UNIVERSITÀ DEGLI STUDI
DI MODENA E REGGIO EMILIA

Dottorato di ricerca in
"Models and methods for material and environmental sciences"

Ciclo XXXVII

Novel approaches for the exploration of hyperspectral images

in co-tutela con la Université de Lille

Candidata: OLARINI Alessandra

Relatore italiano (Tutor): Prof.ssa Marina Cocchi

Relatore francese (Tutor): Prof. Cyril Ruckebusch

Correlatore francese (Cotutor): Prof. Ludovic Duponchel

Coordinatore del Corso di Dottorato: Prof. Stefano Lugli

Université de Lille

Ecole doctorale Science de la Matière, du Rayonnement et de l'Environnement

Laboratoire de Spectroscopie pour les Interactions, la Réactivité et l'Environnement

---

Novel approaches for the exploration of hyperspectral images

Nouvelles approches pour l'exploration des images hyperspectrales

---

Thèse préparée et soutenue publiquement par Alessandra Olarini le 06/05/2025,

pour obtenir le grade de Docteur en Chimie théorique, physique, analytique.

Thèse dirigée par:

Prof. Cyril Ruckebusch - Université de Lille

Prof. Ludovic Duponchel - Université de Lille

Prof. Marina Cocchi - Università di Modena e Reggio Emilia

Composition du jury:

Dr. Jean-Michel Roger - *Rapporteur*, INRAE

Prof. Anna De Juan - *Rapporteuse,* Universitat de Barcelona

Prof. Federico Marini - *Examinateur,* Sapienza Università di Roma

Prof. Paolo Oliveri - *Examinateur,* Università degli studi di Genova

Prof. Ludovic Duponchel - *Directeur de thèse,* Université de Lille

Prof. Marina Cocchi - *Directrice de thèse,* Università di Modena e Reggio Emilia

Acknowledgments

This academic journey was very rich in so many different experiences and it allowed me to grow scientifically but also greatly on a human level. I owe this growth to my supervisors, Cyril, Marina, and Ludovic, each of whom supported me in their own way through doubts and challenges. I feel incredibly lucky to have been guided by such brilliant scientists and exceptional people. I cannot fully express my gratitude for all the work you did with me.

Cyril, your support went far beyond science. You knew how to lift my spirit when I needed it the most, and you believed in me even when I doubted myself. Your empathy and encouragement were a true anchor throughout this journey.

Marina, your kindness, clarity, and availability created a space where I always felt welcomed and understood. I learned so much from your scientific perspective and your warm, human approach to mentoring.

Ludovic, your scientific insight and rigor pushed me to elevate my thinking and work. I truly appreciated your expertise and the precision you brought to it.

Working with the teams in Lille and Modena was a real privilege. Thank you to the LASIRE group, especially Raffaele, Adri, Laureen, Alessandro, and Xingjie and to the Modena Chemometrics group, Caterina, Lorenzo, and Samuele, for your valuable input and for all the laughs and chats beyond the science.

A special thanks goes to my office buddies, Dani, Momo, and Zaza, who made sure I did not quit after day two. I am lucky to call you not just colleagues, but friends. You helped me take my first real steps in chemometrics and imaging, and your support meant the world to me.

I am also deeply grateful to Professors Malagoli, Oliveri, Gowen, and Criquet for being part of my comité de suivi de thèse and for their helpful and constructive feedback throughout. Thanks also to Vicky Caponigro and Riccardo Melis for sharing their ideas and enthusiasm with me.

To Marica, Chiara, Martina thanks for making my time in Lille so special and for becoming such dear friends. As soon as I arrived, Ale told me: “In the North, you cry

twice, when you arrive, and when you leave." Well... he was right.

And then Modena welcomed me like home. Bella Modena, eh... ma Alessio, Lindi, Bibi, Pups, Manu, Simo, Vale, Mirco, Vale, Lori, Lore, Dani siete voi che l'avete resa ancora più bella.

Thank you to Eugenio and the research group in Turin for giving me a new space to grow in a stimulating and supportive environment, and to Alberto, Carolina, and Giovanni for helping take some of the stress off during the final part of the PhD.

Thank you to my lifelong friends, Laura and Nora, Cate and Nina, Lau, Ila, Gaia, and Tonio and to Pezz, Bea and Gabri. Even with distance and different paths, you have always been there. Your presence, quiet or loud, close or far, made all the difference, even when life gave you a thousand other things to think about. To Chiara, a heartfelt thank you for creating the cover for this thesis: your touch always turns the ordinary into a gem! And thank you for believing in the value of what I do.

Thanks to GlennE, for handling an Italian drama queen like me with so much patience, for embracing the challenges of a different culture, and, above all, for always waiting for me.

Il mio più grande ringraziamento va ai miei genitori. Mi avete insegnato a seguire ciò che mi rende felice, anche quando non è la strada più facile. Forse è così che ho deciso di intraprendere un dottorato in un momento della vita meno convenzionale. Grazie a Ines e Umberto e a tutta la mia famiglia allargata, per essere le persone su cui posso sempre contare. È stata una lunga corsa, a volte in salita, spesso faticosa, ma sempre trasformativa. E Alessandro e Jacopo, siete stati il respiro che mi ha dato la forza di andare avanti.

Alessandra

Contents

Page

Abstract **i**

Riassunto **iii**

Résumé **v**

Thesis overview **vii**

List of abbreviations **ix**

I GENERAL INTRODUCTION

1 Context **1**

1.1 Knowledge mining from data 1

1.2 From pixels to insights 2

2 State of the art **3**

2.1 Hyperspectral imaging: merging chemical insight and spatial detail . . 3

2.1.1 Where spectroscopy meets imaging 4

2.1.2 Hyperspectral image 5

2.1.3 Image acquisition 6

2.1.4 Vibrational spectroscopy 9

2.1.5 Raman spectroscopy 11

2.1.6 LIBS spectroscopy 12

2.2 From pixels to insights: transforming data into knowledge 13

2.2.1 Clustering 16

2.2.2 Spectral Unmixing 23

2.2.3 Tensor decomposition methods 30

3 Research objectives **33**

3.1 Addressing key gaps in clustering and spectral unmixing methods . . . 33

3.2 Enhancing unsupervised hyperspectral analysis through tensor-based methods . . . 35

II ADDRESSING KEY GAPS IN CLUSTERING AND SPECTRAL UNMIXING METHODS

1 Introduction **39**

1.1 Clustering and unmixing . . . 39

1.2 Structure of the data . . . 40

1.3 HSI data: clustering or spectral unmixing? . . . 41

2 Methods **49**

2.1 Selection of the most relevant archetype points for exploratory analysis 49

2.2 MCR-ALS . . . 50

2.3 K-means clustering . . . 51

3 Datasets and software **53**

3.1 Raman powder dataset . . . 53

3.2 LIBS mineral dataset . . . 54

3.3 Software . . . 54

4 Exploratory analysis of hyperspectral images **57**

4.1 Raman powder dataset . . . 57

4.2 LIBS mineral dataset . . . 60

5 Final considerations and perspectives **67**

5.1 Some considerations . . . 67

5.2 Perspectives . . . 68

III ENHANCING HYPERSPECTRAL IMAGES ANALYSIS THROUGH TENSOR-BASED METHODS

1 Introduction **73**

1.1 Rationale . . . 73
1.2 Tensor rank . . . 75
1.3 Tensor decomposition methods . . . 75
1.3.1 CANDECOMP/PARAFAC . . . 75
1.3.2 Tucker family . . . 77
1.4 Block term decomposition . . . 80

2 The rank-(Lr, Lr, 1) decomposition **83**

2.1 Decomposition . . . 83
2.2 Algorithm . . . 85
2.3 Uniqueness . . . 85
2.4 Methodology . . . 86
2.4.1 Number of components determination . . . 86
2.4.2 Number of subfactors determination . . . 86
2.4.3 Model selection . . . 87
2.4.4 Validation . . . 88
2.5 Method outline . . . 89

3 Datasets and software **93**

3.1 Simulated data . . . 93
3.2 Stained fabric . . . 94
3.3 Remote sensing data . . . 95
3.4 Data preprocessing . . . 96
3.5 Software . . . 97

4 Exploratory analysis of hyperspectral images by decomposition in rank-(Lr, Lr, 1) terms **99**

4.1 Simulated data . . . 99
4.1.1 Validation . . . 101
4.1.2 Shuffled dataset . . . 102

4.2 Stained fabric . . . 103
4.2.1 Validation . . . 105
4.3 Remote sensing data . . . 106
4.3.1 Validation . . . 108

5 Final considerations and perspectives **111**
5.1 Some considerations . . . 111
5.2 Perspectives . . . 112

IV CONCLUSIONS

1 Final remarks and perspectives **117**
1.1 Addressing key gaps in clustering and spectral unmixing methods . . . 117
1.2 Enhancing unsupervised hyperspectral analysis through tensor-based methods . . . 118

V APPENDICES AND BIBLIOGRAPHY

A Supplementary material for Part II **123**

B Supplementary material for Part III **135**

C Scientific contributions **141**

D Publication: primary work 1 **145**

Bibliography **157**

Abstract

Imaging techniques have made significant advances, allowing for detailed analyses of material composition and structural or morphological features. Among these, hyperspectral imaging (HSI) stands out as a powerful analytical tool, capturing a spectroscopic measurement for each pixel in an image. This generates high-dimensional data, presenting challenges in analysis, particularly in managing the interplay between spatial and spectral information. In unsupervised analyses, clustering and spectral unmixing methods are commonly used, but they face limitations in scenarios with intricate spatial structures, severe spectral overlap, and complex chemical compositions. Moreover, spatial-spectral correlation is often overlooked or poorly addressed. To face these issues, this thesis proposes a novel exploratory approach based on a geometric interpretation of normalized scores and loadings obtained from Singular Value Decomposition (SVD) analysis. The objective of this method is to extract the most essential information, in terms of linear mixture analysis, without the need for optimization or complex calculations, particularly in scenarios with significant spectral and spatial overlap. The first main point of this thesis is to evaluate the advantages and limitations of clustering and unmixing, establishing a domain of applicability of these methods in relation to the data structure. It also explores scenarios where the data structure deviates from ideal conditions for these techniques, leading to challenges in effective analysis. Traditionally, the hyperspectral data cube is unfolded pixel-wise, converting the three-dimensional data into a two-dimensional matrix for chemometric analyses. While effective, this approach risks losing valuable spatial-spectral relationships. The second main point of this thesis is the exploration of tensor-based decomposition techniques. Although tensor decomposition has been widely applied in chemometrics, its use in HSI remains limited. By treating the data as a third-order tensor, this approach preserves their structure, allowing for simultaneous analysis of both spatial and spectral dimensions. This is particularly advantageous in complex cases where traditional methods struggle to capture the full information content due to the overlooking of spatial-spectral relationships. The thesis then introduces an unsupervised approach for HSI analysis based on the rank-(Lr, Lr, 1) block term decomposition (BTD). This tensor-based method demonstrates its potential

in handling complex cases where traditional techniques, which require the reshape of the data cube, are inadequate. The proposed approaches were first evaluated on simulated datasets to define their field of application. They were then applied to benchmark hyperspectral datasets and real case studies, demonstrating their utility in addressing complex spatial structures and significant spectral overlap across a broad range of applications. Overall, novel chemometric approaches that expand the palette of exploratory tools for hyperspectral images analysis and improve sample characterization in complex analytical scenarios were developed.

Keywords: spectral imaging, image analysis, clustering, spectral unmixing, BTD decomposition.

Riassunto

Le tecniche di *imaging* hanno compiuto progressi significativi, consentendo analisi dettagliate della composizione dei materiali e delle caratteristiche strutturali e morfologiche. Tra queste, l'*imaging* iperspettrale (HSI) si distingue come valido strumento analitico, in quanto cattura una misura spettroscopica per ogni pixel di un'immagine. Questo genera dati ad alta dimensionalità, che presentano sfide nell'analisi, in particolare nella gestione dell'interazione tra informazione spaziale e spettrale. Nelle analisi non supervisionate, i metodi di *clustering* e di *unmixing* spettrale sono comunemente utilizzati, ma incontrano limitazioni in scenari con strutture spaziali complicate, sovrapposizioni spettrali significative e composizioni chimiche complesse. Inoltre, la correlazione spaziale-spettrale è spesso trascurata o poco considerata. Per affrontare questi problemi, questa tesi propone un nuovo approccio esplorativo basato su un'interpretazione geometrica di *scores* e *loadings* normalizzati ottenuti dall'analisi di Decomposizione ai Valori Singolari (SVD). L'obiettivo di questo metodo è quello di estrarre le informazioni più essenziali, in termini di analisi di una miscela lineare, senza la necessità di ottimizzazione o di calcoli complessi, in particolare in scenari con una significativa sovrapposizione spettrale e spaziale. Il primo punto principale di questa tesi valuta i vantaggi e i limiti del *clustering* e dell'*unmixing*, stabilendo un dominio di applicabilità di questi metodi in relazione alla struttura dei dati. Inoltre, esplora gli scenari in cui la struttura dei dati si discosta dalle condizioni ideali per queste tecniche, portando a sfide per un'analisi accurata. Tradizionalmente, il cubo di dati iperspettrali viene spiegato pixel per pixel, convertendo i dati tridimensionali in una matrice bidimensionale per le analisi chemiometriche. Pur essendo efficace, questo approccio rischia di perdere preziose relazioni spazio-spettrali. Il secondo punto principale di questa tesi è l'esplorazione di tecniche di decomposizione basate sui tensori. Sebbene la decomposizione tensoriale sia stata ampiamente applicata in chemometria, il suo utilizzo in HSI rimane limitato. Trattando i dati come un tensore del terzo ordine, questo approccio preserva la loro struttura, consentendo l'analisi simultanea delle dimensioni spaziali e della dimensione spettrale. Ciò è particolarmente vantaggioso nei casi complessi in cui i metodi tradizionali faticano a catturare l'intero contenuto informativo, a causa della mancata considerazione delle relazioni

spazio-spettrali. La tesi introduce quindi un approccio non supervisionato per l'analisi delle immagini iperspettrali basato sulla decomposizione *block term* (BTD) con rango-(Lr, Lr, 1). Questo metodo basato su tensori dimostra il suo potenziale nella gestione di casi complessi in cui le tecniche tradizionali, che richiedono la riorganizzazione del cubo di dati, sono inadeguate. Gli approcci proposti sono stati prima valutati su dati simulati per definire il loro campo di applicazione. Sono stati poi applicati a set di dati iperspettrali di riferimento e a casi di studio reali, dimostrando la loro utilità nell'affrontare strutture spaziali complesse e sovrapposizioni spettrali significative in un'ampia gamma di applicazioni. Nel suo insieme, sono stati sviluppati nuovi approcci chemiometrici che ampliano la gamma di strumenti esplorativi per l'analisi delle immagini iperspettrali e migliorano la caratterizzazione dei campioni in scenari analitici complessi.

Parole chiave: *imaging* spettrale, analisi di immagini, *clustering*, *unmixing* spettrale, decomposizione BTD.

Résumé

Les techniques d'imagerie ont connu des avancées significatives, permettant des analyses détaillées de la composition matérielle et des caractéristiques structurelles et morphologiques. Parmi celles-ci, l'imagerie hyperspectrale (HSI) se distingue comme un puissant outil analytique, capturant un spectre pour chaque pixel d'une image, générant ainsi des données multivariées. Ces données posent des défis d'analyse, notamment dans la gestion de l'interaction entre l'information spatiale et spectrale. Dans les analyses non supervisées, les méthodes de partitionnement de données (*clustering*) et de démélange spectral (*unmixing*) sont couramment utilisées, mais elles présentent des limites dans les scénarios où les structures spatiales sont complexes, les recouvrements spectraux sévères et les compositions chimiques complexes. De plus, la corrélation spatiale-spectrale est souvent ignorée ou mal prise en compte. Pour faire face à ces problèmes, le premier point de cette thèse présente une approche exploratoire novatrice fondée sur une interprétation géométrique des *scores* et des *loadings* normalisés obtenus à partir de l'analyse par décomposition en valeurs singulières (SVD). L'objectif de cette méthode est d'extraire les informations les plus essentielles, en termes d'analyse de mélange linéaire, sans recourir à une optimisation ou des calculs complexes, en particulier dans des scénarios présentant un chevauchement spectral et spatial significatif. Cette thèse évalue les avantages et les limites du *clustering* et de l'*unmixing*, en établissant un domaine d'applicabilité de ces méthodes en fonction de la structure des données. Elle explore également des scénarios où la structure des données s'écarte des conditions idéales pour ces techniques, ce qui entraîne des difficultés pour l'analyse et l'interprétation. Traditionnellement, le cube de données hyperspectrales est déplié pixel par pixel, convertissant les données tridimensionnelles en une matrice bidimensionnelle pour les analyses chimiométriques. Bien qu'efficace, cette approche ne prend pas en compte les relations spatiales-spectrales. Le second point principal de cette thèse est l'exploration des techniques de décomposition de tenseurs. Bien que la décomposition par tenseur soit largement appliquée en chimiométrie, son utilisation en HSI reste limitée. En traitant le cube de données comme un tenseur d'ordre trois, cette approche préserve sa structure, permettant une analyse simultanée des dimensions spatiales et spectrales.

Cela s'avère particulièrement avantageux dans le cas où les méthodes traditionnelles peinent à capturer l'intégralité du contenu informationnel en raison de la non prise en compte des relations spatial-spectrales. La thèse introduit ensuite une approche non supervisée pour l'analyse des images hyperspectrales, basée sur la décomposition *block term* (BTD) de rang-(Lr, Lr, 1). Cette méthode démontre son potentiel pour gérer des cas complexes où les techniques traditionnelles, nécessitant une réorganisation du cube de données, sont inadéquates. Les approches proposées ont d'abord été évaluées sur des ensembles de données simulées afin de définir leur champ d'application. Elles ont ensuite été appliquées à des ensembles de données hyperspectrales de référence ainsi qu'à des études de cas réelles, démontrant leur utilité dans le traitement de structures spatiales complexes et de recouvrements spectraux significatifs dans un large éventail d'applications. Dans l'ensemble, des approches chimiométriques novatrices ont été développées, élargissant la palette d'outils exploratoires pour l'analyse des images hyperspectrales et améliorant la caractérisation des échantillons dans des scénarios analytiques complexes.

Mots-clés: imagerie spectrale, analyse d'images, partitionnement de données, démélange spectral, décomposition BTD.

Thesis overview

This thesis has been submitted to the "*Science de la Matière, du Rayonnement et de l'Environnement*" Doctoral School of the University of Lille (France) and the "*Models and Methods for Material and Environmental Sciences*" Doctoral School of the University of Modena and Reggio Emilia (Italy), under a cotutelle agreement. The work presented in this thesis was carried out in 20 months at the LASIRE laboratory of the University of Lille, and in 16 months at the Department of Chemical and Geological Sciences of the University of Modena and Reggio Emilia. The thesis is divided into four parts and incorporates the work of two papers, one of which has already been published (Appendix D). Additional articles were published during this period, but are not directly related to the core research and are listed separately in Appendix C for reference. The central theme of the thesis is the development and application of hyperspectral imaging approaches, focusing on unsupervised methods of analysis. The first part provides a comprehensive introduction, outlining the research project's framework and defining its objectives. The second part includes the work primarily conducted at the University of Lille, focusing on two key techniques for hyperspectral image analysis: clustering and spectral unmixing. The third part shifts the focus to spectral unmixing using an alternative approach based on the tensor decomposition method. This last work was conducted mainly at the University of Modena and Reggio Emilia. Both the second and third parts involve the analysis of hyperspectral datasets with different characteristics and spectroscopic techniques. Each part is systematically organized into sections on introduction, methodology, datasets, results, and conclusions. Finally, the fourth part is dedicated to final remarks and perspectives.

List of abbreviations

ALS: Alternating Least Squares

BTD: Block Term Decomposition

CANDECOMP: Canonical Decomposition

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

EI: Essential Information

FIR: Far Infrared

FLIM: Fluorescence Lifetime Imaging Microscopy

FNNLS: Fast Non-Negative Least Squares

HSI: Hyperspectral Imaging

LIBS: Laser-Induced Breakdown Spectroscopy

LOF: Lack of Fit

MCR-ALS: Multivariate Curve Resolution - Alternating Least Squares

MIR: Mid Infrared

MRI: Magnetic Resonance Imaging

MSI: Mass Spectrometry Imaging

NIR: Near Infrared

NMF: Non-Negative Matrix Factorization

NNLS: Non-Negative Least Squares

OPTICS: Ordering Points To Identify the Clustering Structure

PARAFAC: Parallel Factor Analysis

PARALIND: Parallel Factor Linear Dependency

PBM: Pakhira-Bandyopadhyay-Maulik Index

PCA: Principal Component Analysis

SIMPLISMA: Simple-to-use Interactive Self-Modeling Mixture Analysis

SSE: Sum of Squared Errors

SVD: Singular Value Decomposition

VCA: Vertex Component Analysis

VIS: Visible

I

GENERAL INTRODUCTION

1. Context

> *Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone as the first step.*
>
> **— John Tukey**

1.1 Knowledge mining from data

Since the 1970s, John Tukey, chemist, mathematician and statistician who significantly contributed to shaping modern data analysis, encouraged scientists to explore the data and formulate hypotheses that could lead to new data collection and experiments [1]. Data exploration is a preliminary step for identifying the most appropriate preprocessing, statistical methods, tools, and even sampling strategies for further data analysis. The aim of exploratory data analysis is to better understand the characteristics of the data, enabling the formulation of hypotheses based on observed patterns, rather than imposing *a-priori* hypotheses onto the data. Exploratory data analysis represents a first critical step in any research context, setting the foundation for all subsequent stages of analysis [2]. Although exploratory data analysis has long been an implicit component of scientific investigation, its formalization as a distinct phase of analysis highlights its importance in modern data-driven disciplines, enabling researchers to bridge the gap between raw data and meaningful scientific insights. Data exploration is a subset of the broader field of data mining, which integrates supervised and unsupervised methods from statistics, mathematics, and machine learning to extract information from large data sets [3]. Supervised methods rely on labeled data and require *a-priori* information of the sample, in contrast to unsupervised methods, which seek to reveal patterns and relationships without the need for prior knowledge of classes, revealing information through visualization and expert-driven interpretation. Unsupervised methods are used extensively to provide a first overview of the main sources of variability in the data [4]. In chemistry, the

process of analyzing and interpreting data is known as chemometrics, a fundamental branch of this scientific field [5]. This thesis focuses on unsupervised data exploration tools, which form the core of data mining strategies and play a critical role in guiding further analysis.

1.2 From pixels to insights

Hyperspectral imaging (HSI) provides an ideal context to explore the challenges and opportunities of unsupervised data techniques. The acquisition of hyperspectral images generates a three-dimensional data cube, which captures spatial and spectral information [6]. Measured pixels rarely contain selective wavelengths for a specific component, as they are more often multicomponent systems that contain mixed signals, noise, and other artifacts [4]. Exploratory analysis is of particular importance in HSI, aiding researchers in extracting meaningful insights and guiding the selection of appropriate models to address the high dimensionality and multivariate interactions inherent in these datasets, in order to get a comprehensive understanding of the data properties and complexity. Since its introduction into remote sensing in the late 1970s, HSI has evolved rapidly, driven by advancements in sensing technologies, computational power, and algorithmic innovation [7]. Initially applied to mineralogy, environmental monitoring, vegetation studies, and precision agriculture, it has since expanded to biochemistry, food processing, pharmaceutical research, and forensic investigations. This transition "from major to minor" applications, underscores the adaptability of HSI technology to diverse fields [4]. The increasing development of faster algorithms, improvements in hyperspectral camera, and the introduction of innovative data analysis methodologies have certainly expanded its analytical potential and applications. This thesis builds on these advancements, focusing on the development of unsupervised chemometric methods for HSI. By exploiting exploratory data analysis techniques, this research aims to improve our understanding of hyperspectral data and contribute to more effective data interpretation in various scientific disciplines. With this foundation set, we embark on an exploration of unsupervised chemometric methods for HSI.

2. State of the art

This chapter examines recent advancements in HSI, providing technical details on hyperspectral images and the methodologies central to this thesis. Hyperspectral image analysis generally follows two main approaches: one focusing on the spatial dimension, for example, through pattern recognition, object detection, gradient and contour analysis, texture analysis, edge detection, and the other considering the spectral dimension, which captures the chemical composition of each pixel through its spectral signature [8, 9]. The integration of information from both spatial and spectral dimensions remains one of the main challenges in hyperspectral image analysis. In this thesis, the spectral perspective is taken as a starting point. Specifically, the focus is on clustering techniques, spectral unmixing, and tensor-based methods, which are extensively analyzed to assess their current state of the art and to evaluate opportunities for further development.

2.1 Hyperspectral imaging: merging chemical insight and spatial detail

In the context of spectroscopy, significant advances in instrumentation and acquisition techniques have highlighted the need for corresponding advances in data analysis methodologies [10, 11]. This is particularly evident in chemistry, where sample investigations used to rely on univariate measurements to quantify and examine the properties and characteristics of some/few specific components of the samples. While this approach is effective in some cases, it is insufficient for addressing the complexity of chemical systems. Univariate methods fail to capture the heterogeneity of the samples as they provide only an averaged representation. The transition from univariate to multivariate analysis has increasingly improved analytical capabilities and, together with spectroscopy, has become a fundamental tool for scientists.

2.1.1 Where spectroscopy meets imaging

Chemical analysis took another step forward when attention was given to understanding not only the chemical composition of a sample, but also its spatial structure and the distribution of its constituents [12, 13]. For this scope, imaging techniques have also evolved, from grayscale to spectral images. This evolution can be followed in Fig. 2.1. An image can be defined as a two-dimensional representation of a surface created by any device able to capture the correlated information along the two directions. The smallest spatial entity of a digital image is called *pixel* and contains unique information that is also influenced by its neighbors [4]. Images can be classified in different ways, but the one that is useful in this context is a classification according to the amount of information contained in each pixel [4]. An image that contains intensity information for each pixel, corresponding to the amount of light reflected or emitted from that specific area of the image, is called *grayscale image* or *single channel image* [14]. Advancements in imaging technology led to the adoption of *color-based images*, which try to mimic the human vision. They are generally composed of three independent channels, for example, red (around 700 nm), green (around 550 nm), and blue (around 450 nm) that combined recreate the real colors, in this case, in the RGB color space [15].

But in the last 50 years, when spectroscopic techniques started to be integrated with spatially resolved information, chemical analysis has been elevated to a new level, allowing for a deeper and more comprehensive understanding of complex systems. The progress was from color based images to spectral images, covering broader spectral ranges, like *multispectral images*, which capture individual images at specific wavenumbers or wavelengths [16], to arrive at images that capture a portion of the electromagnetic spectrum, according to the technique used, in each pixel [6, 17]. This spatial and spectral information, associated with the location of compounds and their chemical composition, respectively, merges in the domain of *hyperspectral imaging*. HSI represents a significant advancement in science, offering detailed spectral and spatial information of samples, allowing the study of chemical systems in all their heterogeneity. For example, in the biomedical field, it can differentiate between healthy and diseased tissue or identify regions with distinct properties [18]. Similarly,

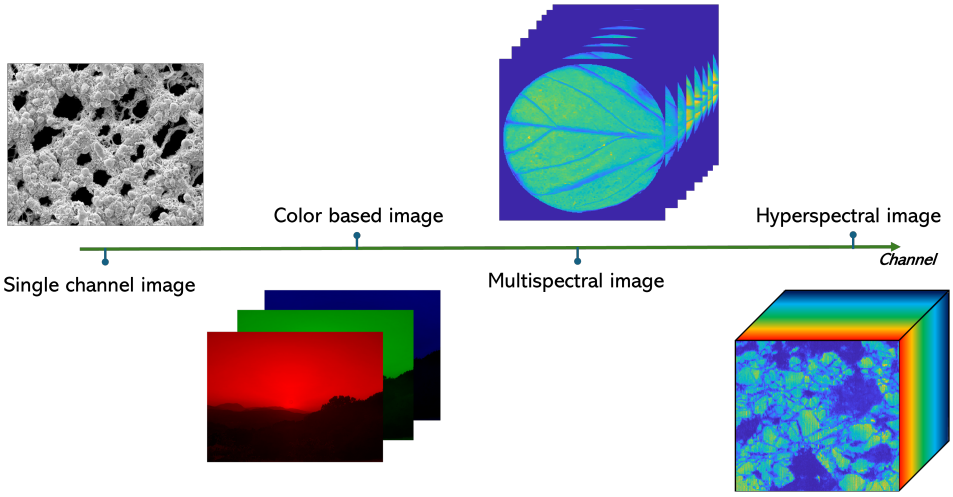

Figure 2.1: Image classification based on the number of channels contained in each pixel. Moving from single-channel images with one value associated to each pixel, to hyperspectral images involves an extension of the image in the third dimension, related to the spectral information.

in precision agriculture, HSI is used to monitor crop health, detect infected leaves, and to localize affected regions [19]. Initially developed for remote sensing in the 1970s [7], HSI has expanded into various fields over time, including cell analysis in biology [20], medicine [21], food science [22], agriculture [23], environmental monitoring [24] and galaxy exploration in astronomy [25]. These examples highlight the broad domain of applicability and practical advantages of this technique in solving analytical challenges [12, 26, 27]. Its capabilities range from analyzing pixels at nanometer scale to covering areas spanning several square kilometers. Innovations in sensors, faster acquisition methods, and algorithms for data analysis have made HSI more accessible and versatile over the years.

2.1.2 Hyperspectral image

HSI data are typically represented as a cube with two spatial dimensions, corresponding to the horizontal and vertical pixel axes, and a spectral dimension that reflects the spectral range (e.g. wavelengths, wavenumbers) of the spectroscopy used for image acquisition. This hyperspectral data cube can be visualized in different ways depend-

ing on the perspective taken, and it is represented in Figure 2.2. In one approach (Fig. 2.2 A), the data is viewed as a series of images, registered at each spectral channel. Alternatively (Fig. 2.2 B), the data can be seen as a cube where each pixel contains a full spectrum across the entire acquisition spectral range. An image in a specific spectral channel and the spectrum of a selected pixel can be extracted from the data cube. The extracted image provides spatial information about the localization of compounds in the chosen spectral channel, while the spectrum reveals the chemical composition of a given pixel. These two visualizations can be provided by different acquisition modalities.

2.1.3 Image acquisition

Hyperspectral images can be acquired using four main modalities. In *plane scanning mode* [28], the entire sample area is scanned in a single shot at a fixed spectral channel (which corresponds to Fig. 2.2 A), while the spectral range is scanned sequentially. Alternatively, the configuration in Fig. 2.2 B is a *point scanning* mode, where the illumination source is directed pixel by pixel. Point scanning offers high spatial and spectral resolution, but it is typically slower and more expensive. Plane scanning, on the other hand, is faster but provides lower spatial resolution, making it suitable for applications where rapid acquisition is essential, such as capturing chemical dynamics over time or when spatial distribution is particularly significant [29]. Other acquisition modes are *line scanning* [30], where a grid of sensor that form the detector records simultaneously a line of pixels, and *snapshot acquisition*, which captures the hyperspectral image in a single shot. While line scanning is faster than point scanning, it has lower spatial and spectral resolution. This method is widely used in remote sensing devices equipped with Vis-NIR imaging platforms [31], and nowadays, is one of the preferred technologies for bench-top instruments but also in industrial applications. Snapshot acquisition, has the great advantage of efficiency in light collection, as it avoids photons loss from unscanned lines, as in the case of line scanning mode, or from other wavelengths as in case of plane scanning mode. This type of HSI is still in evolution, and the high speed of acquisition, like 20 multispectral images per second, makes it worth exploring this technology for real-time monitoring

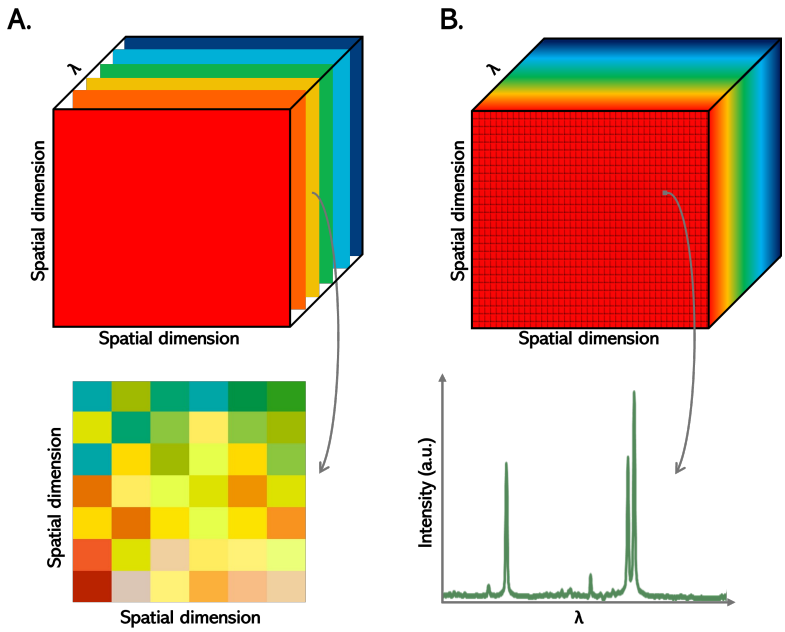

Figure 2.2: Schematic representation of a hyperspectral image data cube. The cube extends across two spatial dimensions and a spectral dimension. **A.** Each spectral channel can be viewed as a 2D images, while in **B.** each pixel contains a spectrum that reflects its chemical composition.

[32, 29]. Image acquisition is compatible with nearly all spectroscopy techniques, ranging from providing limited chemical details to offering very specific information, like in mass spectrometry imaging (MSI) [33, 34]. All acquisition modes are based on three key principles which depend on the interaction of light with the sample. These interactions provide the spectral data used to analyze the chemical composition and properties of the sample. The light from the source can be:

- absorbed by the sample, and it corresponds to specific wavelengths that are directly related to the chemical composition of the sample;
- reflected from the surface of the sample, which relates to both the chemical and morphological properties of the surface;
- scattered, when it is deflected in random directions due to particles or imperfections in the sample, providing information about its chemical and morphological properties;

- transmitted, which represents the light that is neither absorbed nor reflected and relates to the chemical composition and thickness of the sample.

Depending on the instrument setup, three main acquisition modalities are commonly used [35], and are graphically represented in Figure 2.3:

- *Reflectance mode*: where both the light source and the camera are positioned on the same side of the system, capturing light reflected from the sample's surface. It is the most used modality,
- *Transmittance mode*: used less frequently, this modality requires light to pass through the sample, with the hyperspectral camera positioned opposite the light source.
- *Interactance mode*: the light source and the camera are on the same side of the sample and aligned parallel to each other, with a light seal to prevent external light interference. It enables deeper penetration into the sample, reducing the surface effects compared to reflectance mode, and the sample thickness influence compared to transmittance mode.

The choice of the interaction mode influences the type of spectral information collected and the type of data analysis that can be conducted. Many studies in the literature compare results from different acquisition techniques to determine the best equipment for specific types of samples [36, 37, 38, 39]. The discussion of the technical characteristics of the instrumentation is not extended further in this thesis, as the images were not directly acquired during the PhD work. Instead, the focus of this research was on the development and application of chemometric approaches. The most commonly used spectroscopic techniques for acquiring hyperspectral images are in the domain of vibrational spectroscopy, including near-infrared (NIR) and mid-infrared (MIR) regions, as well as Raman spectroscopy [40]. Vibrational spectroscopy is a technique used to study the internal motion of atoms in molecules. Other techniques, such as mass spectrometry imaging (MSI), laser-induced breakdown spectroscopy (LIBS), and fluorescence lifetime imaging microscopy (FLIM) are gaining much popularity in the HSI context. In the following section, the spectral images analyzed in this thesis are briefly discussed.

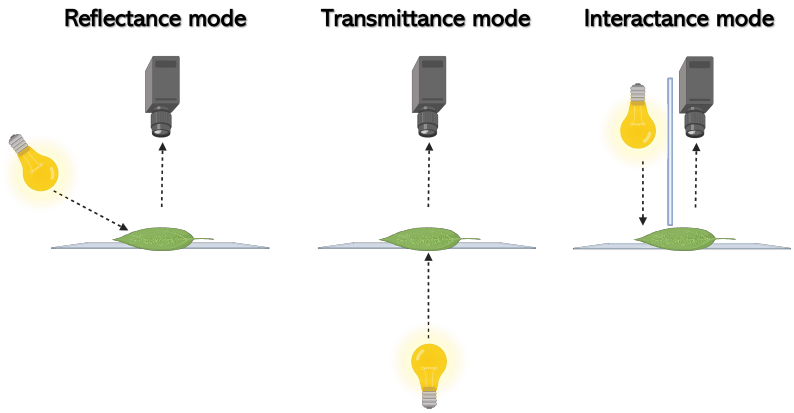

Figure 2.3: Hyperspectral image acquisition mode. Three different modes are defined based on the positioning of the light source relative to the detector: on the same side as the detector (as in reflectance and interactance modes) or on the opposite side (as in transmittance mode). In the interactance mode, a light seal is required to prevent external light interference. A light bulb is shown here for illustration, the light source can be of different type, including a laser beam.

2.1.4 Vibrational spectroscopy

Vibrational spectroscopy is a non-destructive and non-invasive technique used for the characterization of molecular composition of samples. When a molecule absorbs infrared light, it can undergo vibrational transitions, where its atoms move in characteristic ways, such as stretching or bending of chemical bonds, changing its dipole moments. The absorption patterns resulting from the light-matter interactions are characteristic of the types of bonds of the molecule, leading to spectra that can be interpreted. This spectroscopy can be associated with all the scanning modalities of image acquisition. Infrared radiation was discovered in 1800 by the astronomer Herschel while using a prism to refract sunlight and measure the temperature of different colors of light with a thermometer [41]. He observed that the temperature increased beyond the red part of the spectrum, where there was no visible light, leading him to conclude that there were wavelengths beyond visible light [42]. From that point on, the application of vibrational spectroscopy has grown significantly not

only in chemistry [43, 44], but also in biology [45, 46], medicine [47, 48], food science [49, 50], cultural heritage [51, 52], environmental science [53, 54].
Vibrational spectroscopy covers three spectral regions (see Fig. 2.4) [55, 56]:

- Near-infrared (NIR): typically from 780–2500 nm, mainly identifies overtones and combinations of vibrations, offering information on functional groups and chemical bonds.
- Mid-infrared (MIR): in the range of 2500–25000 nm, focuses on fundamental vibrations of molecular bonds, providing a more precise "chemical fingerprint" related to the structure and composition of chemical compounds.
- Far-infrared spectroscopy (FIR): spectroscopy typically operates in the wavelength range of approximately 25000 nm to 1 mm, providing insights into lower energy transitions, for the analysis of a variety of materials, mainly for organized solids.

In this thesis only images in the NIR range are used. The spectra acquisition in infrared spectroscopy measurements is generally done in transmission and reflectance mode [57]. In the transmission mode, infrared radiation passes through the sample, and the transmitted intensity is measured to calculate absorbance (A) using the equation:

$$A = -\log\left(\frac{I_0}{I}\right) \tag{2.1}$$

where I_0 represents the initial intensity and I is the transmitted intensity. This mode is suitable for analyzing transparent or very thin samples, as thicker samples may block most of the light, leading to signal saturation. On the other hand, the reflectance mode directs the infrared radiation onto the surface of the sample, and the detector collects the radiation reflected from the sample. Similar to the transmission mode, the reflectance mode measures the intensity of a reference beam and a reflected beam, calculating the absorbance based on the reflected intensity. This mode is ideal for thick or opaque samples, as it overcomes the limitations of light transmission in these materials. The typical spatial resolution of infrared imaging is around 10–50 µm, which is lower compared to other imaging systems such as Raman [58]. This resolution is sufficient for many applications, but improvements are needed, like

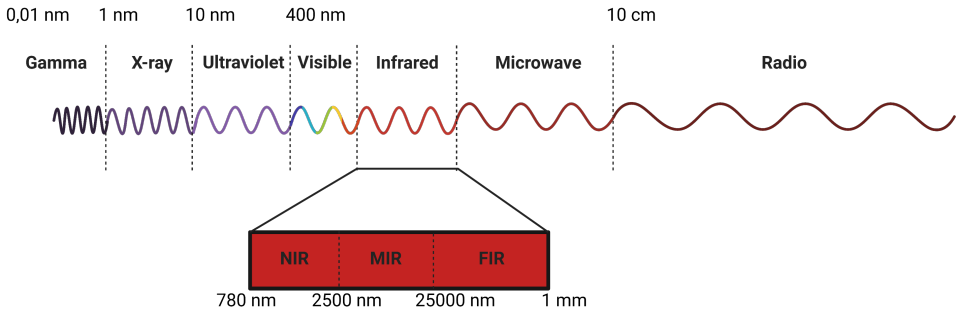

Figure 2.4: The electromagnetic spectrum, showing wave types and corresponding wavelengths, with the detail of the infrared portion of the spectrum.

using the synchrotron radiation source, for the analysis of biological tissues at cell level [59].

2.1.5 Raman spectroscopy

Raman spectroscopy is a powerful and non-destructive vibrational spectroscopy technique that provides detailed molecular and structural information about a sample, generally associated with a point scanning modality of image acquisition. It is based on the inelastic scattering of light, which occurs when a monochromatic light source, typically a laser in the visible (Vis) or NIR range, interacts with molecular vibrations in the sample. This scattering is known as Raman scattering, named after the Indian physicist Raman, who discovered this phenomenon in 1928 and who was awarded with the Nobel prize in Physics for his discovery in 1930 [60]. During the interaction between the laser beam and the sample, the incident photons excite the molecules of the sample to a virtual state. During this interaction a fraction of the incident photons is scattered inelastically [61, 62], producing a characteristic spectrum for the chemical bonds and the molecular structure of the material. In the elastic scattering, named Rayleigh scattering, the energy of the incident photons remains unchanged. The scattered photons are at the same energy level as the incident beam and the Rayleigh scattering does not provide molecular vibrational information. In the inelastic scattering, the scattered photons lose energy. This occurs when the molecules are in a lower vibrational state, and after the photons interactions they are excited to the

virtual state, but when they lose energy can end up to a vibrational state higher than they were before. It corresponds to emission of photons with low energy and this energy loss is represented in the Raman spectrum as peaks to the right of the Rayleigh scattering line (Stokes scattering). It can also happen that the scattered photons gain energy, anti-Stokes scattering, corresponding to the molecule of the samples ending up in a lower vibrational state than before excitation. This results in photons having shorter wavelengths and higher energy than the incident photons. Anti-Stokes scattering is less intense because higher vibrational states are less populated at room temperature. Raman spectroscopy is particularly advantageous for its specificity in differentiating chemically similar compounds because of the richness of information that the Raman technique provides. Also, the spatial resolution that can reach is around 500 nm of pixel size, depending on the laser employed [63]. However Raman signal are often weak limiting the detection of low concentration compounds [64]. It is extensively applied across various fields, including material science [65, 66], chemistry [67, 68], biology [69, 70, 71], pharmacy [72, 73], and quality control [74, 75].

2.1.6 LIBS spectroscopy

LIBS is an atomic emission spectroscopy, that requires minimal sample ablation. LIBS works by using a focused laser, typically generated by a Nd:YAG crystal, to remove a small amount of material from the sample surface, producing a vaporous plume above the sample. This vapor interacts with the laser and forms a plasma of ions on the sample surface, which prevents from further penetration of the laser beam. The amount of material affected is extremely small. This process also raises the temperature, and so ionizes the plasma and creates an emission of electromagnetic radiation including various processes for each species. This radiation carries information about the elements present in the sample, and it is captured through an optical fiber and transmitted to a spectrometer, that will generate a spectrum to facilitate element identification. A significant advantage of LIBS is that each element produces emission at distinct wavelengths, making it a useful technique for the identification of specific atoms. Already developed in the 1960s by Brech [76], it gained more popularity in the 1980s due to advancements in laser and detector technologies [77, 78, 79] and

LIBS has demonstrated several key advantages over the years. It eliminates the need for sample preparation, enables rapid multi-elemental analysis within fractions of a second, and is particularly effective in detecting light elements that other atomic emission methods struggle to observe [80]. Additionally, it can analyze all states of matter and can also be integrated with complementary techniques, such as Raman spectroscopy, to provide both elemental and molecular information [81, 82]. Despite these strengths, LIBS does have limitations, including a detection limit in the parts-per-million range, self-absorption effects that can reduce spectral signal accuracy, and matrix effects influenced due to the sample's physical and chemical properties [83]. Anyway, LIBS is a highly versatile and effective technique in determining the elemental composition of the material in numerous scientific fields [84, 83], including chemistry [85, 86], quality control [87, 88], material identification [89], metallurgy [90, 91], aerospace [92, 93], and environmental monitoring [94, 95].

2.2 From pixels to insights: transforming data into knowledge

The power of HSI to capture spatial and spectral information is a key advantage, but it also presents challenges. In fact, hyperspectral images contain a huge amount of data, for example, Raman and LIBS images typically have several thousand spectra and thousands of spectral variables. This vast amount of data introduces new exciting opportunities for research, particularly in terms of efficiently extracting the most meaningful and interpretable information from raw data. One of the key strategies for addressing these challenges involves the use of chemometric tools. There is a wide range of methods, tailored to different analysis objectives and data characteristics. Multivariate analysis can take two main approaches: supervised and unsupervised analysis. The first approach includes both quantitative and qualitative analyses. Quantitative analysis aims to quantify specific sample properties, such as concentrations or physical characteristics, based on spectral data. Techniques like regression [96, 97] are used to develop predictive models that estimate these properties for new samples [98, 99, 100]. Qualitative analysis, on the other hand, requires prior information about

the samples, which is actively used in model building. For example, in classification analysis, a subset of the data is labeled according to predefined categories and used to train a model that can then predict the class membership of new unlabeled samples [101, 102, 103].

The second approach, followed in this thesis, is an unsupervised analysis. This includes methods used when prior knowledge of the samples is unavailable. Unsupervised methods [104] aim to group or distinguish samples based on spectral similarities, detect specific components, and provide information about the composition and structure of the data [2]. Several methods are available for unsupervised analysis, and the choice of a particular approach depends on the measured data and the specific objectives of the analysis. In unsupervised analysis, the role of the data analyst is crucial. As no references are available, the workflow and results of the analysis should be carefully evaluated, with critical thinking and chemical knowledge. Although the validation of the model is an important step in the analysis and helps to assess its goodness, a critical spirit is always required for its evaluation.

An analysis of the existing literature on HSI shows that more than half of the publications focus on supervised techniques, primarily classification and regression, while only a small portion addresses unsupervised approaches. This trend is clearly illustrated in Figure 2.5, which shows the number of publications per year on HSI in general and those that specifically combine HSI with unsupervised analysis. The graph also highlights the exponential growth of publications on HSI, with a clear but much more modest increase in publications on unsupervised approaches. This trend is likely driven by the fact that industrial applications require predictive models to meet expectations, while fundamental research often deals with unknown samples, which is the essence of scientific inquiry. For this reason, we dedicate this doctoral thesis to the study of unsupervised analysis methods. We believe that the first step in any data analysis process should be unsupervised, even when some prior knowledge about the dataset is available. This exploratory phase provides valuable information to guide subsequent supervised analyses [2]. During recent years, our research groups have focused on developing methods to extract meaningful information from HSI

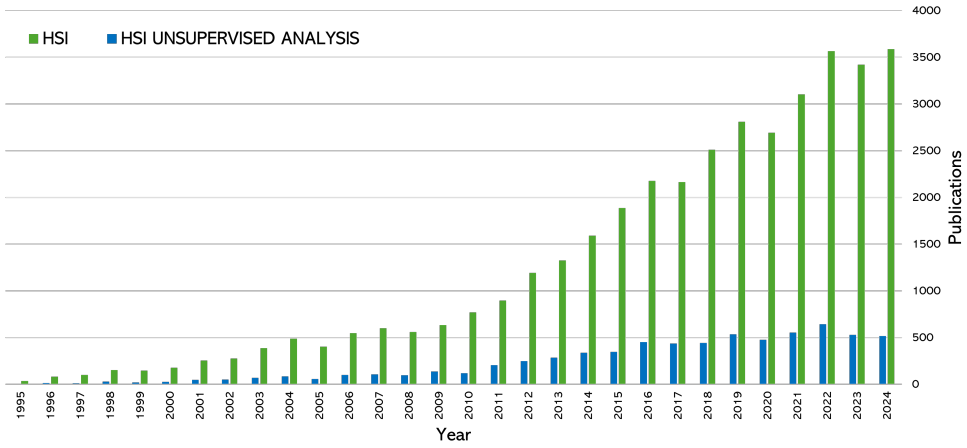

Figure 2.5: Number of publications per year. Publications related to HSI are shown in green, while those specifically focused on unsupervised analysis in HSI are shown in blue. Data source: Web of Science.

data [105, 106, 107] and have worked to combine spatial and spectral information encoded in the hyperspectral data cube [108, 109]. Moreover, we see great potential in further exploring this relatively underrepresented field to uncover new methodologies.

Among unsupervised techniques, Principal Component Analysis (PCA) [110, 111, 112] is the method of choice when it comes to multivariate exploratory data analysis. Firstly developed by Pearson in 1901 and then extended by Hotelling in 1933 in the psychology field, it is today used in many different applications [113, 114, 115, 116] and has become an essential tool also in the analysis of hyperspectral data [4, 6, 104, 117, 118, 119]. The idea behind PCA is to find directions in the multivariate space that provide the best approximation of the data, based on maximizing their variance and to minimize their residuals, that are the average distance between the coordinates of the approximated points and the original ones. These directions are orthogonal to each other and identifies a low-dimensional subspace to achieve the best possible approximation of the data for the chosen dimensionality. The assumption of orthogonal directions of the data is closely linked to two fundamental concepts of bilinear modeling and latent (or abstract) variables [120, 121]. In PCA, the interpretation is done through two matrices, scores which represent the coordinates of the original data points

when projected onto the new directions, and loadings, the coefficients (weights) that describe how much each original variable contributes to a given principal component. Two other methodologies that are the core of the unsupervised methodologies and find numerous applications in HSI [122, 123, 124, 125] are cluster analysis (clustering) and spectral unmixing, and are detailed in the following paragraphs.

2.2.1 Clustering

The guiding principle in the search for useful information is, in general, to identify the presence of non-random structures within the data. This is commonly pursued, in this context, by associating the concept of non-random structure with that of grouping, and investigating the presence of groups within the data space. Cluster analysis methods provide *possible answers* on the presence of groupings (clusters) using the concept of *similarity*. Let's clarify the concept of *possible answers* with a toy example. Consider a number of objects distributed in a space, with their top view shown in Fig. 2.6. Is it possible to identify groups? If so, how many? Each of us might give a different answer: some might say 2, others 3, 4, or even more, while some might say only 1. This diversity of responses is not surprising, as the problem does not have a single definitive solution.

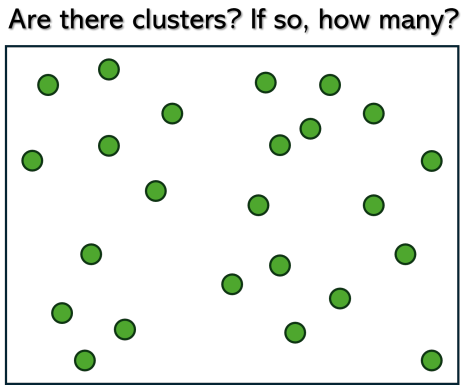

Figure 2.6: Distributions of objects in a space, top view. Answering the questions of whether there are clusters, and if so, how many, is not a trivial task.

Identifying groups, although conceptually simple, is a complex task because the answer typically depends on the perspective and objectives of the data analyst [126].

Now, consider that instead of a few objects, we have thousands of data points, each representing a pixel of the hyperspectral image. It becomes immediately apparent how complicated it is to unambiguously discriminate groups in HSI data. To do so, the concepts of *similarity* and *density* are of great importance in cluster analysis, and all clustering methods are based on these two concepts.

Distance measures

To quantify similarity/dissimilarity, the concept of distance is used: the greater the distance between two objects in a given context, the more dissimilar they are. Vice versa, the closer the objects are, the more similar they can be considered. A distance measure can be transformed into a similarity measure through a normalization of the data according to a standard scale:

$$s_{st} = 1 - \frac{d_{st}}{d_{max}}, \tag{2.2}$$

where d_{st} is the distance between objects s and t, d_{max} is the maximum distance between the objects or a distance used as a reference, and $s_{st} \to 1$ for similar objects and $s_{st} \to 0$ for dissimilar object. Some of the most commonly used distances in chemometrics for HSI data are reported below, where x_{sj} and x_{tj} are the j^{th} components of the vectors $\mathbf{x}_s$ and $\mathbf{x}_t$. The Minkowski distance, defined in Eq. 2.3 as:

$$d_{st} = \left(\sum_j |x_{sj} - x_{tj}|^r \right)^{\frac{1}{r}} \quad \text{Minkowski distance} \tag{2.3}$$

generalizes other distances, in fact for r = 2 the Euclidean distance is obtained, and for r = 1 the Manhattan distance [127]. Maximizing this distance ($r \to \infty$) the Lagrange (or maximum) distance is obtained [128]. Often clustering methods used the Euclidean distance by default. Euclidean distance considers the diagonal distance between the two measurements, Lagrange distance considers the largest distance between the measurements. The Manhattan distance is a metric used to measure the distance between two points in a space, considering only horizontal and vertical paths, without diagonals. Its name derives from the idea of movement on a street grid, like the one in Manhattan, where it is only possible to travel along orthogonal streets.

Compared to the Euclidean distance, the Manhattan distance can be more representative for high-dimensional or sparse data, as it considers each dimension independently. Additionally, it is particularly useful in images analysis where pixels are organized in grids, and provides a more accurate representation than the Euclidean distance in such cases. In addition, the calculation of this distance is faster than that of other distances. However, the Manhattan distance may be less suitable than more complex metrics, such as the Mahalanobis distance, in cases where clusters have irregular shapes [129, 130]. A very practical measure for spectral images is the correlation distance, that is based on the Pearson correlation coefficient between two points, $corr(x_s, x_t)$, from one [131, 132]:

$$corr(x_s, x_t) = \frac{\sum_{j=1}^{J}(x_{sj} - \bar{x}_s)(x_{tj} - \bar{x}_t)}{\sqrt{\sum_{j=1}^{J}(x_{sj} - \bar{x}_s)^2}\sqrt{\sum_{j=1}^{J}(x_{tj} - \bar{x}_t)^2}} \tag{2.4}$$

the correlation distance is defined as,

$$d_{st} = 1 - corr(x_s, x_t) \quad \text{Correlation distance} \tag{2.5}$$

with

$$\bar{x}_s = \frac{1}{j}\sum_{j=1}^{J} x_{sj}, \quad \bar{x}_t = \frac{1}{j}\sum_{j=1}^{J} x_{tj} \tag{2.6}$$

being the mean, over all the measured variables J, for each object. This measure helps to quantify how dissimilar the two points are based on their linear relationship. Many other distances are discussed and compared in the literature [126, 133, 134, 135], and the choice of one or the other is not trivial and can lead to very different results.

Density based methods

Density is another important concept in cluster analysis. A cluster is defined as a "dense" area that is evaluated by examining the neighborhood of each data point. Usually a radius is defined, and if the number of neighbors within this radius is lower than a given value (to be set), then that data point will be considered as an outlier. There are two common ways to define the neighborhood. The first approach

defines the neighborhood radius as the Euclidean distance to the j^{th} nearest neighbor, adjusting this radius to data density. In this case, the neighborhoods are smaller in dense regions and larger in sparse regions of the data space. OPTICS (Ordering Points to Identify the Clustering Structure) [136] is an example. The second approach assumes a fixed neighborhood radius for all points, regardless of data density [137], as in DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [138]. OPTICS is less practical for image analysis due to its computational cost. On the other hand, DBSCAN offers several advantages in identifying clusters of varying shapes and effectively handling noise, but requires the estimation of the radius and the number of minimal neighbors within it, which are difficult to assess, especially for large datasets [137, 139, 140].

Similarity based methods

Similarity-based clustering methods are divided into hierarchical and partitioning approaches. Hierarchical clustering provides, as a result, a dendrogram showing at which level of similarity samples are clustered, based on a linkage function that defines the criterion with which samples are assigned (or split from) to already formed clusters (once the first cluster has been defined) [141]. However, it is computationally intensive to be used efficiently on hyperspectral images. Hierarchical methods are classified into agglomerative and divisive types. Agglomerative methods, which are more commonly used, start with individual samples and iteratively merge them into clusters. In contrast, divisive methods begin with a single large cluster containing all samples and progressively divide it into groups at each step [97, 142].
An ideal clustering technique should [143]:

- have few parameters to tune;
- handle large datasets efficiently;
- identify clusters of any shape, aligning with natural cluster structures.

From this perspective, partitioning clustering methods are more useful for hyperspectral data [140]. With this approach, the dataset is divided into a predefined number of clusters, optimizing some clustering criteria to determine the best partitioning. These

methods assign each data point to exactly one cluster, after an iterative optimization. The most popular partitioning method is K-means [144]. A modification of K-means, named fuzzy, involves modulated response rather than a binary one, allowing data points to belong to multiple clusters simultaneously, with varying degrees of membership [145, 146]. This means that a point can partially belong to one cluster to a certain extent, and to another cluster to a different extent [147].

K-means clustering

K-means clustering only requires as input the number of clusters (K) to be retrieved, i.e. this can be specified by the user or selected by trying different numbers and using a cluster validity criterion to assess the best one; the output will be the vector containing the cluster membership of each sample.
The algorithm proceeds as outlined below:

1. initialization of K centers (initial guess for the cluster centroids) randomly;
2. the objects are assigned to the K clusters (at starts randomly);
3. computation of each formed cluster's centroid;
4. computation of the distance of each object from each defined cluster centroid;
5. re-assignment of the object to the cluster defined by the closest centroid;
6. back to step 3 if at least one object switches cluster. Otherwise, it stops.

During this iterative process, K-means minimizes the sum of squared error (SSE) [148] as in Equation 2.7:

$$SSE = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{s \in C_K} \left(d_{sj} - \overline{d}_j^{(k)} \right)^2 \tag{2.7}$$

where d_{sj} is the data entry of a single data point for the j^{th} variable , $\overline{d}_j^{(k)}$ holds the data value of the centroid for the j^{th} variable in cluster C_K, and J is the total number of variables.
K-means clustering performs well when clusters have a globular shape, similar densities (i.e., a homogeneous dispersion within each cluster) and contain a comparable number of objects (i.e., clusters are of similar size). However, despite its strengths,

K-means clustering is subject to certain limitations, mainly in two aspects. The first concerns the algorithm's tendency to converge to a local, rather than a global, optimum. The final results heavily depend on the initial centroids selection, which can lead to suboptimal solutions. To address this, several strategies have been proposed [149, 148, 150, 151, 152]. The random initialization implies that different solutions are obtained by restarting the algorithm, this issue can be mitigated by running multiple iterations of the algorithm and selecting the configuration with the lowest SSE. When the cluster structure departs from the globular shape and clusters are not well separated, K-means often fails to retrieve the correct clusters. A common approach to overcome this issue consists in starting with a larger number of clusters and after grouping those that are less distinct. This is a variant of K-means in which the clusters are grouped using a hierarchical method. Alternative strategies for K-means initialization (step 1 above) are:

- using the results of hierarchical clustering as a starting point for K-means, with the hierarchical method determining an initial cluster configuration for K-means to refine.
- random partitioning of the data into K groups, calculating their means, and using these as starting centroids. Despite these efforts, the sensitivity of K-means clustering to the random initial step remains a drawback.

The second challenge lies in determining the optimal number of clusters [153]. Some clustering optimization methods allow the number of clusters to be dynamically adjusted based on predefined distance metrics or other criteria. This flexibility should slightly improve the partitioning of the data overcoming the limitation of the prefixed number of clusters [154]. The Gap Statistic evaluates the total within-cluster variation for various values of K and compares them to the expected values under a null reference distribution of the data. The optimal value of K is the one that produces the largest gap statistic [155]. The use of graphical methods, like the elbow method is a popular approach for determining the optimal number of clusters in K-means clustering. It involves plotting the SSE curve, against the number of clusters, as shown in Fig. 2.7 . The ideal number of clusters is indicated by the point where the curve begins to flatten, signaling that adding more clusters yields minimal reduction in

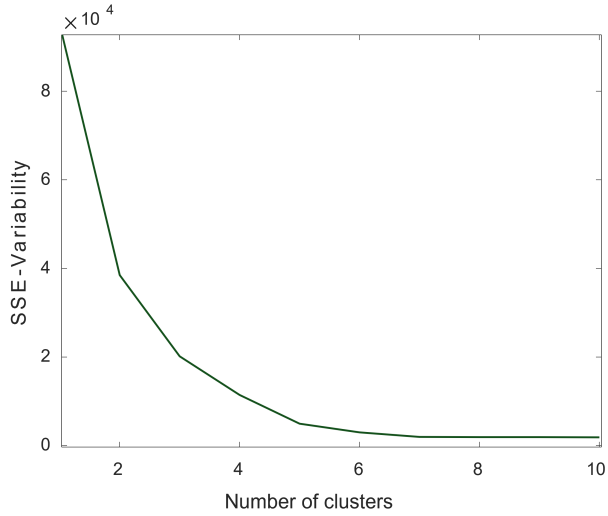

Figure 2.7: SSE curve as a function of the number of clusters. SSE is the sum of the squared differences between each observation and its group's mean. When the curve starts to flatten, adding more clusters results in minimal variance reduction. In this plot, the 'elbow' occurs at five clusters, suggesting it as a possible choice for the optimal number of clusters.

variance. However, picking the elbow of the curve, is considered to be too subjective [156, 157]. Scientific literature offers also several indices that provide information on the goodness of clusters. Each index has its own domain of applicability and limitations [158]. Many of these indices rely on distance measures to evaluate both cluster separation and cohesion. Indices such as the Silhouette [159] or the Pakhira-Bandyopadhyay-Maulik (PBM) [160], are often applied when working with HSI data and partitioning clustering [161]. However, validating a cluster analysis is not an easy task. Indices generally evaluate the variance within and between clusters.

It has also to be considered, that this method conducts data analysis without taking into account both the spatial and spectral information. Pixels or spectral channels are in fact analyzed individually, and an approach is proposed in the literature to couple the spectral and the spatial dimensions [162]. Despite the various proposals and the different algorithm implementations [85, 163, 164, 165, 166], the results of K-means clustering should be carefully analyzed by the data scientist, as the method is sensitive to the distance metric used and the choice of the expected number of clusters, which can significantly influence the results. These decisions require critical

thinking and expertise, as they directly affect the interpretation of the data and the quality of the clustering. Therefore, a thorough understanding and careful application of K-means clustering is essential to gain valuable knowledge from the data.

2.2.2 Spectral Unmixing

Spectral unmixing approaches aim to recover pure component spectra when analyzing mixture data, assuming a bilinear model. This is desirable in many applications, and in particular in HSI, when imaged samples are analyzed in terms of its constituents composing a mixture [17, 167, 168].

The bilinear model

The advantages of using a bilinear model include:

- compliance of spectroscopic data with linear models;
- simplicity in the interpretation of complex data;
- data compression by reducing a large dataset to a small number of meaningful sample constituents (referred to as *components*), which is especially important for large datasets like HSI.

A linear model assumes a linear relationship between the concentration of a chemical compound and its absorbance. This is described by the Beer-Lambert law [169]:

$$A_\lambda = \varepsilon_\lambda c l \tag{2.8}$$

where A is the absorbance at a specific wavelength λ, ε is the molar absorption coefficient at that λ, c is the concentration, and l is the pathlength. This relationship can be extended to mixtures of multiple chemical compounds, denoted by N, across a set of samples i. It is expressed as:

$$d_{i,\lambda} = \sum_{n=1}^{N} c_{i,n} \varepsilon_{\lambda,n} l \tag{2.9}$$

where $d_{i,\lambda}$ is the absorbance for sample i, at a specific wavelength λ, $c_{i,n}$ is the concentration of compound n in sample *i*, and $\varepsilon_{\lambda,n}$ is the molar absorption coefficient

for compound n at wavelength λ. Thus, the total absorbance of a sample is the sum of the absorbances of each compound in the mixture. If all wavelengths λ, are considered, the absorbance spectrum d_i of sample i becomes:

$$d_i = \sum_{n=1}^{N} c_{i,n} \mathbf{s}_n^\top \tag{2.10}$$

where $\mathbf{s}_n^\top$ is obtained considering $\varepsilon_{\lambda,n}$ over all wavelengths, and represents the pure absorbance spectrum for compound n. The bilinear model can be further written in matrix notation as:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \tag{2.11}$$

where $\mathbf{D}$ $(I \times \lambda)$ contains the spectra for I samples related with mixtures of N components, across λ spectral channels, and $\mathbf{C}$ $(I \times N)$ and $\mathbf{S}$ $(\lambda \times N)$ are the matrices containing the concentrations and the spectra profiles for each of the N components in the sample, respectively. The matrix E $(I \times \lambda)$ represents the unexplained variance, often attributed to noise. Because of the bilinear nature of spectroscopic techniques, this relationship can effectively describe hyperspectral images. To achieve this, the hyperspectral data cube of size (X, Y, λ) is reshaped into a matrix $\mathbf{D}$ of size $(X \times Y, \lambda)$, where each pixel is arranged one below the other. As shown in Figure 2.8, this configuration allows the data to be analyzed as a combination of spectra $\mathbf{S}$, and their respective proportions $\mathbf{C}$. This matrix $\mathbf{C}$ can be reshaped back into spatial distribution maps that represent the spatial abundance of each component in the sample, providing both spatial and concentration information. With a simple bilinear model, valuable information about the chemical composition and distribution of components in a sample are obtained [17, 137].

Unmixing methods

In real case applications, the pure spectral profiles $\mathbf{S}$ of the components are often unknown. Therefore, unmixing methods aim to decompose the hyperspectral data cube into its underlying components using only the information available in the data itself [170, 171, 172]. These methods are widely used in chemistry and HSI. Unmixing methods can be:

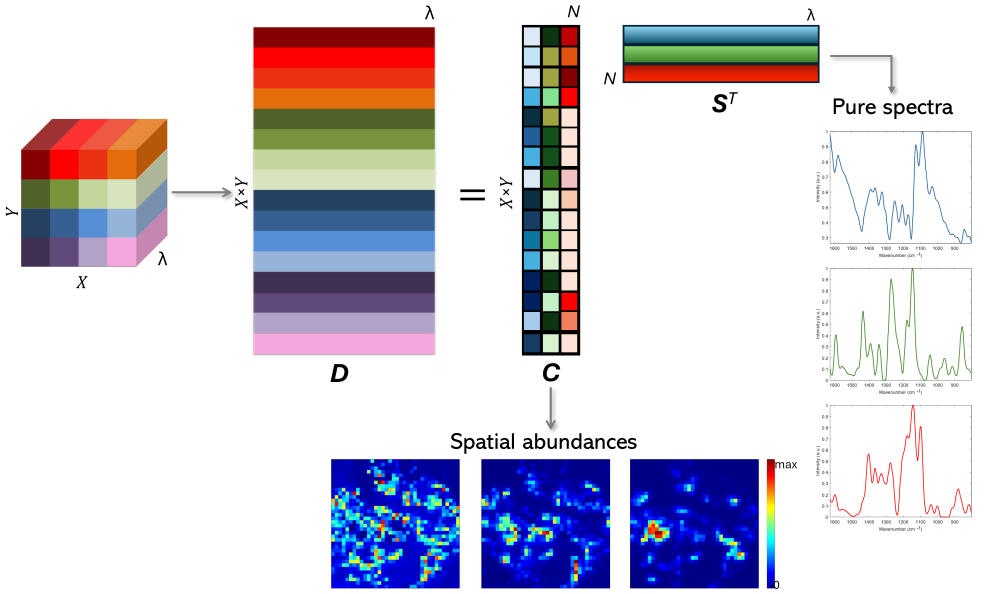

Figure 2.8: Bilinear model of a hyperspectral image. First, the hyperspectral data cube is unfolded into a matrix $\mathbf{D}$ $(X \times Y, \lambda)$. Then, $\mathbf{D}$ is expressed as a bilinear model, $\mathbf{CS}^T$. By refolding the pure concentration profiles $\mathbf{C}$ $(X \times Y, N)$, the spatial abundance for each component n is retrieved. The N pure spectral profiles are the column vectors of the matrix $\mathbf{S}$ (λ, N).

- Two-way methods, that work on the unfolded data matrix and include techniques such as Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) [173, 174, 175, 176]. In remote sensing applications, nonlinear spectral unmixing methods are sometimes employed to account for nonlinear interactions, such as light scattering [177, 178, 179].

- Higher-way methods, that analyze the hyperspectral data cube without the need for reshaping [180, 181, 182]. They are particularly advantageous for preserving the spatial structure of the data, and are discussed further in this chapter.

Among all these approaches, MCR-ALS is one of the most versatile and widely applied techniques in spectroscopy (e.g. NIR, Raman), chromatography, and HSI, due to its ability to produce chemically/physically meaningful results [86, 173, 183, 184, 185, 186, 187, 188].

Multivariate Curve Resolution - Alternating Least Squares

MCR-ALS method follows four main steps that are summarized below:

1. Determination of the number of components that describe the dataset **D**.
2. Initial estimation of the pure components spectra matrix, $\mathbf{S}$, or the concentration matrix, $\mathbf{C}$, is selected.
3. Iterative least squares optimization of $\mathbf{C}$ and $\mathbf{S}$ as:

$$\mathbf{C} = \mathbf{D}\mathbf{S}(\mathbf{S}^\top\mathbf{S})^{-1}, \tag{2.12}$$

$$\mathbf{S}^T = (\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{C}^\top\mathbf{D}. \tag{2.13}$$

   Alternate between updating $\mathbf{C}$ and $\mathbf{S}^T$, applying constraints after each step.
4. Stop the iterative process when convergence is achieved.

Each phase is extremely important in producing a valuable analysis, and important aspects are detailed for each step.

1. Determination of the number of components

The first key aspect of MCR is estimating the appropriate number of components needed to describe the data under investigation. In other words, this means estimating the chemical rank of the data. Chemical and mathematical ranks are equal in case of absence of noise and bilinearity is fulfilled. The most common method to evaluate the rank of a data matrix is based on the observation of a scree plot obtained through PCA [110, 111, 112]. Introduced by Cattell [189], this method visualizes the eigenvalues for each principal component in a plot that typically shows a steep decline, where the "elbow"point separate significant to less significant components. Components to the left of this point are considered to carry meaningful chemical information, while those to the right are associated with noise. This method is simple in principle, but its application can not be suitable where noise is significant. In fact, the discrimination of pure spectra that are highly correlated and the detection of minor compounds in close proximity to the noise level remains difficult, particularly in complex data matrices with low variance explained by these compounds. That is why, knowledge about the sample and the chemistry of the dataset should always be considered by the data analyst together with the chemometric tools.

2. Initial estimation of $\mathbf{C}$ *or* $\mathbf{S}$

The initial estimation of the matrices $\mathbf{C}$ or $\mathbf{S}$ is a critical step in MCR, as it significantly impacts the convergence of the model. If prior knowledge about the sample is available, such as a ground truth, it should be utilized as initial estimates. When such information is not available, various methods can be employed to estimate the purest spectral profiles or concentration values in the dataset. These methods typically involve selecting rows (spectra) with the highest presence of a single component, or identifying specific columns (wavelengths) where the signal is predominantly related to one component. Several approaches are commonly used for this purpose, focusing on identifying the most dissimilar rows and columns in a data matrix. Examples include the SIMPLe-to-use Interactive Self-modelling Mixture Analysis (SIMPLISMA) method [190, 191, 192], sometimes combined with more recent techniques for selecting essential spectra and variables. The latter aims to identify spectra or wavelengths that comprehensively describe the entire dataset [193, 194, 195]. While SIMPLISMA has practical applications for various data types [196, 197, 198], the selection of essential spectra and variables is more used in the context of HSI [68, 106, 199]. For the selection of the initial estimates is important to avoid a random selection, as the method optimizes the profiles iteratively, and starting with an estimation far from the true solution may prevent convergence.

3. Alternating least squares optimization

MCR-ALS proceeds with an iterative alternating least squares optimization of $\mathbf{C}$ and $\mathbf{S}$. In hyperspectral image unmixing, this process typically begins by estimating $\mathbf{C}$ from the data $\mathbf{D}$ using $\mathbf{S}$ as an initial estimate, as shown in Eq. 2.12. Then, $\mathbf{S}$ is updated using the newly estimated $\mathbf{C}$, as in Eq. 2.13. This iterative process continues until convergence is achieved. The iteration could start also from $\mathbf{S}$ with initial estimates of $\mathbf{C}$, in this case the calculation of $\mathbf{C}$ and $\mathbf{S}$ is inverted. A very important aspect of this stage is the application of constraints during the iterative process. Constraints impose mathematical or physical conditions on the profiles, improving chemical interpretability [200, 201]. MCR-ALS supports constraints applied to entire matrices, individual elements, or specific components, making it a versatile method suitable for a wide

range of applications and different types of data. Some constraints are developed to reflect the investigated chemical system, while others are more mathematical [167, 173, 202, 203].

Non-negativity constraint was the first being proposed, and ensures that all values in the constrained matrix are positive. This reflects the properties of concentrations and spectral intensities, particularly in spectroscopic techniques that provide positive signals. The constraint works by replacing negative values in the calculated matrices $\mathbf{C}$ and/or $\mathbf{S}$ with zeros or by employing softer algorithms such as non-negative least squares (NNLS) [204] or fast non-negative least squares (FNNLS) [205], both of which are commonly used in HSI.

There are also constraints tailored specifically for image analysis like segmentation or smoothness constraints [206, 207].

Local rank constraint identifies regions where specific components are absent (also referred to as "windows"), enforcing zero values in these regions in either $\mathbf{C}$ or $\mathbf{S}$. This improves the accuracy of the retrieved profiles[201]. A specialized version of the local rank constraint has been adapted for application to the distribution maps of images [208, 209, 210].

4. Convergence criterion

Establishing the convergence criterion in ALS optimization is another important aspect to consider in MCR, as the convergence criterion ensures accurate results while avoiding excessive computations. Too few iterations may lead to suboptimal solutions, whereas too many can waste computational time without significant improvement [211]. Convergence criteria can involve either setting a fixed number of iterations, which must be predefined as a parameter in ALS, or monitoring changes in the lack of fit (LOF) of the model, defined as:

$$LOF(\%) = 100 \times \sqrt{\frac{\sum_{ij} e_{ij}^2}{\sum_{ij} d_{ij}^2}} \tag{2.14}$$

where d_{ij} and e_{ij} are elements of the original matrix $\mathbf{D}$ and residual matrix $\mathbf{E}$ (defined as $\mathbf{E} = \mathbf{D} - \mathbf{C}\mathbf{S}^T$), respectively. This second approach is generally preferred as it is less arbitrary and also allows for tracking the solution's evolution. However, a

convergence threshold, generally indicated as the relative difference in the standard deviations of the residuals between two successive iterative calculations, must be set. However, small fit changes do not always mean negligible profile variations in $\mathbf{C}$ and $\mathbf{S}$, particularly when profiles are highly correlated. In such cases, using a very low convergence criterion ($\leq 10^{-6}\%$) helps prevent premature stopping [212]. That is why researchers should monitor the changes in the shape of the profile along with the convergence of the fit to ensure accurate results.

Despite bilinear decompositions are powerful and flexible method, they may suffer of ambiguity. This means, in general, that for a given rank, there will be sets of $\mathbf{C}$ and $\mathbf{S}$ solutions that are different in their profiles, but that fit the model equally. In theory, in absence of noise and constraints, all solutions are equally probable [213]. There are three main types of ambiguities: scale, permutation, and rotational ambiguities. Scale and permutation ambiguities are generally less problematic from a data analysis perspective, as they do not alter the shape of the profiles or affect result interpretability. Scale ambiguity occurs when profiles have identical shapes but different relative scales and equivalent fits. This can be mitigated through normalization. Permutation ambiguity arises when the extracted profiles appear in a different order, but their shape, as well as the fit, remain unchanged. Rotational ambiguity is the most critical in a bilinear model because it describes the possibility to obtain solutions that are different in their profiles, but still fitting the model in an identical way [214]. This can be visualized, writing the equation for the bilinear model of D as:

$$\mathbf{D} = \mathbf{C}(\mathbf{T}\mathbf{T}^{-1})\mathbf{S}^{T} \tag{2.15}$$

where $\mathbf{C}$ and $\mathbf{S}$ matrices are multiplied by an identity matrix $\mathbf{T}\mathbf{T}^{-1}$. Eq. 2.15 can be rearranged as:

$$\mathbf{D} = (\mathbf{C}\mathbf{T})(\mathbf{T}^{-1}\mathbf{S}^{T}) \tag{2.16}$$

and then:

$$\mathbf{D} = \mathbf{C}'\mathbf{S}'^{T} \tag{2.17}$$

meaning that the data $\mathbf{D}$ can be described using different components that are linear

combinations of the chemical species of the data. However, these solutions may not be physically or chemically meaningful. To mitigate these ambiguities, constraints are applied during the ALS process to reduce the degrees of freedom in the solution space. However, not all constraints are equally effective in addressing ambiguity [201, 215, 202, 216]. The most effective way to resolve ambiguity is through selectivity, meaning that some data points (e.g., pixels in an image or specific wavelengths in a spectrum) should contain information from only a single component. This selectivity will anchor the solution and minimize rotational ambiguity.

In the context of HSI, another limitation of MCR-ALS, and two-way methods in general, is that spatial information is ignored. To cope with this issue, there can be different approaches, one, as mentioned before [206, 207], is to impose constraints on the spatial distribution, another is to incorporate wavelet transform based methods for spatial features enhancement in the MCR framework as described in [108, 217], a further one could be to use trilinearity constraints [218, 219, 220], which have been proposed to deal with excitation-emission landscapes [203, 221], which fulfill this underlined model, but so far have still limited application in the HSI context [222, 223]. The research interest has then shifted towards unmixing methods that analyze the tensor in its original form.

2.2.3 Tensor decomposition methods

The first proposals for three and higher-way generalizations of factor and principal component analysis date back to the 1960s and early 1970s. Over the years, various methods have been developed in different contexts [180, 224, 225, 226], which can be broadly categorized into two main families: Parallel Factor Analysis (PARAFAC) [227], also known as Canonical Decomposition (CANDECOMP) [228], and Tucker models [229, 230]. A comprehensive overview of these methods is provided in Section III. In the context of HSI, tensor approximation in low-rank terms is an emerging technique that has attracted significant attention, mostly supported by a rapidly evolving of theoretical foundation [231, 232, 233, 234]. Ongoing research continues to refine and expand both the methodologies and applications of low-rank tensor approximation,

particularly within the mathematical and remote sensing communities. Notable applications and developments in this field can be found in [235, 236, 237, 238, 239, 240, 241], with further discussion on the strengths and limitations of these methodologies provided in Section III. The final part of this thesis will focus on hyperspectral images decomposition in low-rank terms.

3. Research objectives

HSI has emerged as a powerful analytical tool, capable of acquiring detailed spectroscopic information for each pixel within an image. This generates high-dimensional datasets that allow scientists to analyze both the chemical composition and the spatial distribution of components within a sample. As a result, HSI has become an essential tool in a wide range of applications, such as assessing the homogeneity of chemical formulations in pharmaceutical products [242], identifying and localizing biological fluids for forensic analysis [243], and exploring the composition and spatial distribution of chemicals in raw materials [244], among others. However, the complexity of HSI data also presents significant challenges.

The global aim of the present doctoral thesis is to contribute to the development of novel exploratory approaches for the analysis of hyperspectral images.

In order to achieve the global purpose, gaps have been identified for well-known unsupervised methodologies currently used for the analysis of HSI data: spectral unmixing and clustering. Based on the current scientific literature, hypotheses and corresponding objectives have been raised and two main research directions are considered.

3.1 Addressing key gaps in clustering and spectral unmixing methods

The most common techniques for hyperspectral data, clustering and spectral unmixing, have two very different aims: clustering is used to partition the data to identify groups, while spectral unmixing aims to resolve the data into its pure components. In this regard, two main gaps have been identified and two specific objectives have been addressed:

- **Gap 1:** in the current scientific literature, there is considerable confusion regarding the appropriate application of unsupervised methods such as clustering and

spectral unmixing methods. These two techniques are often used interchangeably or on datasets that may not be well-suited for either approach, leading to suboptimal results.

We argue that a complementary combination of clustering and spectral unmixing methods can improve the discrimination of spectral features, the identification of pure spectral components, and the interpretability of results, depending on the specific aim of the analysis.

The **first objective** is to develop a general framework for applying clustering and spectral unmixing methods in relation to the characteristics and inherent properties of the data, and to evaluate the potential synergy of these unsupervised chemometric techniques, with respect to the aim of the analysis.

- **Gap 2**: when analyzing large datasets with complex spatial and/or spectral overlap and mixed information in each pixel, unsupervised approaches may yield suboptimal results. Clustering methods often involve random initialization steps, which can lead to inconsistent results. Spectral unmixing techniques typically begin with an initial estimation phase of component concentrations or spectral profiles, which can be further refined. In addition, minor components in the data may be overlooked, particularly in the absence of prior knowledge about the sample. These challenges reduce the reliability of the methods in unsupervised contexts.

  The effectiveness and interpretability of clustering and spectral unmixing methods for hyperspectral image analysis can be improved by considering the geometric characteristic of the HSI data structure to guide the direct extraction of spectra and images, rather then focusing directly on optimizing or resolving spectra and images, in other words guiding the selection of methods, or their parameters, by an exploratory step. This approach could better exploit complex data structures and improve the identification of components.

  The **second objective** is to develop novel methods that take into account structure and the geometry of data points in the feature space, reducing the impact of the random initialization step and improving the identification of the components, including minor components.

These objectives are addressed in the second section of the thesis, which focuses on the analysis of clustering and spectral unmixing methods. The gap between these approaches is investigated to explore the potential for a novel method applicable in scenarios where the conditions for traditional clustering and spectral unmixing are not fully met. This section emphasizes a data-driven exploration of hyperspectral images data.

3.2 Enhancing unsupervised hyperspectral analysis through tensor-based methods

In most existing unsupervised approaches, the first step of the analysis involves reshaping the hyperspectral data cube and analyzing it as a matrix. In this regard, one main gap has been identified:

- **Gap 3**: the unfolding of the data cube into a matrix disregards the spatial-spectral relationships within the hyperspectral data. This omission can lead to the loss of important spatial and spectral information.
  We argue that considering the spatial-spectral relationships can improve the extraction of meaningful information from the data cube.
  The **third objective** is to explore tensor decomposition methods that preserve the spatial structure of hyperspectral images, providing a more suitable approach for data with structured spatial information.

This objective is discussed in the third section, which shifts the focus to algorithm development. It presents a tensor decomposition based algorithm to handle data with structured spatial information.

II

ADDRESSING KEY GAPS IN CLUSTERING AND SPECTRAL UNMIXING METHODS

1. Introduction

Characterizing sample composition and visualizing the distribution of its chemical compounds is an important aspect in various research and applied fields. HSI, by integrating spatial and spectral information, plays a pivotal role in this pursuit. The choice of an appropriate data analysis model of hyperspectral images depends on the nature of the chemical dataset. Prior knowledge, combined with mathematical diagnostic tools, aids in identifying the most suitable model and, consequently, selecting the appropriate data analysis method [245]. While research questions and prior knowledge often guide method selection, overlooking the structure of the investigated data, i.e. linearity, geometry, homogeneity, might lead to biased or erroneous results [246]. This section explores the structural characteristics of the data and identifies the most suitable analysis methods based on these characteristics and the research objectives. Specifically, the application domains of K-means clustering and MCR-ALS methods discussed in the state of the art, are evaluated assessing their advantages and limitations in relation to their structural characteristics.

1.1 Clustering and unmixing

Clustering and spectral unmixing are widely used methods in HSI [4]. The choice between these methods depends on the specific objectives of the analysis. On one hand, clustering methods aim at partitioning similar pixels, assuming that the spectral signature measured at one pixel is characteristic of a unique cluster (for those methods that do not follow a probabilistic statistic). On the other hand, spectral unmixing aims to identify individual sources of spectral variation, modeling each pixel as a linear combination of pure spectral components representing unknown sources. Despite their distinct assumptions and objectives, clustering and spectral unmixing can be used together to enhance image analysis or provide complementary results. As an example, clustering has been incorporated as a constraint in MCR-ALS and vertex component analysis (VCA) for analyzing complex samples [247, 248, 249], and they have often been combined also with other multivariate techniques [143, 250], such

as PCA, to assess data homogeneity in terms of chemical composition or properties [251, 252]. It is clear that clustering and spectral unmixing are not interchangeable. However, misconceptions persist, as evidenced by studies that describe using K-means clustering for spectral unmixing [253, 254, 255]:

"...used K-means clustering to unmix individual pixels in two-photon laser scanning microscopy images..." ;

"We investigated clustering-based unsupervised learning in blindly unmixing channels of multi-color two-photon laser scanning microscopy images.";

"...unmixing based on clustering...";

"K-means spectral unmixing for multi-channel imaging and image analysis...".

Given these ambiguities, it is important to clearly define the applicability domain of each method. In this work, K-means clustering and MCR-ALS are chosen as representative techniques of clustering and spectral unmixing, respectively.

1.2 Structure of the data

In this section, the term *structure of the data* refers to the spatial arrangement of data points in a normalized score space. This arrangement can highlight the presence of distinct groups, indicate a more dispersed distribution, or provide insights into data homogeneity. In line with this, Figure 1.1 illustrates three idealized data arrangements. The first structure (Fig. 1.1 A) refers to a typical case where clustering approaches can be used, as the data are clearly grouped and the mean spectrum of each group is the centroid. As these groups become more dispersed in space (Fig. 1.1 B), the data distribution still exhibits a recognizable clustering pattern but does not cover the entire space. In the context of linear mixtures, these groups correspond to specific compositions of the mixture. In such cases, choosing between clustering and spectral unmixing is not straightforward. This scenario lies between the two methods, where both approaches could be considered. Throughout this section, we focus on this particular data structure. Structure C in Figure 1.1 illustrates a typical case where

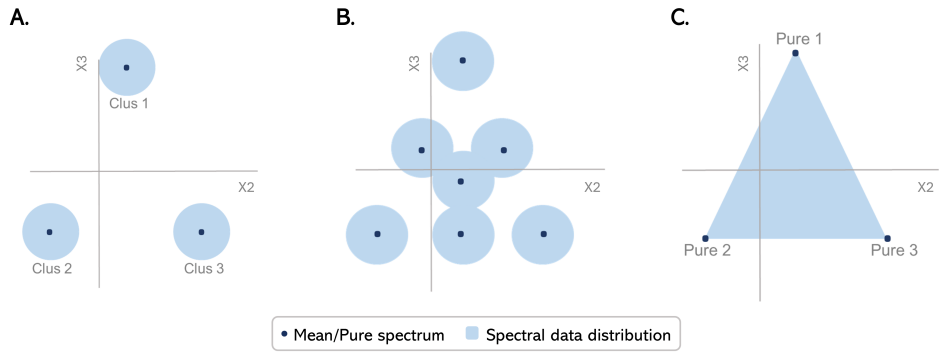

Figure 1.1: Spectral data representation in a normalized space (X2, X3): X2 and X3 are normalized principal components, with X1 normalized to unit value and then excluded from the graphical representation. **A.** The blue dots represent the mean spectra, and the data arrangement (light blue area) in the space corresponds to a typical cluster configuration. In **B.** it is unclear how to interpret the blue data points and the spectral distribution (light blue) in the normalized space. Is this a case of clustering or spectral unmixing? **C.** the blue dots represent pure spectra and a typical mixture distribution (light blue area) it is illustrated.

unmixing methods should be used, as the mixtures of pure compounds (three in this case) span the entire space.

1.3 HSI data: clustering or spectral unmixing?

The type of data as in Fig. 1.1 B can be seen from different perspectives. Let's make this idea clear with a simple simulated sample of three generic compounds A, B, and C, with pure pixels and pure spectral variables for all the three components. The data are built with a bilinear model associating to each spatial distribution a spectral profiles, and adding noise: $\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E}$ (see Fig. 1.2 A). The data arrangement in a score plot resulting from a PCA analysis and normalized to unit first column X1 is represented in Fig. 1.2 B. This problem can be seen from two different perspectives, illustrated in Figure 1.3: clustering (Fig. 1.3 A), where data are grouped in seven clear clusters, and unmixing (Fig. 1.3 B) with the three pure spectra at the vertices of a triangular distribution, and binary and ternary mixtures are also observed. A partitioning method based on distance minimization like K-means performed by setting seven as the number of clusters, with 50 replicates and 200 iterations, resulted in a good estimation of

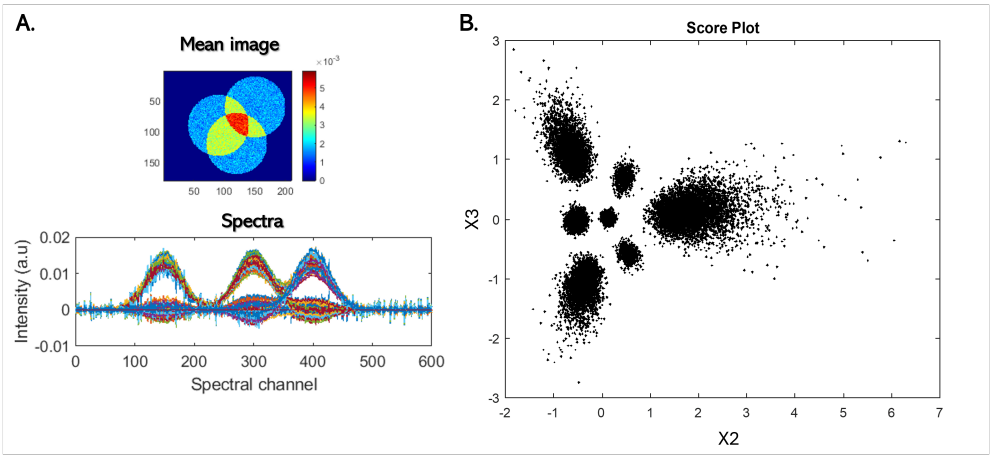

Figure 1.2: Simulation of a three-components dataset: **A.** mean image and spectra; **B.** representation of the scores space (X2,X3) resulted from a normalized PCA, where each data point represent a pixel/spectrum.

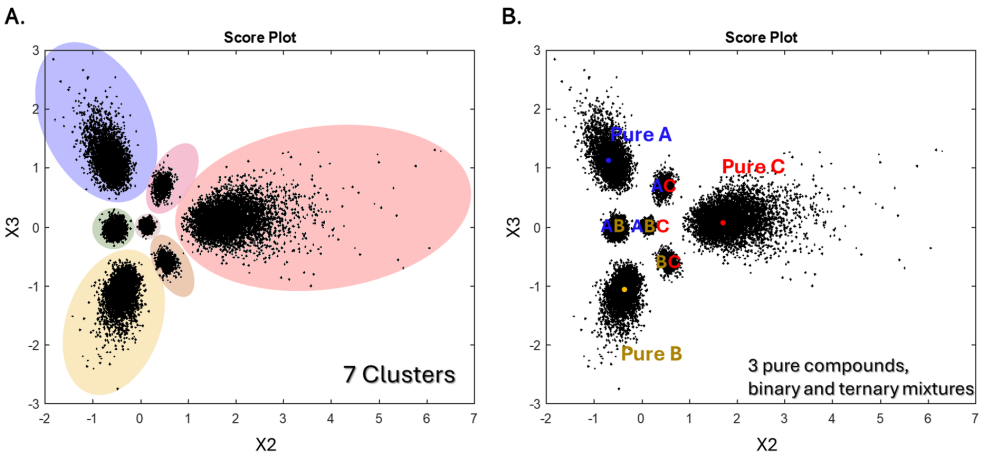

Figure 1.3: The simulated dataset with three components can be seen in two different way: **A.** as a clustering problem, with 7 clear clusters spanning the scores' space, or as in **B.**, spectral unmixing problem, with three pure components A, B, C and some binary and ternary mixtures resulting from a linear combination of the pures are also present.

the spatial contribution (clustering map) (Fig. 1.4 A) and a good estimation of the mean spectrum for each cluster (centroids), identifying the three purest components and the binary and ternary mixtures (Fig. 1.4 B). Figure 1.4 C represents the score plot color coded by cluster membership. Data decomposition methods describing variation in a reduced subspace, like MCR-ALS, computed for three components with non-negativity constraints and SIMPLISMA method for initial estimation, resulted in a good estimation of the spatial distributions as well as an optimal optimization of the spectral profiles. The spatial distribution of each component contains a small part of the other two (see Fig. 1.5), which may be related to dataset noise.

It is clear then, that both the approaches can be used, depending on the research question. In this case they also give comparable results as the presence of pure pixels allows K-means to group the purest three species present.

But what if gradually this selectivity is not a property of the data anymore (Fig. 1.6 A)? In this case, moving to a data arrangement more spread in the score space (see Fig. 1.6 B), a clustering approach, as K-means, will not be able to retrieve the pure components, as it can not unmix (Fig. 1.7), while an unmixing approach will give optimal results. MCR-ALS shows good estimations of spatial distributions and spectral profiles (see Fig. 1.8). MCR-ALS resulted depended on the type and entity of noise present. K-means analysis was not significant in simulations with no selective spectral channels, and the presence of mixed pixels poses a limitation, as it may lead to incorrect data partitioning.

Referring to data that underline the bilinear model, clustering should be applied when imaging data are characterized by the presence of pure pixels distributed all over the imaged field of view; unmixing should be utilized when the image pixels under study are underlined by mixtures of pure spectroscopic fingerprints. Considering the specific problem under study, the application of both methods is only meaningful in certain determined situations.

A different case arises when selectivity is not a property of both spectral variables and pixels. For such cases, we refer to Chapter 4, where an example is introduced. As we have seen, the methodology that can be applied to these types of data is not straightforward and there is not a single method that can deal with all the different cases , which opens up further considerations.

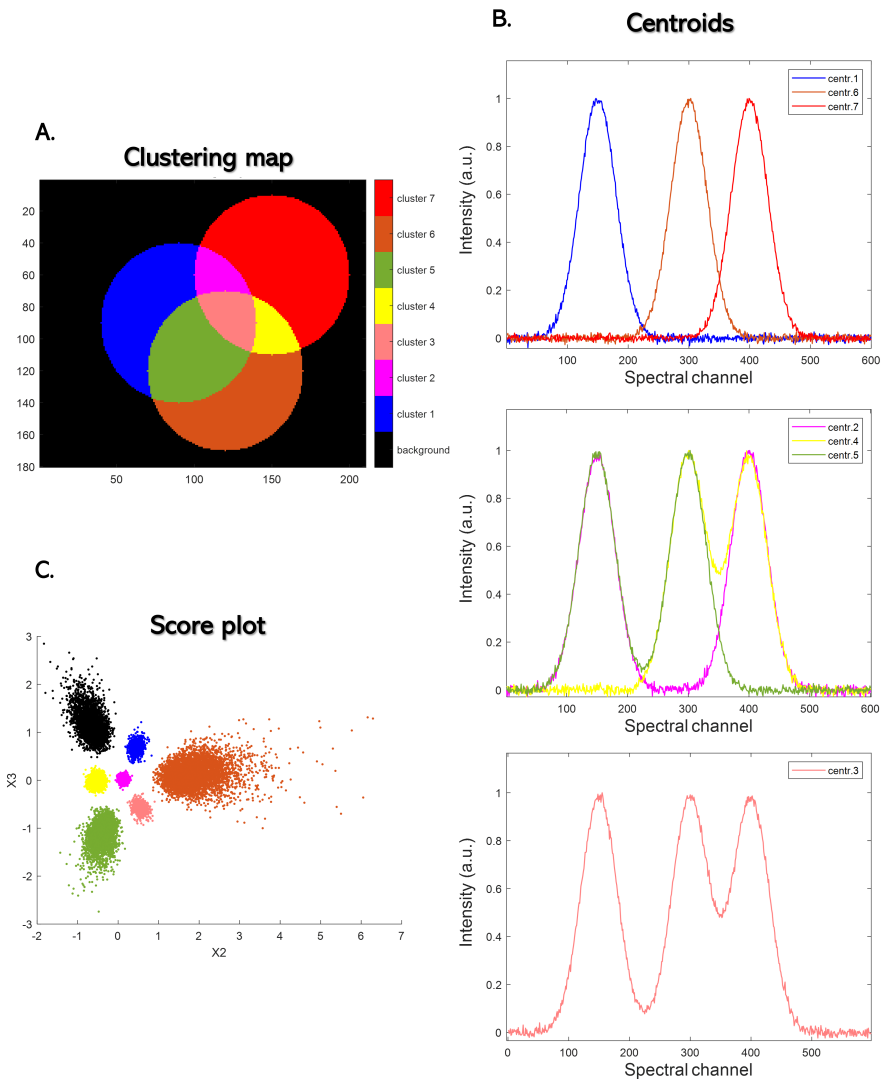

Figure 1.4: K-means results for the simulated dataset. The number of cluster was set to seven. **A.** shown the cluster membership map. In **B.** centroids are shown for the three purest component, the binary and ternary mixtures. C. display the normalized score plot color coded by cluster membership

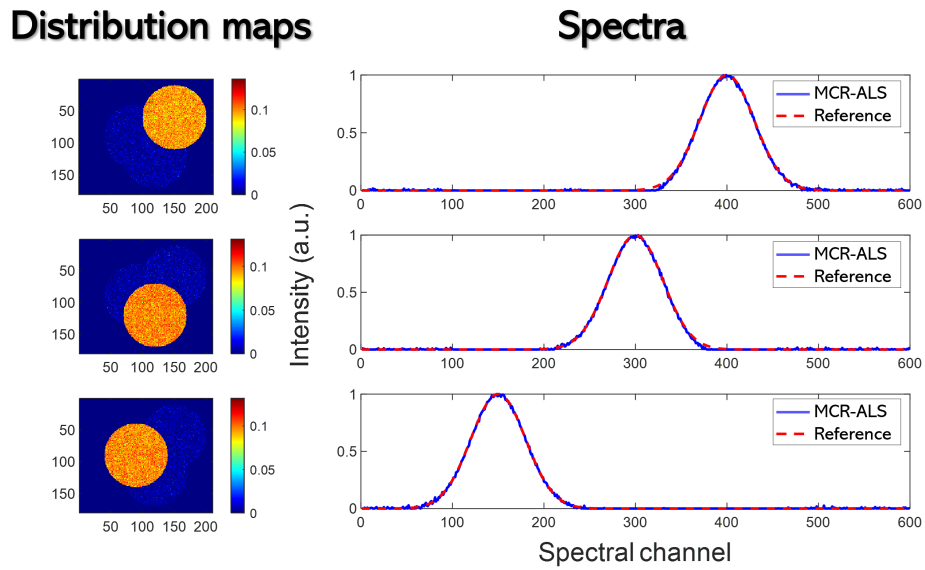

Figure 1.5: MCR-ALS results for the simulated dataset. An unmixing model of three components and constrained with non-negativity resulted in very comparable distribution profiles of the pure. Pure spectra are used as references and are plotted in red, while the resolved MCR-ALS spectra are in blue. SIMPLISMA was used for the initial estimation of the purest spectral profiles. Results after four iterations are provided. The algorithm converged with a threshold fixed at 0.1%. The model has LOF=13% and r^2=98%.

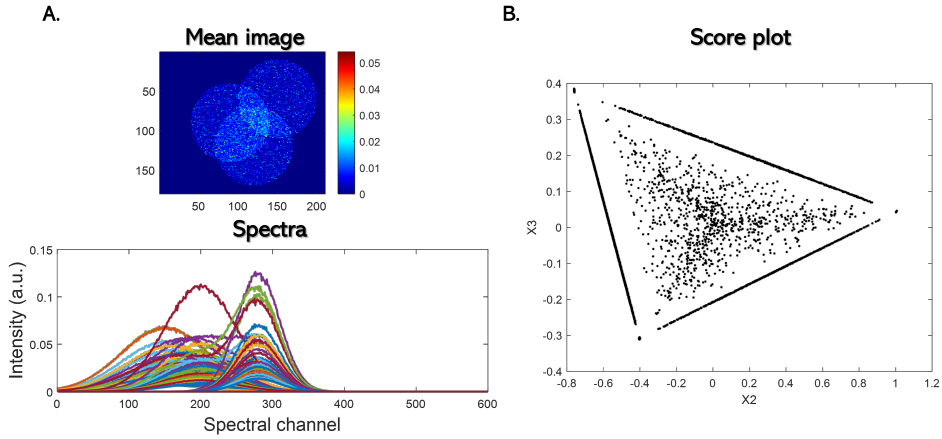

Figure 1.6: Simulation of a three-components dataset with no selective spectral channels: **A.** mean image and spectra; **B.** representation of the scores space (X2,X3) resulted from a normalized PCA, where each data point represent a pixel/spectrum.

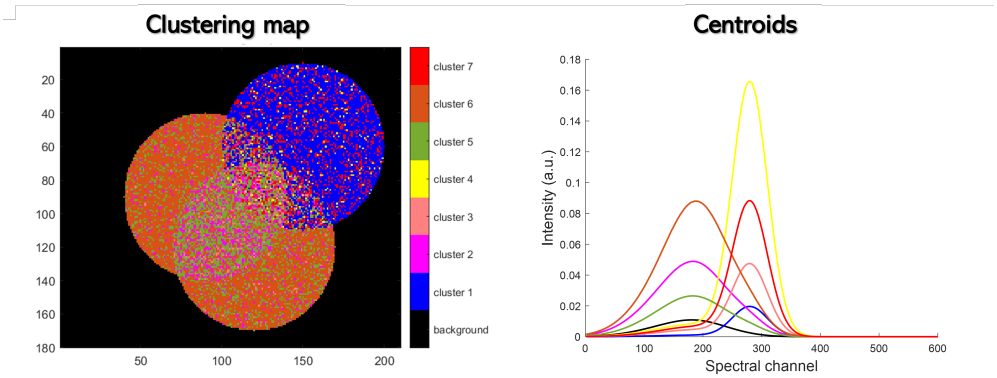

Figure 1.7: K-means results for the simulated dataset with no spectral selectivity. The number of cluster was set to seven. **A.** cluster membership map; **B.** centroids for the seven cluster identified.

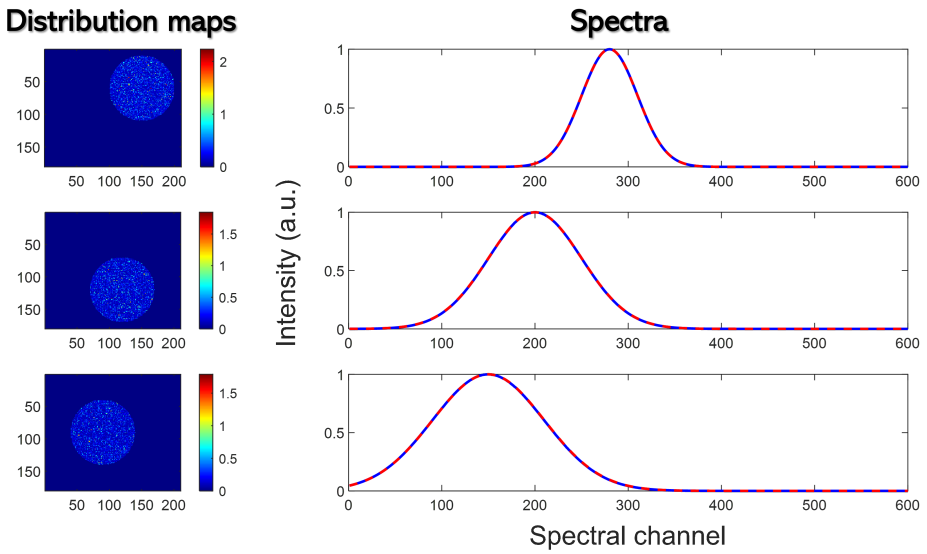

Figure 1.8: MCR-ALS results for the simulated dataset. An unmixing model of three components and constrained with non-negativity resulted in good optimization of the distribution and spectral profiles of the three components. Pure spectra are used as references and are plotted in red, while the resolved MCR-ALS spectra are in blue. SIMPLISMA was used for the initial estimation of the purest spectral profiles. Results after five iterations are provided. The algorithm converged with a threshold fixed at 0.1%. The model has LOF=73,24% and r^2=46,37%.

To address this issue, extracting essential information (EI) can be highly useful, as it relies on the geometric structure of the data cloud rather than on data variance [105, 106, 107, 199, 256, 257]. Essential information consists of archetype points that outline the convex hull of the data points cloud in a normalized abstract data space. Recent studies have highlighted the potential and usefulness of identifying essential rows and columns of a data matrix [193, 194, 258, 259]. A key aspect is that the corresponding samples (spectral pixels) and variables (single-wavelength images) contain all the information needed to reproduce the measured data [260].

For interpretations, users should be aware of the specific characteristics of the structure of the data and carefully consider the assumptions and limitations of the methods to ensure the reliability of their interpretation. A recent study investigating data point importance (DPI) by assessing the volume of data structure revealed that some data points play a more significant role in the preservation of data pattern [258]. Points located within the convex hull exert no influence on altering the overall data structure, resulting in zero change in the calculated volume. In contrast, the vertices of the convex hull show substantial changes in relative hypervolume, indicating a more pronounced impact from some data points. Consequently, crucial data points can be classified according to the strength of their impact signals. In the variable space domain, extreme concentration ratios have special significance. In particular, among the mixture spectra, it is the spectra with extreme concentration ratios that are considered indispensable. The study also shows that the outermost pixels and variables contain all the information needed to describe the data arrangements in the abstract space [246]. Considering all this, a novel approach for the analysis of the geometry of the data cloud is introduced in the next chapter.

2. Methods

In this chapter, an exploratory data analysis approach, based on the geometry of the data point cloud resulting from a singular value decomposition (SVD), is proposed to investigate the structure of HSI datasets and extract their main characteristics. The principle of essential information is used to extract archetype (most linearly dissimilar) spectra and archetype single-wavelength images. This chapter includes the work "Exploratory analysis of hyperspectral imaging data" [246], published as part of this PhD research, with most figures and content drawn from it.

2.1 Selection of the most relevant archetype points for exploratory analysis

The HSI data cube is first unfolded into a matrix $\mathbf{D}$, of dimensions (n, p), with rows corresponding to pixels and columns corresponding to spectral channels. For the sake of clarity, each pixel corresponds to a spectrum, and each spectral channel to an unfolded single-wavelength image. The matrix $\mathbf{D}$ is then decomposed by SVD [261] according to Eq. 2.1:

$$\mathbf{D} = \mathbf{USV}^T + \mathbf{E} \tag{2.1}$$

where $\mathbf{U}$ of dimensions (n, k) is the matrix containing the left singular vectors, $\mathbf{S}$ of dimensions (k, k) is the diagonal matrix of singular values, and $\mathbf{V}^T$ of dimensions (k, p) is the matrix of the right singular vectors transposed. k is the number of factors of the decomposition, and $\mathbf{E}$ of dimensions (n, p) is the error matrix. The matrices $\mathbf{X}$ and $\mathbf{Y}$ of dimensions (n, k) and (p, k), respectively, are calculated as in Eqs. 2.2 and 2.3, containing the coordinates of the data points in the column- and row-vector space:

$$\mathbf{X} = \mathbf{U} \times \mathbf{S} \tag{2.2}$$

$$\mathbf{Y} = \mathbf{V} \times \mathbf{S} \tag{2.3}$$

All column vectors of $\mathbf{X}$ (resp. $\mathbf{Y}$) are then normalized to constant projection on the first column vector of $\mathbf{X}$ (resp. $\mathbf{Y}$) to enforce convexity of the data points cloud [262, 263], as in Eqs. 2.4 and 2.5:

$$\mathbf{X} = \mathbf{X} \oslash \mathbf{x_1}\mathbf{1}^T, \tag{2.4}$$

$$\mathbf{Y} = \mathbf{Y} \oslash \mathbf{y_1}\mathbf{1}^T, \tag{2.5}$$

where for the sake of simplicity the notation of $\mathbf{X}$ (resp. $\mathbf{Y}$) is left unchanged. $\mathbf{x_1}$ (resp. $\mathbf{y_1}$) denotes the first column vector of $\mathbf{X}$ (resp. $\mathbf{Y}$), the operator $\oslash$ is the Hadamard division (element-wise) and $\mathbf{1}$ is a column vector of ones of length K. The archetypes of the data points cloud of $\mathbf{X}$ and $\mathbf{Y}$ can be identified by computing the corresponding convex hulls, as in the equations 2.6 and 2.7 [194]:

$$conv(\mathbf{X}) = \left\{ \mathbf{x} \in \mathbf{X} \;\middle|\; \sum \alpha\mathbf{x}; \alpha \geq 0 \text{ and } \sum \alpha = 1 \right\}, \tag{2.6}$$

$$conv(\mathbf{Y}) = \left\{ \mathbf{y} \in \mathbf{Y} \;\middle|\; \sum \beta\mathbf{y}; \beta \geq 0 \text{ and } \sum \beta = 1 \right\}. \tag{2.7}$$

where α and β are coefficients of the convex linear combinations. They correspond to the most linearly dissimilar spectral pixels and single-wavelength images, respectively. The number of components to consider in the convex hull calculation is left to the user [262, 263]. Analogous to exploratory PCA, the inspection of the information carried by the most dissimilar spectra/images can guide the selection.

The most relevant archetype points are then selected by visual inspection, and the corresponding (essential) spectra and (essential) single-wavelength images are extracted, as illustrated in Fig. 2.1.

2.2 MCR-ALS

The MCR-ALS algorithm provides pure spectral signature and distribution map for each component of the data set. ALS [183] was selected for model optimization, with both concentration and spectral profiles constrained by non-negativity. The SIMPLISMA

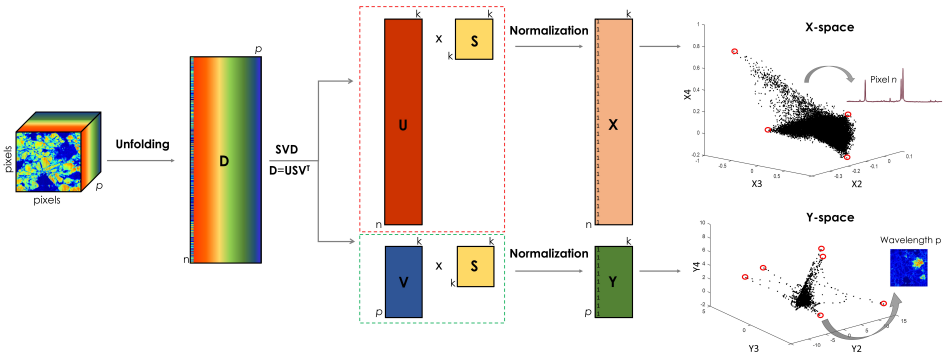

Figure 2.1: Graphical representation of the data exploratory approach: the hyperspectral data cube is unfolded into a matrix $\mathbf{D}$ and decomposed through the SVD algorithm ($\mathbf{D} = \mathbf{USV}^T$). Matrices $\mathbf{X}$ and $\mathbf{Y}$ are calculated and normalized, resulting in a unit first column vector $\mathbf{X1}$ (resp. $\mathbf{Y1}$) to which all other column vectors of $\mathbf{X}$ (resp. $\mathbf{Y}$) are orthogonal. Convex hulls of essential spectra and essential variables are computed for $\mathbf{X}$ and $\mathbf{Y}$, and the most relevant archetype points are identified by visual inspection (red circles).

method [190] was used throughout this study to compute the initial spectral estimates. The convergence criterion threshold was set to 0.1%. The LOF and explained variance, defined in Eqs. 2.8 and 2.9, were used to assess the quality of the MCR model:

$$LOF = 100 \times \sqrt{\frac{\sum e^2}{\sum d^2}}, \tag{2.8}$$

$$r^2 = 100 \times \left(1 - \frac{\sum e^2}{\sum d^2}\right), \tag{2.9}$$

where e and d are elements of $\mathbf{D}$ and $\mathbf{E}$, respectively.

2.3 K-means clustering

The K-means algorithm was applied to $\mathbf{D}$ using the Pearson correlation distance to find c clusters [132]. To stabilize the results, 50 replicate runs of K-means clustering were performed for each analysis, selecting the run with the lowest sum of point-to-centroid distances, summed over all c clusters (*sumd*). The number of iterations was set to 200. Silhouette [159] and PBM [160] indices were used to evaluate the optimal number of clusters. These were compared with the number of most informative pixels/spectral wavelengths suggested by the proposed exploratory approach. For

consistency, the parameter c is used to denote both the number of components in MCR-ALS and clusters in K-means, as the results are presented considering the same number of components and clusters.

3. Datasets and software

Two datasets with varying characteristics and complexities are introduced and described in this chapter: a Raman-HSI powder mixture, collected within our research group, and a LIBS-HSI mineral sample, kindly provided through a collaboration with the Université de Lyon. The first dataset serves as an ideal case for applying unmixing techniques, while the second dataset allows for both unmixing and clustering approaches, depending on the analysis objective. However, the LIBS dataset presents an additional challenge due to the presence of pyrite weathering products and the lack of selectivity, leading to deviations from the ideal models underlying both MCR and clustering methods [264].

3.1 Raman powder dataset

Powders of three salts, i.e., calcium carbonate ($CaCO_3$), sodium nitrate ($NaNO_3$), and sodium sulfate (Na_2SO_4), were mixed and pressed into a tablet, obtaining a three-component system. Sample preparation and Raman imaging acquisition features were described by Coic et al. in [199]. The sample was investigated in the range 901.2 cm^{-1} to 1280.5 cm^{-1} with a spectral resolution of 1.11 cm^{-1}. A 101 $\times$ 101 pixels image was mapped using a point-by-point raster-scanning mode with a 1 μm step between successive acquisitions. The dataset corresponds to an hyperspectral image of dimensions 101 $\times$ 101 $\times$ 341, which was subsequently analyzed without any spectral pretreatment. The three-component Raman-HSI dataset (Fig. 3.1), exhibits well-defined characteristics [246].:

- clear distribution of the salts building up the chemical composition of the sample (Fig. 3.1 A);
- good signal-to-noise ratio data (Fig. 3.1 B);
- selective spectral regions (Fig. 3.1 C).

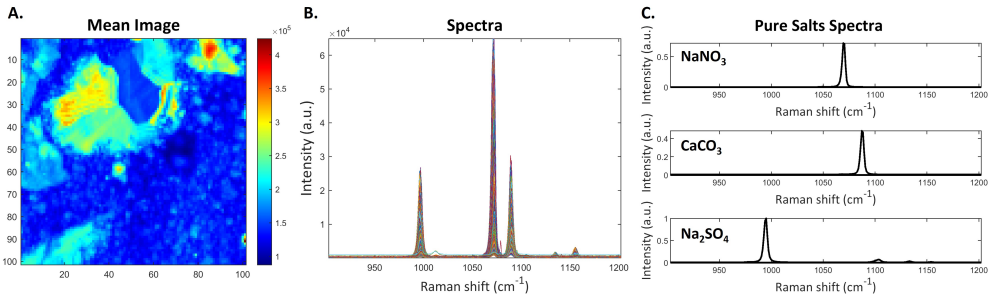

Figure 3.1: Raman powder dataset: **A.** Mean image **B.** spectra and **B.** spectral profiles of the three pure salts.

3.2 LIBS mineral dataset

A thin section of a mineral sample from the Nishapur turquoise deposit (Iran) was prepared and polished for LIBS imaging. Sample preparation, equipment, and LIBS acquisition are detailed in Moncayo et al. in [265]. The sample consists of three main mineral phases: pyrite FeS_2, silica (mainly quartz) SiO_2, and turquoise $CuAl_6(PO_4)_4(OH)_8{\cdot}4H_2O$. The LIBS image was recorded considering a 15 μm step between successive acquisitions over 2048 spectral channels in the spectral range from 250 to 330 nm. From the full acquired dataset, only a region of interest has been selected (Fig. 3.2), resulting in a data cube of dimensions 300 $\times$ 300 $\times$ 1930, which was then analyzed without spectral preprocessing. In Fig. 3.2 A it is important to note that, considering the scale of mineral phases and the pixel size of the image (15 μm), the presence of many pure spectral pixels is, therefore, not expected. Fig. 3.2 B highlights data characterized by low spectral selectivity [246].

3.3 Software

All calculations were performed with MATLAB© 2022a (MathWorks Massachusetts, USA). For the cluster analyses the K-means function of the Statistical and Machine Learning Toolbox was used, with the addition of the MATLAB© Parallel Computing Toolbox to improve the speed of the algorithm. *Pure* and *als* routines from Tauler and De Juan (2003) were used for the MCR-ALS analysis and *convhulln* is the built in

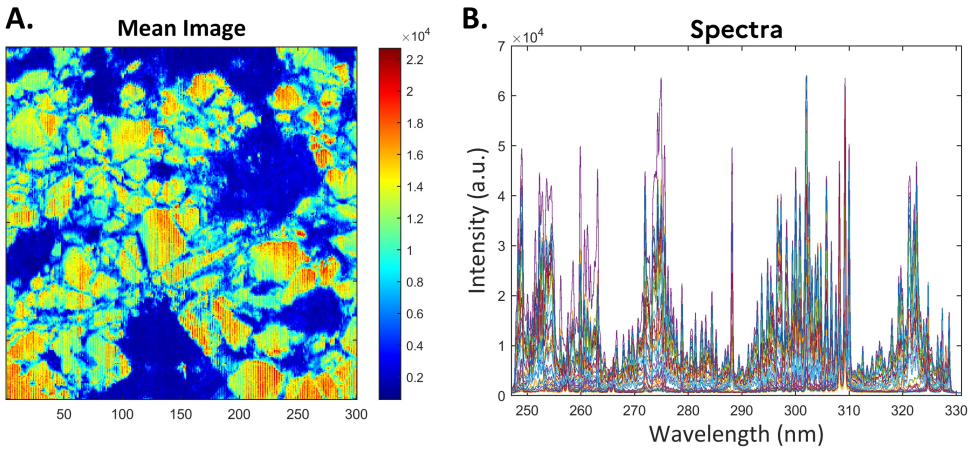

Figure 3.2: LIBS mineral dataset: **A.** Mean image and **B.** overlapped spectra.

MATLAB© function used to compute convex hulls.

4. Exploratory analysis of hyperspectral images

MCR-ALS and K-means clustering are widely used methods for HSI data analysis, with their specific applicability domains, as shown in the introduction chapter of this section 1). However, their effectiveness in complex scenarios, where the structure of the data deviates from the models' assumptions, requires further investigation. To address this, the proposed exploratory data analysis approach was applied to the previously introduced datasets, the Raman powder dataset and the LIBS mineral sample. In this chapter, results are presented and organized by datasets. The information extracted from the exploratory analysis is put in perspective with the results obtained by applying both MCR-ALS and K-means. This chapter includes the work "Exploratory analysis of hyperspectral imaging data" [246], published as part of the PhD research, with most figures and content drawn from it.

4.1 Raman powder dataset

A representation of the (X2, X3) data points clouds is shown in Fig. 4.1 A, where the number 2 and 3 refer to the second and third column of the normalized matrix $\mathbf{X}$. As expected, the observed data structure corresponds to a triangular geometry (as would be obtained for simplex data), with the three vertices being expected to correspond to the pure compounds [263]. Similarly, the data points representation of the second and third column of the normalized matrix $\mathbf{Y}$, (Y2, Y3), (Fig. 4.2 A), enables the identification of vertices pointing in clearly distinct directions. Convex-hull computation provided 14 archetypes points in the (X2, X3) space corresponding to essential spectra and three archetype points in the (Y2, Y3) space corresponding to essential single-wavelength images (black circles in Figs. 4.1 A and 4.2 A). Considering that the number of components is known, three archetype points were selected in both sub-spaces (filled green circles in Figs. 4.1A and 4.2 A, respectively), which are expected

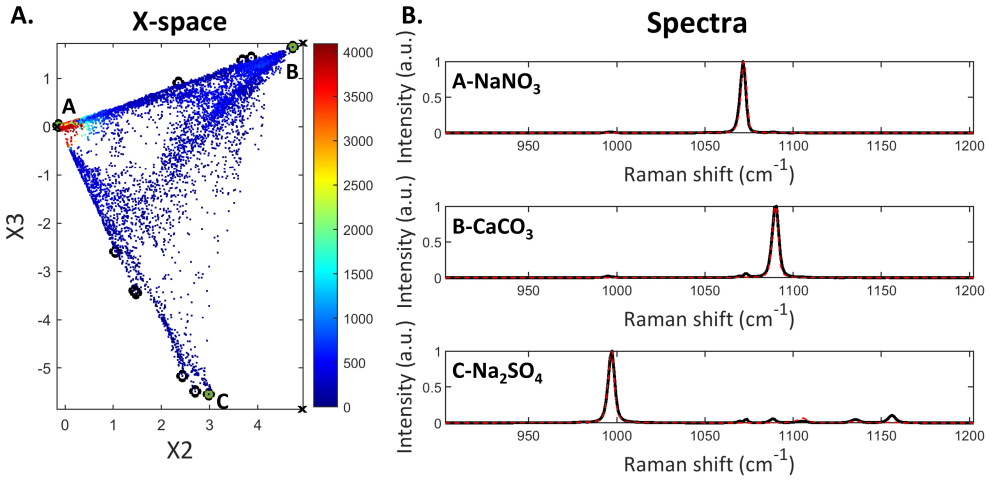

Figure 4.1: Raman powder dataset: **A.** 2-D representation of the X-space color-coded by point density. Black circles mark the archetypes points (some are close and result overlapped in the plot) at the vertices of the convex hull computed in the (X2, X3) normalized space. Filled green circles are the selected points and black crosses are the projection of the pure reference spectra in the (X2, X3) normalized space. **B.** spectra corresponding to the green points (black line) with overlapped the pure spectrum (red line) of the corresponding component.

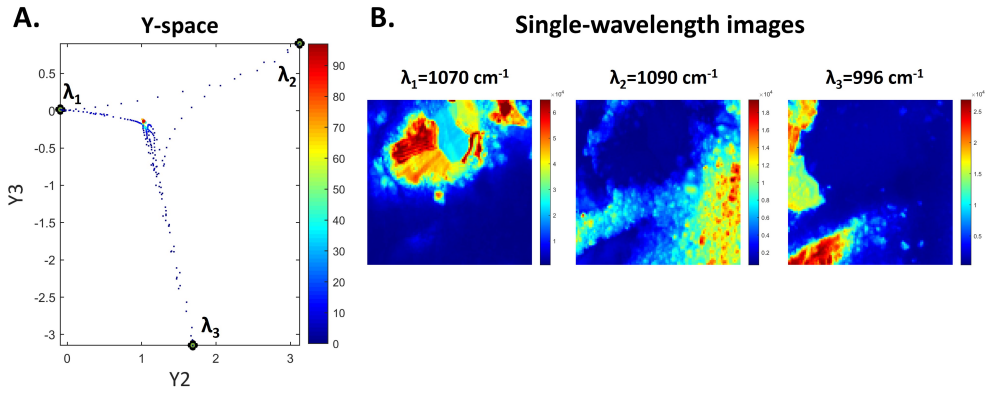

Figure 4.2: Raman powder dataset: **A.** 2-D representation of the Y-space color-coded by point density. Black circles mark the archetypes points of the convex hull computed in the (Y2, Y3) normalized space. Filled green circles are the points identified looking at the structure of the data. **B.** the essential single-wavelength images corresponding to the three identified selective wavelengths.

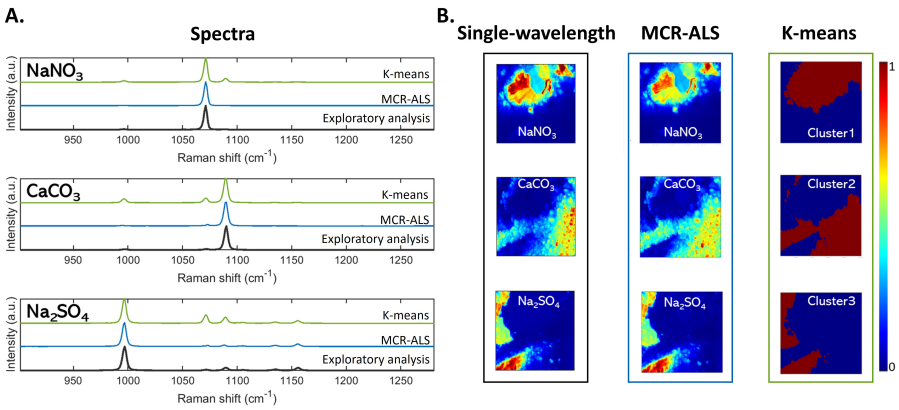

Figure 4.3: Raman powder dataset: **A.** shows the spectra for the purest pixels obtained by the exploratory analysis (black line), the purest components resolved spectra by MCR-ALS (blue) and the K-means centroids spectra (green). Centroids spectra are calculated as the average of the spectra of all the pixels belonging to a given cluster. For sake of clarity an arbitrary vertical offset was added to the MCR-ALS and K-means results. **B.** shows the single-wavelength images extracted with the exploratory approach, the concentration distribution maps retrieved by MCR-ALS and the clustering maps obtained by K-means clustering.

to correspond to the purest spectral pixels and most selective wavelengths measured (see Figs. 4.1 B and 4.2 B, respectively). The provided spectral and image information can be readily interpreted for this simple data set (1070 cm^{-1} maximum selective peak for $NaNO_3$, 1090 cm^{-1} maximum selective peak for $CaCO_3$, 996 cm^{-1} maximum selective peak for Na_2SO_4). For the sake of comparison, the results obtained by SIMPLISMA are provided (Fig. A.1 in Appendix A). Figure 4.3 shows the results obtained for a three-component MCR-ALS model (LOF=10%, r^2=99%) and for the application of K-means considering three clusters. The selection of the number of clusters was set as three according to the mixture composition. For each cluster, the class assignation vector has been refolded in the original image dimensions and shown with the pixels recognized as cluster members colored in brown. For the sake of comparison, the results obtained from the previous archetype identification are also reported. The similarity between the essential spectra and essential single-wavelength images obtained from our approach and the spectra and component distribution maps obtained by applying MCR-ALS is striking. Focusing on the spectra provided in Fig. 4.3 A, the ones shown for K-means correspond to “centroid” spectra and are, as expected, not the pure ones, though in quite good agreement. It is worth noting that the centroid

spectrum corresponding to the $NaNO_3$ salt is more similar to the pure one than for the 2 other salts. This can be explained by considering the density of points for each of the 3 clusters modeled by K-means (see Fig. A.2 in Appendix A). As for K-means, the maps (Fig. 4.3 B) obtained for each cluster are also very comparable (considering that the information is segmented). This dataset was introduced to clearly show that in cases in which we have prior information on the number of components, high spectral and spatial selectivity, as well as a high number of pure pixels, MCR-ALS and K-means solutions are very comparable, with selection of the method depending on specific analysis goals. Also, the information retrieved with the two approaches can be readily extracted from the analysis of the geometry of the data.
We now extend the exploratory analysis to a more complex scenario.

4.2 LIBS mineral dataset

This dataset is characterized by low spectral selectivity, as well as the presence of many pure spectral pixels is not expected, because in the pixel of the image (size: 15 μm), different mineral phases are expected to be found. An additional complexity of this sample arises from its composition, which includes iron. Iron has numerous emission lines across the entire spectral range. Additionally, pyrite typically exists in various oxidative forms [264, 265, 266] and the iron ions within pyrite can easily exchange with copper or aluminum ions present in turquoise [267]. In fact, this kind of rocks are often referred to as “solid mixtures” [268]. Furthermore, within quartz, the predominant silica phase in this sample, aluminum impurities are quite common, while iron inclusions are also possible, albeit less frequent [269]. All these peculiarities translate into a very challenging LIBS HSI dataset to analyze and investigate with classical chemometric tools. Indeed, this scenario is not ideal for approaches such as MCR-ALS as pure pixels may not be present, spectral selectivity is low, and different phases with very similar spatial distribution are present. Similarly, K-means clustering is not ideal as it may have difficulty assigning different minority phases to distinct clusters, as pixels may belong to multiple clusters due to low spectral selectivity. The geometry of the data in the (X2, X3) and (Y2, Y3) spaces is illustrated in Figs. 4.4 A and 4.4 B, respectively. While more complex than the geometry observed in the

previous example, the observed data points clouds exhibit some degree of structure. However, determining the appropriate number of components to consider is not straightforward given the absence of clear a priori information with this dataset. Convex-hull computation provided 19 archetype points in (X2, X3), and 13 archetype points in (Y2, Y3), (black circles in Figs. 4.4 A and 4.4 B, respectively). Considering the geometry of the data observed in Fig. 4.4 A, six archetype points were selected (filled green circles) and the corresponding essential spectra are shown. However, considering Fig. 4.4 B, it is clear that some relevant points, corresponding to clear directions, were not identified as archetypes, as they are not found at vertices of the data points cloud in the two-dimensional Y-space. It should be noted that by applying convex hull calculation to a six-dimensional $\mathbf{Y}$ matrix (see Appendix A Fig. A.3), these points could be selected, but the total number of archetypes would be very large. This is not really needed, though, since they can be manually pointed out in the (Y2, Y3) plot, resulting in the extraction of seven essential single-variable images. The spectra corresponding to points labeled A, C, and D in Fig. 4.4 A correspond to the main mineral phases of pyrite, silica, and turquoise, respectively. Spectrum B shows spectral features corresponding to a phase where silica has iron inclusions, somehow in between the pyrite and the main phase of silica. Spectrum F features another pyrite phase, different from the one observed in A. Lastly, the spectrum corresponding to pixel E characterizes an intermediate phase between turquoise and pyrite, where iron and mainly aluminum exchanges occur. The spectral regions used for the identification are highlighted in blue, referring to Kurucz LIBS database [270], following the assessed procedure of Moncayo et al. [265]. These spectra are the purest spectra identified and can be interpreted as such without further analysis of the data. In the same way, the essential single-wavelength images extracted correspond to the information obtained at the most selective wavelengths. Images λ_1, λ_3, and λ_7, which are linked to point A, C, and D in Fig. 4.4 B, respectively, describe pyrite, silica, and turquoise. Image λ_2, corresponding to point B, describes a situation where both pyrite and silica are present, and image λ_4 shows the distribution of pyrite, turquoise, and silica. It is worth noting that when looking at image λ_5, which does not show any correspondence in the (X2, X3) plot, it could be hypothesized that it represents a mineral phase where both silica and turquoise are present. In fact, it lies between

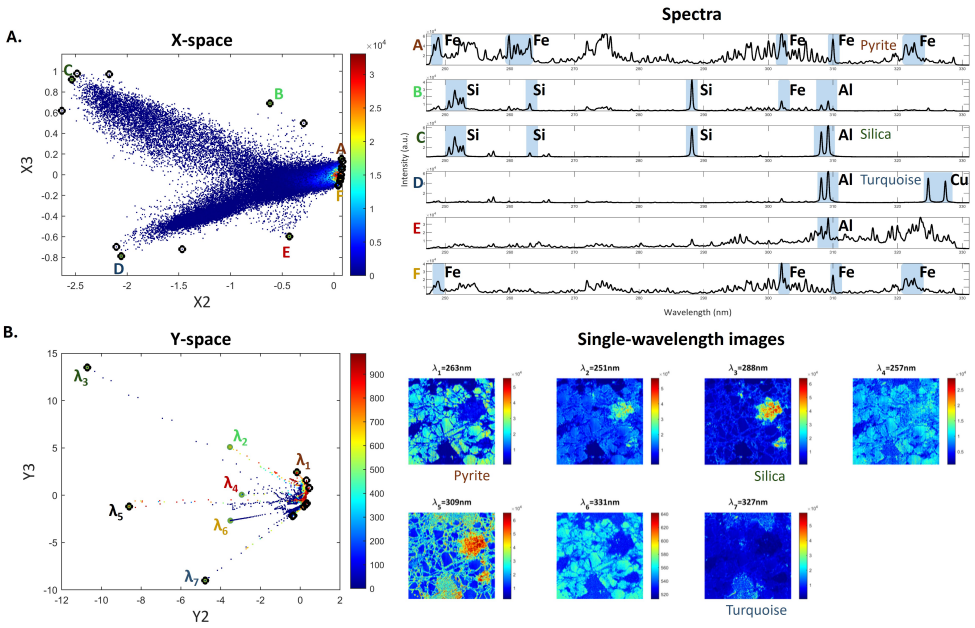

Figure 4.4: LIBS mineral dataset: **A.** 2-D representation of the X-space of the mineral sample dataset, color-coded by point density. Black circles mark the archetypes points at the vertices of the convex hull computed in the (X_2, X_3) normalized space. Letters and filled green circles represent the selected points, while the corresponding spectra are shown in the right panel. **B.** 2-D representation of the Y-space, color-coded by point density. Black circles mark the archetypes points of the convex hull computed in the (Y_2, Y_3) normalized space. Filled green circles represent the selected wavelengths, the corresponding refolded images are shown in the right panel

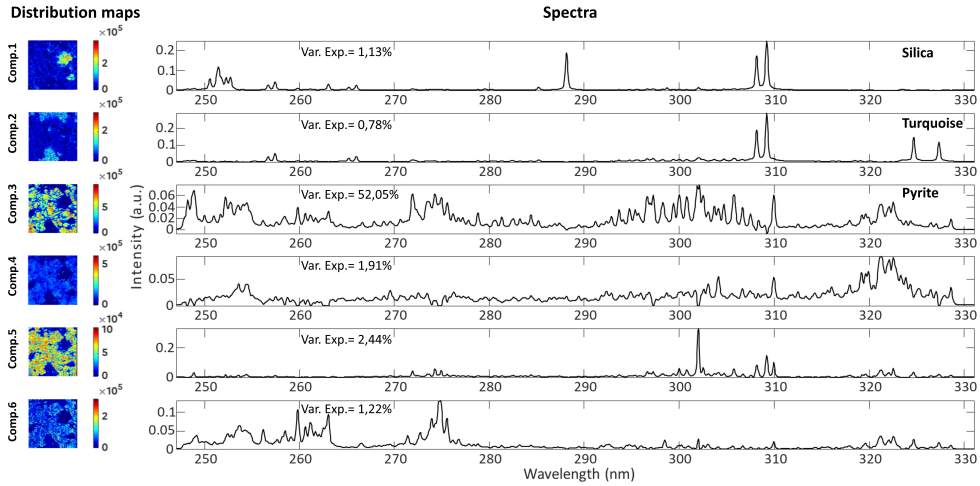

Figure 4.5: LIBS mineral dataset: MCR-ALS solutions. Refolded concentration profiles (left) and resolved spectra (right) are shown for each of the 6 components with the corresponding data variance.

image λ_3 and λ_7 in the (Y2, Y3) plot. Image λ_6, linked to point F, is identified as another form of pyrite. In addition, it can be noticed that in the right area of both the (X2, X3) and (Y2, Y3) plots, there is a higher density of points (either pixels or spectral wavelengths). Since points A and F correspond to spectra that are associated with pyrite phases, it can be concluded that pyrite is identified as the major phase in this mineral sample. For comparison purposes, the results obtained by SIMPLISMA are also provided (Fig. A.4 in Appendix A). A six-component MCR-ALS model could then be fitted (LOF=3%, r^2=99%) and the results are shown in Fig. 4.5. The spectra of the first 2 components of the MCR-ALS model are identified as silica and turquoise phases, respectively. The corresponding concentration distributions are in agreement with the images retrieved by the exploratory analysis. Pyrite is identified primarily in the spectral profile of the third component. However, the spectral profiles observed in the remaining 3 components suggest the potential occurrence of different pyrite phases characterized by ion exchanges, which present challenges for interpretation. This is further complicated by the fact that the corresponding concentration distribution maps show very similar distributions. The third MCR component is the one explaining most of the data variance (52%) confirming that pyrite is the major phase. By contrast the variance explained by other components, is very low, less than 3% for silica,

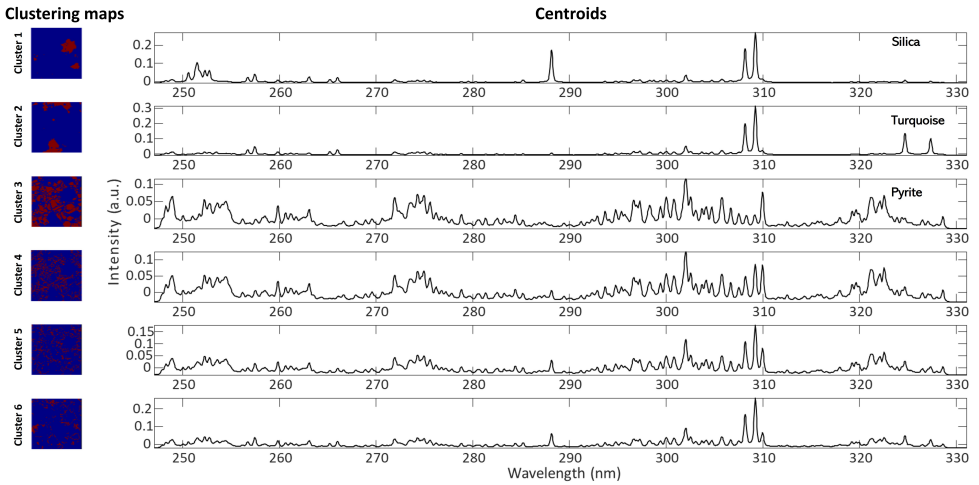

Figure 4.6: LIBS mineral dataset: K-means results for the mineral sample dataset. Cluster membership maps (left) and the mean spectra (right) are shown for each of the six clusters.

turquoise and the other phases of pyrite. For the sake of comparison, a K-means model was computed setting the number of clusters to 6. The results are shown in Fig. 4.6. Clusters 1, 2, and 3 can be associated with silica, turquoise, and pyrite phases, respectively. The clustering maps for clusters 4 to 6 reveal distributions spanning the boundaries between pyrite and the phases described by the first two clusters. The centroid spectra of these clusters are challenging to interpret, suggesting possible exchanges between iron and aluminum. MCR-ALS and K-means clustering provide complementary information that leads to a more complete understanding of the sample. The proposed methodology allows for observing the potential complexity of data exploration prior to implementing MCR and/or K-means. It is important to note that the exploratory approach not only provides the same information as the one obtained from data modeling, but also enables to extract the spectral and spatial features related to the presence of minority components resulting from ion exchanges between the main mineral phases. The results obtained for silica, turquoise, and pyrite are comparable. The centroid spectra obtained by K-means and identified as pyrite is very comparable with the one extracted exploring the X-space and MCR-ALS, again because of the high number of pure pixels in that cluster (being pyrite the major phase, high number of pixels correspond only to pyrite). The concentration

maps of three of the MCR-ALS components and the clustering maps of three of the clusters show the same distribution observed in the purest images extracted from the Y-space, while the other differ and as discussed above, are not easily interpretable. Overall, we may remark that in this challenging scenario, that deviates from the ideal model underlying both MCR and clustering techniques, exploratory analysis driven by archetypes identification can provide insight into the number of components (when going for an unmixing approach) or clusters (when using clustering) to select. In fact, traditional methods such as eigenvalues, scree plots, and cluster indices may not provide unambiguous answers, as illustrated in Appendix A, Fig. A.5. The exploratory approach employed in this study offers notable advantages, particularly in the extraction of spectra and images without the need for complex modelling. Also, convex hulls need to be calculated for more than two components, to retrieve the archetype points for each direction in the Y-space. These findings emphasize the feasibility and efficiency of our methodology in obtaining informative data without excessive computational load.

5. Final considerations and perspectives

This chapter provides a summary of key considerations after the presentation of the proposed approach and the results of its application.

5.1 Some considerations

The exploratory approach has been shown to be able to guide the extraction of useful information encrypted in the spectral image of complex samples and to provide a comprehensive understanding of the investigated system. The shared information, among K-means, MCR-ALS, and the proposed exploratory approach, in terms of distribution and spectral signature of retrieved common components, concerned major phases and/or the one with selective spectral profile. In cases where the application of very well-known methodologies revealed its limits, observing the geometry of the data resulted in an extremely easy and fast way to have better and more complete insights, with respect to the MCR-ALS and/or K-means ones. The analysis of the structure of the data could be considered, as any exploratory tool, as preliminary to allow a more rational choice of the next steps of data analysis. Additionally, it helps address limitations of the two methods, such as deviations from the linear model and lack of selectivity. Moreover, it aids in deciding the appropriate number of components and clusters, as well as in retrieving and identifying the purest species and minor components.

Other datasets with severe spatial and spectral overlap were investigated, demonstrating very good results in retrieving the most distinct component of the sample. An interesting case concerns the semen stain on cotton dataset, which is shown in Appendix A (Figs. A.6, A.7, A.8).

This exploratory approach may have limitations when the data present a quite uniform distribution with no clear structures, thus rendering difficult finding the archetype points. However, to the best of our knowledge, applying appropriate spectral preprocessing could remove those effects, such as baseline, scatter, etc., that go into making the data less geometrically structured. In this way, a change in the “data shape” can be

obtained making this approach therefore applicable. Furthermore, while automation of this process could be considered, it bears the risk of yielding inaccurate results, as extreme points may also include noise points requiring visual inspection before selection. Moreover, any automated implementation must carefully consider relevant parameters and considering convex hull algorithm proves significantly more reliable in this regard.

Moreover, this chapter examined the applicability domains of MCR-ALS and K-means, highlighting also their potential synergy. During the research, a density-based clustering method, DBSCAN, was also tested under the hypothesis that density variations that was observed in the normalized scores could be more effectively clustered using a density-based approach. However, the results were unsatisfactory due to the complexity of selecting the appropriate parameters, radius and the number of neighboring points. This likely arises from the nature of the data distribution, which lacks clearly defined empty spaces, making it unsuitable for this type of algorithm.

It is also important to highlight that a major limitation of this research was the conceptualization process. Simulating scenarios in which both K-means and MCR-ALS encounter challenges was difficult, and such cases were equally hard to identify in the available datasets.

5.2 Perspectives

Current research on this topic is now focusing on the practical implementation of this approach in different imaging techniques and on specific datasets that face limitations with clustering and spectral unmixing approaches. Collaboration with the Department of Pharmacy at the University of Salerno (Italy) has led to the application of this exploratory method to mass spectrometry images of cancerous tissue, for the characterization of healthy and unhealthy pixels' tissue. A collaboration with the Porto Conte Research Centre (Sassari, Italy) has supported the analysis of magnetic resonance images of several food matrices to identify distinct regions characterized by different chemical compositions. Both collaborations are ongoing and have shown promising preliminary results that are being further investigated. Some of these results are included Appendix A.

Future research may also deepen the investigation of spectral preprocessing techniques to further improve the effectiveness of the exploratory approach.
To conclude, the further exploration of the combined application of K-means and MCR-ALS remains a promising line of research, as it has not yet been widely explored by the scientific community. This line of investigation has the potential to refine and improve existing methods, eventually contributing to more effective data analysis strategies in spectral imaging.

III

ENHANCING HYPERSPECTRAL IMAGES ANALYSIS THROUGH TENSOR-BASED METHODS

1. Introduction

In the third section of the thesis, the focus is on the exploration of methods based on the decomposition of a tensor. A *tensor* is basically a multidimensional array, which is a higher-order generalization of a vector and a matrix: a first-order tensor is a vector, a second-order tensor is a matrix, and a tensor of the order of three or higher is called higher-order tensor [225]. The terms *order*, *mode*, and *way* are used more or less interchangeably, but a distinction is sometimes made in the literature between *way*, the geometrical dimension of the array, *order*, number of independent ways and *mode*, the basic entity building an array [234, 181]. This nomenclature convention is adopted in this section. Tensor decomposition methods are called *multi-way* and are the natural extension of multivariate analysis, when data are arranged in three, as the case of hyperspectral images or higher-order array. The work related to these methodology has resulted in the production of a paper that is in the process of being submitted for publication.

1.1 Rationale

The analysis of a tensor as it is, requires more complex mathematical models and, often, more computational costs compared to the analysis of two-way data [271]. A few words on the reason why we are embarking this research are thus needed. Let us begin with an example. In remote sensing, the interaction between sunlight and the Earth's surface is often very complex, and observed spectra are then composed of nonlinear mixing terms. In fact, as shown in Chapter 4, complex data often deviate from the Beer's law [169], and in such scenarios, also the bilinear model can not be the proper one to approximate the data. Solutions have been proposed in the literature to improve the bilinear model over the years [272] and in numerous research areas, including imaging [222], neuroscience [273], process analysis [274], social networks [275], and, in general, in context where the underlying information content of the data may not be captured accurately or identified uniquely by two-way data analysis. Two-way analysis methods, by which we refer to those based on factor models here, suffer from rotational freedom

unless specific constraints, such as statistical independence, orthogonality, selectivity and local rank are enforced [201, 208, 209, 210]. These constraints, requiring prior knowledge or unrealistic assumptions, are not often necessary for multi-way models. Consequently, some multi-way analysis offers advantages over two-way models in terms of uniqueness, robustness to noise and, possibly, facility of interpretation. Multi-way analysis has become popular e.g. as exploratory analysis tool in a variety of application areas in the last 50 years [224]. While two-way methods may seem conceptually simpler than multi-way methods, this simplicity vanishes in HSI context. Unfolding the tensor into a matrix for two-way analysis do not consider the spatial structure in the image, sacrificing interpretability. The principle of unfolding, can lead to model that are less interpretable, less robust, and nonparsmonious. This is particularly pronounced when the multi-way structure of the data is ignored, treating the data with an ordinary two-way approach. Also, the more structured is the method, the poorest is the fit compare to methods that uses much more degrees of freedom and consequentially better fit the data. However this can be also a risk because a two-way model may overfit by using extra degrees of freedom to capture noise or redundant variation. As the principle of parsimony of Occam's razor suggests [276], it is preferable to use the simplest model possible: the tensor method is the most constrained model, while the two-way method is the most flexible and complex in terms of interpretability and conditions to be applied. Therefore, the goal of multi-way methods is not necessarily to improve fit but to provide more robust, adequate and interpretable models [181]. Moreover, spatial and/or spectral selectivity are often not fulfilled in complex HSI data, which drastically complicates the analysis in an unsupervised unmixing context. HSI provides us with spatial and spectral information that can be combined instead of being analyzed separately, for instance considering the spatial distribution can help in discriminating the spectral components and getting insights on the sample. The decision of moving to a more complex model is, then, guided by the interest in exploring the hyperspectral tensor keeping its original structure.

1.2 Tensor rank

Before introducing the main tensor decompositions currently used in the context of HSI, it is necessary to present some basic definitions. In all this section, the term tensor is referred to a third-order tensor, and is indicated as $\boldsymbol{\mathcal{T}}$ of dimensions $(X \times Y \times Z)$. Each element of the tensor has three indices and is indicated as t_{xyz}. Tensor *rank* is a key concept of tensor decomposition methods, and it is defined as:

- The rank of a tensor $\boldsymbol{\mathcal{T}}$ (as opposed to mode-n rank) is the minimal number of rank-1 tensors necessary to describe the array in a linear combination [232, 277].
- A third-order tensor is rank-(L, M, N) if its factor-1 rank, factor-2 rank, and factor-3 rank are equal to L, M, and N, respectively [277]. The word *factor* refers to each term of the decomposition.
- A third-order tensor $\boldsymbol{\mathcal{T}}$ has rank-1 if it equals the outer product of 3 vectors. A rank-$(1, 1, 1)$ tensor is briefly called rank-1 [232, 277].

The rank of an array is extremely important to create a multi-way tensor in a parsimonious way but with sufficient dimensions to describe the data under investigation. Unfortunately, there are not specific rules to determine the maximal rank of arrays in general [181].

1.3 Tensor decomposition methods

Tensor decomposition techniques aim to identify low-rank components in multidimensional datasets. Their advantages and progress in HSI were already discussed in the state of the art chapter. The focus in this paragraph is on the fundamental models used in three-way analysis which are the Parallel Factor Analysis (PARAFAC) model, which is also called Canonical Decomposition (CANDECOMP) and Tucker models.

1.3.1 CANDECOMP/PARAFAC

The models proposed independently by Carroll and Chang [228] and Harshman [227] in the 1970's, CANDECOMP and PARAFAC respectively, are considered equivalent and

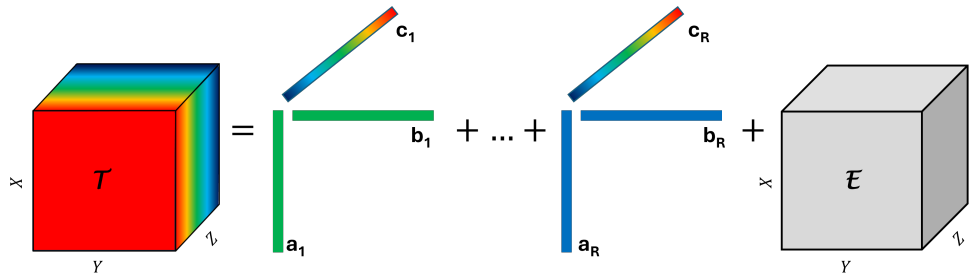

Figure 1.1: The R components CANDECOMP/PARAFAC model of a hyperspectral image $\boldsymbol{\mathcal{T}}$ $(X \times Y \times Z)$: vectors $\mathbf{a}$ and $\mathbf{b}$ are related with the two spatial dimensions of the image, X and Y respectively, while the vector $\mathbf{c}$ is related with the spectral dimension Z. $\boldsymbol{E}$ $(X \times Y \times Z)$ is the tensor containing the residuals of the model.

are the fundamental methods of most currently used three-way models and their multi-way generalizations. CANDECOMP/PARAFAC is considered the simplest multi-way method and can be seen as an extension of bilinear factor models to multilinear data. The model, is represented as the decomposition of a tensor $\boldsymbol{\mathcal{T}}$ in a linear combination of rank-1 tensors (see Fig. 1.1). A rank-1 tensor for a third-order array, is a tensor that can be written as the outer product of three vectors. Then, a R components CANDECOMP/PARAFAC model for a tensor $\boldsymbol{\mathcal{T}}$ of dimensions $(X \times Y \times Z)$ can be expressed as in equation:

$$\boldsymbol{\mathcal{T}} = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r + \boldsymbol{E} \tag{1.1}$$

where $\mathbf{a_r}$, $\mathbf{b_r}$ and $\mathbf{c_r}$ are the column vectors of size $(X \times 1)$, $(Y \times 1)$, $(Z \times 1)$ respectively, and are the columns of the factor matrices $\mathbf{A}$ $(X \times R)$, $\mathbf{B}$ $(Y \times R)$ and $\mathbf{C}$ $(Z \times R)$. $\mathbf{E}$ is the tensor containing the residuals and $\otimes$ indicates the outer product. Having already defined the column vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, the vector outer product is:

$$(\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c})_{xyz} = a_x b_y c_z \tag{1.2}$$

CANDECOMP/ PARAFAC aims to minimize the sum of squares of the residuals obtaining a solution such that factor matrices are determined uniquely. Unique solution means that the estimated model cannot be rotated without loss of fit. Unique solutions can be expected if the vectors of the decomposition are linear independent in two of the

modes. In the third mode, a weaker condition is applied, it is sufficient that no two vectors are linearly dependent. This means the vectors in the third mode do not need to be fully independent, but each pair should still provide a unique contribution. This was shown by Harshman in [227] and Leurgans in [278]. The model uses in fact, factor matrices for all the three ways, but in their model each factor is related to only one factor of each of the other two ways [181]. The CANDECOMP/PARAFAC model in matrix notation makes this concept more clear:

$$\mathbf{I}_z = \mathbf{A}\mathbf{D}_z\mathbf{B}^T + \mathbf{E}_z, \tag{1.3}$$

with $z = 1, ..., Z$. $\mathbf{I_z}$ is the the z-th frontal slice of a three-way tensor, and $\mathbf{A}$ and $\mathbf{B}$ are the factor matrices in the first and second modes, respectively. $\mathbf{D_z}$ is a diagonal matrix, whose diagonal elements are the z-th row of the third factor matrix $\mathbf{C}$. $\mathbf{E_z}$ contains the error terms corresponding to the entries in the z-th frontal slice. This aspect, based on an idea of Cattel in 1944 [279], guarantees that solution is unique up to a permutation or scaling of columns [277, 280, 281]. In those cases, fitting a CANDECOMP/PARAFAC model would give rank deficient solutions and would not guarantee meaningful uniqueness. Parallel Factors with Linear Dependency (PARALIND) is proposed as an approach to modeling such cases [282]. The uniqueness property, has made CANDECOMP/PARAFAC a useful and popular model in different fields, i.e., in chemistry, modeling fluorescence excitation-emission data [283], in food science [284, 285] and in neuroscience [286] among other domains. Even in hyperspectral images analysis, where it is often employed for a denoising of the image [287], or together with other tools, as in [180], where image segmentation techniques are applied to the abundances map to distinguish materials in an area of interest. However, in certain situations, the model may be too restrictive or the uniqueness conditions may not be satisfied. In such situations, the model proposed by Tucker provides a useful alternative [229, 230].

1.3.2 Tucker family

CANDECOMP/PARAFAC can be considered a constrained version of the more flexible Tucker models, which can be seen as an SVD extended to higher order tensor [229],

but without orthogonality constraints on factor matrices [288]. Tucker family includes Tucker1, Tucker2, and Tucker3 models, which refer to specific versions of the Tucker model, based on how many modes are reduced during the decomposition. Among these three methods, Tucker3 is commonly used in hyperspectral images analysis. In literature it can also be referred to as "three-mode principal component analysis", which is actually a proper name to understand what it does. This version of the name was proposed in the work of Kroonenberg [289] who conceptualized an algorithm for fitting the least squares of this model. He proposed this name to distinguish it from the factor analysis, which is usually fitted by means of fitting covariances, using assumptions of uncorrelated unique factors. The advantage of the Tucker3 method is its flexibility, which allows for better modeling of the data. This flexibility is due to the core array $\boldsymbol{\mathcal{G}}$, of dimensions $(I \times J \times K)$ (see Fig. 1.2), which allows an interaction between a factor $\mathbf{A}$ $(X \times I)$, $\mathbf{B}$ $(Y \times J)$ or $\mathbf{C}$ $(Z \times K)$ with any factor in the other modes. However, the price to pay for this flexibility is that a Tucker3 model cannot uniquely determine factor matrices. The Tucker3 model of a tensor $\boldsymbol{\mathcal{T}}$ of dimensions $(X \times Y \times Z)$ can be written as:

$$t_{xyz} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} g_{ijk} a_{xi} b_{yj} c_{zk} + e_{xyz}, \tag{1.4}$$

where g is element of the tensor $\boldsymbol{\mathcal{G}}$ and a, b and c are the elements of the matrices $\mathbf{A}$ $(X \times I)$, $\mathbf{B}$ $(Y \times J)$ and $\mathbf{C}$ $(Z \times K)$ respectively, and e_{xyz} denotes the error term and is an element of $\boldsymbol{E}$. Analogously, the model is written in matrix notation as:

$$\boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{G}} \bullet_1 \mathbf{A} \bullet_2 \mathbf{B} \bullet_3 \mathbf{C} + \boldsymbol{E}, \tag{1.5}$$

were $\bullet_n$ indicates the n-mode product of a tensor with a matrix. This operation means multiplying the core tensor along its *n*-mode with a matrix [232] as:

$$(\boldsymbol{\mathcal{G}} \bullet_1 \mathbf{A})_{xjk} = \sum_{i=1}^{I} g_{ijk} a_{xi}, \tag{1.6}$$

$$(\boldsymbol{\mathcal{G}} \bullet_2 \mathbf{B})_{iyk} = \sum_{j=1}^{J} g_{ijk} b_{yj}, \tag{1.7}$$

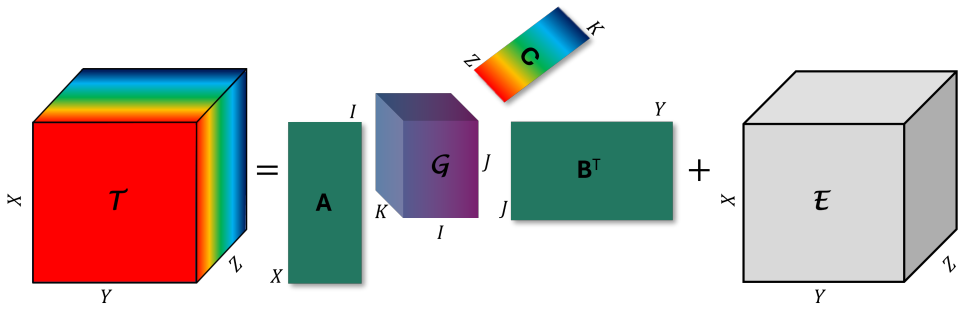

Figure 1.2: The Tucker3 model of a hyperspectral image $\mathcal{T}$ $(X \times Y \times Z)$: $\mathcal{G}$ $(I \times J \times K)$ is the core array that allows the interactions between the factors **A** $(X \times I)$ and **B** $(Y \times J)$ related with the spatial modes of the image, and **C** $(Z \times K)$ related with the spectral mode. $\boldsymbol{E}$ $(X \times Y \times Z)$ is the tensor of the residuals.

$$(\mathcal{G} \bullet_3 \mathbf{C})_{ijz} = \sum_{k=1}^{K} g_{ijk} c_{zk}, \tag{1.8}$$

In many practical scenarios, the goal is to approximate a given tensor by another tensor with a predetermined multilinear rank. This problem can be viewed as a natural extension of finding the best rank approximation for matrices. While matrices can be optimally approximated by truncating their SVD, a similar approach unfortunately does not generally yield the best approximation for tensors when using the Tucker decomposition [226]. Estimation of the multilinear rank of the tensor is still ongoing research [225]. In the literature there are some approaches, mostly developed by De Lathauwer and collaborators, where a truncation of a specific constrained version of the Tucker decomposition [231] yields to a reasonable approximation [233, 290, 291]. It has also been shown by Kiers et al. in [292], that the mathematical structure of the model is so redundant that, even forcing to zero more than half of the elements of the core, the fit does not change. This clearly shows that the Tucker model can be overly complex and explains why it gives ambiguous solutions. Tucker decomposition is often used for dimensionality reduction of large tensor datasets [293], for image processing and analysis [294, 295, 296], harmonic retrieval [297] and for classification and pattern recognition tasks [298], to give just a few examples.

1.4 Block term decomposition

In this section, a different class of tensor decomposition is introduced, which actually leads to a framework that unifies the Tucker and CANDECOMP/PARAFAC models [232]. In between these two methods, there is, in fact, the Block Term Decomposition (BTD), a less popular multi-way model conceptualized by De Lathauwer in 2008 in [232]. BTD can be seen as an extension of the Tucker model for each component, but also as a CANDECOMP/PARAFAC model. In fact, it decomposes the multi-way array into the sum of tensors that have a lower multilinear rank, but not rank-1 like CANDECOMP/PARAFAC. A graphical representation of the model is given in Figure 1.3.

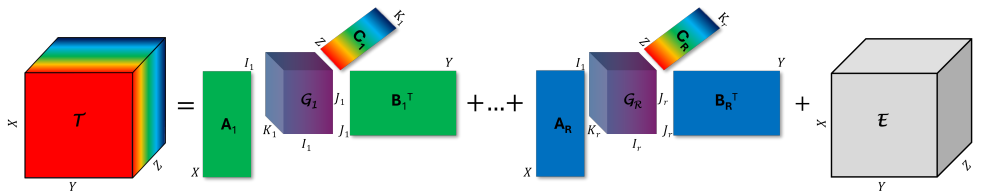

Figure 1.3: The R components BTD model of a hyperspectral image $\boldsymbol{\mathcal{T}}$ $(X \times Y \times Z)$: $\boldsymbol{\mathcal{G}_r}$ $(I_r \times J_r \times K_r)$ is the core array and allows the interactions between the factors matrices $\mathbf{A_r}$ $(X \times I_r)$ and $\mathbf{B_r}$ $(Y \times J_r)$, related with the spatial modes of the image, and $\mathbf{C_r}$ $(Z \times K_r)$, related with the spectral mode. $\boldsymbol{E}$ $(X \times Y \times Z)$ is the tensor of the residuals.

The BTD decomposition of a tensor $\boldsymbol{\mathcal{T}}$ of dimensions $(X \times Y \times Z)$ can be written then as:

$$\boldsymbol{\mathcal{T}} = \sum_{r=1}^{R} \boldsymbol{\mathcal{G}_r} \bullet_1 \mathbf{A_r} \bullet_2 \mathbf{B_r} \bullet_3 \mathbf{C_r} + \boldsymbol{E}, \tag{1.9}$$

where R is the number of components, $\boldsymbol{\mathcal{G}_r}$, of dimensions $(I_r \times J_r \times K_r)$ is the core array for each decomposition and is full rank-(I_r, J_r, K_r). $\mathbf{A_r}$ $(X \times I_r)$, $\mathbf{B_r}$ $(Y \times J_r)$ and $\mathbf{C_r}$ $(Z \times K_r)$ are the factor matrices for each r component and $\boldsymbol{E}$ is the tensor containing the residuals of the model. The symbol $\bullet_n$ indicates the n-mode product of a tensor with a matrix. This operation means multiplying the core tensor along its n-mode with a matrix [232] as described in equations 1.6, 1.7, 1.8. The idea is that the tensor is decomposed into a sum of blocks of inherently smaller size $(I_r \times J_r \times K_r)$ [226]. Basically, if there is only one component $(R = 1)$, then equation 1.9 is a Tucker decomposition, if all the blocks have a scalar core $(I = J = K = 1)$, then

the model is a CANDECOMP/PARAFAC decomposition. The advantage of BTD over CANDECOMP/PARAFAC is that the terms are more general, and could then potentially allow to model more complicated data structures. The advantage of BTD over the Tucker model is that BTD models are essentially unique under some conditions. Uniqueness conditions are discussed extensively mathematically in [232, 291]. A more chemical approach at this regard is discussed in the next chapter, paragraph 2.4. Based on the rank of the decomposition, distinct types of BTD can be considered [232]. During the exploration of the tensor decomposition methods for hyperspectral images, the one characterized by a multilinear rank-$(L_r, L_r, 1)$ for each term of the decomposition, appeared to be of great interest for the phd project and will be the focus of the next chapter. Rank-$(L_r, L_r, 1)$ indicates that the factors have rank-L_r in the first ad second mode and rank-1 in the third mode for each r-component. This approach seemed to be worthwhile because it provides flexibility in modeling the complexity of the spatial distribution of each component. In particular, each component can be modeled on the basis of its spectral signature, but taking into account its spatial complexity. This makes the model highly suitable for the analysis of hyperspectral images. Despite its potential, there are currently few studies in the literature that employ this method for image analysis [236, 299, 300, 301, 302] and key aspects, such as the accurate estimation of the decomposition parameters, have only been explored by a small subset of the scientific community. These challenges will be addressed in the next chapter of this thesis.

2. The rank-(Lr, Lr, 1) decomposition

This chapter provides a detailed description of a tensor decomposition in multilinear rank-$(L_r, L_r, 1)$ terms. The methodology and algorithm are presented, with a particular focus on the criteria and estimation of parameters for using this method. This aspect represents one of the main challenges within the scientific community. The validation of the model and its uniqueness properties are also discussed.

2.1 Decomposition

The rank-$(L_r, L_r, 1)$ terms tensor decomposition have been proposed by De Lathauwer in the 2008 in his important work on the higher-order tensor decomposition in block terms [232]. The tensor $\boldsymbol{\mathcal{T}}$, of dimension $(X \times Y \times Z)$, is decomposed as a linear combination of R terms. Each r-term of the combination is described as the sum of the outer product of a matrix, resulting from $(\mathbf{A_r}\mathbf{B_r}^T)$, with a vector $\mathbf{c_r}$:

$$\boldsymbol{\mathcal{T}} = \sum_{r=1}^{R} (\mathbf{A_r}\mathbf{B_r}^T) \otimes \mathbf{c_r} + \boldsymbol{E}. \tag{2.1}$$

$\mathbf{A_r}$ $(X \times L_r)$ and $\mathbf{B_r}$ $(Y \times L_r)$ are the factor matrices, $r = 1, 2, ..., R$ indicates the terms of the linear combinations, and L_r is the number of subfactors for each factor matrix. $\boldsymbol{E}$ $(X \times Y \times Z)$ is the tensor of the error, which describes the portion of variability in the original data that is not explained by the model. The decomposition in rank-$(L_r, L_r, 1)$ terms, represented in Figure 2.1, indicates that the rank in two modes is the same and equal to L_r (and must be different from 1), while the rank in the third mode is equal to 1. This can lead to the idea of considering the matrices $\mathbf{A_r}$ and $\mathbf{B_r}$ describing the variability expressed in the two spatial dimensions X and Y, while the third mode, of rank-1, describes the variability in the spectral dimension Z through the vector $\mathbf{c_r}$. It is important to note that the rank-L_r of the factor matrices can be different for each r-term of the decomposition. This is an important property because, relating each term of the decomposition with a system component, the

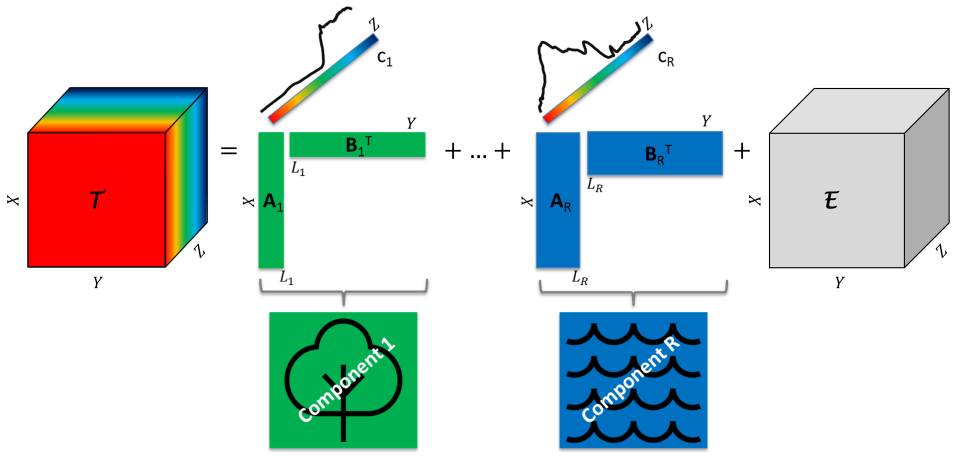

Figure 2.1: Scheme of the decomposition in multilinear rank-$(L_r, L_r, 1)$ terms of the hyperspectral data cube $\boldsymbol{\mathcal{T}}$. The tensor is decomposed as the sum of the outer product of a matrix resulting from $(\mathbf{A_r}\mathbf{B_r}^T)$ with a vector $\mathbf{c_r}$. The dimension of the factor matrices $\mathbf{A_r}$ and $\mathbf{B_r}$ is related with the number of subfactors L_r. The factor matrices described the spatial contribution of the r component while the vector $\mathbf{c_r}$ it is related with the spectral profile of that component. Each component r, in this case vegetation and water, are represented by a different color.

spatial complexity of each component is taken into account, and can be described with the proper amount of information represented by the subfactors L_r. However, the estimation of the proper number of subfactors and components is not trivial and is object of discussion in this chapter (see paragraph 2.4). The idea behind this special case of BTD is that, the spatial distribution sharing the same spectral profile can be considered as a component. The matrix $(\mathbf{A_r}\mathbf{B_r}^T)$ captures the variability or information related to the spatial structure of a component, while the vector $\mathbf{c_r}$ associates this spatial distribution with its corresponding spectral profile, and vice versa. By considering this spatial and spectral information together, the unmixing process is improved, particularly for components with structured spatial distributions, as well as for minority components and those with no selectivity in the spatial and/or spectral domains.

2.2 Algorithm

The algorithm is partially provided by Tensorlab 3.0, a Matlab toolbox for tensor computations and optimization from De Lathauwer and collaborators [303]. It provides algorithm for different decomposition with the possibility of imposing structure on the factors. The structured data fusion by nonlinear least squares algorithm, sdf_nls [304], is used together with the $ll1$ factorization routine, that computes the decomposition of a tensor $\boldsymbol{\mathcal{T}}$ in multilinear rank - $(L_r, L_r, 1)$ terms using a multistep approach. The first initialization step is performed through a generalized eigenvalue decomposition with the $ll1\_gevd$ routine, and the decomposition is performed by minimizing the difference between the data $\boldsymbol{\mathcal{T}}$ and the reconstructed tensor $\boldsymbol{\mathcal{U}}$. The optimization step proceeds with a non-linear least squares Gauss-Newton with dogleg trust region, as suggested in [305], which attempts to find a local minimizer of the real-valued function with the nls_gndl routine [306]. Since the spatial distributions and spectra of the data of interest are assumed to be positive, the factors have been constrained to be non-negative in both the spectral and spatial modes. The routines are compatible with full, sparse, and incomplete datasets. The algorithm operates in a non-sequential way.

2.3 Uniqueness

This decomposition is considered unique under the condition in Eq. 2.2. Following the work of De Lathauwer in the 2008, in which the mathematical theorems regarding the uniqueness property of the tensor decomposition methods are extensively discussed (section 4 in [232]), the uniqueness condition can be written as:

$$\sum_{r=1}^{R} L_r \leq \min(X, Y). \tag{2.2}$$

L_r are the subfactors of the factor matrices, R the number of components and X and Y the spatial modes of the tensor. This condition essentially means that the model is unique if the factor matrices $\mathbf{A_r}$ and $\mathbf{B_r}$ are full column rank. For a unique structural model the parameters, R and L_r, cannot be changed without changing the fit of the

model, in other words, there is only one solution giving the minimal loss function value.

2.4 Methodology

In order to properly consider this multi-way method, a decision must be made regarding the dimensionality of the decomposition terms. To determine the appropriate dimensionality for a model, considerations of parsimony, stability, and interpretability should ideally be taken into account [307]. Although several strategies have been proposed in the literature, the field remains in a research phase [300, 302, 305, 308, 309]. In fact, the findings in the literature are not always consistent and may not be applicable to all cases [310]. The parameters that need to be determined and optimized are R, the number of components, and L, the number of sub-factors for each r component. The selection of these two parameters, particularly L, is non-trivial and significantly impacts the results. To derive the values of R and L, various methodologies were tested, and the one that, to the best of our knowledge, is the most effective is detailed here.

2.4.1 Number of components determination

The choice of the number of components R to describe the system under investigation is essential. To determine R, exploratory analysis is commonly used as in two-way methods. For example, observing the scree plot resulting from a PCA or exploring the data in a normalized space derived from an SVD analysis, as demonstrated in Section 2 of this thesis, are effective strategies. In some cases, *a-priori* knowledge or ground truth may be available and should be considered.

2.4.2 Number of subfactors determination

The estimation of the number of subfactors necessary to describe each component is a key step in determining the correct dimensionality of the factor matrices involved in the decomposition. To choose the number of subfactors, L, a criterion proposed in the

literature associates its value with the spatial (X and Y) and spectral (Z) dimensions of the image [236]:

$$L_{opt} = \frac{\min(X, Y)^2}{R \times Z}. \tag{2.3}$$

In this work, L is weighted by the ratio of the spatial size to the spectral size. This weight adjusts L, increasing it when the spatial size is large relative to the spectral size, and reducing it in the opposite case. The literature also suggests a uniqueness criterion [232, 237] that must be satisfied to ensure the derivation of a unique model, as presented in equation 2.2. If this condition is not met, the solution is non-unique. From the uniqueness criterion, a maximum value for the number of subfactors, L_{max}, can be determined. This value is assumed to be the same for all components, though this is not strictly necessary. In practice, one component may have a higher number of subfactors, while another component may have fewer, with the total number of subfactors adjusted accordingly, as long as the uniqueness condition is not violated. According to most of the literature, the number of subfactors should not be excessively high. Considering also this aspect then, the L_{max} is derived from the uniqueness condition as:

$$L_{max} = \frac{\min(X, Y)}{R}. \tag{2.4}$$

In this sense, the model will undoubtedly satisfy the uniqueness condition, while keeping the number of subfactors not too high. An attempt was also made to impose a condition on the sum of the subfactors, ensuring it did not exceed the smallest spatial size of the dataset. However, this approach resulted in an increase in the number of subfactors, a significant increase in computational time, and no discernible differences in the models. The approach to estimate the number of subfactors involves both the L_{max} and L_{opt} values, using them as guidelines to explore all possible combinations of subfactors.

2.4.3 Model selection

Once values for L_{max} and L_{opt} are determined, all possible values between 2 and either L_{max} or the L_{opt} limit can be considered for each component. Subsequently, all combinations of these values across components are evaluated. The choice of

the upper limit depends on L_{max}, if it is excessively high it is preferable to use L_{opt}. Sometimes, a balance between these two values has to be considered for each dataset to achieve a better result. For each combination, a model is constructed and its performance is visualized using a Tucktest plot [311], which shows the reconstruction error as a function of the number of components used in the decomposition (see Fig. 2.2). The reconstruction error for a 3rd-order tensor $\boldsymbol{\mathcal{T}}$ of dimensions $(X \times Y \times Z)$ is calculated as the Frobenius norm of the difference between $\boldsymbol{\mathcal{T}}$ and the reconstructed tensor $\boldsymbol{\mathcal{U}}$, normalized by the Frobenius norm of $\boldsymbol{\mathcal{T}}$:

$$Error = \frac{\|\mathcal{T} - U\|_F}{\|\mathcal{T}\|_F}, \tag{2.5}$$

where

$$\|\mathcal{T} - U\|_F = \sqrt{\sum_{x=1}^{X}\sum_{y=1}^{Y}\sum_{z=1}^{Z}(t_{xyz} - u_{xyz})^2}, \tag{2.6}$$

and

$$\|\mathcal{T}\|_F = \sqrt{\sum_{x=1}^{X}\sum_{y=1}^{Y}\sum_{z=1}^{Z}(t_{xyz})^2}. \tag{2.7}$$

The Tucktest plot displays all the computed models, with those on the lowest front representing the best-fit models for a given $\sum_{r=1}^{R} L_r$. The model with the absolute minimum error (see Fig. 2.2) is the selected one. Other points of interest, such as models at elbow points or those with errors close to the minimum but requiring fewer subfactors, are also considered and analyzed.

2.4.4 Validation

Validation is a critical step in any methodology. In an unsupervised context, where the ground truth of the image may not be available, the intuition is to use a split-half analysis, as suggested by Kiers et al. [312] for the Tucker3 model. A split-half procedure [313] evaluates the stability of the model by dividing the dataset into two halves and analyzing them independently. Due to the uniqueness of the rank-$(L_r, L_r, 1)$ decomposition, the error for the full dataset should be comparable to the errors of the two halves. Some differences are expected, as the errors depend on the specific sampling of the halves. If the model is stable in the split-half sense, it

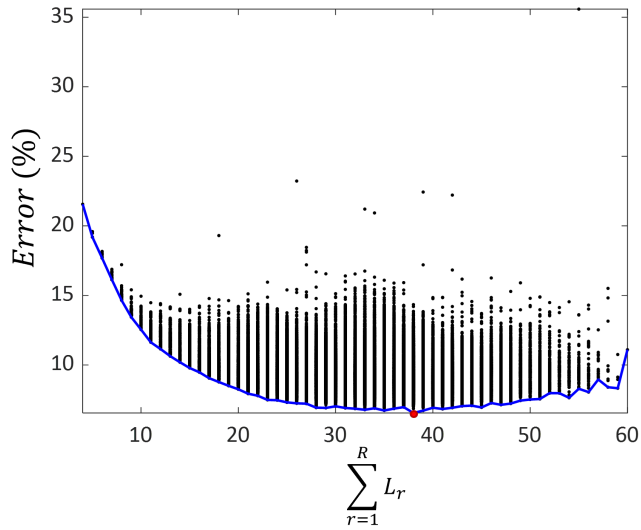

Figure 2.2: Tucktest plot. It shows the decomposition error as a function of $\sum_{r=1}^{R} L_r$. Each black point represents a computed model. Models along the front in blue correspond to L-combinations with the lowest error for a given number of subfactors. The model with the absolute minimum error (red circle) is selected, representing a specific combination of L-subfactors.

can be assumed to be real, capturing the essential variation in the data and being unique [314]. When performing a split-half analysis, the mode along which to split the data must be carefully chosen. Components need to be equally represented in each half, ensuring a balanced number of pixels for each component in both halves. Harshman and DeSarbo [313], also, indicates that both splits must be large enough to minimize the influence of the characteristics of specific components. However, how much should be "large enough" is not possible to say in general, because it depends on the noise level of the data and their underlying structure [307]. The global error is then compared with the errors of the split datasets, which should be similar.

2.5 Method outline

The proposed methodology is graphically represented in Figure 2.3 and summarized in the following steps:

1. **Estimation of the number of components $\mathbf{R}$:** determine the number of compo-

nents in the system under analysis.

2. **Calculation of $\mathbf{L_{max}}$ and $\mathbf{L_{opt}}$:** identify the limits for the number of subfactors and generate all possible combinations of L.

3. **Tucktest analysis:** evaluate the reconstruction error to estimate the best combination of subfactors.

4. **Visualization of the selected model:** visualize the spatial distribution and the spectrum for each component resulting from the chosen model.

5. **Validation of the decomposition:** use a split-half approach to assess the stability and reliability of the selected model.

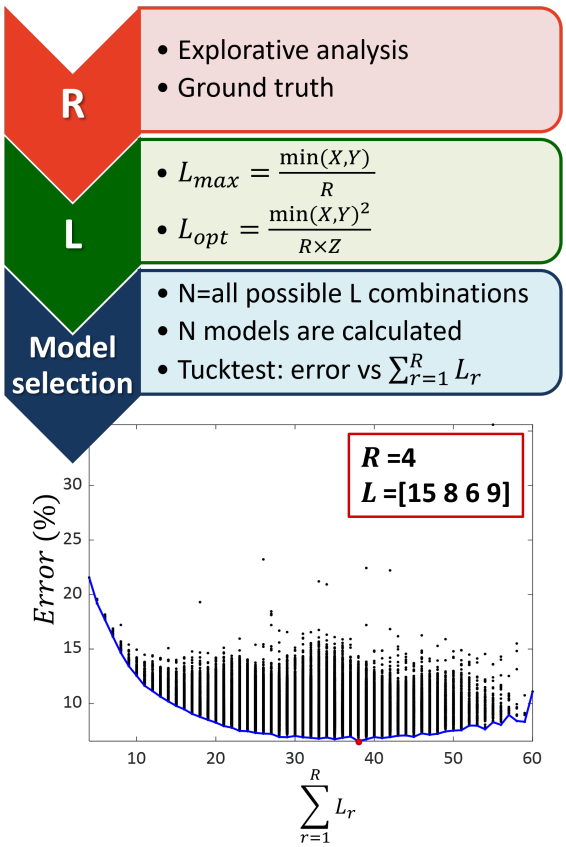

Figure 2.3: Outline of the methodology. Estimation of the number of components (R) and their respective subfactors (L), followed by model selection using a Tucktest plot. In this example, the model selected for a decomposition in 4 components (red circle) corresponds to a combination of subfactors L = [15 8 6 9].

3. Datasets and software

The algorithm and the method described in the previous chapter were tested both on a simulated dataset and on real datasets available to the scientific community or under a collaboration agreement. This chapter includes four representative datasets, among others analyzed, selected for specific purposes. These datasets exhibit different levels of spatial and spectral complexity and are described in each subsection.

3.1 Simulated data

The dataset is a simulation of a hyperspectral image of three components with a structured spatial distribution. The three components were simulated as a rhombus, a square, and a triangle with a significant overlap in the spatial and spectral domains, as shown in Fig. 3.1 A, B. The rhombus and the triangle have no selective pixels in both domains, and the square has selective and mixed pixels but with no selectivity in the spectral domain. Noise is not considered. The spectral image size is 30×30 pixels and has 200 spectral channels. A different image was constructed by shuffling all the pixels. It resulted in an image with the same size of $30 \times 30 \times 200$, with the same spectral profile but with a different spatial distribution (Fig. 3.1 C). These datasets were simulated with the specific objective of showing how taking into account the spatial information in the dataset with a structured spatial domain is of consider importance in the data analysis.

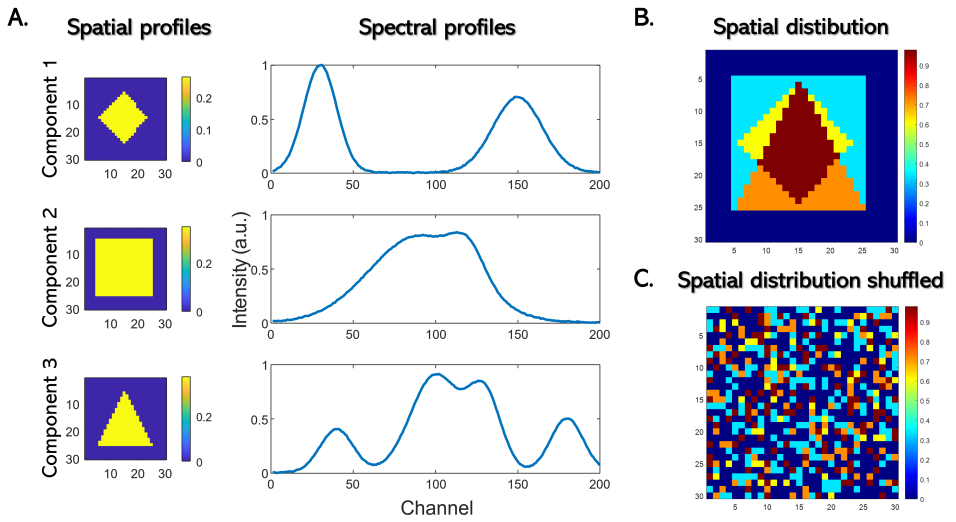

Figure 3.1: Simulated dataset: **A.** the spatial and spectral profiles of the three spatially structured components; **B.** the spatial distribution of the data; **C.** the spatial distribution of the data after pixel shuffling.

3.2 Stained fabric

A stained fabric analyzed with HSI-Near Infrared (HSI-NIR) was made available by the Federal University of Pernambuco, Brazil, and its data acquisition details are given in the work of Silva et al. [315]. It consists of a black cotton fabric 100% with a mixture of semen from 4 different donors and animals. The spectral range is 928 – 2524 nm with a spectral resolution of 6.3 nm and a pixel size of 156 μm^2. The black fabric was chosen among other available color-stained fabric because it shows a more heterogeneous texture compared to white/beige fabrics [315], which should make the analysis even more complex. In addition, it is important to highlight another complexity of the dataset, the high absorbance of black fabric in the range of 900 to 1900 nm, which can be related to the dye used in black fabric manufacturing. A selection of the full dataset (Fig. 3.2) was made, which resulted in a hyperspectral image of size 60 $\times$ 111 $\times$ 256 (Fig. 3.2 B). These data have a complete spatial and spectral overlap between the two components, semen and cotton; there is no selectivity in both the spatial and spectral domains for semen. The spatial structure of the fabric is characterized by a distinct vertical pattern across the image given by the texture of the fabric and a

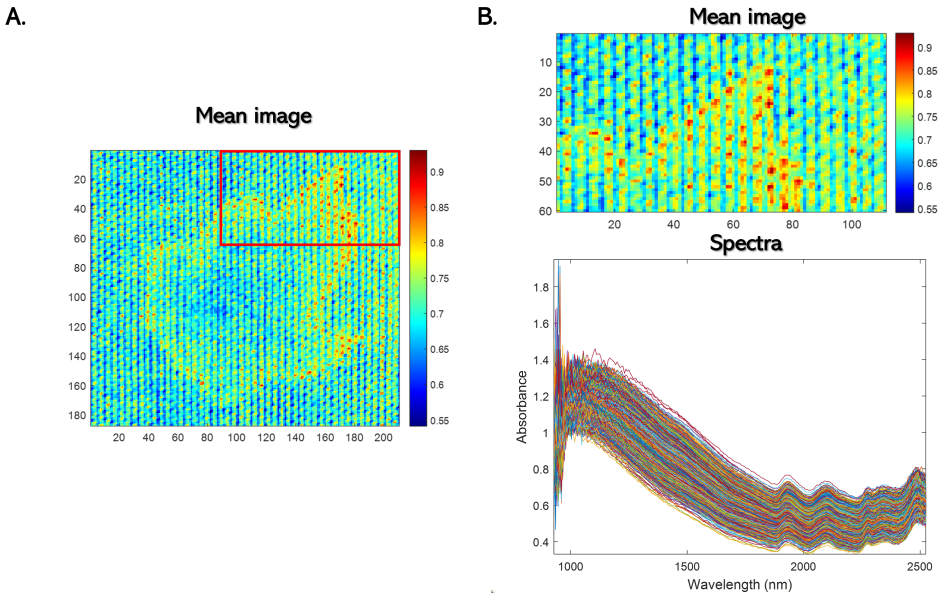

Figure 3.2: The stained fabric data: **A.** mean image of the full dataset with a region of interest selected in red; **B.** mean image ad spectra of the selected dataset.

region where the semen stain is present.

3.3 Remote sensing data

Samson [316] is a dataset in the Visible-NIR (Vis-NIR) spectral range, consisting of three components: soil, vegetation and water, owned by Oregon State University. The hyperspectral dataset was acquired by the SAMSON instrument and was atmospherically corrected using TAFKAA, a hyperspectral atmospheric correction algorithm. Samson contains 95 $\times$ 95 pixels and 156 spectral channels ranging from 401 to 889 nm with a spectral resolution of 3.13 nm (Figs. 3.3 A, B). This dataset is not affected by severe noise. Even if in the literature it is always referred to only three components [317], as shown in the ground truth image (Fig. 3.3 C), investigating the area of acquisition of the dataset (Elkhorn Slough, California), the water component contains a large amount of photosensitizing species [318, 319], which also originate mixed pixels. In fact, the water is part of an estuarine system [320] and this is clear from the satellite

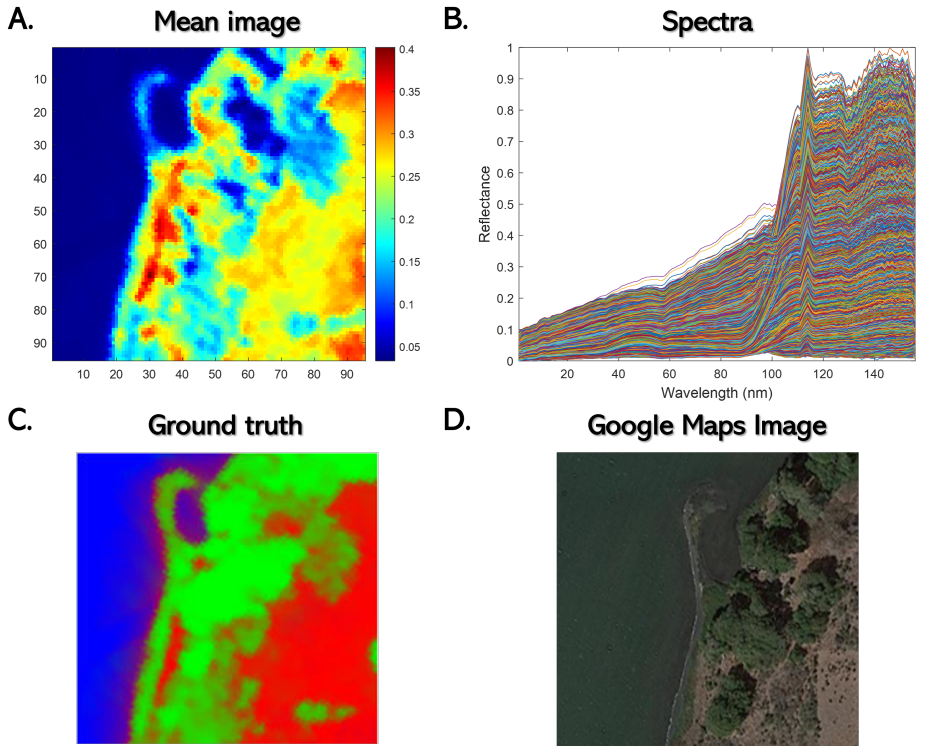

Figure 3.3: Samson dataset: **A.** mean image; **B.** spectra; **C.** ground truth with 3 components: water (in blue), vegetation (in green), soil (in red); **D.** satellite image from Google maps.

image captured from Google Maps and shown in Fig. 3.3 D.

3.4 Data preprocessing

Preprocessing higher-order arrays is significantly more complex than preprocessing 2nd-order data. Unlike matrices, it is not always straightforward to determine how to handle each mode of a tensor [234, 314]. In fact, attempts to preprocess certain datasets resulted in outputs that were no longer interpretable. Someone could think of arrange the tensor in a matrix and preprocess this matrix as in two-way analysis. This, however, will destroy the structure of the data, and the model will no longer be able to reproduce them. It is important to highlight that the two-way preprocessing tools are generally not appropriate for three-way arrays and should be avoided [314]. However, this thesis will not focus further on the preprocessing of tensor data. For

the datasets analyzed in this work, no preprocessing was applied. Despite noise and baseline are present in the real data sets spectra to some extent, we chose not to apply preprocessing to avoid introducing artifact in the three way structure, as well as to evaluate to which extent the tensor approach could cope with noisy data and background contributions, which should ideally be captured by a distinct factor, without hampering recovering the other "chemical" components.

3.5 Software

Analyses were performed using MATLAB© 2022a (MathWorks Massachusetts, USA) and the Tensorlab toolbox [303] for tensor computations and complex optimization together with the MATLAB© Parallel Computing Toolbox.

4. Exploratory analysis of hyperspectral images by decomposition in rank-(Lr, Lr, 1) terms

This chapter presents and discusses the results of decomposing the datasets introduced in Chapter 3 using the multilinear rank-$(L_r, L_r, 1)$ model. At first, the number of components R is properly chosen for each dataset, and the values of L_{opt} and L_{max} is calculated (as explained in Chapter 2) and they are reported in Table 4.1, together with the spatial and spectral dimensions of the datasets.

HSI dataset	$\mathbf{X}, \mathbf{Y}$	$\mathbf{Z}$	$\mathbf{R}$	$\mathbf{L_{opt}}$	$\mathbf{L_{max}}$
Simulated	30, 30	200	3	1.6	10
Stained fabric	60, 111	256	3	4.7	20
Samson	95, 95	156	4	14.4	24

Table 4.1: Parameters of the datasets: spatial (X, Y) and spectral dimension (Z), number of components (R), optimal value L_{opt} and maximum value L_{max} of subfactors.

4.1 Simulated data

The simulated dataset was introduced to highlight both the potential and the limitations of the decomposition in rank-$(L_r, L_r, 1)$ terms. It represents a challenging unmixing problem where two of the three components lack spatial and spectral selectivity. Since the goal of the analysis is to retrieve the three pure components, R is set to 3. All 729 possible models, resulting from combinations of the number of subfactors ranging between 2 and 10 for each component, were calculated to record their errors. The decomposition errors are plotted in Figure 4.1 as a function of $\sum_{r=1}^{R} L_r$. The model with the lowest error (3.09%) is marked with a red circle and corresponds to a model that uses 10 subfactors for describing each component, L=[10 10 10]. The resulting spatial distribution maps and their corresponding spectra are shown in Figure 4.2.

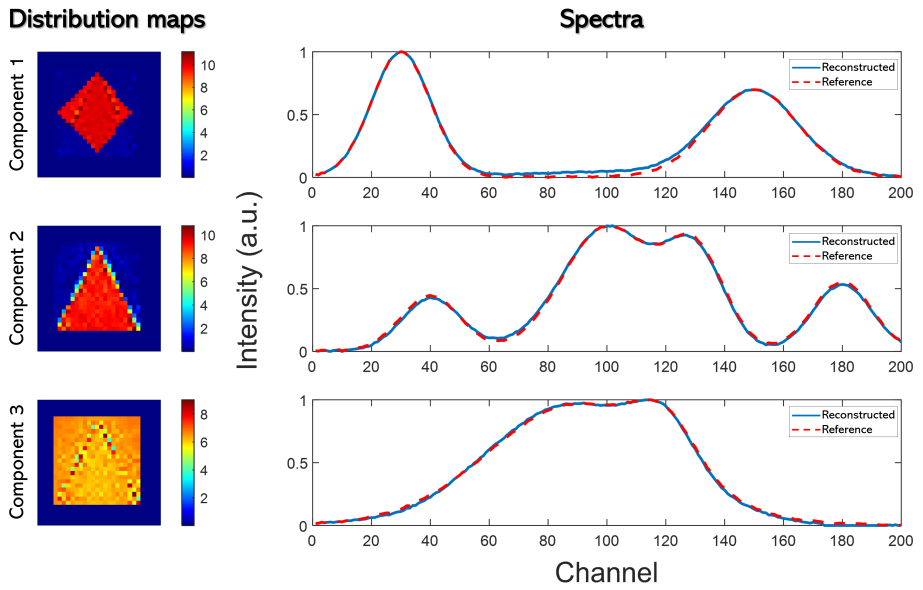

Figure 4.2: Simulated dataset results. On the left are the distribution maps for each of the three components, and on the right are the corresponding spectral profiles reconstructed by the model (blue) and the reference pure spectra (red dotted line).

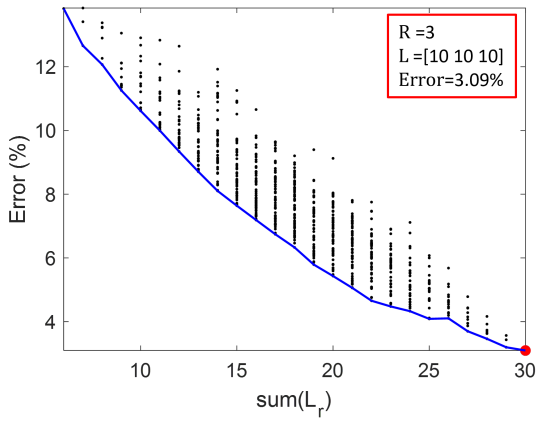

Figure 4.1: Tucktest plot for the simulated data. The model with the highest number of subfactors achieved the lowest error (red circle), corresponding to an approximation of each component with 10 subfactors.

It is interesting to note how the model successfully recovered all the three components, despite the complete absence of spatial and spectral selectivity for the first and the second component (rhombus and triangle). The spatial distributions are nearly

perfectly recovered. The reconstructed spectra (blue line) overlap almost perfectly with the pure reference spectra (red dotted line). It is important to note that MCR-ALS failed to retrieve the first component and the spatial distribution of the third one (square) in this case, due to the lack of selectivity (see Fig. B.1 in Appendix B).

4.1.1 Validation

The model was then validated with a split-half analysis. The dataset was equally divided in two subsets, Datasets 1 and 2, along the (X, Z)-plane fixed the Y-dimension, ensuring an equal number of pixels in each subset (see Fig. B.2 in Appendix B). The models derived from the two halves are shown in Figure 4.3 and the decomposition errors together with the size of the subsets are reported in Table 4.2. The spatial distributions and the spectra of Dataset 1 and 2 are perfectly comparable with the maps and profiles of the full dataset's model. The errors of Dataset 1 and the general model are significantly comparable. The error for Dataset 2 is a bit lower but still comparable with the the fit of the other two models, demonstrating that the model is generally robust. Differences in error are expected because the spatial dimensions of the two halves are relatively small to be described by 10 subfactors for each component, and this could lead to overfit the model.

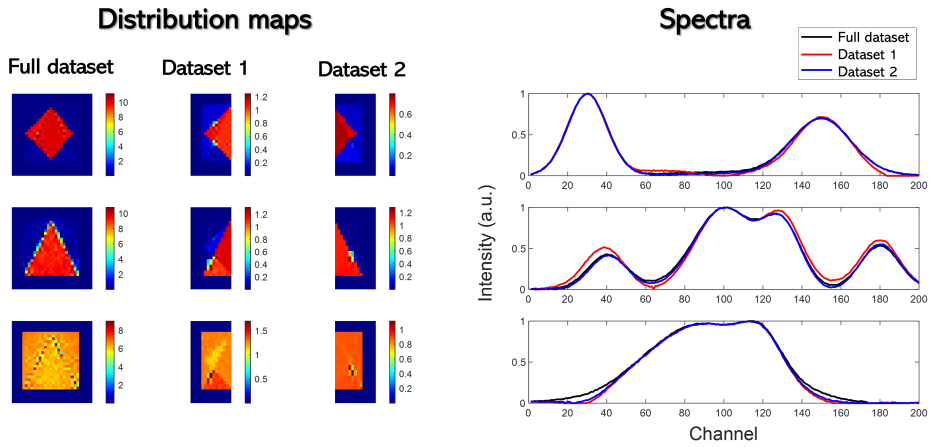

Figure 4.3: Split-half analysis results for the simulated dataset. The general model is compared with the two halves. The distribution maps are shown in the left panel, while the reconstructed spectra for the general model (black) and subsets 1 (red) and 2 (blue) are displayed in the right panel.

	Full dataset	Dataset 1	Dataset 2
Size	30 × 30 × 200	30 × 15 × 200	30 × 15 × 200
Error (%)	3.09	2.85	1.7

Table 4.2: Reconstruction error and dimensions of the simulated dataset and the two dataset subsets resulting from the split-half analysis.

4.1.2 Shuffled dataset

But what if the structured spatial distribution is completely broken? To explore the effects of disrupting the structured spatial distribution, the model was tested on a dataset where the pixel positions were randomly shuffled. In this scenario, the rank-$(L_r, L_r, 1)$ decomposition failed to reproduced correctly the tensor. The spectral profiles are significantly distorted and cannot be recovered (Figure 4.4). This outcome was expected and contrasts with results obtained using unmixing methods for two-way data, where the spatial information is not considered. In those cases, the results remained consistent with those of the original dataset (see Fig. B.3 and Fig. B.4 in Appendix B). The error for the shuffled dataset model was substantially higher, 26.46%, as expected. However, it must be said, that with this tensor decomposition approach, the primary interest lies not in minimizing the error but in the interpretability of the model.

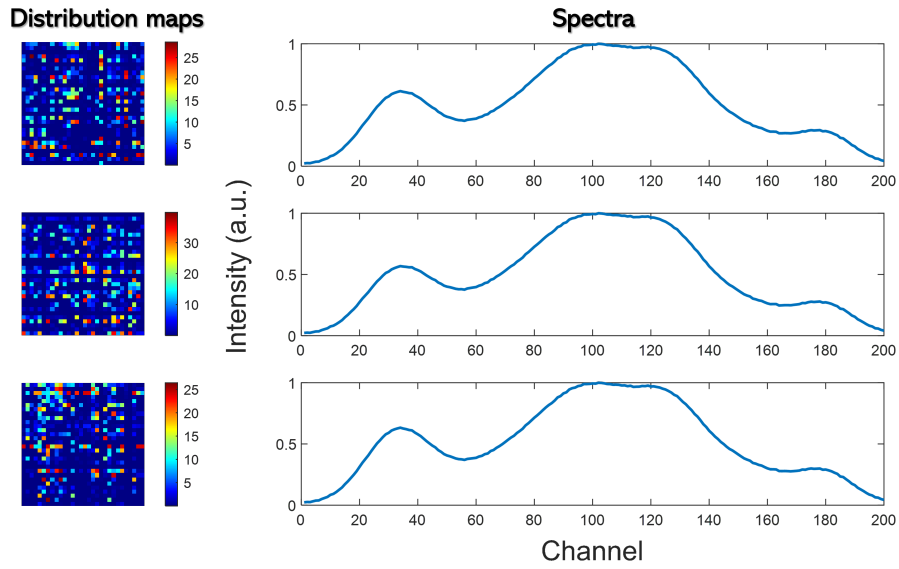

Figure 4.4: Distribution maps (left) and spectral profiles (right). The broken spatial structure precludes the reconstruction of the spatial structure of the three components.

4.2 Stained fabric

This dataset presents a highly complex scenario with significant spatial and spectral overlap between the semen and cotton components. As a stain on fabric, there are no distinct regions selective to semen, and the spectra exhibit overlapping absorbance bands at the same wavelengths. Cotton is always detected in regions where semen is also present, making this dataset particularly challenging to analyze. The number of components was set to three, based on prior knowledge on the dataset which comprises a semen stain, fabric, and background related to the fabric texture. For the subfactors, the maximum number is relatively high, 20, compared to the optimal value, 4.7. A balanced selection of subfactors was then made to reduce the number of model and computational cost. To achieve this, the singular values resulting from the SVD analysis of each z^{th}-image of the tensor were calculated and plotted as a function of the number of principal components (see Fig. B.5 A in Appendix B). The plateau region following the "elbow" in the curve, between 2 and 7, indicated the range over which the value of L was varied for each component. All the 216 models for the possible combinations of subfactors were then computed, varying

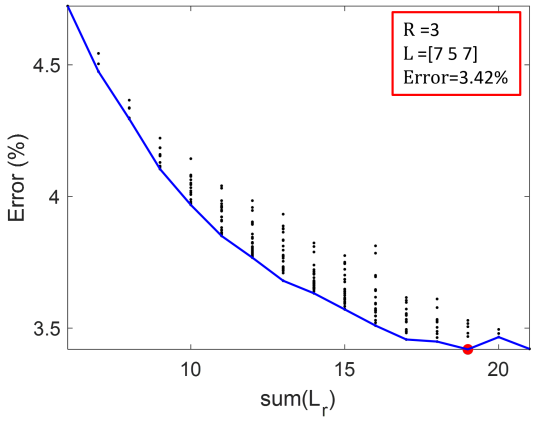

Figure 4.5: Tucktest plot for the stained fabric dataset. The model with the highest number of subfactors achieved the lowest error (red circle), corresponding to an approximation with seven subfactors for the first and third components, and five subfactors for the second.

each component between two and seven subfactors. The model with the lowest error (3.42%) is the one with a combination of seven subfactors for the first and third component and five for the second component (Fig. 4.5). Despite the complexity of NIR spectra complicates the identification, the visualization of the resulting model shows clear identification of the absorption bands and the spatial distributions. The first component is representative of semen, the second of cotton, and the third component, with mixed spectral features, reflects the fabric texture and can be considered as background (Fig. 4.6). The absorption bands used for the component identification are highlighted in light blue. Semen exhibits weak absorption bands between 1850–1950 nm and 2170–2180 nm. According to the literature, these bands are associated with proteins, specifically the vibrational modes of C–O stretching, N–H bending, and C–N stretching combination bands [217, 321]. The cotton fabric spectrum is dominated by absorption bands characteristic of cellulose. It is evident the high absorbance of black fabric between 900–1900 nm, related to the dyes used in black fabric manufacturing [315]. Absorption at 1940 nm is associated to the deformation of the stretching of the 1st overtone of O-H. Absorption bands at 2110 nm from O–H stretching plus C–O stretching and at 2276 nm for the O–H, C–C, and C–H stretching and 2450 nm for the C–H deformation are also considered [315, 322].

In the work of Silva et al. [315], where the MCR-ALS algorithm was used as a two-way

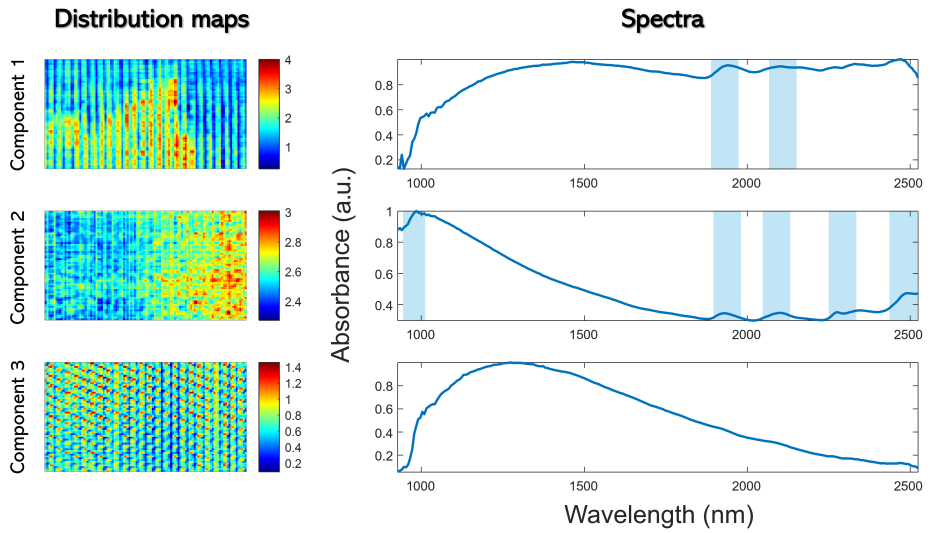

Figure 4.6: Stained fabric dataset results. On the left are the distribution maps for each of the three components, corresponding to semen, cotton and a mixed component respectively. On the right are shown the corresponding spectral profiles reconstructed by the model with the bands of the absorption utilized for the identification in light blue.

unmixing method, components were only partially resolved after extensive preprocessing. In contrast, this method provided excellent results for both spatial and spectral distributions without any preprocessing, demonstrating its effectiveness in handling such complex datasets.

4.2.1 Validation

The model was validated using a split-half analysis, dividing the hyperspectral cube along the (Y, Z)-plane fixed the X-dimension (see Fig. B.2 in Appendix B). In Table 4.3 the dimensions of the two subsets are reported. Each half of the dataset was compared with the overall model (Fig. 4.7). The errors, reported in Table 4.3, are consistent between the general model and the two halves. However, the background component for Dataset 2 shows some association with semen and a contribution from cotton. Additionally, the spectra for components two and three in Dataset 2 are less resolved and appear mixed. It should be noted that splitting a sample like this one into two halves, cannot ensure an equal distribution of pixels between the two datasets. Consequently, differences in the maps or spectra are to be expected.

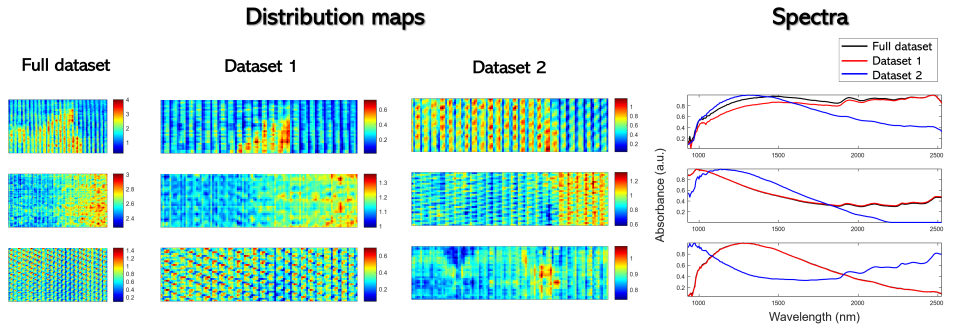

Figure 4.7: Split-half analysis results for the semen dataset. The general model is compared with the two halves. The distribution maps are shown in the left panel, while the reconstructed spectra for the general model (black) and subsets 1 (red) and 2 (blue) are displayed in the right panel.

Despite these minor discrepancies, the interpretation of the two halves is clear, and the error remains highly consistent.

	Full dataset	**Dataset 1**	**Dataset 2**
Size	$60 \times 111 \times 256$	$30 \times 111 \times 256$	$30 \times 111 \times 256$
Error (%)	3.42	3.12	3.13

Table 4.3: Reconstruction error and dimensions of the stained fabric dataset and the two dataset subsets resulting from the split-half analysis.

4.3 Remote sensing data

Samson dataset consists of three ground-truth components: soil, vegetation, and water. However, the unmixing of the three-components dataset using the tensor decomposition method, as well as MCR-ALS applied to the unfolded tensor, yielded suboptimal results, particularly in retrieving the water component. Further investigation of the image acquisition area revealed a mixed water-vegetation component that made isolating the water spectrum particularly challenging. This was primarily due to the water component being poorly selective in both the spectral and spatial domains. Adding an additional component allowed for proper reconstruction of the system. All possible models were computed by varying the number of subfactors from two to $L_{opt} = 15$, resulting in 38416 models. The maximum subfactor value

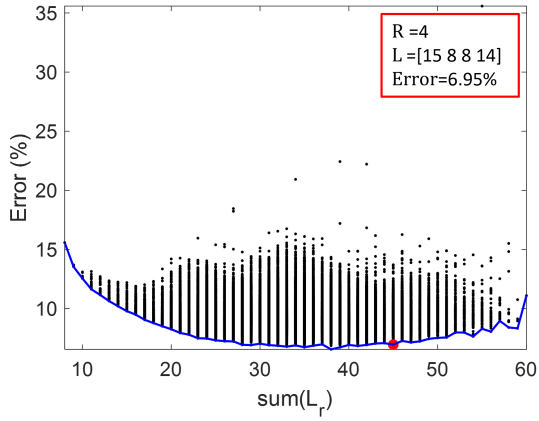

Figure 4.8: Tucktest plot for the remote sensing dataset Samson. The model (red circle), corresponding to an approximation with 15 subfactors for the first component, 8 for the second and the third component and 14 subfactors for the fourth component.

of 24 was also considered, but this led to an exponential increase in the number of models without improving interpretability or achieving a better fit. The selected range of L is also in accordance with what is observed by plotting the singular values S for each image of the tensor as a function of the principal components (see Fig. B.5 B in Appendix B), and with the rank analysis for the first and second mode of the tensor, as shown in Appendix B Fig. B.6. The best model has an error of 6.95% (marked with a red circle in Fig. 4.8), which is very close to the absolute error of 6.45%, which corresponds to a model with L=[7 11 7 13] subfactors. However, seven subfactors are not enough to retrieve the water component and all the model with an error comparable to the absolute error and in a local minimum were observed. The model describing the components with a combination of L=[15 8 8 14] resulted to be the one that better approximates the original data. Results are reported in Figure 4.9. The first reconstructed distribution map shows a profile in the region of water, although some mixed pixels with vegetation are present. The spectrum, however, is perfectly retrieved and correspond to the water spectral profile. The second component is associated with the vegetation and is well reconstructed in both the domains [323, 324], the third component represents a mixture of vegetation, soil, and water. The spectrum combines features of all three, as evidenced by the hump observed between 550 nm and 670 nm. Soil is well reconstructed in the fourth component, both in the

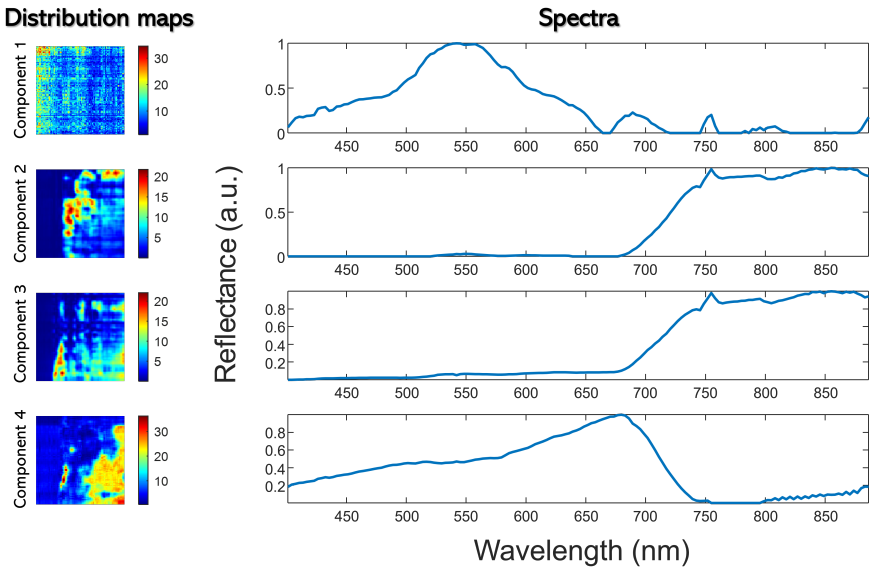

Figure 4.9: Samson remote sensing dataset results. On the left are the distribution maps for each of the four components, corresponding to water, vegetation, a mixed vegetation-soil-water component and soil respectively. On the right are shown the corresponding spectral profiles reconstructed by the model.

spatial and spectral profiles [325, 326].

4.3.1 Validation

The model was split in half along the (Y, Z)-plane, with the X-dimension fixed (see Fig. B.2 in Appendix B). It resulted in two datasets with different X-dimension (Fig. 4.10). The split was performed to include all components within both halves. Due to differences in the pixel distribution of each component, variations in fit, as well as spatial and spectral profiles, are expected. The errors are comparable, as shown in Table 4.4, with differences well within acceptable limits. The distribution maps and

	Full dataset	**Dataset 1**	**Dataset 2**
Size	$95 \times 95 \times 156$	$59 \times 95 \times 156$	$36 \times 95 \times 156$
Error (%)	6.95	8.05	4.23

Table 4.4: Reconstruction error and dimensions of the remote sensing dataset and the two dataset subsets resulting from the split-half analysis.

spectra for both halves are highly consistent, even for the tricky water component,

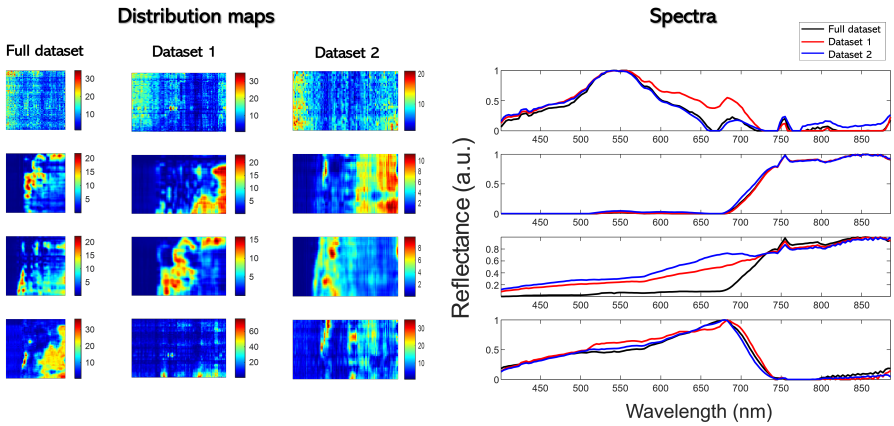

Figure 4.10: Split-half analysis results for the remote sensing Samson dataset. The general model is compared with the two halves. The distribution maps are shown in the left panel, while the reconstructed spectra for the general model (black) and subsets 1 (red) and 2 (blue) are displayed in the right panel.

indicating that the model is particularly stable. Most differences are observed in the mixed component, which was expected, as it is the most affected by splitting the full dataset.

5. Final considerations and perspectives

This chapter provides a summary of key considerations after the presentation of the proposed methodology and results.

5.1 Some considerations

The results obtained using the rank-$(L_r, L_r, 1)$ decomposition approach were highly satisfactory in all the datasets introduced. These datasets were carefully chosen to highlight both the potential and the limitations of the methodology, as they represent complex unmixing problems, where established two-way techniques face significant challenges. When the spatial distribution is structured, the data analysis is significantly improved with this approach. Advantage are also its ability to perform effectively without the need for preprocessing. Additionally, the most important property is that the model offers unique solutions under specific conditions and allows for direct analysis of the data cube without the need for reshaping.

However, an aspect that has to be considered for this approach, is the high computational cost associated with calculating combinations of subfactors to find the optimal configuration that accurately describes all components in a sample. A balance between computational cost, error, and model interpretability can often be achieved by carefully evaluating the values of the optimal and maximum subfactors. User discernment and expertise play an important role in this process, as highlighted in the literature, on estimating the correct dimensionality of factor matrices [314, 237, 327, 328, 329, 310]. This aspect has motivated the exploration of alternative approaches for determining the number of subfactors and the appropriate dimensionality for each component.

An alternative approach that was tested was a superabundant PARAFAC model with at least the maximum value of subfactors L_{max}. From this, similar loading profiles were grouped to determine the number of components. Once R was fixed, the values of the subfactors were derived from the number of PARAFAC components with identical profiles for each R group. This approach did not provide any more interpretable

results than the proposed methodology, although sometimes comparable. However, it is a very cumbersome approach in which the influence of the analyst is very significant. Another approach considered was to analyze the SVD for each z^{th}-image in the hyperspectral tensor ($X \times Y \times Z$) by plotting the individual s-vectors as a function of the number of principal components. The plateau zone after the "elbow" in the curve indicates the range for L (see Fig. B.5 in Appendix B). This approach provided consistent results across datasets and was useful in balancing the optimal and maximum L values to find a range of possible subfactors combinations, when it was necessary (as shown in paragraph 4.2).

A rank analysis was also used, unfolding the tensor in a matrix, for the first ($X \times YZ$) and second mode ($Y \times XZ$). The rank of the matrix of each mode was then estimated, revealing the number of independent phenomena in the first and second modes. This often gave results in agreement with the one proposed by the methodology and helped in estimating the proper range of L values. In Appendix B Fig. B.6 the rank analysis for the remote sensing dataset is reported.

5.2 Perspectives

Some approaches and ideas are here introduced to overcome the problem of computational time and to encourage future work in this area. The concept of essential information in three-order data, as recently proposed by Vitale et al. [330], is considered as a promising field for reducing tensor size. Testing models on reduced data could significantly alleviate computational burdens. However, reconstructing the original tensor remains a non-trivial challenge, since the essential pixels are retrieved in unfolded data to relocate correctly them in the tensor is not straightforward, and it is still on ongoing research. Another aspect is under analysis, which can be useful for the decomposition of large datasets with spatially structured information only in specific regions of the data. It is based on the idea of analyzing only the structured areas, exploring subfactors combinations considering only these regions (whose dimensionality is much less, giving smaller L_{opt} and L_{max}) and once the subfactors corresponding to minima are found, then use the original tensor for the analysis of the full dataset. A further idea involves applying a sparsity constraint to the matrices $\mathbf{A_r}$,

$\mathbf{B_r}$, to emphasize the modes with more information. To achieve this, the optimization routine $nls_g ndl$ should incorporate this information as regularization terms. This approach has already been developed for tensor decomposition methods [331, 237], and we believe that extending it to the decomposition in rank-$(L_r, L_r, 1)$ terms could bring improvements for the determination of components, better generalization and probably speed up the analysis.

As the limitations of traditional techniques became evident, the methodology here proposed offers a fresh perspective for complex HSI data. The development of novel methods, particularly for estimating subfactors for each component, represents the central contribution of this research. This work lays the foundation for further advancements in multi-way analysis and its application to complex HSI datasets.

IV

CONCLUSIONS

1. Final remarks and perspectives

Throughout this thesis, the power and versatility of HSI have been demonstrated. HSI enables the acquisition of detailed spectroscopic information for each pixel within an image, providing insights into both the chemical composition and spatial distribution of a sample. This capability makes HSI invaluable across a wide range of fields. The diverse datasets selected for this thesis highlight the vast applicability of HSI and its potential in different analytical contexts. Despite its many advantages, HSI also presents significant analytical challenges, necessitating novel considerations and the development of new exploratory approaches. This doctoral research has aimed to address these challenges by refining unsupervised methodologies and introducing new strategies to enhance the interpretability of HSI images. The PhD journey followed mainly two two lines of research to address the challenges outlined below.

1.1 Addressing key gaps in clustering and spectral unmixing methods

The first major contribution of this work has been the clarification of the applicability domain of clustering and spectral unmixing methods. Despite these techniques have distinct purposes, partitioning data into meaningful clusters versus resolving pure spectral components, they are still used interchangeably in some research areas. Among the different methods, K-means and MCR-ALS were chosen as core methods for clustering and spectral unmixing, as outlined in the general introduction section. To clarify the two different domains of applicability, K-means clustering is most suitable when imaging data contain pure pixels distributed throughout the field of view, whereas unmixing should be applied when image pixels represent mixtures of pure spectroscopic fingerprints. The simultaneous application of both methods is meaningful only in specific intermediate cases, depending on the nature of the dataset. Intermediate scenarios, where some degree of spectral mixing is observed, present an important question: which method is more appropriate? The answer

depends on the specific objective of the analysis and the type of the data. The Raman powder dataset introduced in Section II has clearly shown how and when K-means and MCR-ALS could lead to comparable results, depending on data characteristics. A combined approach of K-means and MCR-ALS was shown with the LIBS mineral dataset introduced in Section II, where it improved the interpretability of complex datasets. However, in some cases where the data deviate from the assumptions of the methods, neither K-means nor MCR-ALS is sufficient for a successful analysis. To address this, an alternative exploratory method was introduced, which relies on the geometric structure of the data cloud rather than traditional variance-based techniques.

A second key contribution of this research has been the development of an exploratory strategy that account for the geometric structure of hyperspectral data. Traditional unsupervised methods often struggle with datasets exhibiting significant spectral and spatial overlap. Additionally, random initialization steps in clustering can lead to inconsistent results, while the selection of the number of clusters in K-means or the number of components in MCR-ALS can introduce subjectivity. Furthermore, minor spectral components may be overlooked in the absence of prior knowledge. For all this reason, the extraction of the most different spectra and single wavelength images have been show in this thesis, helping in discriminating all the components without the need of optimizations steps and getting a comprehensive information on the investigated sample. This approach offers a more informed and data-driven pathway to hyperspectral image analysis, addressing key limitations of conventional methodologies.

1.2 Enhancing unsupervised hyperspectral analysis through tensor-based methods

The third major contribution has been the exploration of tensor decomposition techniques as an alternative to matrix-based approaches. Most of the methods for HSI analysis often involve unfolding the data cube into a two-dimensional matrix,

disregarding the structure of the hyperspectral data, leading to a loss of valuable spatial-spectral relationships. The PhD research has proposed the rank-(Lr,Lr,1) terms decomposition as a method to preserve the inherent structure of hyperspectral images, describing the spatial distribution with two modes and the spectral profile of each component with one mode, resulting particularly suited for data with structured spatial organization.
Despite these advancements, some limitations remain. First, the estimation of parameters that is still an ongoing research and improvements in this regard could bring to optimize computational efficiency, particularly for large hyperspectral datasets, which is the second limitation of the method.

Possible future research coming from this thesis, could be the integration of the exploratory approach in the rank-(Lr,Lr,1) terms decomposition, to overcome the aforementioned limitations. In conclusion, this doctoral research has contributed to the advancement of exploratory hyperspectral image analysis by addressing critical methodological gaps in clustering and spectral unmixing. By developing new exploratory approaches to the investigation of hyperspectral images, this work hopes to pave the way for future research in this interesting and evolving field.

V

APPENDICES AND BIBLIOGRAPHY

A. Supplementary material for Part II

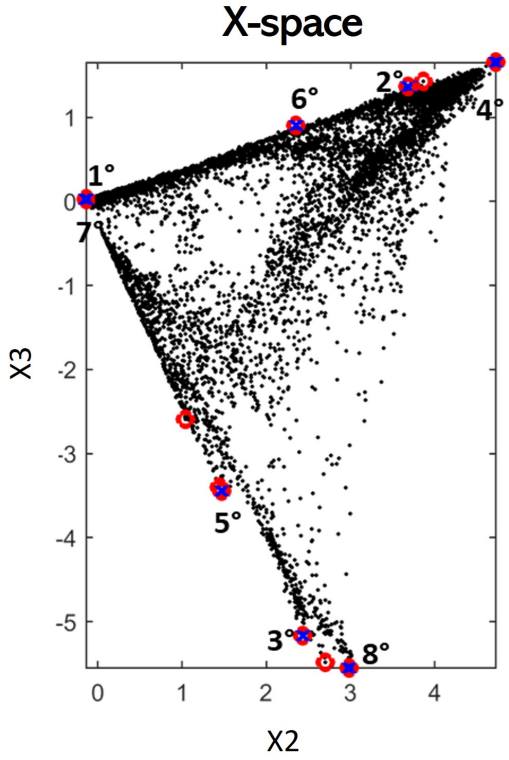

Figure A.1: Raman powder dataset representation in the two-dimensional X-space. The first eight spectra (marked blue) identified by SIMPLISMA computed on the essential information (14 pixels in red) are shown. It should be noted that some pixels are very close and overlapping. When using well-known approaches to prioritize the convex hull points, as SIMPLISMA, is not possible to retrieve all the archetype points identified by looking at the structure of the data, as the first three pure spectra. In fact, in presence of noise, a proper setting of threshold is needed to retrieve them, and more pure profiles need to be extracted.

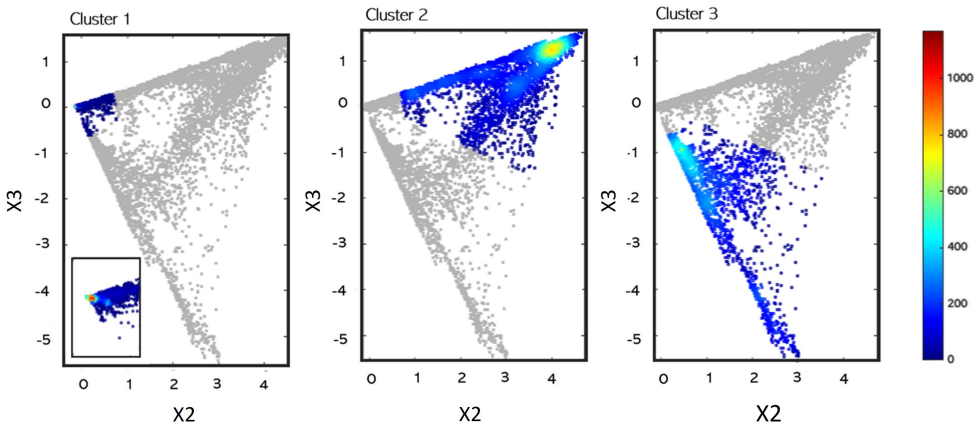

Figure A.2: K-means pixels clusters of the powder mixture dataset, represented in the two-dimensional X-space. Data points are color-coded by cluster density. Cluster 1 has the highest number of pixels focused on the left corner (small distance/high similarity) of the cluster, where the purest pixels are concentrated.

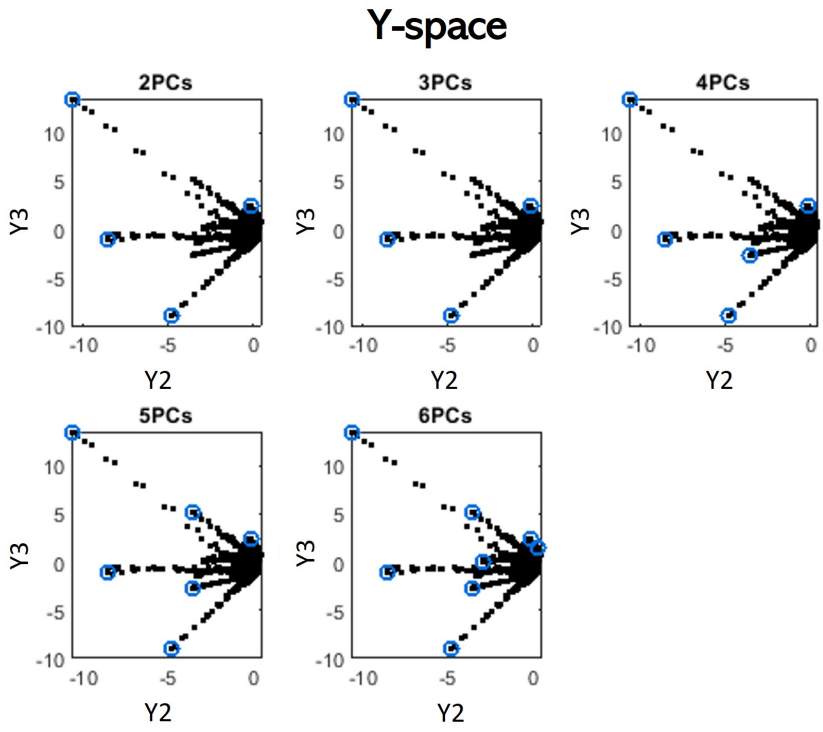

Figure A.3: Projection in the two-dimensional Y-space of the archetype points (blue circles) identified using convex hull computation from two up to six dimensions. Only the archetype points of the convex hull that are not part of the dense data cloud are highlighted for clarity.

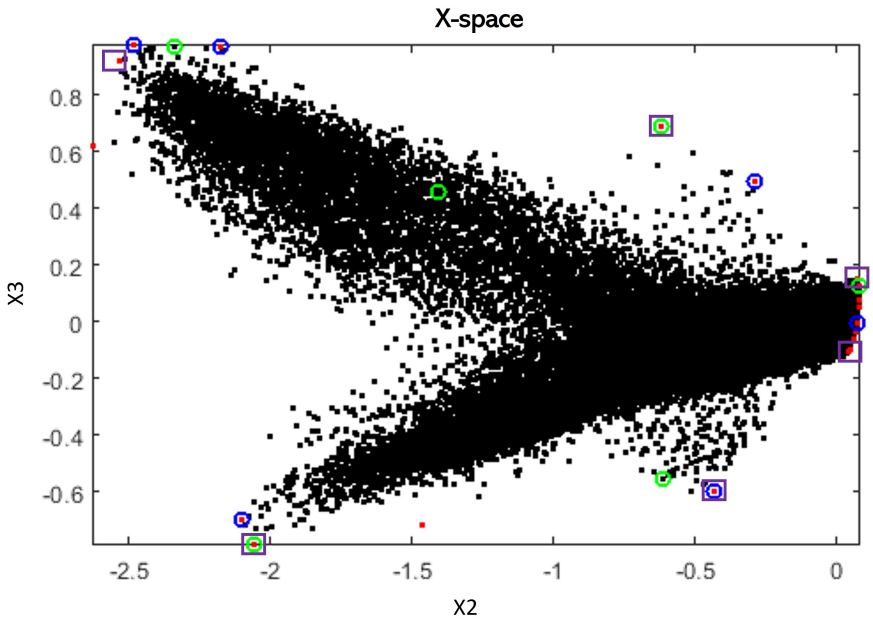

Figure A.4: Projection in the two-dimensional X-space of the mineral sample dataset. In red the 19 pixels identified through convex hull computed for two dimensions (red points). In blue circles are the first six pixels identified by SIMPLISMA using the essential information only and the full dataset (green circles). Purple squares indicate archetype pixels identified through analysis of the structure of the data. By prioritizing pixels and wavelength using SIMPLISMA, it is not possible to recover at the highest position all the points selected from the inspection of the structure of the data. This could result in suboptimal information retrieval, making worth to point out that this difference might become relevant, e.g. using them as first guess in MCR.

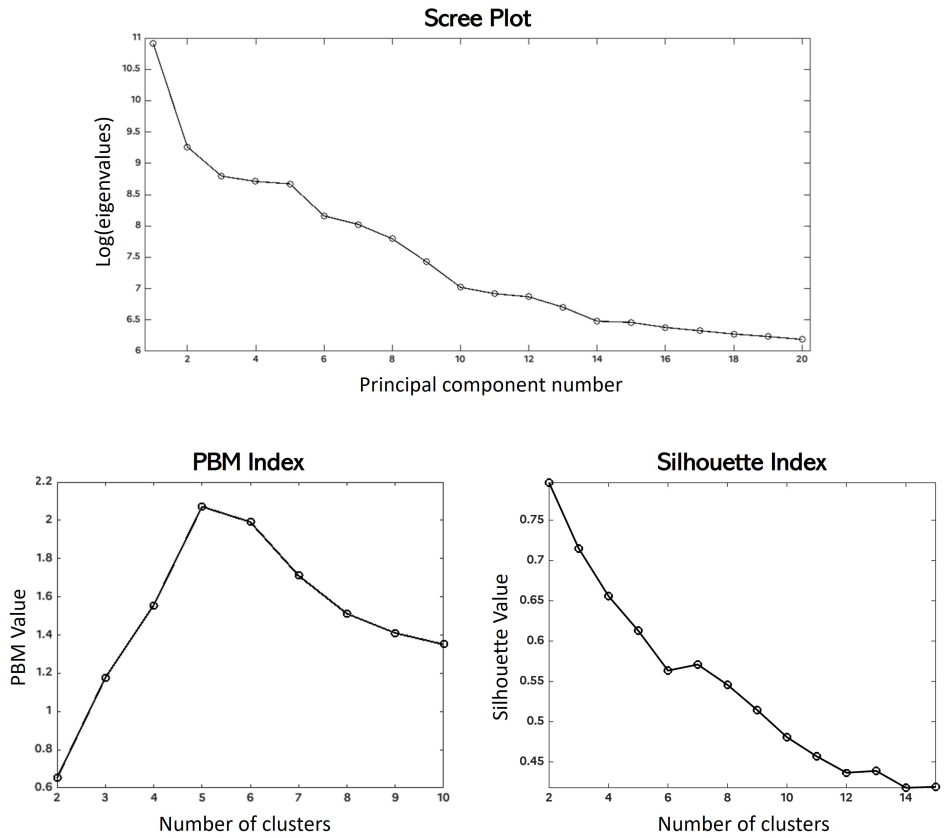

Figure A.5: Scree plot, PBM and Silhouette indices for K-means clustering of the mineral sample. Optimal number of clusters is five or six looking at the scree plot, five according to PBM index, and six according to Silhouette.

Semen droplet on cotton fabric

The proposed exploratory approach was applied to a HSI-NIR image of a semen droplet on white cotton fabric (dimensions: 222 × 220 × 191). Additional details regarding data acquisition are available in Silva et al. [315]. The mean image, shown in Figure A.6, exhibits a complex structure characterized by: (i) a distinct horizontal pattern across the entire image, attributed to the rough texture of the cotton fabric; (ii) a barely perceptible shadow outlining the oval-shaped border of the semen droplet; and (iii) an extraneous fiber filament in the lower central region. This case study presents a notably high level of complexity. The cotton fabric is present throughout

the image, making it impossible to isolate a region exclusively representative of semen. Additionally, the spectral signatures of the different components in the scene exhibit significant overlap [109]. Data were preprocessed with a baseline correction.

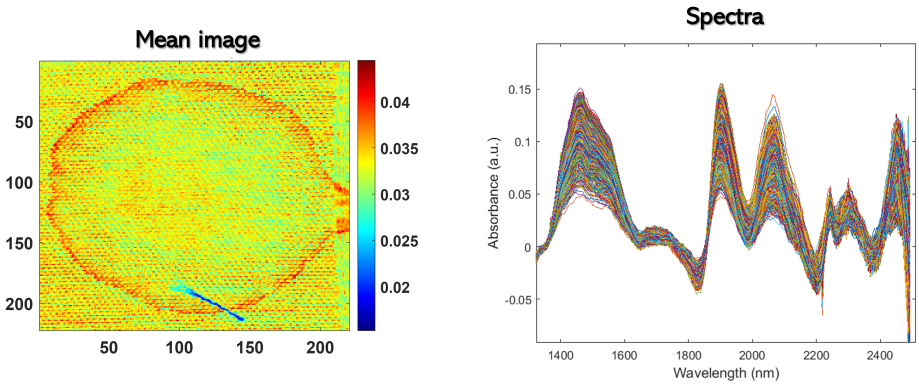

Figure A.6: Semen droplet on cotton fabric dataset. Mean image (left) and spectra (right).

The 2-D representation of the X-space of the dataset reveals two groups of pixels (Fig. A.7). By selecting these two groups, it is possible to differentiate between the inner region of the semen droplet and the outer areas where the semen has spread toward the borders. These regions exhibit variations in drying time, justifying their consideration as two distinct semen groups. The area between these two groups is mainly represented by cotton. By selecting three points from the archetype calculated in the (X2, X3) normalized space, the three extracted purest spectra show significant overlap. However, they can be distinguished based on spectral differences in the 1600–1800 n range, as previously reported in the literature [109, 315].

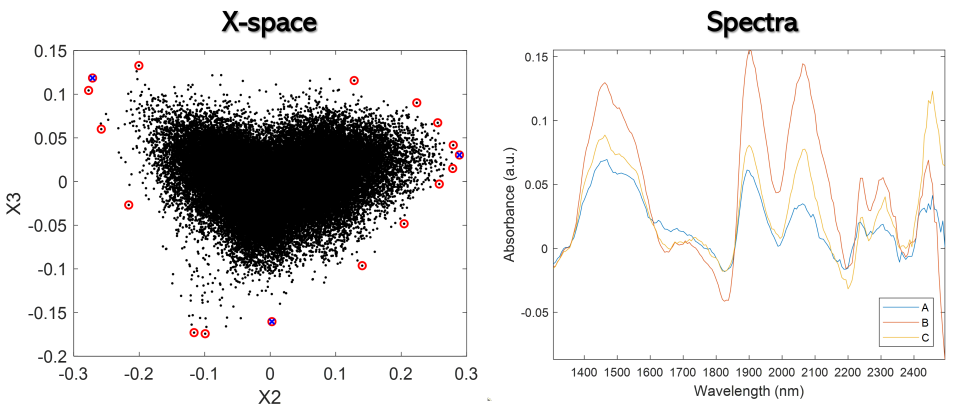

Figure A.7: Semen droplet on cotton fabric dataset. 2-D representation of the X-space of the dataset is represented in the left panel. Red circles mark the archetypes points at the vertices of the convex hull computed in the (X2, X3) normalized space. The letters and blue marks indicate the selected points, while the corresponding spectra are displayed in the right panel.

The Y-space exhibits four distinct directions, represented by different colors in Figure A.8. The extracted images are consistent with those obtained in a previous study [109]. In the upper part of the image, the single-wavelength images associated with semen are represented, while the lower part corresponds to wavelengths related to cotton.

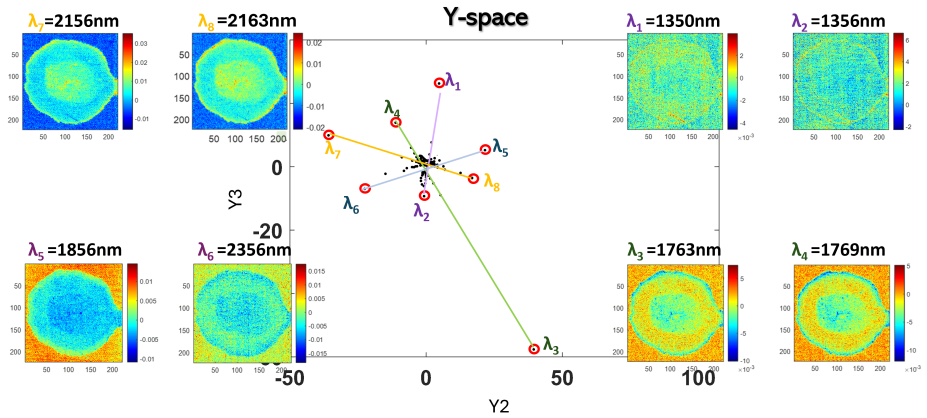

Figure A.8: Semen droplet on cotton fabric dataset. 2-D representation of the Y-space, Red circles mark the archetypes points of the convex hull computed in the (Y2, Y3) normalized space. Red circles represent the selected wavelengths, the corresponding refolded images are shown in the right panel.

Parotid gland tissue

The dataset was kindly provided by the Department of Pharmacy at the University of Salerno. Further details can be found in [33]. Briefly, the MALDI TOF imaging parotid tissue dataset consists of 44 images from 11 patients, including samples of both healthy and pathological tissue in positive and negative ion modes. The approach presented in this work is demonstrated on a single image (Fig. A.9), but it has been applied to the entire dataset. Here, only preliminary results are reported, as this remains part of ongoing research. First, the X- and Y-spaces were analyzed to characterize the data structure for both healthy and pathological tissue. Differences in selected m/z values were observed between healthy and pathological samples .

X-space

- Healthy samples: show a uniform spatial distribution, primarily characterized by a single group.
- Pathological samples: exhibit 2–3 distinct groups with a more globular shape.
- These differences are more pronounced in the MALDI negative TOF-MSI data.
- MALDI negative TOF-MSI data display a consistent distribution across all healthy samples, while a similar pattern is observed for all pathological samples. The same trend is observed also in the MALDI positive TOF-MSI data.
- Patients 2 and 11 present a more complex distribution, differing from those observed in other patients.

Y-space

- Both healthy and pathological samples show m/z data points radiating outward from a central region containing the majority of the data points.

From the different data structures observed in the X-space, the extreme points of each group were selected and analyzed (Fig. A.10). Healthy pixels were identified by comparing the extracted pixels with those of the healthy tissue of the patient. Among

the selected pixels, one from each group can be considered representative, as they exhibit the same spectral profile (Fig. A.11). In the Y-space, the extreme points along each direction were selected, and single-wavelength images were extracted (Fig. A.12). The characterization of m/z is challenging and a comparison with the histological image and a previous study on the dataset was considered [33]. The goal of this project is to develop a computationally efficient approach to gain insight into the tissue under investigation. These findings can then be further utilized for unmixing and classification purposes.

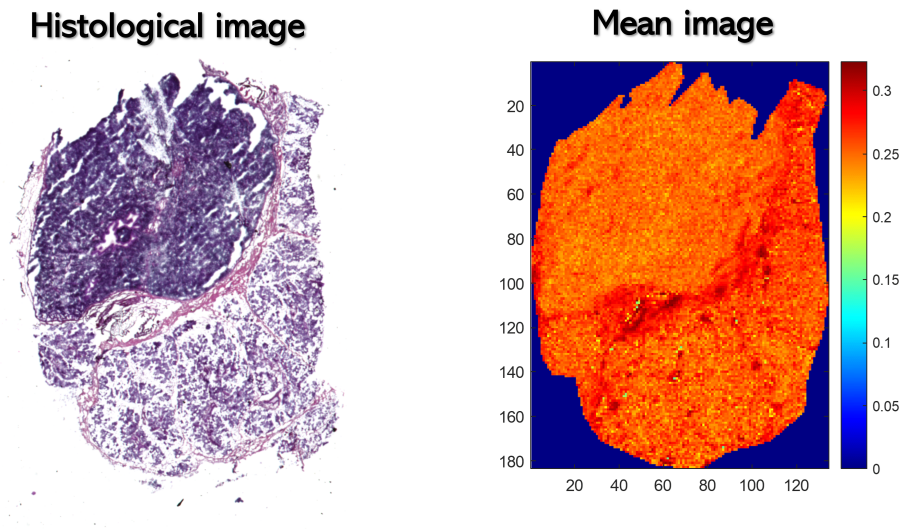

Figure A.9: MALDI-TOF image of parotid gland tissue. Histological image is on the left, and the mean image of the dataset is on the right.

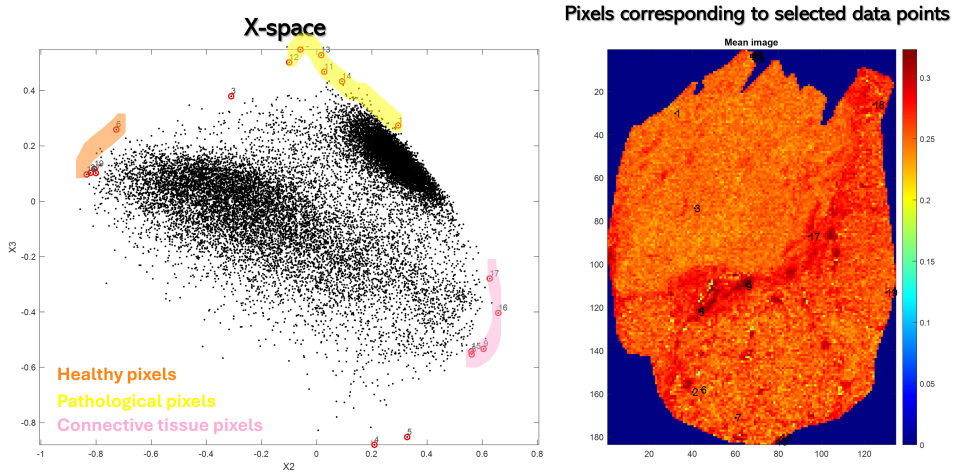

Figure A.10: The 2-D representation of the X-space from the mass spectrometry imaging (MSI) dataset is shown in the left panel, where three groups can be identified, red circles mark the selected pixels. On the right, the selected pixels are highlighted in the mean image.

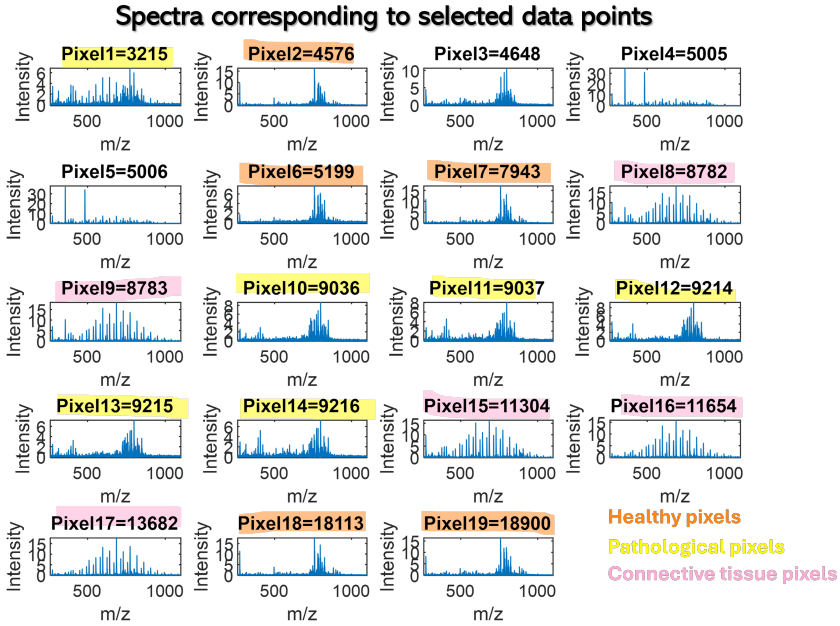

Figure A.11: The corresponding spectra of the selected pixels from the MSI dataset are shown. Orange and pink represent healthy spectra, while yellow highlights indicate pathological pixels. Pixels 3, 4, and 5 are not associated.

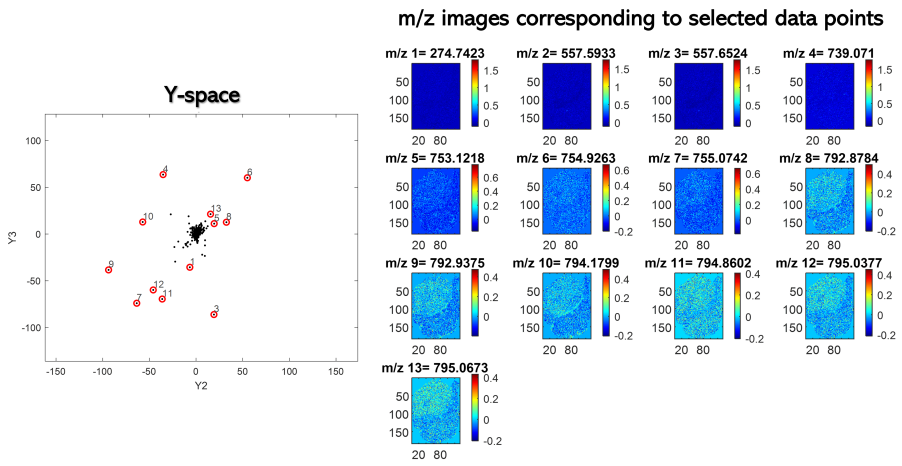

Figure A.12: The 2-D representation of the Y-space from the MSI dataset is shown in the left panel. Red circles represent the selected wavelengths, and the corresponding single-wavelength images are shown in the right panel. The m/z values 10 and 12 are also found in healthy tissue.

Magnetic resonance imaging data

HSI and MRI (Magnetic Resonance Imaging) datasets share several similarities, particularly in terms of their multidimensional characteristics. Both HSI and MRI capture data across multiple spectral bands or channels: HSI records information across hundreds of contiguous spectral bands, while MRI acquires multiple image slices corresponding to different tissue contrasts (e.g., T1-weighted, T2-weighted or diffusion-weighted imaging). Furthermore, both HSI and MRI datasets are considered high-dimensional, often represented as 3D or 4D tensors: HSI data is structured as a spatial-spectral cube (height × width × spectral bands), while MRI consists of multiple spatial slices across different imaging time points (echo times). Accordingly, even though they are used in different fields, HSI and MRI datasets structural similarities make them suitable for feature extraction algorithms, by means of same computational techniques [332]. The proposed approach was tested on a MRI relaxometric dataset acquired on a raw fish fillet kindly provided by the NMR and Imaging Laboratory sited at Porto Conte Ricerche Srl (Sassari, Italy). Further details on the MRI acquisition modalities can be found in [333]. This dataset is representative of challenging scenario for unmixing

of T2 MRI multiexponential decay signals into monoexponential components. The dataset presents different anatomical regions shown in Fig. A.13A: horizontal septum (HS), red muscle (RM), subcutaneous fat (SF), white muscle (WM). The dataset (Fig. A.13 B) has a size of 128 × 128 pixels with a slice thickness of 1.5 mm and 64 echo times that represent the transverse relaxation time (T2).

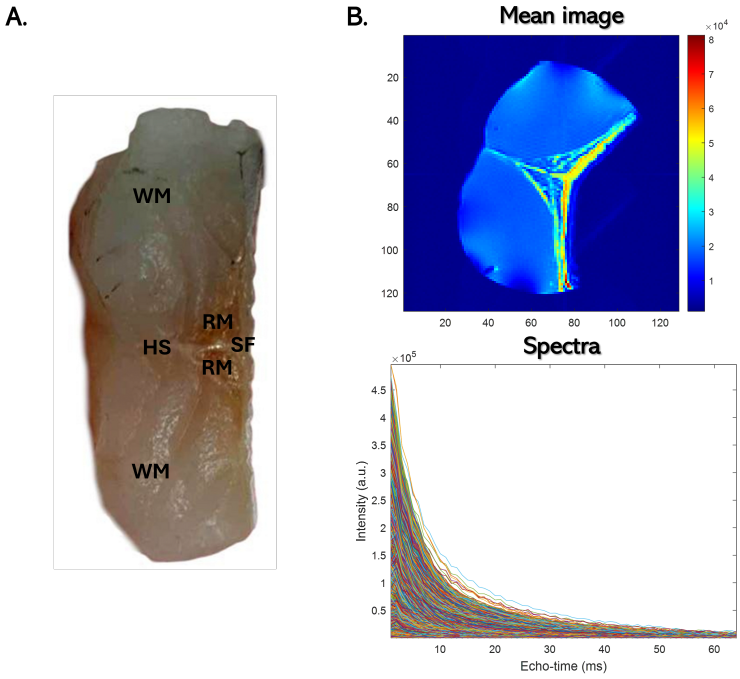

Figure A.13: MRI dataset: **A.** fish filet image with different parts of the sample highlighted in black: horizontal septum (HS), red muscle (RM), subcutaneous fat (SF), white muscle (WM). **B.** mean image ad spectra of the dataset.

The analysis allowed us to distinguish between noisy pixels, primarily caused by the magnetic field and constituting the majority of the dataset, and a smaller subset of pixels representing the decay of fish fillet components. The area of interest, highlighted in Figure A.14, includes the decay of compounds associated with both muscle and fat. The two components can be extracted as single-channel images by analyzing the Y-space, as shown in Figure A.15.

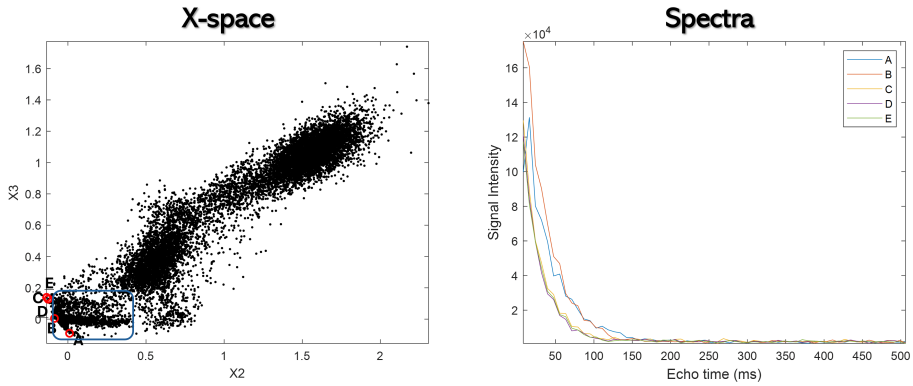

Figure A.14: MRI dataset: the left panel presents a 2-D representation of the X-space, where three distinct groups can be identified. The region corresponding to component decay is highlighted in blue, while red circles indicate selected pixels. The right panel displays the spectra of these selected pixels, revealing two different decay patterns: one associated with fat compounds (A and B) and another linked to muscle components (C, D, and E).

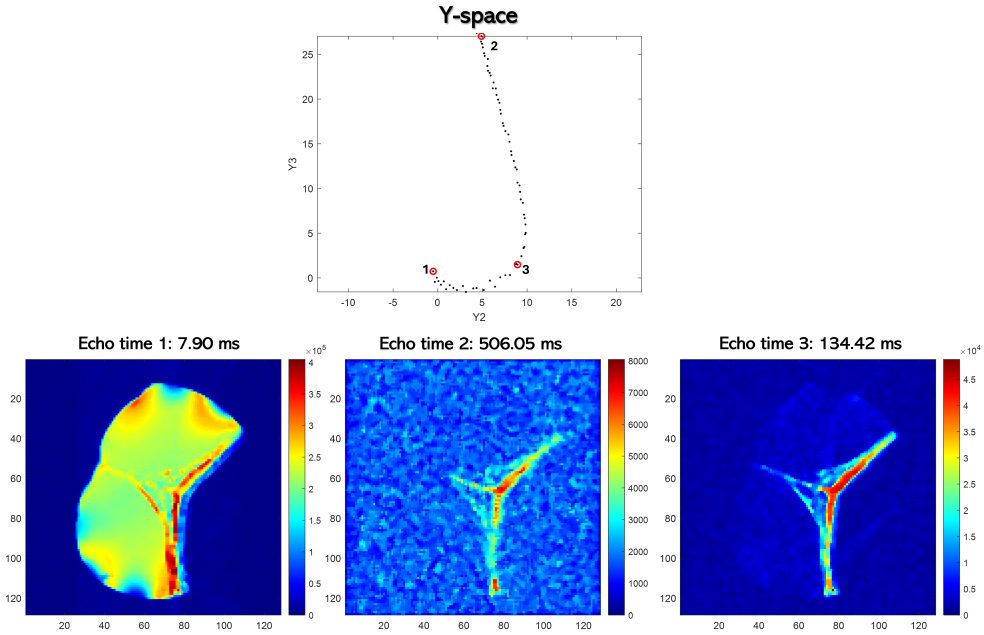

Figure A.15: MRI dataset: The top panel presents a 2D representation of the Y-space. The short echo time corresponds to the muscle region of the sample, while echo time 2 exhibits a noisy distribution. At an echo time of 134.42 ms, the signal corresponds to subcutaneous fat and the horizontal septum.

B. Supplementary material for Part III

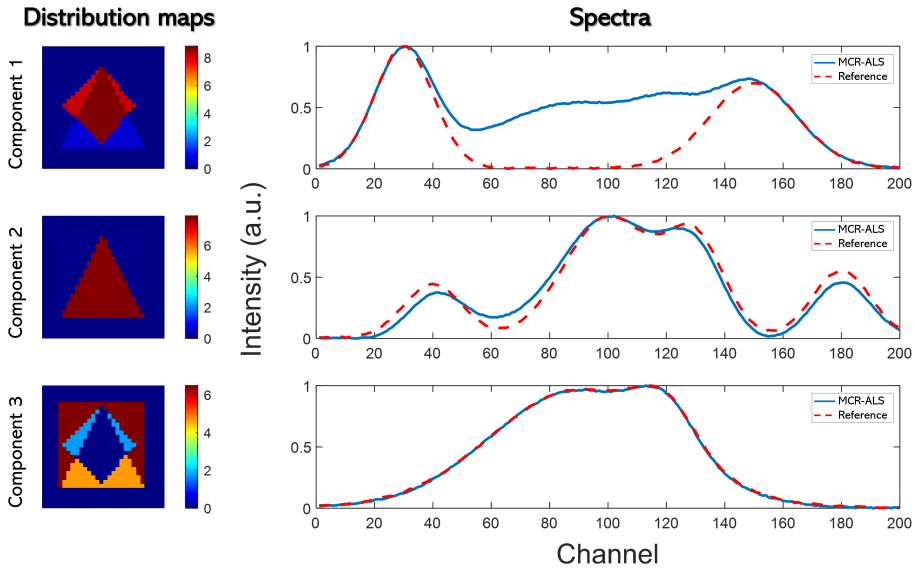

Figure B.1: MCR-ALS results for the simulated dataset. An unmixing model of 3 components and constrained with non-negativity resulted in suboptimal distribution profiles (left panel) for the first and third component, and suboptimal spectrum for the first component (right panel). Pure spectra are used as references and are plotted in red, while the resolved MCR-ALS spectra are in blue. Pure [190] and als [183] routine for the estimation of the initial pure profiles and optimization were used. Results after 500 iterations are provided. The algorithm could not converge, with a threshold fixed at 0.1%. The model has a LOF=0,004% and r^2=100%.

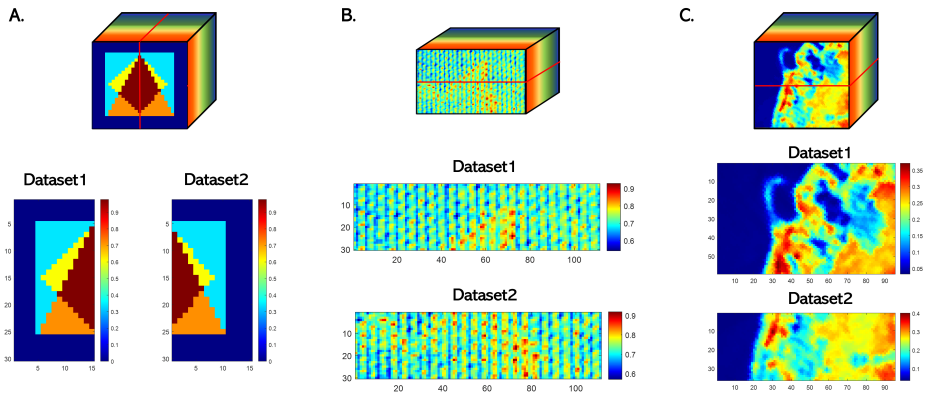

Figure B.2: Split-half of the datasets: **A.** simulated data, **B.** stained fabric, **C.** Samson dataset.

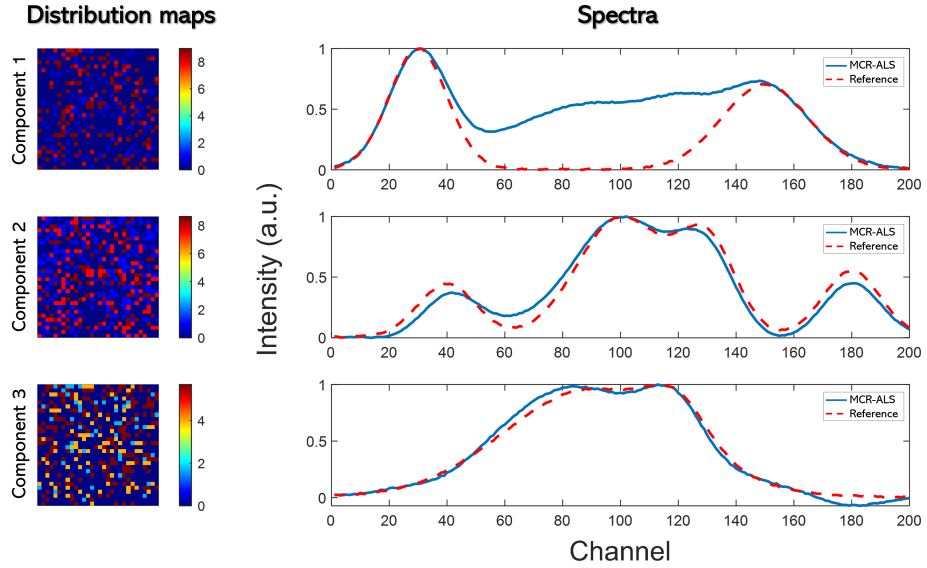

Figure B.3: MCR-ALS results for the simulated dataset with shuffled pixels. An unmixing model of 3 components and constrained with non-negativity resulted in exactly the same distribution profiles of the dataset where the spatial structure is preserved. Pure spectra are used as references and are plotted in red, while the resolved MCR-ALS spectra are in blue. Pure [190] and als [183] routine for the estimation of the initial pure profiles and optimization were used. Results after 92 iterations are provided. The algorithm converged with a threshold fixed at 0.1%. The model has LOF=0% and r^2=100%.

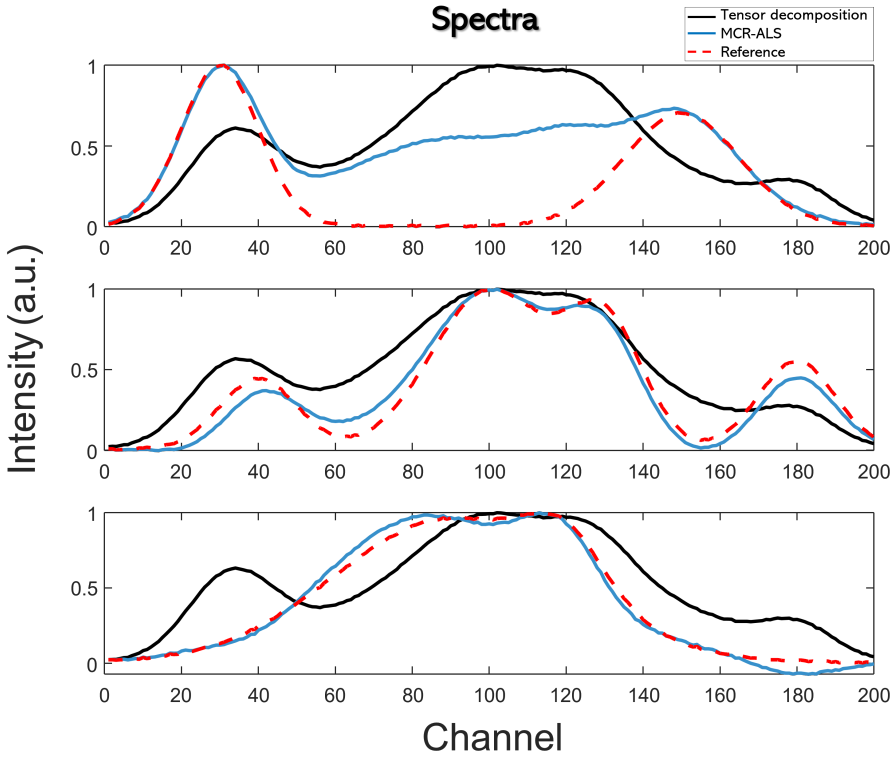

Figure B.4: The rank-(Lr,Lr,1) decomposition (in black) and MCR-ALS (in blue) are plotted together with the pure reference spectra for comparison (in red) of the model performances when the spatial structure distribution of the component is destroyed. It is clear that the tensor decomposition needs a structured spatial dimension to well approximate the data, while MCR-ALS does not consider it.

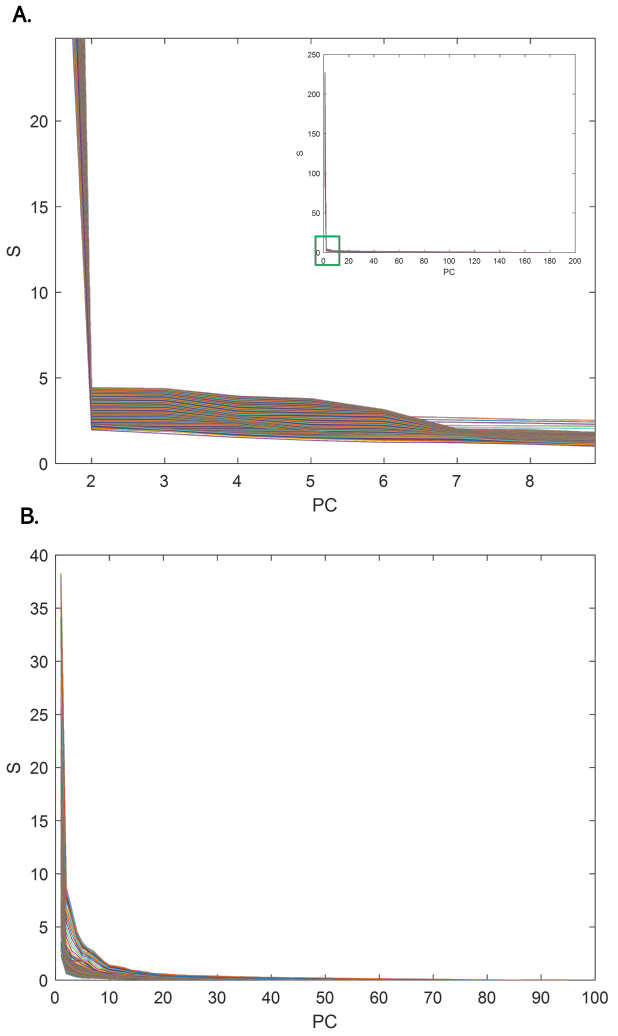

Figure B.5: Plot of the singular values vector S, obtained from the SVD analysis of each z^{th}-image of the hyperspectral data cube ($X \times Y \times Z$) as a function of the number of principal components. **A.** Stained fabric dataset.The plateau zone after the "elbow" in the curve indicates the range for L, that is between 2 and 7 principal components. This range was then selected to vary the value of L for each component. The image is an enlargement of the green region in the top right graph. **B.** Remote sensing dataset. The plateau zone after the "elbow" in the curve indicates the range for L, that is between 10 and 15 principal components.

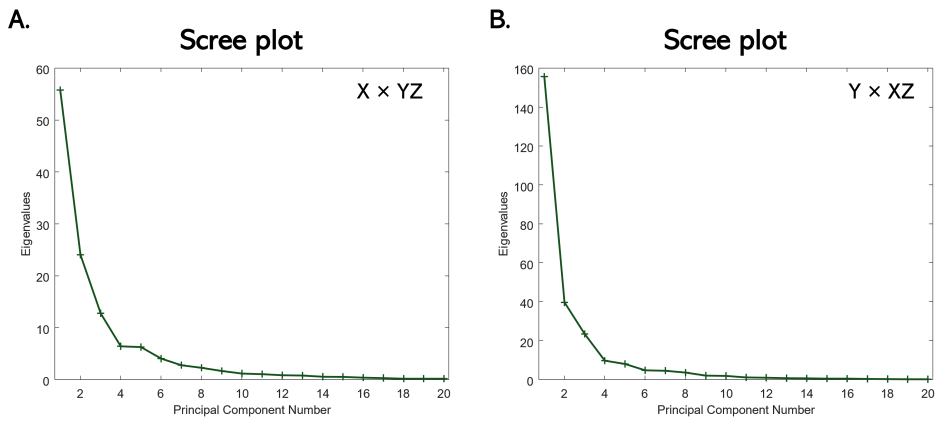

Figure B.6: Remote sensing dataset. Rank analysis of the first and second modes of the tensor. The data cube is unfolded and arranged as a matrix ($X \times YZ$) and as a matrix ($Y \times XZ$). The scree plots for the first and second modes are reported in **A**, for the first mode, and in **B**, for the second mode. Observing the two plots, 15 components can be considered for both modes. Therefore, a combination of L between 2 and 15 is chosen. This is in agreement with the methodology, which identified an optimal L_{opt} of 15.

C. Scientific contributions

Publications

Primary work

1. Olarini A., Cocchi M., Motto-Ros V., Duponchel L., Ruckebusch C. (2024). Exploratory analysis of hyperspectral imaging data. *Chemometrics and Intelligent Laboratory Systems*, 252, 105174.
   https://doi.org/10.1016/j.chemolab.2024.105174

2. Olarini A., Ruckebusch C., Duponchel L., Cocchi M. Exploring Block Term Decomposition for enhanced hyperspectral images analysis.
   Close to submission

Auxiliary work

1. Caponigro V., Salviati E., Olarini A., Campiglia P. (2024). Mass Spectrometry Imaging (MSI). In *Non-invasive and Non-destructive Methods for Food Integrity* (pp. 203-227). Springer, Cham.
   https://doi.org/10.1007/978-3-031-76465-3_10

2. Strani L., Farioli G., Cocchi M., Durante C., Olarini A., Pellacani S. (2024). Chemical Characterization and Temporal Variability of Pasta Condiment By-Products for Sustainable Waste Management. *Foods*, 13(18), 3018.
   https://doi.org/10.3390/foods13183018

3. Brustad N., Olarini A., Kim M., Chen L., Ali M., Wang T., Cohen A., Ernst M., Hougaard D., Schoos A.M., Stokholm J., Bønnelykke K., Lasky-Su J., Rasmussen M.A., Chawes B. (2023). Diet-associated vertically transferred metabolites and risk of asthma, allergy, eczema, and infections in early childhood. *Pediatric Allergy and Immunology*, 34(2), e13917.
   https://doi.org/10.1111/pai.13917

4. Olarini A., Ernst M., Gürdeniz G., Kim M., Brustad N., Bønnelykke K., Cohen A., Hougaard D., Lasky-Su J., Bisgaard H., Chawes B., Rasmussen, M.A. (2022). Vertical transfer of metabolites detectable from Newborn's dried blood spot samples using UPLC-MS: a chemometric study. *Metabolites*, 12(2), 94. https://doi.org/10.3390/metabo12020094

Contributions at national and international conferences

Oral contributions

1. Olarini A., Duponchel L., Ruckebusch C., Cocchi M. Chemometric approaches in hyperspectral imaging. *Workshop I giovani e la chimica in Abruzzo*, on-line, Jul 12th - 13th 2022.
2. Olarini A., Cocchi M., Duponchel L., Ruckebusch C. Using data geometry to highlight the necessity of bridging the gap between clustering and spectral unmixing in complex samples. *XI Colloquium ChemoMetricum Mediterraneum*, Padua-Italy, Jun 27th - 30th 2023.
3. Olarini A., Cocchi M., Duponchel L., Ruckebusch C. Data geometry to solve clustering and unmixing issues. *XXX Congresso della Divisione di Chimica Analitica*, Vasto-Italy, Sep 17th - 21st 2023.
4. Olarini A., Ruckebusch C., Duponchel L., Cocchi M. Exploring tensor-based decomposition methods for hyperspectral image analysis. *Workshop del Gruppo Italiano di Chemiometria*, Ravenna-Italy, May 27th - 29th 2024.

Poster contributions

1. Olarini A., Cocchi M., Duponchel L., Ruckebusch C. Hyperspectral imaging data: clustering or spectral unmixing? *XVIII Conference of Chemometrics in Analytical Chemistry*, Rome-Italy, Aug 29th – Sep 2nd 2022.
2. Olarini A., Duponchel L., Ruckebusch C., Cocchi M. Data geometry approach to extract essential information from hyperspectral imaging data. *XXII Giornata della Chimica dell'Emilia-Romagna 2023*, Parma-Italy, Dec 18th 2023.

3. Olarini A., Cocchi M., Duponchel L., Ruckebusch C. Block Term Decomposition as a promising method for hyperspectral image analysis. *X Simposio Nazionale di Spettroscopia NIR*, Turin-Italy, Dec 18th 2023.

4. Olarini A., Duponchel L., Ruckebusch C., Cocchi M. Exploring Block Term Decomposition for enhanced hyperspectral images analysis. *IX International Conference in Spectral Imaging*, Bilbao-Spain, Jul 6th – 10th 2024. **This contribution won one of the three Best Poster Awards at the IX Meeting of the International Association of Spectral Imaging (IASIM).**

5. Olarini A., Ruckebusch C., Duponchel L., Cocchi M. Exploring tensor-based decomposition methods for hyperspectral image analysis: a comparative study with spectral unmixing techniques. *XXVIII Congresso Nazionale della Società Chimica Italiana*, Milan-Italy, Aug 26th – 30th 2024.

6. Olarini A., Ruckebusch C., Duponchel L., Cocchi M. Block term decomposition: a breakthrough in hyperspectral image analysis. *XXIII Giornata della Chimica dell'Emilia-Romagna 2024*, Modena-Italy, Dec 19th 2024.

D. Publication: primary work 1

Alessandra Olarini, Marina Cocchi, Vincent Motto-Ros, Ludovic Duponchel, Cyril Ruckebusch

Exploratory analysis of hyperspectral imaging data

Chemometrics and Intelligent Laboratory Systems, 2024

https://doi.org/10.1016/j.chemolab.2024.105174

Chemometrics and Intelligent Laboratory Systems 252 (2024) 105174

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Exploratory analysis of hyperspectral imaging data

Alessandra Olarini [a,b,*], Marina Cocchi [b], Vincent Motto-Ros [c], Ludovic Duponchel [a], Cyril Ruckebusch [a]

[a] Université de Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, F-59000, Lille, France
[b] University of Modena and Reggio Emilia, Department of Chemical and Geological Sciences, Via Campi 103, 41125, Modena, Italy
[c] Université Claude Bernard Lyon 1, Institut Lumière Matière, CNRS, UMR 5306, Villeurbanne, 69622, France

ARTICLE INFO

Keywords:
Spectral imaging
Essential information
Clustering
Spectral unmixing
Raman
LIBS

ABSTRACT

Characterizing sample composition and visualizing the distribution of its chemical compounds is a prominent topic in various research and applied fields. Integrating spatial and spectral information, hyperspectral imaging (HSI) plays a pivotal role in this pursuit. While self-modelling curve resolution techniques, like multivariate curve resolution - alternating least squares (MCR-ALS), and clustering methods, such as K-means, are widely used for HSI data analysis, their effectiveness in complex scenarios, where the structure of the data deviates from the models' assumptions, deserves further investigation. The choice of a data analysis method is most often driven by research question at hand and prior knowledge of the sample. However, overlooking the structure of the investigated data, i.e. linearity, geometry, homogeneity, might lead to erroneous or biased results. Here, we propose an exploratory data analysis approach, based on the geometry of the data points cloud, to investigate the structure of HSI datasets and extract their main characteristics, providing insight into the results obtained by the above-mentioned methods. We employ the principle of essential information to extract archetype (most linearly dissimilar) spectra and archetype single-wavelength images. These spectra and images are then discussed and contrasted with MCR-ALS and K-means clustering results. Two datasets with varying characteristics and complexities were investigated: a powder mixture analyzed with Raman spectroscopy and a mineral sample analyzed with Laser Induced Breakdown Spectroscopy (LIBS). We show that the proposed approach enables to summarize the main characteristics of hyperspectral imaging data and provides a more accurate understanding of the results obtained by traditional data modelling methods, driving the choice of the most suitable one.

1. Introduction

Understanding the composition and distribution of the chemical compounds within a sample stands as a priority in many research and applicative fields [1,2]. In this respect, hyperspectral imaging (HSI) is a key analytical tool as it combines spatial information about the distribution of the chemicals across the image pixels with the corresponding spectral signatures. HSI finds applications throughout a wide range of scientific disciplines, spanning from remote sensing to macro- and micro-imaging [3–5]. The information provided by HSI is usually organized in a third-order tensor with two spatial dimensions and a spectral one. To identify individual sources of spectral variation and determine their respective contribution to the mixed signal in each pixel, self-modelling curve resolution techniques are among the most popular approaches [6–8]. One of the principal algorithms in this category is multivariate curve resolution - alternating least squares (MCR-ALS) [9]. Based on the matrix formulation of Beer-Lambert's law, the results of the data decomposition provided by MCR-ALS can be interpreted as concentration distribution maps (spatial distributions) and spectral signatures of the individual components of the spectral mixture. The MCR-ALS algorithm minimizes the difference between the reconstructed data and the original data, by iteratively optimizing the concentration profiles and the spectra profile in each least square iteration, until convergence is achieved. Constraints can be imposed on the components profiles to enforce physically or chemically meaningful solutions. Imposing constraint would also contribute to reduce rotational ambiguity which, except in very specific conditions, remains inevitable [10,11]. MCR-ALS initially found extensive application in spectroscopic data analysis, particularly in fields such as analytical chemistry and process analysis. Later, its utility expanded to include image analysis

* Corresponding author. Université de Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, F-59000, Lille, France.
E-mail address: alessandra.olarini@unimore.it (A. Olarini).

https://doi.org/10.1016/j.chemolab.2024.105174
Received 17 May 2024; Received in revised form 4 July 2024; Accepted 4 July 2024
Available online 9 July 2024

and a broad range of other applications [12–14]. However, in complex scenarios, the MCR bilinear decomposition may not fully capture the complexity of the physics/chemistry underlying the analyzed data, due to e.g. interactions between individual species. This and other effects might result in a deviation from the ideal linear mixture model [15].

Another approach for the analysis of HSI datasets can be found in the framework of clustering techniques, aiming at grouping pixels based on their spectral similarity (hierarchical and partitional clustering) [16] or on density criteria [17], and highlighting spatial patterns in the image. Unlike spectral unmixing which aim at identifying the contributions of the individual mixture components for each pixel, clustering methods, such as K-means [18] assigns each pixel to one cluster only, characterized by a centroid, serving as a prototype of the cluster, and results should be interpreted as spectral pixel classification or image segmentation approaches. The determination of the number of clusters, as well the algorithm initialization step's requiring a random selection of the mean spectrum of each cluster, represent the major challenges for this method. Even though solutions trying to overcome these issues, such as the use of indices and replicates, have been proposed over the years, these questions remain open in the field [19,20].

In practice, although the assumptions and goals of MCR and K-means approaches are different, both can provide complementary results when applied to HSI data, the first can be used for data decomposition and the second for clustering, to obtain interpretable factors or clusters albeit in different ways [21]. Clustering techniques have been used in conjunction with other multivariate analysis methods [22,23], for instance, as a powerful tool for examining data homogeneity, in terms of chemical composition or properties, together with Principal Component Analysis (PCA) [24,25]. Other works have explored the use of clustering as a constraint in unmixing methods such as MCR-ALS or vertex component analysis (VCA) for complex samples [26,27].

While MCR and clustering methods are powerful tools, the first step of any multivariate data analysis should be exploratory [28,29], summarizing the main characteristics of the investigated data set, and this is even more the case for HSI data. Users should be aware of the specific characteristics of the structure of the data and carefully consider the methods' assumptions and limitations in order to ensure the reliability of their interpretation. To this aim, the extraction of the essential information (EI) can reveal very useful as it is not based on data variance but on the geometry of the data points cloud [15,30–34]. Essential information consists of archetype points that outline the convex hull of the data points cloud in a normalized abstract data space. Recent studies have highlighted the potential and usefulness of identifying essential rows and columns of a data matrix [32,35–37]. A key aspect is that the corresponding samples (spectral pixels) and variables (single--wavelength images) contain all the information needed to reproduce the measured data [38].

This paper introduces an exploratory approach for analysing HSI data of complex samples, especially for scenarios where the results obtained from MCR-ALS and K-means are difficult to obtain and interpret. Through identification of the archetype points of the data cloud, we aim to extract some of the most linearly dissimilar spectra and single-wavelength images measured. A powder mixture analyzed with Raman spectroscopy and a mineral sample, characterized mostly by pyrite, analyzed by Laser Induced Breakdown Spectroscopy (LIBS) were investigated. The first dataset is a perfect example for using unmixing approach, the second is a dataset where both the unmixing and clustering approaches could be used considering the analysis task [39]. However, this dataset has an intrinsic complication due to the weathering products of pyrite in LIBS technique and the lack of selectivity, making the hyperspectral imaging data at hand deviating from the ideal model underlying both MCR and clustering techniques. Comparison to the result obtained by applying MCR-ALS [40,41] and K-means [18] is also provided and discussed. We argue that this approach is very useful to extract the main characteristics of a hyperspectral imaging dataset and provide accurate information to be used for spectral unmixing and clustering. Moreover, it has been observed that spatial distributions and spectral signatures extracted by this approach are not always retrievable using conventional methods like MCR-ALS and K-means.

2. Materials and methods

2.1. Datasets

2.1.1. Raman powder dataset

Powders of three salts i.e. calcium carbonate ($CaCO_3$), sodium nitrate ($NaNO_3$) and sodium sulfate (Na_2SO_4) were mixed and pressed in a tablet, obtaining a three-component system. Sample preparation and Raman imaging acquisition features were described by Coic et al. in Ref. [31]. The sample was investigated in the range 901.2 cm^{-1} to 1280.5 cm^{-1} with a spectral resolution of 1.11 cm^{-1}. A 101 × 101 pixels image was mapped using point-by-point raster-scanning mode with a 1 μm step between successive acquisitions. The dataset corresponds to a third-order tensor of dimensions 101 × 101 × 341, which was subsequently analyzed without any spectral pretreatment.

2.1.2. LIBS mineral dataset

A thin section of a mineral sample from the Nishapur turquoise deposit (Iran) was prepared and polished for LIBS imaging. Sample preparation, equipment and LIBS acquisition are detailed in Moncayo et al. in Ref. [42]. The sample is constituted by three main mineral phases: pyrite FeS_2, silica (mainly quartz) SiO_2 and turquoise $CuAl_6(PO_4)_4(OH)_8{\cdot}4H_2O$. The LIBS image was recorded considering a 15 μm step between successive acquisitions over 2048 spectral channels in the spectral range from 250 to 330 nm. From the full acquired dataset, only a region of interest has been selected, resulting in a third-order tensor of dimensions 300 × 300 × 1930, which was then analyzed without spectral preprocessing.

2.2. Data analysis

The data analysis methodologies employed in subsequent sections of the paper are here introduced. Section 2.2.1 provides a detailed description of the proposed data analysis approach, which investigates the geometry of the data point cloud resulting from a singular value decomposition (SVD). Sections 2.2.2 and 2.2.3 describe well-known chemometric methods: Multivariate Curve Resolution - Alternating Least Squares (MCR-ALS) and K-means clustering, respectively. These methods are employed for data analysis to compare with the proposed approach.

2.2.1. Selection of the most relevant archetype points for exploratory analysis

The HSI tensor is first unfolded into a matrix **D** of dimensions (*n*, *p*) with rows corresponding to pixels and columns corresponding to spectral channels (unfolded single-wavelength images). The matrix **D** is then decomposed by SVD [43] according to Eq. (1):

$$\mathbf{D} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

where **U** of dimensions (*n*, *k*) is the matrix containing the left singular vectors, **S** of dimensions (*k*, *k*) is the diagonal matrix of singular values and $\mathbf{V}^{\mathrm{T}}$ of dimensions (*k*, *p*) is the matrix of the right singular vectors transposed, *k* is the number of factors of the decomposition and **E** of dimensions (*n*, *p*) the error matrix. The matrices **X** and **Y** of dimensions (*n*, *k*) and (*p*, *k*) are calculated as in Eqs. (2) and (3), respectively, and contain the coordinates of the data points in the column- and row-vector space, respectively:

$$\mathbf{X} = \mathbf{U} \times \mathbf{S} \tag{2}$$

$$\mathbf{Y} = \mathbf{V} \times \mathbf{S} \tag{3}$$

All column vectors of **X** (resp. **Y**) are then normalized to constant projection on the first column vector of **X** (resp. **Y**) to enforce convexity of the data points cloud [44,45]. The archetypes of the data points cloud of **X** and **Y** can be identified by computing the corresponding convex hulls, as in Eqs. (4) and (5) [35]. They correspond to the most linearly dissimilar spectral pixels and single-wavelength images, respectively.

Convex hulls of matrices **X** and **Y** are computed:

$$conv(\mathbf{X}) = \left\{ \mathbf{x} \in \mathbf{X} \mid \begin{matrix} \sum \alpha \mathbf{x}; \alpha \geq 0 \\ and \sum \alpha = 1 \end{matrix} \right\} \tag{4}$$

$$conv(\mathbf{Y}) = \left\{ \mathbf{y} \in \mathbf{Y} \mid \begin{matrix} \sum \beta \mathbf{y}; \beta \geq 0 \\ and \sum \beta = 1 \end{matrix} \right\} \tag{5}$$

where α and β are coefficients of the convex linear combinations. The number of components to consider into convex hull calculation is left to the user [44,45]. Analogous to exploratory PCA, inspection of the information carried by the most dissimilar spectra/images can guide the selection.

The most relevant archetype points are then selected by visual inspection and the corresponding (essential) spectra and (essential) single-wavelength images extracted, as illustrated in Fig. 1.

2.2.2. Multivariate curve resolution - alternating least squares (MCR-ALS)

The MCR-ALS algorithm provides pure spectral signature of the components and their corresponding component distribution maps, as in Eq. (6):

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{T} + \mathbf{E} \tag{6}$$

where **D** of dimensions (n, p) is the unfolded tensor, **C** (n, c) is the pure concentration matrix, with component distribution maps of the c components as columns, and $\mathbf{S}^{T}$ of dimension (c, p) is the matrix of pure spectra, with spectra profiles of the c components as rows. **E** (n, p) contains the variation unexplained by the MCR model. To solve Eq. (6), alternating least-squares (ALS) optimization [46] is used as well-established approach, and both concentration and spectra profiles are constrained with non-negativity. The first step of the optimization requires spectra or distribution profiles that will be implemented and optimized during the iteration process. Simple to use interactive self-modelling mixture analysis (SIMPLISMA) [47] was used throughout this work to calculate initial spectral estimates. The optimization procedure stops when the convergence criterium is reached, expressed as a threshold (0.1 %) based on the relative difference of the lack of fit (LOF) during consecutive iterations. The LOF and the explained variance, defined in Eqs. (7) and (8), are used as parameters to evaluate the quality of the MCR model:

$$\mathrm{LOF} = 100 \times \sqrt{\frac{\sum e^2}{\sum d^2}} \tag{7}$$

$$\mathrm{r}^2 = 100 \times \left(1 - \frac{\sum e^2}{\sum d^2}\right) \tag{8}$$

where e and d are elements of **D** and **E** respectively.

2.2.3. K-means clustering

As one of the most used partitioning clustering techniques in image analysis, the K-means algorithm can be applied to **D**. In K-means, once the number of clusters is defined (c), the first iteration selects c clusters randomly, then at each iteration samples are reassigned to minimize the sum of point-to-centroid distances, summed over all c clusters (*sumd*). The algorithm stops when clusters assignments do not change, or the maximum number of iterations is reached. As distance measure, the Pearson correlation distance, defined as one minus the correlation coefficient calculated between the point and centroid spectra, has been used [48]. In order to stabilize the results, 50 replicate runs of K-means clustering are performed for each analysis and the run with lowest *sumd* has been selected; the number of iterations was set to 200. Silhouette [49] and Pakhira-Bandyopadhyay-Maulik (PBM) [50] indices were used to evaluate the optimal number of clusters. These were compared with the number of most informative pixels/spectral wavelengths suggested by the proposed exploratory approach. Here, the explicit use of c to denote both the number of components (MCR-ALS) and clusters (K-means), is adopted because the results, for sake of comparison, are presented considering the same number of components and clusters.

2.3. Software

All computations were performed using MATLAB© 2022a (MathWorks Massachusetts, USA). For the cluster analyses the K-means function of the Statistical and Machine Learning Toolbox was used, with the addition of the MATLAB© Parallel Computing Toolbox to improve

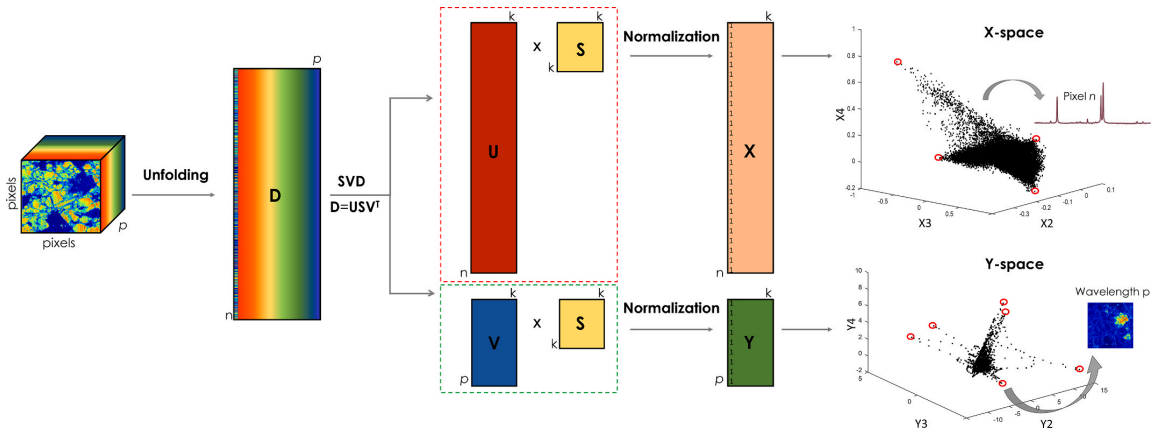

Fig. 1. Graphical representation of the data exploratory approach: third-order tensor is unfolded into a matrix (**D**) and decomposed through SVD algorithm (**D** = **USV**^{T}). Matrices **X** and **Y** are calculated and normalized [45], resulting in a unit first column vector **X1** (resp. **Y1**) to which all other column vectors of **X** (resp. **Y**) are orthogonal. Convex hulls of essential spectra and essential variables are computed for **X** and **Y**, and the most relevant archetype points are identified by visual inspection (red circles). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

the speed of the algorithm. *Pure* and *als* routines from Tauler and De Juan (2003) were used for the MCR-ALS analysis and *convhulln* is the built in MATLAB© function used to compute convex hulls.

3. Results and discussion

For each dataset, the information extracted from exploratory analysis is put in perspective of the results obtained by applying both MCR-ALS and K-means.

3.1. Raman powder dataset

The three-component Raman hyperspectral imaging dataset, described in 2.1.1, exhibits well-defined characteristics: (i) a clear distribution of the salts building up the chemical composition of the sample (Fig. 2A), (ii) good signal-to-noise ratio data (Fig. 2B) and (iii) selective spectral regions (Fig. 2C).

Fig. 3A provides a representation of the (X2, X3) data points clouds, where the number 2 and 3 refer to the second and third column of the normalized matrix **X**. As expected, the observed data structure corresponds to a triangular geometry (as would be obtained for simplex data), the 3 vertices being expected to correspond to the pure compounds [45]. Similarly, the data points representation of the second and third column of the normalized matrix **Y**, (Y2, Y3), (Fig. 4A), enable to identify vertices pointing at clearly distinct directions.

Convex-hull computation provided 14 archetypes points in the (X2, X3) space corresponding to essential spectra and 3 archetype points in the (Y2, Y3) space corresponding to essential single-wavelength images (black circles in Fig. 3A and 4A). Considering that the number of components is known, 3 archetype points were selected in both sub-spaces (filled green circles in Fig. 3A and 4A, respectively), which are expected to correspond to the purest spectral pixels and most selective wavelengths measured (see Fig. 3B and 4B, respectively). The provided spectral and image information can be readily interpreted for this simple data set (1070 cm^{-1} maximum selective peak for $NaNO_3$, 1090 cm^{-1} maximum selective peak for $CaCO_3$, 996 cm^{-1} maximum selective peak for Na_2SO_4). For the sake of comparison, the results obtained by SIMPLISMA are provided (Fig. S1 in Supplementary Material).

Fig. 5 shows the results obtained for a three-component MCR-ALS model (LOF = 10 %, r^2 = 99 %) and for the application of K-means considering 3 clusters. The selection of the number of clusters was set as 3 according to the mixture composition. For each cluster the class assignation vector has been refolded in the original image dimensions and shown with the pixels recognized as cluster member coloured in brown (Fig. 5B third column). For the sake of comparison, the results obtained from the previous archetype identification are also reported (Fig. 5). The similarity between the essential spectra and essential single-wavelength images obtained from our approach and the spectra and component distribution maps obtained applying MCR-ALS is striking.

Focusing on the spectra provided in Fig. 5A, the ones shown for K-means correspond to "centroid" spectra and are, as expected, not the pure ones, though in quite good agreement. It is worth noting that the centroid spectrum corresponding to the $NaNO_3$ salt is more similar to the pure one than for the 2 other salts. This can be explained by considering the density of points for each of the 3 clusters modelled by K-means (see Fig. S2 in Supplementary Material). As for K-means, the maps (Fig. 5B) obtained for each cluster are also very comparable (considering that the information is segmented).

This dataset was introduced to clearly show that in cases in which we have prior information on the number of components, high spectral and spatial selectivity, as well as a high number of pure pixels, MCR-ALS and K-means solutions are very comparable, with selection of the method depending on specific analysis goals. Also, the information retrieved with the 2 approaches can be readily extracted from the analysis of the geometry of the data.

3.2. LIBS mineral dataset

The mean image and the LIBS spectra obtained for the mineral sample are shown in Fig. 6A and B, respectively. In Fig. 6A it is important to note that the pixel size is 15 μm. Considering the scale of mineral phases, the presence of many pure spectral pixels is, therefore, not expected. Fig. 6B highlights data characterized by low spectral selectivity. An additional complexity of this sample arises from its composition, which includes iron. Iron has numerous emission lines across the entire spectral range. Additionally, pyrite typically exists in various oxidative forms [39,42,51], and the iron ions within pyrite can easily exchange with copper or aluminium ions present in turquoise [52]. In fact, this kind of rocks are often referred to as "solid mixtures" [53]. Furthermore, within quartz, the predominant silica phase in this sample, aluminium impurities are quite common, while iron inclusions are also possible, albeit less frequent [54]. All these peculiarities translate into a very challenging LIBS HSI dataset to analyse and investigate with classical chemometric tool. Indeed, this scenario is not ideal for approaches such as MCR-ALS as pure pixels may not be present, spectral selectivity is low and different phases with very similar spatial distribution are present. Similarly, K-means clustering is not ideal as it may have difficulty assigning different minority phases to distinct clusters, as pixels may belong to multiple clusters due to low spectral selectivity.

The geometry of the data in the (X2, X3) and (Y2, Y3) spaces is illustrated in Fig. 7A and B, respectively. While more complex than the geometry observed in the previous example, the observed data points clouds exhibit some degree of structure. However, determining the appropriate number of components to consider is not straightforward given the absence of clear a priori information with this dataset. Convex-

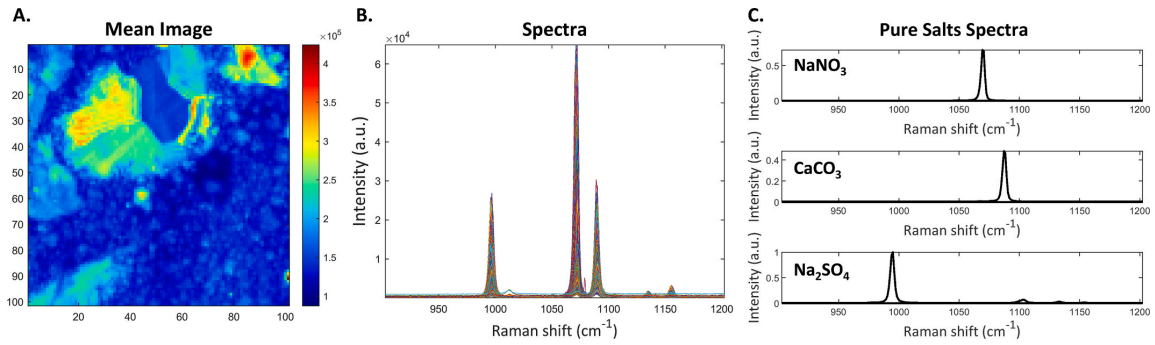

Fig. 2. Raman powder dataset: mean image (A), spectra (B) and spectra of pure salts (C).

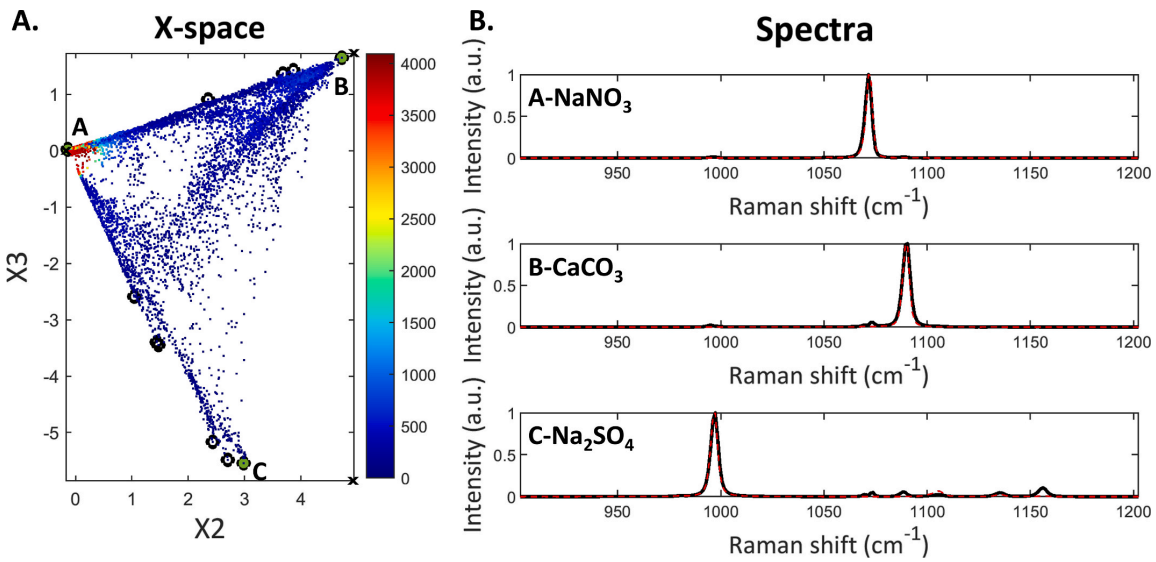

Fig. 3. 2-D representation of the X-space (A) colour-coded by point density. Black circles mark the archetypes points (some are close and result overlapped in the plot) at the vertices of the convex hull computed in the (X2, X3) normalized space. Filled green circles are the selected points and black crosses are the projection of the pure reference spectra in the (X2, X3) normalized space. In panel B, the spectra corresponding to the green points (black line) with overlapped the pure spectrum (red line) of the corresponding component. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

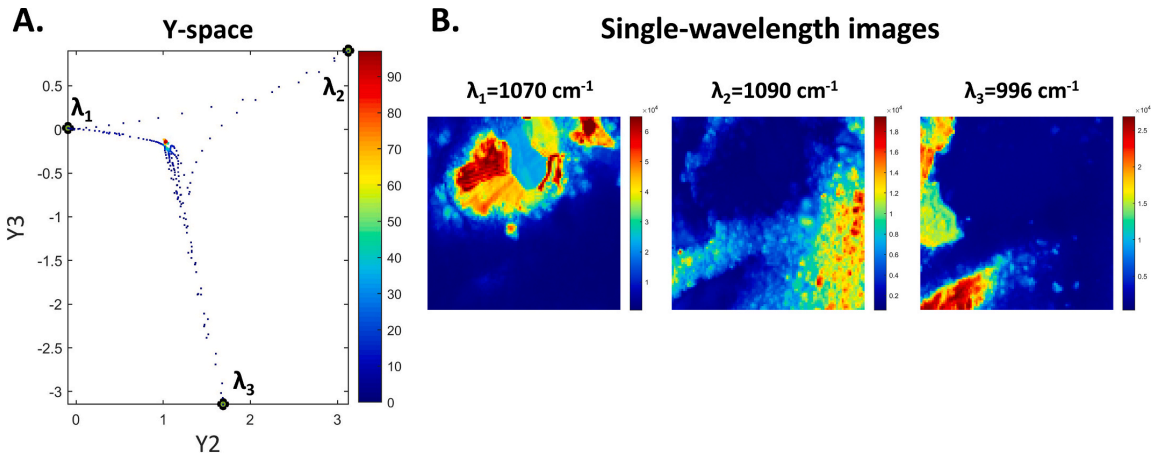

Fig. 4. 2-D representation of the Y-space colour-coded by point density (A). Black circles mark the archetypes points of the convex hull computed in the (Y2, Y3) normalized space. Filled green circles are the points identified looking at the structure of the data. In panel B, the essential single-wavelength images corresponding to the 3 identified selective wavelengths. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

hull computation provided 19 archetype points in (X2, X3), and 13 archetype points in (Y2, Y3), (black circles in Fig. 7A and B, respectively). Considering the geometry of the data observed in Figs. 7A and 6 archetype points were selected (filled green circles) and the corresponding essential spectra are shown. However, considering Fig. 7B–is clear that some relevant points, corresponding to clear directions, were not identified as archetypes, as they are not found at vertices of the data points cloud in the two-dimensional Y-space. It should be noted that by applying convex hull calculation to a six-dimensional **Y** matrix (see Supplementary Material Fig. S3), these points could be selected, but the total number of archetypes would be very large. This is not really needed, though, since they can be manually pointed out in the (Y2, Y3) plot, resulting in the extraction of 7 essential single-variable images.

The spectra corresponding to points labelled A, C and D in Fig. 7A correspond to the main mineral phases of pyrite, silica, and turquoise respectively. Spectrum B shows spectral features corresponding to a phase where silica has iron inclusions, somehow in between the pyrite and the main phase of silica. Spectrum F features another pyrite phase, different from the one observed in A. Lastly, the spectrum corresponding to pixel E characterizes an intermediate phase between turquoise and pyrite, where iron and mainly aluminium exchanges occur. The spectral regions used for the identification are highlighted in blue referring to

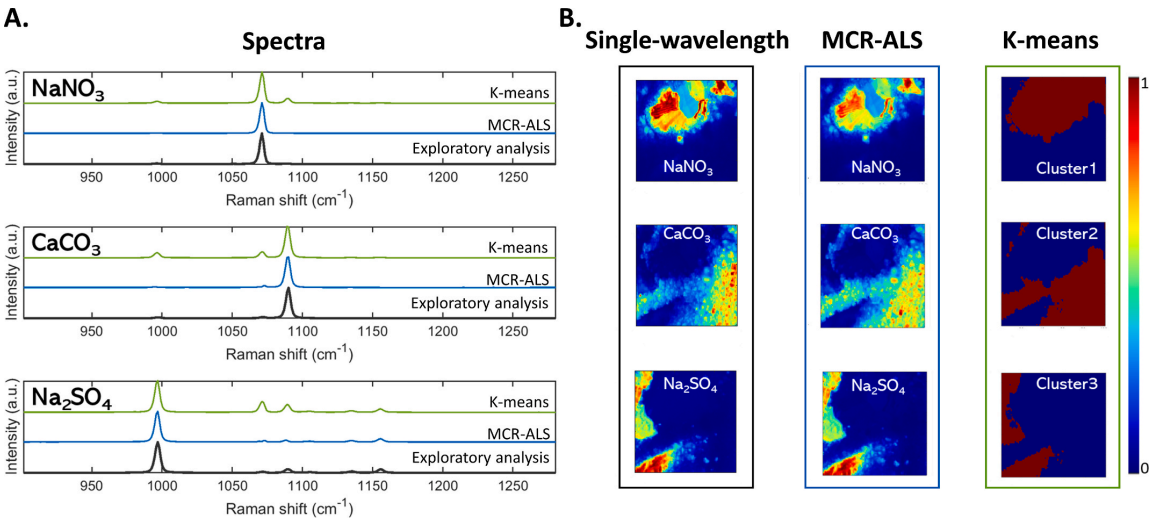

Fig. 5. Panel A shows the spectra for the purest pixels obtained by the exploratory analysis (black line), the purest components resolved spectra by MCR-ALS (blue) and the K-means centroids spectra (green). Centroids spectra are calculated as the average of the spectra of all the pixels belonging to a given cluster. For sake of clarity an arbitrary vertical offset was added to the MCR-ALS and K-means results. Panel B shows the single-wavelength images extracted with the exploratory approach, the concentration distribution maps retrieved by MCR-ALS and the clustering maps obtained by K-means clustering. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

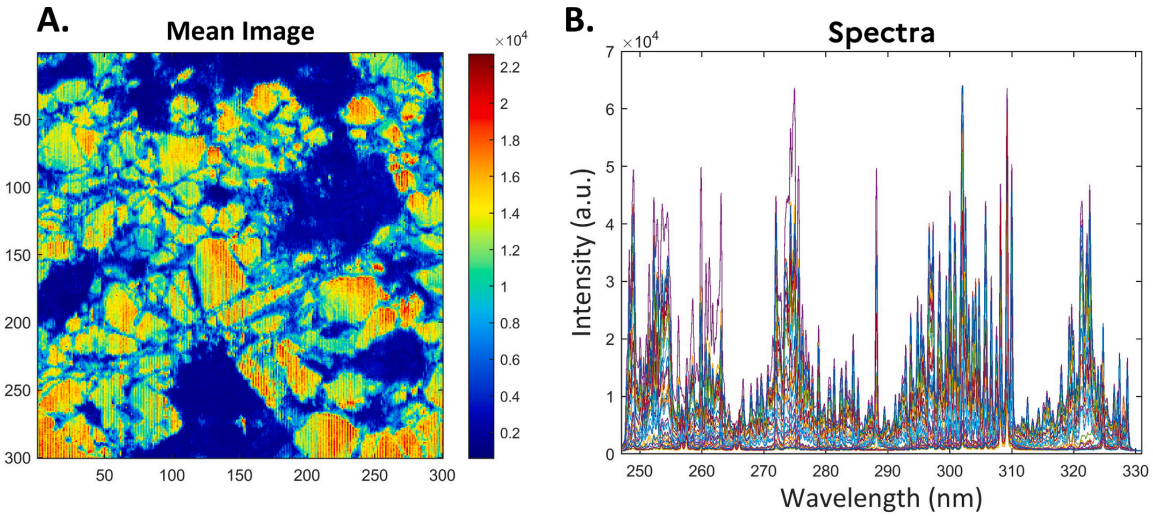

Fig. 6. Mean image (A) and overlapped spectra (B) of the mineral sample dataset.

Kurucz LIBS database [55], following the assessed procedure of Moncayo et al. [42]. These spectra are the purest spectra identified and can be interpreted as such without further analysis of the data.

In the same way, the essential single-wavelength images extracted correspond to the information obtained at the most selective wavelengths. Images λ_1, λ_3 and λ_7 which are linked to point A, C and D in Fig. 7B, respectively describe pyrite, silica and turquoise. Image λ_2, corresponding to point B, describes a situation where both pyrite and silica are present and image λ_4 shows the distribution of pyrite, turquoise and silica. It is worth noting that when looking at image λ_5, which does not show any correspondence in the (X2, X3) plot, it could be hypothesised that it represents a mineral phase where both silica and turquoise are present. In fact, it lies between image λ_3 and λ_7 in the (Y2, Y3) plot. Image λ_6, linked to point F, is identified as another form of pyrite. In addition, it can be noticed that in the right area of both the (X2, X3) and (Y2, Y3) plots, there is a higher density of points (either pixels or spectral wavelengths). Since points A and F correspond to spectra that are associated to pyrite phases, it can be concluded that pyrite is identified as the major phase in this mineral sample. For comparison purposes, the results obtained by SIMPLISMA are also provided (Fig. S4 in Supplementary Material). A six-component MCR-ALS model could then be fitted (LOF = 3 %, r^2 = 99 %) and the results are shown in Fig. 8.

The spectra of the first 2 components of the MCR-ALS model are identified as silica and turquoise phases, respectively. The

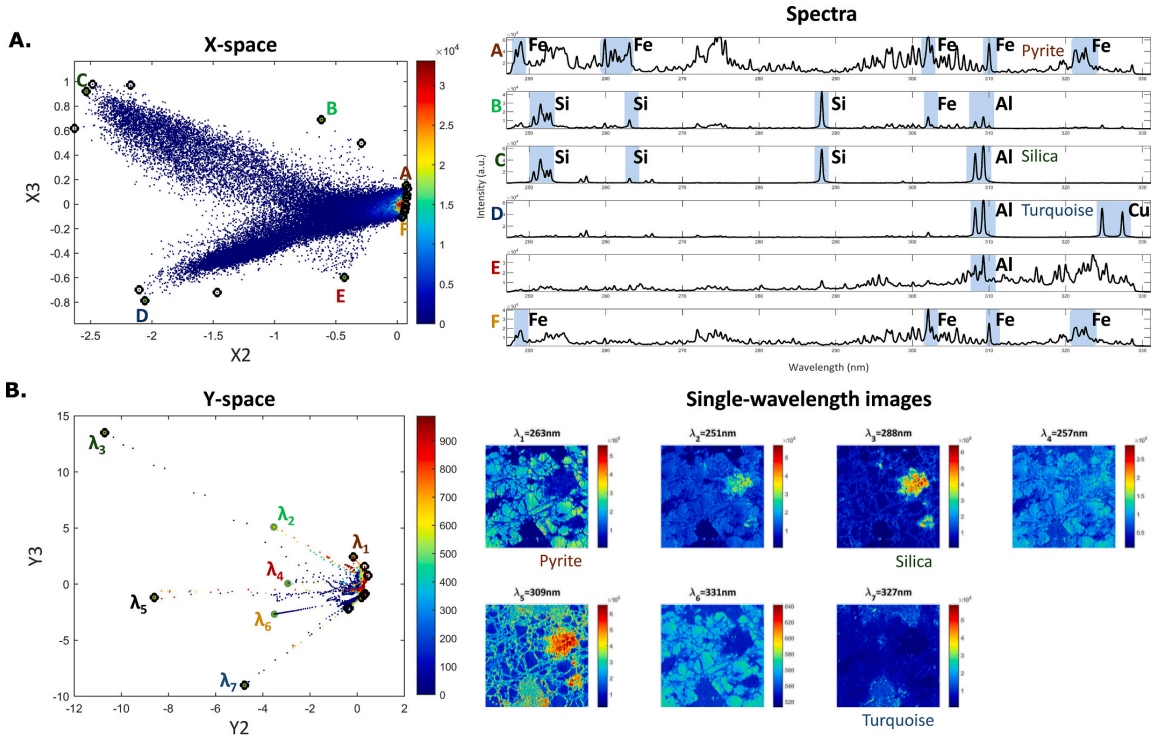

Fig. 7. (A) 2-D representation of the X-space of the mineral sample dataset, colour-coded by point density. Black circles mark the archetypes points at the vertices of the convex hull computed in the (X2, X3) normalized space. Letters and filled green circles represent the selected points, while the corresponding spectra are shown in the right panel. (B) 2-D representation of the Y-space, colour-coded by point density. Black circles mark the archetypes points of the convex hull computed in the (Y2, Y3) normalized space. Filled green circles represent the selected wavelengths, the corresponding refolded images are shown in the right panel. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

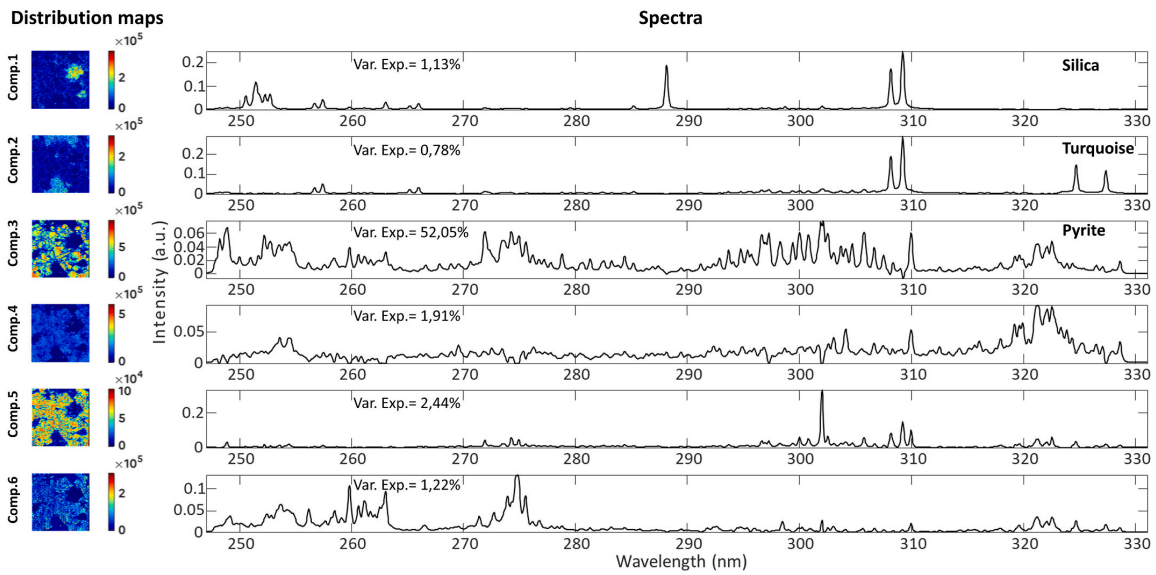

Fig. 8. MCR-ALS solutions for the mineral sample dataset. Refolded concentration profiles (left) and resolved spectra (right) are shown for each of the 6 components with the corresponding data variance.

corresponding concentration distribution are in agreement with the images retrieved by the exploratory analysis. Pyrite is identified primarily in the spectral profile of the third component. However, the spectral profiles observed in the remaining 3 components suggest the potential occurrence of different pyrite phases characterized by ion exchanges, which present challenges for interpretation. This is further complicated by the fact that the corresponding concentration distribution maps show very similar distributions. The third MCR component is the one explaining most of the data variance (52 %) confirming that pyrite is the major phase. By contrast the variance explained by other components, is very low, less than 3 % for silica, turquoise and the other phases of pyrite.

For the sake of comparison, a K-means model was computed setting the number of clusters to 6. The results are shown in Fig. 9. Cluster 1, 2 and 3 can be associated to silica, turquoise and pyrite phases, respectively. The clustering maps for clusters 4 to 6 reveal distributions spanning the boundaries between pyrite and the phases described by the first 2 clusters. The centroid spectra of these clusters are challenging to interpret, suggesting possible exchanges between iron and aluminium.

MCR-ALS and K-means clustering provide complementary information that leads to a more complete understanding of the sample. The proposed methodology allows for observing the potential complexity of data exploration prior to implementing MCR and/or K-means. It is important to note that the exploratory approach not only provides the same information as the one obtained from data modelling, but also enables to extract the spectral and spatial features related to the presence of minority components resulting from ion exchanges between the main mineral phases. The results obtained for silica, turquoise, and pyrite are comparable. The centroid spectra obtained by K-means and identified as pyrite is very comparable with the one extracted exploring the X-space and MCR-ALS, again because of the high number of pure pixels in that cluster (being pyrite the major phase, high number of pixels correspond only to pyrite). The concentration maps of 3 of the MCR-ALS components and the clustering maps of 3 of the clusters show the same distribution observed in the purest images extracted from the Y-space, while the other differ and as discussed above, are not easily interpretable.

Overall, we may remark that in this challenging scenario, that deviates from the ideal model underlying both MCR and clustering techniques, exploratory analysis driven by archetypes identification can provide insight into the number of components (when going for an unmixing approach) or clusters (when using clustering) to select. In fact, traditional methods such as eigenvalues, scree plots, and cluster indices may not provide unambiguous answers, as illustrated in Fig. S5 in the Supplementary Material. The exploratory approach employed in this study offers notable advantages, particularly in the extraction of spectra and images without the need for complex modelling. Also, convex hulls need to be calculated for more than 2 components, in order to retrieve the archetype points for each direction in the Y-space. These findings emphasize the feasibility and efficiency of our methodology in obtaining informative data without excessive computational load.

4. Conclusion

Understanding the structure of the data is a key step in the data analysis workflow of any application. In particular, exploring HSI datasets, because of their nature and dimensionality, is nontrivial. An exploratory approach, like the one proposed in this work, demonstrate to be able to guide extracting the useful information encrypted in the spectral image of complex samples and furnishing a comprehensive understanding of the investigated system.

Two different datasets were analyzed in this work by the exploratory approach and compared with the conventional methods of two widely used approaches in spectral image analysis: spectral unmixing (MCR-ALS) and clustering (K-means), with the aim of envisioning their applicability domain. The shared information, among all methods, in terms of distribution and spectral signature of retrieved common components, concerned major phases and/or the one with selective spectral profile. In cases, where the application of very well-known methodologies revealed its limits, looking at the geometry of the data resulted in an extremely easy and fast way to have better and more complete insights, with respect to the MCR-ALS and/or K-means ones. The analysis of the structure of the data could be considered, as any exploratory tool, as preliminary to allow a more rational choice of the next steps of data analysis and also to help solve all the cases of limitations for the two methods, such as the choice of the number of components and clusters,

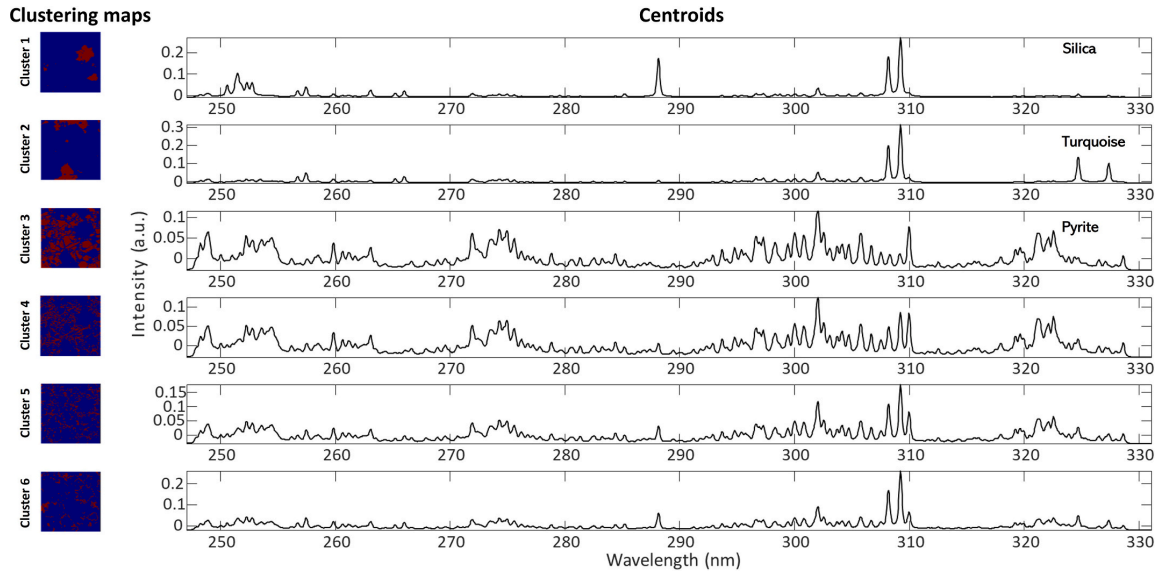

Fig. 9. K-means results for the mineral sample dataset. Cluster membership maps and the mean spectra are shown for each of the 6 clusters.

and in the retrieval and identification of the purest species as well as minor components.

This exploratory approach may have limitations when the data present a quite uniform distribution with no clear structures, thus rendering difficult finding the archetype points. However, to the best of our knowledge, applying appropriate spectral pre-processing could remove those effects, such as baseline, scatter, etc., that go into making the data less geometrically structured. In this way, a change in the "data shape" can be obtained making this approach therefore applicable. Furthermore, while automation of this process could be considered, it bears the risk of yielding inaccurate results, as extreme points may also include noise points requiring visual inspection before selection. Moreover, any automated implementation must carefully consider relevant parameters and considering convex hull algorithm proves significantly more reliable in this regard.

In conclusion, this paper highlights also the potential synergy between the exploratory analysis and the unsupervised methods of clustering and unmixing. Further exploration of their combined application, which remains relatively unexplored in the scientific community, is warranted, thus paving the way for a new research direction.

CRediT authorship contribution statement

Alessandra Olarini: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Marina Cocchi:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Investigation, Funding acquisition, Conceptualization. **Vincent Motto-Ros:** Resources, Data curation. **Ludovic Duponchel:** Writing – review & editing, Visualization, Validation, Supervision, Investigation, Funding acquisition, Conceptualization. **Cyril Ruckebusch:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors acknowledge Laureen Coïc for making data available. A. O. grant received support from Fondo Dipartimentale per la Ricerca (FDR2020) Università degli Studi di Modena e Reggio Emilia and Erasmus+ program University of Modena and Reggio Emilia. A.O. acknowledges LASIRE-DyNaChem research team for fruitful discussion.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2024.105174.

References

[1] J.M. Amigo, Hyperspectral and multispectral imaging: setting the scene, Data Handling Sci. Technol. 32 (2020) 3–16, https://doi.org/10.1016/B978-0-444-63977-6.00001-8.
[2] B. Gaci, F. Abdelghafour, M. Ryckewaert, S. Mas-Garcia, M. Louargant, F. Verpont, Y. Laloum, R. Bendoula, G. Chaix, J.M. Roger, A novel approach to combine spatial and spectral information from hyperspectral images, Chemometr. Intell. Lab. Syst. 240 (2023) 104897, https://doi.org/10.1016/j.chemolab.2023.104897.
[3] L. Coic, P.Y. Sacré, A. Dispas, C. De Bleye, M. Fillet, C. Ruckebusch, P. Hubert, E. Ziemons, Pixel-based Raman hyperspectral identification of complex pharmaceutical formulations, Anal. Chim. Acta 1155 (2021), https://doi.org/10.1016/j.aca.2021.338361.
[4] G. Lu, B. Fei, Medical hyperspectral imaging: a review, J. Biomed. Opt. 19 (2014) 010901, https://doi.org/10.1117/1.jbo.19.1.010901.
[5] B. Lu, P.D. Dao, J. Liu, Y. He, J. Shang, Recent advances of hyperspectral imaging technology and applications in agriculture, Rem. Sens. 12 (2020) 1–44, https://doi.org/10.3390/RS12162659.
[6] N. Keshava, J.F. Mustard, Spectral unmixing, IEEE Signal Process. Mag. 19 (2002) 44–57, https://doi.org/10.1109/79.974727.
[7] V. Olmos, L. Benítez, M. Marro, P. Loza-Alvarez, B. Piña, R. Tauler, A. de Juan, Relevant aspects of unmixing/resolution analysis for the interpretation of biological vibrational hyperspectral images, TrAC, Trends Anal. Chem. 94 (2017) 130–140, https://doi.org/10.1016/j.trac.2017.07.004.
[8] C. Ruckebusch, Resolving spectral mixtures: with applications from ultrafast time-resolved spectroscopy to super-resolution imaging. Data Handling in Science and Technology 30, Elsevier, 2016.
[9] R. Tauler, Multivariate curve resolution applied to second order data, Chemometr. Intell. Lab. Syst. 30 (1995) 133–146, https://doi.org/10.1016/0169-7439(95)00047-X.
[10] H. Abdollahi, M. Maeder, R. Tauler, Calculation and meaning of feasible band boundaries in multivariate curve resolution of a two-component system, Anal. Chem. 81 (2009) 2115–2122, https://doi.org/10.1021/ac8022197.
[11] J. Jaumot, R. Tauler, MCR-BANDS: a user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution, Chemometr. Intell. Lab. Syst. 103 (2010) 96–107, https://doi.org/10.1016/j.chemolab.2010.05.020.
[12] B. Celik, QLSU (QGIS Linear Spectral Unmixing) Plugin: an open source linear spectral unmixing tool for hyperspectral & multispectral remote sensing imagery, Environ. Model. Software 168 (2023) 105782, https://doi.org/10.1016/j.envsoft.2023.105782.
[13] J. Chaumel, M. Marsal, A. Gómez-Sánchez, M. Blumer, E.J. Gualda, A. de Juan, P. Loza-Alvarez, M.N. Dean, Autofluorescence of stingray skeletal cartilage: hyperspectral imaging as a tool for histological characterization, Discov Mater 1 (2021), https://doi.org/10.1007/s43939-021-00015-x.
[14] N. Cavallini, L. Strani, P.P. Becchi, V. Pizzamiglio, S. Michelini, F. Savorani, M. Cocchi, C. Durante, Tracing the identity of Parmigiano Reggiano "Prodotto di Montagna - Progetto Territorio" cheese using NMR spectroscopy and multivariate data analysis, Anal. Chim. Acta 1278 (2023), https://doi.org/10.1016/j.aca.2023.341761.
[15] C. Ruckebusch, R. Vitale, M. Ghaffari, S. Hugelier, N. Omidikia, Perspective on essential information in multivariate curve resolution, TrAC, Trends Anal. Chem. 132 (2020) 116044, https://doi.org/10.1016/j.trac.2020.116044.
[16] R.S. Michalski, Knowledge acquisition through conceptual clustering: a theoretical framework and an algorithm for partitioning data into conjunctive concepts, Int. J. Pol. Anal. Inf. Syst. 4 (1980) 219–244.
[17] T.S. Madhulatha, An overview on clustering methods, IOSR J. Eng. 2 (2012) 719–725, https://doi.org/10.9790/3021-0204719725.
[18] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability vol. 1, 1967, pp. 281–297. Statistics.
[19] U. Maulik, S. Bandyopadhyay, Performance evaluation of some clustering algorithms and validity indices, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 1650–1654, https://doi.org/10.1109/TPAMI.2002.1114856.
[20] B. Desgraupes, Package clusterCrit: clustering indices, CRAN Package (2017) 1–34, cran.r-project.org/web/packages/clusterCrit.
[21] A. Kaarna, P. Zemcik, H. Kälviäinen, J. Parkkinen, Compression of multispectral remote sensing images using clustering and spectral reduction, IEEE Trans. Geosci. Rem. Sens. 38 (2000) 1073–1082, https://doi.org/10.1109/36.841986.
[22] S. Piqueras, C. Krafft, C. Beleites, K. Egodage, F. von Eggeling, O. Guntinas-Lichius, J. Popp, R. Tauler, A. de Juan, Combining multiset resolution and segmentation for hyperspectral image analysis of biological tissues, Anal. Chim. Acta 881 (2015) 24–36, https://doi.org/10.1016/J.ACA.2015.04.053.
[23] L. Massart, D. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Wiley, 1983.
[24] T. Celik, Unsupervised change detection in satellite images using principal component analysis and κ-means clustering, Geosci. Rem. Sens. Lett. IEEE 6 (2009) 772–776, https://doi.org/10.1109/LGRS.2009.2025059.
[25] I.E. Kaya, A.Ç. Pehlivanlı, E.G. Sekizkardeş, T. Ibrikci, PCA based clustering for brain tumor segmentation of T1w MRI images, Comput. Methods Progr. Biomed. 140 (2017) 19–28, https://doi.org/10.1016/j.cmpb.2016.11.011.
[26] P. Firmani, S. Hugelier, F. Marini, C. Ruckebusch, MCR-ALS of hyperspectral images with spatio-spectral fuzzy clustering constraint, Chemometr. Intell. Lab. Syst. 179 (2018) 85–91, https://doi.org/10.1016/j.chemolab.2018.06.007.
[27] D. ChengX, Z. Cai, J. Li, M. Wen, Y. Wang, A. Zeng, spatial-spectral clustering-based algorithm for endmember extraction and hyperspectral unmixing, Int. J. Rem. Sens. 42 (2021) 1948–1972.
[28] J.W. Tukey, Exploratory Data Analysis, vol. 2, Addison-Wesley Publishing Company, 1977.
[29] M. Li Vigni, C. Durante, M. Cocchi, Exploratory Data Analysis, first ed., Elsevier, 2013 https://doi.org/10.1016/B978-0-444-59528-7.00003-X.
[30] M. Ghaffari, N. Omidikia, C. Ruckebusch, Essential spectral pixels for multivariate curve resolution of chemical images, Anal. Chem. 91 (2019) 10943–10948, https://doi.org/10.1021/acs.analchem.9b02890.
[31] L. Coic, R. Vitale, M. Moreau, D. Rousseau, J.H. de Morais Goulart, N. Dobigeon, C. Ruckebusch, Assessment of essential information in the fourier domain to

accelerate Raman hyperspectral microimaging, Anal. Chem. 95 (2023) 15497–15504, https://doi.org/10.1021/acs.analchem.3c01383.
[32] S.V. Zade, K. Neymeyr, M. Sawall, C. Fischer, H. Abdollahi, Data point importance: information ranking in multivariate data, J. Chemom. 37 (2023) 1–15, https://doi.org/10.1002/cem.3453.
[33] V.H.C. Ferreira, V. Gardette, B. Busser, L. Sancey, S. Ronsmans, V. Bonneterre, V. Motto-Ros, L. Duponchel, Enhancing diagnostic capabilities for occupational lung diseases using LIBS imaging on biopsy tissue, Anal. Chem. (2024), https://doi.org/10.1021/acs.analchem.4c00237.
[34] Q. Wu, C. Marina-Montes, J.O. Cáceres, J. Anzano, V. Motto-Ros, L. Duponchel, Interesting features finder (IFF): another way to explore spectroscopic imaging data sets giving minor compounds and traces a chance to express themselves, Spectrochim. Acta Part B At. Spectrosc. 195 (2022), https://doi.org/10.1016/j.sab.2022.106508.
[35] M. Ghaffari, N. Omidikia, C. Ruckebusch, Joint selection of essential pixels and essential variables across hyperspectral images, Anal. Chim. Acta 1141 (2021) 36–46, https://doi.org/10.1016/j.aca.2020.10.040.
[36] S. Khodadadi Karimvand, J. Mohammad Jafari, S. Vali Zade, H. Abdollahi, Practical and comparative application of efficient data reduction - multivariate curve resolution, Anal. Chim. Acta 1243 (2023) 340824, https://doi.org/10.1016/j.aca.2023.340824.
[37] M. Sawall, C. Ruckebusch, M. Beese, R. Francke, A. Prudlik, K. Neymeyr, An active constraint approach to identify essential spectral information in noisy data, Anal. Chim. Acta 1233 (2022) 340448, https://doi.org/10.1016/j.aca.2022.340448.
[38] R. Vitale, C. Ruckebusch, On a black hole effect in bilinear curve resolution based on least squares, J. Chemom. 37 (2023) 1–7, https://doi.org/10.1002/cem.3442.
[39] E.C. Muñoz, F. Gosetti, D. Ballabio, S. Andò, O. Gómez-Laserna, J.M. Amigo, E. Garzanti, Characterization of pyrite weathering products by Raman hyperspectral imaging and chemometrics techniques, Microchem. J. 190 (2023), https://doi.org/10.1016/j.microc.2023.108655.
[40] A. de Juan, R. Tauler, Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – a review, Anal. Chim. Acta 1145 (2021) 59–78, https://doi.org/10.1016/j.aca.2020.10.051.
[41] A. De Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods 6 (2014) 4964–4976.
[42] S. Moncayo, L. Duponchel, N. Mousavipak, G. Panczer, F. Trichard, B. Bousquet, F. Pelascini, V. Motto-Ros, Exploration of megapixel hyperspectral LIBS images using principal component analysis, J Anal At Spectrom 33 (2018) 210–220, https://doi.org/10.1039/c7ja00398f.
[43] C. Golub, G. H Reinsch, Singular value decomposition and least squares solutions, Numer. Math. 14 (1970) 403–420, https://doi.org/10.1007/BF02163027.
[44] R. Rajkó, Studies on the adaptability of different Borgen norms applied in self-modeling curve resolution (SMCR) method, J. Chemom. 23 (2009) 265–274, https://doi.org/10.1002/cem.1221.
[45] B.V. Grande, R. Manne, Use of convexity for finding pure variables in two-way data from mixtures, Chemometr. Intell. Lab. Syst. 50 (2000) 19–33, https://doi.org/10.1016/S0169-7439(99)00041-6.
[46] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS gui 2.0: new features and applications, Chemometr. Intell. Lab. Syst. 140 (2015) 1–12, https://doi.org/10.1016/j.chemolab.2014.10.003.
[47] W. Windig, J. Guilment, Interactive Self-Modeling Mixture Analysis, 1991, pp. 1425–1432.
[48] V. Kumar, J.K. Chhabra, K. Dinesh, Performance evaluation of distance metrics in the clustering algorithms, INFOCOMP J. Comput. Sci. 13 (2014) 38–51.
[49] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65, https://doi.org/10.1016/0377-0427(87)90125-7.
[50] M.K. Pakhira, S. Bandyopadhyay, U. Maulik, Validity index for crisp and fuzzy clusters, Pattern Recogn. 37 (2004) 487–501, https://doi.org/10.1016/j.patcog.2003.06.005.
[51] G. Hu, K. Dam-Johansen, S. Wedel, J.P. Hansen, Decomposition and oxidation of pyrite, Prog. Energy Combust. Sci. 32 (2006) 295–314, https://doi.org/10.1016/j.pecs.2005.11.004.
[52] X. Wang, Y. Guo, The impact of trace metal cations and absorbed water on colour transition of turquoise, R. Soc. Open Sci. 8 (2021), https://doi.org/10.1098/rsos.201110.
[53] J.J. Rushchitsky, Interaction of waves in solid mixtures, Appl. Mech. Rev. 52 (1999) 35–74.
[54] https://www.mindat.org/, last access 18/October/2023.
[55] https://www.atomtrace.com/elements-database/, last access 18/October/2023.

Bibliography

[1] John W Tukey. *Exploratory data analysis*. Addison-Wesley, 1977.

[2] Mario Li Vigni, Caterina Durante, and Marina Cocchi. "Exploratory data analysis". In: *Data handling in science and technology*. Vol. 28. Elsevier, 2013, pp. 55–126.

[3] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education, 2006.

[4] José Manuel Amigo. "Hyperspectral and multispectral imaging: setting the scene". In: *Data handling in science and technology*. Vol. 32. Elsevier, 2019, pp. 3–16.

[5] Svante Wold. "Chemometrics; what do we mean with it, and what do we want from it?" In: *Chemometrics and intelligent laboratory systems* 30.1 (1995), pp. 109–115.

[6] José Manuel Amigo, Hamid Babamoradi, and Saioa Elcoroaristizabal. "Hyperspectral image analysis. A tutorial". In: *Analytica chimica acta* 896 (2015), pp. 34–51.

[7] James B Campbell and Randolph H Wynne. *Introduction to remote sensing*. Guilford press, 2011.

[8] Sergey Kucheryavski. "Extracting useful information from images". In: *Chemometrics and Intelligent Laboratory Systems* 108.1 (2011), pp. 2–12.

[9] John C Russ. *The image processing handbook*. CRC press, 2006.

[10] Vincent Baeten and Pierre Dardenne. "Spectroscopy: Developments in instrumentation and analysis". In: *Grasas y aceites* 53.1 (2002), pp. 45–63.

[11] J Michael Hollas. *Modern spectroscopy*. John Wiley & Sons, 2004.

[12] Hans Grahn and Paul Geladi. *Techniques and applications of hyperspectral image analysis*. John Wiley & Sons, 2007.

[13] Yuval Garini, Ian T Young, and George McNamara. "Spectral imaging: principles and applications". In: *Cytometry part a: the journal of the international society for analytical cytology* 69.8 (2006), pp. 735–747.

[14] Luc Vincent. "Morphological grayscale reconstruction in image analysis: applications and efficient algorithms". In: *IEEE transactions on image processing* 2.2 (1993), pp. 176–201.

[15] M Emre Celebi and Gerald Schaefer. *Color medical image analysis*. Vol. 6. Springer Science & Business Media, 2012.

[16] Paul Geladi and Hans F Grahn. "Multivariate image analysis". In: *Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation* (2006).

[17] Anna de Juan. "Hyperspectral image analysis. When space meets Chemistry". In: *Journal of Chemometrics* 32.1 (2018), e2985.

[18] Sara Piqueras et al. "Resolution and segmentation of hyperspectral biomedical images by multivariate curve resolution-alternating least squares". In: *Analytica chimica acta* 705.1-2 (2011), pp. 182–192.

[19] Daniel Caballero, Rosalba Calvini, and José Manuel Amigo. "Hyperspectral imaging in crop fields: precision agriculture". In: *Data handling in science and technology*. Vol. 32. Elsevier, 2019, pp. 453–473.

[20] Silas J Leavesley et al. "Hyperspectral imaging microscopy for identification and quantitative analysis of fluorescently-labeled cells in highly autofluorescent tissue". In: *Journal of biophotonics* 5.1 (2012), pp. 67–84.

[21] Baowei Fei. "Hyperspectral imaging in medical applications". In: *Data handling in science and technology*. Vol. 32. Elsevier, 2019, pp. 523–565.

[22] Rosalba Calvini et al. "Evaluation of the effect of factors related to preparation and composition of grated Parmigiano Reggiano cheese using NIR hyperspectral imaging". In: *Food Control* 131 (2022), p. 108412.

[23] Alessandro Benelli, Chiara Cevoli, and Angelo Fabbri. "In-field hyperspectral imaging: An overview on the ground-based applications in agriculture". In: *Journal of Agricultural Engineering* 51.3 (2020), pp. 129–139.

[24] Roozbeh Rajabi et al. "Hyperspectral imaging in environmental monitoring and analysis". In: *Frontiers in Environmental Science* 11 (2024), p. 1353447.

[25] E Keith Hege et al. "Hyperspectral imaging for astronomy and space surveillance". In: *Imaging Spectrometry IX*. Vol. 5159. SPIE. 2004, pp. 380–391.

[26] Bing Lu et al. “Recent advances of hyperspectral imaging technology and applications in agriculture”. In: *Remote Sensing* 12.16 (2020), p. 2659.

[27] Laureen Coic et al. “Pixel-based Raman hyperspectral identification of complex pharmaceutical formulations”. In: *Analytica Chimica Acta* 1155 (2021), p. 338361.

[28] Sek M Chai et al. “Focal-plane processing architectures for real-time hyperspectral image processing”. In: *Applied Optics* 39.5 (2000), pp. 835–849.

[29] Jianwei Qin. “Hyperspectral imaging instruments”. In: *Hyperspectral imaging for food quality analysis and control*. Elsevier, 2010, pp. 129–172.

[30] Aoife Gowen, Edurne Gaston, and James Burger. “Hyperspectral imaging”. In: *Process Analytical Technology for the Food Industry*. Springer, 2014, pp. 199–216.

[31] Floyd F Sabins Jr and James M Ellis. *Remote sensing: Principles, interpretation, and applications*. Waveland Press, 2020.

[32] Nathan Hagen and Michael W Kudenov. “Review of snapshot spectral imaging technologies”. In: *Optical Engineering* 52.9 (2013), pp. 090901–090901.

[33] Eduardo Sommella et al. “MALDI mass spectrometry imaging highlights specific metabolome and lipidome profiles in salivary gland tumor tissues”. In: *Metabolites* 12.6 (2022), p. 530.

[34] Hang Hu and Julia Laskin. “Emerging computational methods in mass spectrometry imaging”. In: *Advanced Science* 9.34 (2022), p. 2203339.

[35] Vaibhav Lodhi, Debashish Chakravarty, and Pabitra Mitra. “Hyperspectral imaging system: Development aspects and recent trends”. In: *Sensing and Imaging* 20 (2019), pp. 1–24.

[36] Rajeev Kumar, Vijay Paul, and Rakesh Pandey. “Hyperspectral imaging/reflectance as a tool for assessment of nutritional and quality-related parameters in tomato (Solanum lycopersicum) fruits-a review”. In: *Current Horticulture* 12.1 (2024), pp. 13–22.

[37] Agnese Babini et al. “Comparison of hyperspectral imaging and fiber-optic reflectance spectroscopy for reflectance and transmittance measurements of colored glass”. In: *Heritage* 5.3 (2022), pp. 1401–1418.

[38] Yong-Kyoung Kim et al. “Investigation of reflectance, fluorescence, and Raman hyperspectral imaging techniques for rapid detection of aflatoxins in ground maize”. In: *Food Control* 132 (2022), p. 108479.

[39] Kyung Jo et al. “Hyperspectral imaging–based assessment of fresh meat quality: Progress and applications”. In: *Microchemical Journal* 197 (2024), p. 109785.

[40] Slobodan Sasic and Yukihiro Ozaki. *Raman, infrared, and near-infrared chemical imaging*. John Wiley & Sons, 2011.

[41] Jack R White. “Herschel and the Puzzle of Infrared”. In: *American Scientist* 100.3 (2012), pp. 218–225.

[42] Edward FJ Ring. “The discovery of infrared radiation in 1800”. In: *The Imaging Science Journal* 48.1 (2000), pp. 1–8.

[43] Jack L Koenig and John P Bobiak. “Raman and infrared imaging of dynamic polymer systems”. In: *Macromolecular Materials and Engineering* 292.7 (2007), pp. 801–816.

[44] Eric A Muller, Benjamin Pollard, and Markus B Raschke. “Infrared chemical nanoimaging: Accessing structure, coupling, and dynamics on molecular length scales”. In: *The Journal of Physical Chemistry Letters* 6.7 (2015), pp. 1275–1284.

[45] Cristina Quintelas et al. “An overview of the evolution of infrared spectroscopy applied to bacterial typing”. In: *Biotechnology journal* 13.1 (2018), p. 1700449.

[46] Marena Manley. “Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials”. In: *Chemical Society Reviews* 43.24 (2014), pp. 8200–8214.

[47] Rohit Bhargava. “Towards a practical Fourier transform infrared chemical imaging protocol for cancer histopathology”. In: *Analytical and bioanalytical chemistry* 389 (2007), pp. 1155–1169.

[48] Hamed Akbari et al. “Cancer detection using infrared hyperspectral imaging”. In: *Cancer science* 102.4 (2011), pp. 852–857.

[49] Lei Feng et al. “Application of visible/infrared spectroscopy and hyperspectral imaging with machine learning techniques for identifying food varieties and geographical origins”. In: *Frontiers in Nutrition* 8 (2021), p. 680357.

[50] Giorgia Foca et al. “The potential of spectral and hyperspectral-imaging techniques for bacterial detection in food: A case study on lactic acid bacteria”. In: *Talanta* 153 (2016), pp. 111–119.

[51] James Coddington and Suzanne Siano. “Infrared imaging of twentieth-century works of art”. In: *Studies in Conservation* 45.sup1 (2000), pp. 39–44.

[52] John K Delaney et al. “Visible and infrared imaging spectroscopy of paintings and improved reflectography”. In: *Heritage Science* 4 (2016), pp. 1–10.

[53] Mary B Stuart, Andrew JS McGonigle, and Jon R Willmott. “Hyperspectral imaging in environmental monitoring: A review of recent developments and technological advances in compact field deployable systems”. In: *Sensors* 19.14 (2019), p. 3071.

[54] Bing Zhang et al. “Application of hyperspectral remote sensing for environment monitoring in mining areas”. In: *Environmental Earth Sciences* 65 (2012), pp. 649–658.

[55] William L Wolfe and George John Zissis. *The infrared handbook*. IRIA Center, Environmental Research Institute of Michigan, 1978.

[56] Yukihiro Ozaki. “Infrared spectroscopy—Mid-infrared, near-infrared, and far-infrared/terahertz spectroscopy”. In: *Analytical Sciences* 37.9 (2021), pp. 1193–1212.

[57] José Manuel Amigo and Silvia Grassi. “Configuration of hyperspectral and multispectral imaging systems”. In: *Data handling in science and technology*. Vol. 32. Elsevier, 2019, pp. 17–34.

[58] James Burger and Paul Geladi. “Hyperspectral NIR imaging for calibration and prediction: a comparison between image and spectrometer data for studying organic and biological samples”. In: *Analyst* 131.10 (2006), pp. 1152–1160.

[59] Rosario Brunetto et al. “Hyperspectral FTIR imaging of irradiated carbonaceous meteorites”. In: *Planetary and Space Science* 158 (2018), pp. 38–45.

[60] Chandrasekhara Venkata Raman and Kariamanikkam Srinivasa Krishnan. “A new type of secondary radiation”. In: *Nature* 121.3048 (1928), pp. 501–502.

[61] Robin R Jones et al. “Raman techniques: fundamentals and frontiers”. In: *Nanoscale research letters* 14 (2019), pp. 1–34.

[62] Ewen Smith and Geoffrey Dent. *Modern Raman spectroscopy: a practical approach*. John Wiley & Sons, 2019.

[63] Volker Deckert et al. “Spatial resolution in Raman spectroscopy”. In: *Faraday Discussions* 177 (2015), pp. 9–20.

[64] Satoshi Kawata et al. “Nano-Raman scattering microscopy: resolution and enhancement”. In: *Chemical reviews* 117.7 (2017), pp. 4983–5001.

[65] Xingchen Dong et al. “A review of hyperspectral imaging for nanoscale materials research”. In: *Applied Spectroscopy Reviews* 54.4 (2019), pp. 285–305.

[66] Matthieu Paillet et al. “Graphene and related 2D materials: An overview of the Raman studies”. In: *Journal of Raman Spectroscopy* 49.1 (2018), pp. 8–12.

[67] Sara Piqueras et al. “Monitoring polymorphic transformations by using in situ Raman hyperspectral imaging and image multiset analysis”. In: *Analytica chimica acta* 819 (2014), pp. 15–25.

[68] Valentin Gilet et al. “Superpixels meet essential spectra for fast Raman hyperspectral microimaging”. In: *Optics Express* 32.1 (2023), pp. 932–948.

[69] Joseph P Smith et al. “Raman hyperspectral imaging with multivariate analysis for investigating enzyme immobilization”. In: *Analyst* 145.23 (2020), pp. 7571–7581.

[70] Bianca Durrant et al. “Recent developments in spontaneous Raman imaging of living biological cells”. In: *Current opinion in chemical biology* 51 (2019), pp. 138–145.

[71] Giuseppe Pezzotti. “Raman spectroscopy in cell biology and microbiology”. In: *Journal of Raman Spectroscopy* 52.12 (2021), pp. 2348–2443.

[72] Anja Silge et al. “Trends in pharmaceutical analysis and quality control by modern Raman spectroscopic techniques”. In: *TrAC Trends in Analytical Chemistry* 153 (2022), p. 116623.

[73] Patient Hamuli Ciza et al. “Comparing the qualitative performances of handheld NIR and Raman spectrophotometers for the detection of falsified pharmaceutical products”. In: *Talanta* 202 (2019), pp. 469–478.

[74] Florian Gruber et al. “Hyperspectral imaging using laser excitation for fast Raman and fluorescence hyperspectral imaging for sorting and quality control applications”. In: *Journal of imaging* 4.10 (2018), p. 110.

[75] Yue Sun et al. “Raman spectroscopy for food quality assurance and safety monitoring: A review”. In: *Current Opinion in Food Science* 47 (2022), p. 100910.

[76] Frederick Brech and Lee Cross. “Optical microemission stimulated by a ruby maser”. In: *Appl. Spectrosc.* 16.2 (1962), p. 59.

[77] Edward F Runge, Robert W Minck, and Ford R Bryan. “Spectrochemical analysis using a pulsed laser source”. In: *Spectrochimica acta* 20.4 (1964), pp. 733–736.

[78] Francisco J Fortes et al. “Laser-induced breakdown spectroscopy”. In: *Analytical chemistry* 85.2 (2013), pp. 640–669.

[79] Leon J Radziemski. “From LASER to LIBS, the path of technology development”. In: *Spectrochimica Acta Part B: Atomic Spectroscopy* 57.7 (2002), pp. 1109–1113.

[80] Vincent Gardette et al. “LIBS-based imaging: Recent advances and future directions”. In: *Spectroscopy* 35.2 (2020), pp. 34–40.

[81] Shiv K Sharma et al. “A combined remote Raman and LIBS instrument for characterizing minerals with 532 nm laser excitation”. In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 73.3 (2009), pp. 468–476.

[82] Anastasia Giakoumaki, Iacopo Osticioli, and Demetrios Anglos. “Spectroscopic analysis using a hybrid LIBS-Raman system”. In: *Applied Physics A* 83 (2006), pp. 537–541.

[83] Lina Jolivet et al. “Review of the recent advances and applications of LIBS-based imaging”. In: *Spectrochimica Acta Part B: Atomic Spectroscopy* 151 (2019), pp. 41–53.

[84] Vincent Motto-Ros et al. “LIBS imaging applications”. In: *Laser-Induced Breakdown Spectroscopy*. Elsevier, 2020, pp. 329–346.

[85] Alessandro Nardecchia et al. “Detection of minor compounds in complex mineral samples from millions of spectra: A new data analysis strategy in LIBS imaging”. In: *Analytica Chimica Acta* 1114 (2020), pp. 66–73.

[86] Alessandro Nardecchia et al. “Data fusion of LIBS and PIL hyperspectral imaging: Understanding the luminescence phenomenon of a complex mineral sample”. In: *Analytica Chimica Acta* 1192 (2022), p. 339368.

[87] Miguel FS Ferreira et al. “LIBS imaging as a process control tool in the cork industry”. In: *Optical Sensing and Detection VIII*. Vol. 12999. 2024, pp. 188–192.

[88] Reinhard Noll et al. “Laser-induced breakdown spectrometry—applications for production control and quality assurance in the steel industry”. In: *Spectrochimica Acta Part B: Atomic Spectroscopy* 56.6 (2001), pp. 637–649.

[89] Wilhelm Nikonow et al. “Advanced mineral characterization and petrographic analysis by μ-EDXRF, LIBS, HSI and hyperspectral data merging”. In: *Mineralogy and Petrology* 113 (2019), pp. 417–431.

[90] Ashwin Kumar Myakalwar et al. “LIBS as a spectral sensor for monitoring metallic molten phase in metallurgical applications—a review”. In: *Minerals* 11.10 (2021), p. 1073.

[91] Kristalia Melessanaki et al. “The application of LIBS for the analysis of archaeological ceramic and metal artifacts”. In: *Applied surface science* 197 (2002), pp. 156–163.

[92] Changqing Liu et al. “A stand-off laser-induced breakdown spectroscopy (LIBS) system applicable for Martian rocks studies”. In: *Remote Sensing* 13.23 (2021), p. 4773.

[93] Andrew K Knight et al. “Characterization of laser-induced breakdown spectroscopy (LIBS) for application to space exploration”. In: *Applied Spectroscopy* 54.3 (2000), pp. 331–340.

[94] Christophe Dutouquet et al. “Monitoring of heavy metal particle emission in the exhaust duct of a foundry using LIBS”. In: *Talanta* 127 (2014), pp. 75–81.

[95] Danielle S Francischini and Marco AZ Arruda. “When a picture is worth a thousand words: Molecular and elemental imaging applied to environmental analysis–A review”. In: *Microchemical Journal* 169 (2021), p. 106526.

[96] Svante Wold, Michael Sjöström, and Lennart Eriksson. “PLS-regression: a basic tool of chemometrics”. In: *Chemometrics and intelligent laboratory systems* 58.2 (2001), pp. 109–130.

[97] Desiré Luc Massart et al. *Handbook of chemometrics and qualimetrics*. Elsevier Science Inc., 1998.

[98] Nicola Falco et al. “Supervised classification methods in hyperspectral imaging—recent advances”. In: *Data Handling in Science and Technology* 32 (2019), pp. 247–279.

[99] Neal B Gallagher. “Classical least squares for detection and classification”. In: *Data Handling in Science and Technology*. Vol. 32. Elsevier, 2019, pp. 231–246.

[100] Michael Sjöström et al. “A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables”. In: *Analytica Chimica Acta* 150 (1983), pp. 61–70.

[101] Federico Marini. “Classification methods in chemometrics”. In: *Current Analytical Chemistry* 6.1 (2010), pp. 72–79.

[102] Raffaele Vitale et al. “Class modelling by soft independent modelling of class analogy: why, when, how? A tutorial”. In: *Analytica Chimica Acta* 1270 (2023), p. 341304.

[103] Lorenzo Strani et al. “One class classification (class modelling): state of the art and perspectives”. In: *TrAC Trends in Analytical Chemistry* (2024), p. 118117.

[104] Federico Marini and José Manuel Amigo. “Unsupervised exploration of hyperspectral and multispectral images”. In: *Data handling in science and technology*. Vol. 32. Elsevier, 2019, pp. 93–114.

[105] Cyril Ruckebusch et al. “Perspective on essential information in multivariate curve resolution”. In: *TrAC Trends in Analytical Chemistry* 132 (2020), p. 116044.

[106] Laureen Coic et al. “Essential spectra to improve vibrational imaging of pharmaceutical samples”. In: *Microchemical Journal* (2025), p. 112751.

[107] Qicheng Wu et al. “Interesting features finder (IFF): Another way to explore spectroscopic imaging data sets giving minor compounds and traces a chance to express themselves”. In: *Spectrochimica Acta Part B: Atomic Spectroscopy* 195 (2022), p. 106508.

[108] Mohamad Ahmad et al. “A novel proposal to investigate the interplay between the spatial and spectral domains in near-infrared spectral imaging data by means of Image Decomposition, Encoding and Localization (IDEL)”. In: *Analytica Chimica Acta* 1191 (2022), p. 339285.

[109] Mohamad Ahmad et al. “Exploring local spatial features in hyperspectral images”. In: *Journal of Chemometrics* 34.10 (2020), e3295.

[110] Harold Hotelling. “Analysis of a complex of statistical variables into principal components.” In: *Journal of educational psychology* 24.6 (1933), p. 417.

[111] Karl Pearson. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2.11 (1901), pp. 559–572.

[112] Rasmus Bro and Age K Smilde. “Principal component analysis”. In: *Analytical methods* 6.9 (2014), pp. 2812–2831.

[113] Cristina Malegori et al. “Identification of invisible biological traces in forensic evidences by hyperspectral NIR imaging combined with chemometrics”. In: *Talanta* 215 (2020), p. 120911.

[114] Nicola Cavallini et al. “Tracing the identity of Parmigiano Reggiano “Prodotto di Montagna-Progetto Territorio” cheese using NMR spectroscopy and multivariate data analysis”. In: *Analytica Chimica Acta* 1278 (2023), p. 341761.

[115] Lorraine Awhangbo et al. “Fault detection with moving window PCA using NIRS spectra for monitoring the anaerobic digestion process”. In: *Water Science and Technology* 81.2 (2020), pp. 367–382.

[116] Lorenzo Strani et al. “Chemical Characterization and Temporal Variability of Pasta Condiment By-Products for Sustainable Waste Management”. In: *Foods* 13.18 (2024), p. 3018.

[117] Cristina Malegori and Paolo Oliveri. “Principal component analysis”. In: *Hyperspectral Imaging Analysis and Applications for Food Quality*. CRC Press, 2018, pp. 85–107.

[118] Puneet Mishra et al. “Detection and quantification of peanut traces in wheat flour by near infrared hyperspectral imaging spectroscopy using principal-component analysis”. In: *Journal of Near Infrared Spectroscopy* 23.1 (2015), pp. 15–22.

[119] Aloke Datta, Susmita Ghosh, and Ashish Ghosh. “PCA, kernel PCA and dimensionality reduction in hyperspectral images”. In: *Advances in Principal Component Analysis: Research and Development* (2018), pp. 19–46.

[120] J Edward Jackson. “Principal components and factor analysis: part I—principal components”. In: *Journal of Quality Technology* 12.4 (1980), pp. 201–213.

[121] J Edward Jackson. “Principal components and factor analysis: part II—additional topics related to principal components”. In: *Journal of Quality Technology* 13.1 (1981), pp. 46–58.

[122] Laura Bianca Bilius and Stefan Gheorghe Pentiuc. “Unsupervised clustering for hyperspectral images”. In: *Symmetry* 12.2 (2020), p. 277.

[123] Claude Cariou and Kacem Chehdi. “Unsupervised nearest neighbors clustering with application to hyperspectral images”. In: *IEEE Journal of Selected Topics in Signal Processing* 9.6 (2015), pp. 1105–1116.

[124] Etienne Ducasse et al. “Mapping of Clay Montmorillonite Abundance in Agricultural Fields Using Unmixing Methods at Centimeter Scale Hyperspectral Images”. In: *Remote Sensing* 16.17 (2024), p. 3211.

[125] Gregory P Asner and Kathleen B Heidebrecht. “Spectral unmixing of vegetation, soil and dry carbon cover in arid regions: comparing multispectral and hyperspectral observations”. In: *International Journal of Remote Sensing* 23.19 (2002), pp. 3939–3958.

[126] Roberto Todeschini et al. *Introduzione alla chemiometria*. EdiSES, 1998.

[127] Rizki Suwanda, Zulfahmi Syahputra, and Elvi M Zamzami. “Analysis of euclidean distance and manhattan distance in the K-means algorithm for variations number of centroid K”. In: *Journal of Physics: Conference Series*. Vol. 1566. 1. IOP Publishing. 2020, p. 012058.

[128] Aye Aye Thant, Soe Moe Aye, and Myanmar Mandalay. “Euclidean, manhattan and minkowski distance methods for clustering algorithms”. In: *International Journal of Scientific Research in Science, Engineering and Technology* 7.3 (2020), pp. 553–559.

[129] Renato Cordeiro De Amorim and Boris Mirkin. “Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering”. In: *Pattern Recognition* 45.3 (2012), pp. 1061–1075.

[130] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Desiré Luc Massart. “The mahalanobis distance”. In: *Chemometrics and intelligent laboratory systems* 50.1 (2000), pp. 1–18.

[131] Dominic Edelmann, Tamás F Móri, and Gábor J Székely. “On relationships between the Pearson and the distance correlation coefficients”. In: *Statistics & probability letters* 169 (2021), p. 108960.

[132] Vijay Kumar, Jitender Kumar Chhabra, and Dinesh Kumar. “Performance evaluation of distance metrics in the clustering algorithms”. In: *INFOCOMP Journal of Computer Science* 13.1 (2014), pp. 38–52.

[133] Taher M Ghazal. “Performances of k-means clustering algorithm with different distance metrics”. In: *Intelligent Automation & Soft Computing* 30.2 (2021), pp. 735–742.

[134] Gillian M Mimmack, Simon J Mason, and Jacqueline S Galpin. “Choice of distance matrices in cluster analysis: Defining regions”. In: *Journal of climate* 14.12 (2001), pp. 2790–2797.

[135] Shraddha Pandit, Suchita Gupta, et al. “A comparative study on distance measuring approaches for clustering”. In: *International journal of research in computer science* 2.1 (2011), pp. 29–31.

[136] Mihael Ankerst et al. “OPTICS: Ordering points to identify the clustering structure”. In: *ACM Sigmod record* 28.2 (1999), pp. 49–60.

[137] Steven Brown, Romà Tauler, and Beata Walczak. *Comprehensive chemometrics: chemical and biochemical data analysis*. Elsevier, 2020.

[138] Martin Ester et al. “A density-based algorithm for discovering clusters in large spatial databases with noise”. In: *KDD’96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Vol. 96. 34. 1996, pp. 226–231.

[139] Michal Daszykowski, Beata Walczak, and Desiré Luc Massart. “Density-based clustering for exploration of analytical data”. In: *Analytical and bioanalytical chemistry* 380 (2004), pp. 370–372.

[140] Han Zhai et al. “Hyperspectral image clustering: Current achievements and future lines”. In: *IEEE Geoscience and Remote Sensing Magazine* 9.4 (2021), pp. 35–67.

[141] Frank Nielsen and Frank Nielsen. “Hierarchical clustering”. In: *Introduction to HPC with MPI for Data Science* (2016), pp. 195–211.

[142] Ildiko E Frank and Roberto Todeschini. *The data analysis handbook*. Elsevier, 1994.

[143] Desiré Luc Massart and Leonard Kaufman. *The interpretation of analytical chemical data by the use of cluster analysis*. Wiley, 1983.

[144] James MacQueen et al. “Some methods for classification and analysis of multivariate observations”. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA. 1967, pp. 281–297.

[145] James C Bezdek. “Cluster validity with fuzzy sets”. In: (1973).

[146] Joseph C Dunn. “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters”. In: (1973).

[147] Maria Brigida Ferraro. “Fuzzy k-Means: history and applications”. In: *Econometrics and Statistics* 30 (2024), pp. 110–123.

[148] Ting Su and Jennifer Dy. "A deterministic method for initializing k-means clustering". In: *16th IEEE international conference on tools with artificial intelligence*. IEEE. 2004, pp. 784–786.

[149] M Emre Celebi, Hassan A Kingravi, and Patricio A Vela. "A comparative study of efficient initialization methods for the k-means clustering algorithm". In: *Expert systems with applications* 40.1 (2013), pp. 200–210.

[150] José M Pena, Jose Antonio Lozano, and Pedro Larranaga. "An empirical comparison of four initialization methods for the k-means algorithm". In: *Pattern recognition letters* 20.10 (1999), pp. 1027–1040.

[151] Shehroz S Khan and Amir Ahmad. "Cluster center initialization algorithm for K-means clustering". In: *Pattern recognition letters* 25.11 (2004), pp. 1293–1302.

[152] Avgoustinos Vouros et al. "An empirical comparison between stochastic and deterministic centroid initialisation for K-means variations". In: *Machine Learning* 110 (2021), pp. 1975–2003.

[153] Murilo C Naldi, André Fontana, and Ricardo JGB Campello. "Comparison among methods for k estimation in k-means". In: *2009 Ninth International Conference on Intelligent Systems Design and Applications*. IEEE. 2009, pp. 1006–1013.

[154] Shuyue Zhang and Chao Duan. "Clustering optimization algorithm for data mining based on artificial intelligence neural network". In: *Wireless Communications and Mobile Computing* 2022.1 (2022), p. 1304951.

[155] Robert Tibshirani, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423.

[156] Erich Schubert. "Stop using the elbow criterion for k-means and how to choose the number of clusters instead". In: *ACM SIGKDD Explorations Newsletter* 25.1 (2023), pp. 36–42.

[157] Robert L Thorndike. "Who belongs in the family?" In: *Psychometrika* 18.4 (1953), pp. 267–276.

[158] Roberto Todeschini et al. “Extended multivariate comparison of 68 cluster validity indices. A review”. In: *Chemometrics and Intelligent Laboratory Systems* (2024), p. 105117.

[159] Peter J Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[160] Malay K Pakhira, Sanghamitra Bandyopadhyay, and Ujjwal Maulik. “Validity index for crisp and fuzzy clusters”. In: *Pattern recognition* 37.3 (2004), pp. 487–501.

[161] Huapeng Li et al. “Performance evaluation of cluster validity indices (CVIs) on multi/hyperspectral remote sensing datasets”. In: *Remote Sensing* 8.4 (2016), p. 295.

[162] Nathalie Gorretta et al. “Hyperspectral image segmentation: the butterfly approach”. In: *2009 First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*. IEEE. 2009, pp. 1–4.

[163] Abiodun M Ikotun et al. “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data”. In: *Information Sciences* 622 (2023), pp. 178–210.

[164] Merhad Ay et al. “FC-Kmeans: Fixed-centered K-means algorithm”. In: *Expert Systems with Applications* 211 (2023), p. 118656.

[165] K Krishna and M Narasimha Murty. “Genetic K-means algorithm”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 29.3 (1999), pp. 433–439.

[166] Zhang Chen and Shixiong Xia. “K-means clustering algorithm with improved initial center”. In: *2009 Second International Workshop on Knowledge Discovery and Data Mining*. IEEE. 2009, pp. 790–792.

[167] Cyril Ruckebusch and Lionel Blanchet. “Multivariate curve resolution: a review of advanced and tailored applications and challenges”. In: *Analytica chimica acta* 765 (2013), pp. 28–36.

[168] Thomas G Mayerhöfer, Susanne Pahlow, and Jürgen Popp. "The Bouguer-Beer-Lambert law: Shining light on the obscure". In: *ChemPhysChem* 21.18 (2020), pp. 2029–2046.

[169] Donald F Swinehart. "The beer-lambert law". In: *Journal of chemical education* 39.7 (1962), p. 333.

[170] Jiaojiao Wei and Xiaofei Wang. "An overview on linear unmixing of hyperspectral data". In: *Mathematical Problems in Engineering* 2020.1 (2020), p. 3735403.

[171] Ricardo Augusto Borsoi et al. "Spectral variability in hyperspectral data unmixing: A comprehensive review". In: *IEEE geoscience and remote sensing magazine* 9.4 (2021), pp. 223–270.

[172] José M Bioucas-Dias et al. "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches". In: *IEEE journal of selected topics in applied earth observations and remote sensing* 5.2 (2012), pp. 354–379.

[173] Anna de Juan and Roma Tauler. "Multivariate Curve Resolution: 50 years addressing the mixture analysis problem–A review". In: *Analytica Chimica Acta* 1145 (2021), pp. 59–78.

[174] Anna De Juan. "Multivariate curve resolution for hyperspectral image analysis". In: *Data Handling in Science and Technology*. Vol. 32. Elsevier, 2019, pp. 115–150.

[175] Sara Piqueras et al. "Understanding the formation of heartwood in larch using synchrotron infrared imaging combined with multivariate analysis and atomic force microscope infrared spectroscopy". In: *Frontiers in plant science* 10 (2020), p. 1701.

[176] Ludovic Duponchel et al. "Multivariate curve resolution methods in imaging spectroscopy: influence of extraction methods and instrumental perturbations". In: *Journal of Chemical information and computer sciences* 43.6 (2003), pp. 2057–2067.

[177] Abderrahim Halimi et al. "Nonlinear unmixing of hyperspectral images using a generalized bilinear model". In: *IEEE Transactions on Geoscience and Remote Sensing* 49.11 (2011), pp. 4153–4162.

[178] Rob Heylen, Mario Parente, and Paul Gader. “A review of nonlinear hyperspectral unmixing methods”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7.6 (2014), pp. 1844–1868.

[179] Bin Yang, Bin Wang, and Zongmin Wu. “Nonlinear hyperspectral unmixing based on geometric characteristics of bilinear mixture models”. In: *IEEE Transactions on Geoscience and Remote Sensing* 56.2 (2017), pp. 694–714.

[180] Laura-Bianca Bilius and Stefan Gheorghe Pentiuc. “Improving the analysis of hyperspectral images using tensor decomposition”. In: *2020 international conference on development and application systems (DAS)*. IEEE. 2020, pp. 173–176.

[181] Rasmus Bro. “PARAFAC. Tutorial and applications”. In: *Chemometrics and intelligent laboratory systems* 38.2 (1997), pp. 149–171.

[182] Samiran Das and Sandip Ghosal. “Unmixing aware compression of hyperspectral image by rank aware orthogonal parallel factorization decomposition”. In: *Journal of Applied Remote Sensing* 17.4 (2023), p. 046509.

[183] Joaquim Jaumot, Anna de Juan, and Romà Tauler. “MCR-ALS GUI 2.0: New features and applications”. In: *Chemometrics and Intelligent Laboratory Systems* 140 (2015), pp. 1–12.

[184] Mohammed Alaoui Mansouri, Mourad Kharbach, and Abdelaziz Bouklouze. “Current applications of Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) in pharmaceutical analysis”. In: *Journal of Pharmaceutical Sciences* 113.4 (2024), pp. 856–865.

[185] Xin Zhang and Romà Tauler. “Application of multivariate curve resolution alternating least squares (MCR-ALS) to remote sensing hyperspectral imaging”. In: *Analytica chimica acta* 762 (2013), pp. 25–38.

[186] Mónica B Mamián-López and Ronei J Poppi. “SERS hyperspectral imaging assisted by MCR-ALS for studying polymeric microfilms loaded with paracetamol”. In: *Microchemical Journal* 123 (2015), pp. 243–251.

[187] Judith Felten et al. “Vibrational spectroscopic image analysis of biological material using multivariate curve resolution–alternating least squares (MCR-ALS)”. In: *Nature protocols* 10.2 (2015), pp. 217–240.

[188] Cyril Ruckebusch. *Resolving spectral mixtures: with applications from ultrafast time-resolved spectroscopy to super-resolution imaging*. Elsevier, 2016.

[189] Raymond B Cattell. “The scree test for the number of factors”. In: *Multivariate behavioral research* 1.2 (1966), pp. 245–276.

[190] Willem Windig and Jean Guilment. “Interactive self-modeling mixture analysis”. In: *Analytical chemistry* 63.14 (1991), pp. 1425–1432.

[191] Arnold Peter Snyder, Willem Windig, and John P Toth. “Interactive self-modeling multivariate analysis of thermolysis mass spectra”. In: *Chemometrics and intelligent laboratory systems* 11.2 (1991), pp. 149–160.

[192] Willem Windig. “Spectral data files for self-modeling curve resolution with examples using the Simplisma approach”. In: *Chemometrics and Intelligent Laboratory Systems* 1.36 (1997), pp. 3–16.

[193] Mathias Sawall et al. “An active constraint approach to identify essential spectral information in noisy data”. In: *Analytica Chimica Acta* 1233 (2022), p. 340448.

[194] Mahdiyeh Ghaffari, Nematollah Omidikia, and Cyril Ruckebusch. “Joint selection of essential pixels and essential variables across hyperspectral images”. In: *Analytica Chimica Acta* 1141 (2021), pp. 36–46.

[195] Laureen Coic et al. “Selection of essential spectra to improve the multivariate curve resolution of minor compounds in complex pharmaceutical formulations”. In: *Analytica Chimica Acta* 1198 (2022), p. 339532.

[196] F Cuesta Sánchez and Desiré Luc Massart. “Application of SIMPLISMA for the assessment of peak purity in liquid chromatography with diode array detection”. In: *Analytica chimica acta* 298.3 (1994), pp. 331–339.

[197] Guoxiang Chen and Peter de Boves Harrington. “SIMPLISMA applied to two-dimensional wavelet compressed ion mobility spectrometry data”. In: *Analytica chimica acta* 484.1 (2003), pp. 75–91.

[198] Ian GM Anthony et al. “Library-Integrated SIMPLISMA-ALS Deconvolution of Gas Chromatography-Vacuum Ultraviolet Absorption Spectroscopy Data”. In: *SSRN Electronic Journal* (2022).

[199] Laureen Coic et al. “Assessment of essential information in the fourier domain to accelerate raman hyperspectral microimaging”. In: *Analytical Chemistry* 95.42 (2023), pp. 15497–15504.

[200] Rasmus Bro and Nicholaos D Sidiropoulos. “Least squares algorithms under unimodality and non-negativity constraints”. In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 12.4 (1998), pp. 223–247.

[201] Romà Tauler, Age Smilde, and Bruce Kowalski. “Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution”. In: *Journal of Chemometrics* 9.1 (1995), pp. 31–58.

[202] Anna De Juan et al. “Assessment of new constraints applied to the alternating least squares method”. In: *Analytica Chimica Acta* 346.3 (1997), pp. 307–318.

[203] Adrian Gomez-Sanchez et al. “The trilinear constraint adapted to solve data with strong patterns of outlying observations or missing values”. In: *Chemometrics and Intelligent Laboratory Systems* 231 (2022), p. 104692.

[204] Charles L Lawson and Richard J Hanson. *Solving least squares problems*. SIAM, 1995.

[205] Rasmus Bro and Sijmen De Jong. “A fast non-negativity-constrained least squares algorithm”. In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 11.5 (1997), pp. 393–401.

[206] Siewert Hugelier, Olivier Devos, and Cyril Ruckebusch. “On the implementation of spatial constraints in multivariate curve resolution alternating least squares for hyperspectral image analysis”. In: *Journal of Chemometrics* 29.10 (2015), pp. 557–561.

[207] Raffaele Vitale et al. “A spatial constraint to model and extract texture components in Multivariate Curve Resolution of near-infrared hyperspectral images”. In: *Analytica chimica acta* 1095 (2020), pp. 30–37.

[208] Anna de Juan et al. “Local rank analysis for exploratory spectroscopic image analysis. Fixed size image window-evolving factor analysis”. In: *Chemometrics and Intelligent Laboratory Systems* 77.1-2 (2005), pp. 64–74.

[209] Xin Zhang, Anna de Juan, and Romà Tauler. “Local rank-based spatial information for improvement of remote sensing hyperspectral imaging resolution”. In: *Talanta* 146 (2016), pp. 1–9.

[210] Anna de Juan et al. “Use of local rank-based spatial information for resolution of spectroscopic images”. In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 22.5 (2008), pp. 291–298.

[211] Anna De Juan, Joaquim Jaumot, and Romà Tauler. “Multivariate Curve Resolution (MCR). Solving the mixture analysis problem”. In: *Analytical Methods* 6.14 (2014), pp. 4964–4976.

[212] Adrián Gómez Sánchez. “Development and application of new strategies for data fusion of hyperspectral images”. PhD thesis. Université de Lille; Universitat de Barcelona, 2024.

[213] Mathias Sawall et al. “Simultaneous construction of dual Borgen plots. I: The case of noise-free data”. In: *Journal of Chemometrics* 31.12 (2017), e2954.

[214] Odd S Borgen and Bruce R Kowalski. “An extension of the multivariate component-resolution method to three components”. In: *Analytica Chimica Acta* 174 (1985), pp. 1–26.

[215] Azadeh Golshan et al. “A review of recent methods for the determination of ranges of feasible solutions resulting from soft modelling analyses of multivariate data”. In: *Analytica Chimica Acta* 911 (2016), pp. 1–13.

[216] Mahdiyeh Ghaffari et al. “Effect of image processing constraints on the extent of rotational ambiguity in MCR-ALS of hyperspectral images”. In: *Analytica Chimica Acta* 1052 (2019), pp. 27–36.

[217] Mohamad Ahmad. “Novel chemometric tools for the unmixing of complex mixtures in spectral imaging considering spatial-spectral information and their interplay”. PhD thesis. Université de Lille, 2023.

[218] Roma Tauler and Damià Barceló. “Multivariate curve resolution applied to liquid chromatography—diode array detection”. In: *TrAC Trends in Analytical Chemistry* 12.8 (1993), pp. 319–327.

[219] Marc Marin-Garcia and Roma Tauler. “Chemometrics characterization of the Llobregat river dissolved organic matter”. In: *Chemometrics and Intelligent Laboratory Systems* 201 (2020), p. 104018.

[220] Marta Alier et al. “Trilinearity and component interaction constraints in the multivariate curve resolution investigation of NO and O3 pollution in Barcelona”. In: *Analytical and bioanalytical chemistry* 399 (2011), pp. 2015–2029.

[221] Adrián Gómez-Sánchez et al. “The MCR-ALS trilinearity constraint for data with missing values”. In: *Journal of Chemometrics* 38.11 (2024), e3584.

[222] Adrián Gómez-Sánchez et al. “3D and 4D image fusion: Coping with differences in spectroscopic modes among hyperspectral images”. In: *Analytical Chemistry* 92.14 (2020), pp. 9591–9602.

[223] Adrián Gómez-Sánchez et al. “Study of the photobleaching phenomenon to optimize acquisition of 3D and 4D fluorescence images. A special scenario for trilinear and quadrilinear models”. In: *Microchemical Journal* 191 (2023), p. 108899.

[224] Evrim Acar and Bülent Yener. “Unsupervised multiway data analysis: A literature survey”. In: *IEEE transactions on knowledge and data engineering* 21.1 (2008), pp. 6–20.

[225] Tamara G Kolda and Brett W Bader. “Tensor decompositions and applications”. In: *SIAM review* 51.3 (2009), pp. 455–500.

[226] Lieven De Lathauwer. “A survey of tensor methods”. In: *2009 IEEE international symposium on circuits and systems*. IEEE. 2009, pp. 2773–2776.

[227] Richard A Harshman et al. “Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multi-modal factor analysis”. In: *UCLA working papers in phonetics* 16.1 (1970), p. 84.

[228] J Douglas Carroll and Jih-Jie Chang. “Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition”. In: *Psychometrika* 35.3 (1970), pp. 283–319.

[229] Ledyard R Tucker. “Some mathematical notes on three-mode factor analysis”. In: *Psychometrika* 31.3 (1966), pp. 279–311.

[230] Ledyard R Tucker. “Implications of factor analysis of three-way matrices for measurement of change”. In: *Problems in measuring change* 15.122-137 (1963), p. 3.

[231] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. “A multilinear singular value decomposition”. In: *SIAM journal on Matrix Analysis and Applications* 21.4 (2000), pp. 1253–1278.

[232] Lieven De Lathauwer. “Decompositions of a higher-order tensor in block terms—Part II: Definitions and uniqueness”. In: *SIAM Journal on Matrix Analysis and Applications* 30.3 (2008), pp. 1033–1066.

[233] Mariya Ishteva et al. “Dimensionality reduction for higher-order tensors: algorithms and applications”. In: *International Journal of Pure and Applied Mathematics* 42.3 (2008), p. 337.

[234] Henk AL Kiers. “Towards a standardized notation and terminology in multiway analysis”. In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 14.3 (2000), pp. 105–122.

[235] Na Liu et al. “A survey on hyperspectral image restoration: From the view of low-rank tensor approximation”. In: *Science China Information Sciences* 66.4 (2023), p. 140302.

[236] William Navas-Auger and Vidya Manian. “Spatial Low-Rank Tensor Factorization and Unmixing of Hyperspectral Images”. In: *Computers* 10.6 (2021), p. 78.

[237] Athanasios A Rontogiannis, Eleftherios Kofidis, and Paris V Giampouras. “Block-term tensor decomposition: Model selection and computation”. In: *IEEE Journal of Selected Topics in Signal Processing* 15.3 (2021), pp. 464–475.

[238] Xu Han et al. “Robust multilinear decomposition of low rank tensors”. In: *International Conference on Latent Variable Analysis and Signal Separation*. Springer. 2018, pp. 3–12.

[239] Xu Han. “Robust low-rank tensor approximations using group sparsity”. PhD thesis. Rennes 1, 2019.

[240] Paris V Giampouras, Athanasios A Rontogiannis, and Konstantinos D Koutroumbas. “Alternating iteratively reweighted least squares minimization for low-rank matrix factorization”. In: *IEEE Transactions on Signal Processing* 67.2 (2018), pp. 490–503.

[241] Mingyi Hong et al. “A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing”. In: *IEEE Signal Processing Magazine* 33.1 (2015), pp. 57–77.

[242] P-Y Sacré et al. “Evaluation of distributional homogeneity of pharmaceutical formulation using laser direct infrared imaging”. In: *International Journal of Pharmaceutics* 612 (2022), p. 121373.

[243] Cristina Manis et al. “Non-destructive age estimation of biological fluid stains: An integrated analytical strategy based on near-infrared hyperspectral imaging and multivariate regression”. In: *Talanta* 245 (2022), p. 123472.

[244] Martin C Schodlok, Michaela Frei, and Karl Segl. “Implications of new hyperspectral satellites for raw materials exploration”. In: *Mineral Economics* 35.3 (2022), pp. 495–502.

[245] Anna de Juan and Romà Tauler. “Introduction to Linear Soft-Modeling”. In: *Comprehensive chemometrics: chemical and biochemical data analysis*. Elsevier, 2020.

[246] Alessandra Olarini et al. “Exploratory analysis of hyperspectral imaging data”. In: *Chemometrics and Intelligent Laboratory Systems* 252 (2024), p. 105174.

[247] Patrizia Firmani et al. “MCR-ALS of hyperspectral images with spatio-spectral fuzzy clustering constraint”. In: *Chemometrics and Intelligent Laboratory Systems* 179 (2018), pp. 85–91.

[248] Xiaoyu Cheng et al. “A spatial-spectral clustering-based algorithm for end-member extraction and hyperspectral unmixing”. In: *International Journal of Remote Sensing* 42.5 (2021), pp. 1948–1972.

[249] Xiang Xu et al. “Regional clustering-based spatial preprocessing for hyperspectral unmixing”. In: *Remote Sensing of Environment* 204 (2018), pp. 333–346.

[250] Sara Piqueras et al. “Combining multiset resolution and segmentation for hyperspectral image analysis of biological tissues”. In: *Analytica Chimica Acta* 881 (2015), pp. 24–36.

[251] Turgay Celik. “Unsupervised change detection in satellite images using principal component analysis and k-means clustering”. In: *IEEE geoscience and remote sensing letters* 6.4 (2009), pp. 772–776.

[252] Irem Ersöz Kaya et al. “PCA based clustering for brain tumor segmentation of T1w MRI images”. In: *Computer methods and programs in biomedicine* 140 (2017), pp. 19–28.

[253] Tristan D McRae et al. “Robust blind spectral unmixing for fluorescence microscopy using unsupervised learning”. In: *Plos one* 14.12 (2019), e0225410.

[254] Vijay Sadashivaiah et al. “SUFI: an automated approach to spectral unmixing of fluorescent multiplex images captured in mouse and post-mortem human brain tissues”. In: *BMC neuroscience* 24.1 (2023), p. 6.

[255] Yurong Gao and Tristan McRae. “K-means Spectral Unmixing for Multi-channel Imaging and Image Analysis Platform at a Core Facility”. In: *Journal of Biomolecular Techniques: JBT* 30.Suppl (2019), S24.

[256] Mahdiyeh Ghaffari, Nematollah Omidikia, and Cyril Ruckebusch. “Essential spectral pixels for multivariate curve resolution of chemical images”. In: *Analytical chemistry* 91.17 (2019), pp. 10943–10948.

[257] Victor HC Ferreira et al. “Enhancing Diagnostic Capabilities for Occupational Lung Diseases Using LIBS Imaging on Biopsy Tissue”. In: *Analytical Chemistry* 96.18 (2024), pp. 7038–7046.

[258] Somaye Vali Zade et al. “Data point importance: Information ranking in multivariate data”. In: *Journal of Chemometrics* 37.1 (2023), e3453.

[259] Somaiyeh Khodadadi Karimvand et al. “Practical and comparative application of efficient data reduction-Multivariate curve resolution”. In: *Analytica Chimica Acta* 1243 (2023), p. 340824.

[260] Raffaele Vitale and Cyril Ruckebusch. “On a black hole effect in bilinear curve resolution based on least squares”. In: *Journal of Chemometrics* 37.2 (2023), e3442.

[261] Gene H Golub and Christian Reinsch. “Singular value decomposition and least squares solutions”. In: *Handbook for Automatic Computation: Volume II: Linear Algebra*. Springer, 1971, pp. 134–151.

[262] Róbert Rajkó. “Studies on the adaptability of different Borgen norms applied in self-modeling curve resolution (SMCR) method”. In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 23.6 (2009), pp. 265–274.

[263] Bjørn-Vidar Grande and Rolf Manne. “Use of convexity for finding pure variables in two-way data from mixtures”. In: *Chemometrics and intelligent laboratory systems* 50.1 (2000), pp. 19–33.

[264] Enmanuel Cruz Muñoz et al. “Characterization of pyrite weathering products by Raman hyperspectral imaging and chemometrics techniques”. In: *Microchemical Journal* 190 (2023), p. 108655.

[265] Samuel Moncayo et al. “Exploration of megapixel hyperspectral LIBS images using principal component analysis”. In: *Journal of Analytical Atomic Spectrometry* 33.2 (2018), pp. 210–220.

[266] Guilin Hu et al. “Decomposition and oxidation of pyrite”. In: *Progress in energy and combustion science* 32.3 (2006), pp. 295–314.

[267] Xueding Wang and Ying Guo. “The impact of trace metal cations and absorbed water on colour transition of turquoise”. In: *Royal Society Open Science* 8.2 (2021), p. 201110.

[268] Jeremiah Rushchitsky. “Interaction of waves in solid mixtures”. In: *Applied Mechanics Reviews* (1999).

[269] URL: https://www.mindat.org. (accessed: 18.10.2023).

[270] URL: https://www.atomtrace.com/elements-database. (accessed: 18.10.2023).

[271] Age K Smilde. "Three-way analyses problems and prospects". In: *Chemometrics and Intelligent Laboratory Systems* 15.2-3 (1992), pp. 143–157.

[272] Fatima Zohra Benhalouche et al. "Hyperspectral unmixing based on constrained bilinear or linear-quadratic matrix factorization". In: *Remote Sensing* 13.11 (2021), p. 2132.

[273] Frederic Estienne et al. "Multi-way modelling of high-dimensionality electroencephalographic data". In: *Chemometrics and Intelligent Laboratory Systems* 58.1 (2001), pp. 59–72.

[274] Sébastien Gourvénec et al. "Monitoring batch processes with the STATIS approach". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 19.5-7 (2005), pp. 288–300.

[275] Evrim Acar et al. "Modeling and multiway analysis of chatroom tensors". In: *Intelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics, ISI 2005, Atlanta, GA, USA, May 19-20, 2005. Proceedings 3*. Springer. 2005, pp. 256–268.

[276] Mary Beth Seasholtz and Bruce Kowalski. "The parsimony principle applied to multivariate calibration". In: *Analytica Chimica Acta* 277.2 (1993), pp. 165–177.

[277] Joseph B Kruskal. "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics". In: *Linear algebra and its applications* 18.2 (1977), pp. 95–138.

[278] Sue E Leurgans, Robert T Ross, and Rebecca B Abel. "A decomposition for three-way arrays". In: *SIAM Journal on Matrix Analysis and Applications* 14.4 (1993), pp. 1064–1083.

[279] Raymond B Cattell. ""Parallel proportional profiles" and other principles for determining the choice of factors by rotation". In: *Psychometrika* 9.4 (1944), pp. 267–283.

[280] Richard A Harshman and Margaret E Lundy. "PARAFAC: Parallel factor analysis". In: *Computational Statistics & Data Analysis* 18.1 (1994), pp. 39–72.

[281] Alwin Stegeman, Jos MF Ten Berge, and Lieven De Lathauwer. "Sufficient conditions for uniqueness in Candecomp/Parafac and Indscal with random component matrices". In: *Psychometrika* 71.2 (2006), pp. 219–229.

[282] Rasmus Bro et al. "Modeling multi-way data with linearly dependent loadings". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 23.7-8 (2009), pp. 324–340.

[283] Charlotte Møller Andersen and Rasmus Bro. "Practical aspects of PARAFAC modeling of fluorescence excitation-emission data". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 17.4 (2003), pp. 200–215.

[284] Beatriz Quintanilla-Casas et al. "Using fluorescence excitation-emission matrices to predict bitterness and pungency of virgin olive oil: A feasibility study". In: *Food Chemistry* 395 (2022), p. 133602.

[285] Marina Cocchi et al. "Analysis of sensory data of Aceto Balsamico Tradizionale di Modena (ABTM) of different ageing by application of PARAFAC models". In: *Food quality and preference* 17.6 (2006), pp. 419–428.

[286] Morten Mørup et al. "Parallel factor analysis as an exploratory tool for wavelet transformed event-related EEG". In: *NeuroImage* 29.3 (2006), pp. 938–947.

[287] Xuefeng Liu, Salah Bourennane, and Caroline Fossati. "Denoising of hyperspectral images using the PARAFAC model and statistical performance analysis". In: *IEEE Transactions on Geoscience and Remote Sensing* 50.10 (2012), pp. 3717–3724.

[288] Jérémy Cohen, Rasmus Bro, and Pierre Comon. "Tensor decompositions: principles and application to food sciences". In: *Source Separation in Physical-Chemical Sensing* (2023), pp. 255–323.

[289] Pieter M Kroonenberg and Jan De Leeuw. "Principal component analysis of three-mode data by means of alternating least squares algorithms". In: *Psychometrika* 45 (1980), pp. 69–97.

[290] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. "On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors". In: *SIAM journal on Matrix Analysis and Applications* 21.4 (2000), pp. 1324–1342.

[291] Dimitri Nion and Lieven De Lathauwer. "A tensor-based blind DS-CDMA receiver using simultaneous matrix diagonalization". In: *2007 IEEE 8th Workshop on Signal Processing Advances in Wireless Communications*. IEEE. 2007, pp. 1–5.

[292] Henk AL Kiers and Richard A Harshman. "Relating two proposed methods for speedup of algorithms for fitting two-and three-way principal component and related multilinear models". In: *Chemometrics and Intelligent Laboratory Systems* 36.1 (1997), pp. 31–40.

[293] Lieven De Lathauwer and Joos Vandewalle. "Dimensionality reduction in higher-order signal processing and rank-(R1, R2,…, RN) reduction in multilinear algebra". In: *Linear Algebra and its Applications* 391 (2004), pp. 31–55.

[294] Fatma Allouche et al. "Coupling hyperspectral image data having different spatial resolutions by extending multivariate inter-battery Tucker analysis". In: *Chemometrics and Intelligent Laboratory Systems* 113 (2012), pp. 43–51.

[295] M Alex O Vasilescu and Demetri Terzopoulos. "Multilinear subspace analysis of image ensembles". In: *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.* Vol. 2. IEEE. 2003, pp. II–93.

[296] M Alex O Vasilescu and Demetri Terzopoulos. "TensorTextures: Multilinear image-based rendering". In: *ACM Transactions on Graphics*. Vol. 23. Association for Computing Machinery, 2004, pp. 336–342.

[297] Jean-Michel Papy, Lieven De Lathauwer, and Sabine Van Huffel. "Exponential data fitting using multilinear algebra: the single-channel and multi-channel case". In: *Numerical linear algebra with applications* 12.8 (2005), pp. 809–826.

[298] Berkant Savas and Lars Eldén. "Handwritten digit classification using higher order singular value decomposition". In: *Pattern recognition* 40.3 (2007), pp. 993–1003.

[299] Clémence Prévost. "Multimodal data fusion by coupled low-rank tensor approximations". PhD thesis. Université de Lorraine, 2021.

[300] Hao Guo et al. "Multispectral and hyperspectral image fusion based on regularized coupled non-negative block-term tensor decomposition". In: *Remote Sensing* 14.21 (2022), p. 5306.

[301] Xing Zhang, Gongjian Wen, and Wei Dai. "Target representation in hyperspectral images based on tensor block term decomposition". In: *2015 8th International Congress on Image and Signal Processing (CISP)*. IEEE. 2015, pp. 793–798.

[302] Guoyong Zhang et al. "Hyperspectral super-resolution: A coupled nonnegative block-term tensor decomposition approach". In: *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE. 2019, pp. 470–474.

[303] Nico Vervliet, Otto Debals, and Lieven De Lathauwer. "Tensorlab 3.0—numerical optimization strategies for large-scale constrained and coupled matrix/tensor factorization". In: *2016 50th Asilomar Conference on Signals, Systems and Computers*. IEEE. 2016, pp. 1733–1738.

[304] Laurent Sorber, Marc Van Barel, and Lieven De Lathauwer. "Structured data fusion". In: *IEEE journal of selected topics in signal processing* 9.4 (2015), pp. 586–600.

[305] Xiao Fu et al. "Nonconvex optimization tools for large-scale matrix and tensor decomposition with structured factors". In: *arXiv preprint arXiv:2006.08183* (2020).

[306] Laurent Sorber, Marc Van Barel, and Lieven De Lathauwer. "Unconstrained optimization of real functions in complex variables". In: *SIAM Journal on Optimization* 22.3 (2012), pp. 879–898.

[307] Pieter M Kroonenberg. *Applied multiway data analysis*. John Wiley & Sons, 2008.

[308] Julio Longina Castellanos, Susana Gómez, and Valia Guerra. "The triangle method for finding the corner of the L-curve". In: *Applied Numerical Mathematics* 43.4 (2002), pp. 359–373.

[309] Eva Ceulemans and Henk AL Kiers. "Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method". In: *British journal of mathematical and statistical psychology* 59.1 (2006), pp. 133–150.

[310] Frederik Van Eeghem and Lieven De Lathauwer. “Tensor Similarity in Chemometrics”. In: *Comprehensive chemometrics: chemical and biochemical data analysis*. Elsevier, 2020.

[311] Claus A Andersson and Rasmus Bro. “The N-way toolbox for MATLAB”. In: *Chemometrics and intelligent laboratory systems* 52.1 (2000), pp. 1–4.

[312] Henk AL Kiers and Iven Van Mechelen. “Three-way component analysis: Principles and illustrative application.” In: *Psychological methods* 6.1 (2001), p. 84.

[313] Wayne S DeSarbo. “An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques”. In: *Research methods for multimode data analysis* (1984), pp. 602–642.

[314] Rasmus Bro. “Multi-way analysis in the food industry, model, algorithms and applications”. PhD thesis. University of Amsterdam, 1998.

[315] Carolina S Silva et al. “Detecting semen stains on fabrics using near infrared hyperspectral images and multivariate models”. In: *TrAC Trends in Analytical Chemistry* 95 (2017), pp. 23–35.

[316] URL: `https://rslab.ut.ac.ir/data`. (accessed: 14.09.2022).

[317] Mingle Zhang et al. “A Global Spatial-Spectral Feature Fused Autoencoder for Nonlinear Hyperspectral Unmixing”. In: *Remote Sensing* 16.17 (2024), p. 3149.

[318] Daria Siciliano et al. “Evaluating hyperspectral imaging of wetland vegetation as a tool for detecting estuarine nutrient enrichment”. In: *Remote Sensing of Environment* 112.11 (2008), pp. 4020–4033.

[319] Heidi M Dierssen et al. “Pushing the limits of seagrass remote sensing in the turbid waters of Elkhorn Slough, California”. In: *Remote Sensing* 11.14 (2019), p. 1664.

[320] Kristin B Byrd. “Remote sensing and spatial analysis of watershed and estuarine processes for conservation planning in Elkhorn Slough, Monterey County, California”. In: *Remote Sensing and Geospatial Technologies for Coastal Ecosystem Assessment and Management*. Springer, 2008, pp. 495–520.

[321] Jerry Workman Jr and Lois Weyer. *Practical guide to interpretive near-infrared spectroscopy*. CRC press, 2007.

[322] Mohammad Ali et al. “Spectroscopic studies of the ageing of cellulosic paper”. In: *Polymer* 42.7 (2001), pp. 2893–2900.

[323] Yelu Zeng et al. “Estimating near-infrared reflectance of vegetation from hyperspectral data”. In: *Remote Sensing of Environment* 267 (2021), p. 112723.

[324] Jiyou Zhu et al. “Response of plant reflectance spectrum to simulated dust deposition and its estimation model”. In: *Scientific Reports* 10.1 (2020), p. 15803.

[325] Keith D Shepherd and Markus G Walsh. “Development of reflectance spectral libraries for characterization of soil properties”. In: *Soil science society of America journal* 66.3 (2002), pp. 988–998.

[326] José AM Demattê et al. “Visible–NIR reflectance: a new approach on soil evaluation”. In: *Geoderma* 121.1-2 (2004), pp. 95–112.

[327] Xu Han et al. “Block term decomposition with rank estimation using group sparsity”. In: *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE. 2017, pp. 1–5.

[328] Tatsuya Yokota, Namgil Lee, and Andrzej Cichocki. “Robust multilinear tensor rank estimation using higher order singular value decomposition and information criteria”. In: *IEEE Transactions on Signal Processing* 65.5 (2016), pp. 1196–1206.

[329] José M Bioucas-Dias and José MP Nascimento. “Hyperspectral subspace identification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 46.8 (2008), pp. 2435–2445.

[330] Raffaele Vitale et al. “Three-Way Data Reduction Based on Essential Information”. In: *Journal of Chemometrics* 38.12 (2024), e3617.

[331] Hao Zhang et al. “Sparsity Regularized Rank-(L, M, N) Block Term Decomposition for Hyperspectral Image Mixed Noise Removal”. In: *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2024, pp. 7614–7617.

[332] Eric Aguado-Sarrió et al. “Virtual biopsies for breast cancer using MCR-ALS perfusion-based biomarkers and double cross-validation PLS-DA”. In: *Chemometrics and Intelligent Laboratory Systems* (2024), p. 105152.

[333] Riccardo Melis, Ilaria Vitangeli, and Roberto Anedda. “Effect of fish diet and cooking mode on the composition and microstructure of ready-to-eat fish fillets of gilthead sea bream (Sparus aurata) juveniles”. In: *Journal of Food Composition and Analysis* 114 (2022), p. 104847.