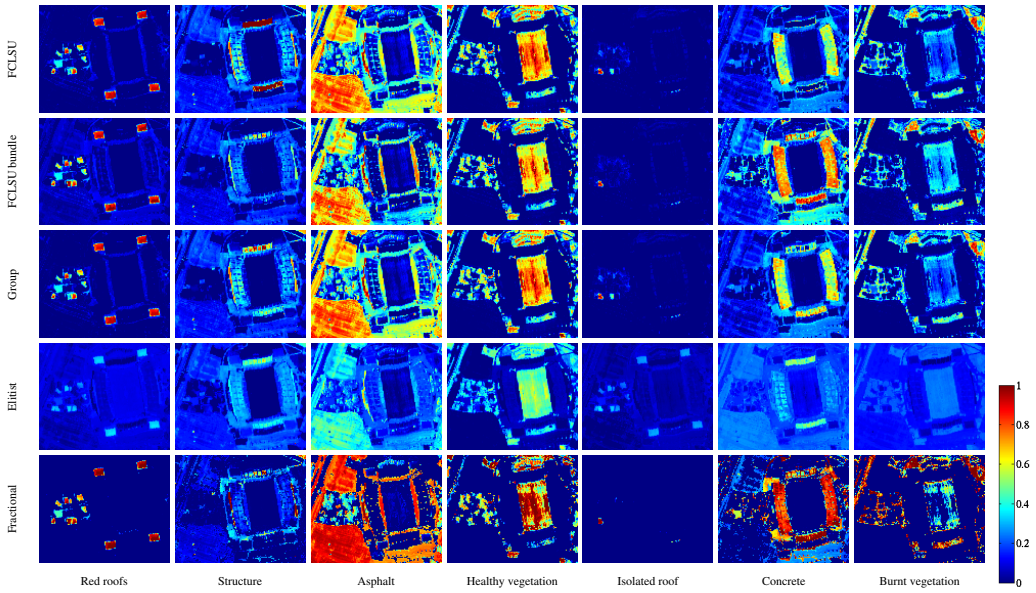


Figure 4.19: Abundance maps for all the compared algorithms for the Houston data. The color scale goes from 0 (blue) to 1 (red).

Fig. 4.20). The elitist penalty does not perform very well since it prefers a dense mixture over the groups, which does not allow to separate the materials well enough. The fractional penalty to produce satisfactory results, by having the sparsest and purest abundance maps for the red roofs and the concrete stands. It also allows to reduce the impact of the duplicate red roof endmember to a minimum by having a very sparse abundance map for it. We note however, that probably because of the nonconvexity of the optimization problem, the abundances are slightly less smooth than with the other approaches, especially for high values of the regularization parameter. This suggests that the results could be further improved by adding a spatial regularization term to the optimization.

4.5 Partial Conclusion

In this Chapter, we have proposed two different approaches in which sparsity can be interesting to handle spectral variability (SV). The first one is using the regularization path of a collaborative sparsity regularized local spectral unmixing (LSU) problem to get rid of the irrelevant endmember signatures extracted in each region of the Binary Partition Tree (BPT) after an overestimation of the intrinsic dimensionality (ID). We have shown that the approach was able to obtain more meaningful local abundance maps than when only FCLSU is used in each region to estimate the abundances. Future work on this part includes adapting the proposed methodology to be able to design a global joint endmember extraction and abundance estimation algorithm for a full HSI, instead of the local regions of a BPT, not unlike what has been done in [8] (this approach is mentioned in section 1.5.3). The major difference is

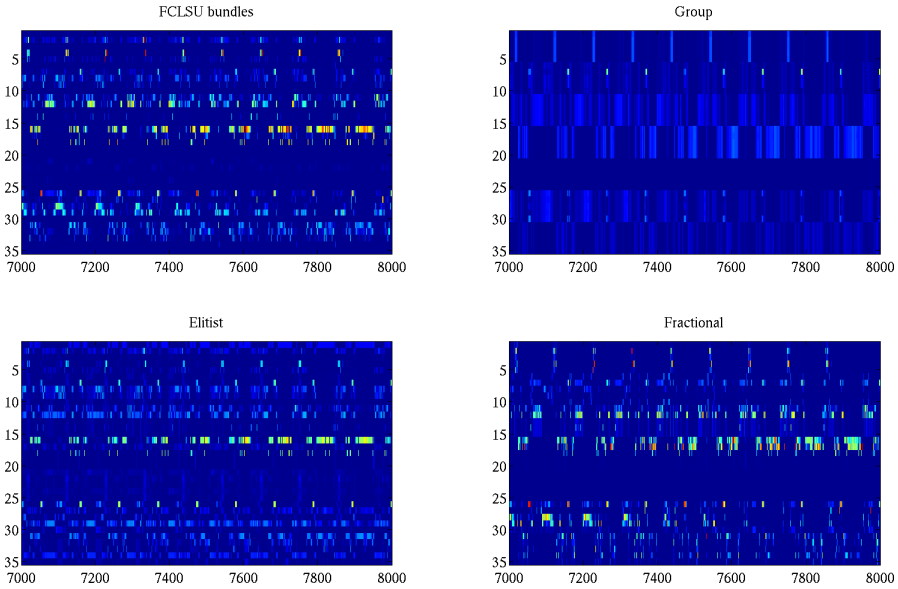


Figure 4.20: Abundance matrices for all the compared algorithms the Houston data. A small part of the pixels (1000) are on the x-axis, and the groups, as well as their representatives are shown on the y-axis. They have been sorted for an easier visualization. Each group comprises 5 instances, which suggests that the endmember extraction and clustering were accurately performed.

that the resulting algorithm would be completely unsupervised, needing no parameter to be tuned by the user (no regularization parameter, or threshold on the abundances to discard). Indeed, we are able to obtain the regularization path of the optimization problem, and the best model is selected by the Bayesian Information Criterion (BIC). Another advantage of such an algorithm is that would be entirely deterministic, unlike most endmember extraction algorithms (EEA) of the literature. A first step in this direction has already been taken in [57], where we extended the proposed algorithm to estimate the appropriate number of endmembers in a whole image from the set extracted by VCA. The LSU part could also be improved by using new region models and merging criteria during the construction of the BPT. For example, using an elevation model in the BPT construction could help defining physically more sound regions, where SV and nonlinear effects are likely to occur. Note that a multimodal LiDAR-Hyperspectral segmentation based on BPTs was recently proposed in [151]. Another possibility would be to integrate directly SV in the region model, e.g. using a SV oriented mixing model instead of the LMM.

The second approach uses different types of sparsity in the bundles approach to deal with SV. The idea is to use sparsity inducing norms which take into account the group structure of the bundle dictionary during the abundance estimation. We have tested three different types of social sparsity: a group penalty (enforcing group sparsity), and elitist sparsity (favoring intra-group sparsity) and a fractional sparsity (combining both in a single penalty), and have shown that these types of sparsity, especially the group and fractional ones, can be

interesting alternatives to simply using FCLSU for SU with bundles. On the theoretical part, a formal proof of convergence of the ADMM for the optimization problem we deal with in the fractional penalty case, using the approximate fractional shrinkage, remains to be found. Another research avenue could be to derive regularization paths for the optimization problems we deal with, in order to avoid having to tune the regularization parameters. Finally, since the extraction of the bundle has a critical impact on the results, introducing new more robust ways to obtain them is another interesting perspective: we can use spatially correlated subsamples instead of random ones (e.g. regions of the optimal partitions of the BPT), or even define the pool of endmembers as the results of the global collaborative unmixing algorithm suggested in the previous paragraph. Indeed, with a lower regularization parameter, several instances of the same material can be retained in the results to incorporate SV, instead of seeking only one signature for each material.

Part III

Extended Linear Mixing Model and applications

Chapter 5

An Extended Linear Mixing Model

Contents

5.1 Introduction

This chapter introduces a mixing model to handle spectral variability (SV), in particular caused by illumination effects and topography. The motivation of such a model is to extend the Linear Mixing Model (LMM) to a variant in which the sources are locally allowed to vary, through pixel and material dependent mappings $f_{pk} : \mathbb{R}^L \to \mathbb{R}^L$ which transform reference endmembers stored in a matrix $\mathbf{S}_0$ into local variants. Then the classical LMM of Eq. (1.3) rewrites:

$$\mathbf{x}_k = \sum_{p=1}^{P} a_{pk} f_{pk}(\mathbf{s}_{0p}) + \mathbf{e}_k. \tag{5.1}$$

Before commenting on the definition of the mappings, we note that the spectral bundles presented in section 2.3.1 can be interpreted in this framework. Indeed, without explicitly defining the functions transforming the references, having at our disposal several instances of each endmember boils down to knowing several possible outcomes of these mappings. In this light, machine learning approaches for spectral bundles can be seen as trying to learn

the mappings from training samples. If SV is to be explicitly modeled, the analytical form of the functions should be flexible and based on physical considerations, so that the additional variability parameters introduced can be physically interpreted. The variations on the spectra induced by illumination and the geometry of the scene affect all materials differently, but in a correlated way since they share one cause (the geometry is the same for all materials). Although as we will see that material specific properties have an influence, this physical phenomenon seems to be a good candidate to design a relatively general model, applicable to all possible materials. This phenomenon is often considered to be well approximated by scaled variants of the spectra [94, 121]. The use of the spectral angle as a metric to compare spectra also has the same rationale: it is insensitive to scalings and hence measures the dissimilarity in the shapes of the spectra, not their magnitude [94]. In the literature, to model shadow and brightness effects, a constant "shade" endmember is often considered, and its abundances are considered in the same way as for other materials [94]. However, using this trick amounts to scaling each pixel with the abundance of this shadow endmember, and then interpreting the remaining abundances for the other materials. In such a case, the ASC is not physically meaningful anymore since this endmember is not an actual material. In addition, to the best of our knowledge, the use of scalings to approximate brightness effects on the spectra has not been related to physical models yet.

5.2 Contributions

The contributions of this Chapter are as follows: First, in order to find a suitable definition for the mappings in Eq. (5.1), we start from the reflectance model of Hapke [73] and simplify it to make it tractable from a blind SU point of view. We obtain a new mixing model taking into account SV under the form of scaling factors affecting each endmember, in every pixel, that we term Extended Linear Mixing Model (ELMM). Then we design an optimization problem aimed at estimating the parameters of this model, adding the usual constraints and useful regularizations. We propose two algorithms to solve this optimization problem. We compare the results of the proposed approach to other techniques of the literature taking SV into account (most of which are reviewed in Chapter 2), on synthetic and real datasets. The explicit use of scaling factors in a mixing model was introduced in [159] and [53], though it had been already suggested in [121, 122]. Most of the results of this Chapter were first reported in [55].

5.3 From the Hapke model to the ELMM

In this section, we briefly describe the Hapke model and try to give some insight on how it models reflectance as a function of various physical parameters. Then we describe standard physical assumptions to simplify its analytical expression, and make a last assumption which motivates the ELMM we propose.

5.3.1 The Hapke model

We already alluded to the Hapke model in section 2.3.4. We present it here in more details. However, the complete analytical expressions of all the terms involved, and let alone a detailed derivation of the Hapke model are far beyond the scope of this thesis. We refer to [73] for the original derivation, and to [34] and [80] for introductions to the model. From a spectral unmixing point of view, this model is very hard to use because of the limited avaibility in practice of the material dependent photometric parameters and of the albedo spectra, and because of its complexity.

Reflectance, the physical quantity usually used to work with hyperspectral remote sensing images (after atmospheric correction of radiance units), is dependent on the geometry of the acquisition. Depending on the incidence and viewing angles, the measured reflectance can significantly differ. The reflectance of a material is also influenced by its photometry, that is to say the way light interacts with the material. Photometry can be modeled through some optical parameters (surface roughness, scattering behavior...) of the materials. We will briefly describe the photometric parameters involved in the model, but for a thorougher description of their physical and geological interpretations we refer to [63]. The albedo of material, contrary to its reflectance, is truly characteristic of the material and depends neither on the geometry of the scene nor on the photometry of the considered material. Note that we assume the mixture of the materials occurs at the macroscopic level, and hence we do not consider intimate mixing, which can also be explained by Hapke's model. Therefore, in this context, the LMM remains a valid assumption. The equations below are to be understood to be applied separately to each endmember, using its pure albedo spectrum (they have to be applied using the albedo value for each wavelength to obtain reflectance spectra). This defines local endmember variants in each pixel, which are then linearly mixed.

The local (i.e. in each pixel) geometry of the scene is determined by several factors (c.f. Fig. 5.1) [34]. The zenith is defined as the direction of the normal vector to the tangent plane to the surface observed. This means that depending on the topography, this plane can be different for each pixel. The angle between the zenith and the sun is called the sun zenith angle, or incidence angle, θ_0. The angle between the zenith and the sensor is called the view zenith angle, or emergence angle, θ. The angle between the sun and sensor directions (with the origin on the FOV of the current pixel) is called the phase angle g. It is practically constant throughout the observed scene, since the distances between the surface and the sun (in all cases) or the sensor (in most cases, except maybe with sensors mounted on UAVs) are far more important than the difference in elevation along the scene, and its spatial extent. Finally, the angle between the projection of the sun on the tangent plane and the projection of the sensor on the tangent plane is called the azimuthal angle ϕ. These four angles are then pixel-dependent and completely characterize the geometry of a pixel's acquisition. They are not completely independent since for instance, the phase angle can be recovered from the other three. Hapke's model can be expressed as [73, 80]:

$$\rho(\omega, \mu, \mu_0, g) = \frac{\omega}{4(\mu + \mu_0)}((1 + B(g))P(g) + H(\omega, \mu)H(\omega, \mu_0) - 1), \tag{5.2}$$

where ρ is the reflectance for a given wavelength range, $\mu = \cos(\theta)$, $\mu_0 = \cos(\theta_0)$, ω is the

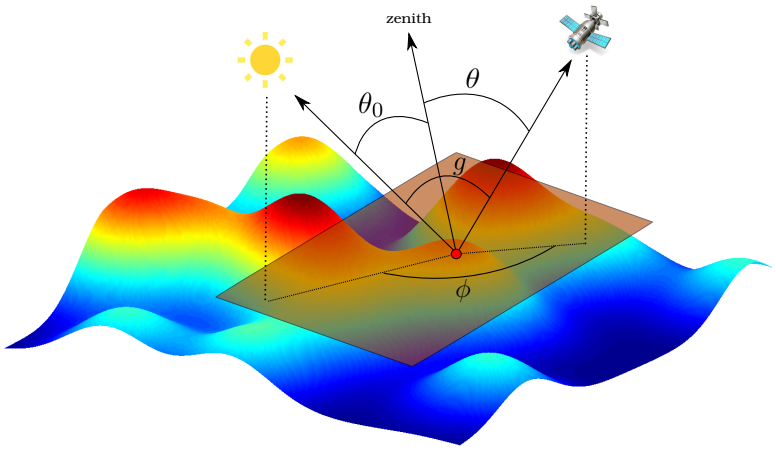


Figure 5.1: Acquisition angles for a given spatial location (red dot). The tangent plane at this point of the surface is in brown. The incidence angle is θ_0, the emergence angle is θ, and the angle between the projections of the sun and the sensor is the azimuthal angle, denoted as ϕ. g is the phase angle. θ_0 and θ are defined w.r.t. the zenith, which is defined locally (in each point of the observed surface) as the normal to the observed surface at this point.

single scattering albedo (SSA) of the material, P is the so-called phase function, which models the angular scattering distribution of the material, B is a function related to the opposition effect (brightening of the observed surface when the illumination comes from behind the sensor, i.e. for small g values), and H is the isotropic multiple scattering function. The functions B, P and H are additionally parametrized by photometric parameters of a material. For B, the parameters used are h and b_0, accounting for the angular width and the strength of the opposition surge. For the phase function P, the photometric parameters used are the asymmetry parameter of the scattering lobes b ($0 \leq b \leq 1$, higher values meaning narrower lobes and higher scattering intensity), and the backward scattering fraction c ($0 \leq c \leq 1$; $c < 0.5$ means that the material mainly backscatters the incoming light towards the incidence direction, and $c > 0.5$ means that the material has a predominantly forward scattering behavior). As examples of particular behaviors of the phase function, we can cite specular reflection, characterized by $b = 1$ and $c = 1$, or Lambertian (isotropic) scattering, characterized by $b = 0$ and $c = 0.5$. A refined version of the model taking into account the macroscopic roughness of the materials is also used:

$$\rho(\omega, \mu, \mu_0, \phi, g) = \frac{\omega}{4(\mu_e + \mu_{0e})} S(\mu, \mu_0, \phi)((1 + B(g))P(g) + H(\omega, \mu_e)H(\omega, \mu_{0e}) - 1), \quad (5.3)$$

where μ_{0e} and μ_e are the cosines of the modified incidence and emergence angles, because of the surface roughness. $S(\mu_0, \mu, \phi)$ is a shadowing function which reduces the total reflectance when surface roughness hides parts of the observed surface from the sensor, or shadows some fraction of the observed surface. This function and the modified incidence and emergence angles are parametrized by an angle $\bar{\theta}$ ($0° \leq \bar{\theta} \leq 45°$) accounting for the average roughness of the materials in the field of view of the sensor.

5.3.2 Simplifying assumptions

Here, using simplifying assumptions, we go from the general Hapke model of Eq. (5.3) to a special case of the ELMM presented in [161, 53, 55].

First, if we assume the surface of the materials to be smooth ($\bar{\theta} = 0$), then there is no shadowing effect, and the emergence and incidence angles are not modified. This is of course not so realistic an assumption in practice, but it considerably simplifies the model and makes it free from the roughness parameter which is not accessible in practice, let alone for each material in each pixel. Then $S(\mu_0, \mu, \phi) = 1$ and $\mu_{0e} = \mu_0$, and $\mu_e = \mu$. Then the model reduces to this of Eq. (5.2).

As explained in [80], assuming a Lambertian scattering; the phase function reduces to $P(g) = 1$. Besides, for Lambertian surfaces, there is no opposition surge ($h = 0$ and $b_0 = 0$). In any case, even for non Lambertian photometries, for large enough phase angles (in practice more than 5° [63]), the opposition effect is negligible and $B(g) \approx 0$ anyway.
The multiple isotropic scattering function H can be approximated by:

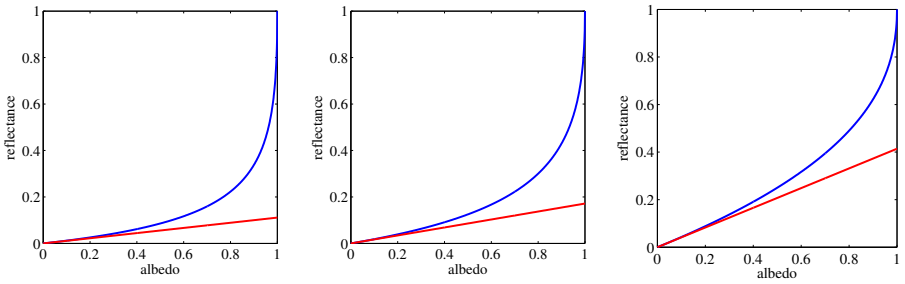
$$H(\omega, \mu) \approx \frac{1 + 2\mu}{1 + 2\mu(\sqrt{1-\omega})}, \tag{5.4}$$

and hence, incorporating all these assumptions, the model for relative (bidirectional) reflectance (relative to a case where $\omega = 1$) becomes:

$$\rho(\omega, \mu, \mu_0) = \frac{\omega}{(1 + 2\mu\sqrt{1-\omega})(1 + 2\mu_0\sqrt{1-\omega})}. \tag{5.5}$$

This expression is already much simpler than the full model of Eq. (5.3). The approximation eliminates all the photometric effects, in particular because of the Lambertian photometry assumption. The model is still material dependent, because the albedo spectrum depends on the material. The only other parameters left are the sun zenith angle, and the view zenith angle. Still, the model is still not suitable for a least squares estimate of its parameters for two reasons. The first is that the albedo spectrum is not available in practice. A workaround for this is to numerically invert the model (the full model for a more precise estimate) if all the parameters but the albedo are known in a pixel. In such a case, the reflectance-albedo relation is bijective. However, there is no simple way to assess the results of this method in practice, especially in real scenarios, and the incertitudes on the results could be very important. The principles of this strategy are applied to controlled lab measurements in [112]. The second reason is that the model is still relatively complex, highly nonlinear, especially for high albedos, and it its not identifiable when no parameters are known, since it is symmetric w.r.t. μ and μ_0. There would be no way to discriminate between the two angles in an estimation problem.

For small SSA values, this relationship is practically linear, while important nonlinearities appear for large albedo values. In Fig. 5.2, the function defined by Eq. (5.5) is plotted for three values of the acquisition angles. On the left is the case when both the sensor and the sun are at nadir. On the middle is a case where the sensor and the sun both make an



(a) (b) (c)

Figure 5.2: Reflectance plotted as a function of the albedo according to Eq. (5.5) (blue), and first order Taylor expansion in $\omega = 0$ (red), for three geometries ($\theta_0 = \theta = 0°$ (a), $\theta_0 = \theta = 45°$ (b), $\theta_0 = 90°$ and $\theta = 45°$ (c)).

angle of 45 degrees with respect to the normal to the surface, and on the right is a case with raking incident light ($\theta_0 = 90°$), and the sensor making an angle of 45 degrees with the nadir direction. Note that if both angles are equal to 90 degrees, the resulting reflectance equals the albedo. Here, because of these considerations, we propose to approximate further the relationship between albedo and reflectance by performing a first order Taylor expansion in 0 (in practice approximately valid for "small" albedos):

$$\rho(\omega, \mu, \mu_0) = \frac{\omega}{4\mu\mu_0 + 2\mu + 2\mu_0 + 1} + o(\omega). \tag{5.6}$$

The coefficient of the expansion only depends on the geometry of the acquisition. This means that it affects an albedo spectrum in the same way for any wavelength. Now let us assume that for a given material p, we have at our disposal a reference endmember $\mathbf{s}_{0p}$ (usually extracted from the data), with a geometry defined by the angles μ and μ_0. Then with the first order model of Eq. (5.6), we can write that for the representative of this endmember in pixel k, we have, for small albedos:

$$\mathbf{s}_{kp} \approx \frac{4\mu_k\mu_{k0} + 2\mu_k + 2\mu_{k0} + 1}{4\mu\mu_0 + 2\mu + 2\mu_0 + 1}\, \mathbf{s}_{0p} = \psi_k \mathbf{s}_{0p}. \tag{5.7}$$

From this equation, we see that now the link between the local representative of an endmember in a pixel and a reference signature for this material is a positive scaling factor incorporating the information about the geometry in the considered pixel. Note that we could have obtained the scaling factor model by approximating the albedo to reflectance relationship by a linear one in any other way (for instance by fitting a straight line to the whole curve, and not only considering a first order Taylor expansion for small albedos). Interestingly, with this approximation, we make the connection between the semi-empirical model of Hapke and the well known fact in the remote sensing community that illumination effects can be well approximated by scaling variations of the spectra (hence the frequent use of the SAM as a distance between spectra). Some works in the HSI processing community had already suggested that spectral variability could be well modeled as a scaling factor [122, 121], but we propose here to define a new spectral variability-based mixing model using this.

5.4 Model description

5.4.1 Model Formulation

The considerations of the previous sections lead us to define the mappings of Eq. (5.1) as $f_{pk}(\mathbf{s}_0) = \boldsymbol{\psi}_k \mathbf{s}_{0p}$, so that the model becomes:

$$\mathbf{x}_k = \sum_{p=1}^{P} a_{pk}\psi_k \mathbf{s}_{0p} + \mathbf{e}_k = \psi_k \sum_{p=1}^{P} a_{pk}\mathbf{s}_{0p} + \mathbf{e}_k = \psi_k \mathbf{S}_0 \mathbf{a}_k + \mathbf{e}_k = \mathbf{S}_0 \psi_k \mathbf{a}_k + \mathbf{e}_k. \tag{5.8}$$

The LMM is simply scaled in each pixel by a different nonnegative scaling factor. In order to estimate this scaling factor, we can take advantage of the nonnegative least squares, or partially Constrained Least Squares Unmixing(CLSU) problem, which was presented in section 1.3.3 (Eq. (1.6)) as a solution to estimate the abundances in a LMM framework, without the ASC, but only considering a priori the ANC. Indeed, we can show that if the model of Eq. (5.8) holds, then this algorithm, which assumes the data lie in a convex cone spanned by the endmembers, does not really estimate the abundances, but a factor incorporating SV. To see this, let us call for now $\hat{\boldsymbol{\alpha}}_k$ the quantity estimated in one pixel by CLSU (which, incidentally, is the Maximum Likelihood Estimator (MLE) for the product $\boldsymbol{\alpha}_k$, assuming the noise is Gaussian with an equal variance in each band):

$$\begin{aligned} \hat{\boldsymbol{\alpha}}_k = \arg\min_{\boldsymbol{\alpha}_k} \frac{1}{2}||\mathbf{x}_k - \mathbf{S}\boldsymbol{\alpha}_k||_2^2 \\ \text{s.t. } \boldsymbol{\alpha}_k \geq \mathbf{0}. \end{aligned} \tag{5.9}$$

Then by identifying $\hat{\alpha}_{kp}$ with $\psi_k a_{pk}$, we have:

$$\sum_{p=1}^{P} \hat{\alpha}_{pk} = \sum_{p=1}^{P} \psi_k a_{pk} = \psi_k \sum_{p=1}^{P} a_{pk} = \psi_k. \tag{5.10}$$

The right hand side of the equation is obtained by reintroducing the ASC, but on the *actual* abundances, rather than on the quantity estimated with CLSU, which absorbs SV effects. The abundances can then simply be estimated by $\mathbf{a}_k = \frac{1}{\psi_k}\boldsymbol{\alpha}_k$. This provides a very convenient way to estimate the parameters of the model using only nonnegative least squares, summing the estimated quantity over all materials to get the scaling factor, followed by a normalization step to obtain the abundances. We call this technique a Scaled (partially) Constrained Least Squares Unmixing (S-CLSU).

However, in practice we are going to allow the scaling factor to vary for each material:

$$\mathbf{x}_k = \sum_{p=1}^{P} a_{pk}\psi_{pk}\mathbf{s}_{0p} + \mathbf{e}_k = \mathbf{S}_0 \boldsymbol{\psi}_k \mathbf{a}_k + \mathbf{e}_k = \mathbf{S}_0(\text{diag}(\boldsymbol{\psi}_k) \odot \mathbf{a}_k) + \mathbf{e}_k, \tag{5.11}$$

where the ψ_{kp} are now pixel and material dependent scaling factors, $\boldsymbol{\psi}_k \in \mathbb{R}^{P\times P}$ is a diagonal matrix, containing the scaling factors for each material on its diagonal ($\text{diag}(\boldsymbol{\psi}_k)$ is a vector

containing its diagonal elements), and $\odot$ is the Schur-Hadamard (termwise) product between two matrices of the same size. The scaling factors can also be rearranged into a matrix $\mathbf{\Psi} \in \mathbb{R}^{P \times N}$, which has the same size as the abundances. This allows the model (5.11) to be rewritten globally for the whole image:

$$\mathbf{X} = \mathbf{S}_0(\mathbf{\Psi} \odot \mathbf{A}) + \mathbf{E}. \tag{5.12}$$

We will refer to this equation as the Extended Linear Mixing Model (ELMM). Of course, the ELMM reduces to the LMM if all the scaling factors are equal to 1 (no variability w.r.t. the reference endmembers).

The main reason behind the introduction of a scaling factor for each pixel and material is that it will make the model more flexible, allowing to model material dependent SV, be it related to material dependent photometric phenomena or more pragmatically to intrinsic variability of each material. Indeed, the scaling factor is able to capture other more complex variations of the spectra by compensating the scale of the reference signature, even if there is a modeling error (that is if the variability cannot be solely explained by scaling variations). The other reason for allowing the scaling factors to be material dependent is related to the interpretability of the results. Indeed, in the original formulation, there is one scaling factor per pixel, which allows to obtain a spatial map of scaling factors. However, the fact that this map is the same for all materials can make the results harder to interpret. For pure pixels, this is not really a problem since the scaling factor will be related to the active material, but for mixed pixel, there is no way to tell which material is subject to SV, and in which proportions.

The interest of introducing one scaling factor for each material also becomes clear when we have a look to the geometric interpretation of the model. The model enjoys a simple geometric interpretation, as can be seen in Fig. 5.3 (in a case where there are three endmembers). The data points are assumed to lie in a cone spanned by the reference endmembers. The scaling factors, combined with the ASC and ANC, constrain each pixel to lie in a simplex whose vertices are variants of the reference endmembers, situated on straight lines joining the origin and each of the reference endmembers, thus defining the simplex orientation in the cone. For the S-CLSU version of the model, the parametrization with the scaling factor is the same for all endmembers. This amounts to assume that SV affects all materials in the exact same way (w.r.t. the reference endmembers). As a consequence, it will restrict the possible simplex configurations of the simplex. In that case, the pixelwise simplices will be linked to the LMM simplex through a homothetic transformation, whose homothetic center is the origin. Introducing a different scaling factor for each material will of course mostly impact mixed pixels, and we expect the scaling factor of S-CLSU and the refined version to be similar in pure or close to pure pixels. In order to further motivate the model advocated by Eq. (5.12), we manually selected two pure pixels of the same material in the image of the Houston dataset used in the experiments of section 5.6.2, and already used in this thesis in section 4.4.2.2 (Fig. 4.18). These pixels are part of the red rooftop on the northwestern part of the football field. The two pixels are part of two distinct facets of this pyramidal roof, which are very differently lit, since their orientation w.r.t. the sun is different. The spectral signatures are shown in Fig. 5.4. We performed a least squares regression in order to approximate the blue

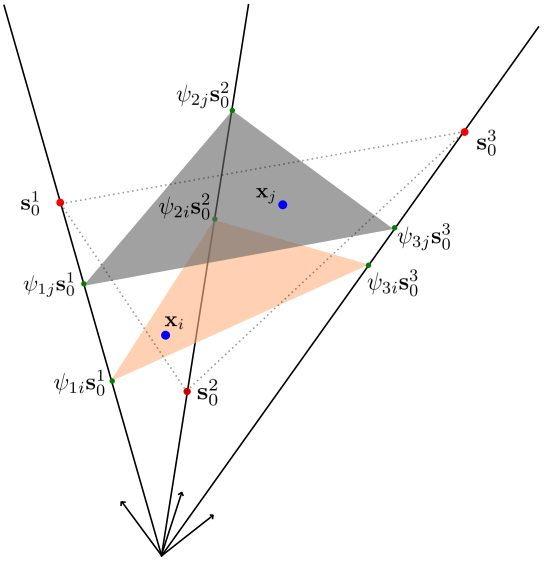


Figure 5.3: Geometric interpretation of the ELMM in the case of three endmembers. In blue are two data points, in red are the reference endmembers and in green are the scaled versions for the two considered pixels. The simplex used in the LMM is shown in dashed lines.

spectrum $\mathbf{x}_1$ by a scaled version of the red spectrum $\mathbf{x}_2$:

$$\hat{\psi} = \arg\min_{\psi} ||\mathbf{x}_1 - \psi\mathbf{x}_2||_2^2. \tag{5.13}$$

The result of this regression (i.e. $\hat{\psi}\mathbf{x}_2$) is shown in green in Fig. 5.4. The obtained value for the scaling factor is $\hat{\psi} = 1.108$. We can see that the fit is almost perfect, meaning that the variability between these two pure pixels can be approximated very well by a scaling factor, which is confirmed by the very high Pearson correlation coefficient between the original blue spectrum and its green regression ($r = 0.9994$).

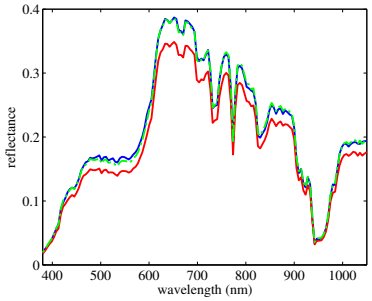


Figure 5.4: Two pixels on different facets of the same red roof in the image of Fig 5.17 (red and blue). The result of the linear regression of the red spectrum on the blue one, using only a scaling factor (dashed green).

5.4.2 Solving the ambiguity between the abundances and the scaling factors

The introduction of material dependent scaling factors refines the model, but as written in Eq. (5.12), it is easy to see that the model is not identifiable, since inverting the roles of $\mathbf{A}$ and $\mathbf{\Psi}$ leaves the equation unchanged. There is an ambiguity between the scaling factors and the abundances. The ASC is the only difference between the two variables, but it is not sufficient to solve this problem completely, even though it has been shown to help calibrating similar models in other applications [16]. Trying to estimate the model parameters using a MLE would result in the abundance update being inversely proportional to the scaling factors, and vice versa. This would be a problem in low abundance areas, where there is very little information to estimate SV in addition to the abundances, and would result in numerical instability.

In order to solve the issue, we introduce a new set of variables: pixel dependent endmember matrices $\mathbf{S}_k \in \mathbb{R}^{L\times P}$. We also define the third order tensor $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{L\times P\times N}$ such that $\boldsymbol{\mathcal{S}} \equiv \{\mathbf{S}_k\}_{k=1,\dots,N}$, and define the following cost function:

$$\mathcal{J}(\mathbf{A},\boldsymbol{\mathcal{S}},\mathbf{\Psi}) = \frac{1}{2}\sum_{k=1}^{N}\left(||\mathbf{x}_k - \mathbf{S}_k\mathbf{a}_k||_2^2 + \lambda_S||\mathbf{S}_k - \mathbf{S}_0\boldsymbol{\psi}_k||_F^2\right) + \mathcal{I}_{\Delta_P}(\mathbf{A}) + \mathcal{I}_{\mathbb{R}_+^{L\times P\times S}}(\boldsymbol{\mathcal{S}}), \quad (5.14)$$

where λ_S is a regularization parameter. The first term is a simple usual data fit term, very similar to the LMM, with the notable exception that the endmember matrix is pixel-dependent. The second term is a regularization forcing the local endmembers to be close the the scaled variations of reference endmembers advocated by the ELMM. The last two terms simply enforce the constraints on the abundances (the indicator function of the simplex has to be understood as being applied separately to each column of the abundance matrix) and on the sources, which have to remain nonnegative. The interest of introducing the second term is twofold: it allows the actual local endmembers to drift away from the exact ELMM of Eq. (5.12) if needed, in particular if SV cannot only be explained by scaling factors. Besides, on a more algorithmic point of view, the additional variables allow us to decouple the scaling factors from the abundances, thus getting rid of the ambiguity between them.

From a statistical point of view, minimizing the proposed criterion w.r.t. the three blocks of variables can be interpreted as the MAP estimator for $\boldsymbol{\mathcal{S}}$, $\mathbf{\Psi}$ and $\mathbf{A}$, with Gaussian i.i.d. spectrally white Gaussian noise, a Gaussian prior on the endmembers, whose mean is $\mathbf{S}_0\boldsymbol{\psi}_k$ for pixel k (and whose variance is linked to the regularization parameter λ_S), with a uniform prior on the nonnegative orthant for the scaling factors, and finally a uniform prior on the simplex for the abundances. In this sense, regarding the classification of the approaches dealing with SV of section 2.3, minimizing this criterion is a hybrid approach, halfway between a computational model (introducing pixel dependent endmembers $\mathbf{S}_k \in \mathbb{R}^L$, free to fluctuate around references) and the estimations of physics related parameters from a specific model (the scaling factors).

5.4.3 Regularization Terms

In order to enforce desirable properties to the parameters, we incorporate two additional regularization terms to the objective function, one for the abundances and one for the scaling factors, each weighted by a regularization parameter:

$$\begin{aligned}\mathcal{J}(\mathbf{A},\boldsymbol{\mathcal{S}},\boldsymbol{\Psi}) =& \frac{1}{2}\sum_{k=1}^{N}\left(||\mathbf{x}_k-\mathbf{S}_k\mathbf{a}_k||_2^2+\lambda_S||\mathbf{S}_k-\mathbf{S}_0\boldsymbol{\psi}_k||_F^2\right)+\mathcal{I}_{\Delta_P}(\mathbf{A})+\mathcal{I}_{\mathbb{R}_+^{L\times P\times N}}(\boldsymbol{\mathcal{S}})\\ &+\lambda_{\mathbf{A}}\mathcal{R}_{\mathbf{A}}(\mathbf{A})+\lambda_{\boldsymbol{\Psi}}\mathcal{R}_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}).\end{aligned} \tag{5.15}$$

As we have seen in Chapter 1, the abundances often exhibit strong spatial correlations, but there can be abrupt discontinuities in their spatial distributions. Similarly, the scaling factors are likely to be spatially coherent to some extent, since the scaling factors are connected to topographic information. Hence, we define at Total Variation (TV) [136] on the abundances:

$$\mathcal{R}_{\mathbf{A}}(\mathbf{A})=TV(\mathbf{A})=\sum_{p=1}^{P}\sum_{k=1}^{N}\sqrt{\mathcal{H}_h(\mathbf{A})_{pk}^2+\mathcal{H}_v(\mathbf{A})_{pk}^2}=||\mathcal{H}(\mathbf{A})||_{2,1,1}, \tag{5.16}$$

where $\mathcal{H}_h : \mathbb{R}^{P\times N}\to\mathbb{R}^{P\times N}$ and $\mathcal{H}_v : \mathbb{R}^{P\times N}\to\mathbb{R}^{P\times N}$ are linear operators computing the horizontal and vertical gradients (first order derivatives) of each band of the image, respectively. The operator $\mathcal{H} : \mathbb{R}^{P\times N}\to\mathbb{R}^{P\times N\times 2}$ computes the complete gradient for each entry of $\mathbf{A}$. It operates in the same way for all materials. The three level mixed $\mathcal{L}_{p,q,r}$ norm is defined for a third order tensor $\mathcal{T}\in\mathbb{R}^{P\times N\times Q}$ and values of $p,q,r\geq 1$ or equal to infinity (taking limits), as:

$$||\mathcal{T}||_{p,q,r}=\left(\sum_{i=1}^{P}\left(\sum_{j=1}^{N}\left(\sum_{k=1}^{Q}|t_{ijk}|^p\right)^{\frac{q}{p}}\right)^{\frac{r}{q}}\right)^{\frac{1}{r}}. \tag{5.17}$$

The TV term then amounts to sum the $\mathcal{L}_2$ norm of the gradient in all pixels and for all materials, which is expressed by the mixed $\mathcal{L}_{2,1,1}$ three-level norm. In this formulation, the regularization of Eq. (5.16) is an isotropic TV term. Several variants can be defined for the TV, using an anisotropic version $\mathcal{R}_{\mathbf{A}}(\mathbf{A})=||\mathcal{H}(\mathbf{A})||_{1,1,1}=||\mathcal{H}_h(\mathbf{A})||_{1,1}+||\mathcal{H}_v(\mathbf{A})||_{1,1}$ (which as less attractive properties but can be simpler to minimize), or more refined vectorial total variations [67], possibly using different combinations of three level mixed norms [5]. which can have nice properties but are harder to handle. Finally, a Tikhonov-like (without the square) spatial regularization can be used by simply changing the norm, using $\mathcal{R}_{\mathbf{A}}(\mathbf{A})=||\mathcal{H}_h(\mathbf{A})||_{2,1}+||\mathcal{H}_v(\mathbf{A})||_{2,1}$. This version has less interesting edge preserving properties and its proximal operator is not cheaper to compute than that a regular TV, but the convergence rate is faster in practice. It is used in the results of [55]. For more information about gradient operators and TV, see Appendix B.

For the scaling factors, we chose a differentiable spatial regularization out of simplicity:

$$\mathcal{R}_{\boldsymbol{\Psi}}(\boldsymbol{\Psi})=\frac{1}{2}(||\mathcal{H}_h(\boldsymbol{\Psi})||_F^2+||\mathcal{H}_v(\boldsymbol{\Psi})||_F^2), \tag{5.18}$$

Here, the gradient operators transform matrices into matrices of the same size, but they act independently on each column of the $\boldsymbol{\Psi}$ matrix. With these additional regularization terms, solving the optimization problem:

$$\{\hat{\mathbf{A}}, \hat{\boldsymbol{\mathcal{S}}}, \hat{\boldsymbol{\Psi}}\} = \underset{\mathbf{A}, \boldsymbol{\mathcal{S}}, \boldsymbol{\Psi}}{\operatorname{argmin}} \; \mathcal{J}(\mathbf{A}, \boldsymbol{\mathcal{S}}, \boldsymbol{\Psi}) \tag{5.19}$$

can still be interpreted as computing a MAP estimate of $\mathbf{A}$, $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\Psi}$ with similar hypotheses as before, except for the abundances which are now associated to a TV prior on the simplex and the scaling factors are associated to a spatially correlated prior.

5.5 Optimization

In this section, we describe how we are going to solve the optimization problem (5.19). This problem is very challenging, for several reasons. First, it can be seen as a NMF problem with three blocks of variables to optimize. Hence, it is not convex w.r.t. the three blocks simultaneously, but it is w.r.t. to each of the individual blocks (regardless of the choice of the regularization chosen for the abundances among those presented, and also regardless of the definition of the linear gradient operators). A triconvex objective function suggests the use of alternating minimizations w.r.t. each block of variables, in order to find a stationary point of the objective function and to obtain a local minimum.

Since the problem is not convex with possibly many local minima, the initialization of the variables is important. Fortunately, in our case, we have seen that S-CLSU is a suitable initialization since it is able to provide a nice initial guess of the abundances and scaling factors in a reasonable amount of time.

In the following sections, we are going to present two algorithms to solve problem (5.19). The first one is based on an alternating minimization, using ADMM for the abundances, and the other one is based on a Coordinate Descent (CD) scheme, using a primal-dual algorithm for the abundances update (see Appendix A for more details on these algorithms).

5.5.1 Alternating Least Squares (ALS) algorithm

In the Alternating Least Squares algorithm, we will alternatively solve the subproblem associated with one block of variables, while keeping the other blocks constant. The outline of the ALS algorithms is given in Algorithm 5. The iterations terminate when the relative variations (measured using Frobenius norms) between consecutive iterates of $\mathbf{A}$, $\boldsymbol{\mathcal{S}} = \{\mathbf{S}_k\}$ and $\boldsymbol{\Psi}$ are below three tolerances ϵ_A, ϵ_S and ϵ_Ψ, respectively. The convergence to the minimum of each convex subproblem is guaranteed. However, the ALS strategy does not theoretically guarantee that we obtain a stationary point of the objective function. However, in practice, the algorithm does converge to a local minimum.

Data: $\mathbf{X}$, $\mathbf{S}_0$
Result: $\hat{\boldsymbol{\mathcal{S}}}$, $\hat{\boldsymbol{\Psi}}$, $\hat{\mathbf{A}}$
Initialize $\boldsymbol{\mathcal{S}}$, $\boldsymbol{\Psi}$, $\mathbf{A}$ and choose λ_S, λ_Ψ and $\lambda_A \geq 0$;
while *ALS termination criterion is not satisfied* **do**
  $\boldsymbol{\mathcal{S}} \leftarrow \underset{\boldsymbol{\mathcal{S}} \geq \mathbf{0}}{\arg\min}\ \mathcal{J}(\mathbf{A}, \boldsymbol{\mathcal{S}}, \boldsymbol{\Psi})$;
  $\boldsymbol{\Psi} \leftarrow \underset{\boldsymbol{\Psi} \geq \mathbf{0}}{\arg\min}\ \mathcal{J}(\mathbf{A}, \boldsymbol{\mathcal{S}}, \boldsymbol{\Psi})$;
  $\mathbf{A} \leftarrow \underset{\mathbf{A} \geq \mathbf{0}}{\arg\min}\ \mathcal{J}(\mathbf{A}, \boldsymbol{\mathcal{S}}, \boldsymbol{\Psi})$;
end

Algorithm 5: ALS scheme to find a local minimum of (5.19).

5.5.1.1 Optimization w.r.t. $\boldsymbol{\mathcal{S}}$

Rewriting the terms of Eq. (5.19) depending on $\boldsymbol{\mathcal{S}}$, we have to solve:

$$\boldsymbol{\mathcal{S}} = \underset{\boldsymbol{\mathcal{S}} \geq \mathbf{0}}{\arg\min}\ \frac{1}{2} \sum_{k=1}^{N} \left(||\mathbf{x}_k - \mathbf{S}_k \mathbf{a}_k||_F^2 + \lambda_S ||\mathbf{S}_k - \mathbf{S}_0 \boldsymbol{\psi}_k||_F^2 \right). \tag{5.20}$$

This problem is completely separable over the N pixels, and has a closed form solution which can be computed separately for each of them. By nulling the gradient of Eq. 5.20 w.r.t. $\mathbf{S}_k$, we get:

$$(\mathbf{S}_k \mathbf{a}_k - \mathbf{x}_k)\mathbf{a}_k^\top + \lambda_S (\mathbf{S}_k - \mathbf{S}_0 \boldsymbol{\psi}_k) = \mathbf{0}. \tag{5.21}$$

After some algebra, we obtain the update rule for $\mathbf{S}_k$:

$$\mathbf{S}_k \leftarrow (\mathbf{x}_k \mathbf{a}_k^\top + \lambda_S \mathbf{S}_0 \boldsymbol{\psi}_k)(\mathbf{a}_k \mathbf{a}_k^\top + \lambda_S \mathbf{I}_P)^{-1}, \tag{5.22}$$

The solution $\mathbf{S}_k$ is then projected onto the nonnegative orthant $\mathbb{R}_+^{L \times P}$ by thresholding the negative entries to 0. This thresholding is not useful in pratice, due to the Gaussian prior for the endmembers around (nonnegative) scaled reference endmembers.

5.5.1.2 Optimization w.r.t. $\boldsymbol{\Psi}$

Rewriting the terms of the criterion (5.19) depending only on $\boldsymbol{\Psi}$, we get:

$$\boldsymbol{\Psi} = \underset{\boldsymbol{\Psi}}{\arg\min}\ \frac{\lambda_S}{2} \sum_{k=1}^{N} ||\mathbf{S}_k - \mathbf{S}_0 \boldsymbol{\psi}_k||_F^2 + \frac{\lambda_\Psi}{2} (||\mathcal{H}_h(\boldsymbol{\Psi})||_F^2 + ||\mathcal{H}_v(\boldsymbol{\Psi})||_F^2). \tag{5.23}$$

First, let us remark that if there is no spatial regularization (*i.e.* $\lambda_\Psi = 0$), then there is a simple closed form update in each pixel, which guarantees the nonnegativity of the scaling factors:

$$\hat{\psi}_{pk} \leftarrow \frac{\mathbf{s}_{0p}^\top \mathbf{s}_{pk}}{\mathbf{s}_{0p}^\top \mathbf{s}_{0p}}. \tag{5.24}$$

When $\lambda_\psi \neq 0$, a way to simplify the problem is to see $\boldsymbol{\mathcal{S}} \equiv \{\mathbf{S}_k\}$ as an $L \times N \times P$ cube, with P slices of size $L \times N$ corresponding to the source matrices in every pixel. Using this description, we can rewrite Eq. (5.23) in a way that is separable w.r.t. the different materials:

$$\boldsymbol{\Psi} = \arg\min_{\boldsymbol{\Psi}} \frac{1}{2} \sum_{p=1}^{P} \Big(\lambda_S ||\mathbf{S}^p - \mathbf{s}_0^p (\boldsymbol{\psi}^p)^\top ||_F^2 + \lambda_\Psi (||\mathcal{H}_h(\boldsymbol{\psi}^p)||_2^2 + ||\mathcal{H}_v(\boldsymbol{\psi}^p)||_2^2) \Big), \tag{5.25}$$

where $\mathbf{S}^p$ is a $L \times N$ slice of the cube, $\mathbf{s}_0^p$ is a column of $\mathbf{S}_0$ (representing one reference endmember). $\boldsymbol{\psi}^p$ is the p^{th} column of $\boldsymbol{\Psi}$ (a $N \times 1$ vector containing the scaling factors for all the pixels for one material). By nulling the gradient of the expression (5.25) w.r.t. $\boldsymbol{\psi}^p$, we get:

$$\lambda_S \boldsymbol{\psi}^p \mathbf{s}_0^{p\top} \mathbf{s}_0^p + \lambda_\psi (\mathbf{H}_h^\top \mathbf{H}_h + \mathbf{H}_v^\top \mathbf{H}_v) \boldsymbol{\psi}^p = \lambda_S \mathbf{S}^{p\top} \mathbf{s}_0^p. \tag{5.26}$$

In all generality, if s_0^p was a matrix, we would have a Sylvester equation to solve, which has to be done numerically and which is very costly. Fortunately, here $s_0^p \in \mathbb{R}^L$ is a column vector, and thus the quantity $\mathbf{s}_0^{p\top} \mathbf{s}_0^p$ is a scalar. Therefore, we can factor this scalar on the left of $\boldsymbol{\psi}^p$, and obtain an expression which can be factorized on the left. Finally, the update for $\boldsymbol{\psi}^p$ is:

$$\boldsymbol{\psi}^p \leftarrow ((\lambda_S \mathbf{s}_0^{p\top} \mathbf{s}_0^p) \mathbf{I}_N + \lambda_\psi (\mathbf{H}_h^\top \mathbf{H}_h + \mathbf{H}_v^\top \mathbf{H}_v))^{-1} (\lambda_S \mathbf{S}^{p\top} \mathbf{s}_0^p). \tag{5.27}$$

We use the matrix representations of the linear gradient operators here. The $N \times N$ matrix inversion is intractable as such in most cases, but as the matrix to invert is circulant, the update can be very efficiently computed in the Fourier domain (assuming periodic boundaries for each $\boldsymbol{\psi}^p$ image) by:

$$\boldsymbol{\psi}^p \leftarrow \mathcal{F}^{-1} \left(\frac{\mathcal{F}(\lambda_S \mathbf{S}^{p\top} \mathbf{s}_0^p)}{(\lambda_S \mathbf{s}_0^{p\top} \mathbf{s}_0^p) \mathbb{1}_{m \times n} + \lambda_\Psi (|\mathcal{F}(\mathbf{h}_h)|^2 + |\mathcal{F}(\mathbf{h}_b)|^2)} \right), \tag{5.28}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Discrete 2D Fourier Transform and its inverse, m and n are the spatial dimensions of the image, such that $m \times n = N$, and $\mathbf{h}_h$ and $\mathbf{h}_v$ are convolution masks for the gradient operators. More details on solving circulant linear systems can be found in Appendix B.

5.5.1.3 Optimization w.r.t. A

The optimization problem w.r.t. $\mathbf{A}$ is:

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}} \frac{1}{2} \sum_{k=1}^{N} ||\mathbf{x}_k - \mathbf{S}_k \mathbf{a}_k||_2^2 + \lambda_A (||\mathcal{H}_h(\mathbf{A})||_{1,1} + ||\mathcal{H}_v(\mathbf{A})||_{1,1}) + \mathcal{I}_{\mathbb{R}_+^{P \times N}}(\mathbf{A}) + \boldsymbol{\mu}^\top (\mathbf{A}^\top \mathbb{1}_P - \mathbb{1}_N). \tag{5.29}$$

In order to take into account the ASC constraint (an equality) constraint, we have replaced the indicator function of the simplex in Eq. (5.15) by a Lagrangian, introducing a vector $\boldsymbol{\mu} \in \mathbb{R}^N$ of Lagrange multipliers. The ANC is enforced through the indicator function of the nonnegative orthant. Here, we are using the anisotropic TV for simplicity. This problem is

neither separable w.r.t. the pixels nor to the different endmembers, and it is not differentiable due to the presence of the $\mathcal{L}_{1,1}$ norm. There are several nondifferentiable terms, which suggests the use of the ADMM. Here we use the scaled version of the ADMM [22]. It will allow us to decompose the hard problem of Eq. (5.29) into iterations of a sequence of easier subproblems with closed form solutions. For more details on the ADMM, see Appendix A. In addition, by an appropriate choice of split variables (namely the definition of $\mathbf{V}_1$, see below), it will allow us to decouple the optimization in the spectral domain (related to the term in which the endmembers appear) to the optimization in the spatial domain (related to the terms in which the gradient operators appear), in a way similar to [88]. Note that by removing the regularization term $\mathcal{R}_{\mathbf{A}}(\mathbf{A})$ of Eq. (5.16), but keeping the ANC and the ASC, the problem becomes the simple FCLSU and can be solved separately in each pixel using for instance the algorithm of [76].

We introduce the splitting variables $\mathbf{V}_i$, $i = 1, \cdots, 4$, and express the problem of Eq. (5.29) as:

$$\begin{aligned}\hat{\mathbf{A}} =& \arg\min_{\mathbf{A}} \frac{1}{2}\sum_{k=1}^{N} ||\mathbf{x}_k - \mathbf{S}_k\mathbf{a}_k||_2^2 + \lambda_A(||\mathbf{V}_2||_{1,1} + ||\mathbf{V}_3||_{1,1}) + \mathcal{I}_{\mathbb{R}_+^{P\times N}}(\mathbf{V}_4) + \boldsymbol{\mu}^\top(\mathbf{A}^\top \mathbb{1}_P - \mathbb{1}_N) \\ &\text{s.t.} \\ &\mathbf{V}_1 = \mathbf{A} \\ &\mathbf{V}_2 = \mathcal{H}_h(\mathbf{V}_2) \\ &\mathbf{V}_3 = \mathcal{H}_v(\mathbf{V}_2) \\ &\mathbf{V}_4 = \mathbf{A}, \end{aligned} \tag{5.30}$$

Now the optimization problem in Eq. (5.30) can be expressed in the framework of the ADMM. To do so, we have to rewrite the problem of Eq. (5.30) in the following form:

$$\{\hat{\mathbf{u}}, \hat{\mathbf{v}}\} = \arg\min_{\mathbf{u},\mathbf{v}} \quad f(\mathbf{u}) + g(\mathbf{v}) \quad \text{s.t.} \quad \boldsymbol{\Gamma}\mathbf{u} + \boldsymbol{\Lambda}\mathbf{v} = \mathbf{0}, \tag{5.31}$$

where $\mathbf{u}$ and $\mathbf{v}$ are vector variables such that:

$$\mathbf{u} = \text{vec}(\mathbf{A}) \quad \text{and} \quad \mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \\ \mathbf{v}_4 \end{bmatrix}, \tag{5.32}$$

where $\mathbf{u} = \text{vec}(\mathbf{A})$, and $\mathbf{v}_i = \text{vec}(\mathbf{V}_i)$ are the vectorized versions of $\mathbf{A}$ and $\mathbf{V}_i$, respectively. Here, we let

$$f(\mathbf{u}) = \frac{1}{2}||\mathbf{x} - \text{vec}(\boldsymbol{\Sigma})||_2^2 + \boldsymbol{\mu}^\top(\mathbf{K}\mathbf{u} - \mathbb{1}_N), \tag{5.33}$$

with $\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{S}_1\mathbf{u}_1 & \cdots & \mathbf{S}_N\mathbf{u}_N \end{bmatrix}$ (where $\mathbf{u}_k \triangleq \mathbf{a}_k$). The function g is a closed proper convex function defined as:

$$g(\mathbf{v}) = \lambda_A(||\mathbf{v}_2||_1 + ||\mathbf{v}_3||_1) + \mathcal{I}_{\mathbb{R}_+^{PN}}(\mathbf{v}_4). \tag{5.34}$$

$\mathbf{K}$ is the $N \times PN$ matrix of the linear operator summing the entries of $\mathbf{a}$ corresponding to the same pixel and putting each of these sums in one entry of a vector. Finally, we have the

following definitions for $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$:

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{I}_{PN} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{I}_{PN} \end{bmatrix}, \ \mathbf{\Lambda} = \begin{bmatrix} -\mathbf{I}_{PN} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{H}_h & -\mathbf{I}_{PN} & \mathbf{0} & \mathbf{0} \\ \mathbf{H}_v & \mathbf{0} & -\mathbf{I}_{PN} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{I}_{PN} \end{bmatrix}, \tag{5.35}$$

using the fact that in the vector spaces of vectorized matrices of the appropriate sizes, the linear operators $\mathcal{H}_h, \mathcal{H}_v : \mathbb{R}^{P\times N} \mapsto \mathbb{R}^{P\times N}$ and $\mathcal{S} : \mathbb{R}^{P\times N} \mapsto \mathbb{R}^{L\times N}$ can be described by their matrices (in the canonical bases of the corresponding vector spaces) $\mathbf{H}_h$, $\mathbf{H}_v \in \mathbb{R}^{PN\times PN}$ and $\mathbf{\Sigma} \in \mathbb{R}^{LN\times PN}$, respectively.

In this framework, the problem we want to solve falls into the category of those which the ADMM can tackle. We have introduced two equivalent representations of the variables we manipulate: in a matrix form and in a vector form. The matrix form is more compact and often convenient to use, but the vector form is the only one allowing us to express linear operators as matrices. The two are completely equivalent (up to an isomorphism) and during the optimization process, we will use either of them depending on which one is the most convenient in the context. The Lagrange multipliers are denoted by $\mathbf{d}$ in a vector form, possibly indexed with the pixels (and possibly with the index of the appropriate split variable), or $\mathbf{D}$ in a matrix form. The ADMM procedure to solve Problem (5.31) is summarized in Algorithm 6.

Data: $\mathbf{X}$, $\boldsymbol{\mathcal{S}}$
Result: $\hat{\mathbf{A}} = \text{vec}^{-1}(\mathbf{u})$
Choose $\rho \geq 0$ and initialize $\mathbf{u}$,$\boldsymbol{\mu}$,$\mathbf{v}$ and $\mathbf{d}$;
while *ADMM termination criterion is not satisfied* **do**
  $\mathbf{u}, \boldsymbol{\mu} \leftarrow \arg\min_{\mathbf{u}} \mathcal{L}(\mathbf{u}, \boldsymbol{\mu}, \mathbf{v}, \mathbf{d})$;
  $\mathbf{v} \leftarrow \arg\min_{\mathbf{v}} \mathcal{L}(\mathbf{u}, \boldsymbol{\mu}, \mathbf{v}, \mathbf{d})$;
  $\mathbf{d} \leftarrow \mathbf{d} - \mathbf{\Gamma}\mathbf{u} - \mathbf{\Lambda}\mathbf{v}$;
end

Algorithm 6: ADMM process to solve problem (5.31).

The augmented Lagrangian for our problem, to be minimized w.r.t. $\mathbf{u}$, $\boldsymbol{\mu}$, $\mathbf{v}$ and $\mathbf{d}$ is:

$$\begin{aligned} \mathcal{L}(\mathbf{u}, \boldsymbol{\mu}, \mathbf{v}, \mathbf{d}) &= f(\mathbf{u}) + g(\mathbf{v}) + \frac{\rho}{2}(||\mathbf{\Gamma}\mathbf{u} + \mathbf{\Lambda}\mathbf{v} - \mathbf{d}||_2^2 - ||\mathbf{d}||_2^2) \\ &= \frac{1}{2}||\mathbf{x} - \text{vec}(\mathbf{\Sigma})||_2^2 + \boldsymbol{\mu}^\top(\mathbf{K}\mathbf{u} - \mathbb{1}_N) + \lambda_A(||\mathbf{v}_2||_1 + ||\mathbf{v}_3||_1) + \mathcal{I}_{\mathbb{R}_+^{PN}}(\mathbf{v}_4) \\ &+ \frac{\rho}{2}||\mathbf{u} - \mathbf{v}_1 - \mathbf{d}_1||_2^2 + \frac{\rho}{2}||\mathbf{H}_h\mathbf{v}_1 - \mathbf{v}_2 - \mathbf{d}_2||_2^2 + \frac{\rho}{2}||\mathbf{H}_v\mathbf{v}_1 - \mathbf{v}_3 - \mathbf{d}_3||_2^2 \\ &+ \frac{\rho}{2}||\mathbf{u} - \mathbf{v}_4 - \mathbf{d}_4||_2^2 - \frac{\rho}{2}||\mathbf{d}||_2^2. \end{aligned} \tag{5.36}$$

The full ADMM optimization procedure is described in Appendix D.

5.5.2 Coordinate Descent (CD) algorithm

Here we propose to reach a local minimum of the objective function by a coordinate descent (CD) scheme [126]. Exactly solving the subproblems can be costly and the abundance update needs to be carried out through an iterative algorithm. In coordinate descent, instead of exactly solving each subproblem, alternating the minimizations with respect to one block, we only perform one (or few) iteration(s) of each subproblem resolution (except if there is a closed form update for a subproblem), and cycle through the three blocks of variables. This scheme leads to faster iterations of the algorithms, and can avoid wasting computation time exactly solving a subproblem to get small improvements. Here, we are using the isotropic TV on the abundance for its properties, but also because using an anisotropic version would result in one more nondifferentiable term acting on the abundances, while we will see that there cannot be more than two for the algorithm we use to be applicable. The whole optimization procedure is summarized in Algorithm 7.

Data: $\mathbf{X}$, $\mathbf{S}_0$
Result: $\hat{\boldsymbol{\mathcal{S}}}$, $\hat{\boldsymbol{\Psi}}$, $\hat{\mathbf{A}}$
Initialize $\boldsymbol{\mathcal{S}}$, $\boldsymbol{\Psi}$, $\mathbf{A}$, choose λ_S, λ_Ψ and $\lambda_A \geq 0$;
while *CD termination criterion is not satisfied* **do**
  · update $\boldsymbol{\mathcal{S}}$ using Eq. (5.38) ;
  · update $\mathbf{A}$ using Eq. (5.41) (or Eq. (5.40) if $\lambda_A = 0$) ;
  · update $\boldsymbol{\Psi}$ using Eq. (5.28) (or Eq. (5.24) if $\lambda_\Psi = 0$) ;
end

Algorithm 7: Coordinate Descent scheme to find a local minimum of Eq. (5.19).

At each cycle, we start by updating $\boldsymbol{\mathcal{S}}$ because this variable interacts with both the abundances and the scaling factors. Note that the linear operators used here are also computed in the Fourier domain.

We briefly comment on the convergence of the proposed algorithm. We are using a block coordinate descent scheme to optimize the objective function. In the case where there is no more than one nondifferentiable term in for each block of variables (i.e. $\lambda_A = 0$), then the global convergence of the algorithm is proven, even if convergence acceleration is included. The proof can be found in [168]. However, if the regularization on the abundance is used, then we lose global convergence guarantees, even though each update has convergence guarantees for the subproblem addressed. In practice the algorithm does converge in any case to a local minimum of the objective function.

5.5.2.1 Endmembers update

Let us rewrite the terms of the global objective function of Eq. (5.15) depending only on $\boldsymbol{\mathcal{S}}$:

$$\begin{aligned}\mathcal{K}(\boldsymbol{\mathcal{S}}) &= \frac{1}{2}\sum_{k=1}^{N}\left(||\mathbf{x}_k - \mathbf{S}_k\mathbf{a}_k||_2^2 + \lambda_S||\mathbf{S}_k - \mathbf{S}_0\boldsymbol{\psi}_k||_F^2\right) + \mathcal{I}_{\mathbb{R}_+}(\mathbf{S}) \\ &= \sum_{k=1}^{N}\left(f_k(\mathbf{S}_k) + \mathcal{I}_{\mathbb{R}^+}(\mathbf{S}_k)\right). \end{aligned} \tag{5.37}$$

We have a smooth term (the data fit and the modelling term for the endmembers), and a nonnegativity constraint. In addition, Eq. (5.37) is pixel-separable, so we can perform N updates in parallel. We resort to a projected gradient scheme to update the sources [42]. The update is equivalent to minimizing a prox-gradient surrogate function (see [126] for details). In addition, we use an extrapolation in order to speed up the convergence, as described in [168] (see Appendix A for more details). In this case, the update at cycle i is:

$$\mathbf{S}_k^{i+1} = (\check{\mathbf{S}}_k^i - \gamma_k\nabla_{f_k}(\check{\mathbf{S}}_k^i))_+, \tag{5.38}$$

where $\gamma_k = \frac{1}{\beta_k}$ is a step size linked to the Lipschitz constant of ∇_f: $\beta_k = ||\lambda_S\mathbf{I}_p + \mathbf{a}_k\mathbf{a}_k^\top||_F$, and $\check{\mathbf{S}}_k^i$ is an extrapolated version of $\mathbf{S}_k^i$ defined as:

$$\check{\mathbf{S}}_k^i = \mathbf{S}_k^i + \omega^i(\mathbf{S}_k^i - \mathbf{S}_k^{i-1}), \tag{5.39}$$

where ω^i is the i^{th} term of a sequence of carefully chosen weights (see [168] for details).

5.5.2.2 Abundances update

The terms of Eq. (5.15) depending only on the abundances are:

$$\begin{aligned}\mathcal{L}(\mathbf{A}) &= \frac{1}{2}\sum_{k=1}^{N}||\mathbf{x}_k - \mathbf{S}_k\mathbf{a}_k||_2^2 + \mathcal{I}_{\Delta_P}(\mathbf{A}) + \lambda_A TV(\mathbf{A}) \\ &= g(\mathbf{A}) + \mathcal{I}_{\Delta_P}(\mathbf{A}) + \lambda_A TV(\mathbf{A}).\end{aligned}$$

If $\lambda_A = 0$, we are left with a smooth term and the ASC. In this case, the objective becomes separable w. r. t. the pixels, so we can perform updates in parallel. Similarly to the update of the endmembers without the regularization, we could apply a projected gradient scheme (projecting on the unit simplex can be carried out very efficiently using the algorithm of [45]). However, in this case we apply the classical FCLSU algorithm of the literature, in parallel in each pixel (since contrary to the usual unmixing problem, the endmember matrix varies in each pixel here) [76]. This algorithm is as fast as a proximal gradient update but has the advantage of solving exactly the subproblem we are interested in (to the very low cost of having a negligible slackness on the ASC).

$$\mathbf{a}_k^{i+1} = \operatorname{argmin}\frac{1}{2}||\mathbf{x}_k - \mathbf{S}_k^{i+1}\mathbf{a}_k^i||_2^2. \tag{5.40}$$

Otherwise, if $\lambda_{\mathbf{A}} \neq 0$, we have one smooth term, a convex constraint and a nondifferentiable term involving a linear operator. In order to solve this, we perform one iteration of a primal dual algorithm of the literature able to deal with this exact type of objective functions [44]. Details on this algorithm and on some necessary convex analysis concepts can be found in Appendix A. We introduce the dual variable $\mathbf{U} \in \mathbb{R}^{P \times N \times 2}$ and define the updates for cycle i (as before, we introduce two step sizes ρ and μ):

$$\begin{aligned} \mathbf{A}^{i+1} &= \mathbf{proj}_{\text{simplex}}(\mathbf{A}^i - \rho(\nabla g(\mathbf{A}^i) + \mathcal{H}^*(\mathbf{U}^i)) \\ \mathbf{U}^{i+1} &= \mathbf{proj}_{||\cdot||_2 \leq \lambda_{\mathbf{A}}}(\mathbf{U}^i + \mu\mathcal{H}(2\mathbf{A}^{i+1} - \mathbf{A}^i)), \end{aligned} \tag{5.41}$$

where $\mathbf{proj}_{\text{simplex}}$ denotes the projection on the unit simplex, and $\mathbf{proj}_{||\cdot||_2 \leq \lambda_A}$ denotes the projection on the $\mathcal{L}_2$ ball of radius λ_A (a simple normalization). This operator is indeed the proximal operator of the convex conjugate of the $\mathcal{L}_2$ norm (which is in this case the indicator function of the unit $\mathcal{L}_2$ ball), required in the primal dual algorithm. It has to be understood as being applied for each pixel and material to a two dimensional vector (corresponding to the two dimensions of the spatial gradient). Here we are using the full gradient operator, and the isotropic TV formulation of Eq. (5.16). It has better properties than the anisotropic version, and is convenient here because it is essential that there should be only one nondifferentiable term in addition to the ASC for the algorithm to be applied. The rule to guarantee convergence [44] is:

$$\frac{1}{\eta_k} - \gamma_k |||\mathcal{H}|||^2 \geq \frac{\beta_k}{2}, \tag{5.42}$$

where β_k is the Lipschitz constant for the gradient of g.

5.5.2.3 Scaling factors update

This update is unchanged w.r.t the ALS algorithm, and the closed form update, as well as all the details can be found in section 5.5.1.2.

5.6 Experimental Results

In this section, we will compare the proposed model with other approaches designed to tackle spectral variability, namely the AEB approach [145] with sparsity (using the SUnSAL algorithm [87]) to recover the abundances and the FDN approach, both introduced in Chapter 2. We will also compare the results of the ELMM (using the ALS algorithm and the anisotropic TV for the spatial regularization) to the S-CLSU approach. Finally, we will also compare the results to a recent approach, the Perturbed Linear Mixing Model (PLMM) [148]. This algorithm was specifically designed to tackle the spectral variability issue. Following our seminal idea and paper [159], which introduced for the first time a mixing model where SV was explicitly taken into account in each pixels, authors in [148] proposed to model spectral variability with an additive perturbation of some reference endmembers. The PLMM models

spectral variability in each pixel as an additive perturbation of reference endmembers, and hence is able to estimate the variability for each material and each pixel by computing the norm of the perturbation term.

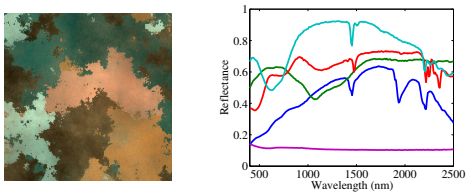
$$\mathbf{x}_k = \sum_{p=1}^{P} (\mathbf{s}_p + \mathbf{ds}_{pk}) a_{pk} + \mathbf{e}_k. \tag{5.43}$$

The term $\mathbf{ds}_{kp}$ accounts for an additive perturbation for each endmember, in each pixel, and for each wavelength. The parameters of this mixing model are then estimated through an optimization problem, in which different constraints and regularizations (spatial regularization on the abundances, proximity of the sources to a reference, limitation of the norm of the additive perturbation...) are added in order to make this NMF problem better posed and more suited to the expected properties of the solution. The algorithm iteratively updates the abundances, the endmembers and the perturbations using ADMM to converge to a stationary point of the objective function (the problem is globally not convex, but each subproblem is). This model can be seen as a particular case of Eq. (5.1), where the mappings are additive perturbations to the endmembers. However, contrary to the approach we propose, which possess physically interpretable parameters, this model is purely computational, and basically sees everything which entails modeling errors on the LMM as SV. It is still able to estimate the variability for each material and each pixel by computing the norm of the perturbation term.

5.6.1 Results on synthetic datasets

We present below the experiments performed on two types of synthetic datasets to validate the proposed approach. In both cases, we will compare the proposed approach with the classical FCLSU and CLSU, but also with the bundles approach combined with both the SUnSAL and FDN algorithms. We also compare our results to those of the PLMM algorithm. Finally, we will also compare the proposed approach with S-CLSU, which follows a particular case of the ELMM. Since the ELMM algorithm makes use of spatial regularization, for a fairer comparison, we also include the results for modified versions of the competing algorithms, in which a TV on the abundances is enforced, using the SUnSAL-TV code of [88]. We only added a termination criterion similar to the one used for the proposed approach, that is when the relative variation (in norm) of the abundance matrix between two consecutive iterations goes below $\epsilon_A = 10^{-3}$. Here, we are using the ALS algorithm, with the anisotropic TV on the abundances.

The proposed approach with both regularizations enforced is denoted by ELMM-Aψ. For the ELMM algorithm, the three tolerances $\epsilon_A = \epsilon_S = \epsilon_\Psi$ were set to 10^{-3}. When no spatial regularization is performed on the abundances or the scaling factors, we simply refer to the algorithm as ELMM. The running times of the different algorithms were measured on a computer using an Intel® Core$^{\text{TM}}$ i7-4770 CPU @ 3.40GHz (except for the PLMM).



(a) (b)

Figure 5.5: A false color representation of the first synthetic dataset (a) and the endmembers used for the simulation (b).

5.6.1.1 First scenario

The first dataset on which we tested the proposed approach was designed to follow the ELMM with some perturbations. The idea is to build a dataset which is halfway between a toy example and a realisitic simulation, in order to compare easily the different algorithms and to explain the properties and particularities of the proposed method. We randomly chose five reference endmembers corresponding to the signatures of minerals from the United States Geological Survey (USGS) spectral library, comprising 224 spectral bands in the visible and near-IR. They are shown in Fig. 5.5. The 200×200 abundance maps used were generated using Gaussian Random Fields, and were designed to comply with the ASC. Note that these abundance maps comprise only one pure pixel for each material, and around 5% of the pixels have an abundance coefficient superior to 0.9 for one material. We also generated spectral variability maps for each endmember using mixtures of Gaussians. The true abundances are shown in Fig. 5.6 (top row) and the true scaling factors are shown in Fig. 5.7 (top row). Then, the dataset was generated as follows: the pixel-dependent endmember instances were generated by multiplying the references by the corresponding spectral variability scaling factors (the achievable values are chosen so that no reflectance value becomes higher than 1, and here the scaling factors range from 0.75 to 1.25), and a white Gaussian noise was added to these endmembers, to obtain a 25dB SNR. Then for each pixel, the mixture was performed using the LMM, and finally we added another white Gaussian noise to the generated pixels, so as to obtain a 25dB SNR. The process then yielded a $200 \times 200 \times 224$ simulated hyperspectral image. A false color representation of the data can be seen in Fig. 5.5. For each algorithm, the used EEA was the Vertex Component Analysis (VCA). We use this algorithm because it is not very affected by scaling variations of the data (as seen in section 1.3.2), which makes it adapted to the geometry of our model. The same set of 5 extracted endmembers was used for all the algorithms which do not require a bundle. For the bundles, we extracted 5 endmembers instances on 50 randomly chosen susbets (without replacement) of the image whose number of pixels was 2% of this of the whole image. The clustering into bundles was performed with the k-means algorithm, with the spectral angle as a similarity measure (it is insensitive to scalings and hence adapted to the problem).

The different regularization parameters used for the tested algorithms were set empirically so as to get the best performance possible. The parameters for the synthetic data are gathered in Table 5.1. $\lambda_{\mathcal{L}_1}$ stands for the sparsity regularization parameter in SUnSAL and SUnSAL-

	BUNDLES + SUnSAL		BUNDLES + SUnSAL-TV		BUNDLES + FDN-TV		CLSU-TV		PLMM		S-CLSU-TV		ELMM		ELMM-Aψ	
$\lambda_{\mathcal{L}_1}$	2.10^{-3}	5.10^{-4}	2.10^{-3}	10^{-5}	×	×	×	×	×	×	×	×	×	×	×	×
λ_{TV}	×	×	2.10^{-3}	5.10^{-4}	3.10^{-3}	10^{-4}	5.10^{-4}	3.10^{-4}	×	×	4.10^{-3}	3.10^{-4}	×	×	×	×
λ_S	×	×	×	×	×	×	×	×	×	×	×	×	7.10^{-2}	×	7.10^{-2}	4
λ_A	×	×	×	×	×	×	×	×	×	×	×	×	0	×	4.10^{-3}	1.10^{-2}
λ_ψ	×	×	×	×	×	×	×	×	×	×	×	×	0	×	3	4.10^{-1}
α	×	×	×	×	×	×	×	×	10^{-5}	10^{-5}	×	×	×	×	×	×
β	×	×	×	×	×	×	×	×	$4.9.10^{-3}$	$4.7.10^{-3}$	×	×	×	×	×	×
γ	×	×	×	×	×	×	×	×	1	1	×	×	×	×	×	×

Table 5.1: Regularization parameters for all the algorithms concerned, for the first synthetic dataset (left cell of each column) and the second synthetic dataset (right cell of each column).

TV. For the PLMM, the parameters α, β and γ are regularization parameters associated to a Tikhonov regularization on the abundances, on a mutual distance penalization on the endmembers (similar to that of the ICE algorithm [15]). and on a penalization of the spectral variability power (see [148] for details).

The initialization of the proposed algorithm is important since the optimization problem we tackle is not convex. We chose to initialize the algorithm using the abundances of S-CLSU, every scaling factor set to one, and the five reference endmembers as well as the initial sources in each pixel were the ones extracted using VCA. In order to assess the performance of the algorithms, we define the abundance overall Root Mean Square Error (aRMSE) as:

$$aRMSE = \frac{1}{N}\sum_{k=1}^{N}\sqrt{\frac{1}{P}\sum_{p=1}^{P}(a_{pk_{true}} - \hat{a}_{pk})^2}, \tag{5.44}$$

and the overall source RMSE (sRMSE) as:

$$sRMSE = \frac{1}{N}\sum_{k=1}^{N}\sqrt{\frac{1}{LP}||\mathbf{S}_{k_{true}} - \hat{\mathbf{S}}_k||_F^2}. \tag{5.45}$$

This metric allows us to measure indirectly how well the spectral variability is recovered by comparing the true sources to the ones extracted by S-CLSU and the proposed approach. A direct comparison using the scaling factors would have been harder to perform since the extracted reference endmembers can differ from the ones actually used to generate the data, and because of the additive perturbation added to the scaled signatures. Finally, we will also compute the usual average RMSE on the whole image, and the average Spectral Angle Mapper (SAM) between the actual and reconstructed data (these two quantities were defined in section 2.3.2, in Eqs. (2.13) and (2.8)).

The quantitative results of this experiment are shown in Table 5.2. A visual representation of the extracted abundances for most algorithms is shown in Fig. 5.6, while the scaling factors extracted by S-CLSU and the proposed approach, as well as the variability estimation from the PLMM are shown in Fig. 5.7.

From the results, we can see that as expected, FCLSU performs rather poorly in a scenario where spectral variability comes into play. Since the endmember signatures are constant throughout the image, a scaled endmember can be easily mistaken for another, for instance

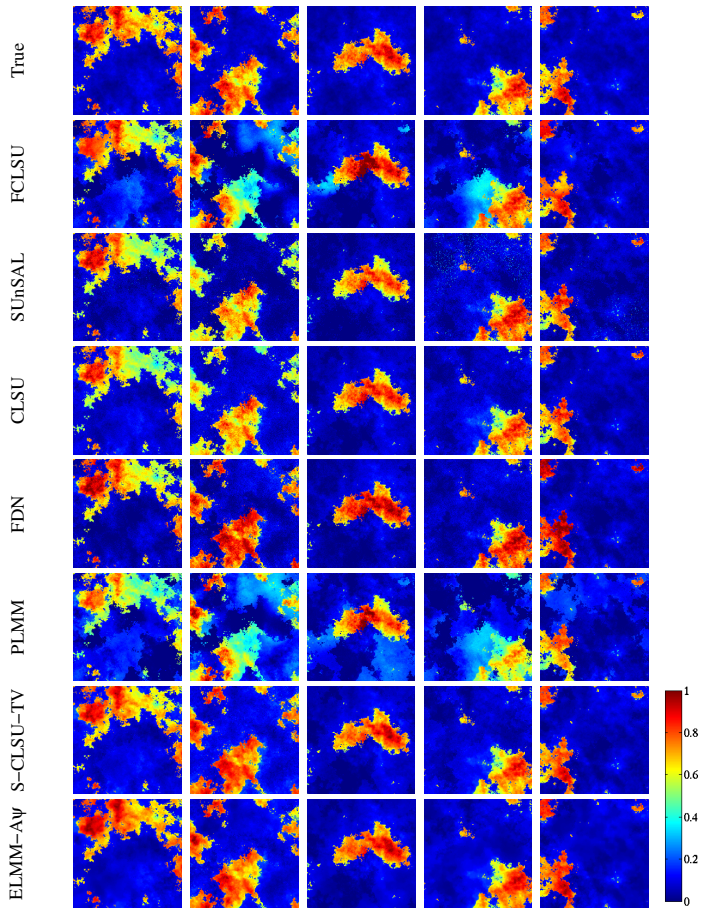


Figure 5.6: The abundances estimated by all algorithms (each column corresponds to one endmember) for the first synthetic dataset, compared to the true ones (first row).

with the endmembers depicted in cyan and red in Fig. 5.5. The bundles approach is able to obtain better results, provided the bundles are balanced and representative of the spectral variability present in the scene. This is not always guaranteed and can lead to erroneous estimations. In these experiments, we show the best result for this approach out of 15 runs. The bundles allow several instances of each endmember to be considered. The sparsity enforced by SUnSAL helps reducing the number of active endmembers per pixel but there can still remain several endmembers of the same endmember class contributing to one pixel value. The FDN approach is allowed to reduce the dimensionality of the dataset such that the impact of spectral variability is lowered.

The results from the PLMM are in this case comparable to those of the algorithms which use bundles. The main advantage of this algorithm is that it is able to estimate spectral variability maps by computing the power of the additive perturbation term in each pixel, although in this case the abundances are relatively close to the ones of FCLSU . We can see that globally, the algorithm is able to roughly identify the regions where most of the spectral variability occurs (corresponding to red or dark blue pixels in the true scaling factors maps

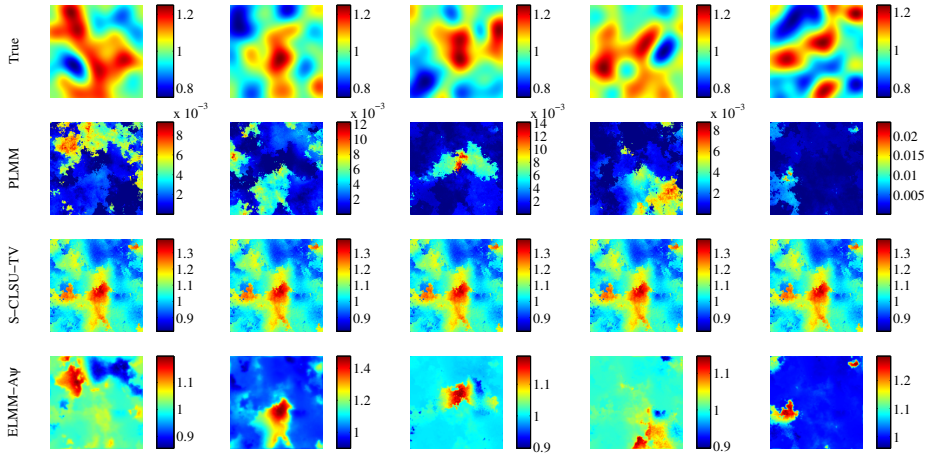


Figure 5.7: The scaling factors estimated by S-CLSU (third row) and by the proposed approach (bottom row), compared to the true ones, and to the power of the variability estimated by the PLMM algorithm (second row) for the first synthetic dataset.

when the true scaling factor is significantly above or below 1, respectively). The main drawback is that the algorithm is not able to extrapolate the information in high abundance pixels to neighboring lower abundance pixels, on which spectral variability is harder to estimate without considering spatial information.

The CLSU algorithm performs better than FCLSU and all the previously mentioned approaches, since dropping the ASC allows to look for the abundances in a cone and not in a simplex. However, the quantity estimated in each pixel by CLSU actually absorbs spectral variability into the abundances. S-CLSU performs much better. It is a very simple approach to adress spectral variability which is well suited in simple cases. The cases in which it performs best are those in which there are few materials in the image, or/and when the scaling factors are either correlated along different materials, or on the contrary if only one material per pixel varies significantly from its reference signature. This approach is also sensitive to deviations from the ELMM such as a noisy perturbation on the scaled signatures.

The TV on the abundances logically improves the results for all the algorithms based on CLSU, but it cannot improve the results for the bundle approaches. Indeed, one drawback

Algorithm	FCLSU	BUNDLES + SUnSAL		BUNDLES + FDN		CLSU		PLMM	S-CLSU		ELMM	ELMM-Aψ
SR on abundances	No	No	Yes	No	Yes	No	Yes	No	Yes	No	No	Yes
aRMSE	0.0629	0.0490	0.0504	0.0407	0.0575	0.0432	0.0413	0.0886	0.0276	0.0269	0.0344	0.0186
sRMSE	×	×	×	×	×	×	×	0.0614	0.0548	0.0545	0.0449	0.0428
xRMSE	0.0119	0.0213	0.0366	0.0331	0.1391	0.0085	0.0100	0.0136	0.0085	0.0100	0.0090	0.0088
xSAM (degrees)	1.4960	1.8752	2.1304	1.8716	7.9886	1.2268	1.2504	1.9231	1.2268	1.2504	1.3002	1.2507
Running Time (s)	14	26	131	18	8	16	6	311	17	7	366	399

Table 5.2: Quantitative results for the first synthetic dataset. The best values in each line is shown in red, and the second best one is shown in blue.

of the SUnSAL-TV algorithm is that it cannot enforce the ASC, since it is not compatible with the $\mathcal{L}_1$ norm minimization (the ASC forces the $\mathcal{L}_1$ norm of the abundance vectors to be constant). What is more, the noisiness of the abundance maps obtained with the bundles is not ideally corrected by the TV, which tends to aggregate "noisy" areas of the abundance maps into patches. However, the combination of the spatial regularizations on the abundances and scaling factors, coupled with the explicit scaling factor estimation is able to improve the results significantly.

Indeed, the proposed approach is much more robust to noise on the measured data as well as on the signatures thanks to both spatial regularizations. Indeed, the TV allows to estimate precisely the spatially correlated abundances, getting rid of the noise and the uncertainty which affects S-CLSU when two endmember variations of two different materials share a common global shape, and can look quite similar after appropriate scalings. The spatial coherency of the abundances and the scaling factors allows to recover the parameters more precisely. Besides, the explicit computation of a different scaling factor for each pixel and material allows to obtain smoother and separated variability maps, which also makes the proposed algorithm much stronger in terms of interpretability of its results. Of course, it is only possible to recover accurately the scaling factors when the abundance contribution of the corresponding material is high enough, or otherwise when the spatial information allows to extrapolate from higher abundance areas. If those two conditions are missing, only the abundance is recovered with precision, while the associated scaling factor tends to be close to one (its initial value) as the abundance decreases. This phenomenon can be interpreted geometrically, with the diagram of Fig. 5.3 in mind. Let us suppose that there are three endmembers in the scene. If in a given pixel, the abundance of one material is low, then a different scaling factor for this material will change the orientation of the simplex related to this pixel, but the edge of the simplex linking the other two (scaled) endmembers will not change, and thus the abundance coefficients for the other two materials will not change much either. In the end, we can say that the proposed approach does not require pure pixels to extract the spectral variability of a material efficiently, but only a significant abundance contribution of this material in the considered pixel, or in the neighboring area.

From a quantitative point of view, we can see that the proposed approach obtains the best results in terms of abundance estimation, as well as spectral variability recovery. An explicit spectral variability map can be only recovered for the S-CLSU, PLMM and ELMM algorithms, since only those algorithms estimate pixel-dependent endmembers, and thus enable us to compute the sRMSE values. Computing those values is theoretically possible for bundles as well, using Eq. (2.6), but the equivalent endmembers are numerically unstable for small values of the abundance coefficients (confirming how hard it is to extract variability when the contribution of a material in a pixel is small), and thus we will not use them here.

It is interesting to note that both spatial regularizations improve the results on their own w.r.t. to the simple ELMM case (the spatial regularization on the abundances (resp. scaling factors) improves the abundance (resp. scaling factors) estimation), but the combination of both improves the results further both for abundance and spectral variability estimation, since a better estimation of the scaling factors allows in turn a better abundance estimation,

and vice versa. The regularizations also improve the conditioning of the problem, and help to solve the ambiguity between abundances and scaling factors. The running time of the proposed algorithm is more important than all others (except the PLMM), but the approach is relatively fast thanks to the favorable initialization chosen.It allows to achieve a good local miminum of the objective function while limiting the number of iterations necessary to reach it with a reasonable precision. However, we can note that S-CLSU performed with SUnSAL-TV (which is based on the ADMM technique) is faster than the usual nonnegative least squares, even with the spatial regularization (at least for $\epsilon_\Psi = 10^{-3}$).

We also compared the reconstruction errors of the different algorithms, in terms of Root Mean Squared Error and in terms of Spectral Angle. These measures are indirect, in the sense that they only show how the mixing model used fits the data, though it is possible to achieve excellent reconstruction errors with a poor abundance and/or spectral variability retrieval. Conversely, accurate parameter estimation entails a good reconstruction if the model is suited to the data. For instance, we see that the bundle-based approaches fit the data worse than FCLSU, although the abundance estimation is significantly improved. We also see that for this data, all the models based on unmixing the data in a cone spanned by three endmembers achieve better reconstruction errors (as well as abundance estimation). For CLSU, S-CLSU, and the ELMM-based algorithms, the reconstruction error is similar, but there are still important differences in the accuracy of the estimation of the parameters. For the case of CLSU, only a simple scaling has a positive effect on the abundance estimation, while the reconstruction errors are of course the same in this case.

5.6.1.2 Interest of smoothing the scaling factors

Here we describe some additional results to show the relevance of smoothing the scaling factors. We show in Fig. 5.8 a plot of the estimated scaling factors against the true ones, for one of the endmembers of the image, and for two algorithms: ELMM and ELMM-Aψ. This direct comparison between estimated and true scaling factors only makes sense if the reference endmember matrix $\mathbf{S}_0$ is the same in both cases. Hence, here we assumed the reference endmembers were known. The red dots in Fig. 5.8 show that without any spatial regularization, the ELMM is able here to estimate spectral variability only when the abundance coefficient in one pixel for the considered material is above a certain value (here around 0.3). Otherwise, the estimated values stay close to their initial value 1. We had already mentioned this phenomenon above, and explained it geometrically. However, with the spatial regularization on the scaling factors, we see that not only are we able to reduce the estimation error for the red dots in most cases, we are also in general able to stir the other pixels with lower abundance values towards a more accurate spectral variability estimation, using the spatial information. As expected, the spatial smoothness of the scaling factor then helps estimating spectral variability in more heavily mixed pixels than with a pixelwise approach. Note that here, only the spatial regularization on the scaling factors impacts the shape of this plot, while the TV on the abundances does not play an important role. Notwithstanding, the latter is still very useful on its own for the abundance estimation, and as shown in the quantitative results, it is complementary to the spatial regularization on the scaling factors since a better

scaling factor estimation allows in turn a better abundance estimation.

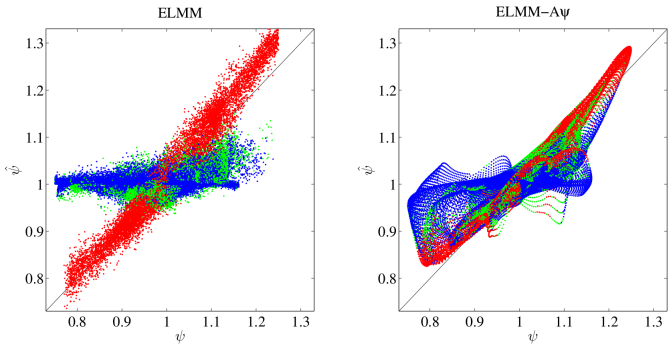


Figure 5.8: Plot of the estimated scaling factors against the true ones (assuming $\mathbf{S}_0$ is known), for ELMM and ELMM-Aψ, corresponding to the one of the endmembers of the results (second column of Figs. 5.6 and 5.7). Each dot represents a pixel in the image, and its color refers to the value of a_{2i}, where i denotes the pixel index. Red corresponds to $a_{2i} \geq 0.3$, green corresponds to $0.1 < a_{2i} < 0.3$, and blue corresponds to $a_{2i} \leq 0.1$.

5.6.1.3 Sensitivity analysis

In this section we show some results regarding the sensitivity of the proposed method w.r.t. the three regularization parameters to tune. We have run the ELMM algorithm for $\lambda_S \in [10^{-2}, 10^{-1}]$ with steps of 10^{-1}, $\lambda_A \in [10^{-3}, 10^{-2}]$ with steps of 10^{-2}, $\lambda_\psi \in [1, 10]$ with steps of 1. In order to visualize the sensitivity of the algorithm to the regularization parameters, we have plotted a well chosen isosurface of the three-variable functions given by aRMSE and sRMSE in Fig. 5.9. The chosen values were 0.0195 for aRMSE and 0.0425 for sRMSE. Inside the volume delimited by the surface, the metric is lower than the chosen value. This delimits a 3D domain for the regularization parameters inside which performance remains much better than the competing algorithms. The intersection of the two surfaces would delimit a volume inside which good performance is guaranteed both for abundance and spectral variability estimation. We see that logically, λ_A (resp. λ_Ψ) is a critical parameter for a good abundance (resp. spectral variability) estimation, while the value of λ_S is important for both, and is probably the most critical parameter overall. Still, it can be chosen in a relatively large domain for close to optimal performance.

5.6.1.4 Second scenario

The second dataset we used was generated in order to mimic the spectral variability induced by changing illumination conditions and topography along the scene, using the Hapke model [73]. For the simulations, we selected 3 endmembers consisting in 16 wavelengths reflectance measurements for materials commonly found on small bodies of the Solar System

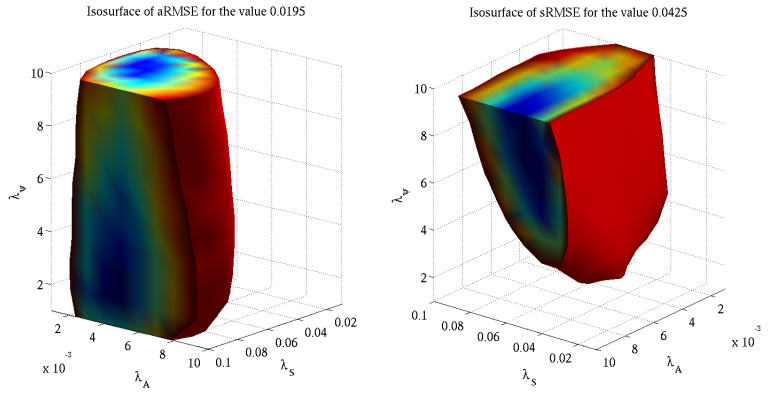


Figure 5.9: Isosurfaces for two values of aRMSE = 0.0195 (left) and sRMSE = 0.0425 (right) (seen as functions of the three regularization parameters), delimiting a domain inside which the metrics are lower than these two values. For the aRMSE, the color scale ranges from 0.0186 (blue) to 0.0195 (red). For the sRMSE, the color scale goes from 0.0423 (blue) to 0.0425 (red).

(basalt, palagonite and tephra), and whose geometry for the acquisition, as well as their photometric parameters are known [46]. We show these reflectance spectral signatures, acquired at nadir with an incidence angle of 30°, in Fig 5.10. Note that palagonite and tephra are spectrally very close (the spectral angle between the two materials is 10 degrees), making the problem harder since the nonlinearities of the Hapke model will have more influence on the abundances for correlated endmembers. From these data, we recovered the single scattering albedo spectra of these materials by inverting the Hapke model [112]. Single scattering albedo is completely characteristic to a material, and unlike reflectance, which is the physical quantity we work with, it depends neither on the geometry of the scene nor on the illumination conditions [34]. Separately, a simulated smooth 200×200 Digital Terrain Model (DTM) was synthesized, assuming a spatial resolution of 1 m. This DTM simulates a hilly region and is shown in Fig. 5.10. From this model and the definition of the position of the sun and the sensor w.r.t. the scene (sun making an angle of 18° with the flat part of the DTM and sensor at nadir), we derived the acquisition angles associated to each pixel. They depend on the position of the sun and sensor, but also on the orientation of the tangent plane to the surface at each location, which itself depends on the topography of the scene. We show the computed angles in Fig. 5.11. Plugging these angles, the single scattering albedos and the photometric parameters into the Hapke model, we simulated the various instances of the reflectance endmembers along the scene. Then we mixed these endmember variants in each pixel using the LMM, using abundances generated in a similar way to the previous section (with the same pixel purity characteristics), providing a $200 \times 200 \times 16$ image, to which 25dB white Gaussian noise was added. The flowchart of the synthetic data generation is shown in Fig. 5.12. A false color composition, and a representation of the dataset (in blue) and the endmembers generated by the Hapke model (in red) using the first three components of a Principal Component Analysis (PCA) are shown in Fig. 5.13. This representation shows that there are strictly speaking very few (one per material, actually) pure pixels in the image.

Algorithm	FCLSU	BUNDLES + SUnSAL		BUNDLES + FDN		CLSU		PLMM	S-CLSU		ELMM-Aψ	ELMM-Aψ-C
TV on abundances	No	No	Yes	No	Yes	No	Yes	No	No	Yes	Yes	Yes
aRMSE	0.133	0.0860	0.0498	0.1330	0.0681	0.0676	0.0601	0.1445	0.0398	0.0300	0.0286	0.0291
sRMSE	×	×	×	×	×	×	×	0.0814	0.0139	0.0135	0.0187	0.0128
xRMSE	0.0131	0.0105	0.0049	0.0716	0.0577	0.0041	0.0045	0.0052	0.0041	0.0045	0.0049	0.0048
xSAM (degrees)	2.0209	1.0906	1.1317	8.7639	7.9573	1.0882	1.1270	1.2035	1.0882	1.1270	1.2720	1.1631
Running Time (s)	10	15	152	10	7	11	6	235	12	7	135	143

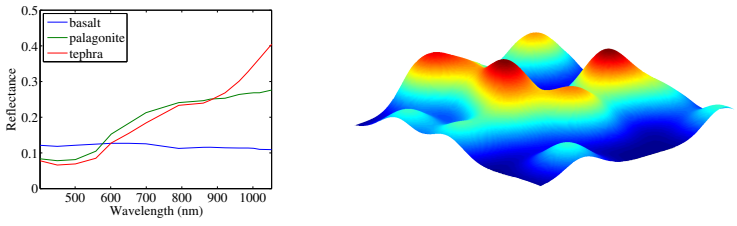
Table 5.3: Quantitative results for the second synthetic dataset. The best values in each line is shown in red, and the second best one is shown in blue.

This figure confirms that palagonite and tephra are spectrally close, when we look at the scale of the second principal component. Furthermore, we can see that the different materials are not equally affected by spectral variability. The endmembers corresponding to basalt are less affected by the nonlinearities of the Hapke model, which have a stronger influence on high albedo materials, whereas the spectrum of basalt is very flat and low. Hence the shape of the variability is almost a straight line. For the other two materials, the manifold of the endmember variants is, however, more complex.

The setup for this dataset is rather similar to this of the first synthetic dataset, with a few notable differences. The regularization parameters for all algorithms are given in the supplementary material file. For this data, we chose to set the mean of each extracted endmember bundle as the reference endmembers. The idea is to obtain representative endmembers to increase the robustness of the algorithm. For a fair comparison, we chose the same set of endmembers for FCLSU, CLSU and S-CLSU. The initial abundances used are those of S-CLSU and the initial scaling factors are either set to one (approach denoted by ELMM-Aψ), or taken from the results of S-CLSU as well (approach denoted by ELMM-Aψ-C).

The quantitative results of this second experiment are shown in Table 5.3. A visual representation of the extracted abundances for most algorithms is shown in Fig. 5.14, while the scaling factors extracted by S-CLSU and the proposed approach (ELMM-Aψ-C) are shown in Fig. 5.15. We also show the endmembers estimated by the proposed approach, compared the true ones using a PCA in Fig 5.16.

From Fig. 5.14 and Table 5.3. we can see that for FCLSU, CLSU, and the combination of the bundles and SUnSAL, the results are similar to those related to the first synthetic



(a) (b)

Figure 5.10: The reflectance endmembers (left) and the Digital Terrain Model used for the second synthetic dataset (right).

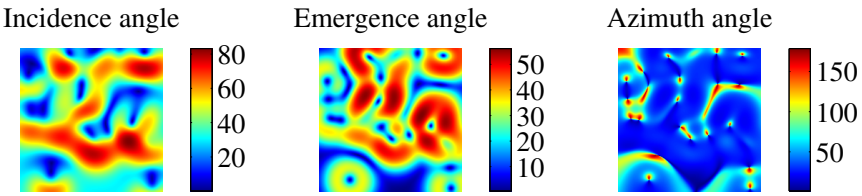


Figure 5.11: The incidence, emergence and azimuth angles computed from the DTM (degrees).

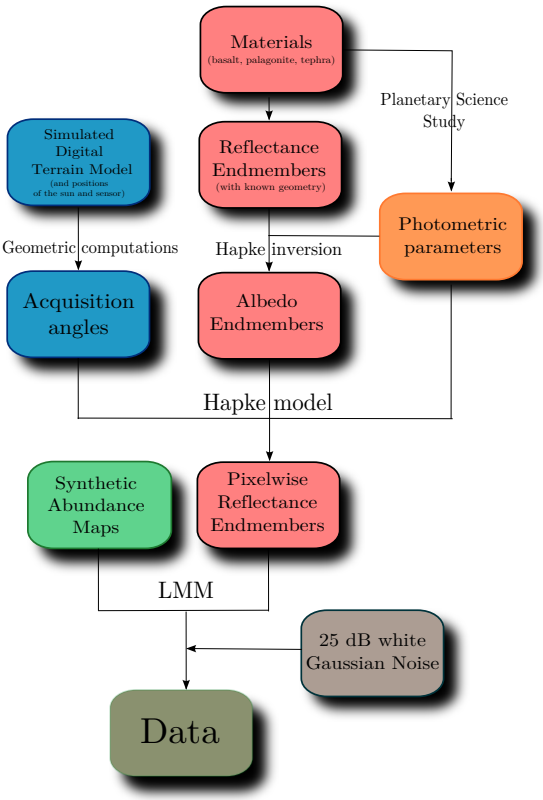
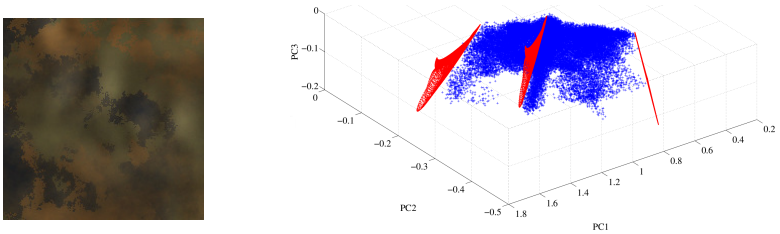


Figure 5.12: Flowchart of the second simulated dataset generation.



(a) (b)

Figure 5.13: A false color representation of the second synthetic dataset. Data cloud (blue) and the endmember variants generated by the Hapke model (red) shown using the first three components of a PCA.

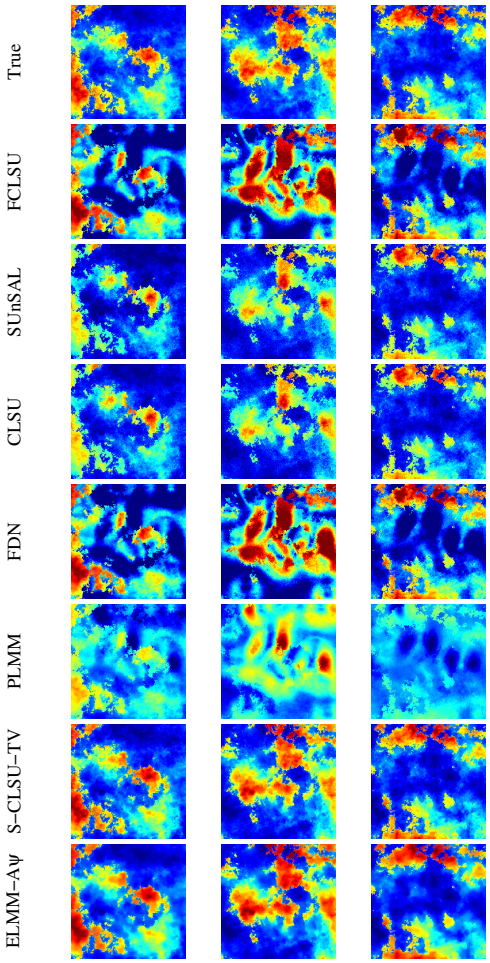


Figure 5.14: The abundances estimated by all algorithms for the second synthetic dataset, compared to the true ones.

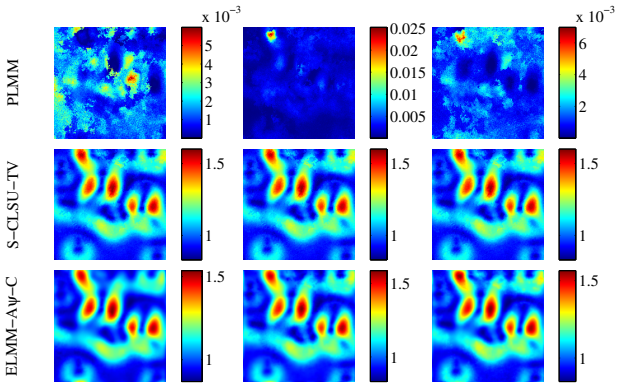


Figure 5.15: Magnitude of the PLMM variability term (top row), the scaling factors estimated by S-CLSU (middle row) and by the proposed approach (bottom row) for the second synthetic dataset.

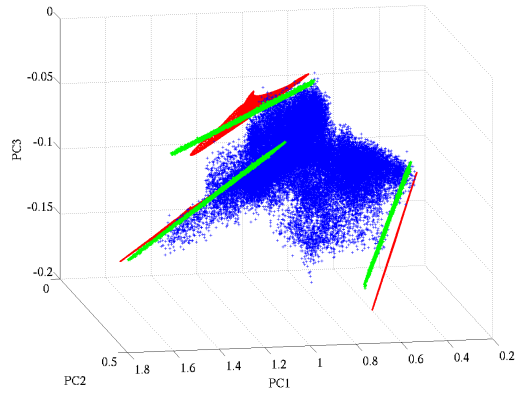
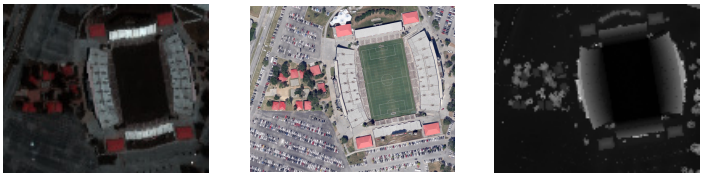


Figure 5.16: The second simulated dataset (blue), the endmember variants generated by the Hapke model (red) and the sources estimated by the proposed algorithm (green), shown using the first three components of a PCA.

dataset. FCLSU gets poor results since it does not take spectral variability into account. CLSU obtains better results, as it does not really estimate the abundances, but their products with the scaling factors. SUnSAL is able to partly explain endmember variability, but the resulting abundance maps are noisy, and the performance is limited by the bundle extraction. The FDN approach only provides slightly better results than FCLSU in this case (we kept the best result over 15 bundle extractions). This might be because the projection performed by the FDN approach suffers from outliers in the bundles. SUnSAL obtains better results with the same bundles because the sparsity constraint helps discarding the outlier endmembers. The S-CLSU approach obtains better results, and is able to recover an average spectral variability map. In this case, its performance in terms of sRMSE (but not aRMSE) is better than the proposed approach initialized with the abundances S-CLSU obtains and the scaling factors set to one. The reason S-CLSU obtains satisfactory results is because the spectral variability in the different materials have the same cause, and hence the scaling factors for each material are correlated. However, the ELMM-Aψ initialization also gets a worse endmember estimation result because it cannot estimate the scaling factor of a material whose abundance is too low. As explained in Sec. 5.6.1.1, the change in orientation of the simplex due to the scaling factor associated to a low abundance material has little impact on the remaining abundance coefficients. As S-CLSU is only able to estimate one scaling factor for each endmember, if we assume it applies to all endmembers, the error committed is less important. A spatial regularization on the abundances if beneficial to the performance of the various algorithms, especially for the bundles approach with SUnSAL, as well as S-CLSU. The PLMM obtains poor results here because its abundances are similar to those of FCLSU, which means the algorithm is probably stuck in a poor local minimum.

The proposed approach, initialized with the abundances and scaling factors of S-CLSU is still able to improve the results thanks to the regularizations, which accommodate the noise, and especially the nonlinearities of the Hapke model, as can be seen in Fig. 5.16. In this figure, we represent the data cloud (in blue) and the endmembers generated by the Hapke model



(a) (b) (c)

Figure 5.17: A RGB representation of the Houston hyperspectral dataset (a). High spatial resolution color image acquired over the same area at a different time (b). Associated Lidar data (c). Black corresponds to 9.6m and white corresponds to 46.2m.

(in red) using the first 3 components of a PCA. The endmembers estimated in every pixel by the proposed algorithm are shown in green. The extracted endmembers are allowed to deviate from the ELMM (as it is defined in Eq. (5.12)), and the endmembers are extracted on a thickened line. This flexibility allows to approximate the endmember manifolds generated by the Hapke Model better than S-CLSU does. Only the basalt endmembers are more or less situated on a straight line, because this material has a lower albedo than the other two, and is then less affected by the illumination changes over the scene.

It is also interesting to note that, as could be expected, the scaling factors extracted by S-CLSU or the proposed approach are very correlated to the DTM, and even more to the spatial maps of the incidence and emergence angles, shown in Fig. 5.11. However, the proposed approach is able to drift away from the the model of Eq. (5.12) and is more robust than S-CLSU here thanks to both spatial regularizations (even when the S-CLSU benefits from an additional spatial regularization on the abundances). In this case, the freedom of the sources to evolve around the straight lines cause a slight increase in the spectral angle between the image and its reconstruction. This could also be because the spatial regularizations denoise the abundance and scaling factor maps, leading to a smoother reconstructed image than the noisy data. These results confirms the potential of the ELMM to deal with illumination and topography induced spectral variability.

5.6.2 Results on real datasets

5.6.2.1 First Dataset

The first real dataset we use here is the subset of the Houston dataset already used in Section 4.4.2.2. We show here once again the RGB representation of the HSI, the high-resolution image of the scene and the LiDAR data in Fig. 5.17. We are comparing the same algorithms as before. In the absence of ground truth, we will only assess the results visually, and give the running times and reconstruction errors of each algorithm. The regularization parameters for this dataset are gathered in the supplementary material. For both datasets, the PLMM algorithm was used with a spatial smoothness constraint on the abundances, and a constraint forcing the endmembers to be close the the reference extracted with VCA. The bundle used

	BUNDLES + SUnSAL		PLMM		ELMM-Aψ	
$\lambda_{\mathcal{L}_1}$	5.10^{-4}	5.10^{-4}	×	×	×	×
λ_S	×	×	×	×	0.5	0.4
λ_A	×	×	×	×	$1.5.10^{-2}$	3.10^{-3}
λ_ψ	×	×	×	×	5.10^{-2}	5.10^{-3}
α	×	×	$1.4.10^{-3}$	$3.1.10^{-4}$	×	×
β	×	×	5.10^{2}	5.10^{2}	×	×
γ	×	×	1	1	×	×

Table 5.4: Regularization parameters for all the algorithms concerned, for the Houston dataset (left cell of each column) and the Cuprite dataset (right cell of each column).

was extracted using 45 subsets of 2 percent of the pixels of the image, without replacement. For the proposed approach, we initialized the algorithm with the abundances of S-CLSU and the scaling factors set to one.

The regularization parameters for the real datasets are stored in Table 5.4. For the PLMM algorithm, the parameters α, β and γ are associated to a spatial regularization on the abundances, to a distance of the endmembers to $\mathbf{S}_0$, and to the spectral variability power, respectively.

The estimated Intrinsic Dimensionality of the dataset using the Hysime algorithm [19] is 17, but we chose to consider only 4 endmembers. The reason for this is twofold: First, when 17 endmembers are extracted, for all algorithms, most abundance maps are very sparse and have very few spatial structure. This is either because outliers are selected as endembers or because a really rare material was chosen (such as an isolated car in a single pixel only). Besides, ID estimation algorithms provide an upper bound on the number of endmember to use, and the definition of an endmember is actually application and context dependent. Results with 4 endmembers are easier to visualize, to interpret and to compare for our endmember variability application. From the reference endmembers selected by VCA, we identified 4 classes: vegetation, concrete stands, asphalt and red metallic roofs. The vegetation endmember could have been split into grass and trees, but we chose to consider only one endmember vegetation to show the capability of the algorithms to recover an interpretable spectral variability. The same goes for the football field, which is actually mixed with soil. However, the soil endmember is very hard to extract, as there is probably no pure soil pixel for this material in the considered dataset.

5.6.2.2 Results

Algorithm	FCLSU	BUNDLES + SUnSAL	BUNDLES + FDN	CLSU	PLMM	S-CLSU	ELMM-Aψ
xRMSE	0.0212	0.0065	0.0645	0.0047	0.0263	0.0047	0.0032
xSAM (degrees)	3.3057	1.0953	4.1885	1.4531	6.6727	1.4531	0.9979
Running Time (s)	4	6	5	4	333	5	432

Table 5.5: Running times and reconstruction errors of the tested algorithms on the Houston dataset.

The results on the real dataset are shown in Fig. 5.18 for the abundances and in Fig. 5.19

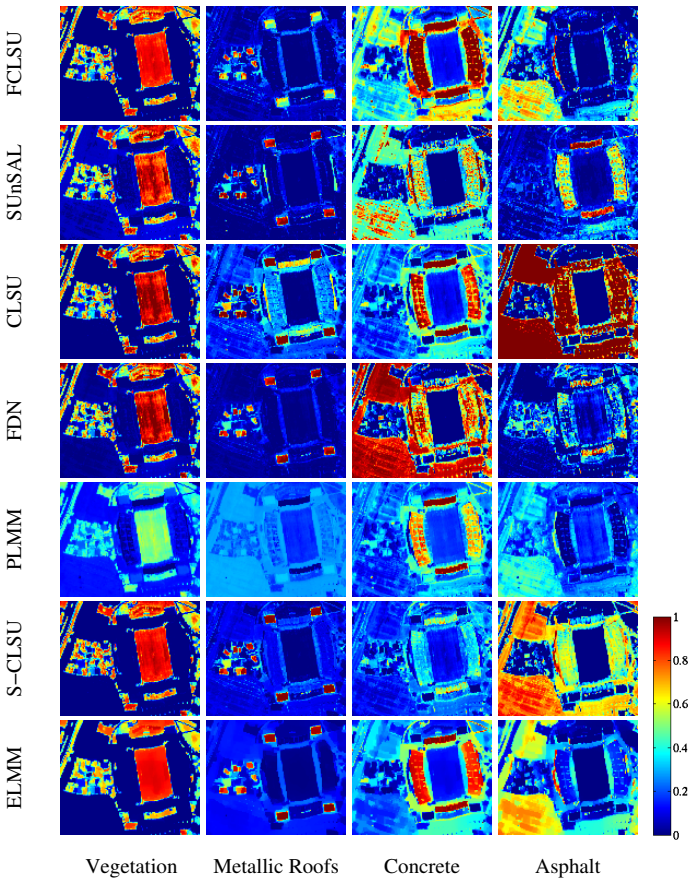


Figure 5.18: The abundance maps estimated by all algorithms for the Houston dataset. The color scale goes from 0 (blue) to 1 (red). For CLSU, all the abundances higher than 1 are shown in red.

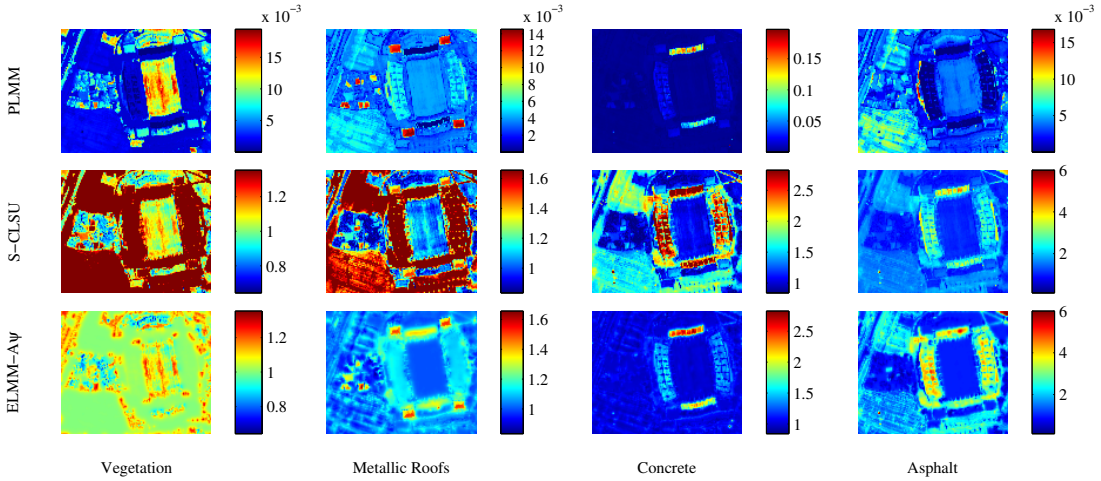


Figure 5.19: Magnitude of the PLMM variability term (top row), the scaling factors estimated by S-CLSU (middle row) and the proposed approach (bottom row) for the Houston dataset.

for the spectral variability estimation. In addition, the reconstruction errors and running times of the different algorithms are shown in Table 5.5. From Fig. 5.18, it seems that overall the abundance distributions of FCLSU follow the visual examination of the image, with very pure areas for the stands, and a good identification of the lawn in the stadium and of the stands (all the stands are indeed made of the same material if we refer to the high resolution RGB image, whereas it is not clear at all from the RGB composition of the hyperspectral data). However, the algorithm fails to consider the red metallic roofs as pure. The bundle approach combined with SUnSAL improves the purity of the red roof areas but the stands are not so well identified and interpreted as a mixture of concrete and asphalt. The CLSU algorithm (without scaling) obtains visually more coherent results, but both the red roofs and the concrete stands exhibit abundances which are significantly higher than 1 for all materials (up to 1.3, corresponding to saturated red values in Fig. 5.18), because CLSU does not actually estimate the abundances but a factor incorporating spectral variability. The FDN approach obtains very clear abundance maps for the vegetation and red roofs, but once again the distinction between concrete stands and asphalt (roads and parking lots) is not so clear. The abundances from the PLMM are visually not very satisfying since there are a lot a pixels which should be pure but are here heavily mixed. However, the algorithm is still able to detect most of the areas where spectral variability occurs (red roofs, stadium stands for instance). S-CLSU, on the other hand, obtains visually good results because it corrects the estimations of CLSU thanks to the scaling. Then, the abundance maps for the red metallic roofs are better defined. The vegetation is also well identified. Asphalt and concrete stands are harder to discriminate. Besides, the scaling factors map is hard to interpret because only one scaling factor is estimated for all four endmembers. In Fig. 5.19, the color scale for each material was chosen using the results of the proposed approach. Otherwise, the results from S-CLSU have a very large dynamic, which makes it hard to visualize the results, added to the fact that there is only a single variability map for all materials. The proposed approach, although computationally more intensive, obtains visually good results as well. The vegetation is well identified, and the football field appears purer than with S-CLSU thanks to the spatial regularization. With our definitions of the endmembers, potential mixtures of grass with soil are interpreted as variability. The distinction between grass and tree leaves is also clearly identifiable in the scaling factors maps because the leaves areas are associated in this case to scaling factors smaller than 1, even though in some cases the pixels can be mixed with red roofs or asphalt (for instance in the area to the left of the stadium). In this case, the model seems to accommodate the intrinsic variability of the materials since it can estimate different scaling factors for different materials in mixed pixels, thanks to the inclusion of one scaling factor for each material and to the spatial regularizations. The red roofs are also well detected, and the corresponding scaling factors significantly differ depending on the orientation of the roof (which we can clearly see with the high resolution image and the LiDAR data). The same phenomenon occurs with the stands: the 4 stands are related to the same endmember class, but they all have significantly different scaling factors, depending once again on their orientation (they also correspond to significant elevation changes, as can be seen on the LiDAR image). These two facts suggest that the ELMM is indeed able to identify variability due to changing illumination conditions. The asphalt abundance map coincides with the location of the parking lots, and also shows local variations in scale. Finally, the

spatial regularization also eliminates outliers (cars) in the spatial distribution of the scaling factors and abundance maps.

5.6.2.3 Second Dataset

The second dataset we consider is a $200 \times 200 \times 186$ subset of the Cuprite dataset, which is shown in Fig. 5.20. The image was acquired by NASA's AVIRIS sensor and covers the Cuprite mining district in western Nevada, USA. We extracted 14 endmembers with the VCA according to the ID value estimated by Hysime on our subset. We compare the same algorithms as before and show in Fig. 5.21 the estimated abundance maps. The results are shown only for some of the extracted endmembers. For the concerned algorithms, we also show in Fig. 5.22 a map of the estimated spectral variability. We also show the reconstruction errors and the running times of all algorithms in Table 5.6. The materials have been identified by visual comparison between the estimated abundance maps and endmember signatures to those recovered in [122].

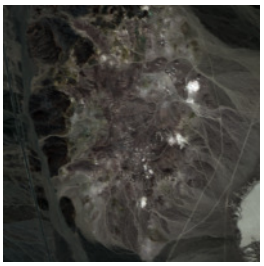


Figure 5.20: A RGB representation of the subset of the Cuprite dataset used.

5.6.2.4 Results

From the visual results, we see for instance that FCLSU detects a near pure area of Alunite in the top of the rightmost part of the image, while this is interpreted as near pure Muscovite with variability by the ELMM and S-CLSU, which shows taking variability into account can significantly change the abundance results. The PLMM algorithm detects more or less the same variability areas than the ELMM, but its abundance maps are in average lower, meaning that is interprets the data as being more mixed. As for the variability maps, for ELMM and S-CLSU, we chose the reference color scale to reflect the dynamic of the map of S-CLSU. This shows that for 14 materials, it can become very hard to interpret visually, and even more so for mixed pixels, while the ELMM results with one scaling factor for each material is much clearer, even at this scale, which favors S-CLSU. Anyway, we see that the materials in the scene seem to be significantly affected by spectral variability, which makes the abundance maps recovered by the algorithms taking it into account very different from the ones recovered by the usual LMM.

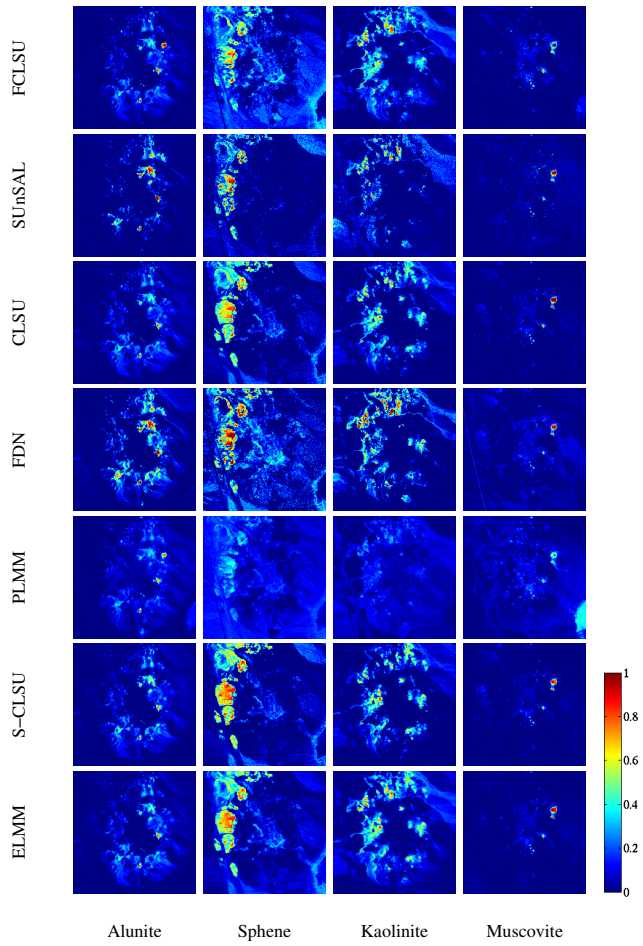


Figure 5.21: The abundance maps estimated by some algorithms for the Cuprite dataset. The color scale goes from 0 (blue) to 1 (red).

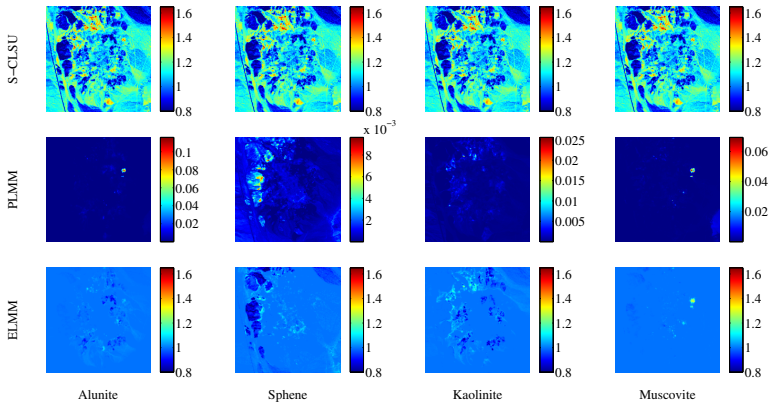


Figure 5.22: Magnitude of the PLMM variability term (top row), the scaling factors estimated by S-CLSU (middle row) and the proposed approach (bottom row) for the Cuprite dataset.

Algorithm	FCLSU	BUNDLES + SUnSAL	BUNDLES + FDN	CLSU	PLMM	S-CLSU	ELMM-Aψ
xRMSE	0.0062	0.0138	0.0361	0.0047	0.0086	0.0047	0.0029
xSAM (degrees)	0.9026	0.4759	1.1954	0.7088	1.2708	0.7088	0.4489
Running Time (s)	26	96	30	27	4.10^3	28	3.10^3

Table 5.6: Running times and reconstruction errors of the tested algorithms on the Cuprite dataset.

5.7 Partial Conclusion

In this chapter, we have proposed a new mixing model for hyperspectral unmixing, specifically aimed at tackling spectral variability. The model approximates the radiative transfer model of Hapke by making simplifying assumptions. In the end, spectral variability is taken in consideration through scaling factors, defining pixelwise endmembers from reference ones, which are extracted using VCA (robust to scaling variations of the data). We incorporate spatial information in the algorithm through spatial regularizations on the abundances and scaling factors. We have also proposed two algorithms to estimate the parameters of the model. We have validated the model on two synthetic datasets, including one generated with the Hapke model, and on two real datasets with different spatial and spectral resolutions, in different contexts (urban scene and natural landscape). We have showed that the algorithm outperforms other approaches of the literature aimed at addressing SV. Note that the scaling factor model was also proven useful to deal with spectral variability in the dynamical unmixing of multitemporal HSIs in [78]. The research perspectives related to this model are numerous, and have started being addressed by the community. The proposed model only affects the endmembers, while the mixing process remains linear. It is theoretically possible to combine the ELMM with any nonlinear model of the literature, so long as the model remains tractable and as there are not too many parameters to estimate. An encouraging step was taken in that direction in [72], where a scaling factor (the same for all materials) is included into a bilinear model, to account for both illumination and multiple reflection effects on the observed reflectance. Taking advantage of an available DTM, and knowing the positions of the sun and sensor would allow to derive the acquisition angles in each pixel of the HSI. This would for example help to detect areas in shadow (when the zenith angle $\theta_0 > 0$). This valuable information could also be included in a refined (nonlinear) mixing model approximating the Hapke model in a less coarse way than the ELMM. Designing more material specific mixing models would also be an interesting possibility. Finally, a comparison of the performance of the two proposed algorithms would be interesting in order to find out which algorithm is the most efficient in terms of complexity and runtime, as well as unmixing performance.

In the next chapter, we are going to make the connection of the ELMM with two different frameworks: tensor decomposition of hyperspectral data and the LSU approach.

Chapter 6

ELMM applications

Contents

6.1 Introduction

In this Chapter, we present two applications of the ELMM introduced in Chapter 5. The algorithms presented in that chapter were using the ELMM in a completely global approach, that is to say that the model was applied to each pixel using the same reference endmembers, and the scaling factors, as well as the abundances and the local endmembers were globally estimated through an optimization process. On the contrary, the Local Spectral Unmixing (LSU) approach is able to deal with spectral variability using a different paradigm, considering sets of local endmembers in each region, possibly representing different macroscopic materials in different regions. We have shown in section 4.3 that this approach could be useful to interpret the unmixing locally in each region, and at different scales using the hierarchy defined by the BPT. However, the main limitation of LSU techniques is that the induced endmembers and abundances are only defined locally in the image, and must somehow be post-processed to be interpretable at the whole image scale. Here, we are going to make the connection between the two approaches, using the ELMM to interpret the LSU results on a global scale.

Besides, we have mentioned in section 2.4 that multitemporal and multiangular HSIs can be viewed as three-way arrays, or tensors, and can be processed using the CP decomposition, in order to obtain spectral factors, spatial factors, and factors related to the third modality, i.e. time or angle. The former two factors can be interpreted as endmembers (and the number

of endmembers can be related to the rank of the tensor) and fractional abundances, but the third factor, albeit physically interpretable in practice, has yet to be linked to a quantity involved in a physics-based mixing model.

6.2 Contributions

This chapter connects the ELMM to the LSU and tensor CP decomposition approaches. In a first step, we propose to combine the LSU approach with the ELMM framework and show that it allows to naturally derive global endmembers and abundances from their local counterparts, while providing additional information related to the spectral variability in each region, thanks to the explicit computation of scaling factors. These results were first reported in [149].

Second, we show that the CP decomposition of hyperspectral data is connected to a regularized version of the ELMM, and that the factors associated to the third modality (that is, not the spatial or spectral one) can be linked to the scaling factors of the ELMM. We illustrate this connection in the multitemporal and multiangular cases, and on a recently proposed representation of conventional HSI images into tensors (we will see that seeing the HSI cube directly as a third order tensor is not a suitable option for SU purposes), for which the CP decomposition provides information on the endmembers and abundances, but also on the spectral variability in the spatial domain.

6.3 LSU and ELMM

6.3.1 Interpreting LSU results on a global scale

Since our objective is to interpret the LSU results on a global scale, the starting point of the proposed methodology is a partition $\pi = \{\mathcal{R}_i\}, i = 1, ..., |\pi|$, of the spatial support of the HSI, where LSU procedures have been conducted on each region $\mathcal{R}_i$ of the partition. Following the approach presented in [161] (and summarized in section 2.3.2), those regions have been designed to have minimal reconstruction errors. Thus, $d_{\mathcal{R}_i}$ local endmembers $\mathbf{s}_1^{\mathcal{R}_i}, \ldots, \mathbf{s}_{d_{\mathcal{R}_i}}^{\mathcal{R}_i}$ are available in each region $\mathcal{R}_i$, as well as their associated local fractional abundances $\mathbf{a}_k^{\mathcal{R}_i} = \left[a_{1,k}^{\mathcal{R}_i}, \ldots, a_{d_{\mathcal{R}_i},k}^{\mathcal{R}_i}\right]$ for all pixel spectra $\mathbf{x}_k$ belonging to $\mathcal{R}_i$. Note that here, $d_{\mathcal{R}_i}$ is the estimated local ID in region $\mathcal{R}_i$, i.e. we do not apply the procedure described in section 4.3 to reduce the number of endmembers in each region, for reasons that we will detail further. Recall that in the LSU framework, each pixel $\mathbf{x}_k$ (k being the index of the position of the pixel in the whole image) associated to a region $\mathcal{R}$ can be expressed only using the information contained in this region:

$$\mathbf{x}_k = \sum_{i=1}^{d_{\mathcal{R}}} a_{i,k}^{\mathcal{R}} \mathbf{s}_i^{\mathcal{R}} + \mathbf{e}_k \ , \tag{6.1}$$

6.3.1.1 Endmember clustering

In a first step, all local endmembers are pooled together in a common set $\mathcal{S}_\pi = \{\mathbf{s}_i^{\mathcal{R}\in\pi}\}_{i=1}^{d_\pi}$ with $d_\pi = \sum_{\mathcal{R}\in\pi} d_\mathcal{R}$ being the total number of local endmembers that have been generated by the LSU approach. Then, this set $\mathcal{S}_\pi$ is clustered into P clusters (where P is the global number of endmembers to consider) $\mathcal{C}_1, \ldots, \mathcal{C}_P$ by means of some clustering algorithm, with cluster $\mathcal{C}_p = \left\{\mathbf{s}_{p,1}^\mathcal{R}, \ldots \mathbf{s}_{p,d_{\mathcal{C}_p}}^\mathcal{R}\right\}$ being composed of $d_{\mathcal{C}_p}$ local endmembers, originating from various regions $\mathcal{R}$ of the partition π. Finally, the centroid $\mathbf{s}_{0p}$ of each cluster $\mathcal{C}_p$ is retrieved

$$\mathbf{s}_{0p} = \frac{1}{d_{\mathcal{C}_p}} \sum_{i=1}^{d_{\mathcal{C}_p}} \mathbf{s}_{p,i}^\mathcal{R} \ , \tag{6.2}$$

and defined as the global endmember representing the p^{th} cluster $\mathcal{C}_p$. As opposed to all local endmembers $\mathbf{s}_{p,i}$ belonging to this cluster, the centroid $\mathbf{s}_{0k}$ is expected to properly describe the macroscopic material associated with cluster $\mathcal{C}_p$ across the whole image.

6.3.1.2 Global abundance retrieval

Once the global endmembers $\mathbf{s}_{01}, \ldots, \mathbf{s}_{0P}$ have been defined, their associated global fractional abundances must be retrieved for all pixels of the HSI. In particular, let $\mathbf{x}_k$, $k \in [\![1, ..., N]\!]$ be such a pixel spectrum contained in region $\mathcal{R}$ of partition π, and let $\left[a_{1,k}^\mathcal{R}, \ldots, a_{d_\mathcal{R},k}^\mathcal{R}\right]$ and $[\alpha_{1,k}, \ldots, \alpha_{P,k}]$ be its local and global fractional abundances, respectively. Then, three possible cases may occur:

- No local endmember $\mathbf{s}_i^\mathcal{R}$ has been clustered in $\mathcal{C}_p$. Thus, the p^{th} macroscopic material represented by $\mathcal{C}_p$ is not contained in $\mathbf{x}_k$ and $\alpha_{p,k} = 0$.
- There is a single local endmember $\mathbf{s}_i^\mathcal{R}$ (for some $i \in [\![1, ..., d_\mathcal{R}]\!]$) belonging to cluster $\mathcal{C}_p$. Therefore, the proportion of this material shall not change within $\mathbf{x}$, hence $\alpha_{p,k} = a_{i,k}^\mathcal{R}$.
- There are several local endmembers $\mathbf{s}_{i_n}^\mathcal{R}, n = 1, \ldots, m$ grouped in the same cluster $\mathcal{C}_p$. In such situation, the material is locally variable within the region $\mathcal{R}$ (healthy and burnt grass for instance), but all contributions sum up with respect to the global instance of the material (being grass in the previous example), as in the case of spectral bundles. Thus, $\alpha_{p,k} = \sum_{n=1}^m a_{i_n,k}^\mathcal{R}$.

All previous cases can be summarized as follows:

$$\alpha_{p,k} = \sum_{i=1}^{d_\mathcal{R}} a_{i,k}^\mathcal{R} \mathbb{1}_{\{\mathbf{s}_i^\mathcal{R} \in \mathcal{C}_p\}} \ , \tag{6.3}$$

where $\mathbb{1}_{\{\mathbf{s}_i^\mathcal{R} \in \mathcal{C}_p\}} = 1$ if $\mathbf{s}_i^\mathcal{R} \in \mathcal{C}_p$ and 0 otherwise. Doing so for all pixels $\mathbf{x}_k$, $(k = 1, ..., N)$ of the HSI allows to reconstruct global fractional abundance maps.

6.3.1.3 Estimation of spectral variability

In the ELMM framework, each endmember $\mathbf{s}_p$ is authorized to vary pixelwise (here in pixel $\mathbf{x}_k$) with respect to some reference endmember $\mathbf{s_0}_p$ according to some local scaling factor ψ_{pk}, as described by Eq. (5.11). Here, we take advantage of this idea by considering cluster centroids to be those reference endmembers and modeling all local endmembers belonging to this cluster $\mathcal{C}_p$ as some scaled versions of $\mathbf{s}_{0p}$:

$$\mathbf{s}_i^{\mathcal{R}} \in \mathcal{C}_p \Rightarrow \mathbf{s}_i^{\mathcal{R}} = \phi_i^{\mathcal{R}} \mathbf{s}_{0p} \ . \tag{6.4}$$

The scaling factor $\phi_i^{\mathcal{R}}$ associated with the local endmember $\mathbf{s}_i^{\mathcal{R}}$ can be recovered in practice by least squares regression between $\mathbf{s}_i^{\mathcal{R}}$ and the centroid $\mathbf{s}_{0k}$ of cluster $\mathcal{C}_p$ it belongs to:

$$\phi_i^{\mathcal{R}} = \left(\mathbf{s}_{0p}^{\top} \mathbf{s}_{0p}\right)^{-1} \mathbf{s}_{0p}^{\top} \mathbf{s}_i^{\mathcal{R}} \ , \tag{6.5}$$

Besides, Eq. (6.5) guarantees the local scaling factor $\phi_i^{\mathcal{R}}$ to be nonnegative.
Plugging Eq. (6.4) into Eq. (6.1) yields

$$\mathbf{x}_k = \sum_{i=1}^{d_{\mathcal{R}}} a_{i,k}^{\mathcal{R}} \phi_i^{\mathcal{R}} \mathbf{s}_{0p_i} + \mathbf{e}_k \ , \tag{6.6}$$

where $p_i \in \{1, \ldots, P\}$ is the index of the cluster $\mathbf{s}_i^{\mathcal{R}}$ belongs to. Eq. (6.6) can be rewritten as

$$\mathbf{x}_k = \sum_{p=1}^{P} \left(\sum_{i=1}^{d_{\mathcal{R}}} a_{i,k}^{\mathcal{R}} \phi_i^{\mathcal{R}} \mathbb{1}_{\{\mathbf{s}_i^{\mathcal{R}} \in \mathcal{C}_p\}} \right) \mathbf{s}_{0p} + \mathbf{e}_k \ . \tag{6.7}$$

On the other hand, pixel $\mathbf{x}_k$ can also be decomposed with respect to the global ELMM framework as described by Eq. (5.11), namely

$$\mathbf{x}_k = \sum_{p=1}^{P} a_{p,k} \psi_{p,k} \mathbf{s}_{0p} + \mathbf{e}_k = \sum_{p=1}^{P} \left(\sum_{i=1}^{d_{\mathcal{R}}} a_{i,k}^{\mathcal{R}} \mathbb{1}_{\{\mathbf{s}_i^{\mathcal{R}} \in \mathcal{C}_p\}} \right) \psi_{p,k} \mathbf{s}_{0p} + \mathbf{e}_k \ , \tag{6.8}$$

with $\psi_{k,p}$ being the global scaling factor associated with centroid $\mathbf{s}_{0p}$ in pixel k. Hence, if we identify the terms in Eqs. (6.7) and (6.8), it is possible to estimate the global scaling factor $\psi_{p,k}$ for pixel $\mathbf{x}_k$ as a weighted average of its local scaling factors $\phi_i^{\mathcal{R}}$ and local abundances $a_{i,k}^{\mathcal{R}}$:

$$\psi_{p,k} = \frac{\sum_{i=1}^{d_{\mathcal{R}}} a_{i,k}^{\mathcal{R}} \phi_i^{\mathcal{R}} \mathbb{1}_{\{\mathbf{s}_i^{\mathcal{R}} \in \mathcal{C}_p\}}}{\sum_{i=1}^{d_{\mathcal{R}}} a_{i,k}^{\mathcal{R}} \mathbb{1}_{\{\mathbf{s}_i^{\mathcal{R}} \in \mathcal{C}_p\}}} \ . \tag{6.9}$$

Note that, in the case where there is a single local endmember $\mathbf{s}_i^{\mathcal{R}}$ belonging to cluster $\mathcal{C}_p$, then $\psi_{p,k} = \phi_i^{\mathcal{R}}$ is constant for all pixels $\mathbf{x}_k$ belonging to region $\mathcal{R}$. As a matter of fact, all pixels in $\mathcal{R}$ appear spatially homogeneous with respect to the material represented by $\mathcal{C}_p$. If there are at least two local endmembers belong to the same cluster $\mathcal{C}_p$ on the other hand, then the global scaling factor $\psi_{p,k}$ varies pixelwise. Finally, in the case where $\alpha_{p,k} = \sum_{i=1}^{d_{\mathcal{R}}} a_{k,i}^{\mathcal{R}} \mathbb{1}_{\{\mathbf{s}_i^{\mathcal{R}} \in \mathcal{C}_p\}} = 0$ (that is, if pixel $\mathbf{x}$ does not contain the p^{th} material), then $\psi_{p,k}$ is set to 1.

6.3.1.4 Geometric interpretation

We have seen that with the proposed strategy, we have been able to reinterpret globally the results of LSU, by fitting the ELMM framework within it. With this, we can now provide a global geometrical interpretation of the results of LSU (initially shown in Fig. 2.8). As Fig. 6.1 now reveals, we have defined lines which account for each endmember in the feature space, and have projected each local endmember onto a line (depending on the cluster it was assigned to), thus deriving local scaling factors for each of them. This way, we can derive a new local simplex in each pixel, and recover the global abundances and scaling factors in the sense of the ELMM.

6.3.2 Results

6.3.2.1 Experimental setup

We apply the proposed methodology to the HSI acquired over the campus of the University of Houston in 2012, already used in sections 4.4.2.2 and 5.6.2.1. The subset we use here is composed of 340×320 pixels in spatial dimension, and comprises 144 bands. The study site features an urban area with a stadium, buildings, parking lots and roads, and some portions of grass and trees. A color composition of this HSI is presented in Fig. 6.2. The partition π, input of the proposed methodology, is obtained following the procedure described in section 2.3.2 and in [161]. First a spatial pre-processing of the HSI is conducted in order to mitigate the effects of potential outliers [175]. Then, we build a BPT representation [155] of the dataset using the spectral region model proposed in [161] (modeling all regions by their local endmembers) and the endmember-based distance as merging criterion [68]. A LSU procedure is conducted over each region of the BPT. The local intrinsic dimensionality $d_{\mathcal{R}}$ of each region is estimated using the RMT algorithm [32], as it proved to be more reliable than most ID estimation algorithms when working over small regions [54] (see Chapter 3).

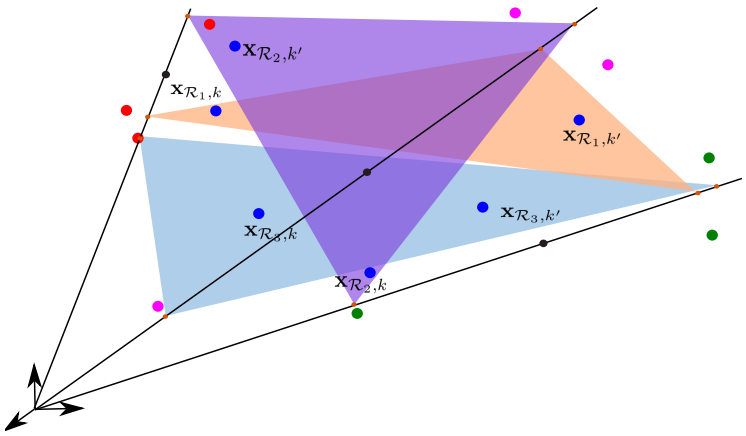
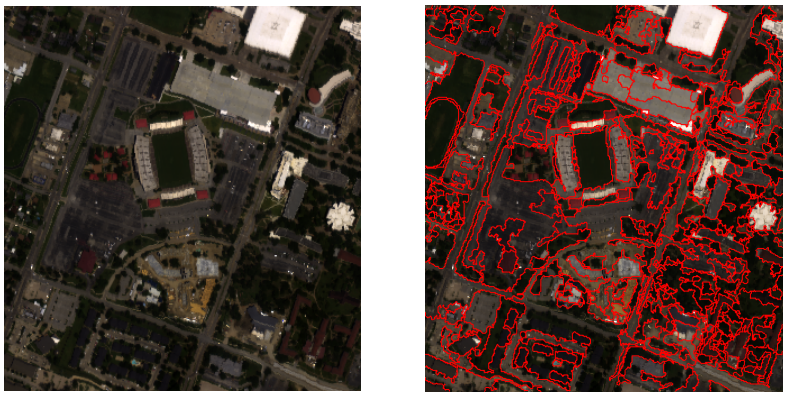


Figure 6.1: Geometric interpretation of LSU in the ELMM framework.



(a) (b)

Figure 6.2: RGB composition of the Houston hyperspectral data set (a), and resulting segmentation composed of 396 regions (b).

Here, we are not using the method proposed in section 4.3 to eliminate the irrelevant endmembers extracted in each region. The reason for this is that here we are not so much interested in the results in local areas as in summarizing this information in global endmembers and abundance maps, using the local endmembers to define the scaling factors of the ELMM. We have seen that when only one local endmember is associated to a certain cluster, there is no spectral variability within the region. In order to allow the scaling factors to be defined pixelwise, and not regionwise, it is interesting to keep several representatives of each cluster in every region. The approach of section 4.3 was more concerned with inter-region spectral variability, while here we are also interested in intra-region spectral variability. Local endmembers are estimated using the vertex component analysis (VCA) algorithm [122] and the fractional abundances are retrieved using the FCLSU algorithm. The partition π extracted from the BPT structure is displayed by Fig. 6.2 and composed of 396 regions. It achieves a trade-off between low region-wise maximal reconstruction errors (penalizing large regions with potential high reconstruction errors, which may be caused by the invalidity of the LMM within the region) and simplicity (penalizing partitions with too many regions). We refer to [161] or [150] for practical details.

All generated local endmembers are grouped in the set $\mathcal{S}_\pi$, eventually composed of 2957 individuals. This set $\mathcal{S}_\pi$ is divided into $P = 12$ clusters following a multivariate Gaussian mixing model hypothesis, by application of the Expectation-Maximization algorithm [147]. Note that the total number of clusters P has been set empirically. According to the proposed methodology, all cluster centroids $\mathbf{s}_p, p = 1, \ldots, P$ are defined to be the ELMM reference endmembers. The associated global abundances $\alpha_{p,k}$ and global scaling factors $\psi_{p,k}$ are retrieved for all pixel spectra in the HSI following the procedures exposed in section 6.3.1.2 and section 6.3.1.3, respectively.

In order to evaluate the performance of the proposed methodology, we also unmix the image following the classical LMM scenario: P "classical" endmembers and associated fractional abundances are globally induced over the image, using the same set-up as the LSU (namely

VCA for the endmember induction FCLSU for the abundances retrieval). In both cases, in the absence of ground truth, the quality of the unmixing any pixel spectrum is evaluated by its RMSE.

6.3.2.2 Results

Fig. 6.3 presents the obtained results for the semantic classes Asphalt, Vegetation and Metallic roofs. The first row of Fig. 6.3 displays the clusters obtained by the proposed strategy, where each blue spectrum depicts a local endmember obtained by LSU, the red spectrum is the cluster centroid, and the black spectrum is the corresponding endmember induced using the classical global approach. The second and third rows of Fig. 6.3 exhibit the fractional abundance maps for the classical global approach, and the proposed approach, respectively (with scales ranging from 0 (blue) to 1 (red)). Finally, the bottom row displays the global scaling factors obtained by the proposed approach. Their values range from 0.5 (in blue) to 1.5 (in red). As remarked in section 5.6.1.1 and in [55], scaling factors are only relevant if the associated fractional abundances are high enough (greater than 0.3 in practice). In the opposite situation, the contribution of the associated endmembers to the pixel spectra cannot be considered significant enough to reliably estimate their variability. In such case, scaling factors have been rounded to 1 for visualization purposes.

As it can be seen on the top row, the obtained clusters are spectrally coherent in the sense that all local endmembers grouped in the same cluster differ only from a scaling factor, which empirically validates the ELMM base assumption. Cluster centroids and classical global endmembers are also similar, up to some scaling factor, which confirms that the former can indeed be considered as global endmember instances. Nevertheless, while the obtained fractional abundance maps appear comparable for the Asphalt and Vegetation classes, which are well present across the image, it is different for the scarce Metallic roofs class. Several metallic roof endmembers have been extracted thanks to the LSU approach and clustered together, allowing to retrieve a clean abundance map, while the global approach leads to an abundance map where other structures are visible.

Finally, the obtained scaling factor maps appear visually consistent, as the observed variations can be linked to the different shades within the parking lots for the Asphalt class for instance, or to the topography of the scene, as it is the case for Vegetation class (where it is possible to distinguish between trees with low scaling factors, and grass with higher scaling factors). It is even clearer when looking at the scaling factors associated to the Metallic roof class, as shown in Fig. 6.4. While the abundances of the classical global approach show some variability due to the topography of the two red roofs, this variability is clearly supported in our approach by the scaling factors, while the abundances remain relatively pure.

Fig. 6.5 displays the reconstruction error maps of the proposed (b) and classical global (a) approaches. As demonstrated in Fig. 6.5 (c), the proposed approach globally yields lower reconstruction errors (all white pixels) than the classical approach. It confirms that processing LSU results within the ELMM framework allows to take advantage of both the local validity

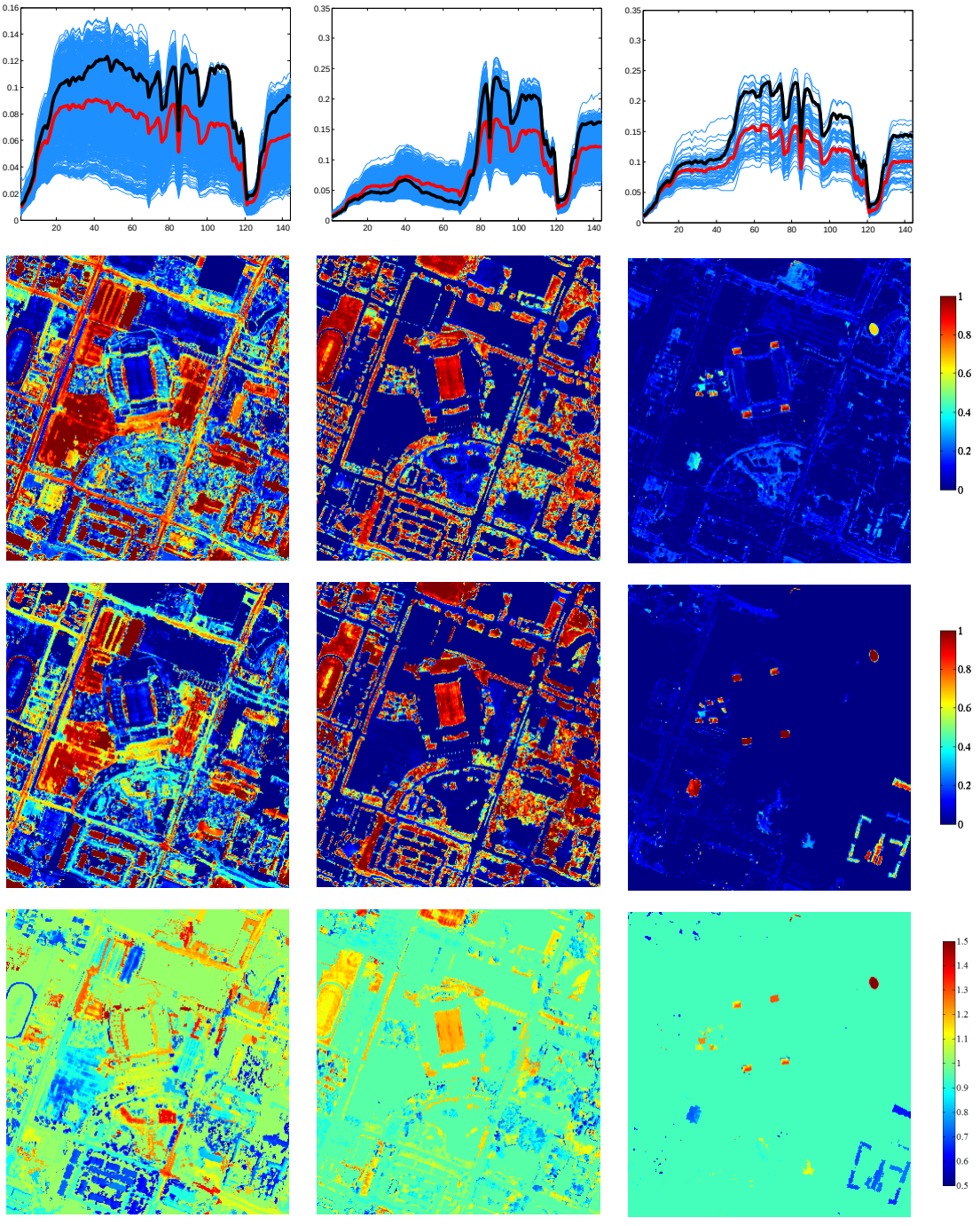


Figure 6.3: First row: obtained clusters (in blue) along with their centroids (in red) and the classical global endmembers (in black). Second and third rows: fractional abundance maps associated to the global endmembers and the cluster centroids using the proposed approach, respectively. Bottom row: scaling factor maps obtained by the proposed methodology.

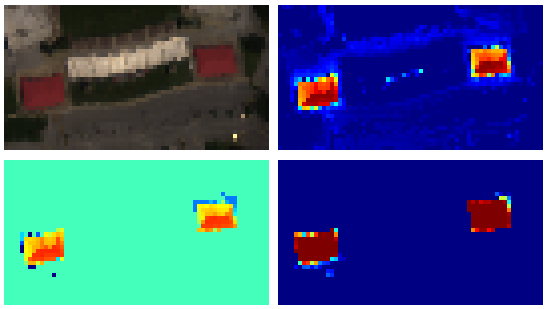
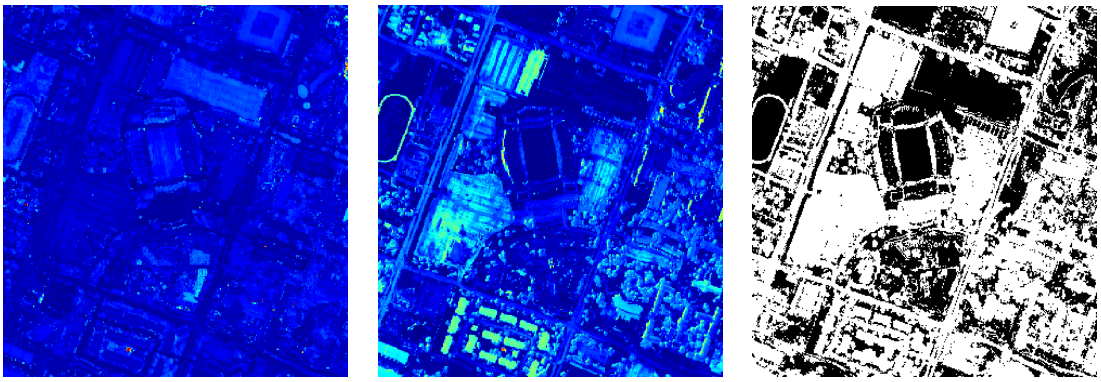


Figure 6.4: Top row: crop of the RGB image (left), and global abundances (right) associated to the Metallic roof class. Bottom row: obtained scaling factors (left) and abundances (right) for the Metallic roof class.



(a) (b) (c)

Figure 6.5: Reconstruction map for (a) the proposed approach and (b) the classical global approach (the scale is saturated between 0 (blue) and 0.1 (red)), and (c) binary comparison of the two (white if the reconstruction is better in (a) and black if it is better in (b)).

of the LMM and the global variability of endmembers to better model the hyperspectral data set.

6.4 Tensor CP decomposition of hyperspectral data

The objective of this section is to make the explicit connection between nonnegative tensor CP decomposition of hyperspectral datasets and spectral variability, through the use of the ELMM. We will first show that performing the nonnegative CP decomposition of a third order hyperspectral tensor is equivalent to solving a regularized version of a SU problem using the ELMM, where the scaling factors are able to account for variability in the third modality (e.g. temporal or angular variability). The derivation of this equivalence is adapted from [156]. Next, we present, as another application of this, a new representation of a regular HSI as a third order tensor, so as to deal with SV in the spatial domain, which has been our main

focus so far. This representation was first introduced in [156].

6.4.1 Connection between the CP tensor decomposition in HSI processing and the ELMM

Here, we assume that we have a third order data tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{L \times N \times T}$ at our disposal, where T is the dimension of the third way of the tensor, representing the third modality (time, angle, or possibly some other relevant modality). The best rank R approximation of $\boldsymbol{\mathcal{X}}$ (see Chapter 2) is the solution to:

$$\begin{aligned} &\underset{\mathbf{S},\mathbf{A},\mathbf{T}}{\arg\min} \; ||\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{I}} \times_1 \mathbf{S} \times_2 \mathbf{A}^\top \times_3 \mathbf{T}||_F^2 \\ &\text{s.t. } \mathbf{S} \geq \mathbf{0}, \mathbf{A} \geq \mathbf{0}, \mathbb{1}_R^\top \mathbf{A} = \mathbb{1}_N^\top, \mathbf{T} \geq \mathbf{0}, \end{aligned} \tag{6.10}$$

where $\mathbf{S} \in \mathbb{R}^{L \times R}$, $\mathbf{A} \in \mathbb{R}^{R \times N}$, $\mathbf{T} \in \mathbb{R}^{T \times R}$. Note that here we have included the ASC in the CP decomposition in the term $\mathbb{1}_R^\top \mathbf{A} = \mathbb{1}_N^\top$. With respect to the decomposition of Eq. (2.22), we have incorporated the scaling indeterminacies in the $\mathbf{T}$ matrix, so that $\boldsymbol{\mathcal{I}} \in \mathbb{R}^{R \times R \times R}$ is a diagonal tensor of ones. If the first two modalities are the spectral and spatial modalities, $\mathbf{S}$ and $\mathbf{A}$ can be interpreted as the endmember spectra and the abundances, respectively. The interpretation of $\mathbf{T}$ depends on the third modality. The underlying multilinear model of the CP decomposition is now:

$$x_{lkm} = \sum_{r=1}^{R} s_{lr} a_{rk} t_{mr}. \tag{6.11}$$

If we denote by $\mathbf{X}_m \in \mathbb{R}^{L \times N}$ a slice of $\boldsymbol{\mathcal{X}}$ for a particular index m in the third way (e.g. one time frame for time series), Eq. (6.11) allows us to decompose it as:

$$\mathbf{X}_m \approx \sum_{r=1}^{R} \mathbf{s}_r t_{mr} \mathbf{a}_r = \mathbf{S} \boldsymbol{\psi}_m \mathbf{A}, \tag{6.12}$$

where $\boldsymbol{\psi}_m \in \mathbb{R}^{R \times R}$ is a diagonal matrix, with the scaling factors t_{mr} on its diagonal. This makes the expression of each slice very similar to the ELMM (Eq. (5.11)), except that the scaling factors can apply to any modality. Note that a related model was used in [78] for the dynamical unmixing of multitemporal images, but with a smooth variation model for the abundances and spectral variations in the temporal domain. With Eq. (6.12), we can rewrite the CP decomposition problem (6.10) as:

$$\begin{aligned} &\underset{\mathbf{S},\mathbf{A},\mathbf{T}}{\arg\min} \; \sum_{m=1}^{T} ||\mathbf{X}_m - \mathbf{S} \boldsymbol{\psi}_m \mathbf{A}||_F^2 \\ &\text{s.t. } \mathbf{S} \geq \mathbf{0}, \mathbf{A} \geq \mathbf{0}, \mathbb{1}_P^\top \mathbf{A} = \mathbb{1}_N^\top, \boldsymbol{\psi}_m \geq 0, \forall m \in [\![1, ..., T]\!]. \end{aligned} \tag{6.13}$$

This equation means that performing the rank R decomposition of a third order hyperspectral tensor is equivalent to solving a blind regularized version of the ELMM, where the abundance matrix is the same for each slice of the data tensor in the third modality, and the endmember matrix is the same in each slice, up to a spectral variability-related correction, encoded in

the scaling factor matrices. In practice, this means that the estimated abundance matrix is an average of the actual abundances in each slice. This explains why the abundance maps obtained by the CP decomposition are usually relatively smooth in the spatial domain [158], since the values incorporate information from all slices. This relationship between the ELMM and the CP decomposition of hyperspectral tensors allows to provide a physical interpretation of the CP decomposition, since it actually performs SU of the data, accounting for SV in its third modality in the form of scaling factors. For time or angular series, these scaling factors model the per-frame variations in illumination or intrinsic variability. For instance, in [158], the temporal factors for different types of snow are shown to be related to seasonal variations, while in [157], the angular factors follow the variations of the viewing angle in each frame. In any case, the main difference with the ELMM is that SV is only taken into account in the domain of the third modality, whereas the ELMM models SV in the spatial domain. The next section will present a representation of HSI data as a third order tensor for which the third modality is connected to SV in the spatial domain.

6.4.2 Hyperspectral Patch Tensor

HSIs are often represented as data cubes, with two spatial dimensions and one spectral dimension. This cube is already a three-way tensor, which has been used as is in the literature [173], for denoising or compression applications mostly. However, this representation of hyperspectral data as a tensor is not well suited to CP decomposition, for two reasons. The first is that the data cube is not a low rank tensor, as we will see below. The second is that the interpretation of the spatial factors of the CP decomposition is not as evident as for the tensor representation of time and angular series we have described so far. In addition, the interpretation of the third way with the ELMM no longer makes sense. It is then preferable to see both spatial dimensions as a single modality. In order to be able to use this interpretation, and to deal with SV in the spatial domain without having an explicit third modality, another representation of the data as a tensor must be used.

In order that the third modality should contain spatial information, we define, for each pixel $\mathbf{x}_k$ of a HSI, a patch comprising all the pixels its neighborhood (defined, for instance as the pixels in a window centered at $\mathbf{x}_k$). Let us denote by $\mathbf{P}_k = [\mathbf{x}_{k_1}, \mathbf{x}_{k_2}, \cdots, \mathbf{x}_{k_T}] \in \mathbb{R}^{L \times T}$ such a patch (stored into a matrix), with the center pixel $\mathbf{x}_k$ being $\mathbf{x}_{k_1}$. T is the number of pixels in the patch. By stacking the obtained patches for each pixel of the image, we obtain a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{L \times N \times T}$, where the third modality is the spatial neighborhood of each pixel. Since the approach is based on sliding windows, border effects have to be handled somehow (using periodic, symmetric or zero-padded boundaries). The construction of this patch tensor is illustrated in Fig. 6.6.

Performing a CP decomposition on this tensor then provides spatial and spectral factors, as usual, but also a matrix $\mathbf{T} \in \mathbb{R}^{T \times R}$ of "neighborhood" factors. With the regularized ELMM formulation of the CP decomposition in mind, these factors account a dictionary of R spatial patches of local scaling factors, representing the atoms of a dictionary of average spectral variability patterns in the spatial domain. In this context, Eq. (6.13) can be reinterpreted as

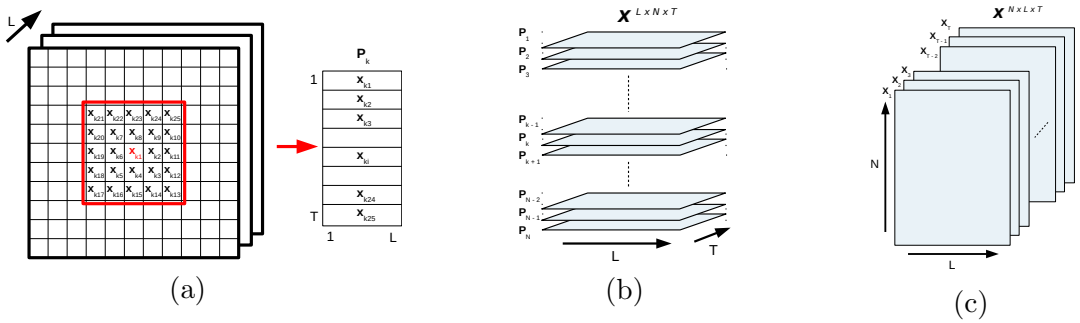


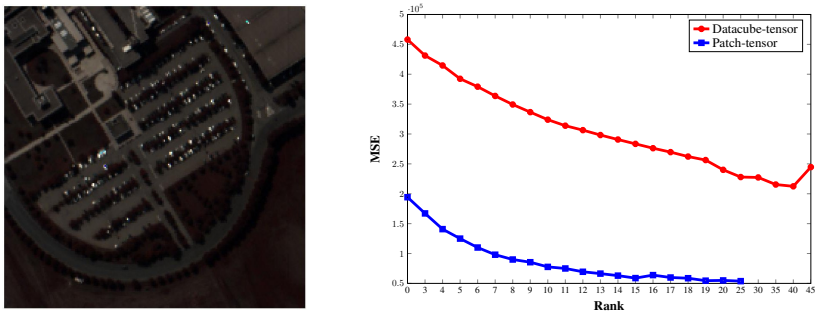
Figure 6.6: Construction of the patches (left), construction of the patch tensor by stacking the patches for each pixel (middle), alternative construction by shifting the image along the neighborhood dimension (right).

performing the SU of the HSI, accounting for SV in the spatial domain thanks to these local scaling factors. The constraint that the abundances are equal in all slices in the third modality means here that abundances within a neighborhood should be correlated. In practice, this acts as a spatial regularization on the abundances. As suggested in [156], this constraint may be a bit harsh in practice, but can be easily relaxed by incorporating weights w_m to each of the T terms, accounting for each pixel of the mask of the spatial neighborhood. For example, these weights can be defined according to a Gaussian kernel of the same size as the sliding window, so as to give more importance to center pixels of the mask in the spatial regularization. With the weighting scheme, the optimization problem becomes:

$$\begin{aligned} &\underset{\mathbf{S},\mathbf{A},\mathbf{T}}{\arg\min} \sum_{m=1}^{T} w_m ||\mathbf{X}_m - \mathbf{S}\boldsymbol{\psi}_m\mathbf{A}||_F^2 \\ &\text{s.t. } \mathbf{S} \geq \mathbf{0}, \mathbf{A} \geq \mathbf{0}, \mathbb{1}_P^\top \mathbf{A} = \mathbb{1}_N^\top, \boldsymbol{\psi}_m \geq 0, \forall m \in [\![1, ..., T]\!]. \end{aligned} \tag{6.14}$$

Here, $\mathbf{X}_m$ is a slice of the patch tensor for the m^{th} element of the mask of the neighborhoods. It is actually the original HSI $\mathbf{X}$, except that it has been shifted vertically and horizontally along the neighborhood: $\mathbf{X}_m = [\mathbf{x}_{1+d_m}\mathbf{x}_{2+d_m}, ..., \mathbf{x}_{N+d_m}]$, where d_m is a spatial displacement from the center of the neighborhood mask to its m^{th} element. This equivalent construction is shown in the right of Fig. 6.6.

In order to justify that this approach is better suited to CP decomposition than directly using the HSI cube, we show in Fig. 6.7 the MSE on the CP decomposition of the HSI represented as the conventional data cube (red) and the patch tensor (blue), as a function of the rank for the decomposition for a subset of the well known Pavia university dataset. This HSI was collected by the ROSIS-03 sensor over the facilities of the University of Pavia in Italy. After discarding pixels with no information and noisy spectral bands, the image has a spatial size of 610 × 340 pixels with a spatial resolution of 1.3 m, and 93 spectral bands in the 430-860 nm range. The subset is a 200 × 200 pixels size crop of the bottom left part of the original image. Fig. 6.7 shows a false color representation of the subset. The scene shows an urban area, comprising a parking lot, buildings, roads and other typical man-made constructions, together with trees, green areas and bare soil.



(a) (b)

Figure 6.7: RGB representation of the chosen subset of the Pavia University dataset (a). MSE of the CP decomposition of the subset of the Pavia dataset as a function of the chosen rank (b), when seen as a data cube (red), or represented as a patch tensor (blue).

First we can notice that regardless of the chosen rank, the MSE is lower for the patch tensor than for the data cube. In addition, the MSE decreases until R = 15 for the patch tensor, and remains almost constant after that, while it decreases until R = 40 for the data cube, suggesting that the data cube is not a low rank tensor, whereas the patch tensor has a lower rank, making it much more suited to CP decomposition.

As an illustration of the connection between the ELMM and the CP decomposition (using unit weights and 5×5 patches), we show some results on a toy example, in which 3 endmember spectra were randomly drawn from the USGS spectral library. The abundance maps are synthesized with Gaussian Random Fields, and a simple spectral variability map is made using 2D Gaussian patterns, following the ELMM. We show in Fig. 6.8 the abundance maps estimated by FCLSU and the CP decomposition of the patch tensor, with a rank $R = 3$. Since there is a significant amount of spectral variability, FCLSU performs rather poorly in this situation, since the scaling factors cause confusion between the endmembers. The CP decomposition of the patch tensor is able to estimate more precise abundance maps, since it is connected to the ELMM, where the scaling factors act on the neighborhoods of each pixel (or equivalently, on each of the T shifted images). We can see that the abundance maps are spatially relatively smooth, in the absence of an explicit spatial regularization. This comes from Eq. (6.14), which actually promotes the abundances to be similar inside a neighborhood. Finally, we show in Fig. 6.9, the extracted neighborhood factors for each material. For each material, the scaling factor patch has a spatial structure which matches a favored direction of the corresponding actual variability pattern. This structure is very directional for the non isotropic Gaussians, while it seems more circular for the last one. More precisely, for each material, the structure of the scaling factor patch reflects the spatial patterns of the high abundance areas of the image for this material where there are in addition significant spectral variability effects. We have already seen that SV properties within a pixel can only be extracted in a dataset for a given material if its abundance contribution is sufficient. The structure of the patches then has to be put in relation with the true SV patterns used, but only with (relatively) high abundance areas. For the leftmost material of Figs. 6.8 and 6.9, the abundance and scaling factors are important only in the area corresponding to the right

of the SV Gaussian pattern. Then the scaling factor patch has a pattern which matches the diagonal edge of the Gaussian in that part of the image. For the material in the middle, the abundance is high for most of the Gaussian pattern, and then the patch is able to recover its orientation. Finally, for the rightmost material, the abundance is significant for most of the area covered by the Gaussian, and hence the scaling factor patch picks up the circularity of the isotropic Gaussian pattern. In the end, the scaling factor patches can be thought of a dictionary of favored spatial variability patterns for each material.

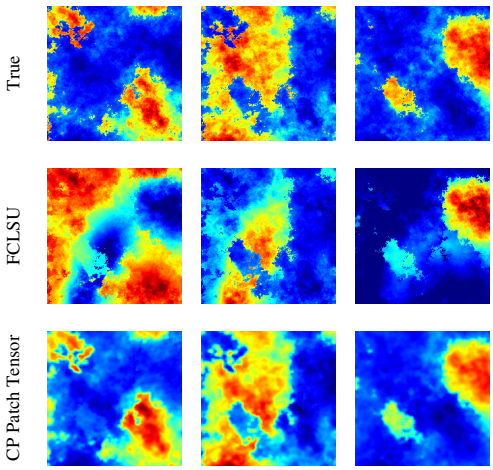


Figure 6.8: True and estimated abundance maps using FCLSU and the CP decomposition of the patch tensor, for the synthetic dataset.

However, since the third modality is not independent from the spatial modality, the patch tensor comprises some redundant information, which can be the source of identifiability issues in complex scenarios. A solution to avoid this would be to define a priori reference endmembers to be the spectral factors (as for the ELMM) in order to avoid confusion between the different factors of the CP model. The patch tensor approach was successfully applied to synthetic datasets [156], but these preliminary results remain to be confirmed on more realistic synthetic

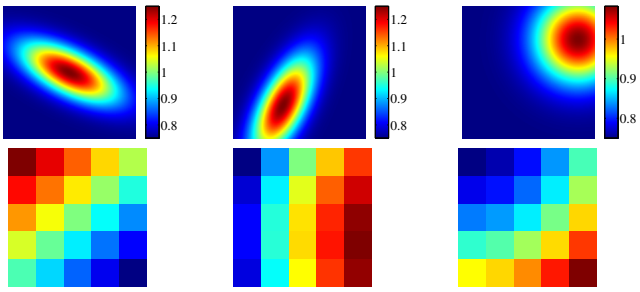


Figure 6.9: Spatial variability patterns (top) for each endmember, used to generate the data. Corresponding scaling factor patches in the CP decomposition of the patch tensor (bottom). The range of values is not relevant because of the scaling indeterminacies of the CP model.

data, as well as on real datasets after this ambiguity issue has been addressed.

6.5 Partial Conclusion

In this Chapter, we have presented two applications of the ELMM. The first one combines the LSU and ELMM approach in order to interpret the local results of the BPT based-LSU at the scale of the global image, taking advantage of the local information to extract spectral variability under the form of the scaling factors of the ELMM. An interesting alternative could be to directly incorporate the ELMM in the region model, and to define a merging criterion taking the local scaling factors into account. The second part of the chapter showed how the CP decomposition of tensors representing multimodal (the additional modality being time, angle or spatial neighborhood) hyperspectral tensors is actually connected to the ELMM, a regularized version of which is equivalent to the CP decomposition problem. The patch tensor approach could be made more robust by using reference endmembers, as was done for the ELMM. The investigation of the effects of weighting of the spatial neighborhoods, as well as the validation of the approach on real datasets need to be addressed in the future. Also, it could be interesting to see if more variability can be explained by increasing the rank of the decomposition, so as to have more variability patches, even though it comes with more spectra for the endmembers, which may require a clustering step to interpret the results, as in the bundles approach.

Conclusion and Perspectives

In this thesis, we have addressed the problem of spectral unmixing (SU) of hyperspectral remote sensing images, through the prism of spectral variability (SV). The SU problem has been extensively studied in the past few decades, going from conventional linear SU to more complex nonlinear models. However, the fact that any endmember always exhibits significant SV, although long known, has only begun to be addressed explicitly by the community in the unmixing problem. We have reviewed in **Part I** the main methods and techniques for linear SU (**Chapter 1**) and the existing methods prior to this work which address SV (**Chapter 2**). Then we have detailed our contributions on this topic in two parts, each exploring a different aspect of the problem. We summarize them here and lay out some research perspectives for this thesis (in italic).

Part II described how sparsity could be of use for the SV problem in the context of the Linear Mixing Model (LMM). First, we have evidenced some limitations of Local Spectral Unmixing (LSU). The first of these, addressed in detail in **Chapter 3**, is related to the problem of estimating the intrinsic dimensionality (ID) (which is linked to the number of endmembers to consider) in local regions of the image, showing that this estimation could be highly impacted by the number of observations and spectral bands in each region, as well as the number of bands, and the noise estimation technique used. The results of this study have been applied in order to improve local ID estimation in LSU. However, the number of endmembers to consider is still often overestimated in many cases.

To alleviate this issue, which impedes the interpretation of LSU results in each region, we have proposed, in the first part of **Chapter 4**, a new algorithm based on collaborative sparsity to eliminate the wrongly estimated endmembers in each region, and we have shown on a real dataset that it allows to eliminate the redundant information in each region, making the results much easier to interpret at the region scale. The remainder of this chapter dealt with the SV problem using spectral bundles, and more precisely on how we could use different forms of sparsity to estimate the abundances, while taking into account the group structure of the bundles. We have tested three variants to do this, a group penalty, favoring a sparse number of materials to be active in each pixel, an elitist penalty, which favors within-group sparsity, and a nonconvex fractional one, which combines the effects of the latter two in a single term, while being compatible with the ASC. The group and elitist penalties are known in the literature, but to the best of our knowledge, this is the first time a mixed $\mathcal{L}_{1,q}$ fractional norm is used for group sparsity. It is also the first time that the group structure of the bundles is explicitly taken into account in the abundance estimation problem. We have shown on real and synthetic datasets that the proposed penalties (except the elitist one on real data) were able to improve the unmixing performance with respect to using FCLSU on the extracted bundle, with the best results occurring for the group and fractional penalties.

The new region model we proposed for the LSU using collaborative sparsity, based on obtaining a regularization path for the collaborative unmixing problem in each region, and the

selection the optimal model based on the BIC, could be easily adapted to a whole HSI to design a completely unsupervised and deterministic linear SU algorithm. The same approach can indeed be applied to the joint endmember extraction and abundance estimation problem alluded to in section 1.5.3, where the image data forms a self-dictionary of endmembers (using the pixel purity assumption). The resulting algorithm would be able to jointly estimate the appropriate number of endmembers to use, the corresponding endmember signatures, and their fractional abundances, in a non-stochastic way, and without any parameter to tune. Depending on the value of the regularization parameter, SV could even be included by voluntarily allowing more atoms of the dictionary to be active, and clustering these active atoms into bundles.

Perspectives for the sparse unmixing on bundles include a theoretical proof of the convergence of the Alternating Direction Method of Multipliers (ADMM) in the nonconvex fractional case, the use of spatial information (under the form of a well chosen TV-like penalty) to be able to sparsify the abundance maps while preventing the abundance maps to become noisy. Obtaining the regularization paths for the optimization problems we solve could also be useful to avoid having to tune the regularization parameters. Finally, using better ways to extract the spectral bundles could also improve the performance of this type of approaches. For instance, clustering the optimal partition of the LSU approach could be a way to define spectral bundles which take into account spatial information. This would also allow to interpret the LSU results at a global scale (in an alternative to the approach of section 6.3).

Part III explored a different avenue to handle SV in the SU problem, by introducing in **Chapter 5** an explicit mixing model taking variability into account. In order to account for the variability induced by illumination changes and uneven topography in the observed scene, we designed an Extended Linear Mixing Model (ELMM), which introduced pixel and material dependent scaling factors to modify locally the endmembers. This model, generalizing the LMM, is both geometrically and physically interpretable, since it can be derived by making simplifying physical assumptions in the Hapke model, used in the physics and planetary science communities to express reflectance as a function of albedo, the geometric parameters of the scene, as well as the photometric parameters of the materials. We included spatial regularizations on the abundances and scaling factors to be able to better estimate SV in the scene, and designed two algorithms to solve the resulting optimization problem. We have shown on two synthetic datasets, including one simulated using the Hapke model that the proposed method obtains state-of-the-art unmixing performance. We also show on real datasets how both the abundance and scaling factor maps are easily interpretable and help refining SU results.

We also presented two applications of the ELMM in **Chapter 6**. The first is a technique which allows to combine all the local information obtained in a LSU approach into global results, extracting global abundances and at the same time explain the local results in terms of SV using the scaling factors advocated by the ELMM. We showed on a real dataset the interest of this approach, compared to conventional SU. The second application is the nonnegative tensor Canonical Polyadic (CP) decomposition of hyperspectral data, which can be shown to be equivalent to solving a blind regularized version of the ELMM, in which the scaling factors explain the variability on the third modality (time, angle, or neighborhood), and can in any

case be physically interpreted.

The ELMM could be combined to nonlinear mixing models of the literature in order to allow the joint extraction of information related to both phenomena. Current nonlinear models can indeed interpret SV as nonlinearities, while the contrary could happen for SV accounting mixing models. The ELMM could be further refined by trying to approximate radiative transfer models such as the Hapke model in a more complex, but still tractable way. Taking advantage of a Digital Terrain Model (DTM) (for example coming from LiDAR) data could help obtain the acquisition angles of in each pixel, which could be fed to a more complex mixing model to estimate other SV related parameters, such as photometry related ones. These angles could also help finding which areas of the imaged scene are in shadow, which can be of use to perform model selection based on this information (areas in shadow are likely to be nonlinearly mixed). Taking example on the ELMM, new more material specific variability models could be designed from physical reflectance models in the literature, in order to estimate precisely physical parameters which influence the spectral signature of said materials.

The ELMM could also be directly integrated to the LSU framework, in order to design region models and merging criteria accounting for SV. This could allow to isolate nonlinear effects in local regions in a more precise way, since the contribution of SV would be directly accounted for in the region model. Finally, although the preliminary results of the patch tensor approach are promising, more work is required to make it more robust. Using reference endmembers, as in the ELMM could be a lead in this direction. Also, the effects of the spatial regularization using a weighting strategy on the neighborhood is also an interesting research perspective.

List of Publications

International Journals

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." IEEE Transactions on Image Processing 25(8), 2016, pp. 3890–3905

L. Drumetz, M. A. Veganzones, R. M. Gomez, G. Tochon, M. Dalla Mura, G. A. Licciardi, C. Jutten, and J. Chanussot. "Hyperspectral Local Intrinsic Dimensionality." IEEE Transactions on Geoscience and Remote Sensing 54(7), 2016, pp. 4063–4078

L. Drumetz, M. Dalla Mura, S. Meulenyzer, S. Lombard and J. Chanussot. "Semiautomatic classification of cementitious materials using scanning electron microscope images." Journal of Electronic Imaging 24(6), 2015, pp. 061109-061109.

International Conferences

L. Drumetz, S. Henrot, M. A. Veganzones, J. Chanussot, and C. Jutten. "Blind hyperspectral unmixing using an Extended Linear Mixing Model to address spectral variability." IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2015). 2015

M. A. Veganzones, **L. Drumetz**, R. Marrero, G. Tochon, M. Dalla Mura, A. Plaza, J. Bioucas-Dias, and J. Chanussot. "A new extended linear mixing model to address spectral variability." Proc. IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS). 2014

L. Drumetz, J. Chanussot, and C. Jutten. "Endmember variability in spectral unmixing: recent advances." Proc. IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS). 2016

T. R. Meyer, **L. Drumetz**, J. Chanussot, A. L. Bertozzi, and C. Jutten. "Hyperspectral unmixing with material variability using social sparsity." 2016 IEEE International Conference on Image Processing (ICIP). 2016, pp. 2187–2191

G. Tochon, **L. Drumetz**, M. A. Veganzones, M. Dalla Mura, and J. Chanussot. "From Local to Global unmixing of hyperspectral images to reveal spectral variability." IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2016). 2016

L. Drumetz, M. A. Veganzones, R. Marrero, G. Tochon, M. Dalla Mura, A. Plaza, and J. Chanussot. "Binary partition tree-based local spectral unmixing." Proc. IEEE Workshop

on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS). 2014

M. A. Veganzones, J. E. Cohen, R. Cabral-Farias, K. Usevich, **L. Drumetz**, J. Chanussot, and P. Comon. "Nonnegative Canonical Decomposition of Hyperspectral Patch Tensors." European Signal Processing Conference. 2016

L. Drumetz, M. Dalla Mura, S. Meulenyzer, S. Lombard and J. Chanussot. "Semi-automatic classification of cementitious materials using Scanning Electron Microscope images". Proc. SPIE 9534, Twelfth International Conference on Quality Control by Artificial Vision 2015, 2015

National Conference

L. Drumetz, S. Henrot, M. A. Veganzones, J. Chanussot, and C. Jutten. "Démélange aveugle d'images hyperspectrales à l'aide d'un modèle linéaire étendu tenant compte de la variabilité spectrale." XXVème colloque GRETSI. 2015

Under Review

L. Drumetz, G. Tochon, M. A. Veganzones, J. Chanussot, and C. Jutten. "Improved local spectral unmixing of hyperspectral data using an algorithmic regularization path for collaborative sparse regression." Submitted to the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2017, pp. 1-5

L. Drumetz, G. Tochon, J. Chanussot, and C. Jutten. "Estimating the number of endmembers to use in spectral unmixing of hyperspectral data with collaborative sparsity." Submitted to the 13th International Conference Latent Variable Analysis and Signal Separation (LVA-ICA). IEEE. 2017, pp. 1-10

Appendices

Appendix A

Convex Optimization Tools

This appendix briefly introduces most of the convex optimization algorithms and notions used in the manuscript. The notations for the variables used here can be unrelated to those of the rest of the manuscript. The material presented here is discussed in greater details in [42, 22, 44, 14, 12]. At some points of this appendix, we derive fixed point equations for the solutions of certain optimization problems, to motivate the introduction of iterative algorithms. However, note that the derivations of the fixed point equations do not suffice to prove the convergence of these algorithms (which can be found elsewhere anyway), since we would still have to prove some additional properties (i.e. the nonexpansiveness and α-averagedness [12] of the operators which are iterated).

A.1 Useful notions in convex optimization

A.1.1 Proximal operators

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be an extended real-valued lower semicontinuous convex function (denoted by $f \in \Gamma_0(\mathbb{R}^n)$) . For any $\mathbf{x} \in \mathbb{R}^n$, the minimization problem:

$$\underset{\mathbf{y}\in\mathbb{R}^n}{\operatorname{argmin}} \quad f(\mathbf{y}) + \frac{1}{2}||\mathbf{x}-\mathbf{y}||_2^2 \tag{A.1}$$

has a unique solution denoted as $\mathbf{prox}_f(\mathbf{x})$. Using this, we can define the operator $\mathbf{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$.

We can interpret this operator using the following example: if $f = \mathcal{I}_C$, where $\mathcal{I}_C$ is the indicator function of a convex set C, defined as:

$$\mathcal{I}_C(\mathbf{x}) = \begin{cases} 0 \text{ if } \mathbf{x} \in C \\ +\infty \text{ otherwise} \end{cases} . \tag{A.2}$$

The projection of $\mathbf{x} \in \mathbb{R}^n$ is the point in C closest to $\mathbf{x}$, i.e. the point $\mathbf{proj}_C(\mathbf{x})$ solution of:

$$\underset{\mathbf{y}\in\mathbb{R}^n}{\operatorname{argmin}} \quad \mathcal{I}_C(\mathbf{y}) + \frac{1}{2}||\mathbf{x}-\mathbf{y}||_2^2 \triangleq \mathbf{prox}_{\mathcal{I}_C}(\mathbf{x}). \tag{A.3}$$

Then $\mathbf{prox}_f$ can then be seen as a generalization of a projection: we want to find a point close to $\mathbf{x}$ which also leads to a small value of $f(\mathbf{x})$. For a nondifferentiable function f, if this

operator can be efficiently computed (we will say that f is "proximable" in such a case), then it provides a way to "project" a point onto a domain where the nondifferentiable function has a small value. This allows to design algorithms to optimize this type of functions, for which the gradient is not defined, and for which alternatives (e.g. a subgradient descent) are very slow.

It is usually even harder to deal with objective functions with several nondifferentiable terms. The reason for this is that even though $\mathbf{prox}_{\tau f}$ is very easy to compute from $\mathbf{prox}_f$, in general $\mathbf{prox}_{f+g}$ is not.

Here are some classical examples of proximal operators:

- indicator function: $\mathbf{prox}_{\mathcal{I}_C}(\mathbf{x}) = \mathbf{proj}_C(\mathbf{x})$ (e.g. positive orthant: $\mathbf{prox}_{\mathcal{I}_{\mathbb{R}^n_+}}(\mathbf{x}) = \mathbf{x}_+$)
- $\mathcal{L}_1$ norm: $\mathbf{prox}_{\tau||\cdot||_1}(\mathbf{x}) = \text{soft}_\tau(\mathbf{x})$ and $\text{soft}_\tau(\mathbf{x})_i = \text{sign}(x_i)(|x_i| - \tau)_+$
- $\mathcal{L}_2$ norm: $\mathbf{prox}_{\tau||\cdot||_2}(\mathbf{x}) = \mathbf{soft}_\tau(\mathbf{x}) = \left(1 - \frac{\tau}{||\mathbf{x}||_2}\right)_+ \mathbf{x}$
- $\mathcal{L}_\infty$ norm: $\mathbf{prox}_{\tau||\cdot||_\infty}(\mathbf{x}) = \mathbf{x} - \tau\ \mathbf{proj}_{(||\cdot||_1 \leq \tau)}(\mathbf{x})$

A.1.2 Convex conjugate

Another useful notion for proximal algorithms and convex optimization in general is the convex conjugate of a lower semi-continuous function f (but not necessarily convex), defined by:

$$\begin{aligned} f^* : \mathbb{R}^n &\to \mathbb{R} \cup \{+\infty\} \\ \mathbf{u} &\mapsto \sup_{\mathbf{x} \in \mathbb{R}^n} (\mathbf{x}^\top \mathbf{u} - f(\mathbf{x})). \end{aligned} \tag{A.4}$$

Geometrically, for a vector $\mathbf{u} \in \mathbb{R}^n$, $f^*(\mathbf{u})$ is the maximum gap between the linear function $\mathbf{x}^\top \mathbf{u}$ and f [23]. f^* is convex, even when f is not. In addition, if f is differentiable, then the conjugate is a point $\mathbf{u}^*$ where $\nabla_f(\mathbf{u}^*) = \mathbf{u}$ (for a convex function, this property is even a characterization of the conjugate, that is there is a unique point verifying this gradient property).

For proximal operators, the convex conjugate of a function $f \in \Gamma_0(\mathbb{R}^n)$ can intervene through the Moreau decomposition property [119]:

$$\forall \mathbf{x} \in \mathbb{R}^n, \forall \tau > 0, \mathbf{x} = \mathbf{prox}_{\tau f}(\mathbf{x}) + \tau\ \mathbf{prox}_{\tau f^*}\left(\frac{\mathbf{x}}{\tau}\right). \tag{A.5}$$

This allows to compute easily the proximal operator of the conjugate of a function whose proximal operator is known. This is especially useful for $\mathcal{L}_p$ norms ($p \geq 1$), since the convex conjugate of $||\cdot||_p$ is the indicator function of the unit ball of the so-called dual norm $||\cdot||_p^* \triangleq ||\cdot||_q$, where q verifies $\frac{1}{p} + \frac{1}{q} = 1$. From this property, we can deduce:

$$\forall \mathbf{x} \in \mathbb{R}^n,\ \mathbf{prox}(\tau||\cdot||_p)(\mathbf{x}) = \mathbf{x} - \tau\ \mathbf{proj}_{||\cdot||_q<\tau}(\mathbf{x}). \tag{A.6}$$

In particular, the proximal operators of the $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_\infty$ norms defined above can be deduced from this property.

A.1.3 Subdifferential

We define the subdifferential ∂f of $f \in \Gamma_0(\mathbb{R}^n)$ as the set valued operator:

$$\begin{array}{rccl} \partial f: & \mathbb{R}^n & \to & \mathcal{P}(\mathbb{R}^n) \\ & \mathbf{x} & \mapsto & \{\mathbf{u} \in \mathbb{R}^n \backslash \forall \mathbf{y} \in \mathbb{R}^n, (\mathbf{y}-\mathbf{x})^\top \mathbf{u} + f(\mathbf{x}) \leq f(\mathbf{y})\}. \end{array} \tag{A.7}$$

At a point $\mathbf{x}$ where f is differentiable, the subdifferential is a singleton, and in addition $\partial f(\mathbf{x}) = \{\nabla_f(\mathbf{x})\}$. Otherwise $\partial f(\mathbf{x})$ is set valued, and its elements are called *subgradients* of at $\mathbf{x}$. Geometrically, the subdifferential at a point is the set of slopes of the affine functions which are always below f, and coincide with f at this point. The subdifferential can then be seen as a generalization of the gradient field for nondifferentiable convex functions. Its main interest is to allow the derivations of first order optimality conditions even for nondifferentiable functions. Indeed, $\mathbf{x}$ is a minimizer of for $f \in \Gamma_0(\mathbb{R}^n)$ if and only if $\mathbf{0} \in \partial f(\mathbf{x})$. This generalizes the usual first order optimality condition for differentiable functions.

The subdifferential allows to derive the important consequence of definition (A.1):

$$\forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n, \mathbf{p} = \mathbf{prox}_f(\mathbf{x}) \Leftrightarrow \mathbf{x} - \mathbf{p} \in \partial f(\mathbf{p}). \tag{A.8}$$

This property allows in particular to prove the Moreau decomposition property (A.5).

A.1.4 Lipschitz continuity

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be β-Lipschitz continuous if and only if:

$$\forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n, ||\nabla_{f_2}(\mathbf{x}) - \nabla_{f_2}(\mathbf{y})||_2 \leq \beta ||\mathbf{x} - \mathbf{y}||_2. \tag{A.9}$$

Lipschitz continuity is a stronger property than usual continuity since a Lipschitz continuous function is in particular (uniformly) continuous. For example, any norm is 1-Lipschitz because of the triangle inequality. Lipschitz continuity is useful to derive bounds for gradient descent steps, or to compute linear operator norms.

A.2 Proximal gradient algorithm

A.2.1 Algorithm

Now suppose we want to solve the following problem, with f_1 and $f_2 \in \Gamma_0(\mathbb{R}^n)$:

$$\underset{\mathbf{x}\in\mathbf{R}^n}{\operatorname{argmin}} \quad f_1(\mathbf{x}) + f_2(\mathbf{x}). \tag{A.10}$$

We assume that f_2 is differentiable and that its gradient ∇_{f_2} is β-Lipschitz continuous (with $\beta \in]0, +\infty[$). However, we do not make any assumptions on the differentiability of f_1, which precludes the use of the simple gradient descent in general. It can be shown that problem (A.10) has at least one solution and for any $\gamma \in]0, +\infty[$, the solutions verify the fixed point equation:

$$\mathbf{x} = \mathbf{prox}_{\gamma f_1}(\mathbf{x} - \gamma\nabla_{f_2}(\mathbf{x})). \tag{A.11}$$

Proof:

$$\begin{aligned}
&\mathbf{x} \quad \text{is a minimizer of Eq. (A.11)} \\
\Leftrightarrow\ &\mathbf{0} \in \partial(f_1 + f_2)(\mathbf{x}) \quad \text{(first order optimality condition)} \\
\Leftrightarrow\ &\mathbf{0} \in \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}) \quad \text{(sum)} \\
\Leftrightarrow\ &\mathbf{0} \in \partial f_1 + \{\nabla_{f_2}(\mathbf{x})\} \quad (f_2 \text{ is differentiable}) \\
\Leftrightarrow\ &-\nabla_{f_2}(\mathbf{x}) \in \partial f_1(\mathbf{x}) \\
\Leftrightarrow\ &-\gamma\nabla_{f_2}(\mathbf{x}) \in \partial\gamma f_1(\mathbf{x}) \quad \text{(nonnegative scaling)} \\
\Leftrightarrow\ &(\mathbf{x} - \gamma\nabla_{f_2}(\mathbf{x})) - \mathbf{x} \in \partial\gamma f_1(\mathbf{x}) \\
\Leftrightarrow\ &\mathbf{x} = \mathbf{prox}_{\gamma f_1}(\mathbf{x} - \gamma\nabla_{f_2}(\mathbf{x})) \quad \text{(using Eq. (A.8))} \\
\Leftrightarrow\ &\mathbf{x} \text{ is a fixed point of the operator } \mathbf{prox}_{\gamma f_1}(\cdot - \gamma\nabla_{f_2}(\cdot)) \ \blacksquare
\end{aligned}$$

Hence we can find the solutions numerically by iterating the proximal gradient (a.k.a. forward backward) update, with $\gamma_k \leq \frac{1}{\beta}$:

$$\mathbf{x}^{k+1} = \mathbf{prox}_{\gamma_k f_1}(\mathbf{x}^k - \gamma_k\nabla_{f_2}(\mathbf{x}^k)). \tag{A.12}$$

In some cases, the proximal gradient reduces to simple well known algorithms:

- if $f_1 \equiv 0$, Eq. (A.12) reduces to the gradient descent:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k\nabla_{f_2}(\mathbf{x}^k). \tag{A.13}$$

- if $f_2 \equiv 0$, Eq. (A.12) reduces to the proximal point algorithm to minimize a nondifferentiable function:

$$\mathbf{x}^{k+1} = \mathbf{prox}_{\gamma_k f_1}(\mathbf{x}^k). \tag{A.14}$$

- if $f_1 = \mathcal{I}_{\mathcal{C}}$, where $\mathcal{C}$ is a convex set, Eq. (A.12) reduces to the projected gradient method:

$$\mathbf{x}^{k+1} = \mathbf{proj}_{\mathcal{C}}(\mathbf{x}^k - \gamma_k \nabla_{f_2}(\mathbf{x}^k)). \tag{A.15}$$

- if $f_1 = ||\cdot||_1$, Eq. (A.12) reduces to:

$$\mathbf{x}^{k+1} = \text{soft}_{\gamma_k}(\mathbf{x}^k - \gamma_k \nabla_{f_2}(\mathbf{x}^k)). \tag{A.16}$$

A.2.2 Convergence acceleration

If the objective function is separable (or nearly separable), splitting the variable $\mathbf{x}$ it into blocks can help to obtain optimal Lipschitz constants in each block $\mathbf{x}_i$. This technique is very convenient in Coordinate Descent (CD) schemes since it allows to adjust the gradient steps to optimal values (because they are bounded by the inverses of the Lipschitz constants) [126].

For the proximal gradient algorithm, it is in addition possible to use extrapolation techniques to get a better convergence rate, e.g. using the Fast Iterative Shrinkage Thresholding Algorithm (FISTA) [14]:

$$\mathbf{x}_i^{k+1} = \mathbf{prox}_{\gamma_k g}(\hat{\mathbf{x}}_i^k - \gamma_k \nabla f(\mathbf{x}_{\neq i}^k, \hat{\mathbf{x}}_i^k)), \tag{A.17}$$

with $\hat{\mathbf{x}}_i^k = \mathbf{x}_i^k + \omega_k(\mathbf{x}_i^k - \mathbf{x}_i^{k-1})$ and the $\omega_k \geq 0$ define a sequence of carefully chosen decreasing weights [14, 168].

For a simple gradient descent with a smooth (strongly) convex objective, this technique allows to go from a complexity of $\mathcal{O}\left(Q\log\left(\frac{1}{\epsilon}\right)\right)$ iterations to $\mathcal{O}\left(\sqrt{Q}\log\left(\frac{1}{\epsilon}\right)\right)$ iterations to reach a precision of ϵ [125]. $Q = \frac{\beta}{\alpha}$ is a condition number for f, where α is a parameter related to the strong convexity of f. The "nicer" the function (strongly convex with a large parameter α, and of Lipschitz continuous gradient with a low Lipschitz constant β), the lower Q is: so the "harder" the problem, the more efficient the convergence acceleration is. For the more general problem, this accelerated scheme improves the convergence rate in a similar way, since the convergence rate is $\mathcal{O}\left(\frac{1}{k^2}\right)$ after k iterations, vs. $\mathcal{O}\left(\frac{1}{k}\right)$ for the standard algorithm [14].

A.3 Alternating Direction Method of Multipliers

A.3.1 One nondifferentiable term

Now suppose we want to solve the more complex problem ($\mathbf{H} \in \mathbb{R}^{m\times n}$ is the matrix of a linear operator $\mathcal{H} : \mathbb{R}^n \to \mathbb{R}^m$, and neither f nor g are assumed to be differentiable):

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{argmin}} \quad f(\mathbf{x}) + g(\mathbf{H}\mathbf{x}). \tag{A.18}$$

As an illustration, $\mathbf{H}$ can be a difference operator and can g be the $\mathcal{L}_1$ norm so that $g(\mathbf{Hx})$ is a Total Variation term, and the problem becomes a TV denoising problem. The proximal gradient algorithm can still be used here (if f is smooth), but the most popular algorithm to solve this, the Alternating Direction Method of Multipliers (ADMM) [22], rewrites (A.18) as an equivalent problem:

$$\underset{\substack{\mathbf{x}\in\mathbb{R}^n,\mathbf{y}\in\mathbb{R}^m \\ \text{s.t. } \mathbf{Hx}=\mathbf{y}}}{\operatorname{argmin}} \quad f(\mathbf{x}) + g(\mathbf{y}). \tag{A.19}$$

The ADMM writes an augmented Lagrangian (AL) for the problem of Eq. (A.19) :

$$\mathcal{L}(\mathbf{x},\mathbf{y},\mathbf{z}) = f(\mathbf{x}) + g(\mathbf{y}) + \frac{1}{\tau}\mathbf{z}^\top(\mathbf{Hx}-\mathbf{y}) + \frac{1}{2\tau}||\mathbf{Hx}-\mathbf{y}||_2^2, \tag{A.20}$$

and then alternatively minimizes it w.r.t. $\mathbf{x}$ and $\mathbf{y}$, and finally updates the dual variable (which is nothing more than a set of Lagrange multipliers) $\mathbf{z}$ by a proximal maximization step (dual update). Indeed ADMM solves the Lagrange dual problem to (A.19): $\max_{\mathbf{z}} \left(\inf_{\mathbf{x},\mathbf{y}} \mathcal{L}(\mathbf{x},\mathbf{y},\mathbf{z})\right)$. It is easy to verify that the alternative minimization problems lead to the following updates for a given iteration:

$$\begin{aligned}
\mathbf{x}^{k+1} &= \mathbf{prox}^{\mathbf{H}}_{\tau f}(\mathbf{y}^k - \mathbf{z}^k) \\
\mathbf{y}^{k+1} &= \mathbf{prox}_{\tau g}(\mathbf{Hx}^{k+1} + \mathbf{z}^k) \\
\mathbf{z}^{k+1} &= \mathbf{z}^k + \mathbf{Hx}^{k+1} - \mathbf{y}^{k+1},
\end{aligned} \tag{A.21}$$

where $\mathbf{prox}^{\mathbf{H}}_{\tau f} : \mathbb{R}^m \to \mathbb{R}^n$ is defined as $\mathbf{prox}^{\mathbf{H}}_{\tau f}(\mathbf{s}) = \underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{argmin}}\ f(\mathbf{x}) + \frac{1}{2\tau}||\mathbf{Hx}-\mathbf{s}||_2^2$. For example, if f is a classical least squares data fit $f(\mathbf{x}) = \frac{1}{2}||\mathbf{u}-\mathbf{x}||_2^2$, then $\mathbf{prox}^{\mathbf{H}}_{\tau f}(\mathbf{s}) = (\mathbf{I}_n + \frac{1}{\tau}\mathbf{H}^\top\mathbf{H})^{-1}(\mathbf{u} + \frac{1}{\tau}\mathbf{H}^\top\mathbf{s})$. Finally, we mention that ADMM is often used in slightly different form [22], which has the advantage of combining the linear and quadratic terms in the augmented Lagrangian. It is easy to see that by letting $\rho = \frac{1}{\tau}$, minimizing Eq. (A.20) is is the same as minimizing:

$$\mathcal{L}(\mathbf{x},\mathbf{y},\mathbf{z}) = f(\mathbf{x}) + g(\mathbf{y}) + \frac{\rho}{2}||\mathbf{Hx}-\mathbf{y}-\mathbf{z}||_2^2 - \frac{\rho}{2}||\mathbf{z}||_2^2, \tag{A.22}$$

where ρ is the so called *barrier parameter* of the AL. This is the form which we use throughout the thesis. It may happen that we do not write explicitly the last term of Eq. (A.22) in the AL for brevity and simplicity. Indeed, this term is related to the dual update, which is always straightforward (the update is shown in Eq. (A.21)).

A.3.2 Several nondifferentiable terms

Now suppose we deal with a problem of the form:

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{argmin}} \quad f_1(\mathbf{x}) + f_2(\mathbf{x}) + \cdots + f_p(\mathbf{x}). \tag{A.23}$$

ADMM can be extended to multiple terms by writing problem (A.23) in an equivalent form using a technique called variable splitting:

$$\underset{\substack{\mathbf{u},\mathbf{v} \\ \text{s.t. } \boldsymbol{\Sigma}\mathbf{u}=\mathbf{v}}}{\operatorname{argmin}} \; f(\mathbf{u}) + g(\mathbf{v}), \tag{A.24}$$

with

$$\mathbf{u} = \mathbf{x}, \mathbf{v} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_p \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{I}_n \\ \mathbf{I}_n \\ \vdots \\ \mathbf{I}_n \end{bmatrix}, f \equiv 0, g(\mathbf{v}) = f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + \cdots + f_m(\mathbf{x}_p). \tag{A.25}$$

Then the augmented Lagrangian is now separable and can be iteratively minimized w.r.t. $\mathbf{u}$ (only smooth terms), and then w.r.t. each $\mathbf{x}_i$ (without forgetting the dual update). Linear operators can even be included (this will only change $\boldsymbol{\Sigma}$ and the dimensions of the $\mathbf{x}_i$), provided they are easily invertible. For example, if for $i = 1, ..., p$, we have the constraint $\mathbf{H}_i\mathbf{x} = \mathbf{x}_i$, then an ADMM iteration writes:

$$\begin{aligned} \mathbf{x}^{k+1} &= \mathbf{prox}_{\tau f}^{\mathbf{H}_1,\dots,\mathbf{H}_p}(\mathbf{v}^k - \mathbf{z}^k) \\ \mathbf{v}_1^{k+1} &= \mathbf{prox}_{\tau f_1}(\mathbf{H}_1\mathbf{x}^{k+1} + \mathbf{z}_1^k) \\ &\dots \\ \mathbf{v}_p^{k+1} &= \mathbf{prox}_{\tau f_m}(\mathbf{H}_p\mathbf{x}^{k+1} + \mathbf{z}_p^k) \\ \mathbf{z}^{k+1} &= \mathbf{z}^k + \boldsymbol{\Sigma}\mathbf{u}^{k+1} - \mathbf{v}^{k+1}, \end{aligned} \tag{A.26}$$

where $\mathbf{prox}_{\tau f}^{\mathbf{H}_1,\dots,\mathbf{H}_p} : \mathbb{R}^M \to \mathbb{R}^n$ is defined as $\mathbf{prox}_{\tau f}^{\mathbf{H}_1,\dots,\mathbf{H}_p}(\mathbf{s}) = \underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{argmin}} \; f(\mathbf{x}) + \frac{1}{2\tau}||\mathbf{H}_1\mathbf{x} - \mathbf{s}_1||_2^2 + ... + \frac{1}{2\tau}||\mathbf{H}_p\mathbf{x} - \mathbf{s}_p||_2^2$, M being the dimension of $\mathbf{v}$.

Depending on the problem, it can be interesting to put in the function f all the "easy" differentiable parts of the objective function (typically the data fit term), so long as $\mathbf{prox}_{\tau f}^{\mathbf{H}_1,\dots,\mathbf{H}_p}$ can be computed with little effort.

A.4 A primal-dual algorithm

ADMM is theoretically very powerful: an arbitrary number of "proximable" terms can be included, and constraints and linear operators can be included in the problem. However complexity increases a lot with each term, and the constraint $\boldsymbol{\Sigma}\mathbf{u} = \mathbf{v}$ is only asymptotically enforced. Let us consider a "simpler" problem involving two potentially nonsmooth terms and a linear operator $\mathbf{H}$:

$$\underset{x\in\mathbf{R}^n}{\operatorname{argmin}} \quad f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{H}\mathbf{x}), \tag{A.27}$$

where f is smooth (with β-Lipschitz gradient), and g and h are proximable. For this problem there exists a nice alternative to the ADMM combined with variable splitting, which is introduced in [44].

The optimality condition for problem (A.27) writes ($\mathcal{H}^*$ is the adjoint operator of $\mathcal{H}$, whose matrix for the canonical bases is $\mathbf{H}^\top$) [126]:

$$\mathbf{0} \in (\nabla f + \partial g + \mathcal{H}^* \circ \partial h \circ \mathcal{H})(\mathbf{x}). \tag{A.28}$$

The Fenchel-Rockafellar duality theorem [84] states that $\mathbf{x}$ is a solution of (A.27), if and only if there exists a solution to a Fenchel dual problem denoted as $\mathbf{s} \in \mathbb{R}^m$:

$$\min_{\mathbf{x}\in\mathbf{R}^n} \quad f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{H}\mathbf{x}) = - \min_{\mathbf{s}\in\mathbf{R}^m} \quad (f+g)^*(-\mathbf{H}^\top \mathbf{s}) + h^*(\mathbf{s}). \tag{A.29}$$

In addition, such a primal-dual solution tuple $(\mathbf{x}, \mathbf{s})$ verifies $\mathbf{H}\mathbf{x} \in \partial h^*(\mathbf{s})$, and $\mathbf{s} \in \partial h(\mathbf{H}\mathbf{x})$ [126]. If $\mathbf{x}$ and $\mathbf{s}$ are solutions of the primal and dual problem, respectively, we have, for any η and $\gamma \in]0, +\infty[$:

$$\begin{aligned}
(A.28) &\Leftrightarrow \mathbf{0} \in \{\eta(\nabla f(\mathbf{x}) + \mathbf{H}^\top \mathbf{s})\} + \partial \eta g(\mathbf{x}) \text{ (substituting } \partial h(\mathbf{H}\mathbf{x}) \text{ by } \mathbf{s} \text{ in (A.28))} \\
&\Leftrightarrow -\eta(\nabla f(\mathbf{x}) + \mathbf{H}^\top \mathbf{s}) \in \partial \eta g(\mathbf{x}) \\
&\Leftrightarrow (\mathbf{x} - \eta(\nabla f(\mathbf{x}) + \mathbf{H}^\top \mathbf{s})) - \mathbf{x} \in \partial \eta g(\mathbf{x}) \\
&\Leftrightarrow \mathbf{x} = \mathbf{prox}_{\eta g}(\mathbf{x} - \eta(\nabla f(\mathbf{x}) + \mathbf{H}^\top \mathbf{s})) \text{ (using Eq. (A.8)),}
\end{aligned}$$

and similarly:

$$\begin{aligned}
\mathbf{H}\mathbf{x} \in \partial h^*(\mathbf{s}) &\Leftrightarrow \gamma \mathbf{H}\mathbf{x} \in \partial \gamma h^*(\mathbf{s}) \\
&\Leftrightarrow (\mathbf{s} + \gamma \mathbf{H}\mathbf{x}) - \mathbf{s} \in \partial \gamma h^*(\mathbf{s}) \\
&\Leftrightarrow \mathbf{s} = \mathbf{prox}_{\gamma h^*}(\mathbf{s} + \gamma \mathbf{H}\mathbf{x}) \text{ (using Eq. (A.8)).}
\end{aligned}$$

We have two fixed point equations, one for the primal variable and one for the dual variable:

$$\begin{aligned}
\mathbf{x} &= \mathbf{prox}_{\eta g}(\mathbf{x} - \eta(\nabla f(\mathbf{x}) + \mathbf{H}^\top \mathbf{s})) \\
\mathbf{s} &= \mathbf{prox}_{\gamma h^*}(\mathbf{s} + \gamma \mathbf{H}\mathbf{x}),
\end{aligned} \tag{A.30}$$

which allow us to iterate the following updates, with appropriate choices of η and γ to guarantee convergence [44]:

$$\begin{aligned}
\mathbf{x}^{k+1} &= \mathbf{prox}_{\eta g}(\mathbf{x}^k - \eta(\nabla f(\mathbf{x}^k) + \mathbf{H}^\top \mathbf{s}^k)) \\
\mathbf{s}^{k+1} &= \mathbf{prox}_{\gamma h^*}(\mathbf{s}^k + \gamma \mathbf{H}(2\mathbf{x}^{k+1} - \mathbf{x}^k)).
\end{aligned} \tag{A.31}$$

The condition on the step sizes γ and η to guarantee convergence is $\frac{1}{\eta} - \gamma |||\mathcal{H}|||^2 \geq \frac{\beta}{2}$, where $|||\mathcal{H}|||$ is the operator norm of $\mathcal{H}$. More details can also be found in [126].

This has several advantages over ADMM for this specific problem: there is no variable splitting, there is no costly update involving the inversion of a matrix of the form $\tau \mathbf{I}_n + \mathbf{H}^\top \mathbf{H}$, we only need to compute the operator and its adjoint, and finally, the constraints (encoded in either g of h if they are indicator functions) are exactly enforced at each iteration if the computation of the proximal operators is exact.

Appendix B

Linear Gradient Operators and Total Variation

This appendix provides some complements on gradient operators for gray level images and on Total Variation (TV) for edge preserving spatial regularization.

B.1 First order gradient operators

For images, we can define horizontal and gradient operators, based on finite difference discretizations of the partial derivatives of an image $\mathcal{I}$, seen as a continuous function:

$$\begin{aligned} \mathcal{I} : \mathbf{\Omega} &\to \mathbb{R} \\ (x, y) &\mapsto f(x, y), \end{aligned} \tag{B.1}$$

where $\mathbf{\Omega}$ is an open bounded set of $\mathbb{R}^2$. In our case, we use first order partial derivatives, which we describe as the most simple forward finite difference operators:

$$\frac{\partial \mathcal{I}}{\partial x} \approx \mathcal{I}(x, y) - \mathcal{I}(x + h, y) \text{ and } \frac{\partial \mathcal{I}}{\partial y} \approx \mathcal{I}(x, y) - \mathcal{I}(x, y + h), \tag{B.2}$$

for a small value of h. For a discretized image I, this simply becomes:

$$\forall x, \nabla_{I,h}(x) = i(x, y) - i(x + 1, y) \text{ and } \forall y, \nabla_{I,h}(y) = i(x, y) - i(x, y + 1). \tag{B.3}$$

Since we are dealing with images with a finite support, assumptions have to be made about boundary conditions. Here, we will assume periodic boundaries for reasons which will become clear later.

For a vectorized m by n gray level image, denoted as $\mathbf{u} = \text{vec}(\mathbf{I})$ (bear in mind that we defined the vectorization operator as stacking the columns of a matrix into a column vector), with $\mathbf{I} \in \mathbb{R}^{m \times n}$, we define the linear operator $\mathcal{H}_h : \mathbb{R}^N \to \mathbb{R}^N$, such that each entry of $\mathcal{H}_h(\mathbf{u})$ is the horizontal first order finite difference of Eq. (B.3) for the corresponding entry of the input vector $\mathbf{u}$. The vertical gradient operator $\mathcal{H}_v$ is defined in a similar way. Finally we define a global gradient operator:

$$\begin{aligned} \mathcal{H} : \mathbb{R}^N &\to \mathbb{R}^{N \times 2} \\ \mathbf{u} &\mapsto \left[\mathcal{H}_h(\mathbf{u}), \mathcal{H}_v(\mathbf{u})\right]. \end{aligned} \tag{B.4}$$

We can use the same notations to denote gradient operators acting in the same way on each band of multivariate images. These operators transform matrices into matrices; for example we can define $\mathcal{H}_h : \mathbb{R}^{P\times N} \to \mathbb{R}^{P\times N}$, which applies the horizontal gradient independently on each row of an abundance matrix.

For now, let us suppose the operators act on gray-level images. Since the horizontal gradient are linear operators, they can be defined by their matrices $\mathbf{H}_h$ and $\mathbf{H}_v \in \mathbb{R}^{N\times N}$ in the canonical basis of $\mathbb{R}^N$:

$$\mathbf{H}_h = \left.\begin{bmatrix} \mathbf{I}_m & -\mathbf{I}_m & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_m & -\mathbf{I}_m & & & \vdots \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{I}_m & -\mathbf{I}_m \\ -\mathbf{I}_m & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{I}_m \end{bmatrix} \right\} n \text{ blocks} \quad , \tag{B.5}$$

and, defining a matrix $\mathbf{J}_n \in \mathbb{R}^{n\times n}$ as

$$\mathbf{J}_n = \begin{bmatrix} 1 & -1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & -1 & & & \vdots \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 & -1 \\ -1 & 0 & \cdots & \cdots & 0 & 1 \end{bmatrix}, \tag{B.6}$$

we have:

$$\mathbf{H}_v = \left.\begin{bmatrix} \mathbf{J}_n & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_n & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{J}_n \end{bmatrix} \right\} m \text{ blocks} \quad . \tag{B.7}$$

These two matrices have numerous interesting properties. First we can notice that the matrices of the horizontal and vertical operators are "dual" to each other. As a matter of fact, $\mathbf{H}_h$ is a block circulant matrix made of identity blocks, possibly with a minus sign (hence it is a circulant matrix as well), and $\mathbf{H}_v$ is a block diagonal matrix made of circulant blocks containing only 1 or -1 (it is then also a circulant matrix as a whole). The circulant property is only possible with a periodic boundary assumption.

In practice, in an optimization algorithm, we will have not only to compute the gradients $\mathbf{H}_h\mathbf{u}$ and $\mathbf{H}_v\mathbf{u}$, but we will also have to deal with the transposed matrices $\mathbf{H}_h^\top$ and $\mathbf{H}_v^\top$ (corresponding to the adjoint operators), and it is possible that we have to invert the matrices $\mathbf{H}_h^\top\mathbf{H}_h$ and $\mathbf{H}_v^\top\mathbf{H}_v$ to solve linear systems. If the computations of the gradients and their transposed versions are easy to vectorize, inverting $N \times N$ matrices can conversely become intractable for large images.

Fortunately, since the matrices are circulant, the linear operators are actually circular convolutions. Indeed, computing the gradients amounts to compute the 2D circular convolutions $\mathbf{K}_h \star \mathbf{I}$ and $\mathbf{K}_v \star \mathbf{I}$, where $\mathbf{K}_h$ and $\mathbf{K}_v \in \mathbb{R}^{m \times n}$

$$\mathbf{K}_h = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ -1 & 0 & \cdots & 0 & 1 \end{bmatrix}, \ \mathbf{K}_v = \begin{bmatrix} 0 & 0 & \cdots & -1 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{B.8}$$

Then, these convolutions can be easily performed in the Fourier domain, where they become pointwise multiplications between the 2D Fourier transforms (denoted by $\mathcal{F}(\cdot)$) of each operand. For instance:

$$\mathcal{H}_h(\mathbf{u}) = \mathbf{H}_h\mathbf{u} \equiv \mathbf{K}_h \star \mathbf{I} = \mathcal{F}^{-1}(\mathcal{F}(\mathbf{K}_h) \odot \mathcal{F}(\mathbf{I})). \tag{B.9}$$

Here $\mathbf{H}_h\mathbf{u} \equiv \mathbf{K}_h \star \mathbf{I}$ means that $\mathbf{H}_h\mathbf{u} = \text{vec}(\mathbf{K}_h \star \mathbf{I})$ (equality up to the vectorization of the result). Similarly, computing the adjoint becomes very simple as well:

$$\mathcal{H}_h^*(\mathbf{u}) = \mathbf{H}_h^\top\mathbf{u} \equiv \mathcal{F}^{-1}(\mathcal{F}(\mathbf{K}_h)^* \odot \mathcal{F}(\mathbf{I})), \tag{B.10}$$

where * is the complex conjugate. Finally, inverting the operator $\mathcal{H}_h^* \circ \mathcal{H}_h$ can be done with the following operation which is much less expensive than a matrix inversion, especially using a Fast Fourier Transform (FFT):

$$(\mathcal{H}_h^* \circ \mathcal{H}_h)^{-1}(\mathbf{u}) = (\mathbf{H}_h^\top\mathbf{H}_h)^{-1}\mathbf{u} \equiv \mathcal{F}^{-1}(\mathcal{F}(\mathbf{I}) \oslash |\mathcal{F}(\mathbf{K}_h)|^2), \tag{B.11}$$

where $|\cdot|$ is the complex modulus.

B.1.1 Operator Norms

Let $\mathcal{K} : \mathbb{R}^p \to \mathbb{R}^q$ (where $\mathbb{R}^p$ are $\mathbb{R}^q$ are endowed with their usual Euclidean norms) be a linear operator. In finite dimension, $\mathcal{K}$ is (uniformly) continuous and

$$\exists M \in \mathbb{R}, \forall \mathbf{x} \in \mathbb{R}^p, ||\mathcal{K}(\mathbf{x})||_2 \leq M||\mathbf{x}||_2. \tag{B.12}$$

The smallest M verifying this relationship is the operator norm $|||\mathcal{K}|||$ of $\mathcal{K}$. Note that M is also the "best" Lipschitz constant of $\mathcal{K}$. Instead of using the linear map $\mathcal{K}$, we can use its matrix $\mathbf{K} \in \mathbb{R}^{q \times p}$. An equivalent definition of $|||\mathcal{K}|||$ is:

$$|||\mathcal{K}||| = |||\mathbf{K}||| = \sup_{||\mathbf{x}||_2=1} \frac{||\mathbf{K}\mathbf{x}||_2}{||\mathbf{x}||_2}. \tag{B.13}$$

When Euclidean norms are used in both the domain and co-domain of $\mathcal{K}$, as in this case, then it can be shown that

$$|||\mathbf{K}||| = \sqrt{\rho(\mathbf{K}^\top \mathbf{K})}, \tag{B.14}$$

where $\rho(\mathbf{K}^\top \mathbf{K})$ is the spectral radius of $\mathbf{K}^\top \mathbf{K}$, i.e. the maximum eigenvalue of $\mathbf{K}^\top \mathbf{K}$ (this matrix is symmetric, and can then be diagonalized in an orthonormal basis). The proof is based on the definition of the adjoint, using the inner product (hence the need of an Euclidean norm), and the decomposition of vectors in $\mathbb{R}^n$ in an orthonormal basis of eigenvectors of $\mathbf{K}^\top \mathbf{K}$ to bound the operator norm by the spectral radius from above. The norm is then bounded from below using the image by $\mathcal{K}$ of the eigenvector associated to the spectral radius, giving equality.

In practice, the computation of $|||\mathcal{K}|||$ can be cumbersome for large matrices, unless they are sparse and can be efficiently stored in the memory of a computer. Otherwise, techniques such as power methods or Rayleigh quotients, which do not require the storage of the matrices but only a way to compute $\mathbf{Kx}$ for any $\mathbf{x}$ can be used to compute the spectral radius in a reasonable amount of time.

Finally, in the case of the gradient operator $\mathcal{H}$, it turns out that $|||\mathcal{H}||| = |||\mathcal{H}_h||| = |||\mathcal{H}_v|||$ (although the symbol used is the same, the norm on the left hand side is not the same as the other two since the co-domains of the operators are different).

B.2 Total Variation

Most of the notions presented in this section can be found in [35]. At first, we will deal with continuous images, represented as functions. For instance, a gray level image u will be described as a locally integrable function: $u : \Omega \to \mathbb{R}$ where Ω is an open (bounded) set of $\mathbb{R}^2$. In this context, the usual denoising problem aims at recovering a denoised image u from noisy observations, possibly with an additional degradation coming from a linear operator $g = \mathcal{A}u + n$. Since trying to recover u by a simple least squares fit is a very ill-posed problem, suitable regularizations are necessary. Natural images often exhibit smooth transitions, hence (for Gaussian noise) it can make sense to look for a differentiable function u with smooth spatial variations and to try to minimize a Tikhonov-like functional:

$$\int_\Omega (\mathcal{A}u(\mathbf{x}) - g(\mathbf{x}))^2 \mathrm{d}\mathbf{x} + \lambda \int_\Omega ||\nabla_u(\mathbf{x})||_2^2 \mathrm{d}\mathbf{x}, \tag{B.15}$$

where the gradient value $\nabla_u(\mathbf{x})$ is a two dimensional vector. However, in addition to the fact that this is only defined for differentiable functions, the squared gradient in the objective penalizes large gradient values very strongly. For instance, step edges (informally functions with "infinite" derivatives on a zero-measure set) commonly found in natural images, are banned with this Tikhonov penalty.

Consequently, other more suited penalizations have to be found in order to enforce smoothness in homogeneous regions of the image, while simultaneously preserving sharp edges. This

is the rationale of the Total Variation [136], which can be defined for any locally integrable function u (but not necessarily differentiable) as:

$$TV(u) = \sup_{\phi \in \mathcal{C}_c^1(\Omega, \mathbb{R}^2)} \left\{ -\int_\Omega u(\mathbf{x}) \mathrm{div}(\phi)(\mathbf{x}) \mathrm{d}\mathbf{x}, ||\phi||_\infty \leq 1 \right\}, \tag{B.16}$$

where $\mathcal{C}_c^1(\Omega, \mathbb{R})$ is the set of continuously differentiable functions, div is the divergence operator, and $||\cdot||_\infty$ is the uniform norm. Note that the TV is defined even for nondifferentiable functions since smoothness is only required for ϕ. A function u is said to have bounded variation when $TV(u) \leq \infty$. This definition could be easily extended to N-D images [35] and multivariate ones [67], in which case it can be interesting to introduce a coupling of the channels, e.g. for color images.

When u is smooth, we can compute an explicit value for the TV. Indeed, in that case, since minus the divergence is the adjoint operator of the gradient, we have:

$$\langle u, -\mathrm{div}(\phi) \rangle = \langle \nabla_u, \phi \rangle, \tag{B.17}$$

where $\langle \cdot, \cdot \rangle$ is, is on the left hand side, the inner product in $\mathcal{C}_c^1(\Omega, \mathbb{R})$ defined by: $\langle u, v \rangle = \int_\Omega u(\mathbf{x}) v(\mathbf{x}) \mathrm{d}\mathbf{x}$, and on the right hand side, the inner product in $\mathcal{C}_c^1(\Omega, \mathbb{R}^2)$ defined by: $\langle \phi, \psi \rangle = \int_\Omega \phi(\mathbf{x}) \cdot \psi(\mathbf{x}) \mathrm{d}\mathbf{x}$, where $\cdot$ is the canonical dot product in $\mathbb{R}^2$. Using Eq. (B.17):

$$TV(u) = \sup_{\phi \in \mathcal{C}_c^1(\Omega, \mathbb{R}^2)} \left\{ \int_\Omega \nabla_u(\mathbf{x}) \cdot \phi(\mathbf{x}) \mathrm{d}\mathbf{x}, ||\phi||_\infty \leq 1 \right\}. \tag{B.18}$$

Besides:

$$\int_\Omega \nabla_u(\mathbf{x}) \cdot \phi(\mathbf{x}) \mathrm{d}\mathbf{x} \leq \left| \int_\Omega \nabla_u(\mathbf{x}) \cdot \phi(\mathbf{x}) \mathrm{d}\mathbf{x} \right| \leq \int_\Omega ||\nabla_u(\mathbf{x})||_2 ||\phi(\mathbf{x})||_2 \mathrm{d}\mathbf{x} \leq \int_\Omega ||\nabla_u(\mathbf{x})||_2 \mathrm{d}\mathbf{x}, \tag{B.19}$$

since $\forall \mathbf{x}, ||\phi(\mathbf{x})||_2 \leq 1$. We have shown that $TV(u)$ is bounded from above by the integral of the $\mathcal{L}_2$ norm of the gradient. This value is attained by the function ϕ defined by $\phi(\mathbf{x}) = \frac{\nabla_u(\mathbf{x})}{||\nabla_u(\mathbf{x})||_2}$, showing that

$$TV(u) = \int_\Omega ||\nabla_u(\mathbf{x})||_2 \mathrm{d}\mathbf{x}. \tag{B.20}$$

The TV functional has nice properties: it is convex and lower semi-continuous.

A TV regularization thus consists (for smooth functions) in penalizing the gradient using its $\mathcal{L}_2$ norm, as before, but without the square. This change is precisely what will allow the TV to preserve edges in images.

The TV is related to summing the local variations of functions over the whole support of the image. These variations can be discontinuous, contrary to a simple Tikhonov regularization. To illustrate this, let us take a 1D step edge function $u : [-1, 1] \to \mathbb{R}$:

$$u(x) = \begin{cases} 0 \text{ if } x < 0 \\ a \text{ if } x \geq 0. \end{cases} \tag{B.21}$$

This function is not differentiable, but it has a finite TV (adapting the definition to 1D) nonetheless. If we take a certain $\phi \in \mathcal{C}_c^1([-1,1],\mathbb{R})$, then:

$$\int_{-1}^{1} u(x)\phi'(x)\mathrm{d}x = a\int_{0}^{1} \phi'(x)\mathrm{d}x = a(\phi(1)-\phi(0)) \leq 2|a|, \tag{B.22}$$

since $||\phi||_\infty \leq 1$ and ϕ is integrable on $[-1,1]$.

This means that edges are penalized by the TV functional in a way proportional to the height of the edge, but not infinitely so and thus not too important edges will be kept in the solution of the TV regularized denoising problem.

Now, if we want to come back to usual discrete digital images, we have to discretize the TV. image $\mathbf{u} \in \mathbb{R}^N$ (in a vectorial form), the discrete TV can be defined similarly, except that the divergence is replaced by the adjoint of the finite difference operator $\mathcal{H}$:

$$TV(u) = \max_{\mathbf{S}\in\mathbb{R}^{N\times 2}} \left\{ \mathbf{u}^\top \mathcal{H}^*(\mathbf{S}), ||\mathbf{S}||_\infty \leq 1 \right\}. \tag{B.23}$$

Then, with $\mathbf{S} \in \mathbf{R}^{N\times 2}$ defined by its rows $\mathbf{s}_k = (\mathcal{H}(\mathbf{u}))_k \oslash ||(\mathcal{H}(\mathbf{u}))_k||_2$, the discrete equivalent of the TV becomes:

$$TV(\mathbf{u}) = \sum_{k=1}^{N} ||\mathcal{H}(\mathbf{u})_k||_2 = ||\mathcal{H}(\mathbf{u})||_{2,1} = \sum_{k=1}^{N} \sqrt{\mathcal{H}_h(\mathbf{u})_k^2 + \mathcal{H}_v(\mathbf{u})_k^2}. \tag{B.24}$$

An anisotropic TV also exists, which despite its weaker properties can be easier to handle in an optimization context:

$$TV_{\mathcal{L}_1}(\mathbf{u}) = \sum_{k=1}^{N} (||(\mathbf{H}_h\mathbf{u})_k||_1 + ||(\mathbf{H}_v\mathbf{u})_k||_1) = ||\mathcal{H}(\mathbf{u})||_{1,1}. \tag{B.25}$$

With this formulation, we can see that the TV, with the $\mathcal{L}_1$ norm here (and the nondifferentiable norm in the isotropic case) can have the side effect of making the gradient sparse, which might be undesirable since it can make the restored image piecewise constant, instead of piecewise smooth. This is the so called "staircasing" effect of the TV.

Appendix C

Complementary results on local Intrinsic Dimensionality estimation

This appendix contains complementary results for Chapter 3 of this manuscript. The contents are organized as follows: section C.1 contains additional results for the competing algorithms on three datasets comprising colored noise, or noise correlation. We show that while noise correlation seems to affect the global ID value, the trends between global and local scales remain the same as with uncorrelated noise. Section C.2 contains results on the synthetic dataset with white Gaussian noise, but in the case where the noise values are assumed to be known. We show that knowing the noise beforehand does not change the results much compared to those where a global noise estimation using the algorithm of [133] is performed, which shows that the local to global trends are much more a result of the algorithms themselves rather than of the noise estimation strategy.

C.1 Results on synthetic datasets with colored and correlated noise

C.1.1 Data generation

Here we explain how the synthetic datasets used were designed. Apart from the noise, the datasets are the same than in Chapter 3.

We first generated colored Gaussian noise values. Here, by colored we imply Gaussian noise with a diagonal covariance matrix whose diagonal elements are not the same. This is the noise model assumed by the noise estimation method we used. In order to make the simulation a bit more realistic in terms of the choice of the variances, we used as noise covariance matrix the one obtained by noise estimation using Roger's method [133] on the Cuprite dataset (which has the same number of bands as a sublibrary of the used USGS spectral library). In the end, the simulated datasets are then of dimensions $350 \times 350 \times 188$. Since the estimated average SNR of the Cuprite dataset (average of the estimated SNRs in each band) is 27 dB, the results can be qualitatively compared to the ones of Chapter 3. (30 dB and 120 bands). For information, the estimated band by band SNR is shown in Fig. C.1.

Then, we address the fully correlated noise case. In that case, the diagonal elements of the

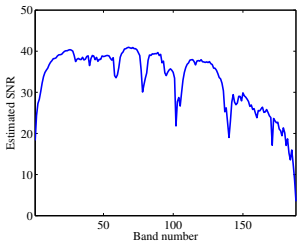


Figure C.1: Estimated SNR for each band of the Cuprite dataset.

noise covariance matrix were chosen in the same way as in the colored case. The off-diagonal elements were all set so as to obtain a correlation coefficient between any two distinct bands of 0.5, in order to get highly correlated noise values. For these three datasets, we have replaced the HFC algorithm with its Noise Whitened version, the NWHFC algorithm, since here a noise whitening should improve the results.

C.1.2 Results

C.1.2.1 Colored noise

A plot of the estimated ID value as a function of the window size in the case of colored noise, for both local and global noise estimation is shown in Fig. C.2. The quality metric μ_i is plotted against the window size i for both local and global noise estimation scenarios in Fig. C.3. These figures are to be compared with the corresponding figures in the case of white noise in section 3.5: Figs. 3.5, 3.6, and 3.7.

For local noise estimations, the same trends appear in the plots as for the white noise, with a window size range with overestimated ID values, especially for local noise estimation, and a slow stabilization after that. It is worthwhile to note that this is not very surprising since the noise estimation method used for the concerned algorithms assumes a diagonal covariance matrix for the noise with different variances in each band. For the algorithms requiring it, the main difference is that colored noise seems to complicate further the noise estimation with very few samples, resulting in a higher overestimation peak, and therefore a lower value of the quality metric for small windows. The other algorithms can have a different behavior: it seems that HFC is sensitive to noise coloration since it systematically provides higher ID values than in the white noise case. The HIDENN algorithm obtains here better results than in the white noise case, although the number of bands is more important in the colored dataset than in the white one. This could be explained by the fact that higher variances in some bands (which is something which happens in the Cuprite data, as seen in Fig. C.1) will push apart (in certain directions) pixels which would have been close, had they been noiseless. This results in a higher Euclidean distance between them than in the noiseless or even white noise case. This means that for the same ϵ, there will be fewer selected neighbors, and hence the correlation integral will be lower, inducing a lower ID. In the end, HIDDEN is a very local (in the feature space) algorithm, so it is likely to be more affected by structure in the noise.

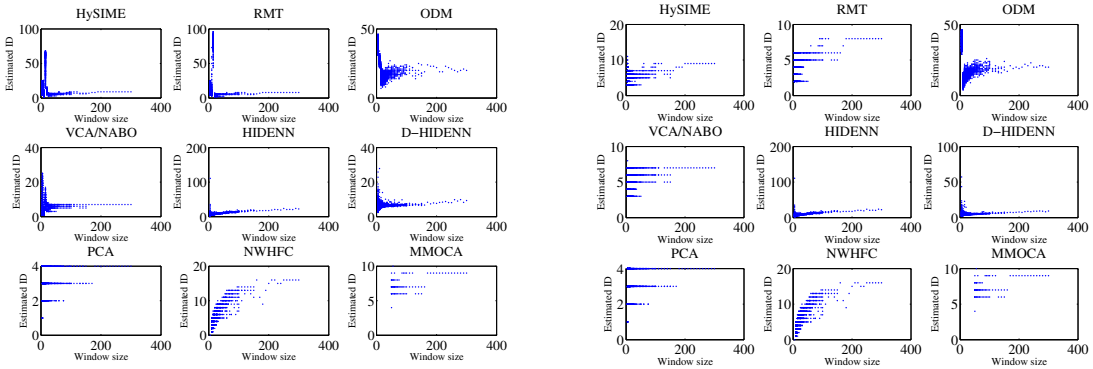


Figure C.2: Estimated ID for all algorithms in the case of local (a) and global (b) noise estimation plotted against window size for all algorithms, for colored noise.

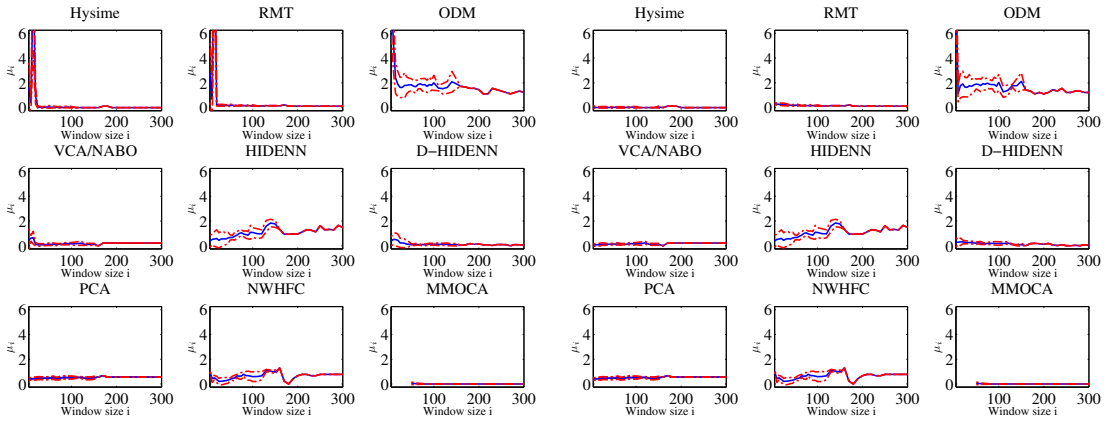


Figure C.3: Quality metric μ_i in the case of local (a) and global (b) noise estimation plotted against window size for all algorithms, for colored noise.

Global noise estimation results are comparable to the ones with white noise, with global ID values (estimated on the whole image) which tend to be higher than in the white noise case, a phenomenon already evidenced in [132].

C.1.2.2 Correlated noise

The results for correlated noise are shown in Figs C.4 and C.5. Here the conclusions are very similar to the previous section, since the diagonal elements of the covariance matrix are unchanged. Besides, the noise estimation algorithm we used, as well as most of the other noise estimation algorithms are not suited to fully correlated noise estimation. The trends of the estimated ID when going from local to global are the same, although a non white noise can cause a decrease in performance, as shown in [132]. In the end, noise correlation and

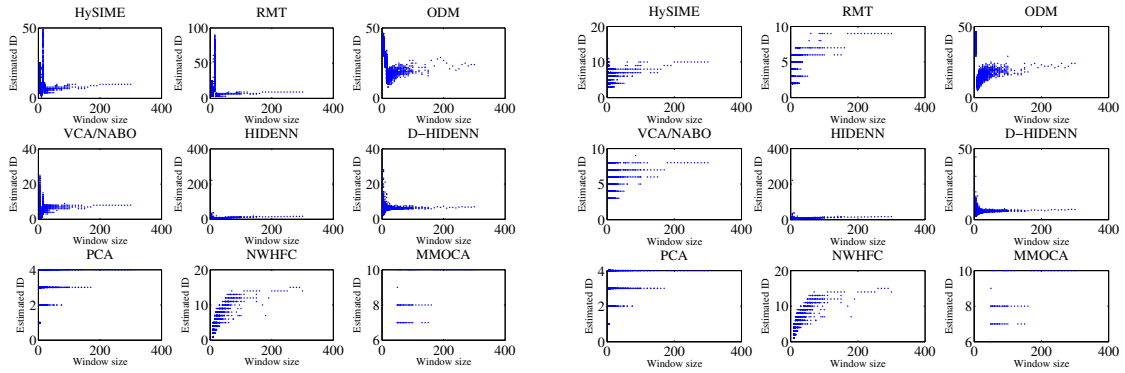


Figure C.4: Estimated ID in the case of local (a) and global (b) noise estimation plotted against window size for all algorithms, for highly correlated noise.

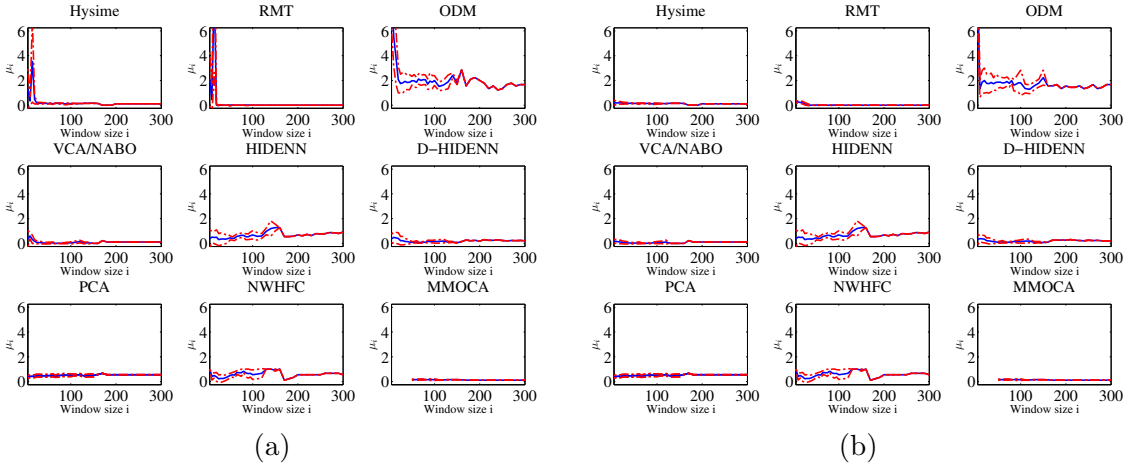


Figure C.5: Quality metric μ_i in the case of local (a) and global (b) noise estimation plotted against window size for all algorithms, for highly correlated noise.

coloration has more influence on the accuracy of the estimation than it has on the local to global behavior.

C.2 Results on synthetic datasets with known noise values

In this section, we show (in Figs. C.6 and C.7) the results of the different ID estimation algorithms for the synthetic dataset with spectrally and spatially white Gaussian noise and a SNR of 25dB, as well as 120 spectral bands. We assume that, for the algorithms requiring it, the noise values and statistics are known (or equivalently, the noise estimation is perfect). Here, we obtain results which are very similar to the white noise case, when the noise is estimated globally. Of course, the results of HFC, PCA, MMOCA and HIDENN are not

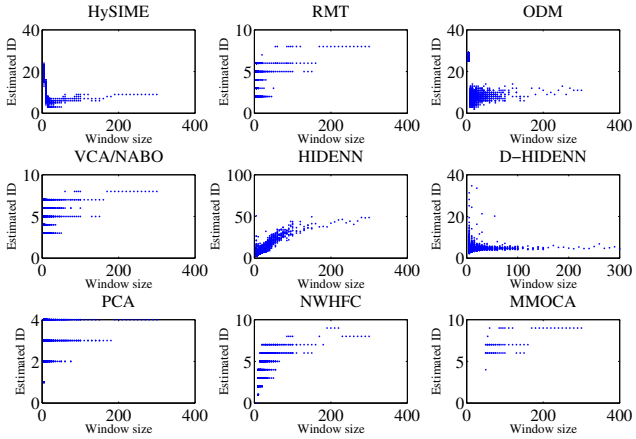


Figure C.6: Estimated ID in the case of known noise values plotted against window size for all algorithms, for 25 dB white noise.

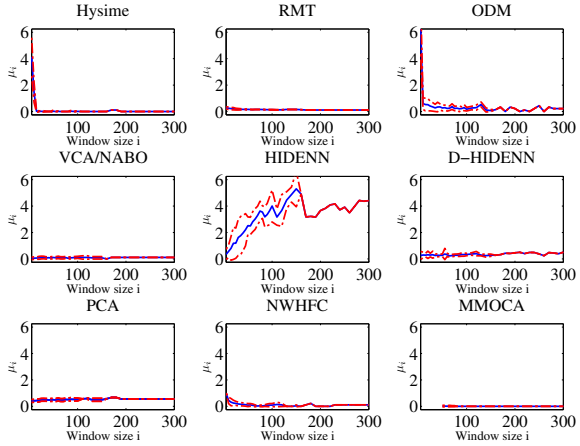


Figure C.7: Quality metric μ_i in the case of known noise values plotted against window size for all algorithms, for 25 dB white noise.

affected since they do not require a noise estimation step. These results simply show that in the global noise estimation case, the chosen noise estimation strategy seems to be reliable enough and is not affecting the local to global behavior of the different algorithms.

Appendix D

ADMM for the update of A in the ELMM-ALS algorithm

We describe below the optimization procedure to minimize the AL of Eq. (D.1) over each variable. The Augmented Lagrangian (in its scaled form [22]) writes:

$$\begin{aligned}\mathcal{L}(\mathbf{u},\boldsymbol{\mu},\mathbf{v},\mathbf{d}) &= f(\mathbf{u})+g(\mathbf{v})+\frac{\rho}{2}(||\boldsymbol{\Gamma}\mathbf{u}+\boldsymbol{\Lambda}\mathbf{v}-\mathbf{d}||_2^2-||\mathbf{d}||_2^2)\\ &= \frac{1}{2}||\mathbf{x}-\mathrm{vec}(\boldsymbol{\Sigma})||_2^2+\boldsymbol{\mu}^\top(\mathbf{K}\mathbf{u}-\mathbb{1}_N)+\lambda_A(||\mathbf{v}_2||_1+||\mathbf{v}_3||_1)+\mathcal{I}_{\mathbb{R}_+^{PN}}(\mathbf{v}_4)\\ &+\frac{\rho}{2}||\mathbf{u}-\mathbf{v}_1-\mathbf{d}_1||_2^2+\frac{\rho}{2}||\mathbf{H}_h\mathbf{v}_1-\mathbf{v}_2-\mathbf{d}_2||_2^2+\frac{\rho}{2}||\mathbf{H}_v\mathbf{v}_1-\mathbf{v}_3-\mathbf{d}_3||_2^2\\ &+\frac{\rho}{2}||\mathbf{u}-\mathbf{v}_4-\mathbf{d}_4||_2^2-\frac{\rho}{2}||\mathbf{d}||_2^2. \end{aligned} \tag{D.1}$$

D.1 Optimization w.r.t u and $\boldsymbol{\mu}$

This subproblem writes:

$$\arg\min_{\mathbf{u},\boldsymbol{\mu}} \sum_{k=1}^{N}\Big(\frac{1}{2}||\mathbf{x}_k-\mathbf{S}_k\mathbf{u}_k||_2^2+\mu_k(\mathbf{u}_k^\top\mathbb{1}_P-1)+\frac{\rho}{2}||\mathbf{u}_k-\mathbf{v}_{1k}-\mathbf{d}_{1k}||_2^2+\frac{\rho}{2}||\mathbf{u}_k-\mathbf{v}_{4k}-\mathbf{d}_{4k}||_2^2\Big), \tag{D.2}$$

and is separable over each pixel. By nulling the gradients of the k^{th} term of Eq. (D.2) w.r.t. to $\mathbf{u}_k$ and μ_k, we get the following system to solve:

$$\begin{bmatrix}\boldsymbol{\Omega}_k & \mathbb{1}_P\\ \mathbb{1}_P^\top & 0\end{bmatrix}\begin{bmatrix}\mathbf{u}_k\\ \mu_k\end{bmatrix}=\begin{bmatrix}\boldsymbol{\delta}_k\\ 1\end{bmatrix}, \tag{D.3}$$

where

$$\boldsymbol{\Omega}_k=\mathbf{S}_k^\top\mathbf{S}_k+2\rho\mathbf{I}_P, \tag{D.4}$$

and

$$\boldsymbol{\delta}_k=\mathbf{S}_k^\top\mathbf{x}_k+\rho(\mathbf{v}_{1k}+\mathbf{d}_{1k}+\mathbf{v}_{4k}+\mathbf{d}_{4k}). \tag{D.5}$$

Finally, by introducing the scalar quantity $s_k=\mathbb{1}_P^\top\boldsymbol{\Omega}_k^{-1}\mathbb{1}_P$ (which is simply the sum of all entries in Ω_k^{-1}), and using the block matrix inversion formula, we get the update rule for $\mathbf{u}_k$ and μ_k:

$$\begin{bmatrix}\mathbf{u}_k\\ \mu_k\end{bmatrix}\leftarrow\frac{1}{s_k}\begin{bmatrix}\boldsymbol{\Omega}_k^{-1}(s_k\mathbf{I}_P-\mathbb{1}_P\mathbb{1}_P^\top\boldsymbol{\Omega}_k^{-1}) & \boldsymbol{\Omega}_k^{-1}\mathbb{1}_P\\ \mathbb{1}_P^\top\boldsymbol{\Omega}_k^{-1} & -1\end{bmatrix}\begin{bmatrix}\boldsymbol{\delta}_k\\ 1\end{bmatrix}. \tag{D.6}$$

D.2 Optimization w.r.t. $\mathbf{v}_1$

The problem to solve for $\mathbf{v}_1$ is:

$$\arg\min_{\mathbf{v}_1} \quad \frac{\rho}{2}||\mathbf{u}-\mathbf{v}_1-\mathbf{d}_1||_2^2+\frac{\rho}{2}||\mathbf{H}_h\mathbf{v}_1-\mathbf{v}_2-\mathbf{d}_2||_2^2+\frac{\rho}{2}||\mathbf{H}_v\mathbf{v}_1-\mathbf{v}_3-\mathbf{d}_3||_2^2, \tag{D.7}$$

which is readily solved by

$$\mathbf{v}_1 \leftarrow (\mathbf{I}_{PN}+\mathbf{H}_h^\top\mathbf{H}_h+\mathbf{H}_v^\top\mathbf{H}_v)^{-1}\left(\mathbf{u}-\mathbf{d}_1+\mathbf{H}_h^\top(\mathbf{v}_2+\mathbf{d}_2)+\mathbf{H}_v^\top(\mathbf{v}_3+\mathbf{d}_3)\right). \tag{D.8}$$

However, in practice this requires an inversion of a $PN \times PN$ matrix, which is intractable in most cases. Fortunately, the matrix $\mathbf{I}_{PN}+\mathbf{H}_h^\top\mathbf{H}_h+\mathbf{H}_v^\top\mathbf{H}_v$ is circulant, and we can use the computation tricks presented in Appendix B. With this in mind, using the basic properties of the Fourier transform, we give the update rule for each band p of the image $\mathcal{V}_1 \in \mathbb{R}^{m\times n\times P}$:

$$\begin{aligned}\mathcal{V}_1^p \leftarrow \mathcal{F}^{-1}\Big(&\big(\mathcal{F}(\mathcal{U}^p-\mathcal{D}_1^p)+\mathcal{F}(\mathbf{h}_h)^*\odot\mathcal{F}(\mathcal{V}_2^p+\mathcal{D}_2^p)+\mathcal{F}(\mathbf{h}_v)^*\odot\mathcal{F}(\mathcal{V}_3^p+\mathcal{D}_3^p)\big)\\ &\oslash\big(\mathbb{1}_{m\times n}+|\mathcal{F}(\mathbf{h}_h)|^2+|\mathcal{F}(\mathbf{h}_v)|^2\big)\Big),\end{aligned} \tag{D.9}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ are the (discrete) 2D Fourier and inverse Fourier transforms, and $\mathbf{h}_h$ and $\mathbf{h}_v$ are as defined in Appendix B. Each script letter corresponds to the p^{th} band of the corresponding variable represented as an $m \times n$ image. Here * is the complex conjugate and $|\cdot|$ is the complex modulus.

D.3 Optimization w.r.t. $\mathbf{v}_2$

Back to a matrix formulation, the optimization w.r.t. $\mathbf{V}_2$ writes:

$$\arg\min_{\mathbf{V}_2} \lambda_A||\mathbf{V}_2||_{1,1}+\frac{\rho}{2}||\mathcal{H}_h(\mathbf{V}_1)-\mathbf{V}_2-\mathbf{D}_2||_F^2. \tag{D.10}$$

Solving problem (D.10) is equivalent to computing:

$$\mathbf{prox}_{(\lambda_A/\rho)||\cdot||_{1,1}}(\mathcal{H}_h(\mathbf{V}_1)-\mathbf{D}_2), \tag{D.11}$$

whose solution involves the *soft thresholding* operator, the proximal operator for the $\mathcal{L}_1$ norm:

$$\text{soft}_\lambda(s)=\left(1-\frac{\lambda}{|s|}\right)_+ s, \tag{D.12}$$

with $(\cdot)_+=\max(\cdot,0)$, and $\text{soft}_\lambda(0)=0\ \forall\lambda$. This leads to the update:

$$\mathbf{V}_2 \leftarrow \text{soft}_{\lambda_A/\rho}(\mathcal{H}_h(\mathbf{V}_1)-\mathbf{D}_2), \tag{D.13}$$

where the soft thresholding has to be understood entrywise. The horizontal gradient of $\mathbf{V}_1$ is computed in the frequency domain as in the previous section.

Note that in order to replace the anisotropic Total Variation (TV) penalization by a simple smoothing penalty, using the $\mathcal{L}_{2,1}$ norm instead of the $\mathcal{L}_{1,1}$ norm, one simply has to replace usual soft thresholding operator by the block soft thresholding, which is the proximal operator for the $\mathcal{L}_2$ norm. Using the isotropic TV operator is also a possibility, which requires some modifications on the algorithm.

D.4 Optimization w.r.t. $\mathbf{v}_3$

Similarly, for $\mathbf{v}_3$ we have to solve:

$$\arg\min_{\mathbf{V}_3} \lambda_A ||\mathbf{V}_3||_{1,1} + \frac{\rho}{2}||\mathcal{H}_v(\mathbf{V}_1) - \mathbf{V}_3 - \mathbf{D}_3||_F^2. \tag{D.14}$$

Using the same update rule as before:

$$\mathbf{V}_3 \leftarrow \text{soft}_{\lambda_A/\rho}(\mathcal{H}_v(\mathbf{V}_1) - \mathbf{D}_3), \tag{D.15}$$

D.5 Optimization w.r.t. $\mathbf{v}_4$

For $\mathbf{v}_4$ the optimization problem is:

$$\arg\min_{\mathbf{V}_4} = \mathcal{I}_{\mathbb{R}_+^{P\times N}}(\mathbf{V}_4) + \frac{\rho}{2}||\mathbf{U} - \mathbf{V}_4 - \mathbf{D}_4||_F^2, \tag{D.16}$$

which is simply solved by

$$\mathbf{V}_4 \leftarrow (\mathbf{U} - \mathbf{D}_4)_+. \tag{D.17}$$

D.6 Dual update

Finally, before going to the next iteration, the Lagrange multipliers have to be updated, giving the dual update:

$$\mathbf{d} \leftarrow \mathbf{d} - \boldsymbol{\Gamma}\mathbf{u} - \boldsymbol{\Lambda}\mathbf{v}. \tag{D.18}$$

Then the algorithm repeats the optimization steps until a termination criterion based on primal and dual residuals (following closely [22]) is fulfilled, The barrier parameter ρ is also updated iteratively to speed up the convergence.

Bibliography

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(11), 2012, pp. 2274–2282 (cit. on p. 38).

[2] N. Acito, M. Diani, and G. Corsini. "A New Algorithm for Robust Estimation of the Signal Subspace in Hyperspectral Images in the Presence of Rare Signal Components." *IEEE Transactions on Geoscience and Remote Sensing* 47(11), 2009, pp. 3844–3856 (cit. on p. 55).

[3] N. Acito, M. Diani, and G. Corsini. "Hyperspectral Signal Subspace Identification in the Presence of Rare Signal Components." *IEEE Transactions on Geoscience and Remote Sensing* 48(4), 2010, pp. 1940–1954 (cit. on pp. 55, 61).

[4] N. Acito, M. Diani, and G. Corsini. "Hyperspectral Signal Subspace Identification in the Presence of Rare Vectors and Signal-Dependent Noise." *IEEE Transactions on Geoscience and Remote Sensing* 51(1), 2013, pp. 283–299 (cit. on p. 55).

[5] P. Addesso, M. Dalla Mura, L. Condat, R. Restaino, G. Vivone, D. Picone, and J. Chanussot. "Hyperspectral pansharpening using covnex optimization and collaborative total variation regularization." *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. 2016, pp. 1–4 (cit. on p. 115).

[6] H. Akaike. "A new look at the statistical model identification." *IEEE Transactions on Automatic Control* 19(6), 1974, pp. 716–723 (cit. on p. 82).

[7] T. Akgun, Y. Altunbasak, and R. M. Mersereau. "Super-resolution reconstruction of hyperspectral images." *IEEE Transactions on Image Processing* 14(11), 2005, pp. 1860–1875 (cit. on p. 3).

[8] R. Ammanouil, A. Ferrari, C. Richard, and D. Mary. "Blind and Fully Constrained Unmixing of Hyperspectral Images." *IEEE Transactions on Image Processing* 23(12), 2014, pp. 5510–5518 (cit. on pp. 26, 80, 99).

[9] C. Andreou and V. Karathanassi. "Estimation of the Number of Endmembers Using Robust Outlier Detection Method." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7(1), 2014, pp. 247–256 (cit. on pp. 55, 60).

[10] P. Bajorski. "Does virtual dimensionality work in hyperspectral images?" *Proc. SPIE*. Vol. 7334. 2009, 73341J–73341J–11 (cit. on pp. 53, 54, 58).

[11] P. Bajorski. "Second Moment Linear Dimensionality as an Alternative to Virtual Dimensionality." *IEEE Transactions on Geoscience and Remote Sensing* 49(2), 2011, pp. 672–678 (cit. on pp. 54, 58).

[12] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Science & Business Media, 2011 (cit. on p. 169).

[13] J. D. Bayliss, J. A. Gualtieri, and R. F. Cromp. "Analyzing hyperspectral data with independent component analysis." *26th AIPR Workshop: Exploiting New Image Sources and Sensors*. International Society for Optics and Photonics. 1998, pp. 133–143 (cit. on p. 12).

[14] A. Beck and M. Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems." *SIAM journal on imaging sciences* 2(1), 2009, pp. 183–202 (cit. on pp. 169, 173).

[15] M. Berman, H. Kiiveri, R. Lagerstrom, A. Ernst, R. Dunne, and J. F. Huntington. "ICE: a statistical approach to identifying endmembers in hyperspectral images." *IEEE Transactions on Geoscience and Remote Sensing* 42(10), 2004, pp. 2085–2095 (cit. on pp. 19, 21, 126).

[16] C. Bilen, G. Puy, R. Gribonval, and L. Daudet. "Convex Optimization Approaches for Blind Sensor Calibration Using Sparsity." *IEEE Transactions on Signal Processing* 62(18), 2014, pp. 4847–4856 (cit. on p. 114).

[17] J. M. Bioucas-Dias. "A variable splitting augmented Lagrangian approach to linear spectral unmixing." *2009 First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*. 2009, pp. 1–4 (cit. on p. 18).

[18] J. M. Bioucas-Dias and J. M. P. Nascimento. "Estimation of signal subspace on hyperspectral data." *Proc. SPIE*. Vol. 5982. 2005, pp. 59820L–59820L–8 (cit. on p. 54).

[19] J. Bioucas-Dias and J. Nascimento. "Hyperspectral Subspace Identification." *IEEE Transactions on Geoscience and Remote Sensing* 46(8), 2008, pp. 2435–2445 (cit. on pp. 18, 54, 59, 64, 84, 98, 138).

[20] J. 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." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5(2), 2012, pp. 354–379 (cit. on pp. 3, 4, 10).

[21] J. W. Boardman, F. A. Kruse, and R. O. Green. "Mapping target signatures via partial unmixing of AVIRIS data." *Fifth JPL Airborne Earth Science Workshop*. Vol. 95. JPL Publication, 1995, pp. 23–26 (cit. on p. 14).

[22] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. "Distributed optimization and statistical learning via the alternating direction method of multipliers." *Foundations and Trends in Machine Learning* 3(1), 2011, pp. 1–122 (cit. on pp. 17, 79, 119, 169, 174, 189, 191).

[23] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004 (cit. on p. 170).

[24] P. Bunting and R. Lucas. "The delineation of tree crowns in Australian mixed species forests using hyperspectral Compact Airborne Spectrographic Imager (CASI) data." *Remote Sensing of Environment* 101(2), 2006, pp. 230 –248 (cit. on p. 3).

[25] F. Camastra. "Data dimensionality estimation methods: a survey." *Pattern Recognition* 36(12), 2003, pp. 2945 –2954 (cit. on pp. 54, 55).

[26] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. Atli Benediktsson. "Advances in Hyperspectral Image Classification: Earth Monitoring with Statistical Learning Methods." *IEEE Signal Processing Magazine* 31(1), 2014, pp. 45–54 (cit. on p. 3).

[27] E. J. Candes and M. B. Wakin. "An Introduction To Compressive Sampling." *IEEE Signal Processing Magazine* 25(2), 2008, pp. 21–30 (cit. on pp. 22, 23).

[28] E. J. Candes, J. K. Romberg, and T. Tao. "Stable signal recovery from incomplete and inaccurate measurements." *Communications on pure and applied mathematics* 59(8), 2006, pp. 1207–1223 (cit. on p. 22).

[29] K. Canham, A Schlamm, A Ziemann, B. Basener, and D. Messinger. "Spatially Adaptive Hyperspectral Unmixing." *IEEE Transactions on Geoscience and Remote Sensing* 49(11), 2011, pp. 4248–4262 (cit. on pp. 37, 54).

[30] W. Cao, J. Sun, and Z. Xu. "Fast image deconvolution using closed-form thresholding formulas of regularization." *Journal of Visual Communication and Image Representation* 24(1), 2013, pp. 31–41 (cit. on p. 93).

[31] K. M. Carter, R. Raich, and A. Hero. "On Local Intrinsic Dimension Estimation and Its Applications." *IEEE Transactions on Signal Processing* 58(2), 2010, pp. 650–663 (cit. on p. 55).

[32] K. Cawse-Nicholson, S. Damelin, A Robin, and M. Sears. "Determining the Intrinsic Dimension of a Hyperspectral Image Using Random Matrix Theory." *IEEE Transactions on Image Processing* 22(4), 2013, pp. 1301–1310 (cit. on pp. 53, 55, 59, 77, 84, 149).

[33] K. Cawse-Nicholson, A. Robin, and M. Sears. "The Effect of Correlation on Determining the Intrinsic Dimension of a Hyperspectral Image." *IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing* 6(2), 2013, pp. 482–487 (cit. on pp. 56, 57, 69).

[34] X. Ceamanos. "Evaluation des performances de l'analyse statistique et physique d'images hyperspectrales de Mars. Application au capteur multi-angulaire CRISM (In English)." PhD thesis. Université de Grenoble, 2011 (cit. on pp. 107, 132).

[35] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock. "An introduction to total variation for image analysis." *Theoretical foundations and numerical methods for sparse recovery* 9(263-340), 2010, p. 227 (cit. on pp. 180, 181).

[36] C.-I. Chang and Q. Du. "Estimation of number of spectrally distinct signal sources in hyperspectral imagery." *IEEE Transactions on Geoscience and Remote Sensing* 42(3), 2004, pp. 608–619 (cit. on pp. 53, 54).

[37] C.-I. Chang, W. Xiong, H.-M. Chen, and J.-W. Chai. "Maximum Orthogonal Subspace Projection Approach to Estimating the Number of Spectral Signal Sources in Hyperspectral Imagery." *IEEE Journal of Selected Topics in Signal Processing* 5(3), 2011, pp. 504–520 (cit. on p. 55).

[38] C.-I. Chang, W. Xiong, and C.-H. Wen. "A Theory of High-Order Statistics-Based Virtual Dimensionality for Hyperspectral Imagery." *IEEE Transactions on Geoscience and Remote Sensing* 52(1), 2014, pp. 188–208 (cit. on p. 55).

[39] C.-I. Chang and A. Plaza. "A fast iterative algorithm for implementation of pixel purity index." *IEEE Geoscience and Remote Sensing Letters* 3(1), 2006, pp. 63–67 (cit. on p. 14).

[40] J. E. Cohen. "About Notations in Multiway Array Processing." *arXiv preprint arXiv:1511.01306*, 2015 (cit. on p. 47).

[41] D. Comaniciu and P. Meer. "Mean shift: A robust approach toward feature space analysis." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(5), 2002, pp. 603–619 (cit. on pp. 38, 77).

[42] P. L. Combettes and J.-C. Pesquet. "Proximal splitting methods in signal processing." *Fixed-point algorithms for inverse problems in science and engineering*. Springer, 2011, pp. 185–212 (cit. on pp. 17, 122, 169).

[43] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010 (cit. on pp. 4, 11).

[44] L. Condat. "A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms." *Journal of Optimization Theory and Applications* 158(2), 2013, pp. 460–479 (cit. on pp. 123, 169, 175, 176).

[45] L. Condat. "Fast projection onto the simplex and the $\mathcal{L}_1$ ball." *Mathematical Programming*, 2014, pp. 1–11 (cit. on pp. 17, 80, 122).

[46] A. M. Cord, P. C. Pinet, Y. Daydou, and S. D. Chevrel. "Experimental determination of the surface photometric contribution in the spectral reflectance deconvolution processes for a simulated martian crater-like regolithic target." *Icarus* 175(1), 2005, pp. 78 –91 (cit. on p. 132).

[47] M. D. Craig. "Minimum-volume transforms for remotely sensed data." *IEEE Transactions on Geoscience and Remote Sensing* 32(3), 1994, pp. 542–552 (cit. on p. 15).

[48] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Pižurica, S. Gautama, W. Philips, S. Prasad, Q. Du, and F. Pacifici. "Hyperspectral and LiDAR Data Fusion: Outcome of the 2013 GRSS Data Fusion Contest." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7(6), 2014, pp. 2405–2418 (cit. on p. 98).

[49] N. Dobigeon, J.-Y. Tourneret, C. Richard, J. Bermudez, S. McLaughlin, and A. Hero. "Nonlinear Unmixing of Hyperspectral Images: Models and Algorithms." *IEEE Signal Processing Magazine* 31(1), 2014, pp. 82–94 (cit. on pp. 10, 11, 27).

[50] N. Dobigeon, S. Moussaoui, M. Coulon, J.-Y. Tourneret, and A. O. Hero. "Joint Bayesian endmember extraction and linear unmixing for hyperspectral imagery." *IEEE Transactions on Signal Processing* 57(11), 2009, pp. 4355–4368 (cit. on p. 20).

[51] D. L. Donoho and M. Elad. "Optimally sparse representation in general (nonorthogonal) dictionaries via $\mathcal{L}_1$ minimization." *Proceedings of the National Academy of Sciences* 100(5), 2003, pp. 2197–2202 (cit. on p. 23).

[52] L. Drumetz, J. Chanussot, and C. Jutten. "Endmember variability in spectral unmixing: recent advances." *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. 2016, pp. 1–4 (cit. on p. 29).

[53] L. Drumetz, S. Henrot, M. A. Veganzones, J. Chanussot, and C. Jutten. "Blind hyperspectral unmixing using an Extended Linear Mixing Model to address spectral variability." *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2015)*. 2015 (cit. on pp. 106, 109).

[54] L. Drumetz, M. A. Veganzones, R. M. Gómez, G. Tochon, M. D. Mura, G. A. Licciardi, C. Jutten, and J. Chanussot. "Hyperspectral Local Intrinsic Dimensionality." *IEEE Transactions on Geoscience and Remote Sensing* 54(7), 2016, pp. 4063–4078 (cit. on pp. 54, 83, 149).

[55] 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." *IEEE Transactions on Image Processing* 25(8), 2016, pp. 3890–3905 (cit. on pp. 106, 109, 115, 151).

[56] L. Drumetz, M. A. Veganzones, R. Marrero, G. Tochon, M. Dalla Mura, A. Plaza, and J. Chanussot. "Binary partition tree-based local spectral unmixing." *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. 2014 (cit. on pp. 54, 76).

[57] L. Drumetz, G. Tochon, J. Chanussot, and C. Jutten. "Estimating the number of endmembers to use in spectral unmixing of hyperspectral data with collaborative sparsity." *Submitted to the 13th International Conference on Latent Variable Analysis and Signal Separation (LVA-ICA)*. 2017, pp. 1–10 (cit. on p. 100).

[58] L. Drumetz, G. Tochon, M. A. Veganzones, J. Chanussot, and C. Jutten. "Improved local spectral unmixing of hyperspectral data using an algorithmic regularization path for collaborative sparse regression." *Submitted to the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 1–5 (cit. on p. 76).

[59] X. Du, A. Zare, P. Gader, and D. Dranishnikov. "Spatial and Spectral Unmixing Using the Beta Compositional Model." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7(6), 2014, pp. 1994–2003 (cit. on p. 43).

[60] O. Eches, N. Dobigeon, C. Mailhes, and J.-Y. Tourneret. "Bayesian Estimation of Linear Mixtures Using the Normal Compositional Model. Application to Hyperspectral Imagery." *IEEE Transactions on Image Processing* 19(6), 2010, pp. 1403–1413 (cit. on p. 42).

[61] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani. "Least angle regression." *The Annals of statistics* 32(2), 2004, pp. 407–499 (cit. on p. 81).

[62] M. Fauvel, Y. Tarabalka, J. Benediktsson, J. Chanussot, and J. Tilton. "Advances in Spectral-Spatial Classification of Hyperspectral Images." *Proceedings of the IEEE* 101(3), 2013, pp. 652–675 (cit. on p. 3).

[63] J. Fernando, F. Schmidt, C. Pilorget, P. Pinet, X. Ceamanos, S. Douté, Y. Daydou, and F. Costard. "Characterization and mapping of surface physical properties of Mars from CRISM multi-angular data: Application to Gusev Crater and Meridiani Planum." *Icarus* 253, 2015, pp. 271–295 (cit. on pp. 107, 109).

[64] K. Fukunaga. *Introduction to Statistical Pattern Recognition, Second Edition*. 2nd ed. Academic Press, Oct. 1990 (cit. on p. 54).

[65] L. Gao, Q. Du, B. Zhang, W. Yang, and Y. Wu. "A Comparative Study on Linear Regression-Based Noise Estimation for Hyperspectral Imagery." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6(2), 2013, pp. 488–498 (cit. on p. 56).

[66] M. Goenaga, M. Torres-Madronero, M. Velez-Reyes, S. Van Bloem, and J. Chinea. "Unmixing Analysis of a Time Series of Hyperion Images Over the Guanica Dry Forest in Puerto Rico." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6(2), 2013, pp. 329–338 (cit. on pp. 37, 46, 54).

[67] B. Goldluecke, E. Strekalovskiy, and D. Cremers. "The natural vectorial total variation which arises from geometric measure theory." *SIAM Journal on Imaging Sciences* 5(2), 2012, pp. 537–563 (cit. on pp. 115, 181).

[68] M. Graña and M. A. Veganzones. "An endmember-based distance for content based hyperspectral image retrieval." *Pattern Recognition* 45(9), 2012. Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011), pp. 3472 –3489 (cit. on pp. 39, 149).

[69] A. A. Green, M. Berman, P. Switzer, and M. D. Craig. "A transformation for ordering multispectral data in terms of image quality with implications for noise removal." *IEEE Transactions on Geoscience and Remote Sensing* 26(1), 1988, pp. 65–74 (cit. on pp. 14, 57, 60).

[70] L. Guigues. "Analyse multi-echelles pour la segmentation d'images." PhD thesis. Université de Cergy-Pontoise, 2003 (cit. on pp. 39, 41, 86).

[71] A. Halimi, N. Dobigeon, and J. Y. Tourneret. "Unsupervised Unmixing of Hyperspectral Images Accounting for Endmember Variability." *IEEE Transactions on Image Processing* 24(12), 2015, pp. 4904–4917 (cit. on p. 43).

[72] A. Halimi, P. Honeine, and J. Bioucas-Dias. "Hyperspectral Unmixing in Presence of Endmember Variability, Nonlinearity or Mismodelling Effects." *arXiv preprint arXiv:1511.05698*, 2015 (cit. on p. 143).

[73] B. Hapke. *Theory of reflectance and emittance spectroscopy*. Cambridge University Press, 2012 (cit. on pp. 46, 106, 107, 131).

[74] M. Hasanlou and F. Samadzadegan. "Comparative Study of Intrinsic Dimensionality Estimation and Dimension Reduction Techniques on Hyperspectral Images Using K-NN Classifier." *IEEE Geoscience and Remote Sensing Letters* 9(6), 2012, pp. 1046–1050 (cit. on p. 56).

[75] H. Hauksdottir, C. Jutten, F. Schmidt, J. Chanussot, J. A. Benediktsson, and S. Douté. "The physical meaning of independent components and artifact removal of hyperspectral data from Mars using ICA." *Proceedings of the 7th Nordic Signal Processing Symposium-NORSIG 2006*. IEEE. 2006, pp. 226–229 (cit. on p. 12).

[76] D. Heinz and C.-I. Chang. "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery." *IEEE Transactions on Geoscience and Remote Sensing* 39(3), 2001, pp. 529–545 (cit. on pp. 15–17, 119, 122).

[77] S. Hemissi, I. R. Farah, K. Saheb Ettabaa, and B. Solaiman. "Multi-spectro-temporal analysis of hyperspectral imagery based on 3-D spectral modeling and multilinear algebra." *IEEE Transactions on Geoscience and Remote Sensing* 51(1), 2013, pp. 199–216 (cit. on p. 46).

[78] S. Henrot, J. Chanussot, and C. Jutten. "Dynamical Spectral Unmixing of Multitemporal Hyperspectral Images." *IEEE Transactions on Image Processing* 25(7), 2016, pp. 3219–3232 (cit. on pp. 143, 154).

[79] R. Heylen, D. Burazerovic, and P. Scheunders. "Non-Linear Spectral Unmixing by Geodesic Simplex Volume Maximization." *IEEE Journal of Selected Topics in Signal Processing* 5(3), 2011, pp. 534–542 (cit. on p. 17).

[80] R. Heylen, M. Parente, and P. Gader. "A Review of Nonlinear Hyperspectral Unmixing Methods." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7(6), 2014, pp. 1844–1868 (cit. on pp. 4, 10, 27, 107, 109).

[81] R. Heylen and P. Scheunders. "A Multilinear Mixing Model for Nonlinear Spectral Unmixing." *IEEE Transactions on Geoscience and Remote Sensing* 54(1), 2016, pp. 240–251 (cit. on p. 27).

[82] R. Heylen and P. Scheunders. "Hyperspectral Intrinsic Dimensionality Estimation With Nearest-Neighbor Distance Ratios." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6(2), 2013, pp. 570–579 (cit. on pp. 55, 61).

[83] Y. Hu, E. Chi, and G. I. Allen. "ADMM Algorithmic Regularization Paths for Sparse Statistical Machine Learning." *arXiv preprint arXiv:1504.06637*, 2015 (cit. on p. 81).

[84] R. Ioan-Bot. *Conjugate duality in convex optimization*. Vol. 637. Springer Science & Business Media, 2009 (cit. on p. 176).

[85] M.-D. Iordache, J. M. Bioucas-Dias, and A. Plaza. "Collaborative Sparse Regression for Hyperspectral Unmixing." *IEEE Transactions on Geoscience and Remote Sensing* 52(1), 2014, pp. 341–354 (cit. on pp. 24, 80).

[86] M.-D. Iordache, J. M. Bioucas-Dias, and A. Plaza. "Hyperspectral unmixingwith sparse group lasso." *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International*. 2011, pp. 3586–3589 (cit. on p. 90).

[87] M.-D. Iordache, J. Bioucas-Dias, and A. Plaza. "Sparse Unmixing of Hyperspectral Data." *IEEE Transactions on Geoscience and Remote Sensing* 49(6), 2011, pp. 2014–2039 (cit. on pp. 22, 24, 123).

[88] M.-D. Iordache, J. Bioucas-Dias, and A. Plaza. "Total Variation Spatial Regularization for Sparse Hyperspectral Unmixing." *IEEE Transactions on Geoscience and Remote Sensing* 50(11), 2012, pp. 4484–4502 (cit. on pp. 25, 119, 124).

[89] M.-D. Iordache, A. Okujeni, S. Van Der Linden, J. Bioucas-Dias, A. Plaza, and B. Somers. "On the use of collaborative sparse regression in hyperspectral unmixing chain." *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. 2014 (cit. on p. 26).

[90] M.-D. Iordache, L. Tits, J. M. Bioucas-Dias, A. Plaza, and B. Somers. "A Dynamic Unmixing Framework for Plant Production System Monitoring." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7(6), 2014, pp. 2016–2034 (cit. on p. 35).

[91] S. Jacquemoud and F. Baret. "PROSPECT: A model of leaf optical properties spectra." *Remote sensing of environment* 34(2), 1990, pp. 75–91 (cit. on p. 45).

[92] J. Jin, B. Wang, and L. Zhang. "A Novel Approach Based on Fisher Discriminant Null Space for Decomposition of Mixed Pixels in Hyperspectral Imagery." *IEEE Geoscience and Remote Sensing Letters* 7(4), 2010, pp. 699–703 (cit. on p. 34).

[93] I. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986 (cit. on p. 57).

[94] N. Keshava and J. F. Mustard. "Spectral unmixing." *IEEE Signal Processing Magazine* 19(1), 2002, pp. 44–57 (cit. on pp. 3, 106).

[95] B. R. Kiran and J. Serra. "Global–local optimizations by hierarchical cuts and climbing energies." *Pattern Recognition* 47(1), 2014, pp. 12–24 (cit. on p. 39).

[96] M. Kowalski. "Sparse regression using mixed norms." *Applied and Computational Harmonic Analysis* 27(3), 2009, pp. 303–324 (cit. on p. 92).

[97] M. Kowalski, K. Siedenburg, and M. Dorfler. "Social sparsity! neighborhood systems enrich structured shrinkage operators." *IEEE Transactions on Signal Processing* 61(10), 2013, pp. 2498–2511 (cit. on pp. 25, 88).

[98] M. Kowalski and B. Torrésani. "Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients." *Signal, image and video processing* 3(3), 2009, pp. 251–264 (cit. on p. 90).

[99] O. Kuybeda, D. Malah, and M. Barzohar. "Rank Estimation and Redundancy Reduction of High-Dimensional Noisy Signals With Preservation of Rare Vectors." *IEEE Transactions on Signal Processing* 55(12), 2007, pp. 5579–5592 (cit. on pp. 55, 61).

[100] D. Landgrebe. "Hyperspectral image data analysis." *IEEE Signal Processing Magazine* 19(1), 2002, pp. 17–28 (cit. on p. 56).

[101] C. L. Lawson and R. J. Hanson. *Solving least squares problems*. Vol. 161. SIAM, 1974 (cit. on p. 16).

[102] É. Lebarbier and T. Mary-Huard. "Une introduction au critère BIC: fondements théoriques et interprétation." *Journal de la Société française de statistique* 147(1), 2006, pp. 39–57 (cit. on pp. 82, 83).

[103] J. Li, A. Agathos, D. Zaharie, J. M. Bioucas-Dias, A. Plaza, and X. Li. "Minimum Volume Simplex Analysis: A Fast Algorithm for Linear Hyperspectral Unmixing." *IEEE Transactions on Geoscience and Remote Sensing* 53(9), 2015, pp. 5067–5082 (cit. on p. 18).

[104] G. Licciardi, M. A. Veganzones, M. Simoes, J. Bioucas-Dias, and J. Chanussot. "Super-resolution of hyperspectral images using local spectral unmixing." *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. 2014 (cit. on p. 54).

[105] J. Liu, J. Zhang, Y. Gao, C. Zhang, and Z. Li. "Enhancing Spectral Unmixing by Local Neighborhood Weights." *IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing* 5(5), 2012, pp. 1545–1552 (cit. on p. 54).

[106] L. Loncan, L. B. de Almeida, J. M. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, S. Fabre, W. Liao, G. A. Licciardi, M. Simoes, J. Y. Tourneret, M. A. Veganzones, G. Vivone, Q. Wei, and N. Yokoya. "Hyperspectral Pansharpening: A Review." *IEEE Geoscience and Remote Sensing Magazine* 3(3), 2015, pp. 27–46 (cit. on p. 3).

[107] W.-K. Ma, J. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A. Plaza, A. Ambikapathi, and C.-Y. Chi. "A Signal Processing Perspective on Hyperspectral Unmixing: Insights from Remote Sensing." *IEEE Signal Processing Magazine* 31(1), 2014, pp. 67–81 (cit. on p. 11).

[108] Z. Ma. "Accuracy of the Tracy–Widom limits for the extreme eigenvalues in white Wishart matrices." *Bernoulli* 18(1), Feb. 2012, pp. 322–359 (cit. on p. 59).

[109] J. Mairal. *Recent Advances in Structured Sparse Models*. `http://lear.inrialpes.fr/people/mairal/resources/pdf/LEAR_seminar.pdf`. [Online; last accessed 10-May-2016]. 2010 (cit. on p. 24).

[110] A. Marinoni, A. Plaza, and P. Gamba. "Harmonic Mixture Modeling for Efficient Nonlinear Hyperspectral Unmixing." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9(9), 2016, pp. 4247–4256 (cit. on p. 27).

[111] R. Marrero, S. Lopez, G. Callico, M. A. Veganzones, A. Plaza, J. Chanussot, and R. Sarmiento. "A Novel Negative Abundance-Oriented Hyperspectral Unmixing Algorithm." *IEEE Transactions on Geoscience and Remote Sensing* 53(7), 2015, pp. 3772–3790 (cit. on pp. 55, 60).

[112] R. Marrero, S. Douté, A. Plaza, and J. Chanussot. "Validation of spectral unmixing methods using photometry and topography information." *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. 2013 (cit. on pp. 46, 109, 132).

[113] P. Meer, J. Jolion, and A. Rosenfeld. "A fast parallel algorithm for blind estimation of noise variance." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12(2), 1990, pp. 216–223 (cit. on pp. 57, 59).

[114] I. Meganem, Y. Deville, S. Hosseini, P. Déliot, and X. Briottet. "Linear-Quadratic Blind Source Separation Using NMF to Unmix Urban Hyperspectral Images." *IEEE Transactions on Signal Processing* 62(7), 2014, pp. 1822–1833 (cit. on p. 27).

[115] L. Meier, S. Van De Geer, and P. Bühlmann. "The group lasso for logistic regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(1), 2008, pp. 53–71 (cit. on p. 89).

[116] T. R. Meyer, L. Drumetz, J. Chanussot, A. L. Bertozzi, and C. Jutten. "Hyperspectral unmixing with material variability using social sparsity." *2016 IEEE International Conference on Image Processing (ICIP)*. 2016, pp. 2187–2191 (cit. on pp. 76, 88).

[117] F. A. Mianji and Y. Zhang. "SVM-Based Unmixing-to-Classification Conversion for Hyperspectral Abundance Quantification." *IEEE Transactions on Geoscience and Remote Sensing* 49(11), 2011, pp. 4318–4327 (cit. on p. 35).

[118] L. Miao and H. Qi. "Endmember Extraction From Highly Mixed Data Using Minimum Volume Constrained Nonnegative Matrix Factorization." *IEEE Transactions on Geoscience and Remote Sensing* 45(3), 2007, pp. 765–777 (cit. on p. 19).

[119] J.-J. Moreau. "Fonctions convexes duales et points proximaux dans un espace hilbertien." *CR Acad. Sci. Paris Sér. A Math* 255, 1962, pp. 2897–2899 (cit. on p. 170).

[120] J. M. P. Nascimento and J. M. Bioucas-Dias. "Hyperspectral unmixing algorithm via dependent component analysis." *Proc. 2007 IEEE International Geoscience and Remote Sensing Symposium*. 2007, pp. 4033–4036 (cit. on p. 20).

[121] J. Nascimento and J. Bioucas Dias. "Does independent component analysis play a role in unmixing hyperspectral data?" *IEEE Transactions on Geoscience and Remote Sensing* 43(1), 2005, pp. 175–187 (cit. on pp. 11, 12, 46, 95, 106, 110).

[122] J. Nascimento and J. Bioucas Dias. "Vertex component analysis: a fast algorithm to unmix hyperspectral data." *IEEE Transactions on Geoscience and Remote Sensing* 43(4), 2005, pp. 898–910 (cit. on pp. 14, 46, 60, 95, 106, 110, 141, 150).

[123] J. M. Nascimento and J. M. Bioucas-Dias. "Nonlinear mixture model for hyperspectral unmixing." *SPIE Europe Remote Sensing*. International Society for Optics and Photonics. 2009, pp. 74770I–74770I (cit. on p. 27).

[124] N. Nasrabadi. "Hyperspectral Target Detection : An Overview of Current and Future Challenges." *IEEE Signal Processing Magazine* 31(1), 2014, pp. 34–44 (cit. on p. 2).

[125] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2013 (cit. on p. 173).

[126] Z. Peng, T. Wu, Y. Xu, M. Yan, and W. Yin. "Coordinate Friendly Structures, Algorithms and Applications." *arXiv preprint arXiv:1601.00863*, 2016 (cit. on pp. 121, 122, 173, 176).

[127] A. Plaza, P. Martinez, R. Perez, and J. Plaza. "A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data." *IEEE Transactions on Geoscience and Remote Sensing* 42(3), 2004, pp. 650–663 (cit. on p. 14).

[128] J. Plaza, E. Hendrix, I. García, G. Martín, and A. Plaza. "On Endmember Identification in Hyperspectral Images Without Pure Pixels: A Comparison of Algorithms." English. *Journal of Mathematical Imaging and Vision* 42(2-3), 2012, pp. 163–175 (cit. on pp. 14, 17).

[129] M. B. Priestley. "Spectral analysis and time series," 1981 (cit. on p. 83).

[130] H. Pu, W. Xia, B. Wang, and G. M. Jiang. "A Fully Constrained Linear Spectral Unmixing Algorithm Based on Distance Geometry." *IEEE Transactions on Geoscience and Remote Sensing* 52(2), 2014, pp. 1157–1176 (cit. on p. 17).

[131] D. Roberts, M. Gardner, R. Church, S. Ustin, G. Scheer, and R. Green. "Mapping Chaparral in the Santa Monica Mountains Using Multiple Endmember Spectral Mixture Models." *Remote Sensing of Environment* 65(3), 1998, pp. 267 –279 (cit. on p. 35).

[132] A. Robin, K. Cawse-Nicholson, A. Mahmood, and M. Sears. "Estimation of the Intrinsic Dimension of Hyperspectral Images: Comparison of Current Methods." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8(6), 2015, pp. 2854–2861 (cit. on pp. 13, 56, 63, 185).

[133] R. Roger. "Principal Components transform with simple, automatic noise adjustment." *International Journal of Remote Sensing* 17(14), 1996, pp. 2719–2727 (cit. on pp. 56, 57, 61, 63, 64, 183).

[134] R. Roger and J. Arnold. "Reliably estimating the noise in AVIRIS hyperspectral images." *International Journal of Remote Sensing* 17(10), 1996, pp. 1951–1962 (cit. on p. 57).

[135] D. Rogge, B Rivard, J Zhang, A Sanchez, J Harris, and J Feng. "Integration of spatial–spectral information for the improved extraction of endmembers." *Remote Sensing of Environment* 110(3), 2007, pp. 287–303 (cit. on p. 36).

[136] L. I. Rudin, S. Osher, and E. Fatemi. "Nonlinear total variation based noise removal algorithms." *Physica D: Nonlinear Phenomena* 60(1), 1992, pp. 259–268 (cit. on pp. 115, 181).

[137] M. Sadeghi, S. B. Jones, and W. D. Philpot. "A linear physically-based model for remote sensing of soil moisture using short wave infrared bands." *Remote Sensing of Environment* 164, 2015, pp. 66 –76 (cit. on p. 46).

[138] P. Salembier and L. Garrido. "Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval." *IEEE Transactions on Image Processing* 9(4), 2000, pp. 561–576 (cit. on p. 37).

[139] D. Scheffler and P. Karrasch. "Destriping of hyperspectral image data: an evaluation of different algorithms using EO-1 Hyperion data." *Journal of Applied Remote Sensing* 8(1), 2014, p. 083645 (cit. on p. 12).

[140] A. Schlamm, D. Messinger, and W. Basener. "Geometric estimation of the inherent dimensionality of single and multi-material clusters in hyperspectral imagery." *Journal of Applied Remote Sensing* 3(1), 2009, pp. 033527–033527–16 (cit. on p. 54).

[141] G. Schwarz. "Estimating the Dimension of a Model." *Ann. Statist.* 6(2), Mar. 1978, pp. 461–464 (cit. on p. 82).

[142] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. "A sparse-group lasso." *Journal of Computational and Graphical Statistics* 22(2), 2013, pp. 231–245 (cit. on p. 90).

[143] B. Somers, G. Asner, L. Tits, and P. Coppin. "Endmember variability in Spectral Mixture Analysis: A review." *Remote Sensing of Environment* 115(7), 2011, pp. 1603 –1616 (cit. on pp. 4, 27, 32, 37, 54, 76).

[144] B. Somers, S. Delalieux, W. Verstraeten, and P. Coppin. "A conceptual framework for the simultaneous extraction of sub-pixel spatial extent and spectral characteristics of crops." *Photogrammetric Engineering & Remote Sensing* 75(1), 2009, pp. 57–68 (cit. on p. 45).

[145] B. Somers, M. Zortea, A. Plaza, and G. Asner. "Automated Extraction of Image-Based Endmember Bundles for Improved Spectral Unmixing." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5(2), 2012, pp. 396–408 (cit. on pp. 33, 54, 123).

[146] Y. Tarabalka, J. Chanussot, and J. Benediktsson. "Segmentation and classification of hyperspectral images using watershed transformation." *Pattern Recognition* 43(7), 2010, pp. 2367 –2379 (cit. on p. 38).

[147] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot. "Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques." *IEEE Transactions on Geoscience and Remote Sensing* 47(8), 2009, pp. 2973–2987 (cit. on p. 150).

[148] P.-A. Thouvenin, N. Dobigeon, and J.-Y. Tourneret. "Hyperspectral unmixing with spectral variability using a perturbed linear mixing model." *IEEE Transactions on Signal Processing* 64(2), 2016, pp. 525–538 (cit. on pp. 123, 126).

[149] G. Tochon, L. Drumetz, M. A. Veganzones, M. Dalla Mura, and J. Chanussot. "From Local to Global unmixing of hyperspectral images to reveal spectral variability." *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2016)*. 2016 (cit. on p. 146).

[150] G. Tochon. "Hierarchical analysis of multimodal images." PhD thesis. Université de Grenoble, 2015 (cit. on pp. 38–40, 150).

[151] G. Tochon, M. Dalla Mura, and J. Chanussot. "Segmentation of Multimodal Images based on Hierarchies of Partitions." *International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing*. Springer. 2015, pp. 241–252 (cit. on p. 100).

[152] J. A. Tropp and A. C. Gilbert. "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit." *IEEE Transactions on Information Theory* 53(12), 2007, pp. 4655–4666 (cit. on p. 23).

[153] D. Tuia, R. Flamary, and M. Barlaud. "Non-convex regularization in remote sensing." *arXiv preprint arXiv:1606.07289*, 2016 (cit. on p. 93).

[154] T. Uezato, R. J. Murphy, A. Melkumyan, and A. Chlingaryan. "A Novel Spectral Unmixing Method Incorporating Spectral Variability Within Endmember Classes." *IEEE Transactions on Geoscience and Remote Sensing* PP(99), 2016, pp. 1–1 (cit. on p. 35).

[155] S. Valero, P. Salembier, and J. Chanussot. "Hyperspectral Image Representation and Processing With Binary Partition Trees." *IEEE Transactions on Image Processing* 22(4), 2013, pp. 1430–1443 (cit. on pp. 37, 38, 149).

[156] M. A. Veganzones, J. E. Cohen, R. Cabral-Farias, K. Usevich, L. Drumetz, J. Chanussot, and P. Comon. "Nonnegative Canonical Decomposition of Hyperspectral Patch Tensors." *European Signal Processing Conference*. 2016 (cit. on pp. 153, 154, 156, 158).

[157] M. A. Veganzones, J. E. Cohen, R. C. Farias, R. Marrero, J. Chanussot, and P. Comon. "Multilinear spectral unmixing of hyperspectral multiangle images." *Signal Processing Conference (EUSIPCO), 2015 23rd European*. 2015, pp. 744–748 (cit. on pp. 48, 155).

[158] M. A. Veganzones, J. E. Cohen, R. C. Farias, J. Chanussot, and P. Comon. "Nonnegative Tensor CP Decomposition of Hyperspectral Data." *IEEE Transactions on Geoscience and Remote Sensing* PP(99), 2015, pp. 1–12 (cit. on pp. 47, 48, 155).

[159] M. A. Veganzones, L. Drumetz, R. Marrero, G. Tochon, M. Dalla Mura, A. Plaza, J. Bioucas-Dias, and J. Chanussot. "A new extended linear mixing model to address spectral variability." *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. 2014 (cit. on pp. 106, 123).

[160] M. A. Veganzones, M. Simoes, G. Licciardi, J. Bioucas-Dias, and J. Chanussot. "Hyperspectral super-resolution of locally low rank images from complementary multisource data." *Proc. IEEE International Conference on Image Processing (ICIP)*. 2014 (cit. on p. 54).

[161] M. A. Veganzones, G. Tochon, M. Dalla Mura, A. Plaza, and J. Chanussot. "Hyperspectral Image Segmentation Using a New Spectral Unmixing-Based Binary Partition Tree Representation." *IEEE Transactions on Image Processing* 23(8), 2014, pp. 3574–3589 (cit. on pp. 37–39, 54, 77, 109, 146, 149, 150).

[162] U. Von Luxburg. "A tutorial on spectral clustering." *Statistics and computing* 17(4), 2007, pp. 395–416 (cit. on p. 95).

[163] Y. X. Wang and Y. J. Zhang. "Nonnegative Matrix Factorization: A Comprehensive Review." *IEEE Transactions on Knowledge and Data Engineering* 25(6), 2013, pp. 1336–1353 (cit. on p. 21).

[164] Y. Wang, W. Yin, and J. Zeng. "Global convergence of ADMM in nonconvex nonsmooth optimization." *arXiv preprint arXiv:1511.06324*, 2015 (cit. on pp. 91, 93, 94).

[165] M. E. Winter. "N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data." *Proc. SPIE*. Vol. 3753. 1999, pp. 266–275 (cit. on p. 14).

[166] J. Woodworth and R. Chartrand. "Compressed sensing recovery via nonconvex shrinkage penalties." *arXiv preprint arXiv:1504.02923*, 2015 (cit. on pp. 93, 94).

[167] M. Xu, L. Zhang, and B. Du. "An Image-Based Endmember Bundle Extraction Algorithm Using Both Spatial and Spectral Information." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8(6), 2015, pp. 2607–2617 (cit. on p. 36).

[168] Y. Xu and W. Yin. "A globally convergent algorithm for nonconvex optimization based on block coordinate update." *arXiv preprint arXiv:1410.1386*, 2014 (cit. on pp. 121, 122, 173).

[169] M. Yukawa and S. I. Amari. "$\mathcal{L}_p$-Regularized Least Squares ($0 < p < 1$) and Critical Path." *IEEE Transactions on Information Theory* 62(1), 2016, pp. 488–502 (cit. on p. 93).

[170] A. Zare and P. Gader. "Sparsity Promoting Iterated Constrained Endmember Detection in Hyperspectral Imagery." *IEEE Geoscience and Remote Sensing Letters* 4(3), 2007, pp. 446–450 (cit. on p. 21).

[171] A. Zare, P. Gader, and G. Casella. "Sampling Piecewise Convex Unmixing and Endmember Extraction." *IEEE Transactions on Geoscience and Remote Sensing* 51(3), 2013, pp. 1655–1665 (cit. on p. 37).

[172] A. Zare and K. Ho. "Endmember Variability in Hyperspectral Analysis: Addressing Spectral Variability During Spectral Unmixing." *IEEE Signal Processing Magazine* 31(1), 2014, pp. 95–104 (cit. on pp. 4, 27, 32, 54).

[173] L. Zhang, L. Zhang, D. Tao, and X. Huang. "Tensor discriminative locality alignment for hyperspectral image spectral–spatial feature extraction." *IEEE Transactions on Geoscience and Remote Sensing* 51(1), 2013, pp. 242–256 (cit. on p. 155).

[174] Y. Zhong, R. Feng, and L. Zhang. "Non-local sparse unmixing for hyperspectral remote sensing imagery." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7(6), 2014, pp. 1889–1909 (cit. on p. 25).

[175] M. Zortea and A. Plaza. "Spatial preprocessing for endmember extraction." *IEEE Transactions on Geoscience and Remote Sensing* 47(8), 2009, pp. 2679–2693 (cit. on p. 149).

[176] A. Zymnis, S. J. Kim, J. Skaf, M. Parente, and S. Boyd. "Hyperspectral Image Unmixing via Alternating Projected Subgradients." *2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*. 2007, pp. 1164–1168 (cit. on p. 21).

Abstract — The fine spectral resolution of hyperspectral remote sensing images allows an accurate analysis of the imaged scene, but due to their limited spatial resolution, a pixel acquired by the sensor is often a mixture of the contributions of several materials. Spectral unmixing aims at estimating the spectra of the pure materials (called endmembers) in the scene, and their abundances in each pixel. The endmembers are usually assumed to be perfectly represented by a single spectrum, which is wrong in practice since each material exhibits a significant intra-class variability. This thesis aims at designing unmixing algorithms to better handle this phenomenon. First, we perform the unmixing locally in well chosen regions of the image where variability effects are less important, and automatically discard wrongly estimated local endmembers using collaborative sparsity. In another approach, we refine the abundance estimation of the materials by taking into account the group structure of an image-derived endmember dictionary. Second, we introduce an extended linear mixing model, based on physical considerations, modeling spectral variability in the form of scaling factors, and develop optimization algorithms to estimate its parameters. This model provides easily interpretable results and outperforms other state-of-the-art approaches. We finally investigate two applications of this model to confirm its relevance.

Keywords: Hyperspectral remote sensing, spectral unmixing, spectral variability, sparsity, convex optimization, hierarchical representation.

Résumé — La finesse de la résolution spectrale des images hyperspectrales en télédétection permet une analyse précise de la scène observée, mais leur résolution spatiale est limitée, et un pixel acquis par le capteur est souvent un mélange des contributions de différents matériaux. Le démélange spectral permet d'estimer les spectres des matériaux purs (endmembers) de la scène, et leurs abondances dans chaque pixel. Les endmembers sont souvent supposés être parfaitement représentés par un seul spectre, une hypothèse fausse en pratique, chaque matériau ayant une variabilité intra-classe non négligeable. Le but de cette thèse est de développer des algorithmes prenant mieux en compte ce phénomène. Nous effectuons le démélange localement, dans des régions bien choisies de l'image où les effets de la variabilité sont moindres, en éliminant automatiquement les endmembers non pertinents grâce à de la parcimonie collaborative. Dans une autre approche, nous raffinons l'estimation des abondances en utilisant la structure de groupe d'un dictionnaire d'endmembers extrait depuis les données. Ensuite, nous proposons un modèle de mélange linéaire étendu, basé sur des considérations physiques, qui modélise la variabilité spectrale par des facteurs d'échelle, et développons des algorithmes d'optimisation pour en estimer les paramètres. Ce modèle donne des résultats facilement interprétables et de meilleures performances que d'autres approches de la littérature. Nous étudions enfin deux applications de ce modèle pour confirmer sa pertinence.

Mots clés : Télédétection hyperspectrale, démélange spectral, variabilité spectrale, parcimonie, optimisation convexe, représentation hiérarchique.