Université de Lille

LASIRE

Université de Lille – Sciences et Technologies

École doctorale des Sciences de la Matière, du Rayonnement et de l'Environnement

Thèse de Doctorat

En vue de l'obtention due grade de

Docteur de l'Université de Lille

Discipline: Optique et Lasers, Physico-Chimie et Atmosphère

Spécialité du doctorat: Chimie Theorique, Physique, Analytique

Cotutelle avec:

L'Université de Modène et de Reggio d'Émilie

Mohamad Ahmad

Soutenance: 5 mai 2023

Novel chemometric tools for the unmixing of complex mixtures in spectral imaging considering spatial-spectral information and their interplay

---

Nouveaux outils chimiométriques pour le démélange de mélanges complexes en imagerie spectrale en tenant compte des informations spatio-spectrales et de leur interaction

Rapporteurs:

(présidente du jury) Mme Beata WALCZAK	Professeur, Uniwersytet Śląski w Katowicach
M. Jose MANUEL AMIGO	Professeur, Euskal Herriko Unibertsitatea, Leioa

Examinateurs:

M. Ludovic DUPONCHEL	Professeur, Université de Lille
Mme Caterina DURANTE	Docteur, Università di Modena e Reggio Emilia
M. Federico MARINI	Professeur, Sapienza Università di Roma

Co-Directeur de Thèse:

M. Cyril RUCKEBUSCH	Professeur, Université de Lille
Mme Marina COCCHI	Professeur, Università di Modena e Reggio Emilia

Preface

I am honoured to present this thesis, which represents the culmination of my three-year journey as a doctoral student in chemometrics at Unimore and the University of Lille. The experience has been enriching, challenging, and transformative, and I am grateful to have had the opportunity to study at both universities, each with its own unique aspects.

Throughout my studies, I have had the privilege of working with brilliant and dedicated researchers and mentors who have provided me with invaluable guidance, support, and inspiration. I would like to specifically name Cyril, Marina and Raffaele for their incredible support, expertise and patience with me. I am incredibly grateful to everybody that has had a hand in helping me grow as a researcher and person during these years.

I have been very lucky to have met some great people during my studies that I can call my close-friends, peers and companions in crime. My two hands and feet are not enough to count the number, and this short preface won't have space for it either. I will leave it with: "You know who you are."

This thesis represents not only my doctoral studies, but also the commencement of a new chapter in my academic and professional career. I am confident that the vast knowledge, skills, and experience that I have gained through this program will serve me well in all my future endeavours.

Mohamad

List of publications

Primary work:

Weighted multivariate curve resolution – alternating least squares based on sample relevance [1]

Journal of Chemometrics, 2023-03

Authors: *Mohamad Ahmad*; Raffaele Vitale; Marina Cocchi; Cyril Ruckebusch

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) [2]

Analytica Chimica Acta, 2022-01

Authors: Mohamad Ahmad; Raffaele Vitale; Carolina S. Silva; Cyril Ruckebusch; *Marina Cocchi*

Exploring local spatial features in hyperspectral images [3]

Journal of Chemometrics, 2020-08

Authors: Mohamad Ahmad; Raffaele Vitale; Carolina S. Silva; Cyril Ruckebusch; *Marina Cocchi*

Auxiliary work:

A New Protocol of Computer-Assisted Image Analysis Highlights the Presence of Hemocytes in the Regenerating Cephalic Tentacles of Adult Pomacea canaliculata [4]

International Journal of Molecular Sciences, 2021-05

Authors: Giulia Bergamini; Mohamad Ahmad; Marina Cocchi; *Davide Malagoli*

Abstract

Hyperspectral imaging is used in many fields of science, for its powerful ability to gather information in both the spatial and spectral domain. Applications range from cell imaging in biology to remote sensing in agricultural sciences. However, in many applications and methods, the information within the spectral image is not fully utilized, leading to sub-optimal results. One of the more prominent concerns is spatial correlation, as it is either completely disregarded or considered in a sub-optimal way. In fact, the spatial dimension is usually unfolded pixelwise to provide a two-dimensional data matrix prior to applying chemometric methods, such as decomposition by Principal Component Analysis (PCA) or Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS). In the thesis project, novel tools have been developed to simultaneously consider the spatial and spectral dimensions, as well as to improve mixture resolution in challenging situations, where current tools might fail to retrieve adequate solutions.
The first developed tool is named "Image Decomposition, Encoding and Localization" (IDEL) which takes a spatial approach to spectral image analysis maintaining the link to the spectral features that are responsible for the observed spatial patterns. IDEL combines an exploitation step based on wavelet filters, with a compression step based on image encoding and multivariate data analysis, and a final reconstruction in the original domain of the most essential spatial-spectral information. IDEL has been applied on several benchmarks representing a varied set of situations, simulated and real. At first, simulations were developed to apply the approach in a controlled manner, then moving to challenging real case studies i.e., recognition of biological traces dispersed on different surfaces in the context of forensic analysis, or food component distributions in food quality control. Results showed the capability of the methodology, obtaining a new perspective on the data that was previously difficult to access. The second tool tackles the

very critical issue of under-represented minor components in mixtures that the standard MCR-ALS algorithm, fails to recover, especially in presence of noise. The proposed method is based on a weighting scheme tuned on sample (of a given mixture composition) relevance with respect to the degree of purity each sample represents. This novel tool, named weighted MCR-ALS has been tested in simulated data, and then applied to a pseudo real image of a pharmaceutical tablet, where considerable gains are obtained, with respect to the current state-of-the art. Both tools have been coded in MATLAB. The IDEL function has been developed including a default set of parameters that are applicable in different situations and ready to be utilized by non-experts of the field. At the same time parameter tuning can be handled by expert users. Both tools are set up to work across a broad range of situations.

Keywords: spectral imaging, wavelet transform, image analysis, decomposition, spectral unmixing, multivariate curve resolution

Riassunto

L'imaging iperspettrale e utilizzato in molti ambiti scientifici, per la sua capacità di acquisire informazioni sia nel dominio spaziale che in quello spettrale. I campi di applicazioni si estendono da tecniche di imaging (Raman, multicanale, etc.) nello studio delle cellule in biologia al telerilevamento nelle scienze agrarie. Tuttavia, in molte applicazioni le metodologie sviluppate non utilizzano completamente le informazioni contenute nell'immagine spettrale, portando a risultati non ottimali. Uno dei problemi più rilevanti e che la correlazione spaziale viene completamente ignorata o considerata in maniera inadeguata. Infatti, la maggiorparte degli approcci prevede l'unfolding pixelwise della dimensione spaziale per fornire una matrice di dati bidimensionale, che sarà poi sottoposta all'applicazione di metodi chemiometrici, come la decomposizione mediante l'analisi delle componenti principali (PCA) o la risoluzione spettrale mediante il metodo Multivariate Curve Resolution (MCR-ALS). Nel progetto di tesi sono stati sviluppati nuovi approcci/ algoritmi per considerare simultaneamente le dimensioni spaziale e spettrale e per migliorare la risoluzione dei componenti puri in miscele complesse, dove i metodi attuali forniscono soluzioni non del tutto adeguate.
Il primo algoritmo sviluppato e stato chiamato "Image Decomposition, Encoding and Localization" (IDEL), e adotta un approccio spaziale nell'analisi delle immagini spettrali, evidenziando la corrispondenza tra i patterns spaziali osservati ed i profili spettrali che ne sono responsabili. IDEL utilizza uno step di esplorazione/espansione basato su filtri wavelet, combinato con uno step di compressione basato sulla codifica delle immagini mediante descrittori. La successiva analisi multi-variata dei dati di imagine compressi consente di catturare l'informazione spaziale-spettrale più essenziale, ed infine la sua ricostruzione nel dominio originale. IDEL e stato applicato a diversi benchmarks che descrivono diverse situazioni, simulate e reali. Inizialmente sono state messe a punto simulazioni per applicare l'approccio in modo controllato,

per poi passare a casi di studio reali impegnativi, come il riconoscimento di tracce biologiche disperse su diverse superfici nel contesto dell'analisi forense o la distribuzione di componenti alimentari nel controllo della qualità degli alimenti. I risultati hanno dimostrato la capacita della metodologia, ottenendo una nuova prospettiva sui dati che in precedenza erano di difficile comprensione. Il secondo approccio/algoritmo risolve il problema molto critico della presenza all'interno di una miscela di una o più componenti minoritarie, nel senso di presenti in numero limitato di pixels rispetto alle altre (caso che corrisponde ad esempio ad un componente in tracce), che l'algoritmo MCR-ALS standard non riesce ad isolare, soprattutto in presenza di rumore. Il metodo proposto si basa su uno schema di ponderazione tarato sulla rilevanza del campione, nel caso delle imagini un pixel, (di una data composizione della miscela) rispetto al grado di purezza che ciascun campione rappresenta. Questo nuovo strumento, denominato "weighted MCR-ALS", e stato testato su dati simulati e poi applicato ad un'immagine in spettroscopia Raman pseudo-reale di una compressa farmaceutica, ottenendo notevoli vantaggi rispetto all'attuale stato dell'arte. Entrambi gli algoritmi sono stati codificati in MATLAB. La funzione IDEL e stata sviluppata includendo un insieme di parametri pre definiti, applicabili in diverse situazioni e pronti per essere utilizzati dai non esperti del settore. Allo stesso tempo, la regolazione dei parametri può essere gestita da utenti esperti. Entrambi gli strumenti sono stati progettati per lavorare su un'ampia gamma di situazioni.

Parole chiave: analisi dell'immagine spettrale, wavelet transform, analisi dell'immagine, decomposizione, scomposizione spettrale, multivariate curve resolution

Résumé

L'imagerie hyperspectrale est utilisée dans de nombreux domaines scientifiques en raison de sa puissante capacité à acquérir des informations dans les domaines spatial et spectral. Ses applications vont de l'imagerie cellulaire en biologie à la télédétection en sciences agricoles. Cependant, dans de nombreuses applications et méthodes, les informations contenues dans l'image spectrale ne sont pas pleinement exploitées, ce qui conduit à des résultats sous-optimaux. L'un des problèmes les plus importants est la corrélation spatiale, qui est soit complètement ignorée, soit insuffisamment prise en compte. En effet, la dimension spatiale est généralement décomposée en pixels pour créer une matrice de données bidimensionnelle, à laquelle sera appliquée des méthodes chimiométriques, telles que la décomposition par l'analyse en composantes principales (ACP) ou la résolution multivariée de courbes par moindres carrés alternés (MCR-ALS). Dans le cadre du projet de cette thèse, de nouveaux outils ont été développés afin de prendre en compte les dimensions spatiales et spectrales de manière simultanée, et d'améliorer la résolution de mélanges dans des situations complexes, où les outils actuels ne sont pas toujours en mesure de trouver des solutions adéquates.
Le premier outil développé est appelé "Image Decomposition, Encoding and Localisation" (IDEL), et adopte une approche spatiale de l'analyse d'images spectrales, une méthode qui conserve le lien des caractéristiques spectrales responsables des modèles spatiaux observés. IDEL combine une étape d'exploitation basée sur les filtres d'ondelettes, à une phase de compression basée sur le codage d'image et l'analyse de données multivariées, et une reconstruction finale dans le domaine original des informations spatiales-spectrales les plus essentielles. IDEL a été appliqué à plusieurs cas de référence décrivant différentes situations, simulées et réelles. Dans un premier temps, des simulations ont été mises en place pour appliquer l'approche de manière contrôlée, avant de passer à des études de cas réels plus complexes, tels que la re-

connaissance de traces biologiques dispersées sur différentes surfaces dans le contexte de l'analyse médico-légale, ou la distribution de composants alimentaires dans le contrôle de la qualité des aliments. Les résultats ont démontré la capacité de la méthodologie, en offrant une nouvelle perspective sur des données qui étaient auparavant difficiles à comprendre. Le second outil résout le problème crucial des composants minoritaires dans les mélanges d'échantillons, que l'algorithme standard MCR-ALS ne parvient pas à récupérer, surtout en présence de bruit. La méthode proposée est basée sur un schéma de pondération calibré à la pertinence de l'échantillon (d'une composition donnée) par rapport au degré de pureté que chaque échantillon représente. Ce nouvel outil, appelé "MCR-ALS pondéré", a été testé sur des données simulées, puis appliqué à l'image pseudo-réelle d'un comprimé pharmaceutique, ce qui a permis d'obtenir d'avantage de données significatives par rapport aux techniques précédentes. Les deux outils ont été codés sur MATLAB. La fonction IDEL a été développée en incluant un ensemble de paramètres prédéfinis, applicables dans différentes situations et prêts à être utilisés à la fois par des experts dans le domaine, et par des non-spécialistes. Ces deux outils sont conçus pour fonctionner dans un large éventail de situations.

Mots clés : image spectrale, wavelet transform, décomposition, déconvolution spectrale, multivariate curve resolution

Samenvatting

Hyperspectral imaging wordt in veel wetenschapsgebieden gebruikt vanwege het krachtige vermogen om informatie te verzamelen in zowel het ruimtelijke als het spectrale domein. Toepassingen variëren van cel-imaging in de biologie tot remote sensing in de landbouwwetenschappen. In veel toepassingen en methoden wordt de informatie binnen het spectrale beeld echter niet volledig benut, wat leidt tot suboptimale resultaten. Een van de meer prominente zorgen is ruimtelijke correlatie, aangezien deze ofwel volledig wordt genegeerd of op een suboptimale manier wordt beschouwd. In feite wordt de ruimtelijke dimensie gewoonlijk pixels-gewijs uitgevouwen om een tweedimensionale datamatrix te verschaffen voordat chemometrische methoden worden toegepast, zoals decompositie door Principal Component Analysis (PCA) of Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS). In het project zijn nieuwe tools ontwikkeld om tegelijkertijd rekening te houden met de ruimtelijke en spectrale dimensies, en om de resolutie van mengsels te verbeteren in uitdagende situaties, waar de huidige tools mogelijk niet in staat zijn om adequate oplossingen te vinden.
De eerste ontwikkelde tool heet "Image Decomposition, Encoding and Localization" (IDEL), die een ruimtelijke benadering hanteert voor spectrale beeldanalyse en de link behoudt met de spectrale kenmerken die verantwoordelijk zijn voor de waargenomen ruimtelijke patronen. IDEL combineert een exploitatiestap op basis van wavelet-filters met een compressiestap op basis van beeldcodering en multivariate data-analyse, en een uiteindelijke reconstructie in het oorspronkelijke domein van de meest essentiële ruimtelijk-spectrale informatie. IDEL is toegepast op verschillende benchmarks die een gevarieerde reeks situaties vertegenwoordigen, gesimuleerd en echt. In eerste instantie werden simulaties ontwikkeld om de aanpak op een gecontroleerde manier toe te passen, en daarna op uitdagende, echte case-studies, zoals herkenning van biologische sporen verspreid over verschillende oppervlakken

in de context van forensische analyse, of distributie van voedselcomponenten bij voedselkwaliteitscontrole. De resultaten toonden het vermogen van de methodologie aan, waardoor een nieuw perspectief werd verkregen op de gegevens die voorheen moeilijk toegankelijk waren. De tweede tool pakt het zeer kritieke probleem aan van ondervertegenwoordigde componenten in mengsels die het standaard MCR-ALS-algoritme niet kan herstellen, vooral in de aanwezigheid van ruis. De voorgestelde methode is gebaseerd op een weging-schema dat is afgestemd op de relevantie van de sample (van een bepaalde mengsels) met betrekking tot de mate van zuiverheid die elke sample vertegenwoordigt. Dit nieuwe hulpmiddel, genaamd weighted MCR-ALS, is getest in gesimuleerde gegevens en vervolgens toegepast op een pseudo-reëel beeld van een farmaceutisch tablet, waar aanzienlijke winst wordt behaald ten opzichte van de huidige stand van de techniek. Beide tools zijn gecodeerd in MATLAB. De IDEL-functie is ontwikkeld met een set van standaard parameters die van toepassing zijn in verschillende situaties en klaar zijn om te worden gebruikt door onervaren gebruikers in het veld. Tegelijkertijd kan het afstemmen van parameters worden afgehandeld door de ervaren gebruikers. Beide tools zijn ingesteld om in een breed scala van situaties te werken.

Sleutelwoorden: spectrale beelvorming, wavelet transform, beeldenanalyse, decompositie, spectrale ontleding, multivariate curve resolution

Contents

Chapter 1

Introduction

This chapter serves as an introduction to the field of spectral image analysis, providing a rudimentary understanding of spectral imaging and data analysis within the context of the thesis.

1.1 Spectral imaging

Spectral analysis by means of imaging is a dominant field within science, and spans a vast amount of disciplines, ranging from cell analysis [5] in biology to entire galaxies in astronomy [6]. Spectral imaging can be viewed differently, depending on the perspective taken. It can be the method of taking spectra of an area at discrete spatial coordinates or the method of taking images at a set of discrete spectral channels. An illustration is presented in figure 1.1A, showing the two perspectives. In either case a data cube is obtained, which is three-dimensional that has two separate spatial and one single spectral dimensions. There are many ways of capturing a spectral image, to stay within the scope of the thesis, we will go over a general methodology. Depending on the apparatus, different approaches to retrieving the signal can be employed [7]. Firstly, the simplest, measuring a spectrum at a pixel position (point scan), and moving towards the next pixel, this generates data illustrated in figure 1.1A (top). Secondly, single images can be taken at discrete spectral channels (wavelength scan), this is visualised in figure 1.1A (bottom). There are more ways of analyses e.g., line scanning or snapshot, where a line of pixels or an entire image, respectively, is analysed simultaneously at a set of spectral channels. For for the sake of understanding spectral imaging,

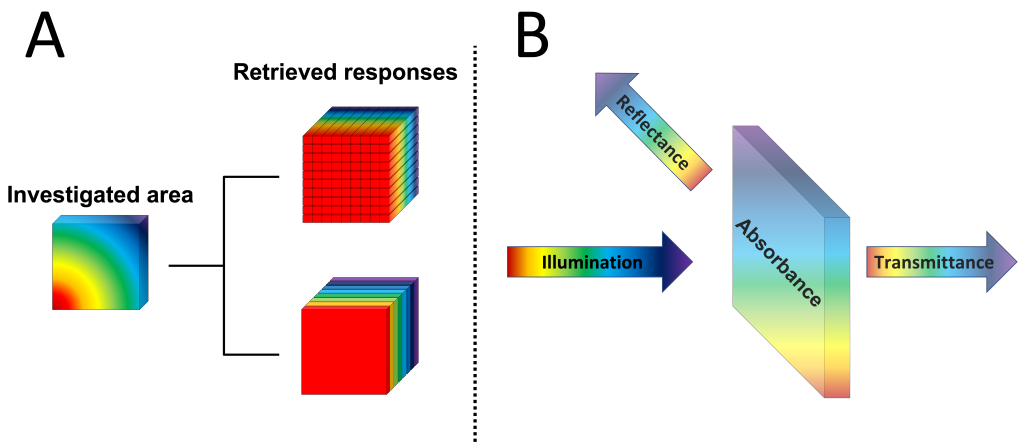


Figure 1.1: Illustration of A) a spectral (top) and imaging (bottom) perspective of a spectral image; B) different measurement units of a spectral signal

the first two perspectives are adequate.

Another parameter of importance is the specific unit of measurement. The three main units are illustrated in figure 1.1B. Starting with absorbance, which is a unit of the amount of light absorbed by the surface and is directly related to the chemical composition and concentration of the compounds within it. Secondly, reflectance, which is the amount of light that is reflected off the surface analysed, this is in relation to the chemical and morphological composition of the surface. The final point is transmittance, which is the amount that is not absorbed, nor reflected, and this is in direct relation to the chemical composition of the surface, as well as the thickness of the sample.

1.2 Data analysis

Staying within the scope of the thesis, spectral image analysis will be confined to decomposing spectral images into their individual sources of variance, extracting the relevant information, and visualising it in an understandable form. Relevant information can be e.g. in the image as distributions, patterns, objects, in the spectrum, as signatures of specific compounds, or in

the link between the two domains. In the following sections the different perspectives in analysing spectral images are briefly illustrated.

1.2.1 Spectral analysis

If a spectral perspective is taken, usually the first step is to unfold the two spatial dimensions to obtain a data matrix that has size pixels × pixels by spectral channels. The pixels are viewed as individual samples and any correlation between pixels is ignored. A brief introduction is provided on one of the more prominent ways of analysis, i.e. the utilisation of the Lambert-Beer law [8–10], which states that, in solution, there is a linear relationship between the concentration of a chemical compound and its absorbance. Equation 1.1 formalises this relationship:

$$A(\lambda) = \epsilon(\lambda)\, c\, l \tag{1.1}$$

where, A is the absorbance, ϵ the absorptivity of a chemical compound at a spectral channel (λ), l the path length and c the concentration. The path length can be assumed to be constant, and is set as 1. This relationship can be expanded to a mixture of multiple chemical compounds n, across a set of samples x, and can be written as:

$$d(x, \lambda) = \sum_{i=1}^{n} (c(x, i) \times \epsilon(\lambda, i)) \tag{1.2}$$

where d is the measurement, at a spectral channel from a specific sample. This can be further simplified with matrix notation:

$$D = CS^T + E \tag{1.3}$$

where $D(I \times J)$ is the absorbance for I samples and J spectral channels, $C(I \times n)$ are the concentrations of the samples for the n different chemical compounds, $S(J \times n)$ are the absorptivity's for the spectral channels of the different chemical compounds and $E(I \times J)$ is the instrumental error. C and S are usually denominated as the concentration and spectral profiles, respectively. With this representation, a bilinear relationship is apparent and the standard analysis methodologies exploit exactly this relationship to recover the underlying sources within the data.

1.2.2 Image analysis

Perpendicular to spectral analysis is viewing the data cube as images at specific spectral channels. Taking this perspective requires tools that can exploit the information within and between pixels. The field of image analysis is incredibly vast, and to stay within the scope of the thesis, a subsection of image processing will be discussed that is within the bounds of image decomposition and encoding. Although fields such as neural networks and machine learning play a big role in image processing, these are not within the scope of the thesis and therefore not discussed.

Within an image is information regarding the morphological aspects of the investigated surface. Morphological aspects consider the presence of objects and the surface on which the objects are. Object/surface shape, roughness, thickness and homogeneity are some aspects to consider in imaging. Different mathematical operations can be applied to determine the various aspects, as well as highlight or enhance specific aspects that are of importance. E.g., enhancing the contrast of objects with respect to the background can increase the visibility within an image. The objective is to recover the objects or textures that are of importance and related directly to the morphology of the surface analysed.

1.2.3 Spectral image analysis

The goal within spectral image analysis is the recovery of chemically and/or physically different objects. The information available is no longer, just spectra or images, but a combination of the two domains. The problems that present an opportunity for spectral imaging, are in situations where it cannot be solved in either domain and utilising the information that is present between the two domains becomes necessary, this is coined the spatial-spectral interplay.

Taking reflectance as an example, the signal is dependent on the morphology and chemical composition of the surface analysed. In such a scenario, the chemical information can be obscured due to the morphological aspects, and vice versa. In this situation the information recovered in either domain, would be affected by the other. This has its opposite effect as well, where if two objects have a similar chemical composition, if physically distinguishable

within the image the two objects can be recovered, independently.

The thesis is presented in the framework of the development of novel tools that are able to exploit the relationships within and between the two domains to extract relevant information within spectral imaging data.

Hyper-

A small note to consider within the context of the thesis, is that the standard nomenclature within the field of spectral imaging for the analysis of images at a number of spectral channels > 20 is usually denominated by the adjective hyper-, as in hyperspectral imaging. However a reference to [11] is presented, where Polder and Gowen advice against such wording. The thesis will not use the hyper- prefix and present it only as spectral imaging.

Chapter 2

State of the art

In this chapter the state of the art in spectral image analysis is discussed, presenting a brief overview of current spectral imaging analyses, and then moving on the the three perspectives of data analysis that are within the scope of the thesis i.e., spectral, spatial and spatial-spectral. A brief discussion on data pretreatment is presented, as well as a final paragraph on the objective of the thesis.

2.1 Spectral imaging

Spectral imaging is an incredibly broad and deep area of science that is applied across numerous disciplines [12] e.g., agricultural [13], forensic [14, 15], medical [5, 16, 17], pharmaceutical [18, 19], conservation and restoration [20], remote sensing [21], quality control [22–24], and astronomy [6] fields just to name a few. The analyses range from cell analysis in biology [25, 26], to atmosphere decomposition in astronomy [27]. The broad nature of the analyses translates to vastly different objectives or data, as e.g. in food quality control, the objective may be real-time analysis of foods [23], while within the conservation sciences, the objective can be to uncover the underlying composition of paintings [28], which might lead to understanding the painter and the timeframe of the work.

2.2 Spectral image analysis

An overview of the methods applied taking the three different perspectives is presented.

2.2.1 Spectral analysis

Firstly, a spectral perspective can be taken, where the data are viewed as spectra, where the correlation between samples is not considered [29]. The methods for data analysis consider the pixels to be a set of samples, and usually by unfolding, a data matrix is generated that has the dimensions pixels by spectral channels that is further analysed. Methods for classification [30–32], unmixing [33] or regression [15] are used to further analyse these data. Within in the context of unmixing, different strategies can be employed, with one of the most well-known techniques being multivariate curve resolution (MCR) [34], where the assumption is that the data are bilinear and non-negative. Which is a fair assumption to make with regards to absorbance spectroscopy, however in some cases e.g., near-infrared imaging, where absorbance can become difficult to measure due to scattering effects dominating the signal, the bilinear relationship might not necessarily hold [35, 36]. Some methods of pre-processing can be applied to remove such effects [37] e.g. multiplicative scatter correction (MSC) [38]. MSC is used to equalise the scattering, however this is not taking into consideration that different pixels might have different scattering contributions [39] and trying to remove any pixel-pixel relationships that are present within the scattering contributions will not always succeed, requiring more complex solutions [35, 40–42]. With regards to reflectance some spatial correlation is present that has to be accounted for [26].

2.2.2 Image analysis

The other side of the coin, with respect to spectral image analysis is the image processing field. Within this field the focus is on the pixel-pixel relationships, with different objectives in mind, e.g. image enhancement, by frequency filtering or contrast enhancement [43, 44]; segmentation by clustering [45] or edge detection [46]; object detection [4]; texture analysis [47–49], by image encoding [50–53]; image decomposition [54]. Different processing techniques allow for various aspects to be highlighted, e.g. with Fourier

transform (FT) [55], where image quality can be significantly improved for a better visualisation of the underlying components. FT is used predominantly in signal processing [56], for its incredible power to decompose a signal into its sine and cosine functions, and isolate the frequency content. It is however, also used in image processing, for filtering noise and increasing image quality. Object detection is used where texture analysis can identify physical aspects of components [4]. Image encoding transforms images into vectors of values that describe spatial features [50]. This allows for data reduction, making multivariate data analysis applicable without the destruction of the spatial information, as well as making classification much easier [57]. Lastly, image decomposition, where single images are decomposed into separate constituents that make up the original. Work-horse algorithms can be applied, such as principal component analysis (PCA) [58], however staying in the context of image analysis, the more prominent texture decomposition techniques are wavelet transform (WT) [59] and Gabor filters (GF) [60], similar to FT. GF uses a complex number that is the multiplication of a sinusoidal and gaussian function, and is able to detect the frequency and orientation of textures in images. It has been compared to human visual prowess. WT decomposes the image into a set of different frequency contents and orientations. Although the methods are used in similar areas and seem similar, WT uses a more methodical approach to extracting spatial information and shows a higher discriminatory power [60]. Similarly to spectral analysis, a disregard for the other side of the coin, in this case spectral correlation, will prevent chemical interpretation.

2.3 Spatial-spectral analysis

Novel developments towards spatial-spectral analyses have become prominent within the different fields [61]. One such approach is multivariate image analysis (MIA), where the unfolded imaging data is augmented with pixel neighbour information, to incorporate local-spatial information before it is analysed with multivariate analysis tools, such as PCA or partial least squares (PLS) regression [52, 62, 63]. MIA has been originally proposed for RGB images [62, 63] then extended to multi-channel images [64] and only recently to spectral images [62]. However, the number of neighbouring pixels increases rapidly with the distance (or window size in pixels) from the centre pixel at which to consider the neighbourhood, and this applies to all spectral

channels, making the data unmanageable in some cases. In this context, a parsimonious solution can be used to employ multivariate curve resolution-alternating least square (MCR-ALS) using image processing constraints to take into account the spatial structure [65–67]. Nonetheless, this does require the data to strictly follow a bilinear model. If different compounds exhibit distinct spatial features (i.e., textural patterns), and/or two (or more) of these different compounds show a significant overlap along both the spectral and the spatial domain, this approach may not be applicable. Some work has also been done on image segmentation, with the integration of the spectral domain [68], as well as utilising the spectral and spatial domain, interactively switching between the two modes [69, 70]. The analysis of textural features in spectral imaging has also been explored, by using the wavelet transform (WT) [64, 71–75]. These analyses i) use a MIA-like approach, where the local spatial information is extracted by WT, and 2D-WT sub-images are then analysed by multivariate analysis [64, 71–73], either on each single sub-image [71, 72] or on the entire data sets [64, 73, 74]; ii) exploit 3D-WT on the imaging data cube [75] or iii) fuse the 2D-WT sub-images obtained at each spectral channel [76]. These approaches also aim at linking the spectral and spatial domains, however in some cases the spectral interpretation is not so straightforward [75], while in others the dimensionality of the data matrix when passing from multispectral [64] to larger spectral images become quite large [74].

2.4 Data pre-treatment

The preprocessing of data has been briefly touched on, but as it is important in any data analysis methodology and inherently linked with the spectral imaging apparatus and the data analysis methodology [77, 78], a brief explanation is provided on the positioning of this aspect. When an imaging perspective is taken, the focus is on the enhancement of the discrete images, by means of e.g., FT, contrast enhancement, where a better distinction of components is visible within the images. Considering a spectral perspective, the importance falls e.g., on generating bilinear data, removal of baseline or spikes [32]. However, when a spatial-spectral perspective is taken, the consideration has to be both on the image as well as on the spectral correlation [79]. Singling out either domain might affect the other.

2.5 Objective

A general trend across the literature leans towards a fusion of the two fields, i.e. spectroscopy and image processing. Recent work [64, 67, 76] shows that incorporating some information from either domain is becoming easier and new developments in the field are not uncommon. However, a note to take is that the fusion of the two fields does not always present stand-alone novel methods, but a fusion of existing methods to generate novel solutions to existing issues. The objective of this thesis is exactly the development of novel solutions that exploit well-established methodologies within the respective communities to tackle prominent issues within spectral image analysis. Simultaneously tackling case-studies that highlight the issues as well as benchmarks to present the novel tools in controlled environments.

Chapter 3

Image Decomposition, Encoding and Localisation

Image processing methods are incredibly powerful [44, 47, 52], and extensions towards spectral images have seen novel developments [64, 75]. The objective of an imaging perspective incorporating spectral information is to utilise current image processing methodologies to extract, highlight or expand the spatial information while exploiting the spectral information.

Image Decomposition, Encoding and Localisation (IDEL) is a general framework developed where images are decomposed to emphasise and expand the spatial information and then encode it to reduce the dimensionality and highlight specific spatial features. A final step is performed to localise the relevant decomposed images within the encoded information. Within this work IDEL relies on wavelet transform [80] to expand the spatial dimension. By applying two-dimensional stationary wavelet transform (2D-SWT) [64], the single images at discrete spectral channels are decomposed into sub-images that contain spatial features of different frequency contents and orientations. This isolates distinct spatial features within an image. An encoding algorithm is applied, that is able to retrieve averaged local spatial features of single images by encoding grey-level co-occurence matrices (GLCM) into a vector of descriptors [50, 52, 53]. This enables the method to reduce every image into a single vector. Lastly to recover specific spectral signatures for each spatial feature, multivariate data analysis is applied in various forms. In this chapter the algorithm will be broken down into its subsections and the data transformations are explained along with the algorithm. Bench-

mark data are used to corroborate the results from IDEL. Different data sets are analysed that are generated across various fields, i.e. an oil in water emulsion, biological liquids on fabrics, and foods. Each data set has different properties that will highlight the benefits and drawbacks of IDEL.

3.1 IDEL – Algorithm

The approach consists of five consecutive steps: (1) 2D-SWT decomposition of the spectral image resulting in wavelet sub-images; (2) from these sub-images, computation of the GLCM; (3) calculation of morphological descriptors from each GLCM; (4) rearrangement of the obtained descriptors into a descriptors matrix (DM); and (5) multivariate modelling of this matrix for the exploration of the variability of the morphological descriptors across spectral channels. The process is illustrated in figure 3.1.

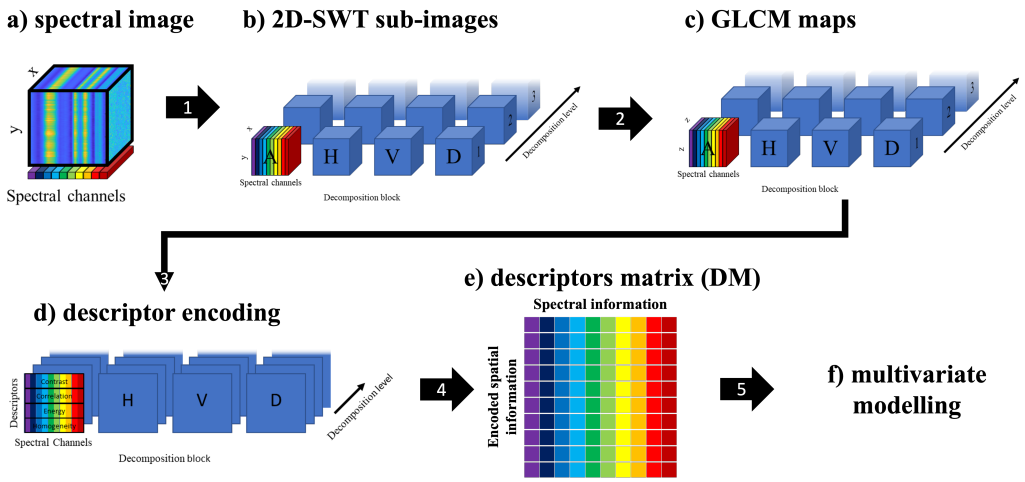


Figure 3.1: Illustration of the IDEL framework. The spectral dimension is coloured after every step for the sake of a better visualisation. a) generic spectral image; b) the sub-images obtained after 2D-SWT decomposition of a; c) GLCM maps obtained from each sub-image of the wavelet decomposition depicted in b; d) the descriptors calculated from the GLCM maps; e) rearrangement of d to obtain a single descriptors matrix (DM); f) generating a multivariate model of DM

These main steps, described in the following three subsections, are decomposition, encoding and localisation. Decomposition (figure 3.1 step 1) exploits the various features that make up the spatial structures within the images of spectral imaging data. Encoding (figure 3.1 step 2-4) serves to reduce the data dimensions without losing the information extracted in the preceding step. Localisation (figure 3.1 step 5) recovers the relevant features and sources of variance from within the encoded data matrix.

3.1.1 Decomposition - 2D-SWT

The first step consists of the decomposition of the individual images corresponding to each spectral channel into a set of sub-images. 2D-SWT is a very powerful filtering method, highlighting the different frequency contents of an image, while maintaining their localisation with respect to the original domain. High and lowpass filters are applied to decompose the signal into two disjoint subspaces holding the sets of details and approximation blocks (high and low frequencies, corresponding to sharp and smooth features, respectively). The decomposition is iterated on the approximation block, obtaining at each level a coarser representation of the image than in the previous approximation block, and the filtered higher frequencies in the details. For image analysis, the same mono-dimensional wavelet filters are recursively applied along the two image directions, i.e. the rows and columns. For each decomposition level, four blocks are obtained: 1) approximation (A): a low-pass filter is applied both row- and column-wise; 2) horizontal details (H): a low-pass filter is applied row-wise, then a high-pass filter, column-wise; 3) vertical details (V): a high-pass filter is applied row-wise, then a low-pass filter, column-wise; 4) diagonal details (D): a high-pass filter is applied both row- and column-wise. The specific direction along which the low- and high-pass filters are alternated, allows for specific textural patterns to be captured e.g., the H decomposition block highlights any pattern which would manifest horizontally, such as stripes (hence the name horizontal details). For the V and D blocks, vertical and diagonal textural patterns are highlighted, respectively, while the A block holds the original image where the details are subtracted. 2D-SWT retains the size of the original image, so that the decomposition blocks (referred to as sub-images A, H, V, and D), for each decomposition level, are equal in size to the raw image. Wavelet filters are grouped in specific families, which differ in shape and symmetry, while amplitude is modulated in each family by the number of vanishing

moments. The choice of an appropriate wavelet filter is data dependent and providing an automatic tool to tackle this task is outside the scope of the chapter. However, there are criteria detailed in literature to guide the choice of suitable wavelet filters [81]. A general recommendation, that can be given is that the simplest Haar wavelet, which comes from the Daubechies - family (Daubechies-1) is usually a good starting point when, as in this case, the aim is exploratory. In fact, Haar can capture general changes present in an image, not focusing on specific spatial features, and disentangle signals that range from sharp contrasting edges to broad structures.

In this work, the Haar wavelet is applied and showed good performance, the maximum decomposition level compatible with the image size is used and periodisation is set as the padding mode by default. All approximation blocks are retained, as they might better retrieve some spatial aspects and the redundancy that is carried with it is handled by the multivariate analysis steps following it. Utilising 2D-SWT allows for the decomposition of a single image into orientation-specific features of different frequency contents, maintaining the spatial localisation of the features. This decomposition method transforms the data cube into a five-dimensional array with dimensions pixels $\times$ pixels $\times$ spectral channels $\times$ decomposition blocks $\times$ decomposition levels.

3.1.2 Encoding

Due to the decomposition methodology a single spectral image is decomposed into a large number of data cubes. For each data cube there are as many images as there are discrete spectral channels, this is an immense amount of data that might not all be relevant. To analyse this data more efficiently, the sub-images are encoded to reduce the data space from two spatial dimensions with pixels to a single vector of descriptors. IDEL uses GLCM and general feature descriptors to describe distinct spatial features within the sub-images, with the assumption that most if not all of the relevant spatial correlation is described within those descriptors.

GLCM

GLCM maps the local textures of a given image by counting how often pairs of pixels with certain normalised integer intensity values occur at a partic-

ular distance. GLCM starts with an image, which is quantised into a set of levels. Within this grey-level image, a counting is done, of the different pairs of pixels present. This counting is mapped into a square grid, with dimensions equal to the number of grey-levels, previously set. GLCM maps every possible pair of pixels.

There are three main parameters that are set, firstly the quantisation, how many grey-levels the intensity is quantised in, secondly, the angle, in which direction pixel-pairs are generated and thirdly, the offset, the distance between said pixel-pairs. The pixels-pairs are determined by the angle and offset, which are dependent on the wavelet decomposition block and level, respectively.

Quantisation The quantisation is dependent on the size of the image, as having more pixels equates to more pixel-pairs, which in turn would give more points to be distributed across the GLCM. This means that for every image with a different number of pixels, there is a different number of grey-levels that are optimal, as a balance is necessary between, the intensity of the map and resolution of the distribution. An example is shown in figure 3.2, where the map and histogram of a GLCM with a range of grey-levels are shown, with the original image have a size of 128×128 pixels. As a rule of thumb, the quantisation is set as $\sqrt{N}$, with N being the number of pixels. This is heuristically determined within the bounds of the datasets analysed. This results in a map being generated with size $\sqrt{N} \times \sqrt{N}$ elements, containing all possible quantised pixel-pairs. In figure 3.2 the optimal map is set at 128 bins, as it has an adequate resolution, without the loss of intensity.

Angle The details related to the orientation of specific patterns that are within the image guide the choice of the angle at which the pixel-pairs should be generated. In figure 3.3 the different orientations are visualised. Specific angles are selected to maintain coherency between the directions of the WT decomposition details H, V, and D and the location of the investigated neighbour within the GLCM. Hence, the angle is set depending on the type of sub-image: H considers the top and bottom neighbours, V, the left and right neighbours, and D, the top-left, and bottom-right neighbours. These neighbours highlight the local differences within the sub-images. While for A, as there is no specific direction, the sum of the three previous directions

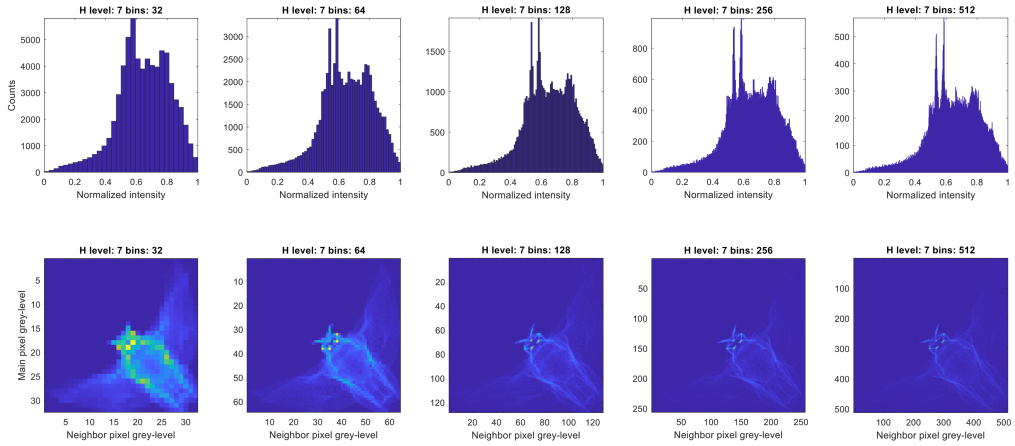


Figure 3.2: Histograms (top) and maps (bottom) of the GLCM from an example image. Varying the number of bins from 32 bins (left) to 512 bins (right)

is considered.

Offset Since at higher decomposition levels the image patterns become smoother, relevant pixels are found at progressively higher distances. An example is visualised in figure 3.4, where the horizontal details are shown at different decomposition levels. To account for this, the offset is set as $2^{level-1}$. This permits GLCM to account for the smoother patterns that are highlighted with increased WT decomposition levels.

Descriptors

GLCM transforms the pixels $\times$ pixels into grey-levels $\times$ grey-levels, maintaining the five-dimensional array size. This can be compressed by calculating descriptors from the GLCM of the A, H, V, and D sub-images. GLCM maps the local structure of an image, while the descriptors calculate specific aspects averaged over these maps. A global description is made from GLCM maps that describe local features. A set of eight descriptors (Energy, Contrast, Correlation, Variance, Inverse difference moment, Sum entropy, Information

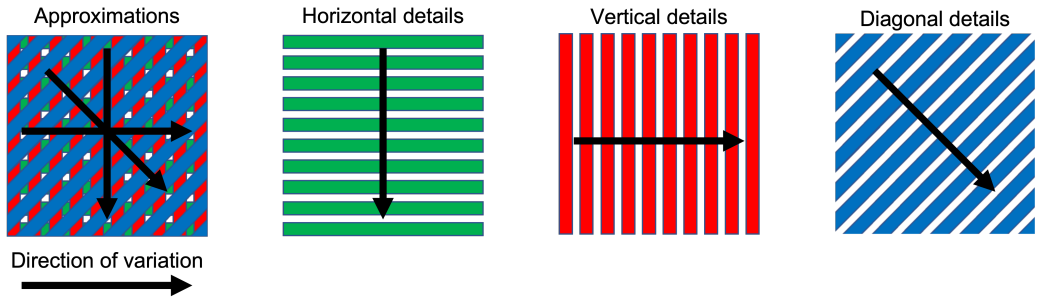


Figure 3.3: Illustration of the wavelet transform orientations of the various details and their respective direction of variation

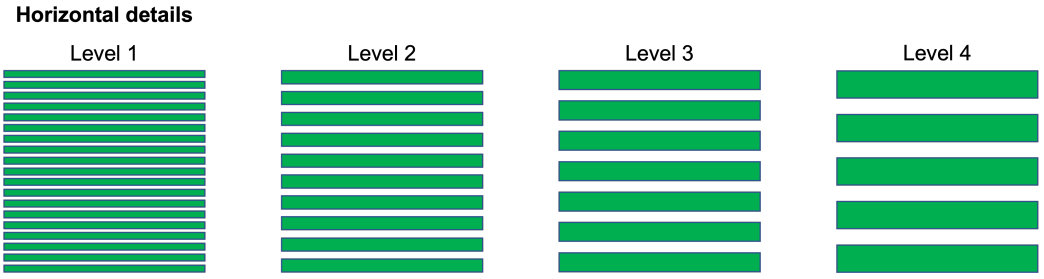


Figure 3.4: Illustration from horizontal details of the wavelet decomposition levels, from low (left) to high (right)

Measure of Correlation 1, and Maximal correlation coefficient) is computed for each GLCM. A brief description of each of these features and their respective formulas are in table 3.1.

These descriptors are selected as they are the least correlated with one another while describing all the relevant spatial features. However, depending on the case study, distinct descriptors can be selected based on prior knowledge or on the necessity of specific image features to be highlighted. Encoding the GLCM into descriptors reduces the data dimensionality from five to four, these being number of descriptors × spectral channels × decomposition blocks × decomposition levels.

Table 3.1: Definition of descriptors [50] and their respective formulas

Descriptor	**Formula**
Energy computes the squared sum of all elements, giving a measure of uniformity of the original image, it is at a maximum when the image is constant.	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(p(i,j)^2)$
Contrast computes a measure of the intensity contrast between a pixel and its neighbour over the whole image, its value is zero for a constant image and is at its highest when the neighbouring pixel intensity is very different.	$\sum_{n=0}^{N_g-1} n^2\{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(p(i,j))\}_{n=\lvert i-j\rvert}$
Correlation computes a measure of how correlated a pixel is to its neighbour over the whole image. Correlation is a measure of grey tone linear dependency in the image, it attains values of 1 or -1 for a perfectly positively or negatively correlated image, respectively. It is undefined for a constant image.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(ij)p(i,j)-\mu_x\mu_y}{\sigma_x\sigma_y}$ Where μ_x, μ_y and σ_x, σ_y are the means and standard deviations of p across rows and columns of GLCM matrix, respectively.

Variance simply calculates the variance of the GLCM, giving information on the distribution of the pairs of pixels.	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-\mu)^2 p(i,j)$
The inverse difference moment returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal, this is also known as homogeneity. This is 1 for a diagonal GLCM.	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{1}{1+(i-j)^2}(p(i,j))$
Sum Entropy uses the same formula as Entropy* but considers, instead of single elements of the GLCM, the sum of elements in its diagonals. Entropy*: $HXY = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(p(i,j)log(p(i,j)))$ It computes a measure of disorder related to the grey-level distribution of the image and it is large when the image is not texturally uniform and many GLCM elements have very small values.	$-\sum_{k=2}^{2N_g} p_{x+y}(k)\log(p_{x+y}(k))$ where $p_{x+y}(k) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\ k = i+j$
Information measure of correlation-1 computes the ratio of the difference between overall entropy and joint entropy (across rows and columns) to the max entropy across rows/columns. It can be interpreted as a measure of texture complexity.	$\frac{HXY - HXY1}{max(HX,HY)}$ where HXY is Entropy* and HX(HY) are rows (columns) Entropy, respectively, i.e.: $HX = -\sum_{j=1}^{N_g}(p_x(i)log(p_x(i)))$ where: $p_x = \sum_{j=1}^{N_g}(p(i,j))$ and $p_y = \sum_{i=1}^{N_g}(p(i,j))$, and $HXY1 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(p(i,j)log(p_x(i)p_y(j)))$

Maximal correlation coefficient computes the maximum correlation between two variables, which gives a numerical measure of dependence between said variables.	It is defined as the square root of the second largest eigenvalue of: $Q(i,j) = \sum_m \frac{p(i,m)p(j,m)}{p_x(i)p_y(m)}$

Transformation

To make the data more manageable, the 4-dimensional array is unfolded into a two-dimensional matrix, where the dimensions of the decomposition and encoding are unfolded into a single dimension and the spectral dimension is retained. The resulting data matrix allows for bilinear modelling to be applied, and to correlate distinct spectral channels to descriptors that highlight specific patterns within the images.

Pretreatment

Although the different descriptors utilise the same information within the GLCM, they will not have the same units e.g., the Energy descriptor, which is the sum of squares, and the information measure of correlation-1 descriptor, which is a ratio will have vastly different values. To give equal contribution to each descriptor, instead of standard column autoscaling of the DM matrix, autoscaling has been applied per descriptor. Mean-centring and scaling has been applied, for each descriptor, all decomposition levels, decomposition blocks and spectral channels, pertaining to that descriptor. This pretreatment allows for the multivariate analysis of the different descriptors, without any particular descriptor dominating.

3.1.3 Localisation

The data matrix obtained after the pretreatment step is called the Descriptor's Matrix (DM), and is used to extract the most relevant information. Although different multivariate strategies can be applied, to remain within the scope of the thesis, the main exploratory tool will be principal component analysis (PCA).

DM is subjected to the exploration of the information it carries and, more specifically, for establishing a link between the spatial and spectral information captured by the variation of the morphological descriptors within the investigated spectral range. A possible pathway to establishing this link in a more systematic way is also explored, that is, the application of k-means clustering to the resulting PCA loadings to get an idea about the spectral channels that exhibit similar variance in the descriptors. DM contains descriptors on sub-images at every spectral channel, encoding spatial information in the rows and retaining spectral information in the columns. Applying PCA to DM, thus scores and loadings relate to the spatial and spectral information, respectively. The number of PCs to consider is of course data dependent and a scree-plot is used as a guideline. Each point in the scores plot corresponds to one descriptor of a sub-image (A, H, V or D) at a specific decomposition level. As such, looking at the scores, the most distinct spatial features can be identified. The loadings plot, in conjunction with the scores plot, enables us to establish a link with the spectral channels. In fact, the loadings plot shows at which spectral wavelengths the largest variation of the descriptors within the different sub-images and decomposition levels is observed. To observe the correlation between the scores and loadings, a bi-plot is generated, where the scores and loadings are in the same space and can be superimposed on the same plot. The relationship is determined by the cosine of the angle between two vectors, in this case the correlation between a score and loading is investigated. This would mean that an angle of $0°$, would have a correlation of 1, while an angle going towards $180°$ would have a correlation of -1. An angle going towards $90°$ is set as having 0 correlation. This is illustrated with a simple two dimensional example in figure 3.5, where A is a score, and B, C and D are loadings. With respect to A, B has a high positive correlation, C a high negative correlation and D, no correlation. This can be scaled up to any number of dimension, which in this case are determined by the number of PCs used.

3.2 Performance parameters

Performance of the methods used will be determined by comparing prior knowledge (from literature), a PCA model on the data, and the ground truth (if available), to an exploratory PCA analysis on DM. A final analysis is performed by using k-means clustering to obtain more systematic results.

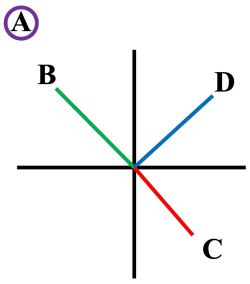


Figure 3.5: Illustration of correlation and angle, with A being a score and B, C and D being loadings

3.3 Benchmarks

3.3.1 Oil in water emulsion

The first case study relates to a Raman imaging dataset of an oil-in-water emulsion [19], which illustrates a situation where the spatial and spectral information are both somehow selective for some compounds, meaning, no severe overlap of the spatial and spectral features is observed. More specifically, the different individual compounds (featured in the spectral domain) are associated to clearly distinguishable shapes/spatial structures. This dataset is relatively simple, serves as a proof of concept for the proposed methodology and has been extensively analysed by the scientific community [62, 65].
The Raman imaging system by which these data are collected has a spatial resolution of around 1 μm, and the image is 60 $\times$ 60 pixels. The spectral resolution is 3.6 cm^{-1}, and the investigated spectral range goes from 953.6 to 1861 cm^{-1} (253 wavelengths). There are at least four compounds present, including the background. These four will be the focus of the analysis, which can be already fully visualised by selecting four pixels for the four spectra, or four spectral channels for the four images that show the concentration profiles. Figure 3.6 highlights these four compounds. Although the spatial structures are distinct, there is some spectral overlap between two compounds (ring and big drop). From the PCA analysis, seen in figure 3.7, we observe three of the four compounds within the scores images of the first three PCs. The background, which does not have any absorbance, as it is water, is not visible in the PCA analysis, as there is no variance to capture. In every PC image, some overlap is visible, PC1 contains all three compounds, PC2 con-

tains the ring and big drop, and PC3 seems to contain a minor contribution form the big drop and mostly the small drop is visible. PC4 contains a combination of the three compounds. The accompanying loadings show overlap with the extracted spectra in figure 3.6, however due to the lack of isolated compounds in the PCs a direct comparison is not made.

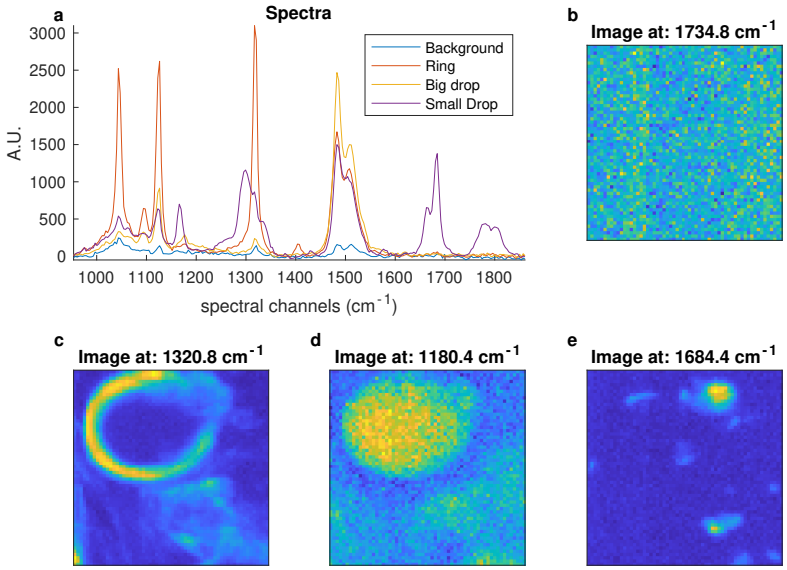


Figure 3.6: Overview of oil in water emulsion data set; a) four spectra at designated locations; b-e) four images at specified spectral channels

3.3.2 Semen stain on cotton

The second case study regards a 222 × 220 pixels NIR image (acquired in the wavelength range 1268.8–2456.2 nm with a spectral resolution of 6.3 nm) of a semen stain on a piece of cotton. Further details on the data acquisition are given in Silva et al. [15]. The data are pre-processed with a weighted least squares baseline correction, more standard pre-processing steps are avoided due to their significant impact on the spatial structures within the images (appendix A). An overview of the data is shown in figure 3.8. Using the same method of presenting the data as the first case study does not prove as fruitful, as there is complete spatial and spectral overlap between the two compounds present. A quite complex structure is observed which is characterised, at least at a first glance, by (i) a distinct horizontal pattern across the

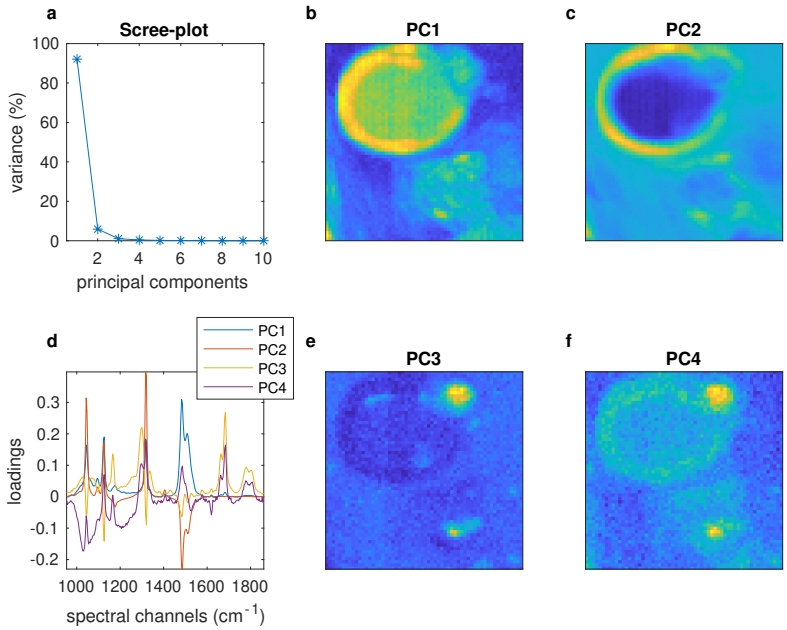


Figure 3.7: Overview of PCA analysis on oil in water emulsion data set; a) scree-plot; b,c,e,f) refolded scores of PCs 1-4, respectively; d) loadings of PCs 1-4

entire image that is due to the rough surface of the cotton fabric (texture); (ii) different discernible structures of the oval-shaped semen stain; and (iii) a spurious fibre filament in the lower middle area of the image. It is worth noting that this case study exhibits a much higher complexity than the previous one: the cotton contribution is present everywhere across the image; thus, there exists no spatial area selective for semen. In addition, the spectral profiles of the different compounds of the captured scene are severely overlapped and semen is not homogeneously distributed over the cotton sample. From the PCA analysis, seen in figure 3.9, no extra relevant information seems to be present, it is worth noting that more than 99% of variance is captured by PC1, this is most likely due to the fact that the difference between the images are minor with respect to the cotton background.

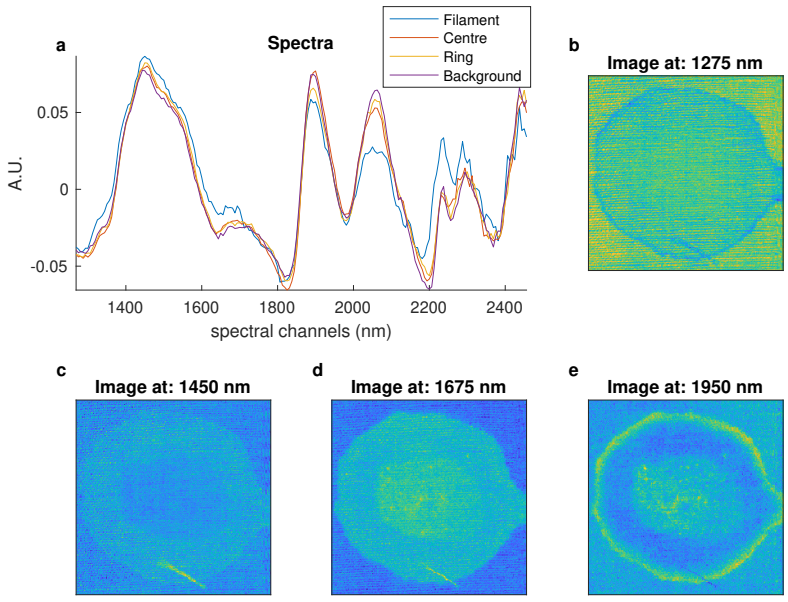


Figure 3.8: Overview of semen data set; a) four spectra at designated locations; b-e) four images at specified spectral channels

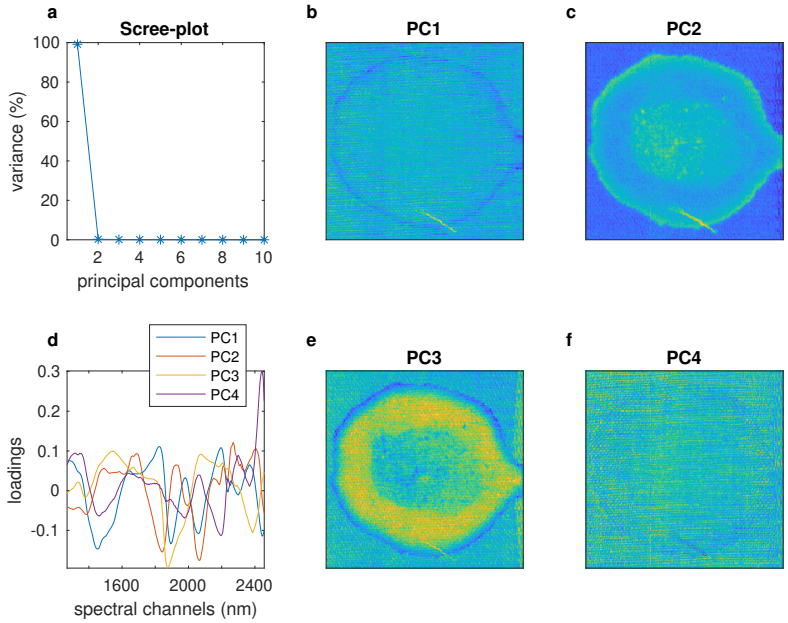


Figure 3.9: Overview of PCA analysis on semen data set; a) scree-plot; b,c,e,f) refolded scores of PCs 1-4, respectively; d) loadings of PCs 1-4

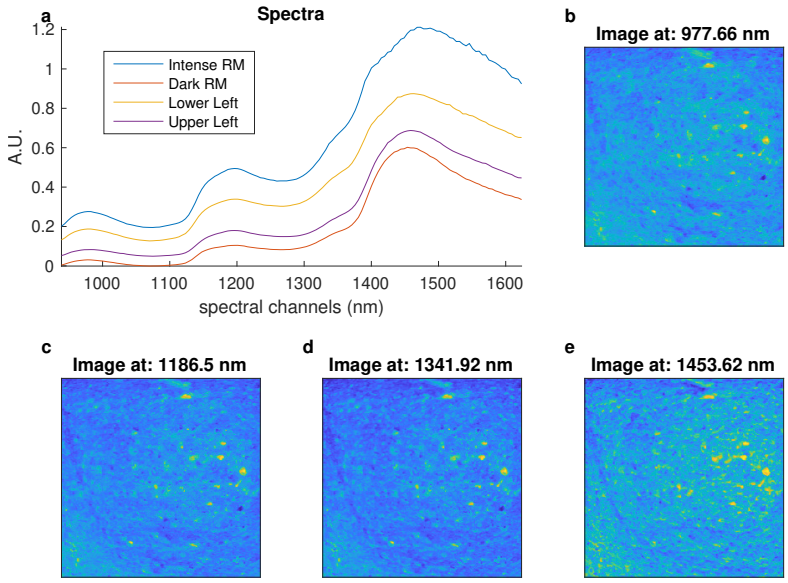


Figure 3.10: Overview of bread data set; a) four spectra at designated locations: 1) intense spot on right middle of the image (blue), 2) dark spot on right middle of the image (orange), 3) lower left corner of the image (yellow), 4) upper left corner of the image (purple); b-e) four images at specified spectral channels

3.3.3 Bread

The third and final case study is an analysis of a slice of bread. This full data set consists of six NIR images, that are taken at six different time points, ranging from 1 to 21 days, however only the first time point is investigated. The NIR image has a size of 144 pixels $\times$ 212 pixels $\times$ 142 spectral channels. The spectral range is 938 to 1630 nm with a resolution of 4.85 nm and each pixel has a size of 320 $\times$ 300μm. A 4$\times$4 binning has been applied to each image, with no further pre-processing. Further details on the experiment are discussed by, Amigo et al. [82–84]. An overview of the data is in figure 3.10, it is clear that there are little to no differences between the images and spectra, other than intensity. The PCA analysis, in figure 3.11 does not provide any further clarification on the data.

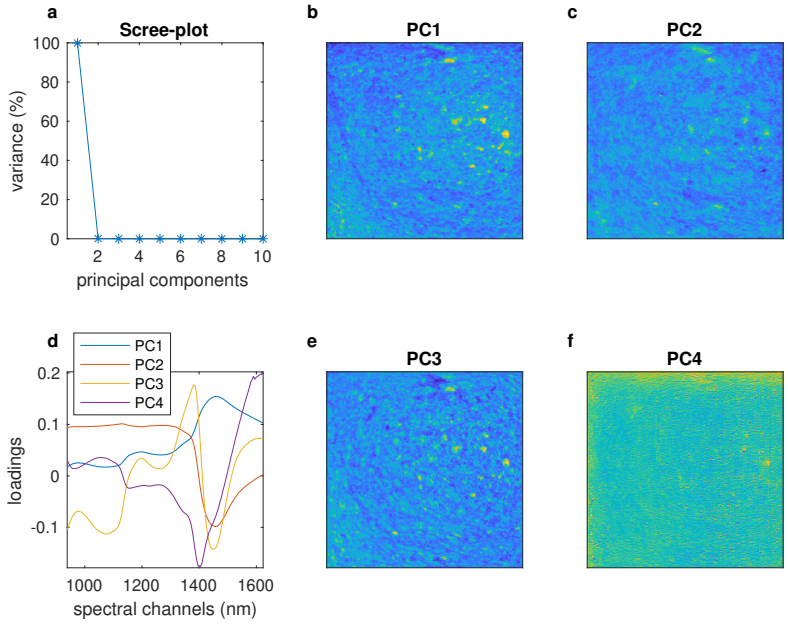


Figure 3.11: Overview of PCA analysis on bread data set; a) scree-plot; b,c,e,f) refolded scores of PCs 1-4, respectively; d) loadings of PCs 1-4

3.4 Results and Discussion

Within this section, the three datasets are investigated with IDEL, with the main goal being data exploration. PCA is applied on DM, after which the most relevant scores, and their correlated loadings are retrieved, by visually inspecting their respective sub-image and wavelength. Elucidations are made by observing the spatial structures and spectral contributions.

3.4.1 Oil in water emulsion

IDEL is applied on the spectral image, generating DM. The outcome resulting from the PCA modelling of DM is shown in figure 3.12, which shows the scree-plot (a), and scores (b and c) and loadings (e and f) plots of the first three principal components. The score plot provides a graphical representation of the variation of the morphological descriptors across the wavelet sub-images whereas the loading plot accounts for their variation across the spectral channels.

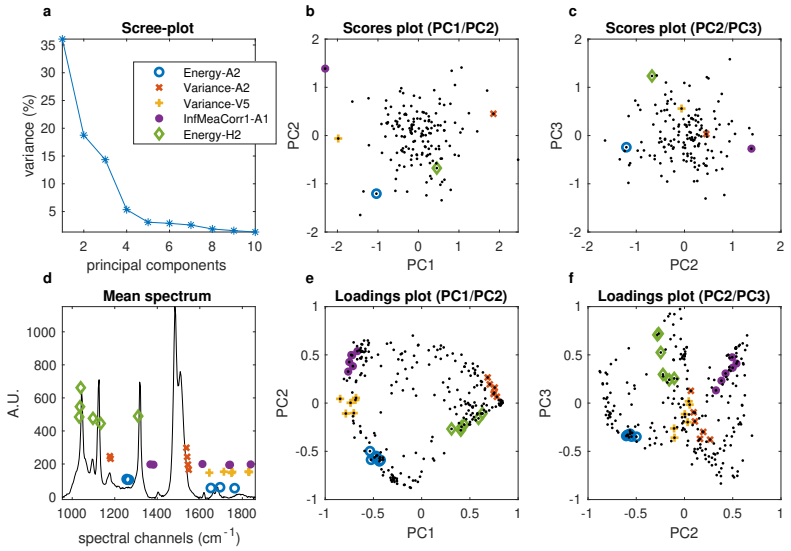


Figure 3.12: Overview of PCA results on DM of the oil in water data set, highlighting the selected scores and their correlated loadings in their respective plots. a) scree plot; b) scores plot of PC1 vs PC2; c) scores plot of PC2 vs PC3; d) mean spectrum of the raw data, with the highlighted loadings superimposed; e) loadings plot of PC1 vs PC2; f) loadings plot of PC2 vs PC3

An exploratory analysis of the most extreme score values is performed. Observing the scree-plot, figure 3.12a, the first three PCs contain a significant portion of the variance, and are the focus of the analysis. A full analysis is done on the twelve most extreme score values of the score plots (see Appendix C.1 and C.2). Each score can be correlated to a set of spectral channels, by determining the angle between the score and any loading value and setting a threshold for correlation ($\pm 9°$). This pair of score and loading corresponds to a particular sub-image within the wavelet decomposition. After an analysis of the first three PCs, observing the twelve highest scores for the PC1/PC2 and PC2/PC3 scores plot, separately, five relevant sub-images are found and further investigated, see figure 3.13. The score of each sub-image is highlighted in the scores plots (figure 3.12b and c), and their five most positively correlated loadings are highlighted in the loadings plots (figure 3.12e and f). Each loading is highlighted with their appropriate symbol on top of the mean

spectrum of the raw data, see figure 3.12d, to correlate the sub-images with their respective spectral contributions. This combination of sub-image and spectral contribution allows for the identification of the various compounds within the data. A comparison with figure 3.6 is made, clearly distinguishing the four compounds from one another. Starting with the sub-image of the first score, Energy-A2 (figure 3.13a) clearly corresponds to the small drop, and the correlated loadings highlighted across the mean spectrum show similar contributions to the observed spectrum in figure 3.6. With the second score, Variance-A2 (figure 3.13b), a similar story is seen with respect the the big drop, where a similar wavelength is highlighted in the mean spectrum of figure 3.12d, in addition to the shoulder of the largest peak at around 1550 cm^{-1}. The third score, Variance-V5 (figure 3.13c), shows a sub-image that is not previously observed, however, observing the image of the background in figure 3.6b, a minor artefact is observed that would seem to relate to an illumination effect, which are low intensity vertical lines in the middle and right of the image. The deepest decomposition levels capture smooth textural features, and this possible illumination effect, has a gradual intensity shift across the image, which is highlighted by the variance descriptor. Observing the spectral contributions, it would seem that the most correlated spectral channels, towards this spatial structure, are low intensity signals. The fourth sub-image, corresponds to the InfMeaCorr1-A1 score (figure 3.13d), which is a measure of texture complexity, this concurs with the structure that is observed, which seems to be background and or noise. Similarly to the third sub-image, the most correlated wavelengths correspond to low intensity signals, which is corroborateb by the reference spectrum and image, seen in figure 3.6a/b, respectively. The final sub-image, corresponding to score Energy-H2, shows the ring structure, which is the fourth and final component that is observed within the data. The most correlated loadings, correspond to the three small peaks on right of the mean spectrum in figure 3.12d. This is concurs with the spectrum observed in 3.6 that corresponds to the ring. A point of interest is the lack of correlation of the highest intensity peak at 1500 cm^{-1}, this is most likely due to the overlap between the three main components (excluding the background), at those wavelengths.

To summarise, IDEL uses descriptors to identify spatial structures within decomposed spectral images, and PCA to correlate and localise the most relevant sub-images. Although just an exploratory analysis is performed, four components, and a possible illumination effect are identified, which corre-

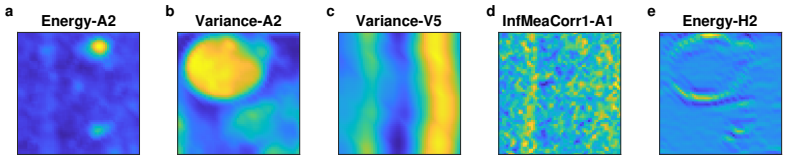


Figure 3.13: Overview of five sub-images from their respective selected scores, at the wavelength with the highest and most correlated loading value

sponds with the analysis of the data set, shown in figure 3.6 as well as being in good agreement with previous findings, [65] which discussed the big droplet as oil and the ring structure being the oil-water interface, while matching the smaller droplet to an oil with a different composition.

To complement the results of this exploratory analysis, k-means clustering is used. For this purpose, DM is compressed by PCA (13 PCs, explaining 90% variance), and clustering is applied on the estimated PCA loadings (four clusters of wavelengths are retrieved). With the goal being, to obtain a better understanding of the spectral contributions and their respective correlated spatial structures, using a more global approach, with respect to the previous exploratory analysis. A more detailed explanation on k-means can be found in Appendix B.

In figure 3.14a, the clustering results are highlighted within the loadings plot, where four distinct groups are visible. In figure 3.14d, the same clusters are highlighted on top of the mean spectrum of the data, while in b, c, e and f the mean images of the clustered wavelengths are plotted. Firstly, the images correspond near identically to the previous results obtained. It is then clear that for the data at hand, contributions showing a distinctive spatial distribution are also associated to rather selective spectral signatures within different wavelength ranges. These spectral signatures are: cluster 1 (b), which is associated to the ring structure, corresponding to three major peaks at 1044, 1126, and 1317 cm^{-1}; Cluster 2 (c), which is associated to the small drop, coinciding with the minor peaks at 1300, 1684, and 1789 cm^{-1}; cluster 3 (e) mainly encompasses the peaks at 1483 and 1508 cm^{-1}. It corresponds to the big drop; cluster 4 (f) corresponds to the baseline regions observed in the mean spectrum.

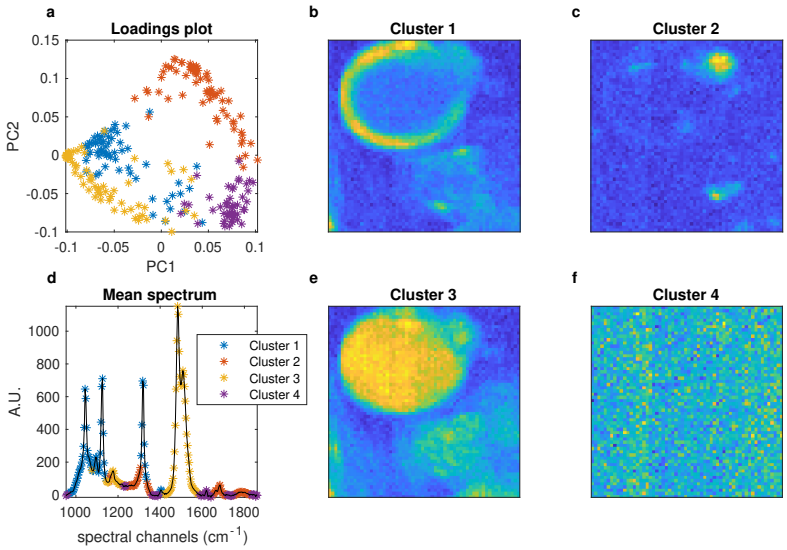


Figure 3.14: Overview of the clustering results on the loadings of the PCA analysis on DM from the oil in water emulsion data set. a) loadings plot of PC1 vs PC2, with the clusters coloured; b,c,e,f) mean images taken from the raw data for each cluster; d) mean spectrum of the raw data, with the clusters coloured

Averaging the spectral image across the four extracted clusters of spectral channels leads to the isolation of the different structures previously unraveled. This corroborates the hypothesis that observing either the spatial or spectral domain, does not negatively impact the retrieval of any of the four components. Due to the straightforward nature of the results, the dataset can been used to assess the relevance of each individual step of the IDEL workflow. Analyses have been performed by removing some of these steps, that is, applying k-means on the PCA of the unfolded spectral image or excluding the wavelet decomposition step and carrying out the GLCM and descriptors calculations directly on the raw images. The obtained results (see appendix B.2) show that the information obtained is less favourable than when applying the full original workflow.

3.4.2 Semen stain on cotton

IDEL is applied to the spectral image, and the results from the PCA analysis are in figure 3.15. From the scree-plot (a) it is clear that more PCs are required to explain a significant part of the data, with respect to the oil in water dataset. This is to be expected as there are more complex structures present. As with the previous analysis the twelve largest scores are investigated, observing the sub-images of the selected scores with their respective correlated wavelength, see Appendix C.3 and C.4. The five most relevant sub-images are retrieved, depicting the different components, see figure 3.16.

Relevant information can be extracted from figures 3.15 and 3.16, by observing them conjointly. With the first sub-image (figure 3.16a), clearly depicting the semen stain, of which the structure seems to be related to a drying effect. From the correlated spectral channels (figure 3.15d), which are around 1900 to 2000 nm, a water contribution can be deduced, as these are combination O-H bands. The descriptor of the observed score is Variance, this is most likely due to the high variance observed between the dried and non-dried semen stain.

With the second sub-image (figure 3.16b), the fabric with the fibre filament is observed, with the correlated spectral channels (figure 3.15d) being around 1500 nm, where the 1st overtone of O-H stretching is present, and 2340 and 2450 nm, where combination bands of C-H and C-C are expected. The descriptor of the selected score is InfMeaCorr1, which is a measure of texture complexity, this makes sense in light of the complex cotton texture

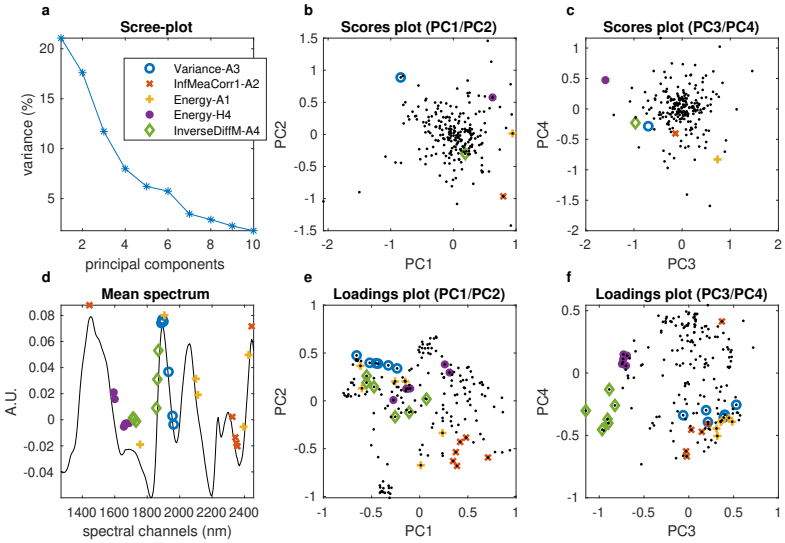


Figure 3.15: Overview of PCA results on DM of the semen data set, highlighting the selected scores and their correlated loadings in their respective plots. a) scree plot; b) scores plot of PC1 vs PC2; c) scores plot of PC3 vs PC4; d) mean spectrum of the raw data, with the highlighted loadings superimposed; e) loadings plot of PC1 vs PC2; f) loadings plot of PC3 vs PC4

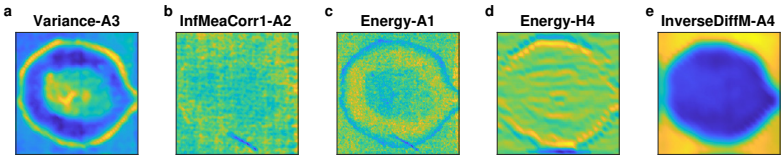


Figure 3.16: Overview of five sub-images from their respective selected scores, at the wavelength with the highest and most correlated loading value

and the presence of the filament.

With the third sub-image (figure 3.16c), the inverse of the first sub-image (a) is observed however, with the presence of the filament. The correlated spectral channels (figure 3.15d) are at 1750, 1900, 2100 and 2400 nm, of which some channels can be elucidated, however it is more likely that a scattering effect is observed, due to the dispersed nature of the correlated spectral channels. It would seem that no particular spectral band is highlighted, if compared to the other investigated scores. The Energy descriptor is connected to this score, which could be due to the intensity dependence of scattering effects.

The fourth sub-image (figure 3.16d) shows the border of the stain, with the correlated spectral channels (figure 3.15d) being, 1600 and 1675 nm. Although no specific band with respect to semen can assigned, from figure 3.8, we can clearly see that it is present in the original spectral image.

The final sub-image (figure 3.16e) shows the mask of the semen stain, in its entirety, with the correlated spectral channels at 1725 and 1850 nm (figure 3.15d), possibly linked to a protein/cellulose band [85] and the second overtone of C=O stretch, respectively. The InverseDiffM represents this particular score, and is linked to the homogeneity of an image, this is most likely due to the size and smoothness of the stain across the image.

In order to corroborate the results after exploration of the PCA analysis, k-means is applied, as previously explained (in this case, the descriptor matrix is compressed to 19 PCs, explaining 90.44% variance), figure 3.17 shows the clustering results. A more ambiguous clustering is obtained here compared with the previous case study, most likely due to the increased complexity and spatial overlap of the different components underlying the spectral image. However the previously determined components are still discernible, cluster 1 (figure 3.17b) is observed in the first sub-image seen in figure 3.16a, cluster 2 (figure 3.17c) is observed in the third image, however it would seem that the filament is not present. Clusters 3 (e) and 4 (f) show a mixture of the cotton fabric, the filament and the semen, however with some vertical artefact being present in cluster 3.

Taking into consideration, the sub-images of figure 3.16, the correlated spectral channels from figure 3.15d, the clustered spectral channels and their respective mean images in figure 3.17, some relevant information can be extracted. Starting with the semen stain, it is deduced that the drying effect

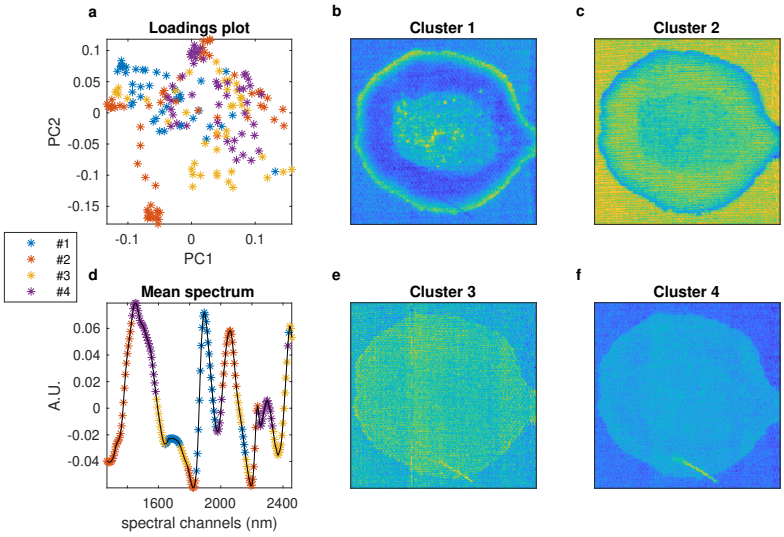


Figure 3.17: Overview of the clustering results on the loadings of the PCA analysis on DM from the semen data set. a) loadings plot of PC1 vs PC2, with the clusters highlighted in colour; b,c,e,f) mean images taken from the raw data for each cluster; d) mean spectrum of the raw data, with the clusters coloured

of semen and its border are present in the first, fourth and fifth sub-images, which correlated to the spectral channels at 1600 - 1700 and 1850 - 1950 nm. These exact spectral channels are highlighted in cluster 1, which shows more prominently the semen. With the third sub-image, it is deduced to be related to scattering effects, as no particular spectral band is highlighted, this is corroborated in cluster 2, where a similar structure is seen. The second sub-image is clearly related to cluster 4, as the spectral channels overlap. Precaution should be taken as the mixture of the different components within the clusters inhibits any further deductions.

It must be emphasised that compared with the emulsion dataset, the results of the clustering are not satisfactory in terms of isolation of the different compounds. However, it has been shown that the sub-images at specific spectral wavelengths and recovered by the PCA of DM do have the ability to isolate some aspects of the different components, with some overlap being present compared to previous work [15].

3.4.3 Bread

This data set adds a layer of difficulty on top of the previous data set, as the single images of the spectral image are very similar to each other. Applying IDEL on this data, one can see from the scree-plot in figure 3.18a, that the first PC explains more than 60% of variance. Following the same procedure as the previous two data, the most significant scores of the first two PCs are investigated, with their correlated loadings. Within this example, negatively correlated loadings are taken into account as well, as not every significant score would have a positively correlated loading, due to the shape of the loadings plot. This does not take away from any possible elucidations that might be done.

Starting with the first sub-image from figure 3.19a, as the decomposition is A1, no particular information can be extracted, as it shows an image, very similar to the original data. Comparing the scores and loadings plots (figure 3.18b and d), there are positive and negative loadings present at 960 - 1000 and 1520 nm, respectively. Some insight can be obtained, when observing the data, in figure 3.10, as going from the image at 977 to 1453 nm, a smoother structure is observed, which could be captured, by the Energy descriptor, which exacerbates these minor differences.

Within the second sub-image (figure 3.19b), although difficult to see, some intense points can be observed, that would seem to coincide with the intense points observed in the middle right part of the original images. The correlated loadings (figure 3.18c) correspond to 940 and 1375 nm, which could be assigned to the second and first overtone of O-H.

The third sub-image (figure 3.19c) is difficult to deduce, as it does not seem to represent any particular spatial structure. The score does not have any positively correlated loadings, however the negatively correlated loadings are at the maximum of the spectra, going from 1420 to 1480 nm.

The fourth sub-image (figure 3.19d) is slightly more clear, and would seem to show the smooth background that is present within the original spectral image. The correlated loadings are at 1250 - 1280 nm, which can assigned to the first overtone of C-H combination. The final sub-image (figure 3.19e), similarly to the third, shows no apparent structure. The correlated loadings (figure 3.18c) are between 960 and 1000 nm, this could be connected to the first sub-image, however, it could also be artefacts from e.g. illumination, that are usually captured by the deepest decomposition levels.

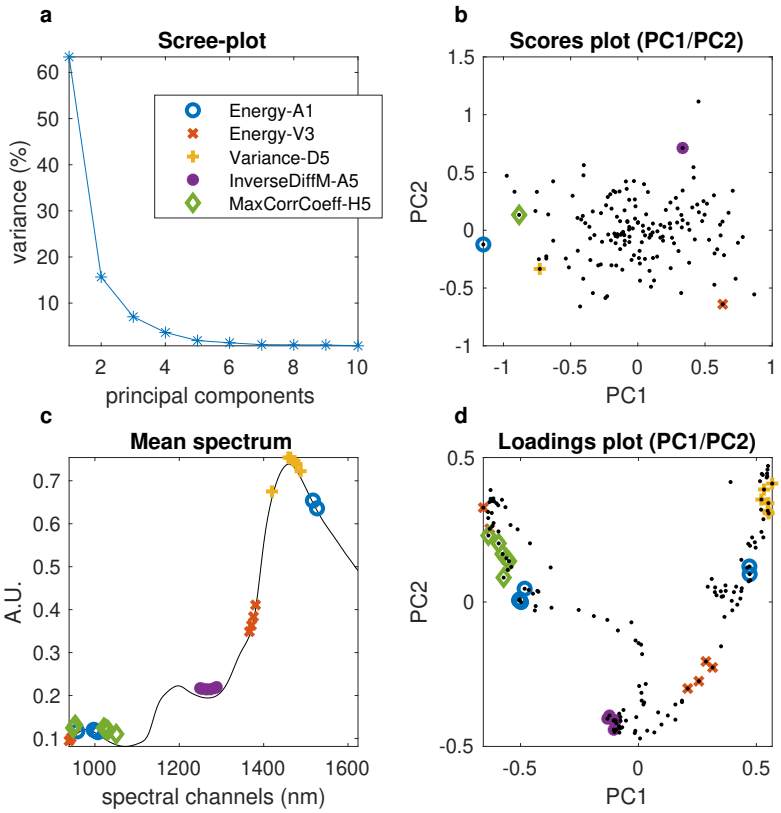


Figure 3.18: Overview of PCA results on DM of the bread data set, highlighting the selected scores and their correlated loadings in their respective plots. a) scree plot; b) scores plot of PC1 vs PC2; c) mean spectrum of the raw data, with the highlighted loadings superimposed; d) loadings plot of PC1 vs PC2

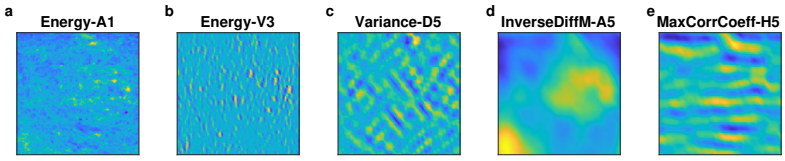


Figure 3.19: Overview of five sub-images from their respective selected scores, at the wavelength with the highest and most correlated loading value

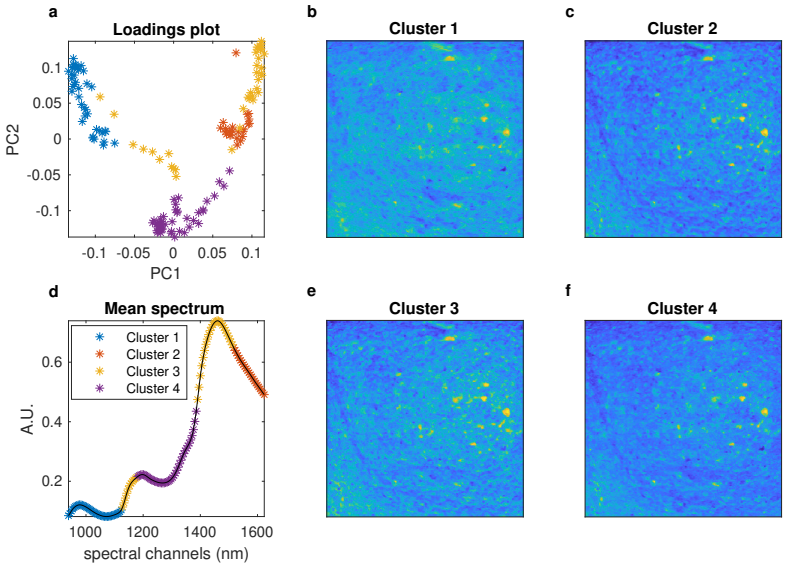


Figure 3.20: Overview of the clustering results on the loadings of the PCA analysis on DM from the bread data set. a) loadings plot of PC1 vs PC2, with the clusters highlighted in colour; b,c,e,f) mean images taken from the raw data for each cluster; d) mean spectrum of the raw data, with the clusters coloured

These results, are significantly more ambiguous, with respect to the previous data. However, some minor deductions are made. A final analysis is performed with k-means, seen in figure 3.20, however the results, reflect much of the original data, and do not show any relevant information. A point of interest with respect to this particular dataset is the presence of smooth and rough structures, however a deeper analysis of the data is required, which is not within the scope of this chapter.

3.5 Conclusion

In this chapter, a methodological framework based on combining both spatial features and spectral information for the analysis of spectral data is presented. The method decomposes every single-wavelength image of a three-dimensional array by 2D-SWT and computes individual GLCM for every

resulting wavelet sub-image. Morphological descriptors are estimated from all the GLCMs. In this way, spatial and spectral information is enhanced and conveyed in a single features matrix, which is finally processed by multivariate data analysis tools like PCA. Depending on the specific tasks the user must address, different multivariate statistical tools can be exploited at this point. Although the presented workflow combines different computational steps, which translates into higher complexity, every one of them is a necessary link in the chain. In fact, when some of these steps are skipped, the information obtained is insufficient to unravel all the distinct spatial features underlying the dataset under study and relate them to specific spectral regions (appendix B.2). According to the results obtained in three different case studies, it can be concluded that the proposed strategy is capable of recovering the main spatial features of a spectral image and of highlighting the most distinctive spectral regions. In particular, the outcomes related to the investigation of the semen stain dataset are found to be particularly promising considering the extreme physicochemical complexity of the examined image. In fact, even though the semen and cotton compounds show highly overlapped spectra, and cotton fabric is present everywhere in the sample, it is possible to highlight and localise the semen stain as well as the spectral channels at which specific forms are present. The bread dataset however lacking, showed the difficulty of working with single sub-images, and indicates the steps necessary to a better representation of the relevant information. In conclusion, this is an exploratory step towards properly defining the workflow of IDEL as well as showing its capabilities.

3.6 Perspectives

Further developments can be foreseen. Firstly, the retrieval of relevant information is insufficient, as a visual inspection is lacking when the number of sub-images becomes overwhelming or if the information of single sub-images do not represent the desired information. Secondly, the single steps can be further optimised, as desired i.e., different decomposition and or encoding methods. However this a significantly time-consuming step due to the broadness of the respective fields. Thirdly, the possibility of utilising other multivariate data analysis tools. Lastly, the improved and at least preliminarily disentangled spatial-spectral information returned by the described approach might constitute a valuable starting point for the design of new

constraints to be applied in the context of multivariate curve resolution [34].

Chapter 4

IDEL-Ω

This chapter is adapted from the IDEL publication [2], and is on the further development of the IDEL algorithm (Chapter 3), the representation of the final solutions, as well as the understanding of the underlying processes. Although IDEL is capable of retrieving an adequate representation of the relevant spatial structures within the data, the eyes of an expert are still required when the data is complex, as it is an exploratory tool. IDEL-Ω aims to simplify and streamline the methodology, by taking a more systematic approach.

The encoded spatial information is more extensively exploited by applying a semi-automatic procedure (that is data driven) that results in a set of distinct spatial features linked to the specific spectral channels at which they are observable. In this way, clear and precise spatial features can be extracted, while chemical interpretability is maintained. The approach is challenged with different data sets, that consist of simulations, controlled images, and sets of images that are relevant in the forensic field.

The increased use of spectral images in forensic applications makes this methodology particularly interesting. Taking body fluid detection in the forensic field as an example [15]. In such a scenario, forensic experts are often searching for compounds (such as blood, semen, and saliva) with specific spectral signatures that can link a crime scene to a victim, an assaulter or even a witness. However, those fluids usually appear on many different substrates, whose composition and texture characteristics can hamper its localisation, making it for example difficult for the analyst to identify its origin and, consequently, to submit them to further DNA analyses. In this chapter IDEL is challenged with a benchmark consisting of complex samples

made of semen and lubricant stains on cotton fabrics analysed with NIR imaging. There is significant spatial and spectral overlap between the stains and fabrics, and strong scattering effects are present. The localisation of the fluid on the substrate is of interest in forensic applications. As such, the segmentation of the biological fluid from the substrate as well as the removal of the significant scattering effects visible in the spectral imaging data is crucial.

4.1 Ω-algorithm

4.1.1 IDEL

The framework of IDEL is described in detail in Chapter 3 (figure 4.1a-d). Principal component analysis (PCA) of the descriptor's matrix (DM) highlights the descriptors (scores, figure 4.1e) that display the most variance, and links them to their corresponding sub-image from the decomposition step. In the first version of IDEL, the wavelengths (loadings, figure 4.1f) that correlate the most with this descriptor and sub-image are retrieved, by calculating the angle between the scores and loadings. However, in Chapter 3 we show that it is not always straightforward how to retrieve in systematic manner this information. Especially in complex cases, IDEL-Ω aims at overcoming this limitation by adopting a semiautomatic procedure to extract only the relevant sub-images at specific spectral channels (generating the Ω-domain) and then fusing them. These steps are detailed in the following paragraphs.

4.1.2 Ω-domain

With the aim of pushing the the algorithm further, a semi-automatic procedure is developed, which looks for relevant points in the scores plot while matching them to the loadings. This is a two-steps procedure.

Score selection

The first step consists of the selection of relevant sub-images from the scores plot. It is based on the estimation of the convex-hull of the scores (figure 4.1e). Convex hull is applied, instead of e.g., a thresholding on scores values, as it depicts the minimum set of distinctive points possibly enclosing all

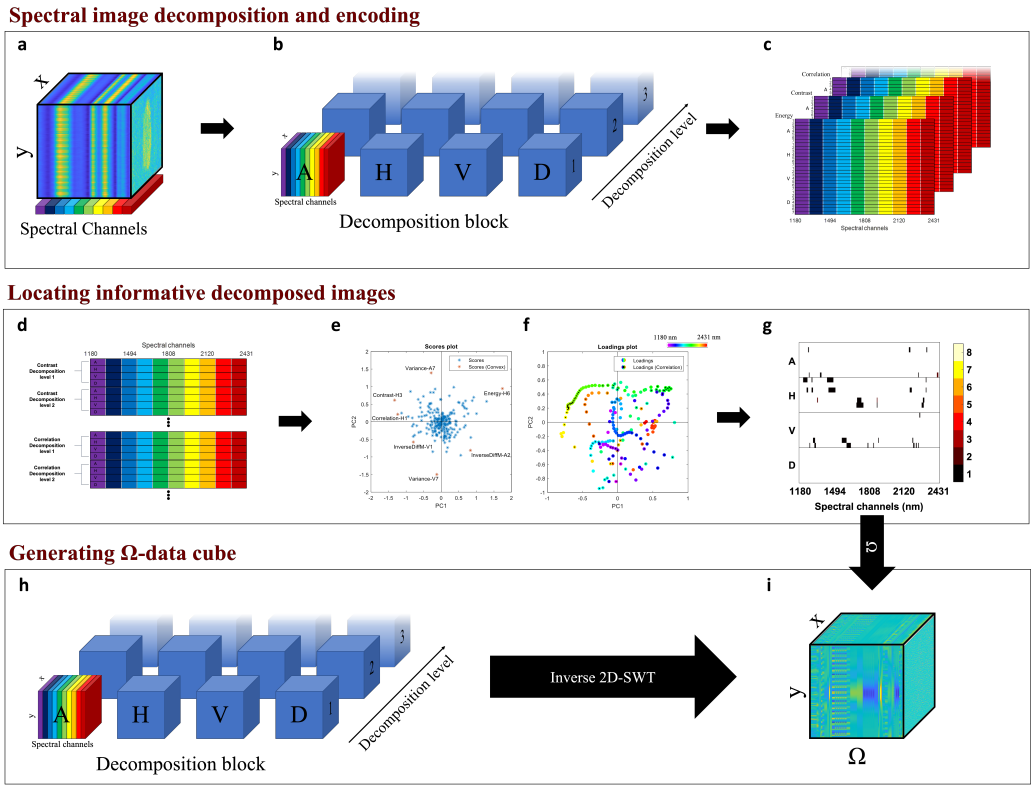


Figure 4.1: Illustration of IDEL-Ω. The methodology consists of three main actions. Firstly, "Spectral image decomposition and encoding", encompassing; a) a spectral image, which is decomposed by means of wavelet transform into; b) blocks containing horizontal (H), vertical (V) and diagonal (D) details, and approximations (A) at different decomposition levels; c) that are then encoded into a set of descriptors. Secondly, "Locating the informative decomposed images": where d) the Descriptors' Matrix (DM) is unfolded descriptor wise, retaining the spectral dimension, and e-f) decomposed by principal component analysis. The convex hull of the resulting scores is highlighted in red and labelled in the scores plot, while the corresponding correlated loadings are highlighted by a black point, inside the coloured point in the loadings plot. g) The scores (on the convex hull) and their respective (correlated) loadings are mapped in the Ω-map. The map reports on the "x-axis" the spectral channels and on the "y-axis" the decomposition block, to which each sub-image belongs, as well as the decomposition levels ordered from first to last (going down). Lastly, "Generating Ω-data cube", where the sub-images that are localised in the loadings map (g) are extracted from the reconstructed wavelet decomposition (h) and assembled into a data cube (i).

information captured in the scores plot. Depending on the complexity of the data a "peeling procedure" is implemented which consists of applying the convex hull twice. The first convex hull removes the first "peel" of the data and the second refines the selection. This accounts for situations where a few quite extreme points may skew the convex geometry too much. This procedure identifies the distinct sub-images that show the highest variation across spectral channels for the descriptors as, e.g., in figure 4.1e, the selected points (marked red, single peel) show significant variation on the first two PCs, meaning that they show high variation for a specific descriptor (within a certain sub-image at a given decomposition level across all spectral channels).

Loading correlation

The second step is to match the relevant spectral channels with the distinct spatial features (see figure 4.1f). To do this, the scores and loadings must be projected in the same space by adequate scaling, as in a bi-plot [86]. The correspondence of the loading points with the selected score points is expressed in terms of angle, which evaluates the location of the scores and loadings with respect to the origin of the PC-space. To identify the loading points that have a correspondence with specific score points, a threshold is set around 22.5 degrees (zero degrees meaning perfect correspondence, and ninety degrees, no correspondence). This was set, as it would cover an area of 25% of a circle as well as maintaining a minimum correlation of 0.92. The sign, of the scores and loadings, is not considered, meaning that a loading point that shows negative correlation (opposite location with respect to the PC origin) to a score point is considered equal to a loading point that shows positive correlation. We assume that positive and negative correlations between the scores and loadings have equal importance.

Ω-map

A single sub-image can be selected multiple times if it showed significant variation across spectral channels in several different descriptors. In fact, there are eight different points in the scores plot corresponding to each sub-image at a specific decomposition level, one for each descriptor. To give a clear overview of the selected sub-images at distinct spectral channels and highlight the sub-images that show significant variation for several descriptors, a representation is generated. This representation, depicted as an Ω-map in

figure 4.1g, maps the number of selected descriptors for every decomposition block and level of all sub-images vs the spectral channels. The Ω symbol indicates the sub-images selected by the procedure. The colour coding on the colour bar depicts the number of descriptors that are selected. In the end, only the sub-images that explain a significant amount of variance are kept. The spatial features within those sub-images that make up the different spatial components in the wavelet decomposition are of interest.

Ω-data cube

The selected sub-images are then reconstructed by inverse stationary wavelet transform (SWT) and reorganised in the so-called Ω-data cube (figure 4.1j). Even if the decomposed sub-images are of congruent size, reconstruction avoids spatial distortion with respect to the original image, which may be introduced at the deepest level of decomposition and brings the decomposed images back to original intensity scale. The Ω-data cube contains the wavelet sub-images at specific spectral channels, that isolated the significant spatial structures determined from a set of chosen descriptors. Also, the values in each of these sub-images, when assembling the Ω-data cube, are auto-scaled, and multiplied by $\sqrt{f}$, with f being the number of descriptors that have been selected for each selected sub-image. In this way, more weight is given to sub-images which show significant variation for more than one descriptor, meaning that different and distinctive spatial features are enhanced/captured by them.

The transformation from the spectral data cube to the Ω-data cube is the crux of the methodology. As the original spectral data consists of spectral correlation within the wavelengths and spatial correlation within the pixels. Ω transforms this to containing spatial-spectral correlation within the Ω-domain, as the dimension contains the most significant spatial information at their correlated wavelengths.

4.2 Image fusion

The Ω-data cube contains a subset of reconstructed wavelet sub-images exploited by 2D-SWT decomposition, however it may still include some redundant spatial information (spatial features visible at two or more spectral channels). Thus, it is desirable to further distill the information. We gener-

ically refer to this task as "image fusion" and different approaches may be used. A simple approach is to decompose the unfolded Ω-data cube by PCA. The refolded scores provide images (figure 4.2b) that combine spatial patterns which show a similar variation. The representation and interpretation of the loadings is slightly more complex, as they do not encompass all channels of the original spectral domain (figure 4.2c). The loadings are organised in such a way that they have the same dimensions as the Ω-map to get a clear overview on their importance (by means of the colour bar) at a specific decomposition and spectral channel. This results in a so-called loadings map. Beside it (figure 4.2d), for each PC, a plot of the mean spectrum of the original data is reported, with only the significant loadings highlighted by using distinct symbols/colours to indicate the corresponding wavelet sub-image, i.e. A, H, V, and D . This is done to visualise any correspondence of the obtained scores images with any spectral bands in the original data set, while the colour/symbol reference to a particular orientation of the relevant spatial features.

4.3 Ω-projection

An advantage of IDEL is that the generated PCA model can be used to project new imaging data, requiring only the 2D-SWT decomposition step to assemble the Ω-cube for the test images. Sub-images at the specific spectral channels (the ones belonging to the Ω of the training image) are calculated, which retrieves the Ω-cube of the new image. This is unfolded and projected onto the original PCA model, obtaining the scores (eq. 4.1), which in turn give the scores images by refolding, as well as the possibility of calculating the mean squared prediction error (MSPE, eq. 4.2). From the scores images, similarities and differences can be observed with respect to the model.

$$T_{test} = X_{test} \times P_{train} \tag{4.1}$$

where X_{test} is the test data, P_{train}, the loadings of the PCA model on the training data and T_{test}, the projected scores.

$$MSPE = \frac{\sum_{i=1}^{n} (X_{test} - (T_{test} \times P_{train}))^2}{n} \tag{4.2}$$

where n is the number of variables in the loadings.

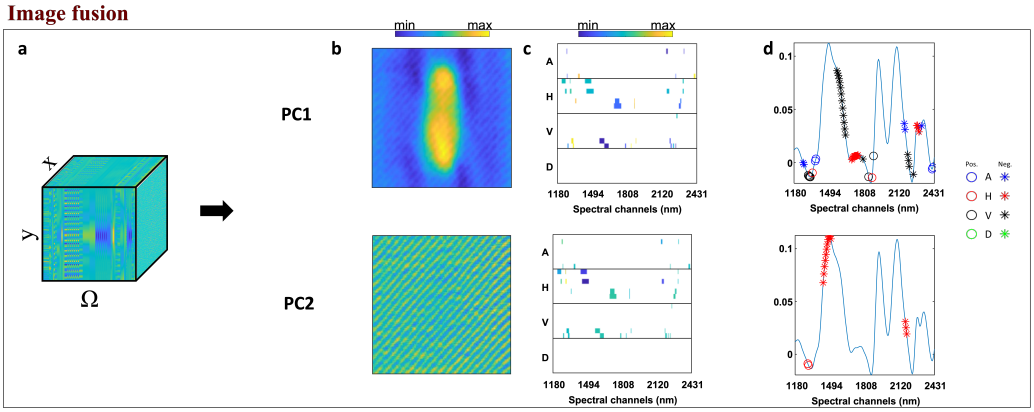


Figure 4.2: Example data set applied on the "Image fusion" methodology. Principal component analysis is applied on the unfolded Ω-data cube (a) and the resulting (refolded) scores images for the first two principal components are shown in (b). The loadings are mapped and visualised in a so-called loadings map (c), where the colour coding is set according to the loadings values. The mean spectrum for the original data is shown in (d), which highlights only loadings with absolute values > 0.075 (to declutter the figure, where negative values are denoted by a $\star$ and positive ones by an O), coloured according to the decomposition block: A (blue), H (red), V (black) and D (green).

4.4 Multivariate Curve Resolution (MCR)

A subsequent step that is taken, is the incorporation of MCR within the IDEL framework. MCR is a linear unmixing methodology that resolves the sources of variance that are present within bilinear data. This is discussed more in-depth in Chapter 5, however, the crux of the methodology is its ability to model bilinear relationships from multivariate data. There are different approaches to MCR, here specifically, alternating least squares (ALS) is utilised. MCR-ALS, uses non-negative least squares [34] to estimate the model parameters. Where the only requirement is setting the number of bilinear components present within the data, as well as an initial estimate for one of the two dimensions. With respect to spectral data, the model parameters would be the concentration profiles across samples and the spectral contributions across wavelengths, of the components. MCR-ALS aims at retrieving the isolated concentration profiles and pure spectral contributions of the various components within the data. The initial estimate can be estimated in a number of ways [33], within the scope of this chapter, only SIMPLISMA [87] and PCA are used.

Incorporating MCR into IDEL branches into two ways, the first is applying MCR on the Ω-cube, to retrieve its underlying sources of variance. By which, the cube is unfolded pixel-wise and resolved with MCR. The second is incorporating the information retrieved, from the PCA analysis on the Ω-cube, into the MCR framework, more specifically, the spatial information that is inferred from the results. Within this work, it conforms to utilising the scores images as initial estimates.

4.5 Data

There are a total of three different analyses performed, that will illustrate the capabilities as well as shortcomings of the methodology. The first data set (Simple simulation) is a simulation of a two component system where the goal is clear, but the data is difficult to analyse with standard linear methodologies as there are non-linear effects simulated. This is used, to illustrate the methodology as well as show its most well suited type of data. The second data set (Texture pack) is from a database of spectral images of various materials that is analysed with the exact same conditions. A superimposed image of vegetation on top of textile is analysed, where the

conditions for retrieving the underlying sources is incredibly difficult, due to the similarities in spectral signatures and complete spatial overlap. The third data set is from the forensics field, and is a set of ten NIR-images of stained fabrics. The goal of these data is consistent analysis, on top spectral elucidation and retrieval of relevant spatial structures.

The first two data considers the default analysis strategy for IDEL-Ω which is generating the Ω-data cube and applying PCA to retrieve the relevant components. For the stained fabric data set, in addition to the PCA step, a modelling approach is applied with MCR-ALS as well, to reflect the initial goal of the data.

4.5.1 Simple simulation

The first data has a reasonably simple premise of one single circular object (figure 4.3a), super imposed on a homogenous background, where both the image and spectral contributions (figure 4.3c) overlap completely. The spectral image size is 56×56 pixels and has 30 spectral channels. A multiplicative effect in the form of horizontal and vertical lines, and random noise with 5% of the maximum intensity of the original data is added. The final image (figure 4.3b) shows these horizontal and vertical lines, as well as some remnant of the initial circular object. The spectra (figure 4.3d) show more clearly the impact of the noise, however, some remnant contributions of the circular object are visible. The objective of the analysis is to retrieve the original spatial and spectral contributions of the circular object.

4.5.2 Texture pack

The second data set is two images artificially superimposed on top of each other, taken from the HyTexiLa database. HyTexiLa is a data base of spectral images that has a spectral range of visible (VIS) to NIR, consisting of 112 textured materials. The spectral images are analysed by using the HySpex VNIR-1800 spectral camera, manufactured by Norsk Elektro Optikk AS. The original data have an image size of 1024×1024 pixels and 186 spectral channels. These channels are associated to spectral bands centred at wavelengths which range from 405.37 nm to 995.83 nm at 3.19 nm intervals. A more extensive description of the full database is in [57].

The analysed images are a smaller squared patch from the original, as the original image size requires a considerable amount of computation. A

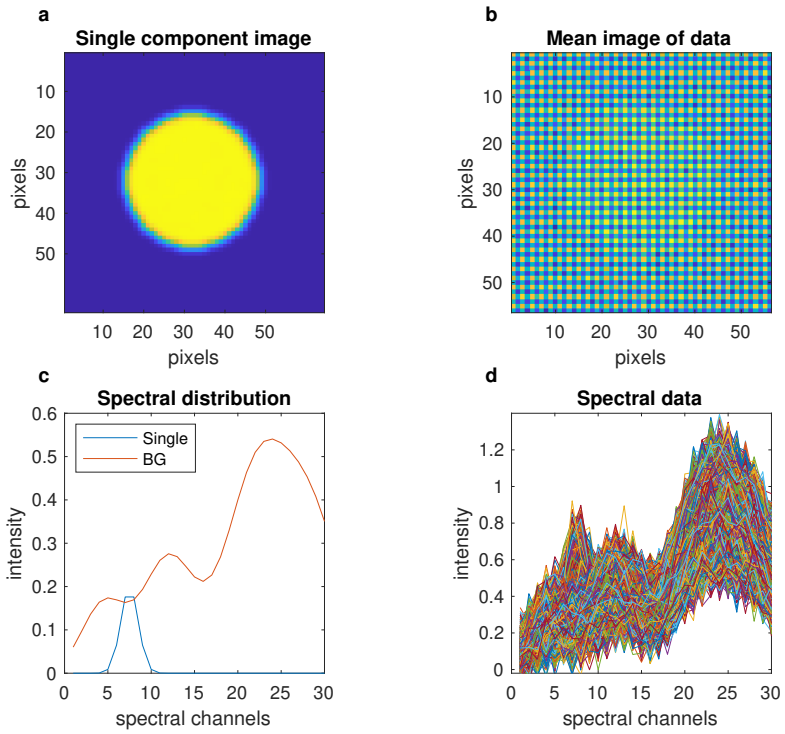


Figure 4.3: Overview of simulated spectral image. a) simulated image of the circular component; b) mean image of the simulated data; c) spectral contributions of the circular component (Single) and background (BG); d) spectra of the spectral image superimposed on one another

360×360 pixels cut is made of two images, firstly an image of vegetation, where green leaves are present, and secondly an image of blue textile. Figure 4.4a shows the superimposed mean image, where the vegetation seems more dominant. The mean spectra (figure 4.4b) are very similar. A PCA analysis is performed to get a better understanding of the data, with the first PC scores image (figure 4.4d, explaining 53.6% of variance) showing something similar to the mean image, and the loadings (figure 4.4c, blue) telling the same story. The second PC scores image (figure 4.4e) shows predominantly the centre of the vegetation, with some remnants of the textile being present. The loadings explain this, as the peak of the loadings of PC2 show positive correlation with the area where vegetation supersedes the textile spectra in intensity (figure 4.4b). PC3 (explaining 3% variance) shows relative high negative loadings in the first spectral channels and high positive loadings at 700 nm. This reflects the ratio of intensity for the two components, as textile is predominantly present in the first spectral channels and vegetation is present at 700 nm. In the scores image (figure 4.4f), the structure of both the vegetation and textile is visible. The objective of this data set is to isolate the two physical component by their spatial structures as well as retrieving their respective spectral channels.

4.5.3 Stained fabric

There are five differently coloured (yellow, white, red, green, and black) cotton fabrics, each with a stain of either lubricant or semen. All semen samples are obtained from the same donor [15], and the lubricant called KY-Jelly, mostly consisting of glycerol and hydroxyethyl cellulose, originates from the Durex© brand. The NIR imaging data is acquired by a Short-Wave Infrared (SWIR) SisuCHEMA imaging system from Specim (Oulu, Finland). The spectral range is 900 to 2500 nm with a spectral resolution FWHM of 10 nm and a spectral step size of 6.3 nm (256 spectral channels). The imaging system uses a lens of 50 mm and a pixel size of $156 \times 156 mm^2$. Squared pieces of fabric are stained with either a droplet of semen or lubricant, and left to dry for a week at room temperature. We refer to Silva et al. [15] for more details on the samples and data acquisition, see appendix D for the mean data of all ten spectral images.

These data are of interest for forensic applications and are used for the purpose of presumptive identification of biological fluids on textile. At least two physical constituents exist for each spectral image, the cotton and the

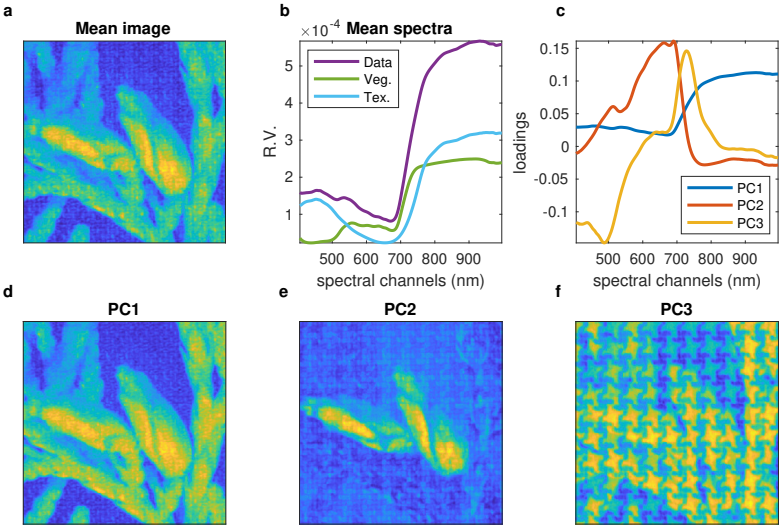


Figure 4.4: Overview of vegetation on textile data set, with a PCA analysis. a) mean image of the data; b) mean spectra of the data (purple), pure vegetation (Veg., green) and pure textile (Tex., blue); c) loadings of the first three PCs of the PCA analysis; d-f) refolded scores images of the first three PCs

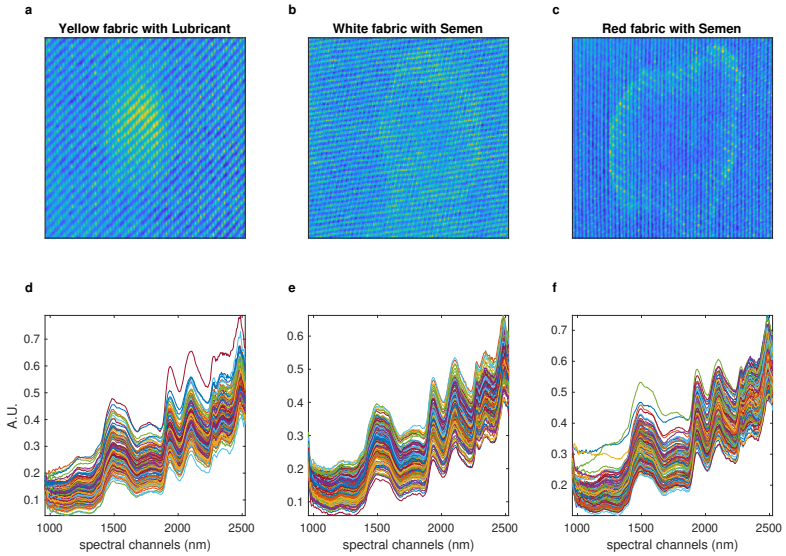


Figure 4.5: Overview on raw data of yellow fabric with lubricant (LY, a and d), white fabric with semen (SW, b and e) and red fabric with semen (SR, c and f), top plot showing the mean images and bottom plots showing 10% of the spectra

stain (lubricant or semen), but there is no spatial region without cotton, and the location of the stain may be detectable at selected spectral channels, but it is not clearly observed in the raw data as it is mixed with the cotton. Moreover, the fabric and stain show overlapped spectral bands.

Although the full data set contains ten spectral images, a subset of three is selected and fully discussed, namely the yellow cotton fabric with a lubricant stain (LY), the white cotton fabric with a semen stain (SW) and the red cotton fabric with a semen stain (SR), see figure 4.5 for an overview. LY and SW are fully elucidated, however SR is used as an example of projection, to gain a better understanding of the model that is generated by Ω. SR is projected onto LY and SW, and examined. Although the full data set is not extensively elucidated, the results from IDEL-Ω are provided, see appendix D.

Data pre-treatment

The angle of analysis for IDEL is image processing while maintaining spectral correlation. As such, the intended purpose of pre-processing is to contrast the spatial features within the images, while maintaining spectral correlation. To achieve this purpose weighted least squares baseline correction is applied, where a baseline is estimated for each pixel. However, firstly, the data was smoothed with a Savitzky-Golay filter [88] (11-point window, 2nd order polynomial) to account for any unwanted spikes in the spectra. Secondly, the first 40 and last 15 spectral channels are removed, as these show only noisy images, containing no significant information, as well as being distorted by the Savitzky-Golay filter. And lastly, weighted least squares (WLS) baseline correction (3rd order) [89] is applied. The pre-processing of the data is not the standard procedure for NIR data [90], as the aim of standard procedures is to generate bi-linearity within the data by harmonising the scattering and removing the variance of the path length (e.g. multiplicative scatter correction [41], MSC and standard normal variate [68], SNV). The preprocessed data are in figure 4.6, and similarly to the unprocessed data, from a spectral perspective the difference are not directly visible. The scores maps and loadings profiles resulting from its PCA analysis after pre-processing by means of WLS and two more standard pretreatment algorithms for NIR data (i.e., MSC, and SNV) are compared in appendix A.

4.6 Results and Discussion

Within this section the results of the analyses on the data are presented and discussed.

4.6.1 Simulations

IDEL is applied on the simulated spectral data, considering only a single decomposition level. The bi-plot of the scores and loadings of PC1/PC2 (15.1/8.9% variance explained) is presented in figure 4.7. The selected scores, by means of convex hull are named, with the most prominent loadings being labelled. Considering the information provided in figure 4.3c, the interesting spectral channels are 6 to 9, as these channels contain the spectral signature of the circular component. These channels have negative values with respect

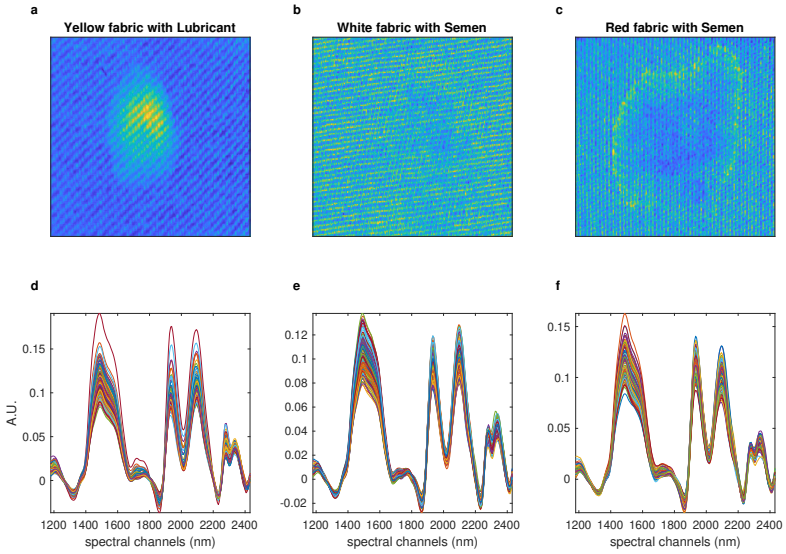


Figure 4.6: Overview on preprocessed data of yellow fabric with lubricant (LY, a and d), white fabric with semen (SW, b and e) and red fabric with semen (SR, c and f), top plots showing the mean images and bottom plots showing 10% of the spectra

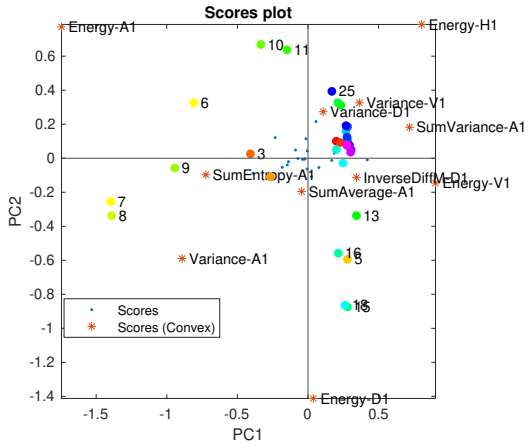


Figure 4.7: Bi-plot (PC1 vs PC2) of the PCA analysis on the DM of the simulation data set. Spectral channels are coloured from red to yellow, green and blue, indicating their respective numbers 1 through 30. Scores selected through the convex peeling procedure are labelled and coloured red (star), while all other are blue (dot)

.

to PC1, and are positively correlated with the Energy-A1, Variance-A1 and SumEntropy-A1 scores. While simultaneously being negatively correlated with Energy-H1, Energy-V1 and SumVariance-A1. This indicates that for the Energy descriptor, A1 is different compared to H1 and V1, and most likely is the indicator that will represent the circular object best. Energy-D1 shows a negative score value along PC2, which is in stark contrast with the other scores. The correlated loadings are 3, 15, 16 and 18, and do not reflect anything informative at the moment.

The overview of the analysis, presented in figure 4.8, shows the scores images(a, d, g and j), loadings maps (b, e, h and k) and highlighted spectral channels (c, f, i and l) of the first four PCs, from the PCA analysis on the Ω - data cube. The scores images of the first two PCs (a and d) show the vertical and horizontal lines that are caused by the simulated multiplicative effect. Their respective loadings maps (b and e) and highlighted spectral channels (c and f) indicate this by firstly being isolated in their respective decomposition and as well, being present across the entire spectral domain, indicating that not a singular spectral channel isolates the component. The third PC (g) shows clearly the circular object, completely isolated from the

multiplicative effects, with some minor amount of noise. Again, the loadings and spectrum highlight exactly the channels at which the simulated object absorbs. The final PC (j), shows only some noise effect that is captured at the A1 level, which is most likely the simulated noise. There is a lack of a background component in the PCs, which is clearly visible in the mean spectrum. This due to the fact that although the amount of variance present due to the background is high in the original image, its spatial distribution is non-existent. It is not isolated by the 2D-SWT as it has no distinct spatial features to capture.

Lastly for comparative purposes a PCA analysis on the unfolded original spectral image is performed, presented in figure 4.9. The first PC scores image (a, 98.29% variance) shows the mean image, while the second PC (b, 0.38% variance) shows the circular object, with the multiplicative effects still being present, indicating that the object cannot be isolated by such means. The third PC (c, 0.06% variance) shows only some random noise. The loadings plot (d), reflects much of the same information, where the first PC (blue) show the mean spectrum, the second (red), mostly the 6th through 10th channel with some negative values in the 20th to 30th channels. The third (orange), showing mostly noise. The results are sub-optimal with respect to IDEL-Ω.

4.6.2 Texture Pack

From the PCA analysis, in figure 4.4, the scores images show that the spectral image cannot be explained with just two principal components. This is an indicator of the level of complexity that data has. This means that more than one component is necessary to explain the two different compounds, i.e. vegetation and textile. For example, with the textile, if within this physical component, different spatial structures are visible at different spectral bands, retrieving and linking these structures to the singular component, becomes quite difficult. This will be highlighted in the results, presented in figure 4.10. The scores images show, for PC1 (a) and PC2 (d), a faint shadow of the vegetation, with an even fainter structured background (textile), the accompanying spectral bands highlighted in c and f, respectively, are around 400, 450 and 650 nm. Most likely textile originates from the 400 nm and 450 nm bands, while the 650 nm band is from the vegetation. No isolation of either the vegetation or textile is observed, and for the scores images of PC3 (g) and PC4 (j), a similar story is present, as mostly a structured background is

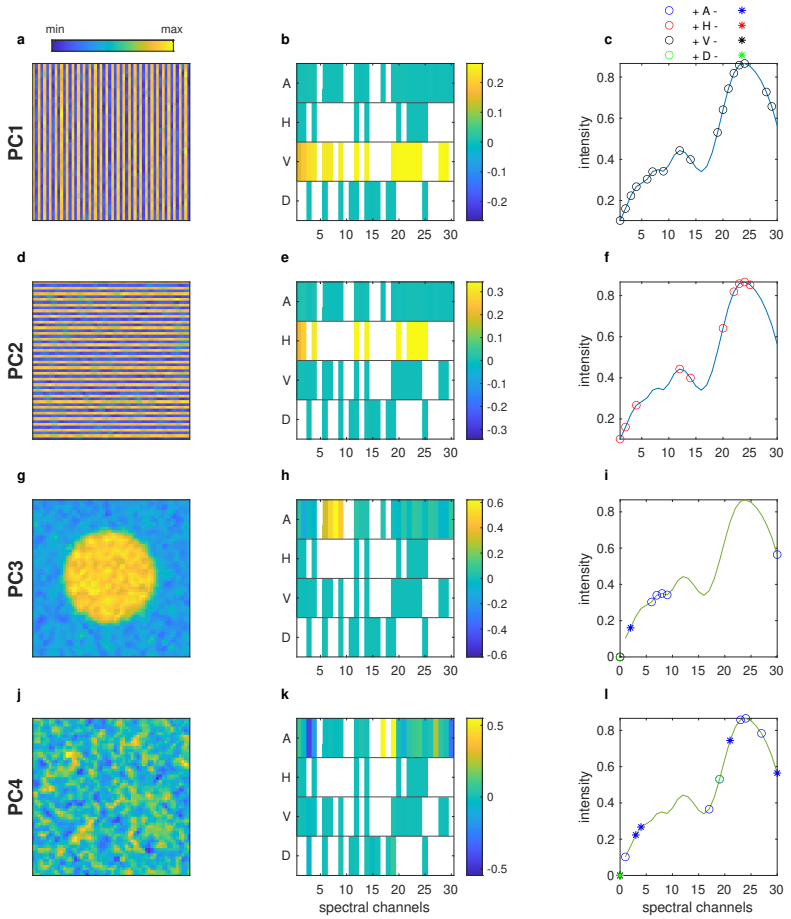


Figure 4.8: Overview of the results from IDEL on the Ω - data cube from the simulation data set, showing PC1 through PC4. scores images, for the respective PCs at a, d, g and j, loadings maps at b, e, h and k, and highlighted spectral channels superimposed on the mean spectrum of the original data at c, f, i and l. The highlighted spectral channels are extracted from the loadings maps and differentiated by colours (blue (A), red (H), black (V) and green (D)) and symbols (circle, positive correlation and star, negative correlation)

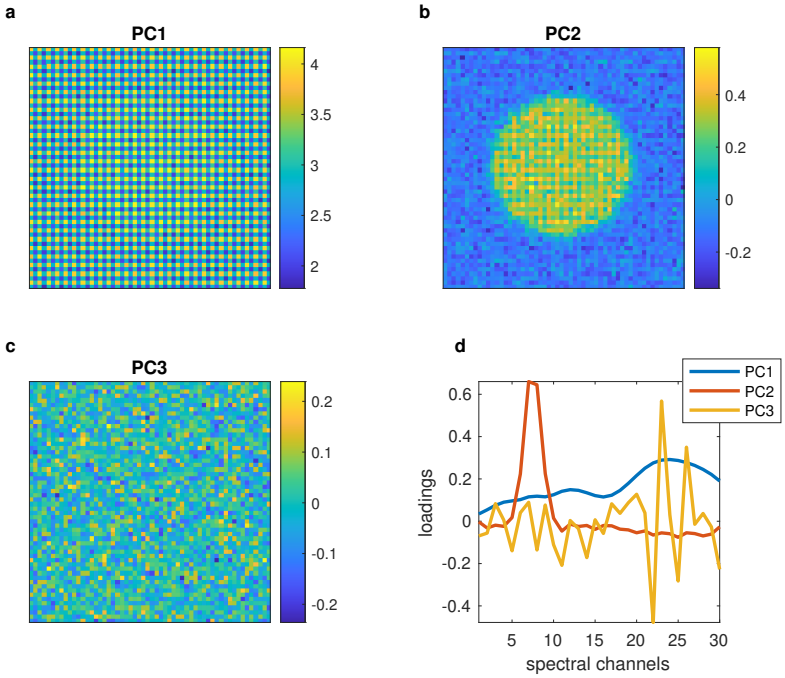


Figure 4.9: Overview of PCA analysis on simulation data set. a-c) refolded scores images of their respective PCs; d) loadings plot of the first three PCs

present, with some faint shadow of the vegetation. Although the highlighted spectral channels are mainly from the 400 to 500 nm bands, where mainly the textile is present, the images show that some of the vegetation is present.

When going deeper into the PCA analysis on the Ω-data cube, at PC12 an isolated structure of the vegetation is found, presented in figure 4.11. The scores image (a), shows the border of the vegetation, with the spectral bands at 600, 700 and 1000 nm being highlighted (c). These results together with the results of the PCA analysis of the original data (figure 4.4), highlight one of the drawbacks of using PCA with IDEL-Ω, which is the fact that when highly complex structures are present within the data, data diving becomes necessary to retrieve the relevant information.

4.6.3 Stained fabric

IDEL is applied on the LY and SW data sets, and the results go over the PCA analysis on the Ω-data cube. The elucidation considers the scores images, loadings maps and highlighted loadings on the mean spectrum, as shown in the example in figure 4.2.

Lubricant on Yellow cotton

The Ω-data cube for the LY data set consists of 140 sub-images, extracted from the wavelet decomposition, and the results are reported in figure 4.12. The resulting scores images (figure 4.12a) of the first 4 PCs (explaining 58.3% variance of the Ω-data cube) are considered, where four distinctive spatial features are clearly recognisable. One can clearly identify the stain and cotton fibre pattern, as is discussed below.

The PC1 scores image (figure 4.12a) mainly shows the presence of an intense spot (almost in the centre) which can be identified as a stain. The corresponding loadings map (figure 4.12b) shows that all the wavelet sub-images from every decomposition block (A, H, V and D) are contributing to the model, however the highest loadings values are mainly associated to approximation sub-images (A), which retain low frequency contributions in the original data set, hence smooth patterns. Figure 4.12c represents the relevant spectral wavelengths on the mean spectrum of the original data with the notable points being: i) approximation sub-images at decomposition levels 2 and 6 (A-2 and A-6), which are linked to positive loadings within the spectral region 1990 to 2060 nm, ii) sub-images A-4 and 5, linked to negative

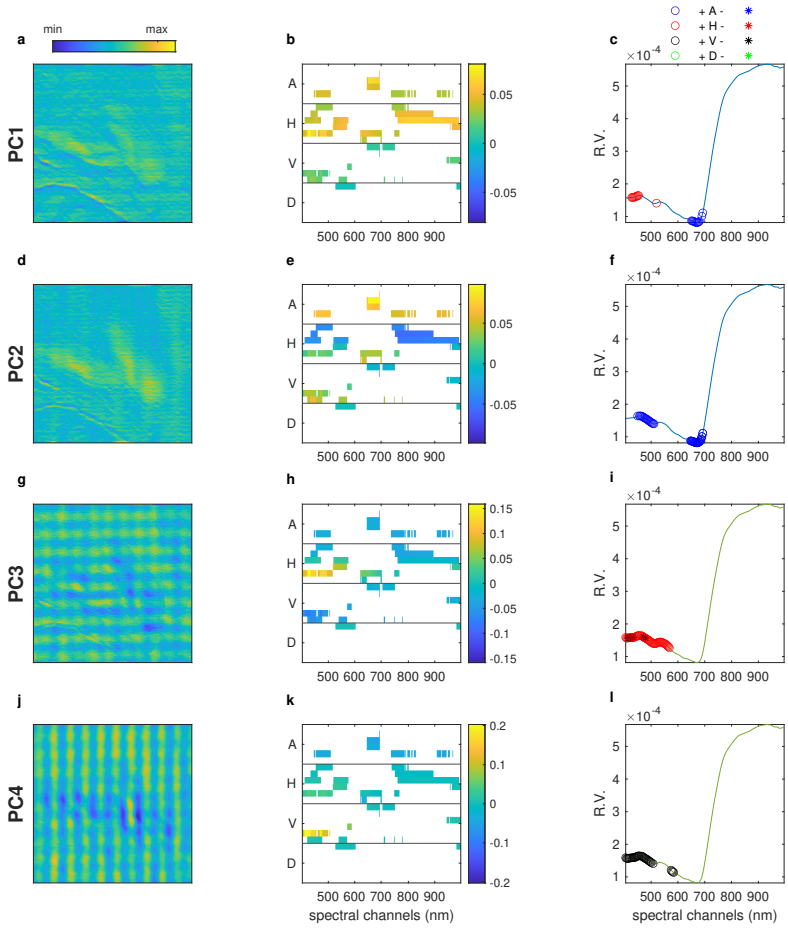


Figure 4.10: Overview of the results from IDEL on the Ω - data cube from the vegetation on textile data set, showing PC1 through PC4. scores images, for the respective PCs at a, d, g and j, loadings maps at b, e, h and k, and highlighted spectral channels superimposed on the mean spectrum of the original data at c, f, i and l. The highlighted spectral channels are extracted from the loadings maps and differentiated by colours (blue (A), red (H), black (V) and green (D)) and symbols (circle, positive correlation and star, negative correlation)

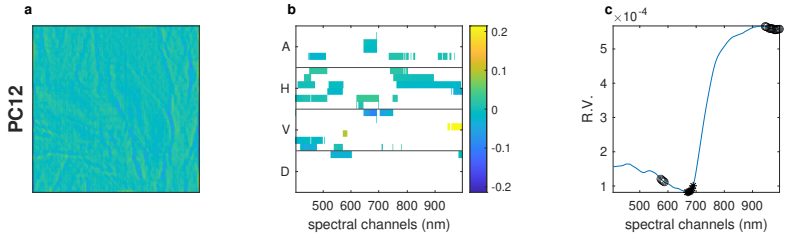


Figure 4.11: Overview of a subset of results from IDEL on the Ω - data cube from the simulation data set, showing PC12. a) scores image; b) loadings maps; c) highlighted spectral channels superimposed on the mean spectrum of the original data. The highlighted spectral channels are extracted from the loadings maps and coloured black (V) with symbols circle for positive correlation and star for negative correlation

loadings around 2400 nm, and iii) A-6 linked to negative loadings around 1300 nm. Although it is extremely difficult to consider band assignment in NIR for such complex matrices, the band around 2000 nm suggests contributions from glycerol [91], one of the main compounds of the lubricant. The negative contribution at 1300 nm is interesting as well, as neither cotton nor glycerol absorb at that wavelength. This contribution could be linked to a solely physical effect, due to the lack of absorbance of either cotton or lubricant, or it could be linked to a third unknown component.

The PC2 scores image (figure 4.12a) clearly shows the diagonal texture associated to the cotton fibres. The loadings map (figure 4.12b) shows that the most relevant contributions are from the H and D sub-images in the spectral range from 1400 to 1550 nm. When looking at the highlighted loadings (figure 4.12c), all these contributions relate to the band centred at 1494 nm. This can be attributed to the first overtone of O-H in cotton [92]. The main contribution comes from the D sub-images, however some minor contributions come from the H sub-images. This can be attributed to the large spacing that is seen between the diagonal fibres, which can be captured in the horizontal details.

The PC3 scores image (figure 4.12a) is not straightforward to interpret. It shows some very smooth patterns, which are usually captured at the deepest decomposition levels (low frequency contributions in the spectral images)

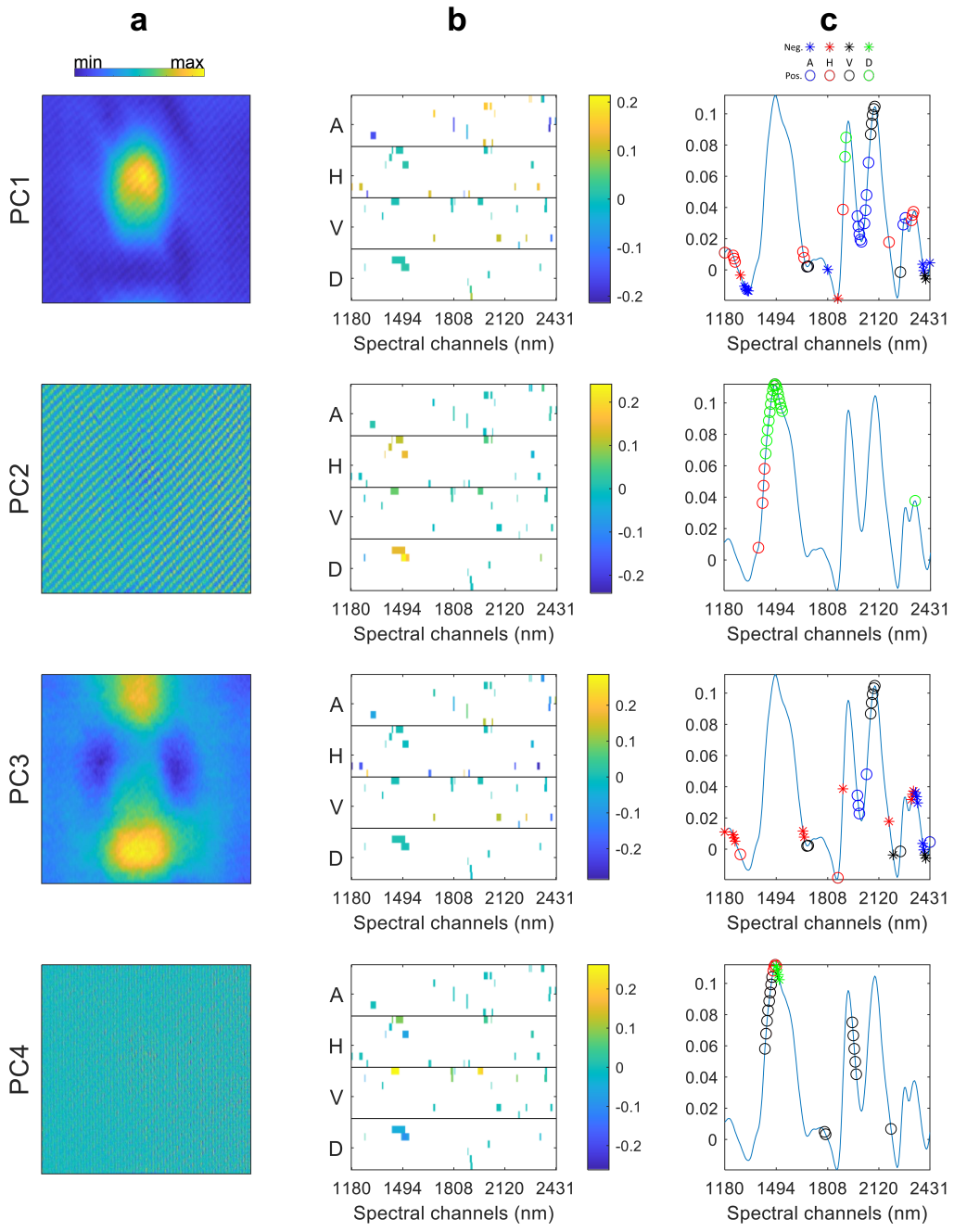


Figure 4.12: Overview of the results from IDEL on the Ω - data cube from the LY data set, showing PC1 through PC4. a) scores images, for the respective PCs; b) loadings maps; c) highlighted spectral channels superimposed on the mean spectrum of the original data. The highlighted spectral channels are extracted from the loadings maps and differentiated by colours (blue (A), red (H), black (V) and green (D)) and symbols (circle, positive correlation and star, negative correlation)

and mainly by the approximations. However, details may also capture this type of information when, as in this case, there could be low frequency orientation specific spatial patterns present (the most intense loadings values are from the H and V sub-images). When looking at the highest values in the loadings map (figure 4.12b) and at their location on the mean spectrum (figure 4.12c), the contributing spectral regions are quite spread and mainly on shoulders or along the spectral baseline. These patterns are quite difficult to interpret, and can originate from various sources, e.g. non-homogeneous illumination of the surface. These points can introduce minor variations in an image that can be seen in the deepest levels of a wavelet decomposition.

Finally, PC4, as for PC2, shows the texture of the cotton textile, however now its pattern has mostly a vertical orientation. The main contributions are the details sub-images (mainly V) and again the relevant spectral region includes the band centred at 1494 nm. Added to this is a contribution from the spectral band at about 2000 nm, which was not captured by PC2. This spectral channel is slightly shifted with respect to the contribution discussed in PC1. This could be attributed to the first overtone of R-CO-R. Possible reasoning for PC4 to be separated from PC2 is that the spatial structure is significantly different and is isolated as a different component. Even though they both originate from cotton, the overlapping fibre structures show significant differences.

Summarising the results for the LY data set, different spatial features could be isolated, segmenting the stain and recovering the cotton fibre patterns across the whole image in the scores images. A possible link to the spectral domain has also been established, where the interplay of chemical and physical information is observed.

Semen on White cotton

The Ω-data cube consists of 491 sub-images, extracted from the wavelet decomposition. The results of the PCA analysis of the Ω-data cube for the SW data set are shown in figure 4.13. The first four PCs (explaining 52.1% variance), which capture the different spatial structures, are discussed below.

The PC1 scores image only shows the semen stain without any pattern related to the texture of the fabric (see figure 4.13a). The loadings map (figure 4.13b) highlights several contributions but the highest (in absolute

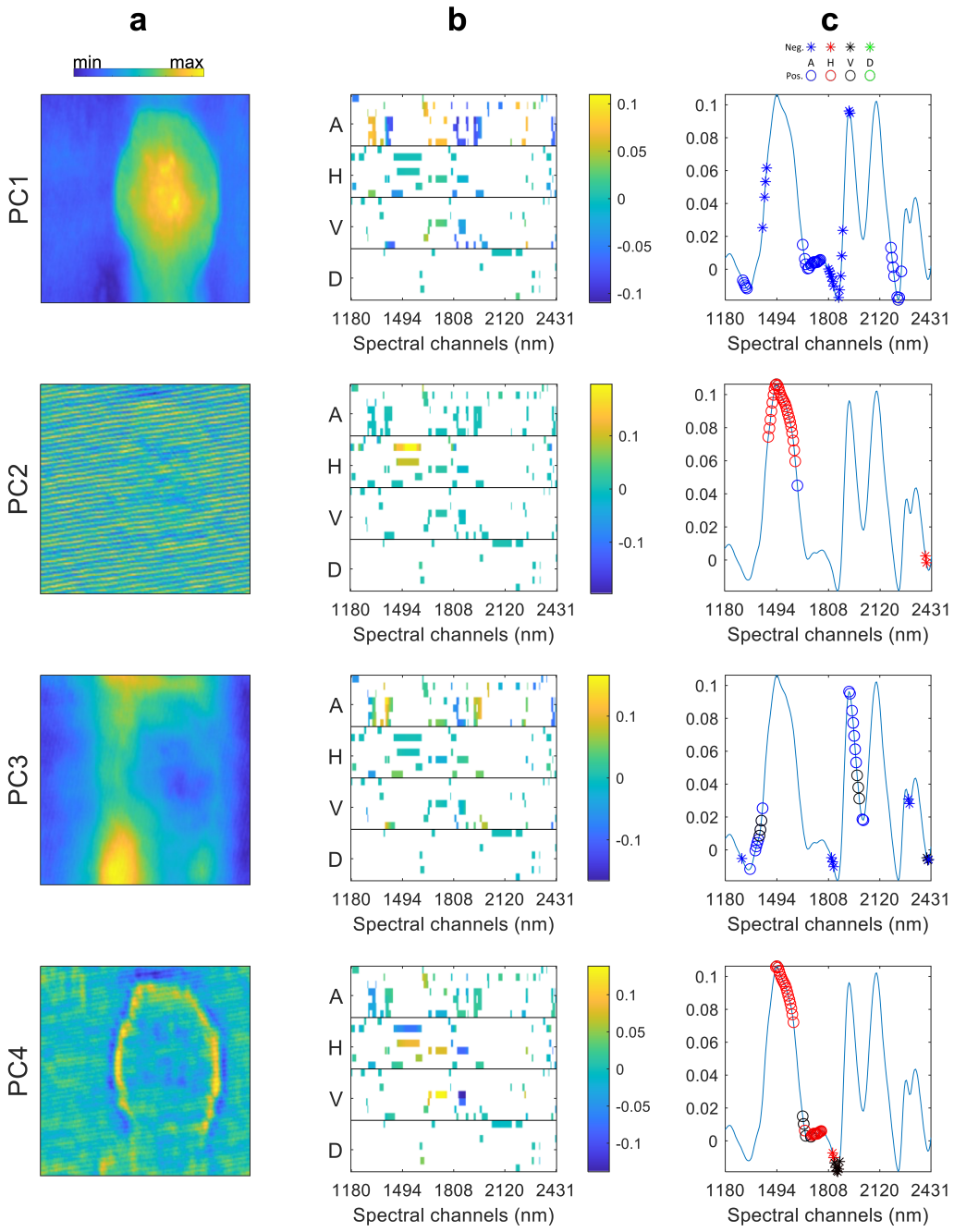


Figure 4.13: Overview of the results from IDEL on the Ω - data cube from the SW data set, showing PC1 through PC4. a) scores images, for the respective PCs; b) loadings maps; c) highlighted spectral channels superimposed on the mean spectrum of the original data. The highlighted spectral channels are extracted from the loadings maps and differentiated by colours (blue (A), red (H), black (V) and green (D)) and symbols (circle, positive correlation and star, negative correlation)

terms) are from the A sub-images across most of the decomposition levels. Reporting the correlated loadings on the mean spectrum (figure 4.13c), the corresponding spectral regions are located at: 1300 nm, 1700 nm and 2200 nm, showing positive loadings values, 1450 nm, 1850 nm and 1940 nm, showing negative loadings values. The contributions at 1700, 1850 and 2200 nm could relate to semen, as they could be attributed to protein bands [85]. However, the band at 1300 nm is not attributable to a specific component: it could be that this is solely associated to physical scattering effects that come into play, as something similar was observed in the lubricant example. The 1450 and 1940 nm bands could be attributed to water bands [93], as the loadings show negative values and a faint negative circle is observable in the lower left part of the corresponding scores image (figure 4.13a, PC1). A similar contribution can be seen on the PC3 scores image (figure 4.13a) but with an inverted sign (positive values of scores and loadings). Even if the samples are dried, it cannot be excluded that water on the border is reabsorbed due to the environmental conditions, since the humidity of the room (where the samples are stored) was not controlled.

As in the lubricant data, the PC2 scores image depicts the texture linked to the cotton fibres. However, here the fibre orientation spatially manifests in the horizontal direction. As such, the H sub-images are mostly selected (figure 4.13b, PC2). The relevant loadings highlighted on the mean spectrum (figure 4.13c) are associated to the absorption band at 1494 nm, which has already been referred to as the first overtone of O-H stretching in cotton.

The PC3 scores image shows, like for the lubricant data, smooth spatial patterns. However, a small intense circle is also visible in the bottom left part of the image. Looking at the relevant loadings, both in the loadings map and reported on the mean spectrum, we see that mostly A and V sub-images have the highest absolute loadings; the relevant spectral channels are in large part the same as for PC1, e.g. 1300, 1450, 1830 and 1940 nm. In fact, the simultaneous absorbance around 1450 and 1940 nm could be linked to water, which could mean that what is observed is due to a drying effect at the border of the semen stain. Analogously, similar, but negative spectral contributions are observed in PC1. In the image, the semen stain border has an elongated form in the vertical direction. As such, it is being captured by the vertical details (V sub-image). On the other hand, the A sub-images capture the small spot, which is linked to semen.

The PC4 scores image shows the border of the semen stain, captured by H and V sub-images, contributing the most to the loadings map. However, the contribution from the texture of cotton is observable. Around the border of the stain, the spectral contributions from the cotton fabric and the semen stain are strongly overlapping. The relevant spectral regions include the 1700 nm band, already discussed for PC1 as connected to semen, and the 1500 nm band connected to the cotton fibres, mentioned with regards to PC2.

The compression (or "fusion") step operated by PCA was extremely efficient at extracting information, separating not only the semen stain from the texture (which consists of the scattering effects of cotton), but also distinguishing the scattering around the border of the semen stain together with a possible drying effect of the semen.

Semen on Red cotton - Projection

We have seen that the proposed approach is very efficient in retrieving spatial information and interpreting it in terms of spectral contributions. In particular, the scores images obtained from the analysis of the Ω-data cube help in discerning the various spatial components, which are not observable separately at any single spectral channel in the original data. The loadings highlight the spectral channels at which those components mostly manifest. A clear next step can be foreseen, which is evaluating if new (test) images projected on the loadings of a reference image can extract the same kind of specific spatial information in the scores images.

The SR data set is investigated (figure 4.14a). The system is sufficiently similar to SW, as both contain cotton and semen, however the shape of the stain, and the orientation of the scattering effects and colour of the fabric are different. With respect to LY, the only similarity is the fact that the fabric is made of cotton. The resulting scores images are shown in figure 4.14b. For the SW-model, in the projected scores images, similar spatial features can be observed: the semen stain is isolated in PC1, the texture of the cotton fibres with some bordering effects is seen in PC2, in PC3 a bordering effect linked to the semen stain is visible, and finally, the joint border and scattering effects are highlighted along PC4. Although the texture of the cotton fibres is not completely isolated from the border effects in PC2, the semen stain has been isolated and identified. These minor differences may

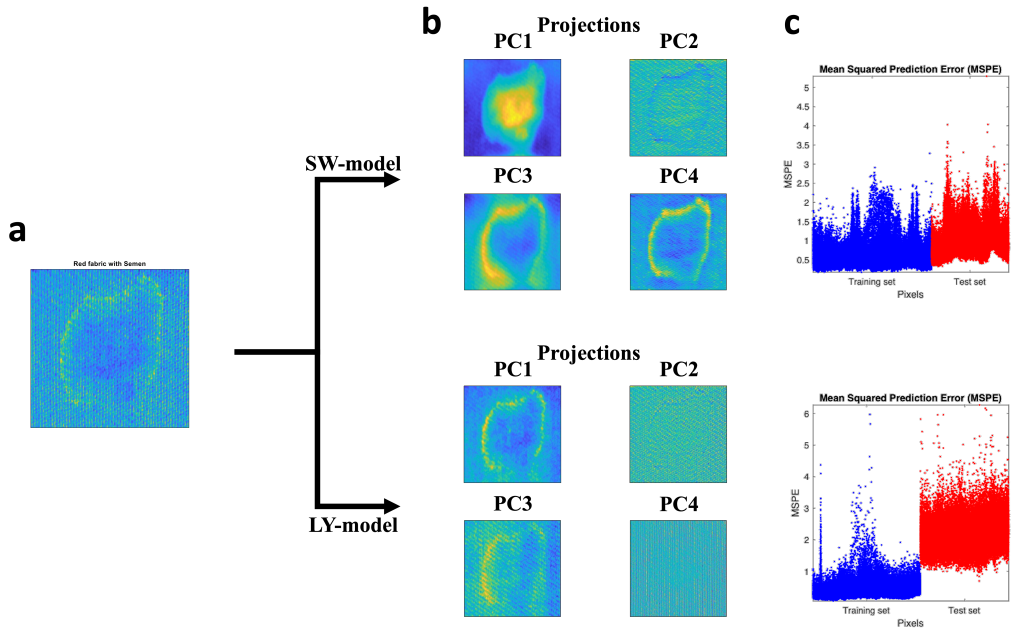


Figure 4.14: Overview of the results of projecting SR onto the SW-model (top) and LY-model (bottom). a) mean image of SR data set; b) refolded projected scores images for PCs 1-4; c) mean squared prediction error (MSPE) for the SW (top) and LY (bottom) projections

be due to spatial and spectral differences between the data sets. The texture of the cotton fibres is not the same, i.e., it is oriented in a different direction with respect to the modelled image. In addition, the colour of the fabric is different: red fabrics might exhibit a different absorption with respect to white within the NIR range. Also, the amount of deposited semen may not be the same, nor its position or shape. In figure 4.14c, the MSPE of both SW (training) and SR (test) exhibit a similar error. Moving towards the LY-model, a stark difference is observed, in the scores images, where for the LY images (figure 4.12a), a stain is observed in PC1, here scattering effects, with the border of the stain is seen. This is also the case for PC3, where with LY a large structure is observed, in the projection, again, scattering effects with the border of the stain is observed. PC2 and PC4, do show similarities with respect to LY. These images could be explained, by the fact that probably with LY, mostly physical effects are isolated and this is observed in the projection. With the SW-model, more chemical effects are seemingly modelled, if the elucidation is to be believed. The MSPE further highlight the differences between the LY and SR data, as there is a more prominent difference than with SW.

Overall, these results seem very promising. The results of the projection of all ten data sets onto the SW-model are in appendix D, and show similar conclusions. However, the projection of the black fabric with a semen stain (SB) and, to a minor extent, of the yellow fabric (SY), while showing similar spatial features on scores images, resulted in high a MSPE. If the spatial structures and/or spectral background (as it is the case for SB) of the test images are very different from the calibration image, some care should be taken in interpreting the scores images, even if interesting spatial structures are retrieved.

MCR-ALS

The SW results are taken as an example for a possible incorporation of the IDEL-Ω algorithm into the MCR-framework. This is compared to a standard MCR-ALS analysis, where the original data are pre-processed, by MSC to generate a bilinear data matrix and SIMPLISMA is used to find the initial estimates.

A general analysis is done, by changing two parameters within the MCR-ALS algorithm. Firstly the data set used within the algorithm, this is either

the data preprocessed with MSC, or the Ω-data cube. Secondly, the initial estimate used, which is set as either the SIMPLISMA initial estimates or the scores images obtained from the PCA analysis on the Ω-data cube. This translates to four analyses in total: 1) MSC data with initial estimates from SIMPLISMA; 2) MSC data with initial estimates from the PCA analysis on the Ω-data cube; 3) Ω-data cube with initial estimates from SIMPLISMA; 4) Ω-data cube with initial estimates from the PCA analysis on the Ω-data cube;

Firstly the standard analysis, where the MSC data is unfolded pixel-wise and standard MCR-ALS is applied, using SIMPLISMA as an initial estimate with four components. Figure 4.15, shows the results, in the form of refolded concentration maps and spectral contributions. A lack of fit (LoF) of 1.07% is obtained, stopped at a $\Delta LoF < 10^{-5}\%$, with the non-negativity constraint active. The concentration maps are reasonably different for the first and second component, and seem to show the semen stain and fabric, respectively. However, there is a significant contribution of the fabric for the first component, as well as a minor contribution of semen in the middle of the map, for the second component. The third map shows the semen border, from which a similar image is observed in the IDEL analysis, indicating a possible drying effect being present. However, again, with clear contributions from the cotton fabric. The fourth map, shows seemingly an inverted image with respect to the first concentration map. There is no single component that shows an isolated contribution. The accompanying spectral contributions do not give us any further elucidation. For the first component, it would seem to have isolated multiplicative effects, due to its shape. The other three spectral contributions show more or less the exact same shape, with differing ratios of spectral peaks, making it difficult to elucidate any specific spectral bands.

The second analysis uses the same MSC data set, however, as an initial estimate, the scores images of the previous PCA analysis on the Ω-data cube are used. The results (figure 4.16) do not significantly change and more or less of the same conclusions can be made with the same LoF of 1.07% being obtained. The only visible change that is seen is in the concentration maps of the third and fourth component, however, these differences seem negligible and, due to the little to no differences in the spectral contribution, no definitive conclusions can be made.

For the third and fourth analysis, instead of the MSC data set, the Ω-

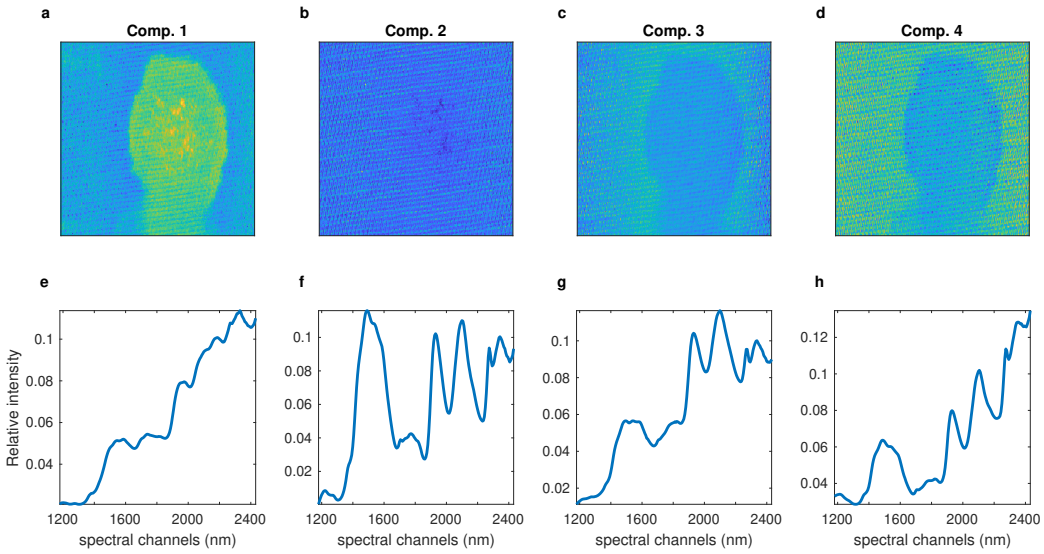


Figure 4.15: Four component MCR analysis on MSC data, taking SIMPLISMA as initial estimates. a-d) refolded concentration maps for the four components, with e-h) their respective spectral profiles

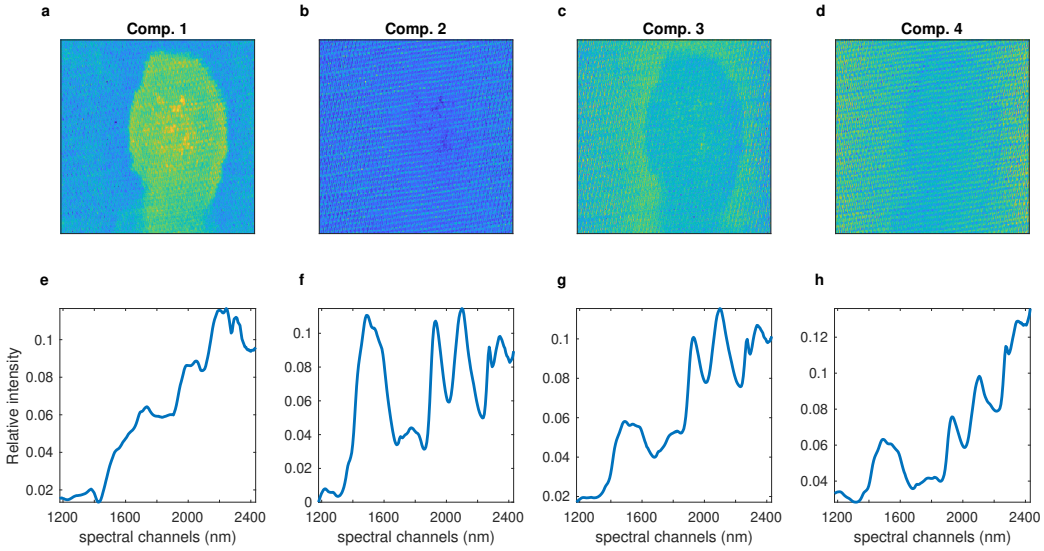


Figure 4.16: Four component MCR analysis on MSC data, taking Ω as initial estimates. a-d) refolded concentration maps for the four components, with e-h) their respective spectral profiles

data cube is used and again, unfolded pixel-wise. Both SIMPLISMA and the scores images of the previous PCA analysis are used in the MCR-ALS analysis, retrieving a LoF of 6.44%. Starting with the SIMPLISMA results, refolded concentration maps are obtained, see figure 4.17. Some overlap is observed with respect to the previous analysis, however for the semen stain (component 1), a much clearer map is obtained. Component 2 and 3, show similar maps, with some combination of fabric, border and drying effect being present. The fourth component, however, does seem to indicate a different structure being present, that has not been previously seen. Moving on towards the spectral contribution, a different approach is required to elucidate the results, as a subset of spectral channels is selected at different decompositions, with the Ω-data cube. The spectral contribution plots are structured in a way that indicate their spectral channel, by position, and its accompanying decomposition from the wavelet transform, by colour. Taking the first spectral contribution plot as an example, mostly, blue dots are visible, indicating approximation sub-images, at the 1300, 1700 and 2200 nm bands. These bands are already elucidated, by the previous IDEL-Ω analysis of this data in section 4.6.3 Semen on White cotton. An interesting point is the inclusion of horizontal and vertical decompositions at the 1700 nm band, which might be the cause of the soft fabric structure that is visible in the concentration map. This indicates that not a completely isolated semen stain is obtained.

For the second and third component, significant overlap is observed, with respect to both the concentration maps and spectral contributions. Both components show the fabric with the border of the semen stain within the maps, that was previously elucidated to be a possible drying effect. And the spectral contributions corroborate this, as approximation points (blue dots) at 1400 and 1900 nm are observed. There are two differences that are observed in the spectral contributions, as component 2 retrieved horizontal details at 2400 nm, which in not present in component 3. And the opposite is true for the 1500 nm band that is present in component 3, but not in component 2. Even taking these points into consideration, it is difficult to determine what these differences mean, as no clear indicator is present within the concentration maps. However, when looking at PC4 of the IDEL-Ω results in figure 4.13a and c, the scores image (a) show the semen stain border, while the spectral bands highlighted (c), show horizontal bands at 1500 nm. A possible explanation is that the drying effect is fused with the scattering effect present at the border, this however requires further investigation.

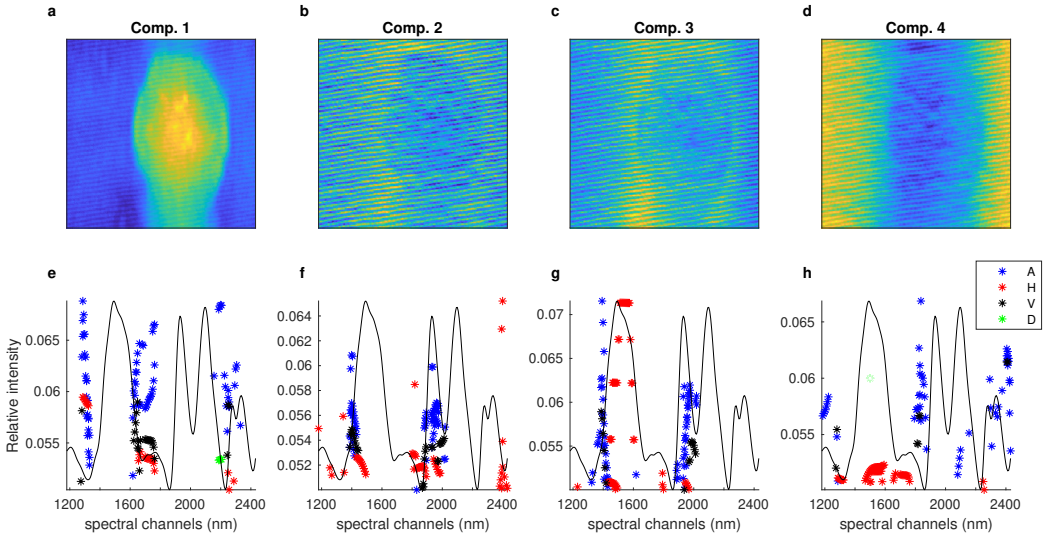


Figure 4.17: Four component MCR analysis on Ω-data, taking SIMPLISMA as initial estimates. a-d) refolded concentration maps for the four components, with e-h) their respective spectral profiles, coloured by decomposition (blue (A), red (H), black (V) and green (D)) with the mean spectrum of the data superimposed

The final component, is the most interesting of the four. The concentration map shows some shadow of the semen stain, combined with the fabric. The spectral contributions highlighted are at 1250 and 1850 nm for approximations, and 1500 and 1700 nm for horizontal details, with some more spread contributions present between 2100 and 2400 nm. Two comparisons can be made with the IDEL analysis from figure 4.13. The 1850 nm band can be attributed to the negative correlation that was observed in PC1, which makes sense, as the scores image of figure 4.13a shows an inverted image with respect to component 4 in figure 4.17. The horizontal details at 1500 and 1700 nm seem present in PC4, from figure 4.13, indicating the border of the semen stain. Most likely some contributions from PC2 are present as well, as there is clearly fabric present within component 4.

Moving on towards to the final analysis, which is the MCR-ALS analysis of the Ω-data cube, using the PC scores images from figure 4.13a as initial estimates. Figure 4.18, components 1, 2 and 3 show very similar concentration maps and spectral contributions, when compared to the scores images

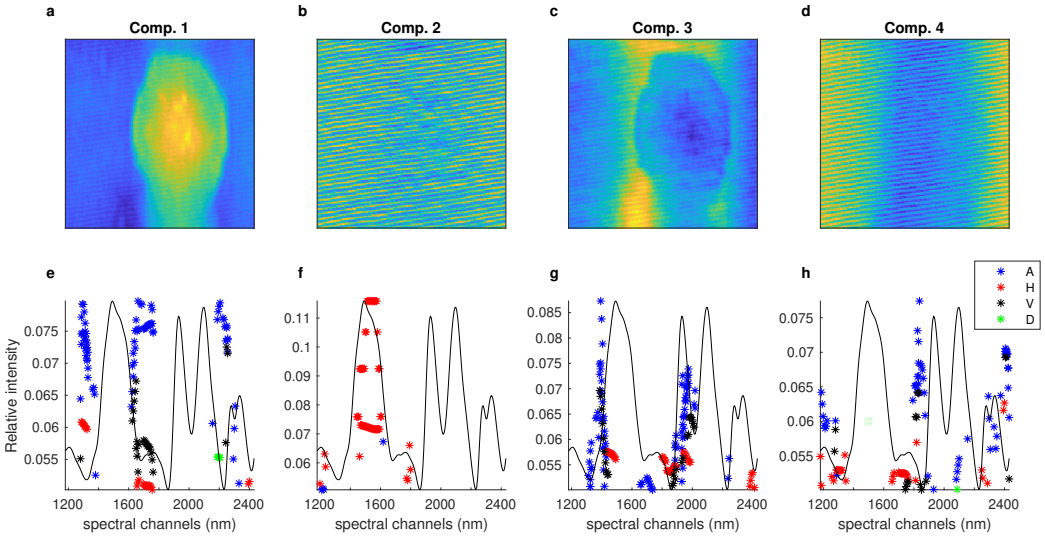


Figure 4.18: Four component MCR analysis on Ω-data, taking Ω as initial estimates. a-d) refolded concentration maps for the four components, with e-h) their respective spectral profiles, coloured by decomposition (blue (A), red (H), black (V) and green (D)) with the mean spectrum of the data superimposed

and highlighted spectral channels, respectively. The main differences are the contributions of fabric being present in components 1 and 3, most likely coming from the horizontal and vertical detail present, as is indicated by the red and black dots (figure 4.18e/g). Of which a possible explanation could be the non-negativity constraint, as that is one of the more crucial difference between PCA and MCR-ALS. Component 4 shows a similar map and spectral contribution with respect to figure 4.17, excluding the horizontal contributions at 1500 nm. This is visible in the concentration maps, as a less significant presence of the fabric is observed.

Although the results are arguably worse than the PCA analysis on IDEL, due to the lack of isolating single physical components, it does open the door to using possible constraints within the MCR-ALS algorithm, to further improve the results e.g., applying a smoothness constraint, a mask or forcing selectivity on specific components. This does add more user interaction, and is not necessarily required, depending on the data. These results also show that, even if the initial estimates isolate certain components, if the data does

not conform to the assumptions made within the specific algorithm used, the results will not significantly change, as is shown in figure 4.16. However, it is also apparent that if there is any relevant variance within the data, the initial estimate can improve the final solutions obtained.

4.7 Conclusion

The IDEL-Ω algorithm shows promise in retrieving semi-automatically spatial structures within a data set with their respective spectral contributions. The simple simulation example results in an "ideal-scenario" where different spatial structures, with different spectral contributions are present and retrieved by IDEL-Ω. The texture pack data set goes into a situation where incredibly complex spatial structures are superimposed, with near identical spectral contributions. Although the entirety of the components are not able to be retrieved, some aspects were isolated.

The stained fabric data set, sets up a situation where there are similar spectral contributions, but seemingly different spatial structures, visually speaking. IDEL-Ω is not only able to retrieve the relevant spectral/spatial combinations, but model it as well, where a new spectral image can be projected onto it and if similar spectral/spatial combinations are present in the newly acquired data, it will be able to retrieve those structures. MCR-ALS is also incorporated within the algorithm, where similar results were retrieved, opening the door to the field of curve resolution with all its benefits.

4.8 Perspectives

Notwithstanding that IDEL-Ω has provided very useful insight with respect to spatial-spectral contributions within a spectral image, the texture pack data set showed the main weak point of the algorithm. The image fusion done by PCA, although incredibly powerful, does come with its drawbacks, these being the dependency on variance, the orthogonality constraint and the bilinear nature of the results. A future work is expected with different fusion techniques, where more consistent results can be obtained, e.g. wavelet fusion [54,94] or independent component analysis [95]. This however is another field onto itself and requires a more in-depth investigation [96]. The parameters are another factor within the IDEL-Ω algorithm, which are within the context

of the thesis fixed, with adequate reasoning, however a proper validation strategy is necessary on this subject.

Chapter 5

Weighted MCR-ALS

In this chapter a proposal to improve the MCR-ALS algorithm is presented, with respect to the difficulty of retrieving the pure profile for a component in a mixture which is strongly underrepresented in the samples (a minor component). To this aim the same concept of using several peels of the convex hull, as resorted to in IDEL-Ω to select the most informative sub-images in the descriptor matrix scores plot (Chapter 4), has been implemented in order to select a subset of informative samples in order to overcome the blurring of the minor component due to the over-representativeness of the major ones. This situation can be very common in spectral imaging data, e.g. a contaminant in food or environmental specimen, or the active compound in a pharmaceutical tablet are typical examples of a very unbalanced presence in the pixels of the acquired image. MCR-ALS has been proven to be effective in many practical scenarios however, within the ALS procedure, some of the well-known limitations of least squares approaches must be considered. One of these limitations in particular is with regards to the presence of non-independent and -identically distributed (non-iid) noise. To cope with this, a weighted MCR-ALS algorithm has been developed by Wentzell et al. based on maximum likelihood projections and applied to different types of data [97,98]. Another limitation relates to the so-called "black-hole" effect as pointed out recently by Vitale and Ruckebusch [99]. This issue is connected to the leverage that some data points may have in the non-negative least squares (NNLS) calculation. In MCR, single data points that are very far from the data cloud are potentially the purest ones. However, when utilising MCR-ALS, their leverage might become too low for guaranteeing the correct solution when the number of data points is very large. Even starting from

the most favourable initial estimates (i.e., the true pure profiles), the solution will in such a situation iteratively move closer to the centre of the data cloud. This problem can prove to be detrimental for imaging, as the analysis of an image can result in the retrieval of a large number of samples (pixels). A solution to overcome this black-hole effect and improve the accuracy of the MCR-ALS output is sample selection, and an efficient way to do so is by selecting only essential samples based on a convex hull criterium [100, 101]. Examples of applications to spectral imaging data show that it is possible to recover similar or sometimes better solutions, with respect to standard MCR-ALS [102–104]. However, two issues come forward, firstly the imaging data is reduced to a incredibly small subset of samples, which has the benefit of incredibly fast analysis, however with the drawback of removing a large chunk of imaging information, which can be recovered by reconstruction. Secondly, when noise comes into play, selecting too few samples can at some point decrease the stability of the model, as the number of points to properly estimate its parameters is greatly reduced [105]. In practical situations, there is a trade-off to be found, as reducing the data down to its most essential information will increase the variance of the estimated parameters while utilising the entire data set might reduce the accuracy of said parameters, as observed in the aforementioned black-hole effect. A weighted MCR-ALS methodology is proposed to balance this trade-off. To put it in perspective, the selection of samples based on essential information can be seen as the most extreme form of weighted analysis i.e., weighting one, essential samples, and zero, others. Here we propose a weighting scheme where sample weights are determined based on their relevance towards the MCR solution. To this aim, convex peeling [106] of the data set is performed. A comparison is made between standard MCR-ALS, weighted MCR-ALS with weights encoding ESP selection and weighted MCR-ALS with weights encoding convex peeling, applied on three different data sets, two simulated and one real.

5.1 MCR - ALS

5.1.1 ALS-algorithm

MCR assumes the data to follow a bilinear model (eq. 5.1) and the ALS-algorithm solves for it by taking the pseudo-inverse (eq. 5.2) and using a non-negative least squares minimisation function (eq. 5.3) to estimate its

parameters:

$$D = CS^T + E \tag{5.1}$$

$$\hat{C} = DS(S^T S)^{-1} \tag{5.2}$$

where, with respect to spectroscopy, $D(I \times J)$ is the absorbance for the I samples and J wavelengths, $C(I \times n)$ are the concentrations of the samples for the n different chemical species, $S(J \times n)$ are the absorptivity's for the wavelengths of the different chemical species and $E(I \times J)$ is the instrumental error. The $\hat{x}$ notation specifies an estimation of x.

$$\phi = min(||y - \beta\hat{x}||^2) \; with \; x \geq 0 \tag{5.3}$$

where ϕ is the minimisation function, y the data vector, β the system parameters and x the variables. The fast non-negative least squares (FNNLS) algorithm, developed by Bro et al. [107] is used within this work.

An initial estimate is required to start the algorithm, where either C or S is set. If S_{ini} is set as an initial estimate, then the algorithm solves eq. 5.2, by least squares and retrieves $\hat{C}$. The procedure continues by solving for $\hat{S}$, by using $\hat{C}$ (eq. 5.4).

$$\hat{S} = (\hat{C}^T\hat{C})^{-1}\hat{C}^T D \tag{5.4}$$

The iterative alternating procedure is stopped when no significant change within the parameters is present, significance is usually determined by the lack of fit (lof) of the estimated parameters (eq. 5.5/5.6).

$$\hat{E} = D - \hat{C}\hat{S}^T \tag{5.5}$$

$$lof = \sqrt{\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{e}_{ij}^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}d_{ij}^2}} \tag{5.6}$$

where d_{ij} and $\hat{e}_{ij}$ are elements ij of D and $\hat{E}$, respectively. When the relative Δlof between two consecutive iterations is below a set threshold, the algorithm is stopped. This threshold is dependent on the problem and within this work it is set at 10^{-8}.

5.1.2 Visualisation

The data and solutions obtained from MCR-ALS are visualised in two ways, firstly as either concentration and spectral profiles, and secondly as points in a $n-1$ dimensional space, with n being the number of components within the system [108]. The second visualisation applies PCA on the data (eq. 5.7), and normalises the scores by dividing element-wise by the first PC (eq. 5.8):

$$D = TP^T + E \tag{5.7}$$

$$T_n = \frac{T}{t_1} \tag{5.8}$$

where T are the scores, P the loadings and t_1 the first PC scores. This method of standardisation allows for a reasonable visualisation of the MCR-problem within in a system of four or less components, assuming the system is non-negative. This is due to the limitations of visualising a subspace with more than three dimensions. Projections onto lower dimensional subspaces can be applied, however the risk of not properly presenting all components is present.

The standardisation transforms the data into a closed system and where the mixture space becomes apparent. An example data set is shown in figure 5.1, where the three components (A/B/C) span the entire ternary/binary mixture space. A simplex is formed, where the vertices (green dots) correspond to the pure components and the other points within the simplex correspond to linear combinations of those pure components. This visualisation transforms the MCR problem into a geometry problem, where the optimal solutions are on the vertices of this simplex, as those vertices can explain all other data points as linear combinations.

5.1.3 Optimal solutions

Another way of explaining the optimal solution is with respect to the FNNLS. The algorithm moves the solutions towards the minimum error (eq. 1.3) and as such, the optimal parameters will give the best estimation of the data. This can be visualised with an error surface, where the error is computed for every set of parameters, and should minimise towards the true solutions, which are the underlying sources of variance that generate the data. An example

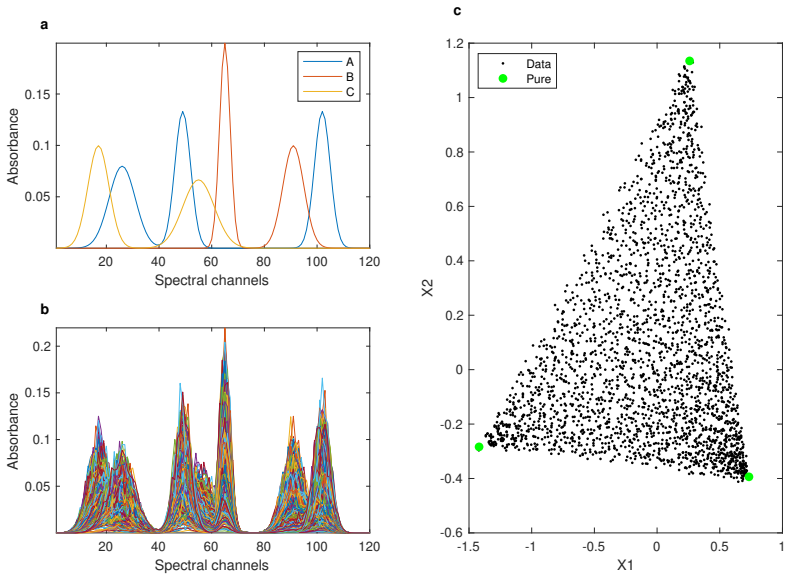


Figure 5.1: Single matrix of an example data set, a) pure spectral profiles (A-C); b) 10% of the spectra from the data matrix; c) normalised scores in the (X1, X2) PC-subspace, each black dot representing a spectrum and the three green dots representing the pure spectra of A-C.

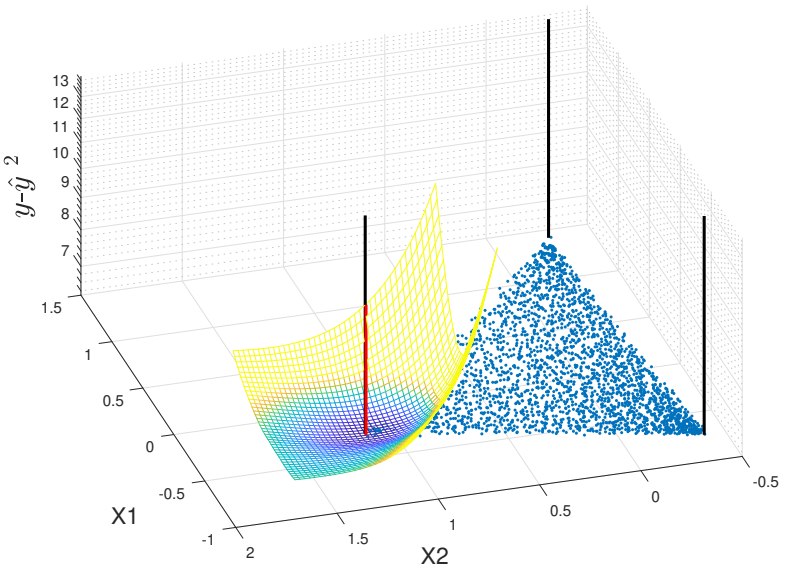


Figure 5.2: Error surface visualisation for an ideal situation in the normalised (X1,X2) PC-subspace, with scores as blue dots, scores with only A, B or C as black rods, minimum error ($(y - \hat{y})^2$) indicated by red rod.

of the error surface is in figure 5.2. As this is a three component system, the error surface is limited to fixing two components to their true solutions and varying the third across the plane, and calculating the corresponding error at that point in the plane. In the plot, the minimum, indicated by the red vertical line, overlaps almost completely with the black vertical line that corresponds to the true solution, for that component. The lack of perfect correspondence is due to the number of points used to estimate the error surface.

In standard conditions, the minimum of the error surface will correspond to the true and correct solution. However the crux of the issue is when standard conditions no longer apply, this is the case in situations where the data deviates from the ideal least squares situation. Imaging is a situation where the possibility of underrepresented components is of concern, due to the sheer number of samples retrieved by the analysis. Figure 5.3 shows a non-ideal situation, where there is a discrepancy within the mixture space. A component is disproportionally underrepresented. In this situation, there is a single pure sample present for one component, and all other samples contain either a very low or zero concentration of that component. The two

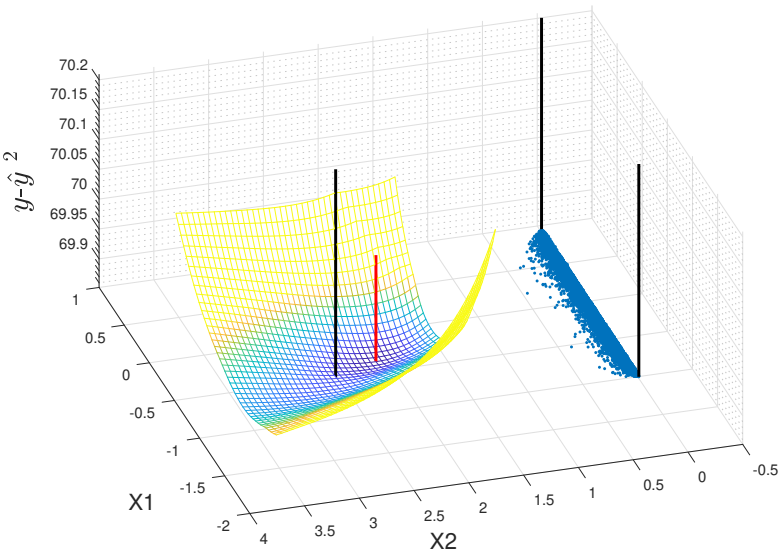


Figure 5.3: Error surface visualisation for a non-ideal situation in the normalised (X1,X2) PC-subspace, with scores as blue dots, scores with only A, B or C as black rods, minimum error ($(y - \hat{y})^2$) indicated by red rod.

other components are fixed and the error surface of the underrepresented component is calculated. It shows that the minimum is not overlapped with the true solution.

5.1.4 Essential Spectral Pixels (ESP)

One of the more impactful developments within MCR is the development of the ESP procedure [100]. It states, that taking only an essential subset of samples is required to retrieve the optimal MCR solution. This subset is the convex hull of the data cloud visualised within the normalised scores plot. This subset can explain every other sample, as a linear combination from itself. An indicator is the error surface. In figure 5.4, the error surface for the same component is calculated, utilising only the ESP subset from the non-ideal situation in figure 5.3. The minimum of the error surface is overlapped with the true solution. This means that the optimal solution in a least squares context is coherent with the true sources of variance that make up the data, this is due to the removal of samples that give a disproportionate amount of leverage to the major components.

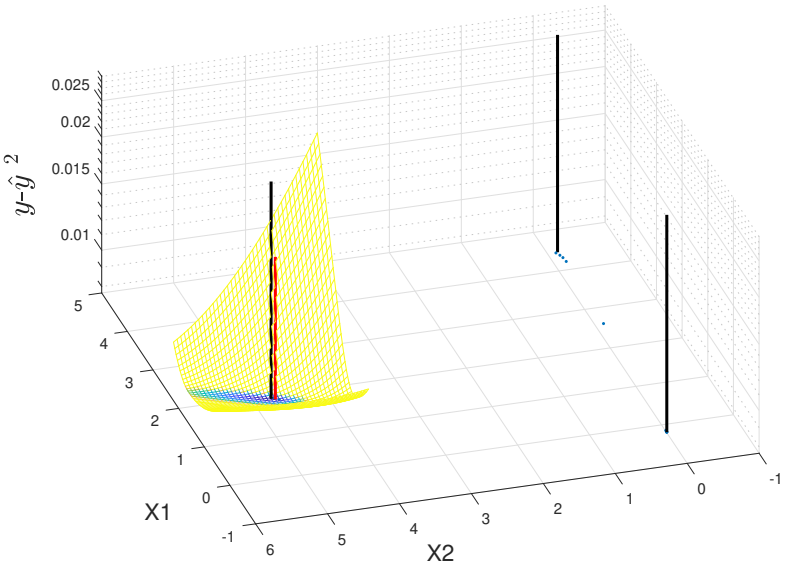


Figure 5.4: Error surface visualisation for a non-ideal situation (selecting only the ESP-subset) in the normalised (X1,X2) PC-subspace, with scores as blue dots, scores with only A, B or C as black rods, minimum error ($(y-\hat{y})^2$) indicated by red rod.

5.2 Weighted MCR-ALS

A modified ALS framework is formulated where, instead of applying a standard non-negative least squares approach to estimate the concentration and spectral profiles, a weighted version of the FNNLS algorithm, developed by Bro et al. is applied. The main differences with respect to standard MCR-ALS is found in eq. 5.9 - 5.13 below:

$$D = \hat{C}\hat{S}^T + E \tag{5.9}$$

$$\hat{C} = D\hat{S}_{ini}(\hat{S}_{ini}^T\hat{S}_{ini})^{-1} \tag{5.10}$$

$$\hat{S} = (\hat{C}^T W_c \hat{C})^{-1}\hat{C}^T W_c D \tag{5.11}$$

$$\hat{C} = D\hat{S}(\hat{S}^T\hat{S})^{-1} \tag{5.12}$$

$$\hat{S} = (\hat{C}^T W_c \hat{C})^{-1}\hat{C}^T W_c D \tag{5.13}$$

Where $W_c(I \times I)$ is the weighting matrix for C, and the spectra used as initial estimates are denoted as S_{ini}^T. W_c is a square matrix with its diagonal containing the weights of all samples, and zeros elsewhere. Eq. 5.12 and 5.13 are alternated until a satisfactory solution is obtained.

The samples are weighted according to their relevance to the MCR solution. With the basis of relevance being dictated by ESP. ESP selection can be encoded in W_c, with weights equal to one, for ESPs, and weights equal to zero, for non-ESPs. However, this is the most extreme form of weighting. We extended this approach by applying convex peeling, where each peel, l, is considered a layer of the data in the normalised scores within the PC-space. Peeling is an iterative process where the most external convex hull (first layer, $l = 1$) is removed and considering the remaining samples a new convex hull is computed (second layer, $l = 2$). The process is repeated, until there are not enough points left to continue. The remaining points, if present, are given a weight of 0. The samples belonging to each convex hull are inversely weighted with their respective peel number (weights equal to $1/l$). In W_c, the samples of the first peel (ESP) have a weight one and for the last and

inner most peel, a weight close to zero is set. The lower the relevance of the sample towards the MCR solution, the lower its weight.

5.3 Validation

Residual bootstrapping and a bias-variance analysis are used to increase the level of confidence in the results as well as highlight the drawbacks/benefits of the methods used.

5.3.1 Residual bootstrapping

Residual bootstrap analysis for regression [109] is a randomised error resampling technique to determine the stability of a model from a single data set. The bootstrap framework is shown below, with eq. 5.14 - 5.17:

The reconstructed data matrix $\hat{D}(I \times J)$ is estimated from a singular value decomposition (SVD) of D with k components.

$$\hat{D} = U\Sigma V^T \tag{5.14}$$

Where the $U(I \times k)$ and $V(J \times k)$ are the left and right singular vectors of D, respectively, and $\Sigma(k \times k)$ is a square matrix with its singular values on the diagonal and zeros elsewhere.

The error $E(I \times J)$ is calculated by subtracting $\hat{D}$ from D.

$$E = D - \hat{D} \tag{5.15}$$

E is resampled to generate a new error matrix $E_{bs}(I \times J)$, by removing a random subset (1%) of samples from E and repopulating it with another random subset of E. In this way a new error matrix is generated, following the same error distribution that is present within E.

$$E \rightarrow E_{bs} \tag{5.16}$$

This resampled error is added back to $\hat{D}$ to generate a residual-bootstrapped data matrix $D_{bs}(I \times J)$ which is further processed or analysed, in this case, by means of MCR-ALS.

$$D_{bs} = \hat{D} + E_{bs} \tag{5.17}$$

To get a proper estimation of the model stability, the bootstrap is repeated 50 times [105], to generate 50 matrices D_{bs}.

5.3.2 Relative goodness of solutions

To determine the goodness of the solutions obtained, the bias of the mean solutions with respect to the ground truth and the variance within solutions is determined. These calculations are performed per component and spectral channel, as these parameters were known within the system, as such the bias and variance between samples is determined.

Eq. 5.18) retrieves the variance and gives information on the dispersion of the solutions, where $\hat{x}$ is a retrieved solution (for a single spectral channel and component), M is the number of analyses and $\bar{x}$ is the mean of $\hat{x}$, across all analyses performed. Eq. 5.19 calculates the Bias^2, where x is the true value (at a single spectral channel and component). It retrieves the squared bias of the solution. To retrieve the total error, the two parameters are summed (eq. 5.20).

$$Variance = \frac{\sum_{m=1}^{M} (\hat{x}_m - \bar{x})^2}{M - 1} \tag{5.18}$$

$$Bias^2 = (\bar{x} - x)^2 \tag{5.19}$$

$$Total\ error = Variance + Bias^2 \tag{5.20}$$

5.4 Datasets

Three data sets are analysed, two resulting from simulations and one from a six-component Raman image of a pharmaceutical tablet.

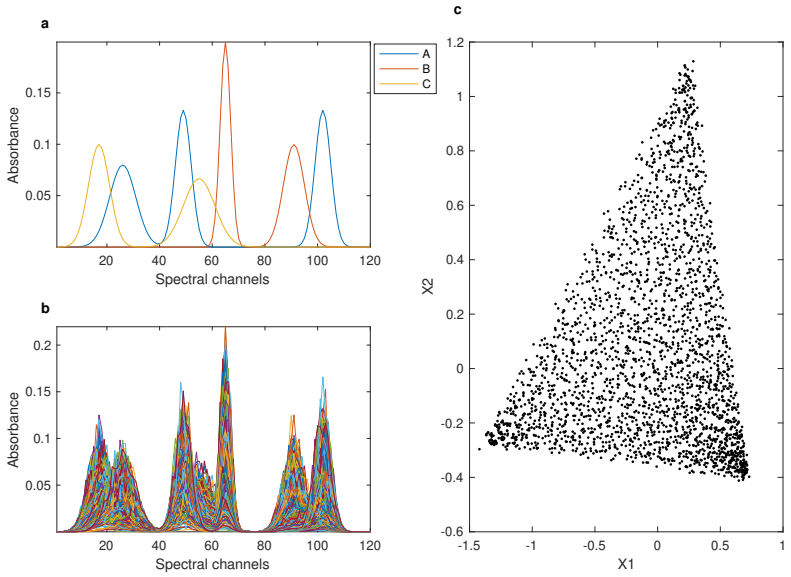


Figure 5.5: Single matrix of Data set 1, a) pure spectral profiles (A-C); b) 10% of the spectra from the data matrix; c) normalised scores in the (X1, X2) PC-subspace.

5.4.1 Data set 1

A set of three spectral profiles (figure 5.5a, 120 variables) and three concentration profiles (2595 samples) are simulated. The concentration profiles (equal for each component) span the entire mixture space, containing a set of pure, binary, and ternary samples. One pure sample per component is present and at least one spectral variable is fully selective. All three components have the exact same concentration distribution. The concentration and spectral profiles are multiplied with each other to obtain a noiseless data matrix. Afterwards, Gaussian noise (15 % of the signal intensity) is added to obtain a final data matrix (figure 5.5b). A clear triangle is observed within the normalised PC-scores plot (figure 5.5c), which indicates that every possible combination of the three components is present within the data matrix. 50 data matrices are generated, with each matrix having an error structure randomly sampled from a Gaussian distribution.

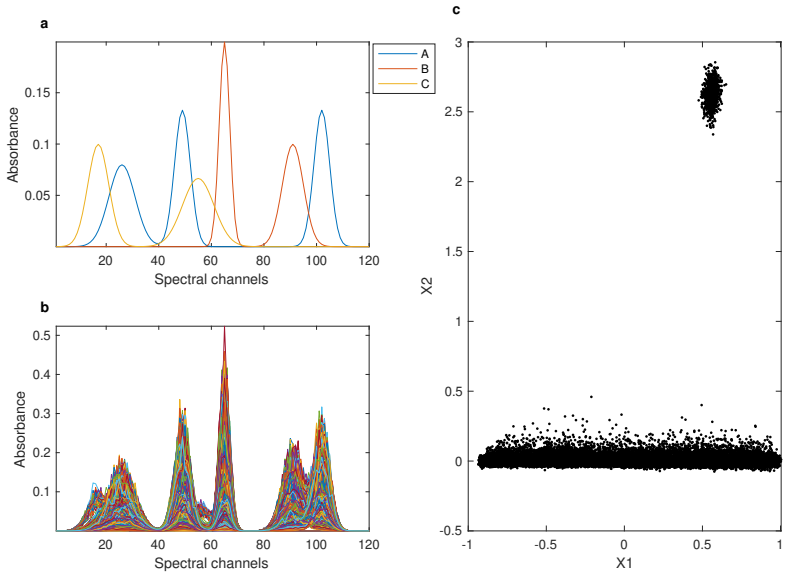


Figure 5.6: Single matrix of Data set 2, a) pure spectral profiles (A-C); b) 10% of the spectra from the data matrix; c) normalised scores in the (X1, X2) PC-subspace.

5.4.2 Data set 2

The simulated data matrix is generated as reported in Vitale et al. [99], which results in a three-component (A, B and C) system and features 56700 samples and 120 spectral-like variables (figure 5.6a). A set of 50 matrices are generated from it, by recalculating the error, but maintaining the exact same concentration and spectral profiles. The relative amount of noise is kept at 15% of the signal intensity, similarly to Data set 1. Component C is set as a minor component, meaning that a big portion of the samples contains mainly components A and B, and component C has a very low concentration across the samples. However, differently from the data matrices generated by Vitale et al., this data set contains 800 pure samples of C, this is because the noise level is three times larger. The spectra of a single data matrix are shown in figure 5.6b.

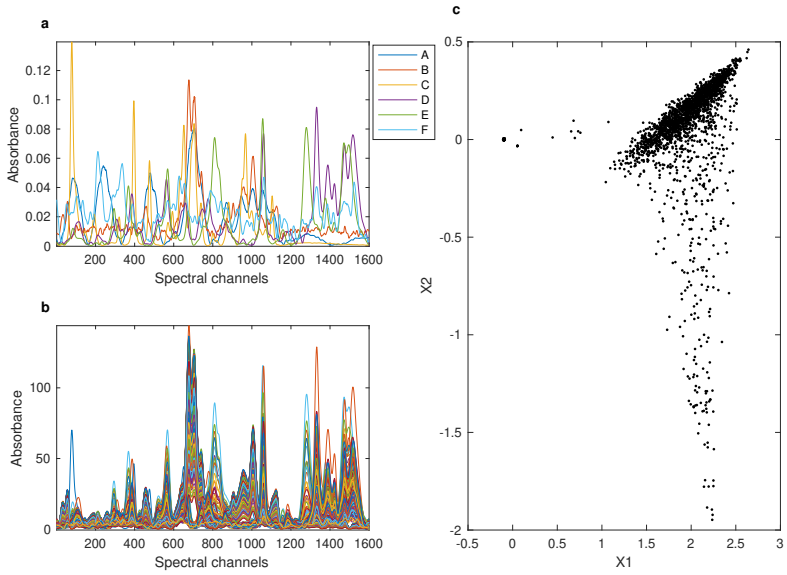


Figure 5.7: Data set 3, a) pure spectral profiles (A-F); b) 10% of the spectra from the data matrix, after pre-processing; c) normalised scores (X1, X2) in the PC-subspace.

5.4.3 Data set 3

This data set relates to a six component semi-synthetic Raman image of a pharmaceutical tablet and consists of 5000 samples and 1600 variables. We refer to Coic et al. [102] for the details on the analysis. The spectra are pre-processed with a Savitzky-Golay filter [88], using a first order polynomial and a window size of 11. The six blended chemical compounds are known, and their corresponding spectral profiles (used as a reference) are taken from an in-house database. See figure 5.7 for an overview of the data. As can be seen in the normalised scores plot (figure 5.7c), the data structure shows that minor components are present, although a higher dimensional representation would be needed for a full visualisation.

5.5 Results and Discussion

For each data set, 50 MCR solutions (every model estimated from a matrix with a different error structure) are obtained by the approaches tested:

1) MCR-ALS on the data set (results denoted as "Full" in the remainder of the text); 2) weighted MCR-ALS with weights encoding ESP selection ("ESP") and 3) weighted MCR-ALS with weights encoding the results of convex peeling ("Weighted"). The dispersion of the solutions obtained from the 50 replicates can be compared among the different approaches to determine the stability of the estimated parameters.

For Data set 1, results are provided in figure 5.8. As expected, those obtained from "Full" show that without any weighting or selection, a good (accurately representing the ground-truth for each component) and stable (no dispersion) estimation of the pure spectral profiles is obtained. The solutions obtained from "ESP" show that weighting the ESPs as one and all others as zeros clearly has an impact on the dispersion of the solutions, indicating an increased variance in the estimates of the MCR-ALS model parameters (profiles). While for "Weighted", the performance is found to be similar to "Full".

In figure 5.9 the results for Data set 2 are shown, which now highlight, differently from Data set 1, the potential impact of an under-represented minor component on the accuracy of the outcomes. The solutions obtained from "Full" show that MCR-ALS is not able to accurately estimate the parameters of component C, even though little to no dispersion is observed. With "ESP", the spectra of component C are estimated properly and point out the importance of selecting relevant samples to drive the MCR-ALS solutions towards the true one in the presence of minor components. However, this comes at the price of a higher dispersion, as already noted for Data set 1. For "Weighted", similarly to "ESP", accurate spectra are obtained, without any bias in the minor component estimate. However, in contrast to "ESP", very little dispersion in the model parameters is seen, since the full data set is used.

Figure 5.10 shows the results for Data set 3. Like in Coic et al., "Full" cannot estimate all components, minor components C and F (which explain 0.05 and 1.77% of the variance of the original data, respectively) are missed. By contrast, "ESP" and "Weighted" can retrieve solutions very close to the reference spectra, with "Weighted" showing a decrease in the dispersion of the solutions compared to "ESP". These results corroborate the ones obtained from Data set 2: a decrease in the dispersion of the parameter estimates of an order of magnitude for component F to around half for component B is observed. Only component C sees no decrease in dispersion, because "ESP"

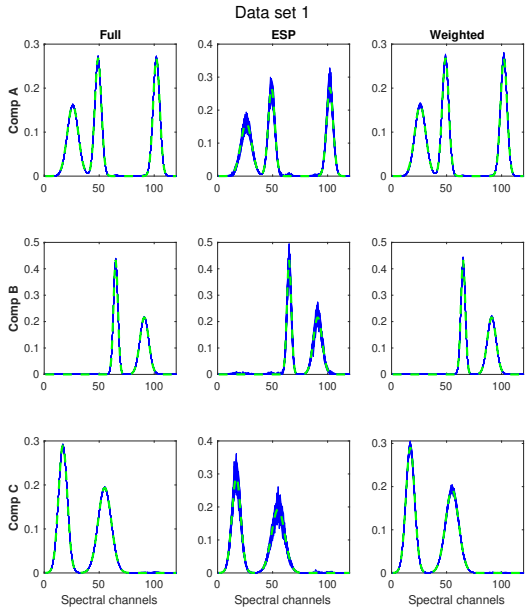


Figure 5.8: MCR-ALS solutions obtained from Data set 1, using MCR-ALS on the data set ("Full", 2595 samples), using weighted MCR-ALS with weights encoding ESP selection ("ESP", 10 samples) and using weighted MCR-ALS with weights encoding the results of convex peeling ("Weighted", 97 peels). The true solutions (dashed green) and the 50 MCR solutions (blue) are plotted separately for each component and analysis method.

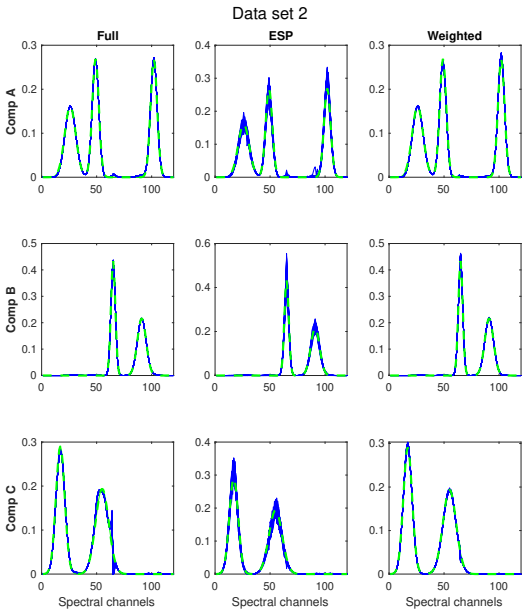


Figure 5.9: MCR-ALS solutions obtained from Data set 2, using MCR-ALS on the data set ("Full", 56700 samples), using weighted MCR-ALS with weights encoding ESP selection ("ESP", 15 samples) and using weighted MCR-ALS with weights encoding the results of convex peeling ("Weighted", 98 peels). The true solutions (dashed green) and the 50 MCR-ALS solutions (blue) are plotted separately for each component and analysis method.

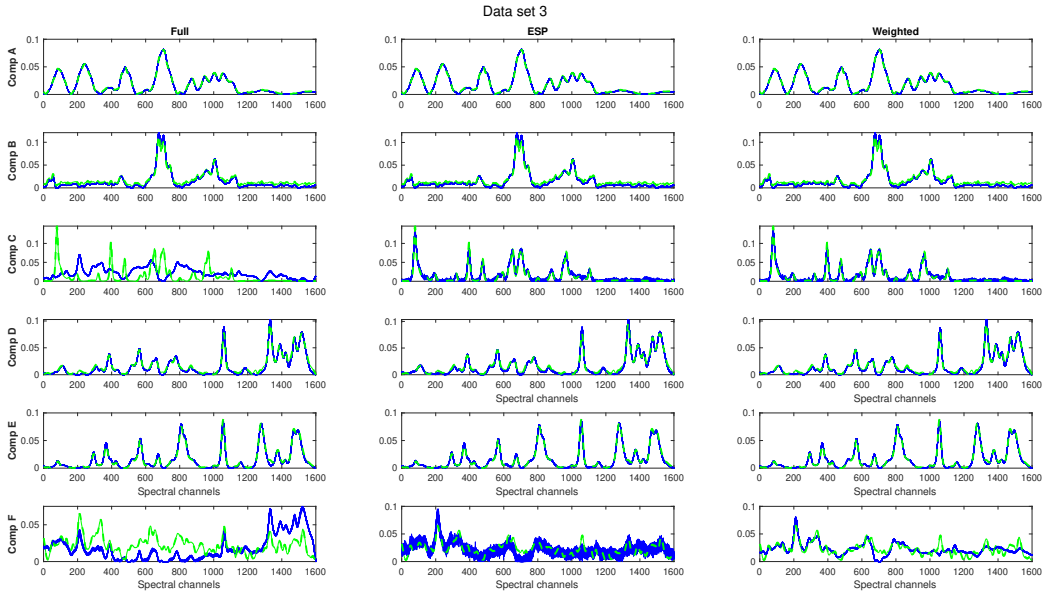


Figure 5.10: MCR-ALS solutions obtained from Data set 3, using MCR-ALS on the data set ("Full", 5000 samples), using weighted MCR-ALS with weights encoding ESP selection ("ESP", 35 samples) and using weighted MCR-ALS with weights encoding the results of convex peeling ("Weighted", 52 peels). The true solutions (dashed green) and the 50 bootstrapped MCR-ALS solutions (blue) are plotted separately for each component and analysis method.

selects all the samples containing C, meaning that using the full data with respect to ESP adds no additional information on C. Concerning component F, "ESP" still selects the purest samples, however the selected samples have a significant noise level, inducing a dispersion in the solutions. Weighting the data set with convex peeling instead of just the ESPs, increases the number of analysed samples containing component F, in turn, reducing the variance in its calculated spectral profile.

When comparing the results of "Full" and "ESP", both Data sets 2 and 3 show that, in the presence of minor components, a trade-off is present between the approaches. One should choose between precise but biased solutions with "Full" or accurate but imprecise solutions with "ESP". "Weighted", instead, takes the middle ground, where the utilisation of the full data com-

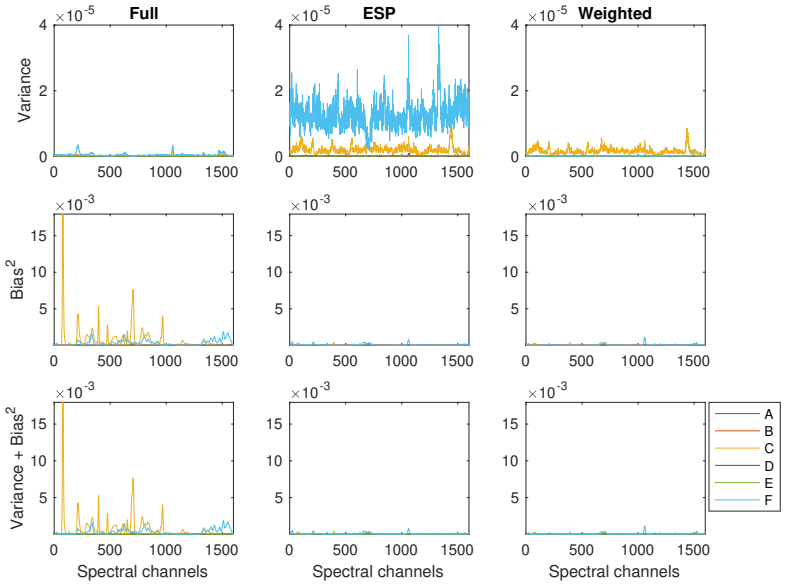


Figure 5.11: Bias/Variance results calculated from the MCR-ALS solutions obtained from Data set 3, using MCR-ALS on the data set ("Full", 5000 samples), using weighted MCR-ALS with weights encoding ESP selection ("ESP", 35 samples) and using weighted MCR-ALS with weights encoding the results of convex peeling ("Weighted", 52 peels). Components A-F coloured.

bined with the knowledge on the essential information they carry, in the presence of noise and minor components gives both more accurate and precise solutions.

To corroborate these results, a bias/variance analysis is performed, with respect to the spectral profiles (figure 5.11). A similar story is told with respect to figure 5.10, where the main concerns lie with component C and F (orange and light blue lines, respectively, in figure 5.11. Variance calculates the dispersion of the solutions where ESP clearly shows the highest values, then Weighted and Full shows the least Variance. The Bias^2 variable has, for Full, the highest values, as is expected, while ESP and Full show little to no differences. From $Variance + Bias^2$, it is clear that the Bias observed in Full has a significantly larger impact on the goodness of the solutions that the Variance that is observed in ESP. The final results remain the same, where Weighted retrieves solutions that are balanced with respect to Full and ESP.

5.6 Conclusion

This chapter highlights an issue within the MCR-ALS methodology. Spectral imaging is put into perspective and is shown that with regards to the presence of minor components ESP selection is required to drive the MCR-ALS solution towards the true one, with the caveat of losing spatial information and parameter stability. An extended weighting scheme within the weighted MCR-ALS framework is proposed that is based on convex hull data peeling and is able to preserve the benefits of ESP selection while mitigating its drawbacks. The weighting scheme utilises sample information, to resolve the leveraging issue that is present, giving weights to samples of more relevance towards the MCR-ALS solution.

5.7 Perspectives

This weighting framework is based on sample relevance, however this can be further developed by introducing spatial information, retrieved by other means e.g., from the IDEL-Ω algorithm or a priori knowledge. A different perspective can be taken as well, instead of weighting the pixels (samples), the spectral channels are weighted, which are linked to the images within a spectral image. The methodology is still in its development stage and further work is expected.

Chapter 6

Conclusion

6.1 Closing remarks

The main goal of the thesis project was to develop novel tools within the spectral imaging fields that would be able to utilise both spatial and spectral information conjointly, while simultaneously exploiting the spatial-spectral interactions to retrieve more informative results, that were previously covered. This problem was highlighted in data that exhibited strong spatial-spectral interactions, e.g. in absorbance spectroscopy of rough surfaces and reflectance spectroscopy in the visible and near-infrared range. This problem was tackled by firstly considering one perspective i.e. either spatial or spectral and then incorporate the other into the pipeline of the methodology.

The first developed tool, IDEL, was extensively discussed in the thesis, by firstly taking an exploratory route in chapter 3, and then augmenting and methodically testing the tool in chapter 4. IDEL takes an imaging perspective towards spectral image analysis where it first uses an image decomposition method to separate spatial structures based on their orientation and smoothness. It then highlights distinct features by encoding the decomposed images into vectors of descriptors. Lastly, it extracts the decomposed images that show the most distinct descriptors, and the spectral channels where the descriptors exhibit the highest variance. A set of images are obtained that are not only at specific spectral channels, but also at specific decompositions. This new dimension of sparse spectral channels and decompositions is designated as Ω and contains both spatial and spectral information. An

image fusion method was applied to recover the components that exhibit a similar variance across this dimension. The tool was applied in different scenarios with simulations, emulsions, textiles, vegetation, food, and forensics. The method showed in particular for forensic analysis unprecedented performance.

The method is not fully finalised, as an extensive validation process of real controlled data is required to obtain a better grasp on the spatial-spectral interactions. As well as streamlining the tool to be more easily implemented within the spectral imaging field, by automatising the parameters as well as incorporating more robust guidelines with respect to the algorithm.

The second tool developed was Weighted MCR-ALS, which was discussed in Chapter 5. It took a spectral perspective to spectral image analysis, utilising the MCR-ALS methodology. The method was augmented with a weighting procedure that takes pixel relevance into consideration. An issue was presented where the amount of data analysed with imaging becomes detrimental towards the MCR method, due to the ALS-optimisation procedure. The issue is resolved by weighting the pixels with a parameter, that dictates the relevance a pixel has towards the MCR-ALS solution. Two simulated and one real data sets were analysed and showed the benefit of the weighted procedure with respect to the current methodologies.

The method is still in its infancy and can be further developed by extending the approach towards different weighting procedures, as well as optimising the methodology to be easily implemented within the spectral imaging field.

6.2 Future developments

Some aspects were not fully realised and are for IDEL-Ω the different applications, in which the algorithm could be applied. Some future work in the conservation sciences can be considered, as spectral imaging is an incredibly dominant tool within the field, due to its powerful non-invasive analyses [20,28]. The complex chemical compositions of e.g. paints combined with the structured spatial information that is painted for which IDEL-Ω could prove to be an incredible boon. A similar story is made for the remote sensing

field where the structured landscapes can be explored with the algorithm.

New angles that can be approached, to further develop the IDEL-Ω algorithm: 1) Automatising the parameters within the pipeline; 2) Generalising the framework of IDEL to expand the reach of the methodology; 3) Setting up an experimental framework to analyse and validate the recovered spatial-spectral interactions in various scenarios; 4) Development of a stand-alone toolbox that incorporates a generalised framework to analyse spectral imaging data and exploit spatial-spectral interactions.

The Weighted MCR-ALS algorithm is equally important as it presents a doorway to perhaps optimising the immensely common ALS-algorithm in more applications. With an investigation into trace-analysis being promising, where small amounts of a subset of components is present within a data set but difficult to isolate and extract. As well as high resolution applications within spectral imaging, as increasingly more data is obtained, obscuring or revealing minor components.

Further development with respect to the Weighted MCR-ALS methodology: 1) Automatising the balance between the amount of data used and goodness of the solutions obtained; 2) Incorporating other forms of spatial information; 3) Development of a stand-alone toolbox to easily implement/automate the analysis of spectral imaging data.

Appendix A

Preprocessing of stained fabrics

In figure A.1, the subtracted baseline from the preprocessing of stained fabrics (Chapter 4) is presented. A non-standard preprocesssing step is taken with the stained fabrics data set, to better contrast the spatial dimensions within the spectral images. A superficial comparison is presented in figure A.2. The semen on white cotton fabric data set is preprocessed with the indicated methods, i.e. weighted least squares baseline correction (WLS, A.2A), standard normal variate (SNV, A.2B), Multiplicative scatter correction (MSC, A.2C) and lastly taking the second derivative of the data (A.2D). In addition, a small analysis is done on the effect of using MSC on spectral imaging data applying the method globally vs locally (figure A.3).

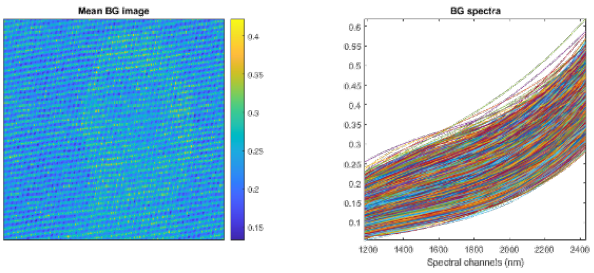


Figure A.1: Subtracted baseline by using WLS on SW, mean image (left) and per pixel (right).

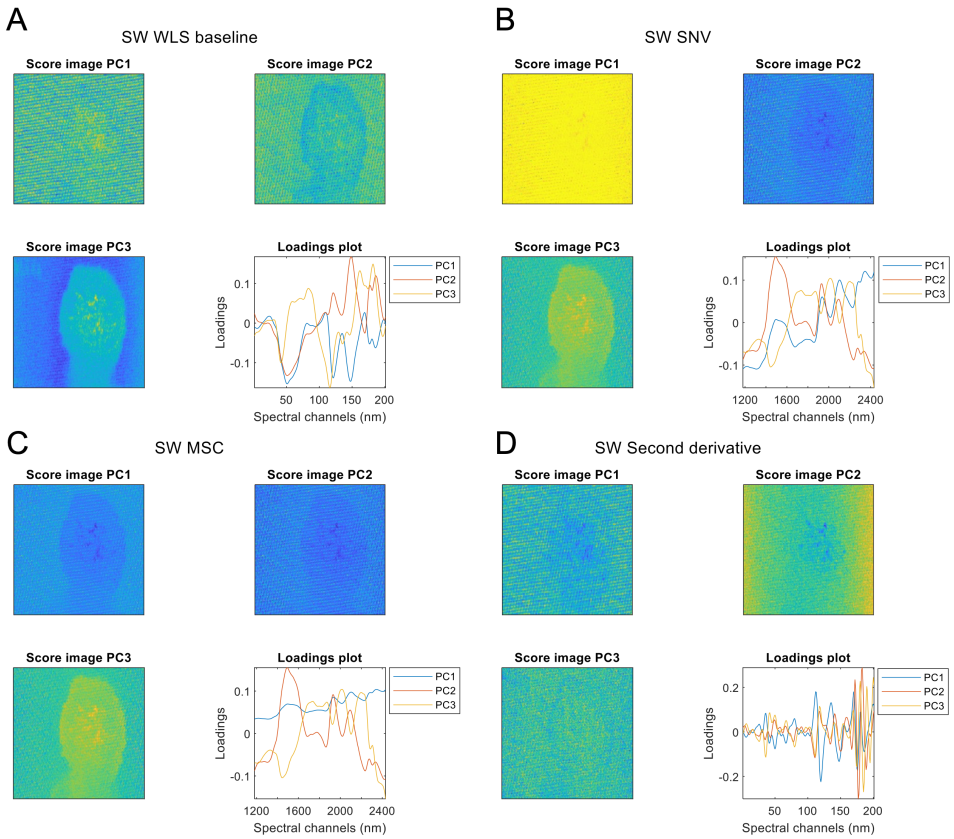


Figure A.2: Different preprocessing methods reviewed, with an unfolded PCA analysis on white fabric with data data with A) WLS; B) SNV; C) MSC; D) Second derivative.

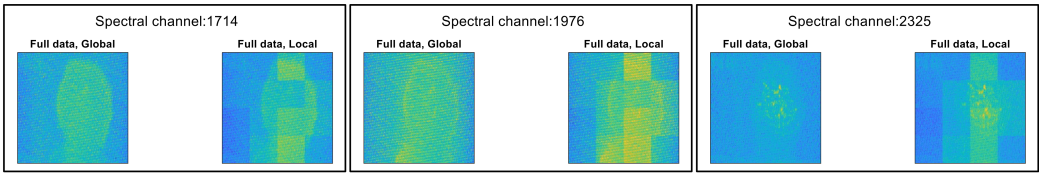


Figure A.3: MSC applied on white fabric with semen data, visualising three spectral channels. The method is either applied globally (using the entire image) or locally (applied on windows of the image, separately).

Appendix B

K-means algorithm

B.1 Augmented algorithm

K-means [110] is a grouping algorithm applied on matrices that determines the groups with the smallest within group distances. When the number of groups is specified a prior, an equal number of nodes are randomly positioned in the data space. The points that are closest to the specified nodes are grouped. Every node is recalculated by taking the mean of the grouped points. The algorithm iteratively calculates the nodes, until no change is observed.

An issue present within the algorithm is the random initialisation of the nodes. The standard k-means algorithm is augmented, by running the algorithm 5000 times and recovering the iteration that best represents the average grouping. The augmented algorithm will not be expanded here due to the scope of the thesis and will be presented in a future work.

B.2 Exclusion of methods

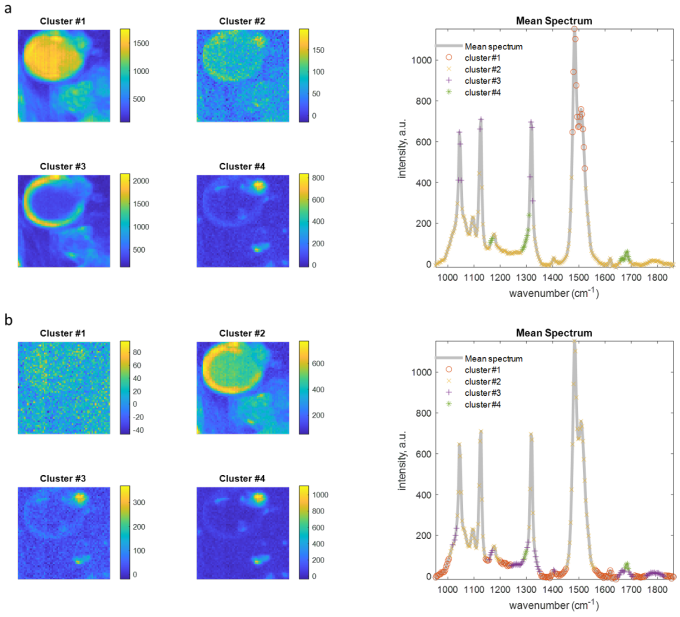


Figure B.1: Results obtained by clustering the loadings of PCA on a) the original spectral image; b) the descriptors matrix of the GLCM, excluding the wavelet decomposition step.

Appendix C

Expanded results - IDEL

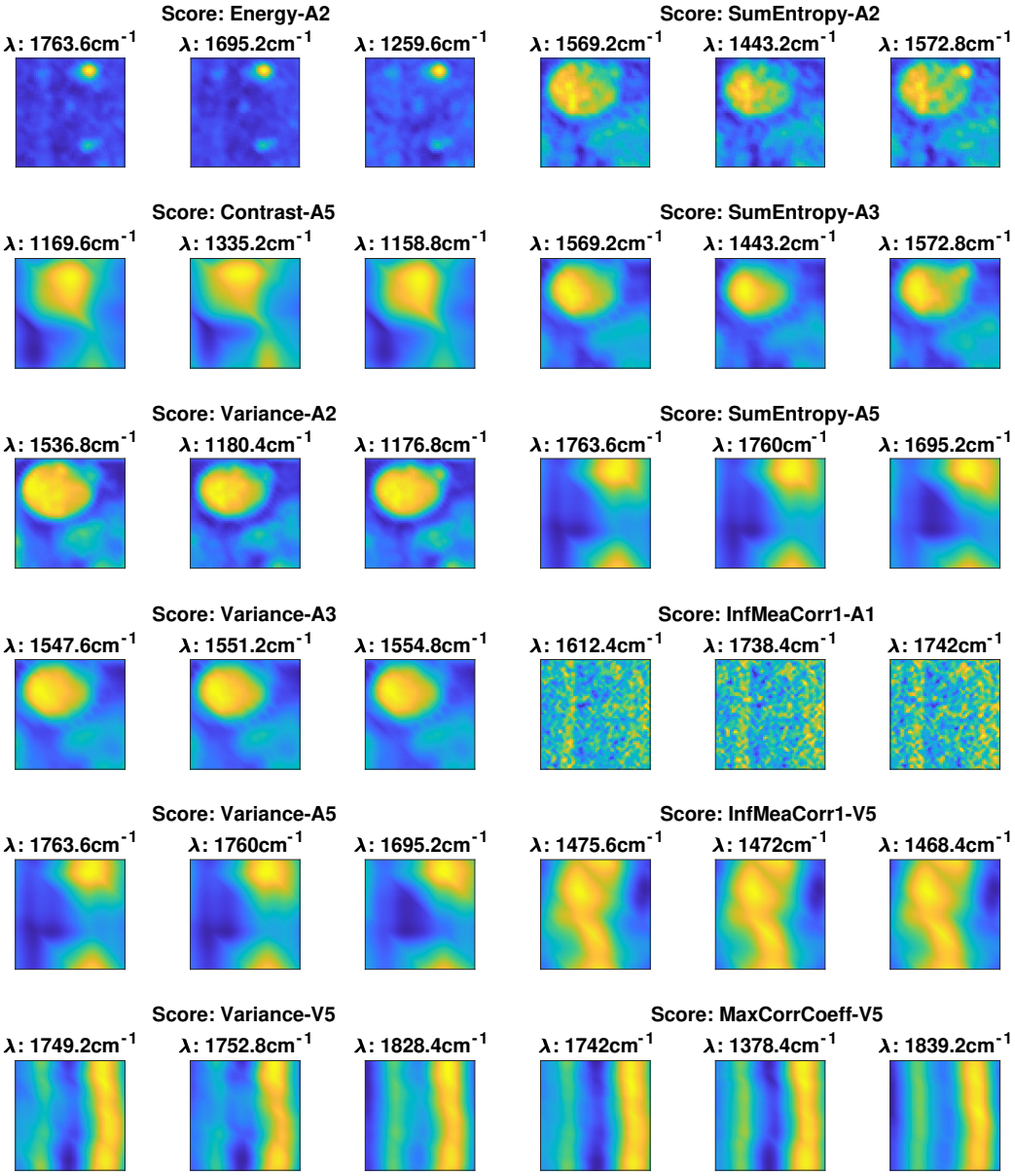


Figure C.1: Expanded results obtained from applying IDEL on the oil in water emulsion data set. The twelve highest scores on the PC1/PC2 scores plot are selected with the three highest loadings values (< 9 degrees correlation).

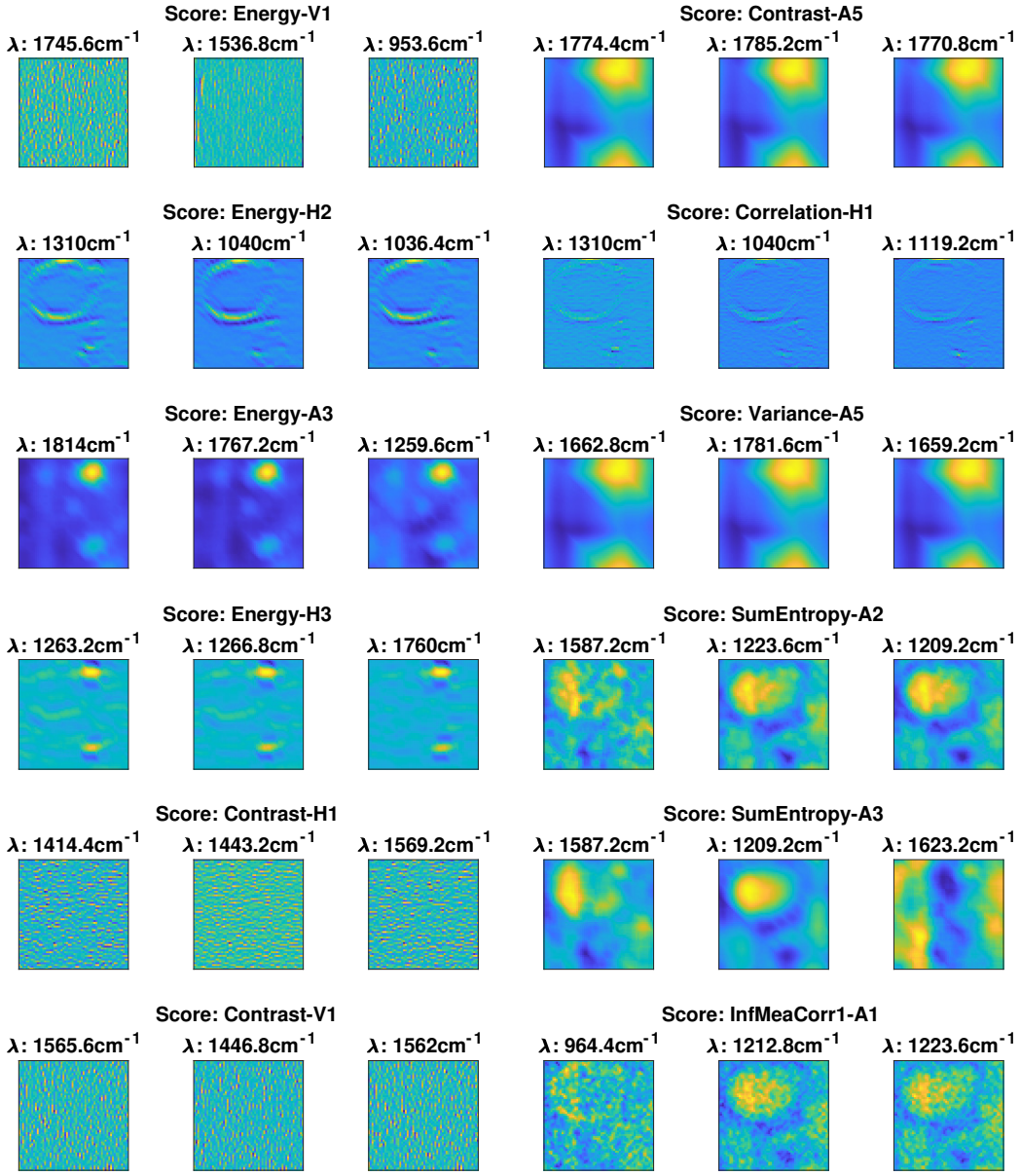


Figure C.2: Expanded results obtained from applying IDEL on the oil in water emulsion data set. The twelve highest scores on the PC2/PC3 scores plot are selected with the three highest loadings values (< 9 degrees correlation).

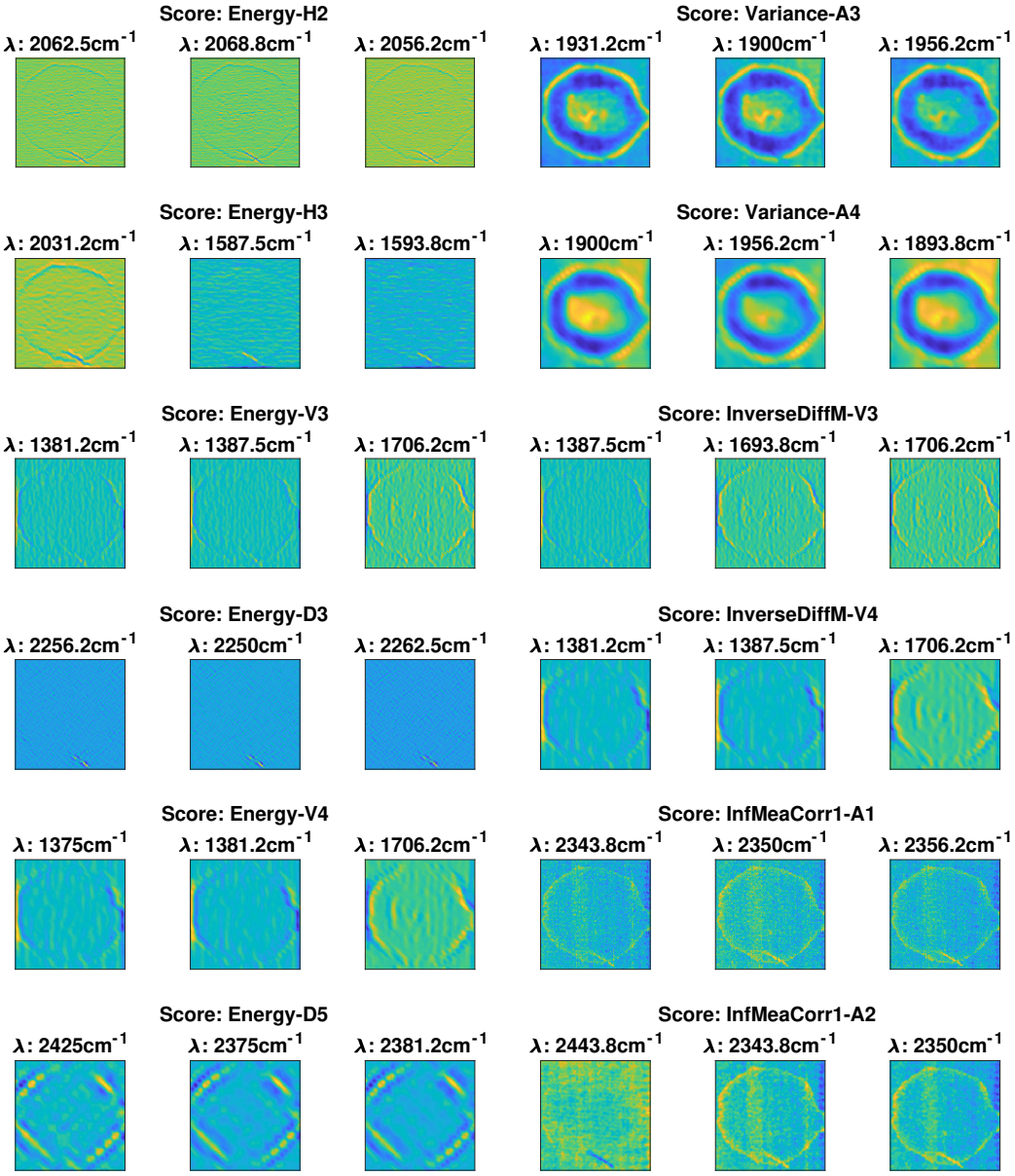


Figure C.3: Expanded results obtained from applying IDEL on the semen data set. The twelve highest scores on the PC1/PC2 scores plot are selected with the three highest loadings values (< 9 degrees correlation).

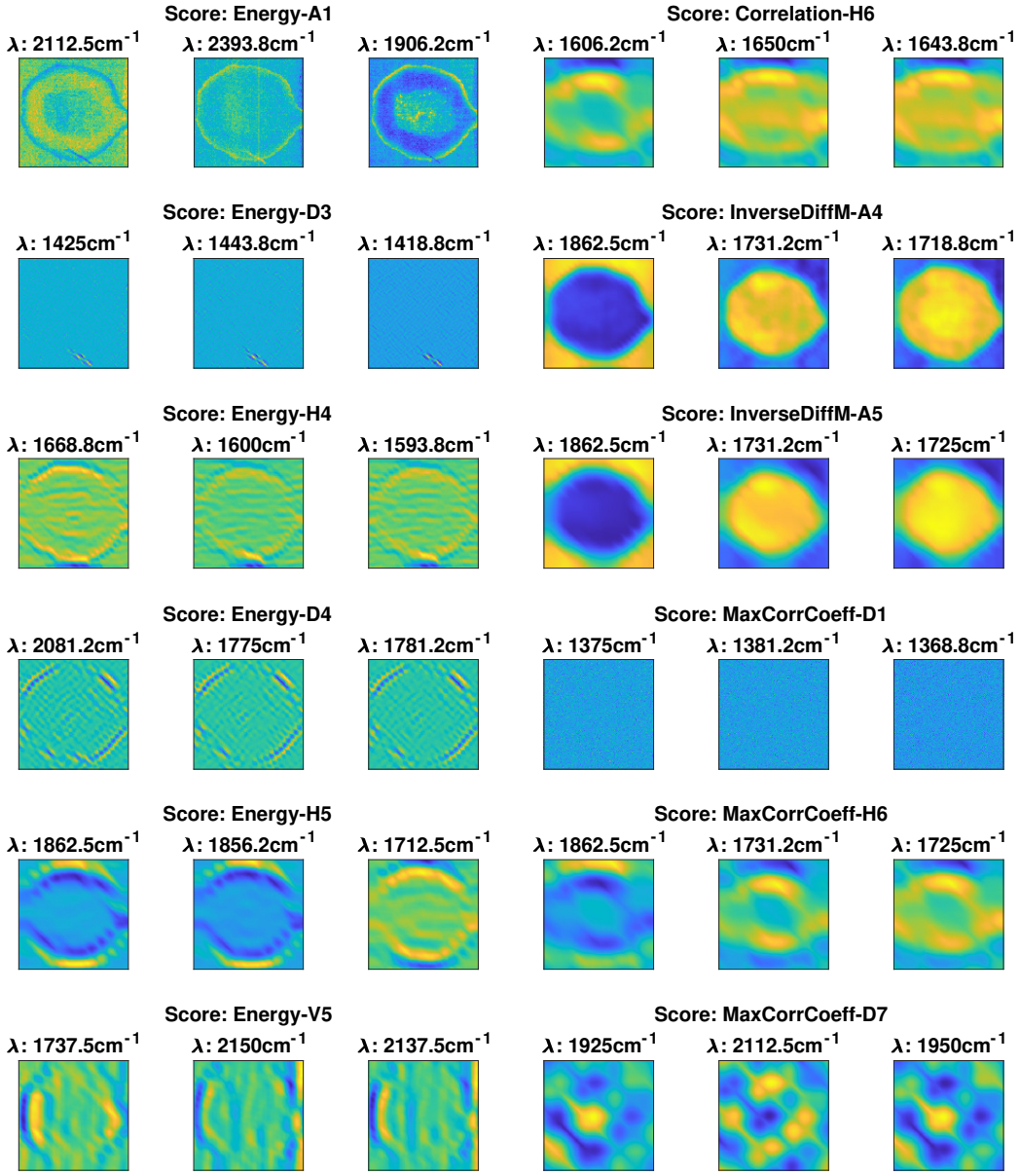


Figure C.4: Expanded results obtained from applying IDEL on the semen data set. The twelve highest scores on the PC3/PC4 scores plot are selected with the three highest loadings values (< 9 degrees correlation).

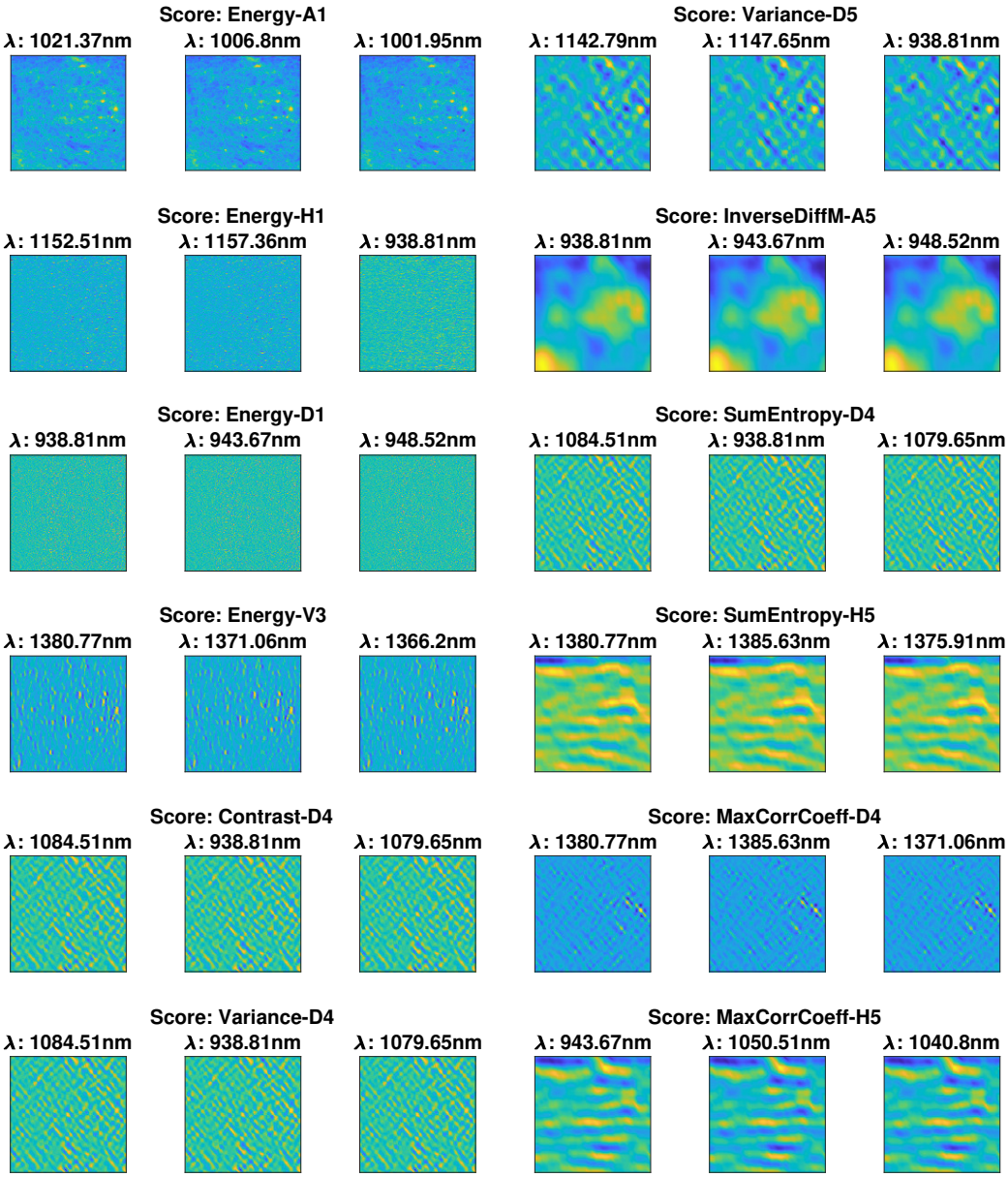


Figure C.5: Expanded results obtained from applying IDEL on the bread data set. The twelve highest scores on the PC1/PC2 scores plot are selected with the three highest loadings values (< 9 degrees correlation).

Appendix D

Expanded results - stained fabrics

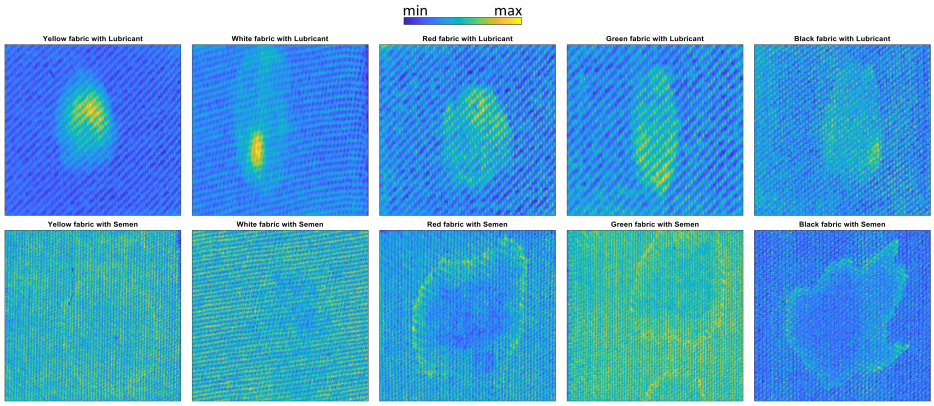


Figure D.1: Mean images (across the spectral dimension) of either lubricant (top) or semen (bottom) stains, on coloured fabrics from left to right, yellow, white, red, green and black.

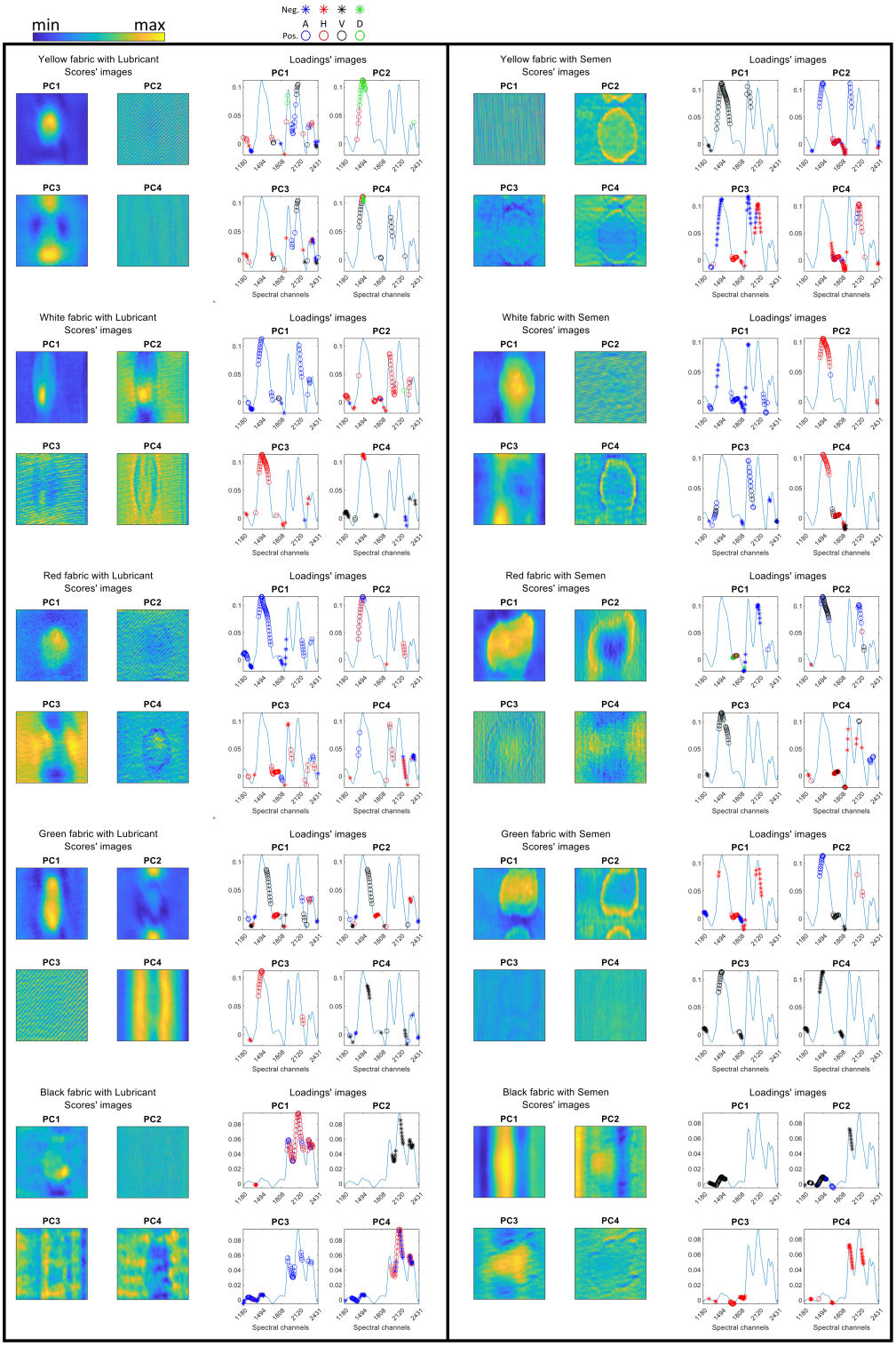


Figure D.2: Outcomes of the analysis of the ten images of coloured cotton fabrics with either lubricant (left) or semen (right) stains. The first four scores images and selected spectral channels on the mean spectrum are shown.

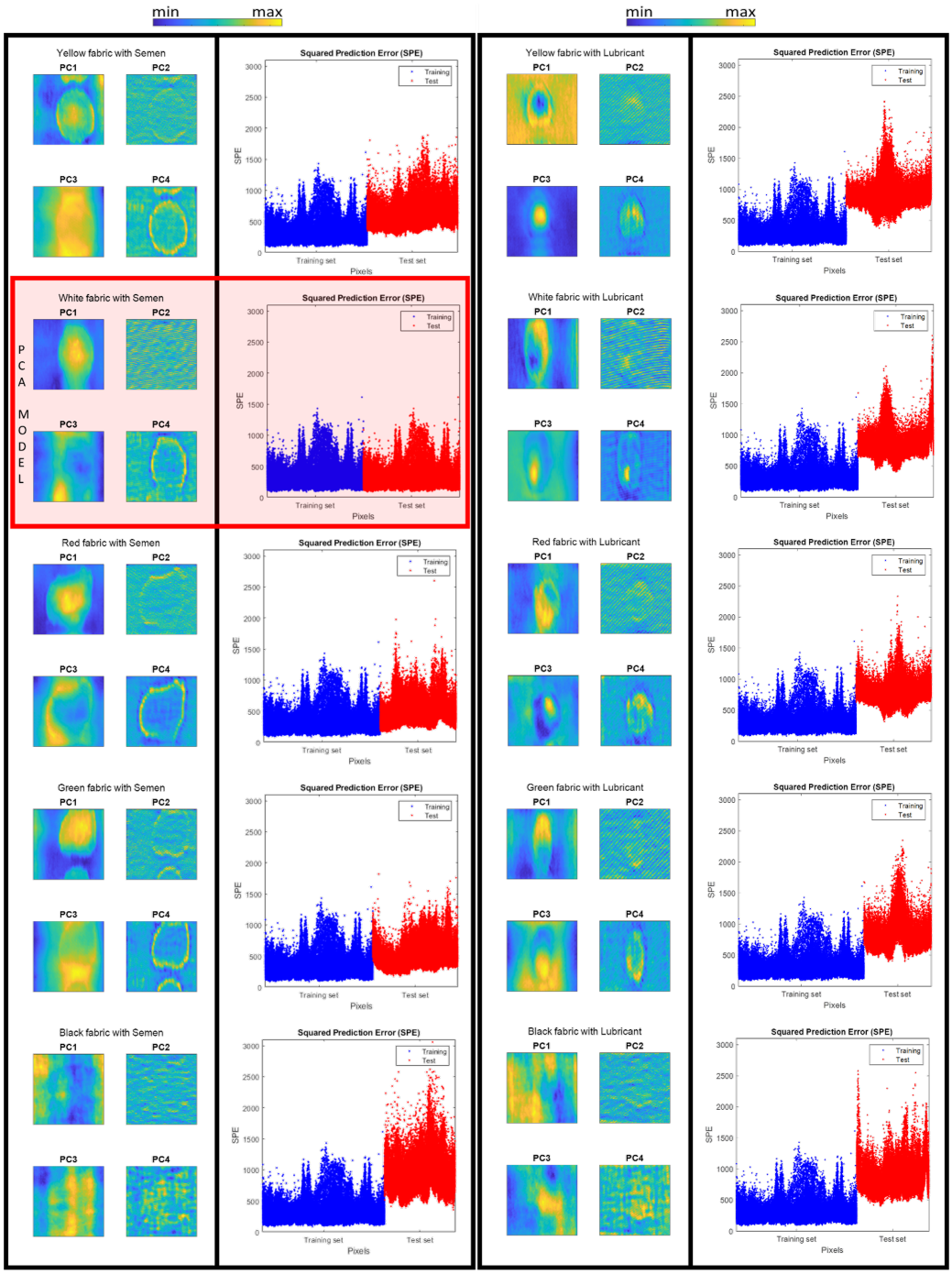


Figure D.3: Outcomes of the projection of the ten images of coloured cotton fabrics with either lubricant (left) or semen (right) stains on the model generated by applying IDEL-Ω on white fabric with semen. The first four scores images and squared prediction error (SPE) of the pixels are shown.

Appendix E

Publications

Paper I

Mohamad Ahmad, Raffaele Vitale, Marina Cocchi and
Cyril Ruckebusch

Weighted multivariate curve resolution – alternating least squares based on sample relevance

published

DOI: 10.1002/cem.3478

Received: 21 November 2022 | Revised: 6 February 2023 | Accepted: 20 March 2023

DOI: 10.1002/cem.3478

SHORT COMMUNICATION

Journal of CHEMOMETRICS WILEY

Weighted multivariate curve resolution—Alternating least squares based on sample relevance

Mohamad Ahmad[1,2] | **Raffaele Vitale**[1] | **Marina Cocchi**[2] | **Cyril Ruckebusch**[1]

[1]Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, Univ. Lille, CNRS, LASIRE (UMR 8516), Lille, France

[2]Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, Modena, Italy

Correspondence
Mohamad Ahmad, Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, Univ. Lille, CNRS, LASIRE (UMR 8516), Lille, France and Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, Modena, Italy.
Email: m.ahmad@live.nl

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-21-CE29-0007

Abstract

Alternating least squares within the multivariate curve resolution framework has seen a lot of practical applications and shows their distinction with their relatively simple and flexible implementation. However, the limitations of least squares should be carefully considered when deviating from the standard assumed data structure. Within this work, we highlight the effects of noise in the presence of minor components, and we propose a novel weighting scheme within the weighted multivariate curve-resolution-alternating least squares framework to resolve it. Two simulated and one Raman imaging case are investigated by comparing the novel methodology against standard multivariate curve resolution-alternating least squares and essential spectral pixel selection. A trade-off is observed between current methods, whereas the novel weighting scheme demonstrates a balance where the benefits of the previous two methods are retained.

KEYWORDS

essential information, multivariate curve resolution–alternating least squares (MCR-ALS), sample selection, spectral pixels, weighted least squares

1 | INTRODUCTION

Multivariate curve resolution (MCR) is a methodology with its fingers in many fields.[1] Its ability to resolve unknown mixtures, combined with simple-to-understand algorithms and interpretable results, makes it a highly performant method. MCR is aimed at resolving the most linearly dissimilar sources of variances (which are assumed to be the purest) underlying bilinear data. One of the most utilised MCR algorithms is multivariate curve resolution–alternating least squares (MCR-ALS). MCR-ALS has been proven to be effective in many practical scenarios; however, within the ALS procedure, some of the well-known limitations of least squares approaches must be considered. One of these limitations, in particular, regards the presence of non-independent and -identically distributed (non-iid) noise. To cope with this, a weighted MCR-ALS algorithm has been developed by Wentzell et al. based on maximum likelihood projections and applied to different types of data.[2,3] Another limitation relates to the so-called 'black-hole' effect, as pointed out recently by Vitale et al.[4] This issue is connected to the leverage that some data points may have in the non-negative least squares (NNLS) calculation. In MCR, single data points that are very far from the data cloud are potentially the

purest ones. However, when utilising MCR-ALS, their leverage might become too low to guarantee the correct solution when the number of data points for other components is very large. Even starting from the most favourable initial estimates (i.e., the true pure profiles), the solution will, in such a situation, iteratively move closer to the centre of the data cloud. A solution to overcome this black-hole effect and improve the accuracy of the MCR-ALS output is sample selection, and an efficient way to do so is by selecting essential samples based on a convex hull criterium.[5–7] Examples of applications to hyperspectral imaging data showed that it is possible to recover similar or sometimes better solutions with respect to standard MCR-ALS.[8–10] However, when noise comes into play, selecting too few samples can, at some point, decrease the stability of the model, as the number of points to properly estimate its parameters is greatly reduced.[11] In practical situations, there is a trade-off to be found, as reducing the data down to its most essential information will increase the variance of the estimated parameters, whereas utilising the entire data set might reduce the accuracy of said parameters, as observed in the aforementioned black-hole effect.

In this short communication, we propose a weighted MCR-ALS methodology to balance this trade-off. To put it in perspective, the selection of samples based on essential information can be seen as the most extreme form of weighted analysis, that is, weighting one essential samples, and zero, others. Here we propose a weighting scheme where sample weights are determined based on their relevance to the MCR solution. To this aim, convex peeling[12] of the data set is performed, that is, repeated convex hull calculations, pruning the data layer by layer until no points remain. For each layer, samples receive the same weight, with weights decreasing for the consecutive layers. A comparison is made between standard MCR-ALS, weighted MCR-ALS with weights encoding ESP selection and weighted MCR-ALS with weights encoding convex peeling, applied to three different data sets, two simulated and one real.

2 | METHODS

2.1 | Weighted MCR-ALS

For the sake of brevity, MCR-ALS will not be detailed, and we refer to de Juan et al.[1] In this work, a modified ALS framework is formulated where instead of applying a standard NNLS approach to estimate the concentration and spectral profiles, a weighted version of the fast NNLS algorithm developed by Bro et al.[13] is applied. The main differences with respect to standard MCR-ALS are presented in Equations (1)–(5):

$$\boldsymbol{D} = \widehat{\boldsymbol{C}}\widehat{\boldsymbol{S}}^T + \boldsymbol{E} \tag{1}$$

$$\text{Weighted MCR} - \text{ALS}$$

$$\widehat{\boldsymbol{C}} = \boldsymbol{D}\boldsymbol{S}_{ini}\left(\boldsymbol{S}_{ini}^T\boldsymbol{S}_{ini}\right)^{-1} \tag{2}$$

$$\downarrow$$

$$\widehat{\boldsymbol{S}}^T = \left(\widehat{\boldsymbol{C}}^T\boldsymbol{W}_c\widehat{\boldsymbol{C}}\right)^{-1}\widehat{\boldsymbol{C}}^T\boldsymbol{W}_c\boldsymbol{D} \tag{3}$$

$$\downarrow$$

$$\widehat{\boldsymbol{C}} = \boldsymbol{D}\widehat{\boldsymbol{S}}\left(\widehat{\boldsymbol{S}}^T\widehat{\boldsymbol{S}}\right)^{-1} \tag{4}$$

$$\downarrow\uparrow \text{ (alternating)}$$

$$\widehat{\boldsymbol{S}}^T = \left(\widehat{\boldsymbol{C}}^T\boldsymbol{W}_c\widehat{\boldsymbol{C}}\right)^{-1}\widehat{\boldsymbol{C}}^T\boldsymbol{W}_c\boldsymbol{D} \tag{5}$$

where $\boldsymbol{D}$ $(I \times J)$ is the data matrix with I samples and J spectral channels, $\widehat{\boldsymbol{C}}$ $(I \times m)$ contains the estimated concentration profiles for m components, $\widehat{\boldsymbol{S}}$ $(J \times m)$ carries the estimated spectral profiles, $\boldsymbol{E}$ $(I \times J)$ is the model error matrix, $\boldsymbol{W}_c(I \times I)$ the weighting matrix for $\widehat{\boldsymbol{C}}$, and the spectra used as initial estimates are denoted as $\boldsymbol{S}_{ini}^T$. $\boldsymbol{W}_c$ is a square matrix with its diagonal containing the weights of all samples and zeros elsewhere.

2.2 | Weighting scheme

The samples are weighted according to their relevance to the MCR solution. In this work, we first used essential spectral points (ESPs) as the basis for the weighting scheme. ESPs can be found by taking the points along the convex hull of the normalised scores of the data set within its principal component subspace (PC-space).[5–7] ESPs carry all the spectral information required for an accurate MCR resolution, and their selection reduces the data set considerably without losing any essential information.[5–7] ESP selection can be encoded in $\boldsymbol{W}_c$, with weights equal to one, for ESPs, and weights equal to zero, for non-ESPs. However, this is the most extreme form of weighting.

We extended this approach by applying convex peeling,[12] where each peel, l, is considered a layer of the data in the normalised scores within the PC-space. Peeling is an iterative process where the most external convex hull (first layer, $l = 1$) is removed, and considering the remaining samples, a new convex hull is computed (second layer, $l = 2$). The process is repeated until there are not enough points left to continue. The remaining points, if present, are given a weight of 0. The samples belonging to each convex hull are inversely weighted with their respective peel number (weights equal to $1/l$). In $\boldsymbol{W}_c$, the samples of the first peel (ESP) have a weight of 1, and for the last and innermost peel, a weight close to zero is set. The lower the relevance of the sample towards the MCR solution, the lower its weight.

2.3 | Residual bootstrap analysis

Residual bootstrap analysis for regression[14] is a randomised error resampling technique to determine the stability of a model from a single data set ($\boldsymbol{D}$). The bootstrap framework is shown below, with Equation (6)–(9):

The reconstructed data matrix $\widehat{\boldsymbol{D}}$ $(I \times J)$ is estimated from a singular value decomposition (SVD)[15] of $\boldsymbol{D}$ with k components.

$$\widehat{\boldsymbol{D}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{6}$$

where the $\boldsymbol{U}$ $(I \times k)$ and $\boldsymbol{V}$ $(J \times k)$ are the left and right singular vectors of $\boldsymbol{D}$, respectively, and $\boldsymbol{\Sigma}$ $(k \times k)$ is a square matrix with its singular values on the diagonal and zeros elsewhere.

The error $\boldsymbol{E}$ $(I \times J)$ is calculated by subtracting $\widehat{\boldsymbol{D}}$ from $\boldsymbol{D}$.

$$\boldsymbol{E} = \boldsymbol{D} - \widehat{\boldsymbol{D}} \tag{7}$$

$\boldsymbol{E}$ is resampled to generate a new error matrix $\boldsymbol{E}_{bs}$ $(I \times J)$ by removing a random subset (1%) of samples from $\boldsymbol{E}$ and repopulating it with another random subset of $\boldsymbol{E}$. In this way, a new error matrix is generated, following the same error distribution that is present within $\boldsymbol{E}$.

$$\boldsymbol{E} \rightarrow \boldsymbol{E}_{bs} \tag{8}$$

This resampled error is added back to $\widehat{\boldsymbol{D}}$ to generate a residual-bootstrapped data matrix $\boldsymbol{D}_{bs}$ $(I \times J)$ which is further processed or analysed, in this case, by means of MCR-ALS.

$$\boldsymbol{D}_{bs} = \widehat{\boldsymbol{D}} + \boldsymbol{E}_{bs} \tag{9}$$

To get a proper estimation of the model stability, the bootstrap is repeated 50 times,[11] to generate 50 matrices $\boldsymbol{D}_{bs}$.

3 | DATA SETS

Three data sets are analysed, two resulting from simulations and one from a six-component Raman hyperspectral image of a pharmaceutical tablet.

3.1 | Data set 1

A set of three spectral profiles (Figure 1A, 120 variables) and three concentration profiles (2595 samples) are simulated. The concentration profiles (equal for each component) span the entire mixture space, containing a set of pure, binary and ternary samples. One pure sample per component is present, and at least one spectral variable is fully selective. All three components have the exact same concentration distribution. The concentration and spectral profiles are multiplied by each other to obtain a noiseless data matrix. Afterwards, Gaussian noise (15% of the signal intensity) is added to obtain a final data matrix (Figure 1B). A clear triangle is observed within the normalised PC-scores plot (Figure 1C), which indicates that every possible combination of the three components is present within the data matrix. Fifty data matrices are generated, with each matrix having an error structure randomly sampled from a Gaussian distribution.

3.2 | Data set 2

The simulated data matrix is generated as reported in Vitale et al.,[4] which results in a three-component (A, B and C) system and features 56,700 samples and 120 spectral-like variables (Figure 2A). A set of 50 matrices are generated from it by recalculating the error but maintaining the exact same concentration and spectral profiles. The relative amount of noise is kept at 15% of the signal intensity, similar to Data set 1. Component C is set as a minor component, meaning that a big portion of the samples contains mainly components A and B, and component C has a very low concentration across the samples. However, different from the data matrices generated by Vitale et al., this data set contains 800 pure samples of C; this is because the noise level is three times larger. The spectra of a single data matrix are shown in Figure 2B.

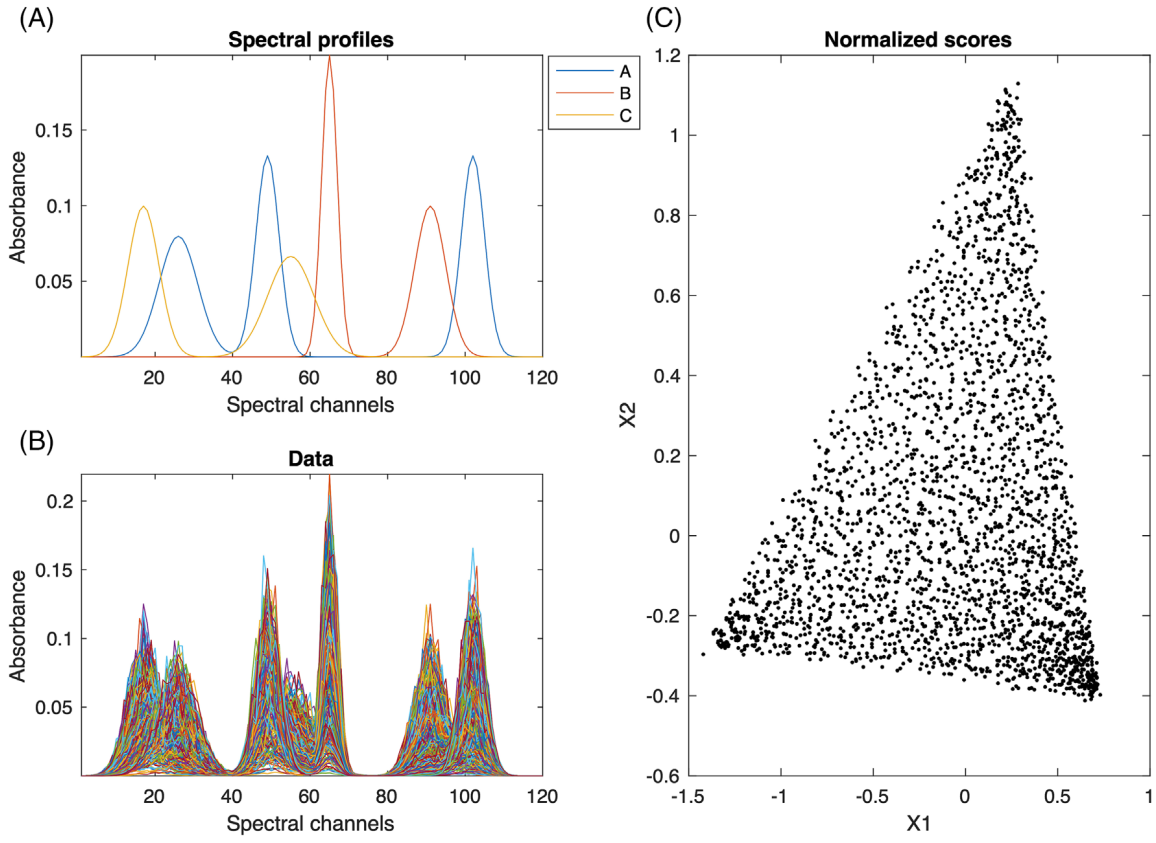


FIGURE 1 Single matrix of Data set 1, (A) pure spectral profiles (A-C); (B) 10% of the spectra from the data matrix; (C) normalised scores in the (X1, X2) PC-subspace.

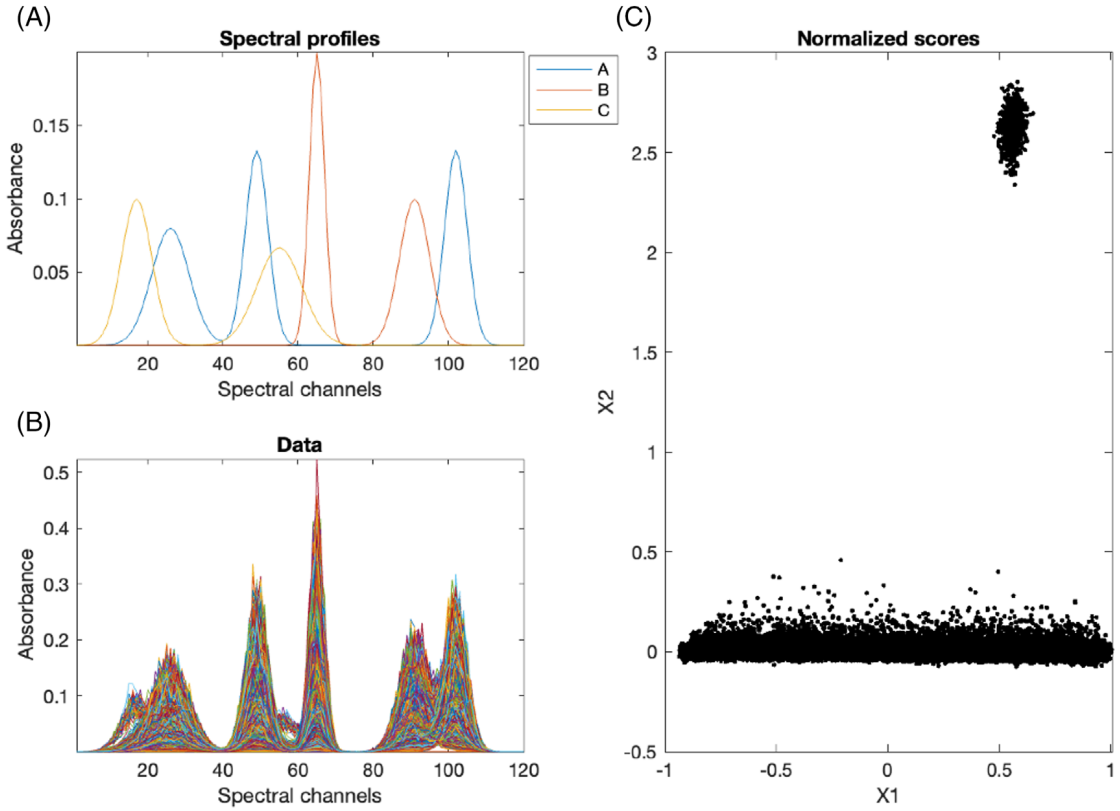


FIGURE 2 Single matrix of Data set 2, (A) pure spectral profiles (A-C); (B) 10% of the spectra from the data matrix; (C) normalised scores in the (X1, X2) PC-subspace.

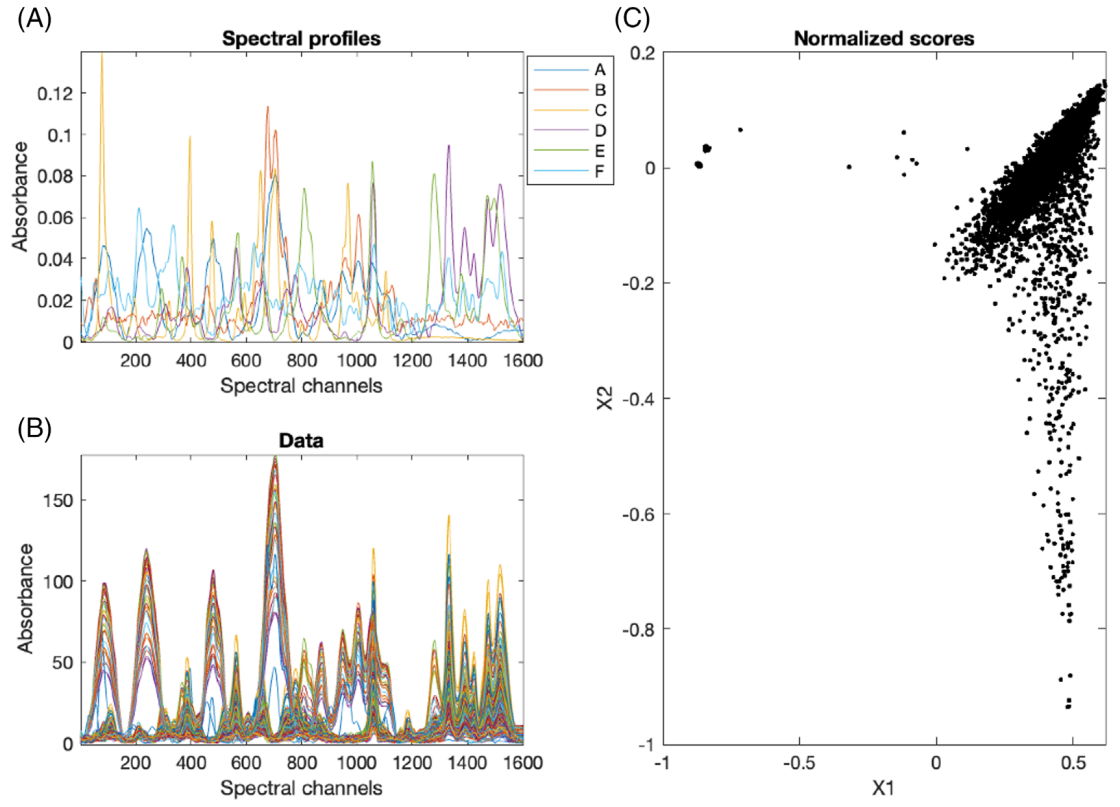


FIGURE 3 Data set 3, (A) pure spectral profiles (A-F); (B) 10% of the spectra from the data matrix, after preprocessing; (C) normalised scores (X1, X2) in the PC-subspace.

3.3 | Data set 3

This data set relates to a six-component semi-synthetic Raman image of a pharmaceutical tablet and consists of 5000 samples and 1600 variables. We refer to Coic et al.[8] for the details of the analysis. The spectra are pre-processed with a Savitzky–Golay filter[16] using a first-order polynomial and a window size of 11. The six blended chemical compounds are known, and their corresponding spectral profiles (used as a reference) are taken from an in-house database. See Figure 3 for an overview of the data. As can be seen in the normalised scores plot (Figure 3C), the data structure shows that minor components are present, although a higher dimensional representation would be needed for full visualisation.

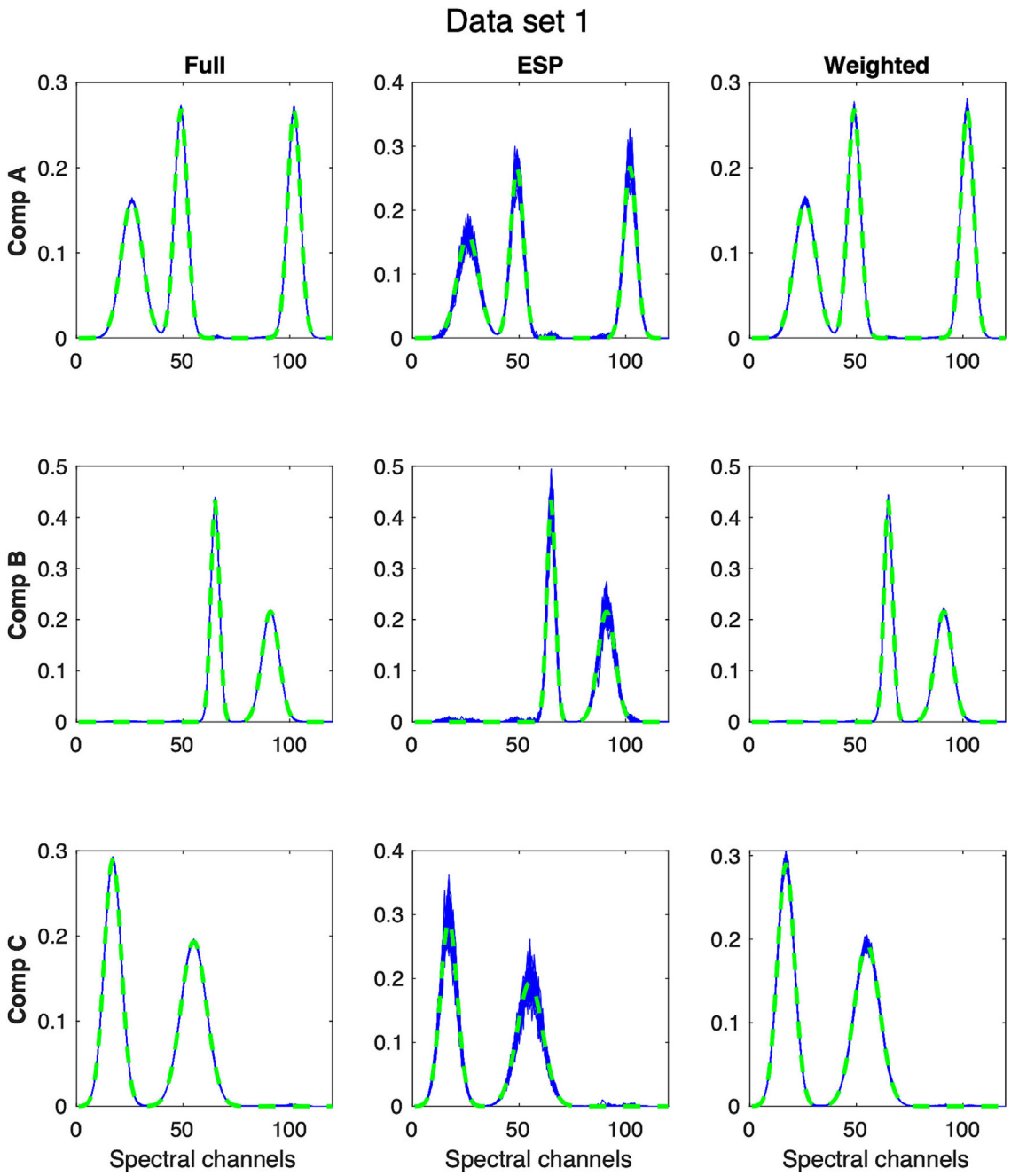


FIGURE 4 Multivariate curve resolution–alternating least squares (MCR-ALS) solutions obtained from Data set 1, using MCR-ALS on the data set ('Full', 2595 samples), using weighted MCR-ALS with weights encoding essential spectral point (ESP) selection ('ESP', 10 samples) and using weighted MCR-ALS with weights encoding the results of convex peeling ('Weighted', 97 peels). The true solutions (dashed green) and the 50 MCR solutions (blue) are plotted separately for each component and analysis method.

A bootstrap analysis is performed on the data to obtain 50 bootstrapped matrices, as described in Section 2.3 with $k = 10$.

4 | RESULTS AND DISCUSSION

For each data set, 50 MCR solutions (every model estimated from a matrix with a different error structure) are obtained by the approaches tested: (1) MCR-ALS on the data set (results denoted as 'Full' in the remainder of the text); (2) weighted MCR-ALS with weights encoding ESP selection ('ESP') and (3) weighted MCR-ALS with weights encoding

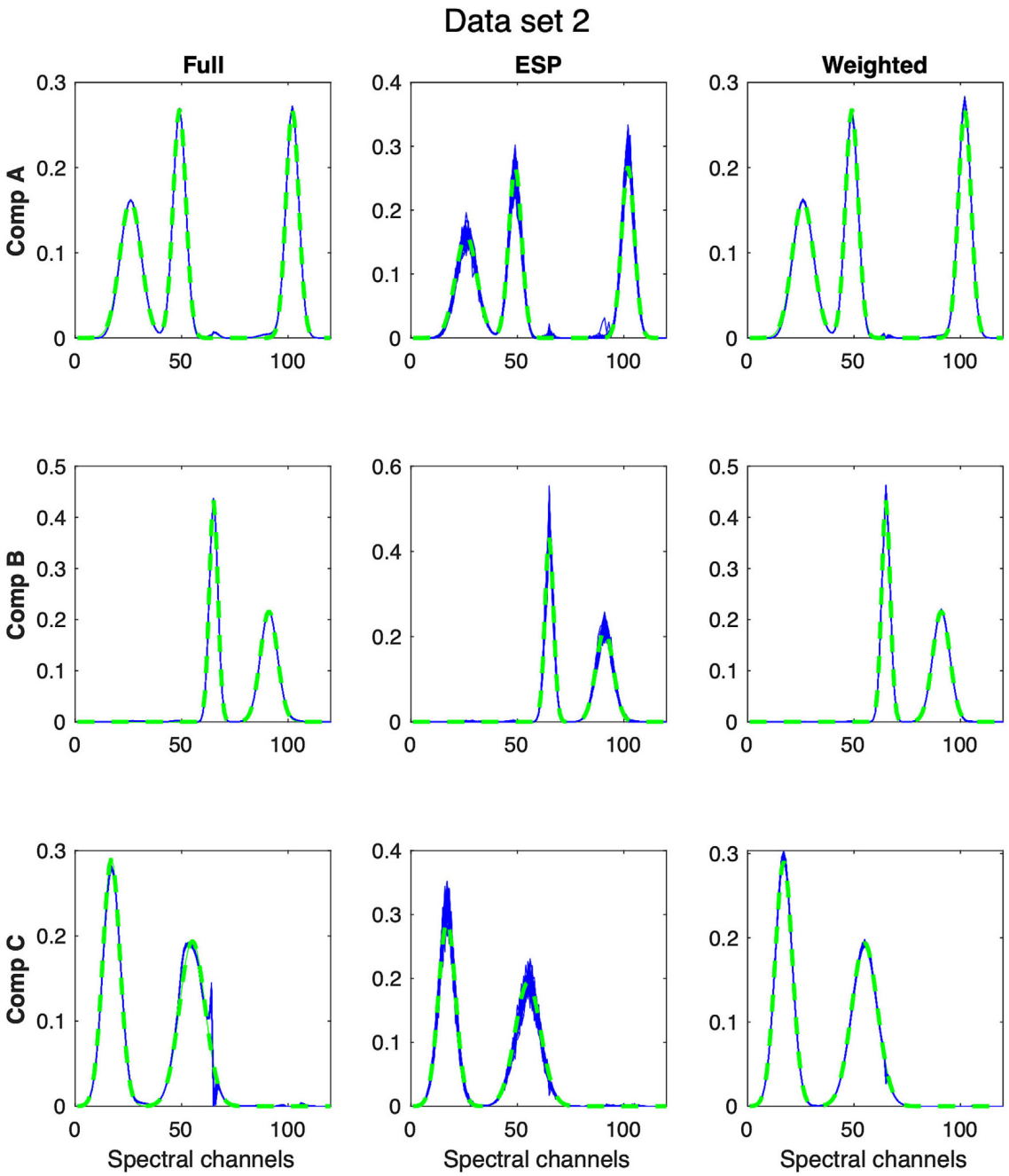


FIGURE 5 Multivariate curve resolution–alternating least squares (MCR-ALS) solutions obtained from Data set 2, using MCR-ALS on the data set ('Full', 56,700 samples), using weighted MCR-ALS with weights encoding essential spectral point (ESP) selection ('ESP', 15 samples) and using weighted MCR-ALS with weights encoding the results of convex peeling ('Weighted', 98 peels). The true solutions (dashed green) and the 50 MCR-ALS solutions (blue) are plotted separately for each component and analysis method.

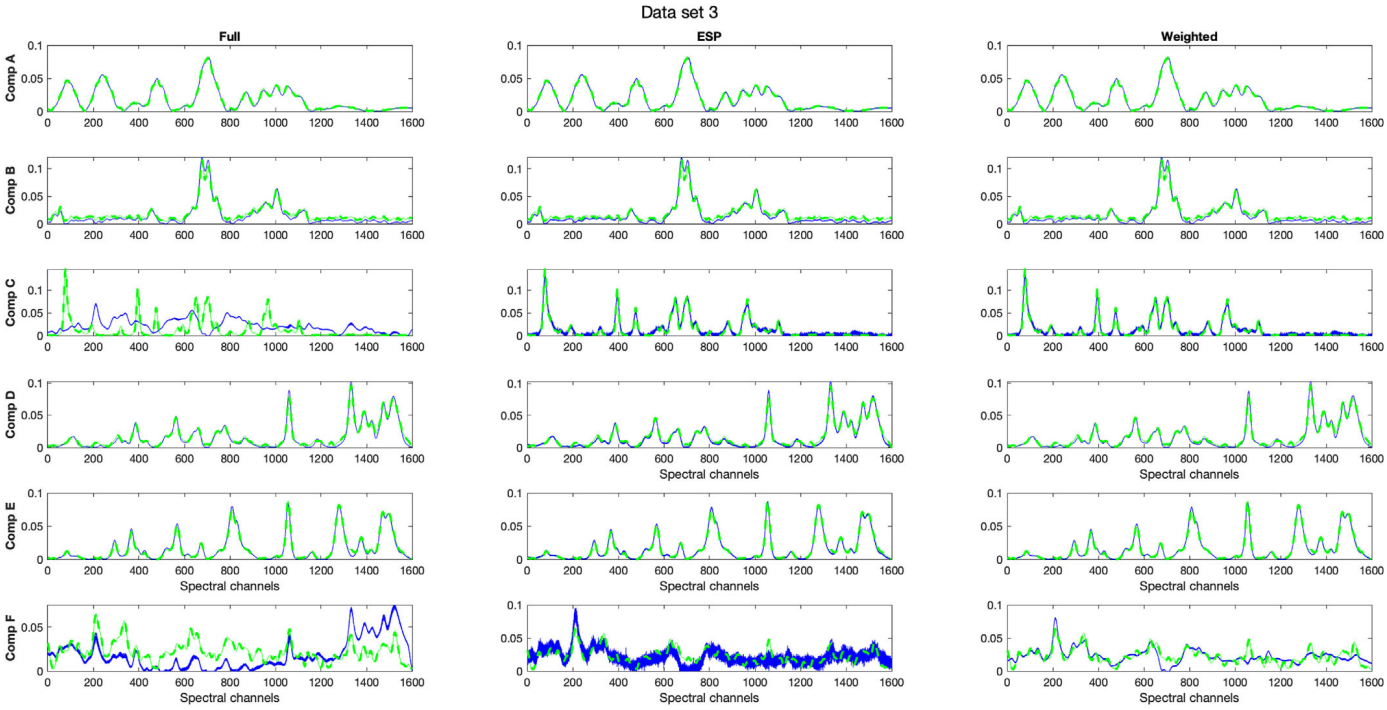


FIGURE 6 Multivariate curve resolution–alternating least squares (MCR-ALS) solutions obtained from Data set 3, using MCR-ALS on the data set ('Full', 5000 samples), using weighted MCR-ALS with weights encoding essential spectral point (ESP) selection ('ESP', 35 samples) and using weighted MCR-ALS with weights encoding the results of convex peeling ('Weighted', 52 peels). The true solutions (dashed green) and the 50 bootstrapped MCR-ALS solutions (blue) are plotted separately for each component and analysis method.

the results of convex peeling ('Weighted'). The dispersion of the solutions obtained from the 50 replicates can be compared among the different approaches to determine the stability of the estimated parameters.

For Data set 1, results are provided in Figure 4. As expected, those obtained from 'Full' show that without any weighting or selection, a good (accurately representing the ground truth for each component) and stable (no dispersion) estimation of the pure spectral profiles is obtained. The solutions obtained from 'ESP' show that weighting the ESPs as one and all others as zeros clearly has an impact on the dispersion of the solutions, indicating an increased variance in the estimates of the MCR-ALS model parameters (profiles). Whereas for 'Weighted', the performance is found to be similar to 'Full'.

In Figure 5, the results for Data set 2 are shown, which now highlight, differently from Data set 1, the potential impact of an under-represented minor component on the accuracy of the outcomes.[4] The solutions obtained from 'Full' show that MCR-ALS is not able to accurately estimate the parameters of component C, even though little to no dispersion is observed. With 'ESP', the spectra of component C are estimated properly and point out the importance of selecting relevant samples to drive the MCR-ALS solutions towards the true one in the presence of minor components. However, this comes at the price of higher dispersion, as already noted in Data set 1. For 'Weighted', similarly to 'ESP', accurate spectra are obtained without any bias in the minor component estimate. However, in contrast to 'ESP', very little dispersion in the model parameters is seen because the full data set is used.

Figure 6 shows the results for Data set 3. Like in Coic et al.,[8] 'Full' cannot estimate all components, minor components C and F (which explain 0.05% and 1.77% of the variance of the original data, respectively) are missed. By contrast, 'ESP' and 'Weighted' can retrieve solutions very close to the reference spectra, with 'Weighted' showing a decrease in the dispersion of the solutions compared with 'ESP'. These results corroborate the ones obtained from Data set 2: A decrease in the dispersion of the parameter estimates of an order of magnitude for component F to around half for component B is observed. Only component C sees no decrease in dispersion because 'ESP' selects all the samples containing C, meaning that using the full data with respect to ESP adds no additional information on C. Concerning component F, 'ESP' still selects the purest samples; however the selected samples have a significant noise level, inducing dispersion in the solutions. Weighting the data set with convex peeling instead of just the ESPs, increases the number of analysed samples containing component F, in turn, reducing the variance in its calculated spectral profile.

When comparing the results of 'Full' and 'ESP', both Data sets 2 and 3 show that, in the presence of minor components, a trade-off is present between the approaches. One should choose between precise but biased solutions with 'Full' or accurate but imprecise solutions with 'ESP'. 'Weighted', instead, takes the middle ground, where the utilisation of the full data combined with the knowledge of the essential information they carry, in the presence of noise and minor components, gives both more accurate and precise solutions.

5 | CONCLUSION

With this work, we show that, in the presence of minor components, ESP selection is required to drive the MCR-ALS solution towards the true one, with the caveat of losing parameter stability due to instrumental noise. We propose an extended weighting scheme within the weighted MCR-ALS framework that is based on convex hull data peeling and is able to preserve the benefits of ESP selection without reducing model parameter stability. This weighting framework is based on the relevance of the entire ensemble of investigated samples towards the MCR-ALS resolution. However, this can be further optimised by, for example, limiting the number of convex peels or applying a threshold on the sample relevance criterion to compress the data more adequately, reducing computation times. Furthermore, other relevance criteria can be applied as well (see the recent work done by Zade et al.[17]).

ACKNOWLEDGEMENTS

The authors acknowledge Laureen Coïc and Eric Ziemons for making their data available and Nematollah Omidikia and Mathias Sawall for fruitful discussion. Raffaele Vitale and Cyril Ruckebusch acknowledge financial support from the 'ANR-21-CE29-0007' project (Agence Nationale de la Recherche).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Mohamad Ahmad https://orcid.org/0000-0001-5127-5707
Raffaele Vitale https://orcid.org/0000-0002-7497-1673
Marina Cocchi https://orcid.org/0000-0001-8764-4981
Cyril Ruckebusch https://orcid.org/0000-0001-8120-4133

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/cem.3478.

REFERENCES

1. de Juan A, Tauler R. Multivariate curve resolution: 50 years addressing the mixture analysis problem—a review. *Anal Chim Acta*. 2021; 1145:59-78. doi:10.1016/j.aca.2020.10.051
2. Wentzell PD, Karakach TK, Roy S, Martinez MJ, Allen CP, Werner-Washburne M. Multivariate curve resolution of time course microarray data. *BMC Bioinformatics*. 2006;7(1):343. doi:10.1186/1471-2105-7-343
3. Blanchet L, Réhault J, Ruckebusch C, Huvenne JP, Tauler R, de Juan A. Chemometrics description of measurement error structure: study of an ultrafast absorption spectroscopy experiment. *Anal Chim Acta*. 2009;642(1-2):19-26. doi:10.1016/j.aca.2008.11.039
4. Vitale R, Ruckebusch C. On a black hole effect in bilinear curve resolution based on least squares. *J Chemometr*. 2023;37(2):e3442. doi: 10.1002/cem.3442
5. Ghaffari M, Omidikia N, Ruckebusch C. Essential spectral pixels for multivariate curve resolution of chemical images. *Anal Chem*. 2019; 91(17):10943-10948. doi:10.1021/acs.analchem.9b02890
6. Ruckebusch C, Vitale R, Ghaffari M, Hugelier S, Omidikia N. Perspective on essential information in multivariate curve resolution. *TrAC Trends Anal Chem*. 2020;132:116044. doi:10.1016/j.trac.2020.116044
7. Sawall M, Ruckebusch C, Beese M, Francke R, Prudlik A, Neymeyr K. An active constraint approach to identify essential spectral information in noisy data. *Anal Chim Acta*. 2022;1233:340448. doi:10.1016/j.aca.2022.340448
8. Coic L, Sacré PY, Dispas A, et al. Selection of essential spectra to improve the multivariate curve resolution of minor compounds in complex pharmaceutical formulations. *Anal Chim Acta*. 2022;1198:339532. doi:10.1016/j.aca.2022.339532

9. Nardecchia A, Duponchel L. Randomised SIMPLISMA: using a dictionary of initial estimates for spectral unmixing in the framework of chemical imaging. *Talanta*. 2020;217:121024. doi:10.1016/j.talanta.2020.121024
10. Ghaffari M, Omidikia N, Ruckebusch C. Joint selection of essential pixels and essential variables across hyperspectral images. *Anal Chim Acta*. 2021;1141:36-46. doi:10.1016/j.aca.2020.10.040
11. González-Martínez JM, Camacho J, Ferrer A. Bilinear modelling of batch processes. Part III: parameter stability. *J Chemometr*. 2013; 28(1):10-27. doi:10.1002/cem.2562
12. Preparata FP, Shamos MI. *Computational Geometry: An Introduction (Monographs in Computer Science) (Softcover reprint of the original 1st ed. 1985)*. Springer; 2012.
13. Bro R, Jong S. A fast non-negativity-constrained least squares algorithm. *J Chemometr*. 1997;11(5):393-401. doi:10.1002/(SICI)1099-128X(199709/10)11:5<393::AID-CEM483003E3.0.CO;2-L
14. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap: Chapter 9 (Chapman & Hall/CRC Monographs on Statistics and Applied Probability)*. 1st ed; 1993.
15. Stewart GW. On the early history of the singular value decomposition. *SIAM Rev*. 1993;35(4):551-566. doi:10.1137/1035134
16. Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem*. 1964;36(8):1627-1639. doi:10.1021/ac60214a047
17. Zade SV, Neymeyr K, Sawall M, Fischer C, Abdollahi H. Data point importance: information ranking in multivariate data. *J Chemometr*. 2023;37(1):e3453. doi:10.1002/cem.3453

How to cite this article: Ahmad M, Vitale R, Cocchi M, Ruckebusch C. Weighted multivariate curve resolution—Alternating least squares based on sample relevance. *Journal of Chemometrics*. 2023;e3478. doi:10.1002/cem.3478

Paper II

Mohamad Ahmad, Raffaele Vitale, Carolina S. Silva,
Cyril Ruckebusch and Marina Cocchi

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)

published

DOI: 10.1016/j.aca.2021.339285

Analytica Chimica Acta 1191 (2022) 339285

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

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)

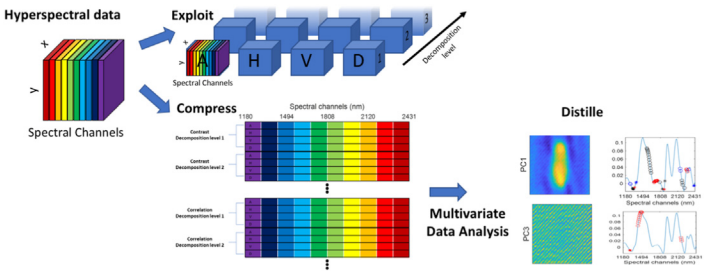
Mohamad Ahmad [a, b], Raffaele Vitale [b], Carolina S. Silva [c], Cyril Ruckebusch [b], Marina Cocchi [a, *]

[a] Università di Modena e Reggio Emilia, Dipartimento di Scienze Chimiche e Geologiche, Via Campi 103, 41125, Modena, Italy
[b] Univ. Lille, CNRS, LASIRE, LAboratoire de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, Cité scientifique, F-59000, Lille, France
[c] Department of Food Sciences and Nutrition, University of Malta, Msida, 2080, Malta

HIGHLIGHTS

- A novel method for unsupervised exploration of hyperspectral imaging data is presented.
- The method is based on Image Decomposition, Encoding and Localization steps.
- It retrieves distinct spatial features while linking them to specific spectral channels.
- The method is tested on data sets of forensic interest.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 7 August 2021
Received in revised form
6 November 2021
Accepted 14 November 2021
Available online 16 November 2021

Keywords:
Spectral imaging
Wavelet decomposition
Near-infrared
Image encoding
Forensics
Cotton
Biological fluids
Multivariate image analysis
Spatial-spectral analysis

ABSTRACT

The emergence of new spectral imaging applications in many science fields and in industry has not come to be a surprise, considering the immense potential this technique has to map spectral information. In the case of near-infrared spectral imaging, a rapid evolution of the technology has made it more and more appealing in non-destructive analysis of food and materials as well as in process monitoring applications. However, despite its great diffusion, some challenges remain open from the data analysis point of view, with the aim to fully uncover patterns and unveil the interplay between both the spatial and spectral domains. Here we propose a new approach, called Image Decomposition, Encoding and Localization (*IDEL*), where a spatial perspective is taken for the analysis of spectral images, while maintaining the significant information within the spectral domain. The methodology benefits from wavelet transform to exploit spatial features, encoding the outcoming images into a set of descriptors and utilizing multivariate analysis to isolate and extract the significant spatial-spectral information. A forensic case study of near-infrared images of biological stains on cotton fabrics is used as a benchmark. The stain and fabric have hardly distinguishable spectral signatures due to strong scattering effects that originate from the rough surface of the fabric and the high spectral absorbance of cotton in the near-infrared range. There is no selective information that can isolate signals related to these two components in the spectral images under study, and the complex spatial structure is highly interconnected to the spectral signatures. *IDEL* was capable of isolating the stains, (spatial) scattering effects, and a possible drying effect from the stains. It was possible to recover, at the same time, specific spectral regions that mostly highlight these isolated spatial structures, which was previously unobtainable.

* Corresponding author.
E-mail address: marina.cocchi@unimore.it (M. Cocchi).

https://doi.org/10.1016/j.aca.2021.339285
0003-2670/

1. Introduction

Near-infrared (NIR) imaging has become a cheap, versatile and very attractive method in many fields of science and diverse industrial applications, for its ability to capture phenomena in both spectral and spatial domains. Examples of applications are remote sensing in agriculture [1], stain analysis in forensics [2,3] and foodstuffs quality control [4]. With hundred-thousands of pixels for which a full NIR spectrum can be registered, the information content available from a spectral imaging data set is potentially overwhelming. This issue is often amplified by the nature of the sample itself, due to the complexity of its chemical composition and/or physical structure [5]. In most studies, it is mainly information extracted from the spectral domain that is exploited by the chemometric analysis, while the spatial (structural, textural) information of the sample is often disregarded.

The analysis of highly scattering materials is challenging in the NIR range and limited by the difficulty to describe the chemical and physical properties of the sample separately [6]. In practice, the separation of absorption and scattering has been the subject of different spectroscopic studies [7,8], with most of them removing the effect of scattering on the measured spectra by adequate preprocessing. Nonetheless, with NIR images of highly diffusive samples, scattering and absorption are entangled by highly non-linear mechanisms, which cannot always be fully eliminated by applying scatter-correction techniques on the individual spectral pixels, without the consideration of the spatial domain. Indeed, dramatic changes in the scattering contribution to the spectral signal can be expected from object borders and texture, which may fully dominate the spectral signal. Strong spectral interferences can be localized in the spatial domain by supervised or first principle modelling [9,10], but these methods require significant a priori information about the scattering behavior of the samples.

Most of the approaches to analyze NIR imaging data solely exploit spectral variation as the two spatial dimensions of the measured data cube are unfolded pixel-wise, ignoring spatial correlation. Still, one possible approach is multivariate image analysis (MIA), where the unfolded imaging data is augmented with pixel-neighbor information, to incorporate local-spatial information before it is analyzed with multivariate analysis tools, such as principal component analysis (PCA) or partial least squares (PLS) regression [11–13]. MIA has been originally proposed for RGB images [11,13] then extended to multi-channel images [14] and only recently to spectral images [11]. However, the number of neighboring pixels increases rapidly with the distance (or window size in pixels) from the center pixel at which to consider the neighborhood, and this applies to all spectral channels, making the data unmanageable in some cases. In this context, a parsimonious solution can be to employ multivariate curve resolution-alternating least square (MCR-ALS) using image processing constraints to take into account the spatial structure [15]. Nonetheless, this does require the data to strictly follow a bilinear model.

Image processing techniques (object detection [16], contrast enhancement [17], etc.) might be used to highlight some features of single images, i.e. at a given spectral channel or the mean image across all spectral channels [18], but disregarding the spectral domain will prevent chemical interpretation. Some work has also been done on image segmentation, with the integration of the spectral domain [19], as well as utilizing the spectral and spatial domain, interactively switching between the two modes [20]. The analysis of textural features in spectral imaging has also been explored, by using the wavelet transform (WT) [14,21–26]. These analyses i) use a MIA-like approach, where the local spatial information is extracted by WT, and 2D-WT sub-images are then analyzed by multivariate analysis [14,21–23], either on each single sub-image [21,22] or on the entire sets [14,23,24]; ii) exploit 3D-WT on the imaging data cube [25] or iii) fuse the 2D-WT sub-images obtained at each spectral channel [26]. These approaches also aim at linking the spectral and spatial domains, in some cases the spectral interpretation is not so straightforward [25] or of no concern [26], while in others the dimensionality of data matrix when passing from multispectral [14] to hyperspectral images become quite huge [24].

We recently proposed a novel approach to highlight the spatial-spectral interplay of the different components underlying a spectral imaging data set of a complex analytical system and published preliminary results [27] concerning a relatively simple Raman spectroscopy case study and a more complex one involving NIR spectral imaging datasets of an oil droplet in water and of biological fluids on cotton fabrics, respectively. However, we noticed that in systems of higher complexity, whose components show strong spectral and spatial overlap the analysis will become increasingly complex. To cope with this kind of situation, we here propose an extension and formalization leading to a novel method, called Image Decomposition, Encoding and Localization (*IDEL*). The method is meant to be unsupervised and exploratory.

IDEL relies on wavelet transform (WT) to resolve spatial features in distinct WT sub-images, then encodes this information in a set of descriptors (by using gray-level co-occurrence matrices [28]), and finally recovers specific spectral signatures for each spatial feature by multivariate data analysis. The encoded spatial information is fully exploited applying a semi-automatic procedure (that is data-driven) furnishing as a result a set of distinct spatial features linked to the specific spectral channels at which they are observable. In this way, clear and precise spatial features can be extracted, while chemical interpretability is maintained.

IDEL is challenged with a benchmark consisting of complex samples made of semen and lubricant stains on cotton fabrics analyzed with NIR imaging. There is significant spatial and spectral overlap between the stains and fabrics, and strong scattering effects are present. The localization of the fluid on the substrate is of interest in forensic applications. As such, the segmentation of the biological fluid from the substrate as well as the removal of the significant scattering effects visible in the spectral imaging data is crucial. *IDEL* was able to isolate the stains from the fabrics, while preserving spectral information, as well as isolating a spatial structure previously unobserved. Moreover, the final obtained model is capable of isolating components also in new images, of similar type, once projected on it.

2. Materials & methods

2.1. Methodology

The framework for WT decomposition and gray-level co-occurrence matrices is described in detail in Ref. [27] and briefly recalled in this section. The main novelty implemented in this work consists of a methodology where: i) only the most relevant spatial information is selected by applying PCA on the descriptors' matrix, which is based on picking the most significant scores (in terms of unique information) by means of convex peeling [29,30] and ii) a semi-automatic procedure to link the spectral information to the relevant spatial information, establishing a correspondence among PCA scores and loadings. As a result, the most relevant wavelet sub-images are extracted. These sub-images form a new data cube that contains the most significant spatial information at specific spectral channels. In more general terms, the spatial structures are firstly resolved, exploiting the original data cube. Then, maintaining a direct link with the spectral signature, a reduced image data cube is retrieved in the WT domain. Subsequently, to interpret the

corresponding information encoded in terms of individual spatial components, a PCA approach is proposed. In Fig. 1, the three main steps of *IDEL* are schematically shown. These steps are explained in detail in the following sub-sections.

2.1.1. Spectral image decomposition and encoding

The first step consists of the decomposition of the individual images corresponding to each spectral channel by 2D-WT (see Fig. 1a). 2D-WT is a very powerful filtering method, highlighting the different frequencies content of an image, while maintaining their localization with respect to the original domain. High- and low-pass filters are applied to decompose the signal into two disjoint subspaces holding the sets of details and approximation blocks (high and low frequencies, corresponding to sharp and smooth features, respectively). The decomposition is iterated on the approximation block, obtaining at each level a coarser representation of the image than in the previous approximation block and the filtered higher frequencies in the details. For image analysis, the same mono-dimensional wavelet filters are recursively applied along the two image directions. For each decomposition level, four blocks are obtained: 1) approximation (A): a low-pass filter is applied both row- and column-wise; 2) horizontal details (H): a low-pass filter is applied row-wise, then a high-pass filter, column-wise; 3) vertical details (V): a high-pass filter is applied row-wise, then a low-pass filter, column-wise; 4) diagonal details (D): a high-pass filter is applied both row- and column-wise. The specific direction along which the low- and high-pass filters are alternated, allows for specific textural patterns to be captured e.g., the H decomposition block highlights any pattern which would manifest horizontally, such as stripes (hence the name horizontal details). For the V and D blocks, vertical and diagonal textural patterns are highlighted, respectively, while the A block holds the original image with the details subtracted. We use the 2D stationary WT (2D-SWT) [31] which retains the size of the original image (see Fig. S1, Supplementary Material), so that the decomposition blocks (from now on, referred to as sub-images A, H, V, and D), for each decomposition level, are equal in size to the raw image.

Wavelet filters are grouped in specific families, which differ in shape and symmetry, while amplitude is modulated in each family by the number of vanishing moments [32]. The choice of an appropriate wavelet filter is data dependent and providing an automatic tool to tackle this task is outside the scope of the paper. However, there are criteria detailed in literature ([33], and references therein) to guide the choice of suitable wavelet filters. A general recommendation, that can be given is that the simplest *Haar* wavelet, which comes from the Daubechies-family (Daubechies-1) is usually a good starting point when, as in this case, the aim is exploratory. In fact, *Haar* can capture general changes present in an image, not focusing on specific spatial features, and disentangle signals that range from sharp contrasting edges to broad structures.

In this work, the simplest *Haar* wavelet is applied, which comes from the Daubechies-family (Daubechies-1) and showed good performance (other wavelet filters, such as Daubechies-2, -5, -7, Symlet-2, -4 and Coiflet-3, -5 were tested, data not shown); the maximum decomposition level compatible with the image size was used. As is illustrated in fig. 1b, 2D-SWT is applied to the spectral imaging data.

To exploit the spatial information, the decomposition blocks are encoded into a set of descriptors that contain information on distinct local spatial features. This is done by calculating descriptors on the gray-level co-occurrence matrices (GLCM) derived from the A, H, V, and D sub-images. The GLCM method maps the spatial dependence of pixel-pairs in quantized gray-level images. The quantization was set at 128. As such, each image intensity is normalized and distributed across 128 Gy-levels. A map is generated with size 128 by 128 elements, containing all possible quantized pixel-pairs. Selecting the appropriate number of gray-levels is similar to selecting the bin size in a mono-dimensional histogram and is always dependent on the size of, and information present in an image. A balance must be found between highlighting the relationships between neighboring pixels and not losing the details in the maps. Choosing a low number of gray-levels (large bin size) will result in a high number of counts over a small number of points, while choosing a high number of gray-levels (small bin size) will yield a low number of counts over a larger number of points (see Fig. S2 in Supplementary Material for a visual representation).

On the pixel-pairs counting, two other parameters are of importance, namely the offset and angle. Both parameters must be attuned to the decomposition blocks and levels, due to the nature of WT. The offset determines the distance at which every neighboring pixel is observed with respect to the main pixel e.g., for a direct neighboring pixel, this distance is 1. This has been set to vary as $2^{\text{level}-1}$, with *level* being the WT decomposition level corresponding to the sub-image being codified. This permits GLCM to account for the smoother patterns that are highlighted with increased WT decomposition levels, due to the removal of higher frequencies. The second parameter is the angle, or neighbors' location, which dictates the direction in which a neighboring pixel is located. We select the angle to maintain coherency between the directions of the WT decomposition details H, V, and D and the location of the investigated neighbor within the GLCM. Hence, the angle is set depending on the type of sub-image: H considers the top and bottom neighbors, V, the left and right neighbors, and D, the top-left, and bottom-right neighbors. These neighbors highlight the local differences within the sub-images. While for A, as there is no specific direction, the neighbors in all directions, are considered.

To encode the information carried by distinct patterns within an image, a set of eight descriptors was calculated from the GLCM, namely Energy, Contrast, Correlation, Variance, Inverse difference moment, Sum entropy, Information measure of correlation 1, and Maximal correlation coefficient. These are a subset of the descriptors proposed by Haralick et al. [28], which were selected as they are not much correlated with one another while describing all the relevant spatial features. We refer to Ref. [27] for a more in-depth survey of the selected descriptors.

As is illustrated in Fig. 1c, a matrix of dimensions: *decomposition blocks x decomposition levels*, in the rows, and *spectral wavelengths* in the columns, is obtained for each descriptor, which is auto-scaled. Appending column-wise the matrices obtained for all descriptors, a so-called Descriptors' matrix (DM) is obtained (see Fig. 1d).

2.1.2. Locating informative decomposed images

The DM contains descriptors on sub-images at every spectral channel, encoding spatial information in the rows and retaining spectral information in the columns. Applying PCA to DM, scores (Fig. 1e) and loadings (Fig. 1f) thus relate to the spatial and spectral information, respectively. The number of PCs to consider is of course data dependent and here we used the scree-plot as a guideline. Each point in the scores plot corresponds to one descriptor of a sub-image (A, H, V or D) at a specific decomposition level. Thus, looking at the scores, the most distinct spatial structures can be identified. The loadings plot, in conjunction with the scores plot, enables us to establish a link with the spectral channels. In fact, the loadings plot shows at which spectral wavelengths the largest variation of the descriptors within the different sub-images and decomposition levels is observed.

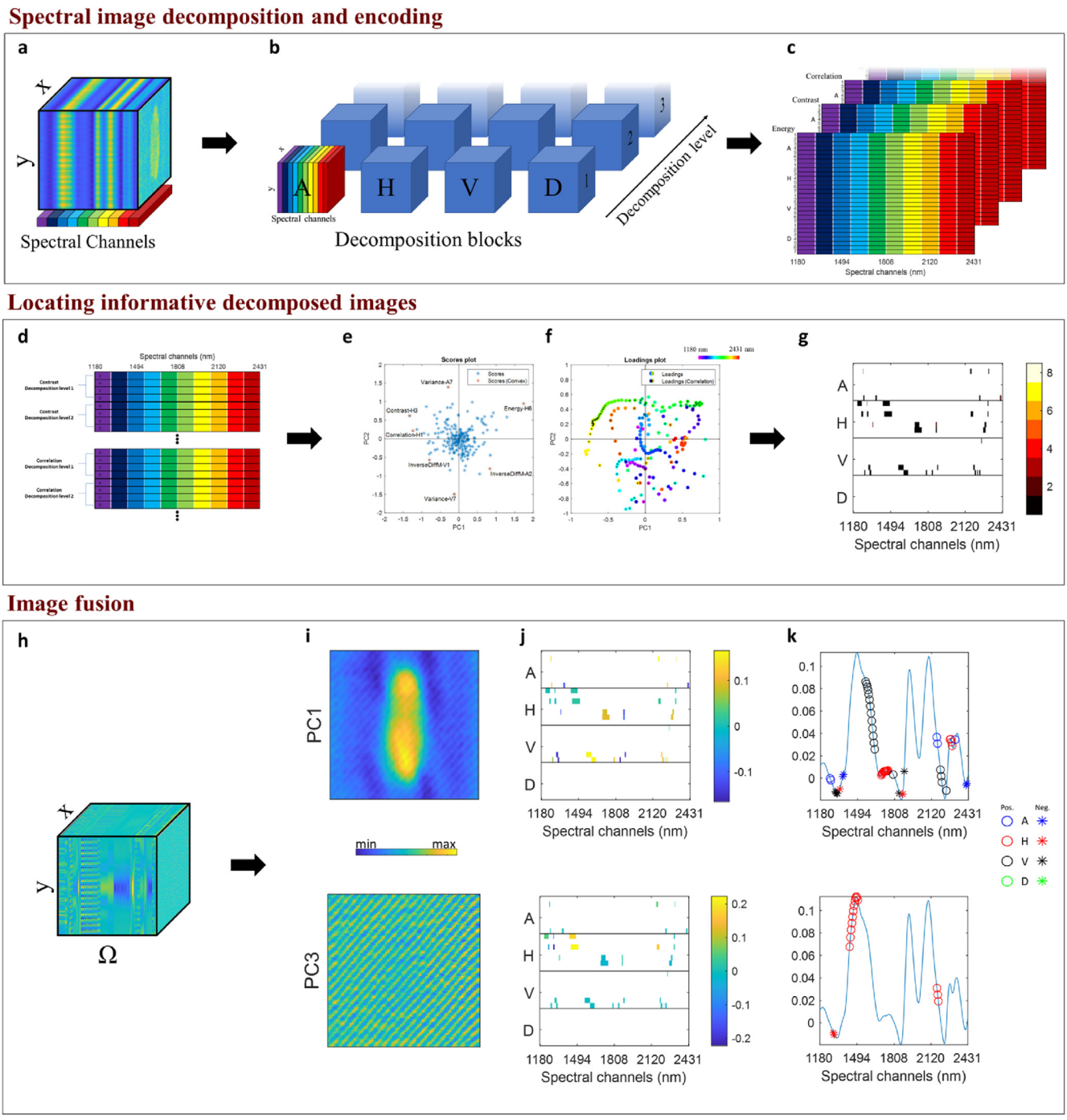


Fig. 1. Illustration of DIEL. The methodology consists of three main steps. Firstly, "Spectral image decomposition and encoding", encompassing; a) a NIR spectral image, which is decomposed by means of wavelet transform into; b) blocks containing horizontal (H), vertical (V) and diagonal (D) details, and approximations (A) at different decomposition levels; c) that are then encoded into distinct descriptors and organized into a Descriptor's matrix. Secondly, "Locating the informative decomposed images": where d) the Descriptors' matrix is unfolded descriptor wise, retaining the spectral dimension, and e)-f) decomposed by principal component analysis. The convex hull of the resulting scores is highlighted in red and labelled in the scores plot, while the corresponding salient loadings are highlighted by a black point, inside the colored point in the loadings plot. g) The scores (on the convex hull) and their respective (salient) loadings are mapped in the Ω-map. The map reports on the "x-axis" the spectral channels and on the "y-axis" the decomposition block, to which each sub-image belongs, as well as the decomposition levels ordered from first to last (going down). Lastly, "Image fusion", where the sub-images that are localized in the Ω-map are extracted from the reconstructed wavelet decomposition and assembled into a Ω-data cube (h). Principal component analysis is applied on the unfolded Ω-data cube and the resulting (refolded) scores' images for the first two principal components are shown in $i_{1\text{-}2}$. The loadings are mapped and visualized in a so-called loadings' map ($j_{1\text{-}2}$), where the color coding is set according to the loadings values. The mean spectrum is shown in $k_{1\text{-}2}$, which highlights only loadings with absolute values > 0.075 (to declutter the figure, where negative values are denoted by a ✶ and positive ones by a O), colored according to the decomposition block: A (blue), H (red), V (black) and D (green) sub-images. The purple to red color coding relates to the spectral dimension throughout the figure. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

To this aim, we developed a semi-automatic procedure, which looks for relevant points in the scores plot while matching them to the loadings. It is a two-steps procedure.

The first step of the procedure consists of the selection of relevant sub-images from the scores plot. It is based on the estimation of the convex-hull of the score points (Fig. 1e). Convex hull is applied, instead of e.g., a thresholding on scores values, as it depicts the minimum set of distinctive points enclosing all information captured in the scores plot. Here, we implemented a "peeling procedure" where the convex hull is applied twice. The first convex hull will remove the first "peel" of the data and the second will refine the selection. This accounts for situations where a few quite extreme points may skew the convex geometry too much [34,35]. This procedure identifies the distinct sub-images that show the highest variation across spectral channels for the descriptors as, e.g., in Fig. 1e, where the selected points (marked red) show significant variation on the first two PCs, meaning that they show high variation for a specific descriptor (within a certain sub-image at a given decomposition level across all spectral channels).

The second step is to match the salient spectral channels with the distinct spatial features (see Fig. 1f). To do this, the scores and loadings must be reconducted in the same space by adequate scaling, as in a biplot [36]. The correspondence of the loading points with the selected score points is expressed in terms of angle, which evaluates the location of the scores and loadings with respect to the origin of the PC-space. To identify the loading points that have a correspondence with specific score points, a threshold is set around 20° (zero degrees meaning perfect correspondence, and ninety degrees, no correspondence). The sign, of the scores and loadings, is not considered, meaning that a loading point that shows negative correlation (opposite location with respect to the PC origin) to a score point is considered equal to a loading point that shows positive correlation. We assume that positive and negative correlations between the scores and loadings have equal importance.

A single sub-image can be selected multiple times if it showed significant variation across spectral channels in several different descriptors. In fact, there are eight different points in the scores plot corresponding to each sub-image at a specific decomposition level, one for each GLCM descriptor. To give a clear overview of the selected sub-images at distinct spectral channels and highlight the sub-images that show significant variation for several descriptors, a representation is generated. This representation, depicted as a Ω-map in Fig. 1g, represents the decomposition blocks and levels of all sub-images *vs* the spectral channels. The Ω symbol indicates the sub-images selected by the procedure. The color coding on the color bar depicts the number of descriptors that were selected. In the end, only the sub-images that are required to explain the spatial features that make up the different spatial components in the wavelet decomposition are kept.

The selected sub-images are then reconstructed by inverse SWT and reorganized in the so-called Ω-data cube (Fig. 1h). Even if the decomposed sub-images are of congruent size, reconstruction avoids spatial distortion with respect to the original image, which may be introduced at the deepest level of decomposition and brings the decomposed images back to original intensity scale. The Ω-data cube contains the wavelet sub-images at specific spectral channels, that isolated the significant spatial structures determined from a set of chosen descriptors. Also, the values in each of these sub-images, when assembling the Ω-data cube, are auto-scaled, and multiplied by $\sqrt{f}$, with f being the number of descriptors that have been selected for each selected sub-image. In this way, more weight is given to sub-image which show significant variation for more than one descriptor, meaning that different and distinctive spatial features are enhanced/captured by them.

2.1.3. Image fusion

Notwithstanding that the Ω-data cube contains a subset of reconstructed wavelet sub-images exploited by 2D-SWT decomposition of spectral imaging data, it still includes some redundant spatial information (the same spatial features are visible at more than one single spectral channel). Thus, it can be desirable to further distill the captured information. We generically refer to this task as "image fusion" and different approaches may be used. The simplest approach is to decompose the data matrix obtained after pixel-wise unfolding of the Ω-data cube by applying PCA. The refolded scores will provide images (Fig. 1i) that combine the spatial patterns that show a similar variation in the Ω-domain. The representation and interpretation of the loadings is slightly more complex, as they do not encompass all channels of the original spectral domain (see Fig. 1j). The loadings are organized in such a way that they will have the same dimensions as the Ω-map to get a clear overview on their importance (by means of the color bar) at a specific decomposition and spectral channel. This results in a so-called loadings map. Beside it, for each PC, a plot of the mean spectrum of the original spectral image is reported, with only the significant loadings highlighted by using distinct symbols/colors to indicate the corresponding wavelet sub-image, i.e. A, H, V, and D (Fig. 1k). This is done to visualize any correspondence of the selected sub-images with any spectral bands in the original data set. The path from here can branch out, as extracting the spatial structures of the Ω-data cube can be done by several different fusion or modelling techniques e.g., one can apply MCR or Independent Component Analysis instead of PCA.

2.1.4. Ω-projection

An advantage of *IDEL* is that the generated PCA model can be used to project new imaging data, requiring only the 2D-SWT decomposition step to assemble the Ω-data for the test images (see section 3.3). Sub-images at the specific spectral channels (the ones belonging to the Ω of the training image) need to be calculated. Then, having the Ω-data of the new image, we can unfold and project it onto the PCA model, obtaining the scores, which in turn give the scores' images by refolding.

2.2. Data and preprocessing

The increased use of spectral images in forensic applications makes this methodology particularly interesting for body fluid detection [2,3,37–39]. In such a scenario, forensic experts are often searching for compounds (such as blood, semen, and saliva) with specific spectral signatures that can link a crime scene to a victim, an assaulter or even a witness. However, those fluids usually appear on many different substrates, whose composition and texture characteristics can hamper its localization, making it difficult for the analyst to identify its origin and, consequently, to submit them to further DNA analyses, for example.

IDEL has been applied on ten spectral images of stained cotton fabrics. There are five differently colored (yellow, white, red, green, and black) cotton fabrics, each with a stain of either lubricant or semen. All semen samples were obtained from the same donor [3,40], and the lubricant called KY-Jelly, mostly consisting of glycerol and hydroxyethyl-cellulose, came from the Durex© brand. The NIR imaging data was acquired by a Short-Wave Infrared (SWIR) SisuCHEMA imaging system from Specim (Oulu, Finland). The spectral range was 900–2500 nm with a spectral resolution FWHM of 10 nm and a spectral step size of 6.3 nm (256 spectral channels). The imaging system used had a lens of 50 mm and a pixel size of 156 × 156 mm^2. The squared pieces of fabric were stained with a droplet of semen and lubricant, and left to dry for a week at room temperature. We refer to Silva et al. [3] for more details on the

samples and data acquisition. The pre-processing of the data in this work is not the standard procedure for NIR imaging data, as the aim of standard procedures is to generate bi-linearity within the data by harmonizing the scattering and removing the variance of the path length (e.g. multiplicative scattering correction [41], MSC and standard normal variate [42], SNV). The angle of analysis for IDEL is image processing while maintaining spectral correlation. As such, the intended purpose of pre-processing is to contrast the spatial features within the images, while maintaining spectral correlation. To achieve this purpose weighted least squares baseline correction is applied, where a baseline is estimated for each pixel. However, firstly, the data was smoothed with Savitzky-Golay [43] (11-point window, 2nd order polynomial) to account for any unwanted spikes in the spectra. Secondly, the first 40 and last 15 spectral channels were removed, as these show only noisy images, containing no significant information. And lastly, weighted least squares (WLS) baseline correction (3rd order) [44] is applied. A summary of the example dataset used in this work is shown in Fig. 2. In addition, the scores maps and loadings profiles resulting from its PCA analysis after preprocessing by means of WLS and two more standard pretreatment algorithms for near-infrared data (i.e., MSC, and SNV) are displayed in Fig. S3 of Supplementary Material.

3. Results and discussion

The stained cotton fabrics data are of interest for forensic applications and were used for the purpose of presumptive identification of biological fluids on textile. These data provide a meaningful benchmark to assess the efficiency of *IDEL*. At least two chemical constituents exist for each image, the cotton and the stain (lubricant or semen), but there is no spatial region without cotton, and the location of the stain may be detectable at selected spectral channels, but it is not clearly observed in the raw data as it is mixed with the cotton [16]. Moreover, the fabric and stain show overlapped spectral bands. The results obtained from three different images will be discussed, namely the yellow cotton fabric with a lubricant stain (LY), the white cotton fabric with a semen stain (SW) and the red cotton fabric with a semen stain (SR) (see Fig. S4). In addition, the results of all ten datasets are provided in Fig. S5.

For the LY and SW data sets, parameters were set as detailed in section 2.1.1. The results will go over the PCA analysis of the Ω-data cube, investigating the scores' images, loadings maps and highlighted loadings on the mean spectrum, as shown in Fig. 1i, j and k.

3.1. Lubricant on yellow cotton fabric

The Ω-data cube for the LY data set consists of 140 sub-images, extracted from the wavelet decomposition, and the results are reported in Fig. 3. The resulting scores' images (Fig. 3a) of the first 4 PCs (explaining 58.3% variance of the Ω-data cube) are considered, where four distinctive spatial features are clearly recognizable. One can clearly identify the stain spot and the cotton fiber pattern, as will be discussed below.

The PC1 scores' image (Fig. 3a) mainly shows the presence of an intense spot (almost in the center) which can be identified as a stain. The corresponding loadings map (see Fig. 3b) shows that all the wavelet sub-images from every decomposition block (A, H, V and D) are contributing to the model, however the highest loadings values are mainly associated to approximation sub-images (A), which retain low frequency contributions in the original data set, hence smooth patterns. Fig. 3c represents the relevant spectral wavelengths on the mean spectrum of the spectral image with the notable points being: i) approximation sub-images at decomposition levels 2 and 6 (A-2 and A-6), which are linked to positive loadings within the spectral region 1990–2060 nm, ii) sub-images A-4 and 5, linked to negative loadings around 2400 nm, and iii) A-6 linked to negative loadings around 1300 nm. Although it is extremely difficult to consider band assignment in NIR for such complex matrices, the band around 2000 nm suggests contributions from glycerol [45], one of the main compounds of the lubricant. The negative contribution at 1300 nm is interesting as well, as neither cotton nor glycerol absorb at that wavelength. This contribution could be linked to a solely physical effect, due to the lack of absorbance of either cotton or lubricant, or it could be linked to a third unknown component.

The PC2 scores' image (Fig. 3a) clearly shows the diagonal texture associated to the cotton fibers. The loadings' map (Fig. 3b) shows that the most relevant contributions are from the H and D sub-images in the spectral range from 1400 to 1550 nm. When looking at the loadings (see Fig. 3), all these contributions relate to the band centered at 1494 nm. This can be attributed to the first overtone of O–H in cotton [46]. The main contribution comes from the D sub-images, however some minor contributions come from the H-sub-images. This can be attributed to the large spacing that is seen between the diagonal fibers, which can be captured in the horizontal details.

The PC3 scores' image (Fig. 3a) is not straightforward to interpret. It shows some very smooth patterns, which are usually

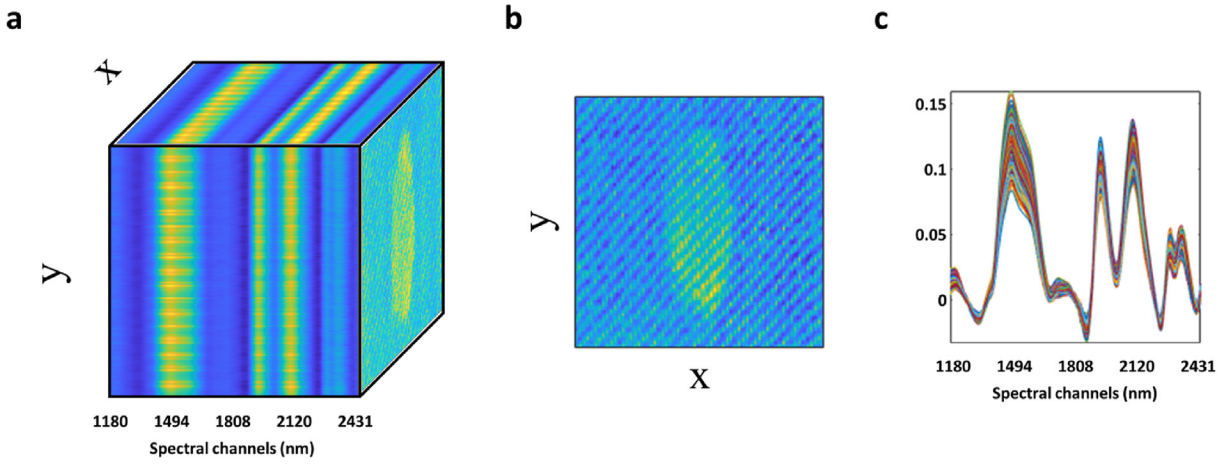


Fig. 2. An illustrative data set is shown: (a) the spectral data cube, (b) the corresponding mean image and (c) 1% of the spectra.

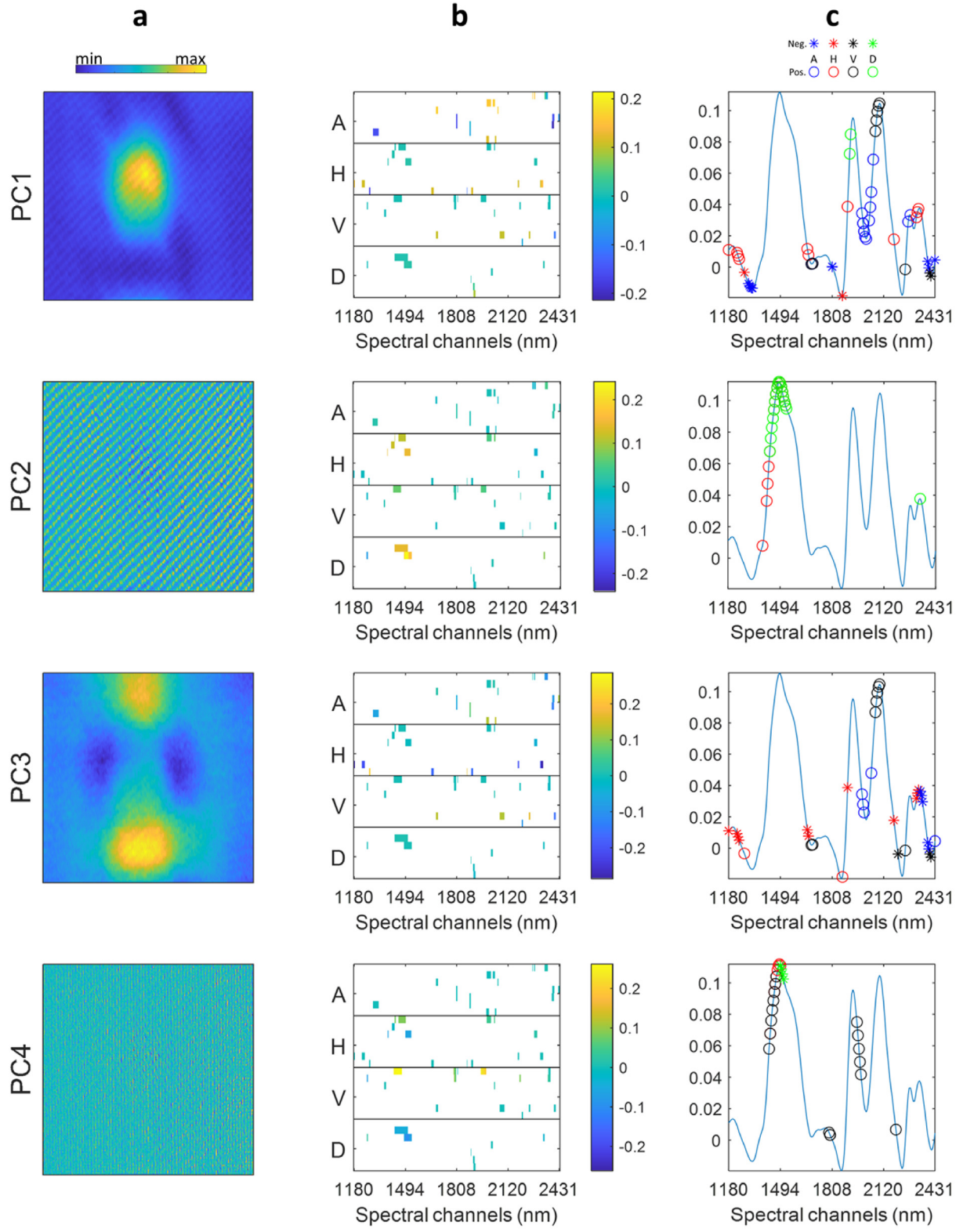


Fig. 3. Results for the LY data set are shown. Scores' images ($a_{1\text{-}4}$), loadings' maps ($b_{1\text{-}4}$) and salient spectral channels on the mean spectrum ($c_{1\text{-}4}$) are shown for the first four principal components extracted from the analysis of the Ω-data cube.

captured at the deepest decomposition levels (low frequency contributions in the spectral images) and mainly by approximations. However, details may also capture this type of information when, as in this case, there could be low frequency directional spatial patterns present (the most intense loadings values are from the H and V sub-images). When looking at the highest values in the loadings map (Fig. 3b) and at their location on the mean spectrum (Fig. 3c), the contributing spectral regions are quite spread and

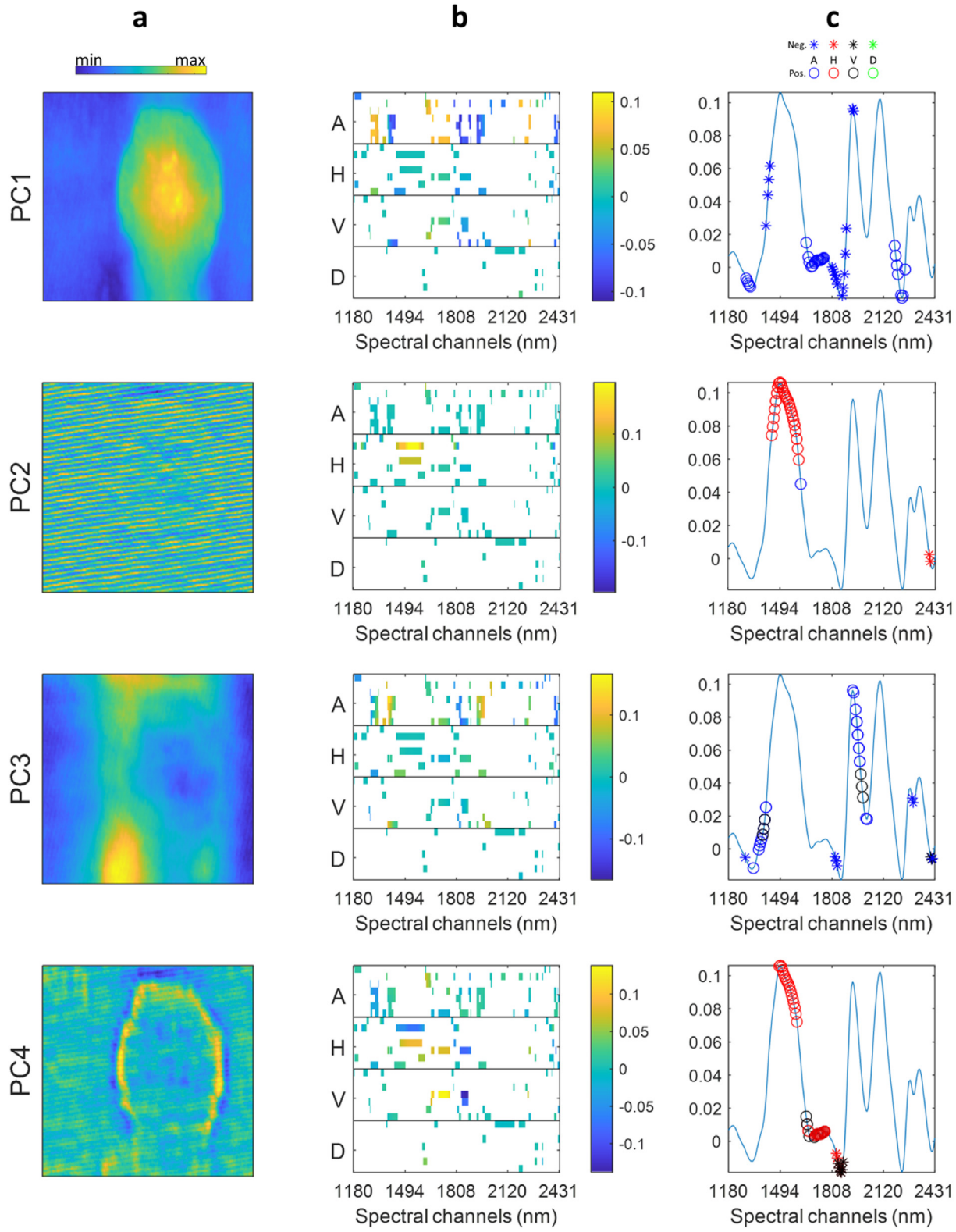


Fig. 4. Results for the SW data set are shown. Scores' images ($a_{1\text{-}4}$), loadings' maps ($b_{1\text{-}4}$) and salient spectral channels on the mean spectrum ($c_{1\text{-}4}$) are shown for the first four principal components extracted from the analysis of the Ω-data cube.

mainly on shoulders or along the spectral baseline. These patterns are quite difficult to interpret, and can originate from various sources, e.g. non-homogeneous illumination of the surface. These points can introduce minor variations in an image that can be seen in the deepest levels of a wavelet decomposition.

Finally, PC4, as for PC2, shows the texture of the cotton textile, however now its pattern has mostly a vertical direction. The main contributions are for details sub-images (mainly V) and again the

relevant spectral region includes the band centered at 1494 nm.

Added to this is a contribution from the spectral band at about 2000 nm, which was not captured by PC2. This spectral channel is slightly shifted with respect to the contribution discussed in PC1. This could be attributed to the first overtone of R—CO—R. Possible reasoning for PC4 to be separated from PC2 is that the spatial structure is significantly different and is isolated as a different component. Even though they both originate from cotton, the overlapping fiber structures show significant differences.

Summarizing the results for the LY data set, different spatial features could be isolated, segmenting the stain and recovering the cotton fiber patterns across the whole image in the scores' images. A possible link to the spectral domain has also been established, where the interplay of chemical and physical information is observed.

3.2. Semen on white cotton fabric

The Ω-data cube consists of 491 sub-images, extracted from the wavelet decomposition. The results of the PCA analysis of the Ω-data cube for the SW data set are shown in Fig. 4. The first four PCs (explaining 52.1% variance of the Ω-data), which capture the different spatial structures, are discussed below.

The PC1 scores' image only shows the semen stain without any pattern related to the texture of the fabric (see Fig. 4a). The loadings map (Fig. 4b) highlights several contributions but the highest (in absolute terms) are from the A sub-images across most of the decomposition levels. Reporting the correlated loadings on the mean spectrum, the corresponding spectral regions are located at: 1300 nm, 1700 nm and 2200 nm, showing positive loadings values, 1450 nm, 1850 nm and 1940 nm, showing negative loadings values.

The contributions at 1700, 1850 and 2200 nm could relate to semen, as they could be attributed to protein bands [47]. However, the band at 1300 nm is not attributable to a specific component: it could be that this is solely associated to physical scattering effects that come into play, as something similar was observed in the lubricant example. The 1450 and 1940 nm bands could be attributed to water bands [48], as the loadings show negative values and a faint negative circle is observable in the lower left part of the corresponding score image (Fig. 4a, PC1). A similar contribution can be seen on the PC3 scores' image (Fig. 4a) but with an inverted sign (positive values of scores and loadings). Even if the samples were dried, it cannot be excluded that water on the border is reabsorbed due to the environmental conditions, since the humidity of the room (where the samples were stored) was not controlled.

As in the lubricant data, the PC2 scores' image depicts the texture linked to the cotton fibers. However, here the fiber orientation spatially manifests in the horizontal direction. As such, the H sub-images are mostly selected (Fig. 4b, PC2). The salient loadings highlighted on the mean spectrum are associated to the absorption band at 1494 nm, which has already been referred to as the first overtone of O—H stretching in cotton.

The PC3 scores' image shows, like for the lubricant data, smooth spatial patterns. However, a small intense circle is also visible in the bottom left part of the image. Looking at the salient loadings, both in the loadings map and reported on the mean spectrum, we see that mostly A and V sub-images have the highest absolute loadings; the relevant spectral channels are in large part the same as for PC1, e.g. 1300, 1450, 1830 and 1940 nm. In fact, the simultaneous absorbance around 1450 and 1940 nm could be linked to water, which could mean that what is observed is due a to a drying effect at the border of the semen stain. Analogously, similar, but negative spectral contributions were observed in PC1. In the image, the semen stain border has an elongated form in the vertical direction. As such, it is being captured by the vertical details (V sub-image). On the other hand, the A sub-images capture the small spot, which is linked to semen.

The PC4 scores' image shows the border of the semen stain, captured by H and V sub-images, contributing the most to the loadings map. However, the contribution from the texture of cotton is observable. Around the border of the stain, the spectral contributions from the cotton fabric and the semen stain are strongly overlapping. The salient spectral regions include the 1700 nm band, already discussed for PC1 as connected to semen, and the 1500 nm band connected to the cotton fibers, mentioned with regards to PC2.

The compression (or "fusion") step operated by PCA was extremely efficient to extract information, separating spatially not only the semen stain from the texture (which consists of the scattering effects of cotton), but also distinguishing the scattering

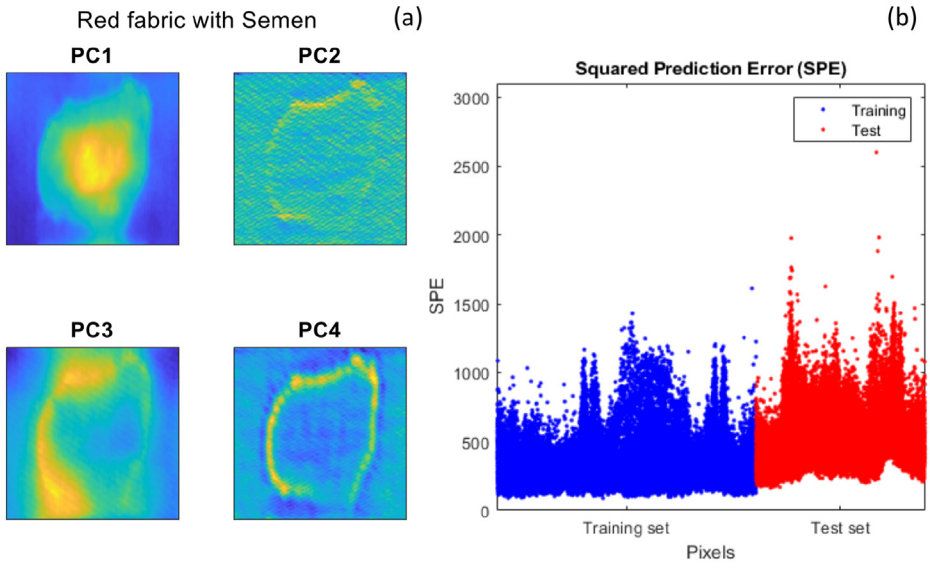


Fig. 5. (a) Results of the projection of the SR data set onto the SW PCA model. Scores' images of the first four principal components are shown; (b) Plot of squared prediction residuals (SPE). SPE for calibration set (SW image) are shown in blue color. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

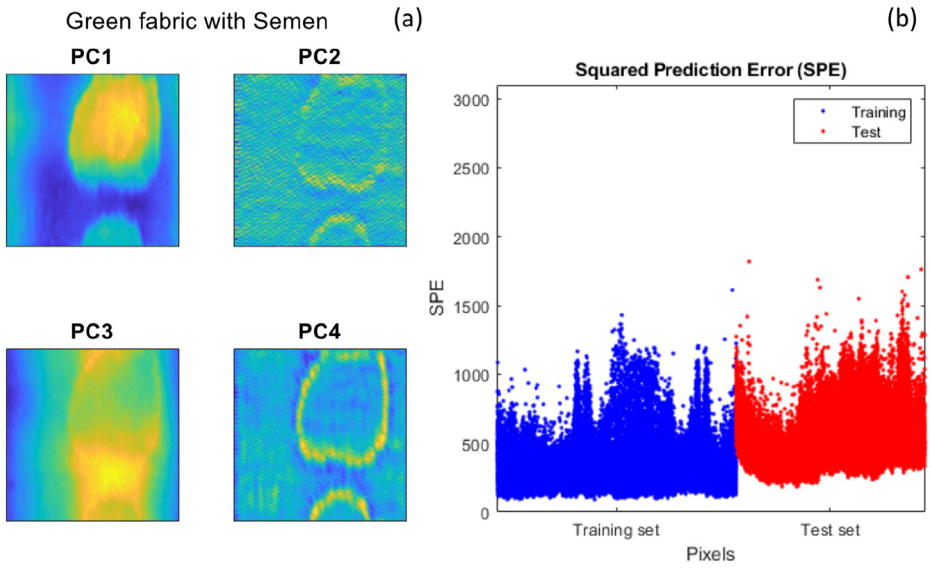


Fig. 6. (a) Results of the projection of the SG (semen stain on green fabric) data set onto the SW PCA model. Scores' images of the first four principal components are shown; (b) Plot of squared prediction errors (SPE). SPE for calibration set (SW image) are shown in blue color. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

around the border of the semen stain together with a possible drying effect of the semen.

3.3. Projection: semen on red cotton fabric

We have seen that the proposed approach is very efficient to retrieve spatial information and interpret it in terms of spectral contributions. In particular, the scores' images obtained from the analysis of the Ω-data help in discerning the various spatial components, which are not observable separately at any single spectral channel in the original data. The loadings highlight the spectral channels at which those components mostly manifest. A clear next step can be foreseen, which is evaluating if new (test) images projected on the loadings of a reference image can extract the same kind of specific spatial information in the scores' images.

The SR data set is investigated. The system is sufficiently similar to SW, but the shape of the stain, the scattering effects and the color of the cotton are different. The resulting scores' images are shown in Fig. 5. In the projected scores' images, similar spatial features can be observed: the semen stain is isolated in PC1, the texture of the cotton fibers with some bordering effects is seen in PC2, in PC3 a bordering effect linked to the semen stain is visible, and finally, the joint border and scattering effects are highlighted along PC4. Although the texture of the cotton fibers is not completely isolated from the border effects in PC2, the semen stain has been isolated and correctly identified. These minor differences may be due to spatial and spectral differences between the data sets. The texture of the cotton fibers is not the same, i.e., it is oriented in a different direction with respect to the modelled image. In addition, the color of the fabric is different: red fabrics might exhibit a different absorption with respect to white. Also, the amount of deposited semen may not be the same, nor its position or shape. Very similar results were obtained by projecting, as test image, the green cotton fabric with a semen stain (SG), as shown in Fig. 6a. It is worth noticing that, for both SR and SG, the squared prediction residuals (SPE) are in the same range of the calibration image (i.e. SW) as shown in Figs. 5b and 6b, respectively.

Overall, these results seem very promising. Nonetheless, the projection (figure not shown for the sake of brevity) of the black fabric image with a semen stain (SB) and, to a minor extent, of the yellow fabric, while showing similar spatial features on scores images, resulted in high SPE signaling that when, the spatial structures and/or the spectral background (as it is the case of SB) of the test images are very different from the calibration image much care should be taken in interpreting the scores maps, even if interesting spatial structure are unveiled.

4. Concluding remarks

IDEL utilizes WT, image encoding and PCA to extract decomposed sub-images that show significant variation across the spectral domain for spatial features related to distinct descriptors. Not only can it extract the distinct spatial-scattering effects present in a NIR spectral image, but also other components that show significant spatial differences between each other, while simultaneously having the capability to retain the spectral information that is linked to such captured spatial components. Thus, *IDEL* seems a very useful and powerful spectral imaging exploratory tool. However, some care must be paid when interpreting the highlighted spectral channels, as the previously discussed physico-chemical effects are difficult to separate from one another.

Once the model is built for components that have distinct spatial-spectral features, test images can be projected onto its space for their direct assessment. Also, the application of PCA to the Ω-data cube showed very promising results for spectral image interpretation. Some future work will be to utilize image fusion techniques to better extract and isolate spatial components.

The results obtained in this work can be generalized to any application field employing spectral imaging for the visualization of materials characterized by high morphological content, such as biological tissues [48], wooden materials [49–51], or remote sensing [52]. The integration of the proposed approach with other data analysis techniques, like multivariate curve resolution (MCR), will also be investigated.

CRediT authorship contribution statement

Mohamad Ahmad: Writing – original draft, Conceptualization, Methodology, Visualization, Software. **Raffaele Vitale:** Supervision, Software, Data curation, Writing – review & editing. **Carolina S.**

Silva: Investigation, Resources, Writing – review & editing, Funding acquisition. **Cyril Ruckebusch:** Supervision, Validation, Writing – review & editing, Project administration. **Marina Cocchi:** Supervision, Validation, Software, Visualization, Writing – review & editing, Project administration, Funding acquisition.

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.

Acknowledgements

Dr. C.S. Silva acknowledges financial support from: FACEPE (BFP-0800-1.06/17 and APQ-0576-1.06/17).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2021.339285.

References

[1] I. Tahmasbian, N.K. Morgan, S. Hosseini Bai, M.W. Dunlop, A.F. Moss, Comparison of hyperspectral imaging and near-infrared spectroscopy to determine nitrogen and carbon concentrations in wheat, Rem. Sens. 13 (2021) 1128, https://doi.org/10.3390/rs13061128.
[2] C. Malegori, E. Alladio, P. Oliveri, C. Manis, M. Vincenti, P. Garofano, F. Barni, A. Berti, Identification of invisible biological traces in forensic evidences by hyperspectral NIR imaging combined with chemometrics, Talanta 215 (2020), https://doi.org/10.1016/j.talanta.2020.120911, 120911.
[3] C.S. Silva, M.F. Pimentel, J.M. Amigo, R.S. Honorato, C. Pasquini, Detecting semen stains on fabrics using near infrared hyperspectral images and multivariate models, Trac. Trends Anal. Chem. 95 (2017) 23–35, https://doi.org/10.1016/j.trac.2017.07.026.
[4] H. Huang, L. Liu, M.O. Ngadi, Recent developments in hyperspectral imaging for assessment of food quality and safety, Sensors 14 (2014) 7248–7276, https://doi.org/10.3390/s140407248.
[5] M. Manley, Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials, Chem. Soc. Rev. 43 (2014) 8200–8214, https://doi.org/10.1039/C4CS00062E.
[6] J. Zhou, L. Yu, Q. Ding, R. Wang, Textile fiber identification using near-infrared spectroscopy and pattern recognition, Autex Res. J. 19 (2019) 201–209, https://doi.org/10.1515/aut-2018-0055.
[7] L.E. Agelet, C.R. Hurburgh, A tutorial on near infrared spectroscopy and its calibration, Crit. Rev. Anal. Chem. 40 (2010) 246–260, https://doi.org/10.1080/10408347.2010.515468.
[8] B. Debus, R. Vitale, S. Sasaki, T. Asahi, M. Sliwa, C. Ruckebusch, A multivariate curve resolution approach to separate UV–vis scattering and absorption contributions for organic nanoparticles, Chemometr. Intell. Lab. Syst. 160 (2017) 72–76, https://doi.org/10.1016/j.chemolab.2016.11.011.
[9] E.A. Magnussen, J.H. Solheim, U. Blazhko, V. Tafintseva, K. Tøndel, K.H. Liland, S. Dzurendova, V. Shapaval, C. Sandt, F. Borondics, A. Kohler, Deep convolutional neural network recovers pure absorbance spectra from highly scatter-distorted spectra of cells, J. Biophot. 13 (2020), e202000204, https://doi.org/10.1002/jbio.202000204.
[10] A. Kohler, J.H. Solheim, V. Tafintseva, B. Zimmermann, V. Shapaval, Model-based pre-processing in vibrational spectroscopy, in: Comprehensive Chemometrics, Elsevier, 2020, pp. 83–100.
[11] F. Jamme, L. Duponchel, Neighbouring pixel data augmentation: a simple way to fuse spectral and spatial information for hyperspectral imaging data analysis, J. Chemometr. 31 (2017), e2882, https://doi.org/10.1002/cem.2882.
[12] M.H. Bharati, J. Liu, J.F. MacGregor, Image texture analysis: methods and comparisons, Chemometr. Intell. Lab. Syst. 72 (2004) 57–71, https://doi.org/10.1016/j.chemolab.2004.02.005.
[13] J.M. Prats-Montalbán, A. de Juan, A. Ferrer, Multivariate image analysis: a review with applications, Chemometr. Intell. Lab. Syst. 107 (2011) 1–23, https://doi.org/10.1016/j.chemolab.2011.03.002.
[14] M. Li Vigni, J.M. Prats-Montalban, A. Ferrer, M. Cocchi, Coupling 2D-wavelet decomposition and multivariate image analysis (2D WT-MIA), J. Chemometr. 32 (2018), e2970, https://doi.org/10.1002/cem.2970.
[15] R. Vitale, S. Hugelier, D. Cevoli, C. Ruckebusch, A spatial constraint to model and extract texture components in Multivariate Curve Resolution of near-infrared hyperspectral images, Anal. Chim. Acta 1095 (2020) 30–37, https://doi.org/10.1016/j.aca.2019.10.028.
[16] Z. Wang, P. Xu, B. Liu, Y. Cao, Z. Liu, Z. Liu, Hyperspectral imaging for underwater object detection, SR 41 (2021) 176–191, https://doi.org/10.1108/SR-07-2020-0165.
[17] G. Maragatham, S. Mansoor Roomi, A review of image contrast enhancement methods and techniques, RJASET 9 (2015) 309–326, https://doi.org/10.19026/rjaset.9.1409.
[18] 2014, Annual IEEE Computer Conference, IEEE International Conference on Image Processing, ICIP, IEEE International Conference on Image Processing (ICIP), IEEE, Piscataway, NJ, 2014, pp. 27–30. Oct. 2014, Paris, France.
[19] J.-L. Xu, A.A. Gowen, Spatial-spectral analysis method using texture features combined with PCA for information extraction in hyperspectral images, J. Chemometr. 34 (2020), https://doi.org/10.1002/cem.3132.
[20] N. Gorretta, J.M. Roger, G. Rabatel, V. Bellon-Maurel, C. Fiorio, C. Lelong, Hyperspectral image segmentation: the butterfly approach, in: Whispers 2009, IEEE, Piscataway, NJ, 2009, pp. 1–4.
[21] J. Liu, J. MacGregor, On the extraction of spectral and spatial information from images, Chemometr. Intell. Lab. Syst. 85 (2007) 119-130, https://doi.org/10.1016/j.chemolab.2006.05.011.
[22] M.S. Reis, An integrated multiscale and multivariate image analysis framework for process monitoring of colour random textures: MSMIA, Chemometr. Intell. Lab. Syst. 142 (2015) 36-48, https://doi.org/10.1016/j.chemolab.2015.01.008.
[23] P. Juneau, A. Garnier, C. Duchesne, The undecimated wavelet transform—multivariate image analysis (UWT-MIA) for simultaneous extraction of spectral and spatial information, Chemometr. Intell. Lab. Syst. 142 (2015) 304-318, https://doi.org/10.1016/j.chemolab.2014.09.007.
[24] A. Nardecchia, R. Vitale, L. Duponchel, Fusing spectral and spatial information with 2-D stationary wavelet transform (SWT 2-D) for a deeper exploration of spectroscopic images, Talanta 224 (2021) 121835, https://doi.org/10.1016/j.talanta.2020.121835.
[25] X. Guo, X. Huang, L. Zhang, Three-dimensional wavelet texture feature extraction and classification for multi/hyperspectral imagery, IEEE geosci, Rem. Sens. Lett. 11 (2014) 2183–2187, https://doi.org/10.1109/LGRS.2014.2323963.
[26] M. Beauchemin, Spatial pattern discovery for hyperspectral images based on multiresolution analysis, Int. J. Image Data Fusion 3 (2012) 93–110.
[27] M. Ahmad, R. Vitale, C.S. Silva, C. Ruckebusch, M. Cocchi, Exploring local spatial features in hyperspectral images, J. Chemometr. 34 (2020), https://doi.org/10.1002/cem.3295.
[28] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, in: IEEE Trans. Syst., Man, Cybern. SMC-3, 1973, pp. 610–621, https://doi.org/10.1109/TSMC.1973.4309314.
[29] H. Caussinus, P. Ettinger, R. Tomassone, Proceedings in Computational Statistics, Physica-Verl., Heidelberg, Wien, 1982.
[30] A. Cutler, L. Breiman, Archetypal analysis, Technometrics 36 (1994) 338, https://doi.org/10.2307/1269949.
[31] G.P. Nason, B.W. Silverman, The stationary wavelet transform and some statistical applications, in: A. Antoniadis (Ed.), Wavelets and Statistics, Springer-Verlag, New York, 1995, pp. 281–299.
[32] A. Cohen, I. Daubechies, B. Jawerth, P. Vial, Multiresolution analysis, wavelets and fast wavelet transform on an interval, CRAS Paris, Ser. A 316 (1993) 417-421.
[33] J.M. Prats-Montalbán, M. Cocchi, A. Ferrer, N-way modeling for wavelet filter determination in multivariate image analysis, J. Chemometr. 29 (2015) 379–388.
[34] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (2014) 2812–2831, https://doi.org/10.1039/C3AY41907J.
[35] J.A. Fernández Pierna, L. Jin, M. Daszykowski, F. Wahl, D.L. Massart, A methodology to detect outliers/inliers in prediction with PLS, Chemometr. Intell. Lab. Syst. 68 (2003) 17–28, https://doi.org/10.1016/S0169-7439(03)00084-4.
[36] I.T. Jolliffe, Principal Component Analysis, Springer International Publishing, Cham, 20.
[37] A. Majda, R. Wietecha-Posłuszny, A. Mendys, A. Wójtowicz, B. Łydżba-Kopczyńska, Hyperspectral imaging and multivariate analysis in the dried blood spots investigations, Appl. Phys. A 124 (2018), https://doi.org/10.1007/s00339-018-1739-6.
[38] M. Romaszewski, P. Głomb, A. Sochan, M. Cholewa, A dataset for evaluating blood detection in hyperspectral images, Forensic Sci. Int. 320 (2021) 110701, https://doi.org/10.1016/j.forsciint.2021.110701.
[39] F. Zapata, F.E. Ortega-Ojeda, C. García-Ruiz, Revealing the location of semen, vaginal fluid and urine in stained evidence through near infrared chemical imaging, Talanta 166 (2017) 292–299, https://doi.org/10.1016/j.talanta.2017.01.086.
[40] D.H. Owen, D.F. Katz, A review of the physical and chemical properties of human semen and the formulation of a semen simulant, J. Androl. 26 (2005) 459–469, https://doi.org/10.2164/jandrol.04104.
[41] P. Geladi, D. MacDougall, H. Martens, Linearization and scatter-correction for near-infrared reflectance spectra of meat, Appl. Spectrosc. 39 (1985) 491–500.
[42] R. Barnes, M. Dhanoa, S. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777.
[43] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639, https://doi.org/10.1021/ac60214a047.

[44] Eigenvector Research, Inc., PLS_Toolbox Function Reference Manual, Eigenvector Research, Inc., 2021 available at: http://wiki.eigenvector.com/index.php?title=Wlsbaseline.
[45] K. Izutsu, Y. Hiyama, C. Yomota, T. Kawanishi, Near-infrared analysis of hydrogen-bonding in glass- and rubber-state amorphous saccharide solids, AAPS PharmSciTech 10 (2009) 524–529, https://doi.org/10.1208/s12249-009-9243-0.
[46] H.W. Siesler, Y. Ozaki, S. Kawata, H.M. Heise (Eds.), Near-Infrared Spectroscopy, Wiley-VCH Verlag GmbH, Weinheim, Germany, 2001.
[47] E.W. Ciurczak, D.A. Burns, Handbook of Near-Infrared Analysis, second ed., Marcel Dekker, New York, 2001.
[48] Y. Ozaki, Applications in chemistry, in: H.W. Siesler, Y. Ozaki, S. Kawata, H.M. Heise (Eds.), Near-Infrared Spectroscopy, Wiley-VCH Verlag GmbH, Weinheim, Germany, 2001, pp. 179–211.
[49] M. Halicek, H. Fabelo, S. Ortega, G.M. Callico, B. Fei, In-vivo and ex-vivo tissue analysis through hyperspectral imaging techniques: revealing the invisible features of cancer, Cancers 11 (2019), https://doi.org/10.3390/cancers11060756.
[50] J. Sandak, A. Sandak, L. Legan, K. Retko, M. Kavčič, J. Kosel, F. Poohphajai, R.H. Diaz, V. Ponnuchamy, N. Sajinčič, O. Gordobil, Č. Tavzes, P. Ropret, Nondestructive evaluation of heritage object coatings with four hyperspectral imaging systems, Coatings 11 (2021) 244, https://doi.org/10.3390/coatings11020244.
[51] R. Vitale, P. Stefansson, F. Marini, C. Ruckebusch, I. Burud, H. Martens, Fast analysis, processing and modeling of hyperspectral videos: challenges and possible solutions, in: Comprehensive Chemometrics, Elsevier, 2020, pp. 395–409.
[52] P. Stefansson, J. Fortuna, H. Rahmati, I. Burud, T. Konevskikh, H. Martens, Chapter 2.12 - hyperspectral time series analysis: hyperspectral image data streams interpreted by modeling known and unknown variations, in: J.M. Amigo (Ed.), Data Handling in Science and Technology Hyperspectral Imaging, Elsevier, 2020, pp. 305–331.

Paper III

Mohamad Ahmad, Raffaele Vitale, Carolina S. Silva,
Cyril Ruckebusch and Marina Cocchi

Exploring local spatial features in hyperspectral images

published

DOI: 10.1002/cem.3295

Received: 16 March 2020 | Revised: 17 June 2020 | Accepted: 6 August 2020

DOI: 10.1002/cem.3295

SHORT COMMUNICATION

Journal of CHEMOMETRICS WILEY

Exploring local spatial features in hyperspectral images

Mohamad Ahmad[1,2] | **Raffaele Vitale**[2,3] | **Carolina S. Silva**[4] | **Cyril Ruckebusch**[2] | **Marina Cocchi**[1]

[1]Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, Via Campi 103, Modena, 41125, Italy

[2]Univ. Lille, CNRS, LASIRE, Lille, F-59000, France

[3]Department of Chemistry, Molecular Imaging and Photonics Unit, KU-Leuven, Celestijnenlaan 200F, Leuven, B-3001, Belgium

[4]Department of Chemical Engineering, Federal University of Pernambuco, Av. Prof. Moraes Rego, 1235, Cidade Universitaria, Recife, Brazil

Correspondence
Marina Cocchi, Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, Via Campi 103, 41125 Modena, Italy.
Email: marina.cocchi@unimore.it

Funding information
FACEPE, Grant/Award Numbers: BFP-0800-1.06/17, APQ-0576-1.06/17; Núcleo de Estudos em Química Forense, Grant/Award Number: CAPES AUXPE 3509/2014; NUQAAPE-FACEPE, Grant/Award Number: APQ-0346-1.06/14

Abstract

We propose a methodological framework to extract spatial features in hyperspectral imaging data and establish a link between these features and the spectral regions, capturing the observed structural patterns. The proposed approach consists of five main steps: (i) two-dimensional stationary wavelet transform (2D-SWT) is applied to a hyperspectral data cube, decomposing each single-channel image with a selected wavelet filter up to the maximum decomposition level; (ii) a gray-level co-occurrence matrix is calculated for every 2D-SWT image resulting from stage (i); (iii) distinctive spatial features are determined by computing morphological descriptors from each gray-level co-occurrence matrix; (iv) the morphological descriptors are rearranged in a two-dimensional data array; and (v) this data matrix is subjected to principal component analysis (PCA) for exploring the variability of the aforementioned descriptors across spectral channels. As a result, groups of spectral wavelengths associated to specific spatial features can be pointed out yielding a better understanding and interpretation of the data. In principle, this information can also be further exploited, for example, to improve the separation of pure spectral profiles in a multivariate curve resolution context.

KEYWORDS

gray-level co-occurrence matrix, hyperspectral images, multivariate image analysis, spatial features, wavelet transform

1 | INTRODUCTION

Hyperspectral imaging (HSI) has numerous possible applications that, depending on the instrumentation and the spectral domain covered, can range from environmental surveillance to cellular monitoring.[1–4] HSI data consist of three-dimensional arrays with two spatial dimensions and one spectral dimension, providing an image for each scanned spectral channel. When HSI is concerned, one is usually interested in retrieving both the pure spectra of the individual components constituting the image and their respective spatial distribution.

In chemistry, one of the most used approaches for achieving this aim is multivariate curve resolution–alternating least squares (MCR–ALS). With MCR–ALS, a hyperspectral image is first unfolded pixel-wise, and afterwards, a bilinear model is fitted to the unfolded data using an ALS approach under appropriate constraints. This permits to unravel the distribution maps and the pure spectral signatures of the physicochemically meaningful constituents of the image.[5] However, unfolding the data results in losing the information on the local spatial features within the dataset. A solution to this issue can be the implementation of spatial constraints as described in Hugelier et al.[6] Nonetheless, if different constituents exhibit distinct spatial features (i.e., textural patterns), and/or two (or more) of

these different constituents show a significant overlap along both the spectral and the spatial domain, this approach may not be actionable. Multivariate image analysis (MIA) is a very useful alternative to investigate spatial features in grayscale, RGB, and, to a lesser extent, hyperspectral images.[7–9] The basic principle of MIA is to analyze the unfolded data by means of multivariate tools like principal component analysis (PCA) or partial least squares (PLS) regression. Spatial features are captured considering the relationships between each pixel and its neighbors in the unfolding step (see Bharati et al.[7] and Prats-Montalbán et al.[8] for details). MIA has also been coupled to wavelet transform (WT) multiresolution analysis,[10] and this combination has been proven effective in resolving spatial features in multispectral[11] and Raman hyperspectral images.[12] In addition, other strategies, like coclustering[13] and gray-level co-occurrence matrices (GLCMs),[14] have recently been applied to examine texture in HSI datasets. Textural features in HSI have also been explored by using three-dimensional discrete WT (DWT)[15] or by fusion of the two-dimensional DWT decomposition images obtained from each spectral channel.[16] All these approaches are capable of enhancing spatial features in HSI data and of establishing some link between these features and the spectral regions responsible for the observed textural patterns. However, their performance has not been satisfactory in situations where both textural/spatial and spectral information are highly mixed, that is, when pure chemical components and/or distinct physical contributions are overlapped both in the investigated spectral range and in their distribution across the image.[17]

Aiming at facing this issue, we propose in this short communication a methodological framework that relies on the capability of two-dimensional stationary wavelet transform (2D-SWT)[18,19] to capture distinct spatial features in disjoint subspaces (different wavelet images can be extracted for every spectral channel) and on the versatility of multivariate data analysis tools to explore the information these spatial features carry. The approach consists of five consecutive steps: (i) 2D-SWT decomposition of the HSI data cube resulting in wavelet subimages; (ii) from these subimages, computation of the GLCM; (iii) calculation of morphological descriptors from each GLCM; (iv) rearrangement of the obtained descriptors into a matrix; and (v) PCA modeling of this matrix for the exploration of the variability of the morphological descriptors across spectral channels.

Such a workflow would allow, for example, spectral wavelengths mostly capturing specific spatial features to be pointed out by the simultaneous investigation of PCA loadings and scores. PCA can also be coupled to other statistical approaches for addressing particular tasks depending on the study at hand. Here, for example, we used the *k*-means algorithm to cluster the loadings obtained from the PCA of the descriptor matrix and determine the spectral regions that show similar variation patterns in terms of spatial descriptors. These regions can be selective for different pure components underlying HSI data and highlighting them can help users to, for example, improve the quality of MCR-ALS solutions.

2 | METHODS

The proposed methodology consists of five main steps which are illustrated in Figure 1 and detailed below.

2.1 | Step 1: 2D-SWT decomposition

A low-pass filter and a high-pass filter are applied to every spectral channel of the analyzed hyperspectral image (see Figure 1A) to obtain four distinct sets of wavelet coefficients, denoted H, V, D, and A. Each set corresponds to a decomposition block and will be referred to as a wavelet subimage. The horizontal set (H) corresponds to the application of both a low-pass filter, row-wise, and high-pass filter, column-wise; the vertical set (V) to the application of a high-pass filter, row-wise, and a low-pass filter, column-wise; the diagonal set (D) to the application of a high-pass filter, both row-wise and column-wise; and the approximation set (A) to the application of a low-pass filter, both row-wise and column-wise. This scheme is iterated on the approximation subimage until a certain decomposition level has been reached (see Figure 1B). In 2D-SWT, at each iteration of the decomposition, the wavelet filter is upsampled, contrary to standard DWT[20] where the wavelet coefficients are downsampled. In this way, congruent wavelet subimages are yielded, and the position of the spatial patterns in the image is preserved.

For the objective of this short communication, the decomposition level is set to the maximum compatible with the size of the original image. Furthermore, the Haar filter was utilized here even though different decomposition filters exist and can be exploited for the same purpose.[21]

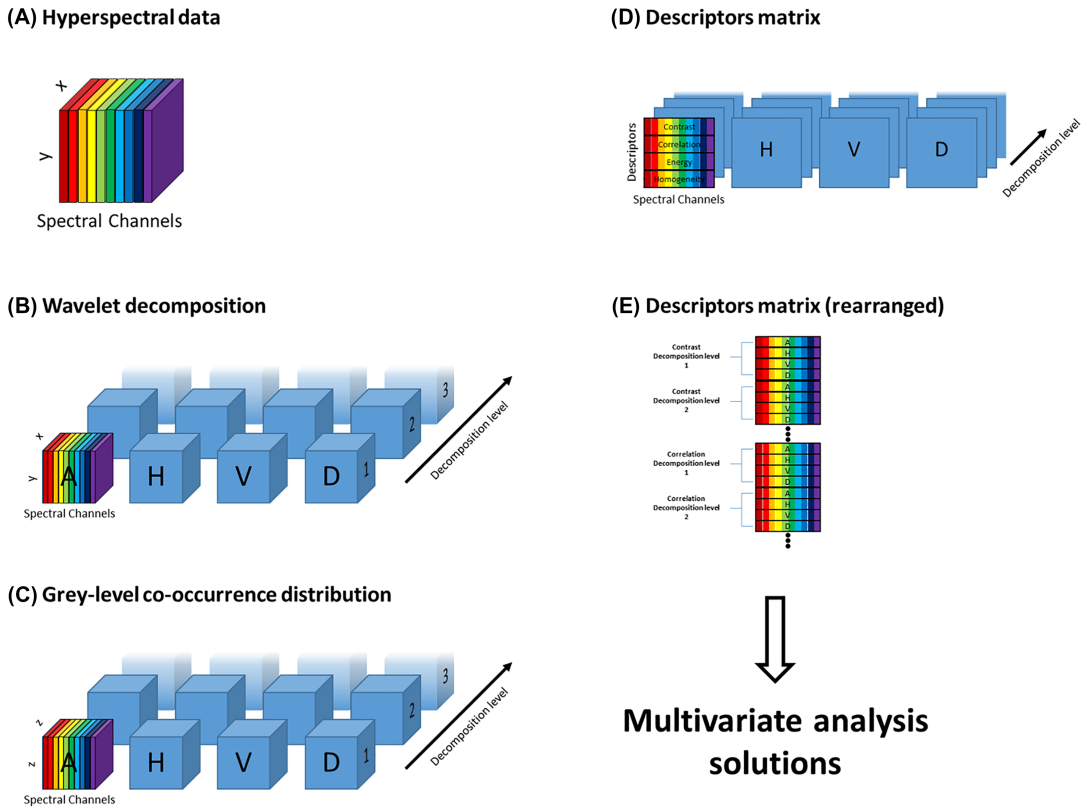


FIGURE 1 The methodological framework. The spectral dimension is colored after every step for the sake of a better and easier visualization. (A) A generic hyperspectral data cube; (B) the data structure obtained after the two-dimensional stationary wavelet transform (2D-SWT) decomposition; (C) gray-level co-occurrence matrix (GLCM) obtained from each slab of the wavelet coefficient three-dimensional arrays depicted in (B); (D) the descriptors calculated from the GLCM; and (E) rearrangement of the descriptor matrix

2.2 | Step 2: GLCM calculation

A GLCM is computed for each slab corresponding to a specific wavelet subimage, that is, for a given block of coefficients (H, V, D, and A) associated to a specific decomposition level and spectral wavelength (Figure 1C). A GLCM maps the local textures of a given image by counting how often pairs of pixels with certain normalized integer intensity values occur at a particular distance.[22,23] The type of normalization, and the direction along which the distance is calculated, needs to be set a priori. Here, we used a 64-integer intensity range and different directions for each wavelet subimage, matching the spatial pattern every wavelet decomposition image highlights: vertical direction for the horizontal coefficients image (the information retained after the decomposition, in fact, reflects the local spatial changes in that direction); horizontal for the vertical coefficients image; and top-left to bottom-right direction for the diagonal coefficients image and the summation of the previous directions for the approximation coefficients image.

2.3 | Step 3: Descriptors calculation

A set of eight descriptors (Energy, Contrast, Correlation, Variance, Inverse difference moment, Sum entropy, Information Measure of Correlation 1, and Maximal correlation coefficient[23]) is computed for each GLCM (see Figure 1D). A brief description of each of these features and their respective formulas are included in Table S1. These descriptors were chosen because they summarize most of the local spatial features one can find in an image. However, depending on the case study, distinct descriptors can be selected based on prior knowledge or on the necessity of specific image features to be highlighted.

2.4 | Step 4: Descriptor matrix rearrangement

All the morphological descriptor values estimated for every spectral channel are organized into individual column vectors subsequently gathered in a single data matrix. Thus, the descriptor matrix (see Figure 1E) features a number of rows equal to the number of descriptors (eight) times the number of decomposition levels (which depends on the size of the hyperspectral image) times the number of wavelet subimages per decomposition level (four). The number of columns corresponds to the number of spectral variables (wavelengths).

2.5 | Step 5: Multivariate analysis

The descriptor matrix is subjected to PCA for the exploration of the information it carries and, more specifically, for establishing a link between the spatial and spectral information captured by the variation of the morphological descriptors within the investigated spectral range. In this work, a possible pathway to establishing this link in a more systematic way is also explored, that is, the application of *k*-means clustering to the resulting PCA loadings to get an idea about the spectral channels associated to similar spatial features.

3 | RESULTS AND DISCUSSION

The aim of this communication is to show how local spatial features extracted with the use of the procedure outlined in the previous section can provide valuable information for HSI data analysis and exploration. For this purpose, the results of two case studies are presented.

3.1 | Oil-in-water emulsion

The first case study[24,25] relates to a Raman HSI dataset of an oil-in-water emulsion, which illustrates a situation where the spatial and spectral information are both somehow selective in their respective domains, that is, no severe overlap of the spatial and spectral features is observed. More specifically, the different individual chemical components of the image (featured in the spectral domain) are associated to clearly distinguishable shapes/spatial structures. This dataset is relatively simple and will serve as a proof of concept for the proposed methodology. The Raman imaging system by which these data were collected has a spatial resolution of around 1 μm, and the image is 60 × 60 pixels. The spectral resolution is 3.6 cm^{-1}, and the investigated spectral range goes from 953.6 to 1861 cm^{-1} (253 wavelengths). In Figure 2A,B, the mean image, averaged over all the spectral channels, and the mean spectrum, averaged over all the pixels, are shown, respectively.

We applied our approach considering eight descriptors (Section 2) and up to five decomposition levels. The outcomes resulting from the PCA modeling of the descriptor matrix are shown in Figure 2C,D. They display the scores and loading plots of the two first principal components, respectively. The score plot provides a graphical representation of the variation of the morphological descriptors across the wavelet subimages whereas the loading plot accounts for their variation across the spectral channels. As it can be assumed that the most extreme score values are the most informative, as recently pointed out in MCR context,[26] we will focus on the 10 most extreme scores values along PC_1 and PC_2 in Figure 2C. The corresponding points are labeled according to their respective descriptor name and to the wavelet subimage from which such a descriptor has been calculated. In Figure 2D, the loadings lying in the same quadrant along the direction determined by each one of these points and the origin of the PCA subspace (±9°) are highlighted accordingly (same colors/symbols). This way, 10 different groups of spectral channels were identified (loadings too close to the origin and, thus, contributing very little to the definition of PC_1 and PC_2 were excluded from this assessment). One wavelet subimage can be associated to each group which corresponds to one of the labeled descriptors in the score plot, as represented in Figure 3 (a maximum number of three wavelengths per group is considered here).

The comparative inspection of Figures 2 and 3 enables the simultaneous exploration of the spatial and spectral characteristics of the investigated HSI data. The loading plot gives insights into the spectral domain whereas the score plot together with the wavelet subimages does so for the spatial domain.

From Figure 3, overall, four main spatial contributions are discernible:

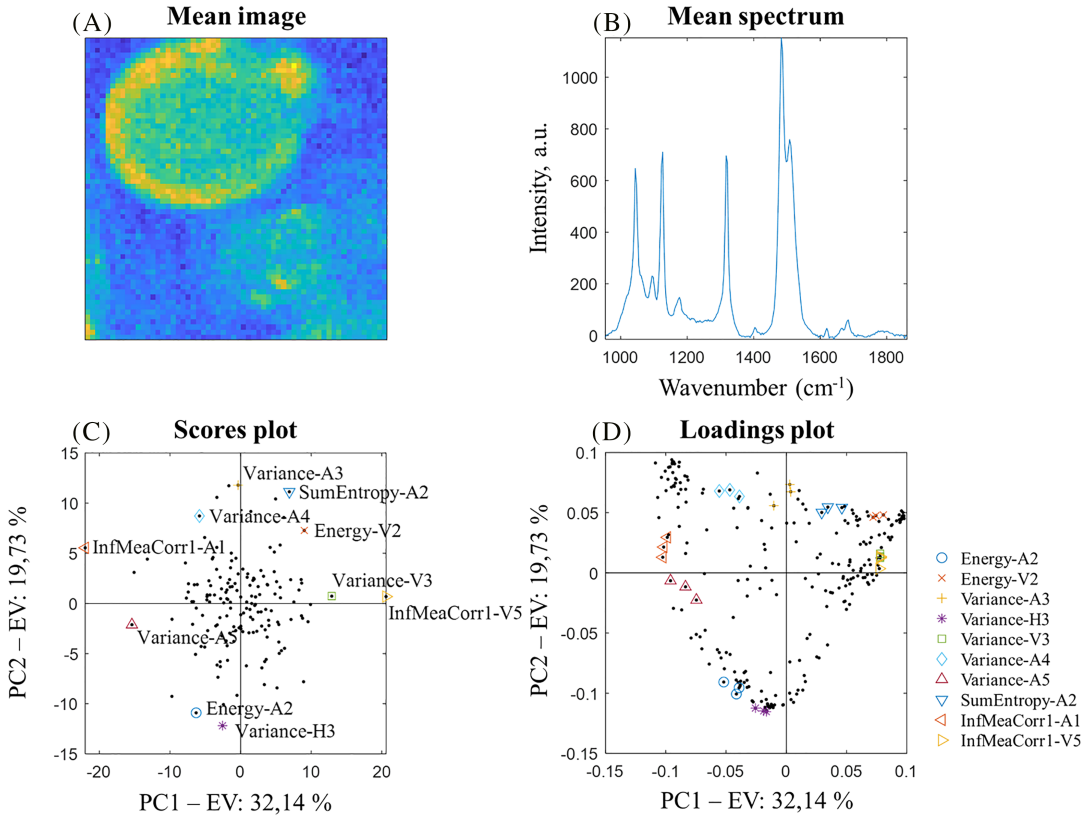


FIGURE 2 Oil in water dataset. (A) Mean hyperspectral image (mean taken across spectral dimension). (B) Mean spectrum (mean taken across pixels after unfolding). (C) PC_1 versus PC_2 score plot (principal component analysis [PCA] of descriptor matrix). The most extreme values in the scores space are highlighted by different colored symbols and labeled. (D) PC_1 versus PC_2 loading plot. The points in the loading plot that correspond (i.e., on the same direction with respect to origin) to the points highlighted in the score plot are shown with the same symbols

- two small droplets (see Figure 3A,D): the descriptors capturing these spatial features were Energy-A2 (i.e., energy calculated on the wavelet subimage corresponding to the approximations coefficients at decomposition level two) and Variance-H3;
- a larger droplet-like structure (see Figure 3C,H), associated to descriptors Variance-A3 and Sum-Entropy-A2;
- a circular border around the larger droplet-like structure (see Figures 3B,E), associated to descriptors Energy-V2 and Variance-V3. It was expected that the vertical details would be able to capture the border shape; and
- a background effect (see Figure 3G,J), associated to descriptors Variance-A5 and Information Mean Correlation-V5. Actually, the deepest decomposition level typically captures very smooth textural features. Overall, we may regard the fifth level of decomposition as the one that captures background and/or illumination effects.

On the other hand, Figure 3F,I seem to result from the overlap of some of these contributions.

To summarize, the proposed approach enabled to identify and distinguish four different components within the oil-in-water scene, spatially unraveled in the wavelet subimages (Figure 3) and spectrally identified in the loading plot (Figure 2D). These results are in good agreement with previous findings,[24,25] which discussed the interior droplet as due to oil and the border structure being the oil/water interface, while matching the smaller droplet to oil with a different composition. However, to complement the results of this preliminary visual inspection, *k*-means clustering was exploited. For this purpose, the descriptor matrix was compressed by PCA (17 PCs, explaining 90.12% variance), and clustering was applied on the estimated PCA loadings (four clusters of wavelengths were retrieved).

As highlighted in Figure 4C, averaging the hyperspectral image across the four extracted clusters of spectral channels led to the isolation of the different structures previously unraveled. It is then clear that for the data at hand, contributions showing a distinctive spatial distribution are also associated to rather selective spectral signatures within different wavelength ranges. These spectral signatures can be inspected in Figure 4A:

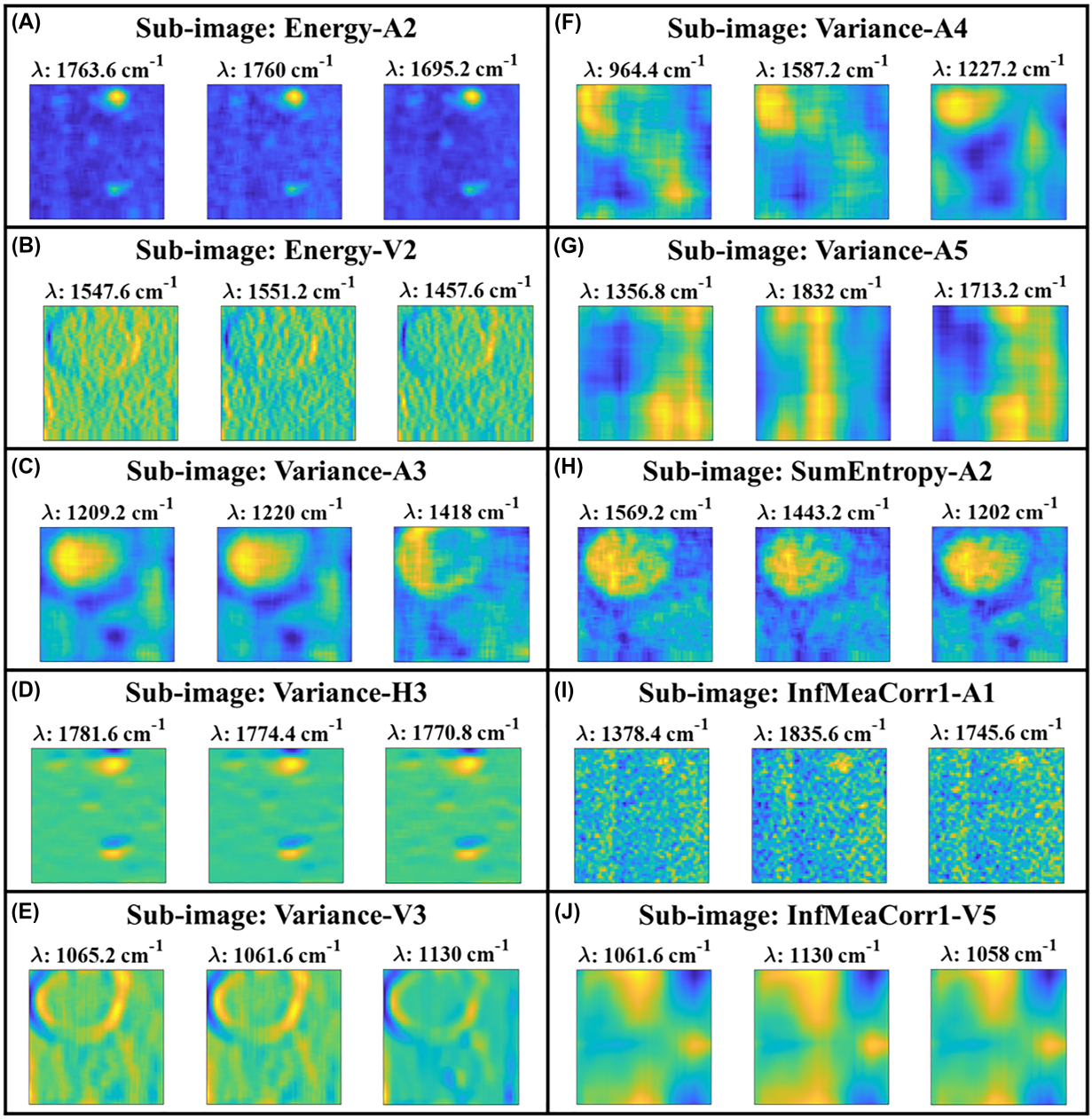


FIGURE 3 Oil in water dataset. The wavelet subimages, at decomposition block and level as reported in the points label in Figure 2C and at the spectral wavelengths corresponding to the ones showing the same symbols in Figure 2D, are shown in separate frames, labeled (A)–(J). The name of the descriptor-block level is reported on top of each frame; for example, the approximation images at the second decomposition level for the three wavenumbers 1695.2, 1760, and 1763.6 cm^{-1} are shown in Figure 3A, and so on for Figure 3B–J. The corresponding wavenumber is reported above each image

- Cluster #1, which is associated to the small droplets, coincides with the minor peaks at 1300, 1684, and 1789 cm^{-1}.
- Cluster #2, which is associated to the border structure, corresponds to three major peaks at 1044, 1126, and 1317 cm^{-1}.
- Cluster #3 mainly encompasses the peaks at 1483 and 1508 cm^{-1}. It corresponds to the interior droplet.
- Cluster #4 corresponds to the baseline regions observed in the mean spectrum.

A point to take note of is the overall correspondence of the spectral channels belonging to the groups identified by exploratory PCA of the descriptor matrix with the clusters found through *k*-means, as can be seen by comparing Figures 2D and 4B. In addition, the spatial features revealed by the wavelet subimages in Figure 3 mostly match those highlighted in the clustered mean images in Figure 4C.

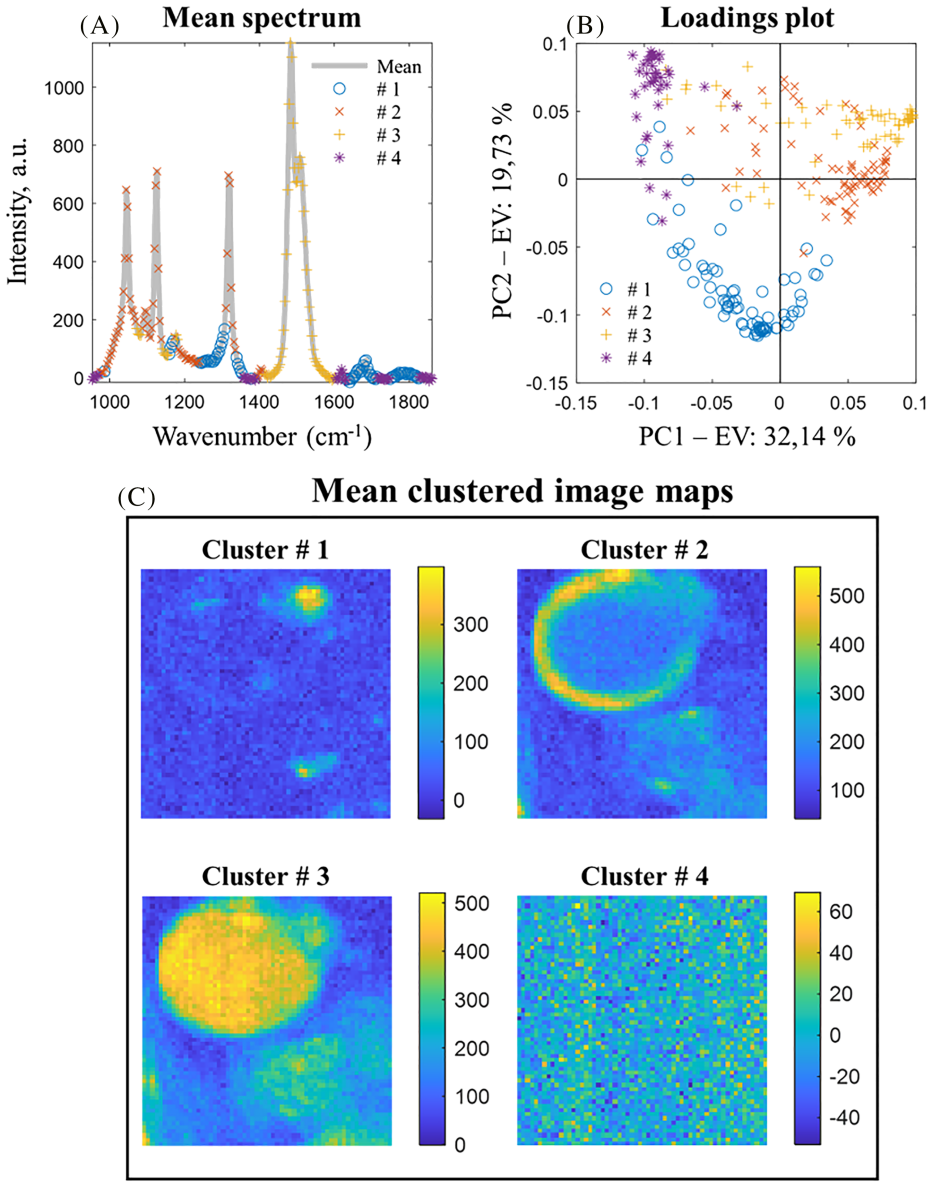


FIGURE 4 Oil in water dataset. Results of the *k*-means clustering algorithm on the loadings matrix. (A) Mean spectrum, with the point at each wavelength colored according to the cluster's number they were assigned to; (B) PC_1 versus PC_2 loading plot, points colored according to clusters; and (C) mean images for each cluster, mean taken across the spectral channels belonging to the same cluster

Due to the straightforward nature of the results, the dataset has been used to assess the relevance of each individual step of the proposed workflow (Figure 1). Analysis have been performed by removing some of these steps, that is, applying *k*-means on the PCA of unfolded HSI data or excluding the wavelet decomposition step and carrying out the GLCM and descriptors calculations directly on the raw images. The obtained results (Figure S1) showed that the information obtained was less significant than when applying the full original workflow.

3.2 | Semen droplet on cotton tissue

The second case study regards a 222 × 220 pixels HSI-near infrared (NIR) image (acquired in the wavelength range 1268.8–2456.2 nm with a spectral resolution of 6.3 nm) of a semen droplet on a piece of white cotton. Further details on the data acquisition are given in Silva et al.[17] The mean image is represented in Figure 5A. A quite complex structure is observed which is characterized, at least at a first glance, by (i) a distinct horizontal pattern across the entire image that is due to the rough surface of the cotton fabric (texture); (ii) an almost indiscernible shadow of the oval-shaped border of the semen droplet; and (iii) a spurious fiber filament in the lower middle area of the image.

It is worth noting that this case study exhibits a much higher complexity than the previous one: the cotton contribution is present everywhere across the image; thus, there exists no spatial area selective for semen. In addition,

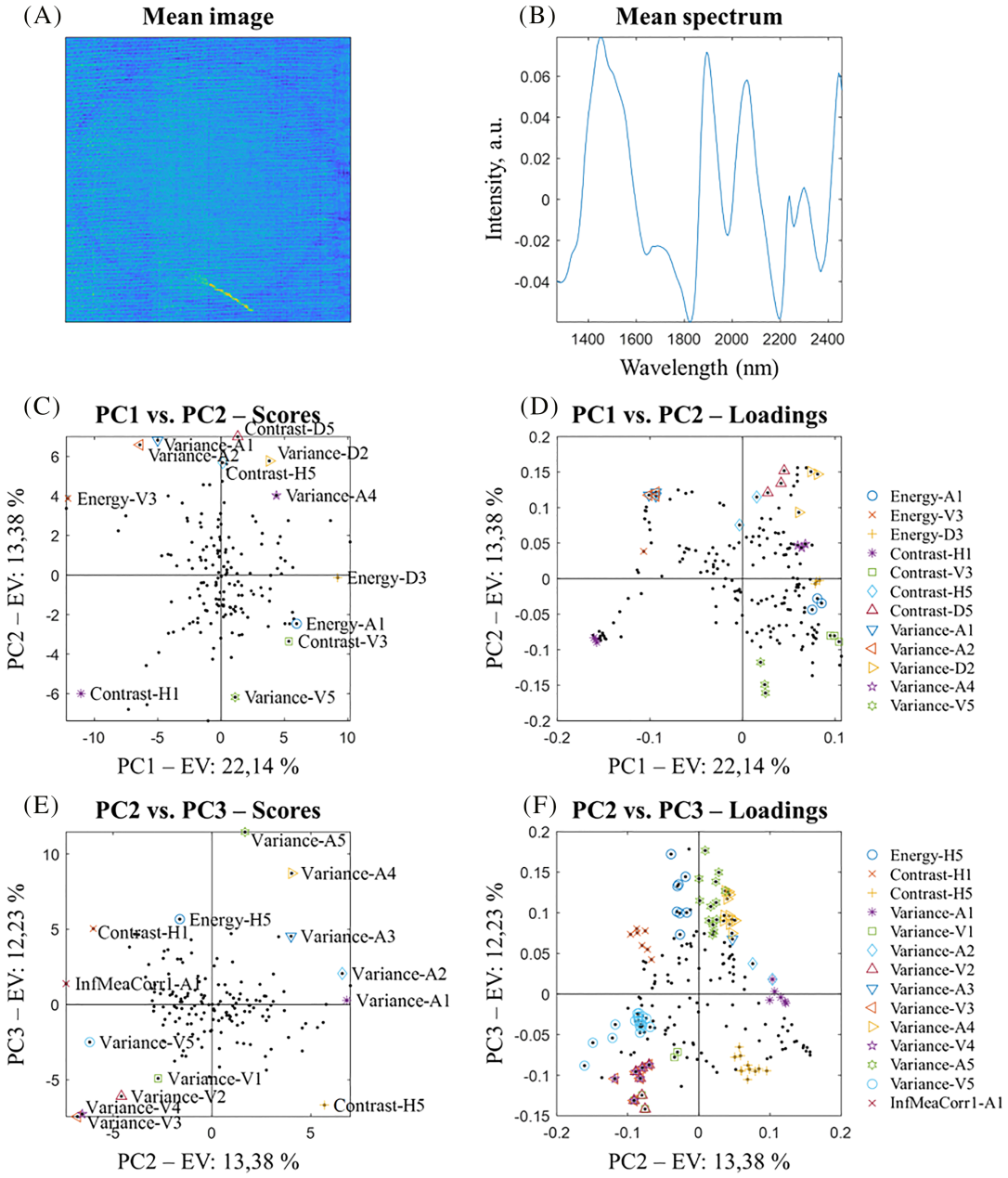


FIGURE 5 Semen dataset. (A) Mean image; (B) mean spectrum; and (C) PC_1 versus PC_2 score plot resulting from the application of principal component analysis (PCA) on the descriptor matrix (the 10 most extreme values are labeled with the descriptor's name and the wavelet subimage on which it has been calculated). (D) PC_1 versus PC_2 loading plot. The points highlighted in the score plot are shown with the same symbols. (E) PC_2 versus PC_3 score plot and (F) PC_2 versus PC_3 loading plot

the spectral profiles of the different constituents of the captured scene are severely overlapped and semen might be not homogeneously distributed over the cotton sample.

In order to explore these data, we applied our approach as detailed in Section 2.

The three first PCs of the descriptor matrix were inspected here. Figure 5 displays the resulting outcomes. Different clusters of wavelength channels were identified within both the PC_1/PC_2 and PC_2/PC_3 subspaces. The 10 most extreme score values in Figure 5C,E were taken into consideration, following the same procedure outlined for the emulsion dataset. For the sake of simplicity, not all the corresponding points in the loading plot (Figures 5D and 5E) were considered to extract the related wavelet subimages, but all were investigated. Among them, only those associated to the wavelet subimages showing easily interpretable spatial patterns were isolated. These wavelet subimages are shown in Figure 6.

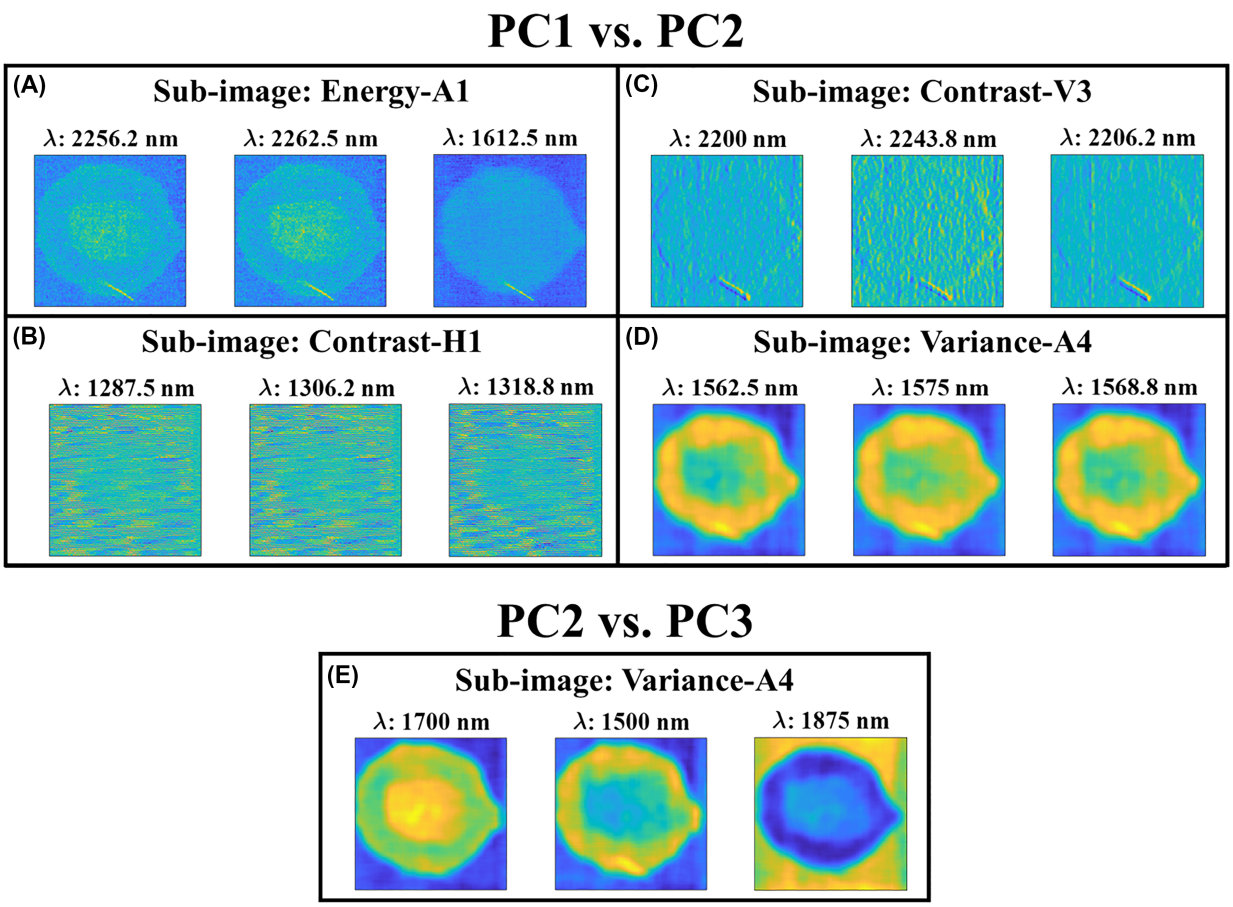


FIGURE 6 Semen dataset. The selected wavelet subimages from principal component analysis (PCA) of descriptor matrix, shown in separate frames labeled (A)–(E). For example, the approximation images at the first decomposition level for the three wavelengths 2256.2, 2262.5, and 1612.5 nm are shown in Figure 6A, and so on for Figure 6B–E

The images shown in Figure 6 can be categorized into three groups, each showing one of the main spatial contributions underlying this dataset:

- The semen stain associated to descriptors Energy-A1 (Figure 6A) and Variance-A4 (Figure 6D,E). Two subgroups seem to be present, as two spatially distinct forms of the stain seem to exist. In Figure 6E, for the images taken at 1700 and 1500 nm, the two forms are apparent. This is in good agreement with previous findings showing how the spatial distribution observed might be generated by complementary semen compounds exhibiting a distinct behavior during the drying process of the biological fluid on the cotton fabric.[17]
- The background (Figure 6B) captured by the descriptor Contrast-H1. This textural pattern that is observed across the entire image is most likely caused by the reflection of light on the cotton fibers.
  The corresponding loadings (purple stars in Figure 5D) are associated to wavelengths 1287.5, 1306.2, and 1318.8 nm. These are located in the range where cotton absorbs (1268.8–1362.5 nm)[27] and would most likely be related to the second overtones and combination of C–H stretching and C–H deformation.
- The fiber filament (Figure 6A,C) captured by descriptor Contrast-V3. The associated spectral wavelengths are in the range between 2000 and 2243.8 nm, where O–H bending and C–O stretching contributions are expected from cotton.

Notice that the fiber filament and the semen stain appear to be slightly overlapped in various wavelet subimages. Another important point to consider is that the best separation of the two semen stain forms resulted from the A4 wavelet subimage (Figure 6E). This highlights the fact that wavelet decompositions can provide a much greater spatial resolution, as they have the ability to unravel distinct spatial structures. This is observed in Figure 6B,D, where the separation between semen and the horizontal spatial interference is evident (such a separation cannot be visualized in any of the single channel images of the original HSI data, not shown). Considering the isolation of the horizontal details

(Figure 6B), the different forms of the semen stain (Figure 6E), the fiber (Figure 6A,C), and the "mask" that excludes the semen stain (Figure 6E, see at 1875 nm), the wavelet subimages feature a high potential for further investigation of the data. For example, considering methods such as MCR–ALS, on one hand, these isolated images can furnish the number of components to use and on the other hand can serve as initial estimates and could, for example, increase the spatial unmixing capability of the current methodology.[28] However, this could come with some ambiguity, as with higher decomposition levels, the wavelet subimages will only retain low frequency signals. This has to be considered carefully and will be explored in a future publication.

In order to corroborate the conclusion drawn after this visual inspection, *k*-means was applied, as previously explained (in this case, the descriptor matrix was compressed to 18 PC models, explaining 90.61% variance).

Figure 7 represents the obtained results. A more ambiguous clustering was obtained here compared with the previous case study, most likely due to the increased complexity and spatial overlap of the different components underlying the HSI dataset. Yet the previously determined components are discernable in Figure 7D. However, they are considerably more mixed. For example, "Cluster #1" encompasses both the fiber and (to a lesser extent) the semen stain. Also, in "Cluster #3" and "Cluster #4," the semen stain, the fiber, and the background are not completely separated from one another. Nonetheless, despite being mixed with the textural background, the two different spatial forms of semen were isolated in Clusters #2 and #3.

The spectral channels corresponding to Clusters #2 and #3 (Figure 7A) include the wavelengths regions around 1700 nm, and in the range 1500–1575 nm, which are selective for the two forms of semen when wavelet subimage A4 is considered (Figure 6D,E). It must be emphasized that compared with the emulsion dataset, the results of the clustering are not satisfactory in terms of isolation of the different components (Figure 7D). However, it has been shown

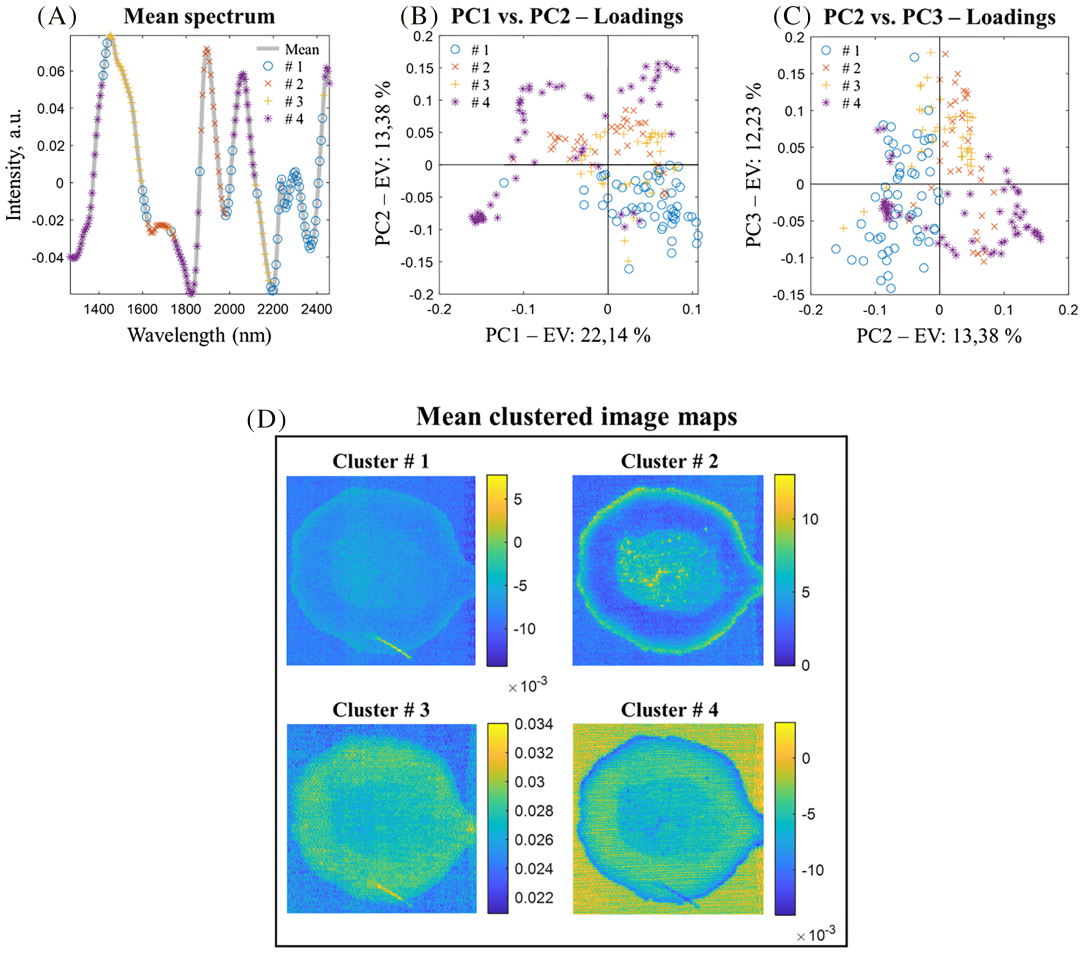


FIGURE 7 Semen dataset. Results of the *k*-means clustering algorithm on the loadings matrix. (A) Mean spectrum, with clustering highlighted on the spectral channels; (B) PC_1 versus PC_2 loading plot of points colored according to clusters; (C) PC_2 versus PC_3 loading plot; and (D) mean image (across the spectral dimension) for each cluster

(in Figure 6) that the wavelet subimages estimated at specific spectral wavelengths and recovered by the PCA of the descriptor matrix do have the ability to isolate the different components.

4 | CONCLUDING REMARKS

In this short communication, a methodological framework based on combining both spatial features and spectral information for the analysis of HSI data is proposed. The method decomposes every single-wavelength image of a three-dimensional HSI array by 2D-SWT and computes individual GLCM for every resulting wavelet subimage. Morphological descriptors are afterwards estimated from all the GLCMs. In this way, spatial and spectral information is enhanced and conveyed in a single features matrix, which is finally processed by multivariate data analysis tools like PCA. Depending on the specific tasks the user must address, different multivariate statistical tools can be exploited at this point.

Although the proposed workflow combines different computational steps, which translates into higher complexity, every one of them is a necessary link in the chain. In fact, when some of these steps were skipped, the information obtained was insufficient to unravel all the distinct spatial features underlying the dataset under study and relate them to specific spectral regions.

According to the results obtained in two different case studies, it can be concluded that the proposed strategy is capable of consistently recovering the main spatial features of a HSI dataset and of highlighting the distinctive spectral regions accounting for them. In particular, the outcomes related to the investigation of the semen droplet dataset were found to be particularly promising considering the extreme physicochemical complexity of the examined image. In fact, even though the semen and cotton components show highly overlapped spectra, and cotton fabric is present everywhere in the sample, it was possible to highlight and localize the two different forms of the semen stain as well as the spectral wavelengths at which this spatial separation is effective.

Nonetheless, further developments can be foreseen. GLCM is just one of the possible ways to compress the spatial information encoded in wavelet subimages, and others will be explored in future research. The possibility of utilizing other multivariate data analysis tools in the developed framework will also be assessed. Finally, the improved and at least preliminarily disentangled spatial/spectral information returned by the described approach might constitute a valuable starting point for the design of new constraints to be applied in the context of MCR–ALS. Additional work is currently in progress towards this direction.

ACKNOWLEDGMENTS

Dr. C.S. Silva acknowledges financial support from NUQAAPE-FACEPE (APQ-0346-1.06/14), Núcleo de Estudos em Química Forense (NEQUIFOR; CAPES AUXPE 3509/2014, Call PROFORENSE 2014), and FACEPE (BFP-0800-1.06/17 and APQ-0576-1.06/17).

ORCID

Raffaele Vitale https://orcid.org/0000-0002-7497-1673
Carolina S. Silva https://orcid.org/0000-0003-3868-3233
Cyril Ruckebusch https://orcid.org/0000-0001-8120-4133
Marina Cocchi https://orcid.org/0000-0001-8764-4981

REFERENCES

1. Guolan L, Baowei F. Medical hyperspectral imaging: a review. *J Biomed Opt*. 2014;19(1):010901-1-010901-23.
2. Gowenl AA, O'Donnell CP, Cullen PJ, Downey G, Frias JM. Hyperspectral imaging—an emerging process analytical tool for food quality and safety control. *Trends in Food Sci Tech*. 2007;18(12):590-598.
3. Liang H. Advances in multispectral and hyperspectral imaging for archaeology and art conservation. *Appl Phys A*. 2012;106(2):309-323.
4. Goetz AFH, Curtiss B. Hyperspectral imaging of the earth: remote analytical chemistry in an uncontrolled environment. *Field Anal Chem Tech*. 1996;1(2):67-76.
5. De Juan A, Tauler R, Dyson R, Marcolli C, Rault M, Maeder M. Spectroscopic imaging and chemometrics: a powerful combination for global and local sample analysis. *Tr AC*. 2004;23(1):70-79.
6. Hugelier S, Devos O, Ruckebusch C. On the implementation of spatial constraints in multivariate curve resolution alternating least squares for hyperspectral image analysis. *J Chemometr*. 2015;29(10):557-561.
7. Bharati MH, Liu JJ, MacGregor JF. Image texture analysis: methods and comparisons. *Chemom Intel Lab Syst*. 2004;72(1):57-71.

8. Prats-Montalbán JM, de Juan A, Ferrer A. Multivariate image analysis: a review with applications. *Chemom Intel Lab Syst*. 2011;107(1):1-23.
9. Jamme F, Duponchel L. Neighbouring pixel data augmentation: a simple way to fuse spectral and spatial information for hyperspectral imaging data analysis. *J Chemometr*. 2017;31:e2882. https://doi.org/10.1002/cem.2882
10. Liu J, MacGregor J. On the extraction of spectral and spatial information from images. *Chemom Intel Lab Syst*. 2007;85(1):119-130.
11. Li Vigni M, Prats-Montalbán JM, Ferrer A, Cocchi M. Coupling 2D-wavelet decomposition and multivariate image analysis (2D WT-MIA). *J Chemometr*. 2018;32:e2970. https://doi.org/10.1002/cem.2970
12. Gosselin R, Rodrigue D, Gonzalez-Nunez R, Duchesne C. Potential of hyperspectral imaging for quality control of polymer blend films. *Ind Eng Chem Res*. 2009;48(6):3033-3042.
13. Jacques K, Ruckebusch C. Model-based co-clustering for hyperspectral images. *J Spectral Imaging*. 2016;5:1-6. https://doi.org/10.1255/jsi.2016.a3
14. Xu JL, Gowen A. Spatial-spectral analysis method using texture features combined with PCA for information extraction in hyperspectral images. *J Chemometr*. 2019:e3132. https://doi.org/10.1002/cem.3132
15. Guo X, Huang X, Zhang L. Three-dimensional wavelet texture feature extraction and classification for multi/hyperspectral imagery. *IEEE Geosci Remote Sens Lett*. 2014;11(12):2183-2187.
16. Beauchemin M. Spatial pattern discovery for hyperspectral images based on multiresolution analysis. *Int J Image Data Fusion*. 2012;3(1):93-110.
17. Silva CS, Pimentel MF, Amigo JM, Honorato RS, Pasquini C. Detecting semen stains on fabrics using near infrared hyperspectral images and multivariate models. *Trends Anal Chem*. 2017;95:23-35.
18. Nason GP, Silverman BW. The stationary wavelet transform and some statistical applications. In: Antoniadis A, ed. *Wavelets and Statistics*. Lecture Notes in Statistics. New York: Springer-Verlag; 1995.
19. Juneau P, Garnier A, Duchesne C. The undecimated wavelet transform–multivariate image analysis (UWT-MIA) for simultaneous extraction of spectral and spatial information. *Chemom Intel Lab Syst*. 2015;142:304-318.
20. Mallat S. A theory for multi-resolution signal decomposition: the wavelet representation. *IEEE Trans Pattern Anal Mach Intell*. 1989;11(7):674-693.
21. Prats-Montalbán JM, Ferrer A, Cocchi M. *N*-way modeling for wavelet filter determination in multivariate image analysis. *J Chemometr*. 2015;29(6):379-388.
22. Haralick RM. Statistical and structural approaches to texture. *Proc IEEE*. 1979;67(5):780-803.
23. Haralick RM, Shanmugan K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern*. 1973;3(6):610-621.
24. Andrew JJ, Browne MA, Clark IE, Hancewicz TM, Millichope AJ. Raman imaging of emulsion systems. *Appl Spectrosc*. 1998;52(6):790-796.
25. De Juan A, Maeder M, Hancewicz T, Tauler R. Use of local rank-based spatial information for resolution of spectroscopic images. *J Chemometr*. 2008;22(5):291-298.
26. Ghaffari M, Omidikia N, Ruckebusch C. Essential spectral pixels for multivariate curve resolution of chemical images. *Anal Chem*. 2019;91(17):10943-10948.
27. Burns DA, Ciurczak EW. Chapter 25: Table 25.2. In: *Handbook of Near-Infrared Analysis*. III ed. Boca Raton, FL: CRC Press; 2008.
28. Vitale R, Hugelier S, Cevoli D, Ruckebusch C. A spatial constraint to model and extract texture components in Multivariate Curve Resolution of near-infrared hyperspectral images. *Anal Chim Acta*. 2020;1095:30-37.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Ahmad M, Vitale R, Silva CS, Ruckebusch C, Cocchi M. Exploring local spatial features in hyperspectral images. *Journal of Chemometrics*. 2020;e3295. https://doi.org/10.1002/cem.3295

Bibliography

[1] M. Ahmad, R. Vitale, M. Cocchi, and C. Ruckebusch. Weighted multivariate curve resolution – alternating least squares based on sample relevance, 2022.

[2] Mohamad Ahmad, Raffaele Vitale, Carolina S. Silva, Cyril Ruckebusch, and Marina Cocchi. 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). *Analytica Chimica Acta*, 1191, 1 2022.

[3] Mohamad Ahmad, Raffaele Vitale, Carolina S. Silva, Cyril Ruckebusch, and Marina Cocchi. Exploring local spatial features in hyperspectral images. *Journal of Chemometrics*, 34, 10 2020.

[4] G. Bergamini, M. Ahmad, M. Cocchi, and D. Malagoli. A new protocol of computer-assisted image analysis highlights the presence of hemocytes in the regenerating cephalic tentacles of adult pomacea canaliculata. *International Journal of Molecular Sciences*, 22, 2021.

[5] Xiangnan Dang, Neelkanth M. Bardhan, Jifa Qi, Li Gu, Ngozi A. Eze, Ching Wei Lin, Swati Kataria, Paula T. Hammond, and Angela M. Belcher. Deep-tissue optical imaging of near cellular-sized features. *Scientific Reports*, 9, 12 2019.

[6] Alexandra Witze. Four revelations from the Webb telescope about distant galaxies. *Nature*, 608(7921):18–19, 7 2022.

[7] Yu Winston Wang, Nicholas P. Reder, Soyoung Kang, Adam K. Glaser, and Jonathan T.C. Liu. Multiplexed Optical Imaging of Tumor-Directed Nanoparticles: A Review of Imaging Systems and Approaches. *Nanotheranostics*, 1(4):369–388, 2017.

[8] Pierre Bouguer. Essai d'optique sur la gradation de la lumière. *Nature*, 111(2784):320–320, 3 1923.

[9] Johann Heinrich Lambert. *Lambert's Photometrie, Vol. 1: Photometria, Sive de Mensura Et Gradibus Luminis, Colorum Et Umbrae (1760); Theil I Und II (Classic Reprint)*. Forgotten Books, 5 2018.

[10] Beer. Bestimmung der Absorption des rothen Lichts in farbigen Flüssigkeiten. *Annalen der Physik und Chemie*, 162(5):78–88, 1852.

[11] Gerrit Polder and Aoife Gowen. The hype in spectral imaging. *Journal of Spectral Imaging*, 2 2020.

[12] Muhammad Jaleed Khan, Hamid Saeed Khan, Adeel Yousaf, Khurram Khurshid, and Asad Abbas. Modern trends in hyperspectral image analysis: A review, 3 2018.

[13] Bing Lu, Phuong D. Dao, Jiangui Liu, Yuhong He, and Jiali Shang. Recent advances of hyperspectral imaging technology and applications in agriculture. *Remote Sensing*, 12(16), 2020.

[14] Cristina Malegori, Eugenio Alladio, Paolo Oliveri, Cristina Manis, Marco Vincenti, Paolo Garofano, Filippo Barni, and Andrea Berti. Identification of invisible biological traces in forensic evidences by hyperspectral nir imaging combined with chemometrics. *Talanta*, 215, 8 2020.

[15] Carolina S. Silva, Maria Fernanda Pimentel, José Manuel Amigo, Ricardo S. Honorato, and Celio Pasquini. Detecting semen stains on fabrics using near infrared hyperspectral images and multivariate models, 10 2017.

[16] Katharina Eberhardt, Clara Stiebing, Christian Matthäus, Michael Schmitt, and Jürgen Popp. Advantages and limitations of Raman spectroscopy for molecular diagnostics: an update. *Expert Review of Molecular Diagnostics*, 15(6):773–787, 4 2015.

[17] Zanyar Movasaghi, Shazza Rehman, and Ihtesham Ur Rehman. Fourier transform infrared (ftir) spectroscopy of biological tissues, 2008.

[18] P. Y. Sacré, C. De Bleye, P. F. Chavez, L. Netchacovitch, Ph Hubert, and E. Ziemons. Data processing of vibrational chemical imaging for pharmaceutical applications, 4 2014.

[19] Jeremy J. Andrew, Mark A. Browne, Ian E. Clark, Tom M. Hancewicz, and Allen J. Millichope. Raman Imaging of Emulsion Systems. *Applied Spectroscopy*, 52(6):790–796, 6 1998.

[20] Haida Liang. Advances in multispectral and hyperspectral imaging for archaeology and art conservation. *Applied Physics A: Materials Science and Processing*, 106:309–323, 2 2012.

[21] Chein-I Chang. *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. Springer, 2003 edition, 7 2003.

[22] Gamal ElMasry and Da-Wen Sun. Principles of Hyperspectral Imaging Technology. *Hyperspectral Imaging for Food Quality Analysis and Control*, pages 3–43, 2010.

[23] Asma Khan, M. T. Munir, W. Yu, and B. R. Young. A Review Towards Hyperspectral Imaging for Real-Time Quality Control of Food Products with an Illustrative Case Study of Milk Powder Production. *Food and Bioprocess Technology*, 13(5):739–752, 3 2020.

[24] Ryan Gosselin, Denis Rodrigue, Rubén Gonzál Núñ Ez, and Carl Duchesne. Potential of hyperspectral imaging for quality control of polymer blend films. *Industrial and Engineering Chemistry Research*, 48:3033–3042, 3 2009.

[25] Timo Zimmermann, Jens Rietdorf, and Rainer Pepperkok. Spectral imaging and its applications in live cell microscopy. *FEBS Letters*, 546(1):87–92, 5 2003.

[26] Barbara Boldrini, Waltraud Kessler, Karsten Rebner, and Rudolf W. Kessler. Hyperspectral imaging: A review of best practice, performance and pitfalls for in-line and on-line applications. *Journal of Near Infrared Spectroscopy*, 20:483–508, 10 2012.

[27] M. K. McClure, W. R. M. Rocha, K. M. Pontoppidan, N. Crouzet, L. E. U. Chu, E. Dartois, T. Lamberts, J. A. Noble, Y. J. Pendleton, G. Perotti, D. Qasim, M. G. Rachid, Z. L. Smith, Fengwu Sun,

Tracy L. Beck, A. C. A. Boogert, W. A. Brown, P. Caselli, S. B. Charnley, Herma M. Cuppen, H. Dickinson, M. N. Drozdovskaya, E. Egami, J. Erkal, H. Fraser, R. T. Garrod, D. Harsono, S. Ioppolo, I. Jiménez-Serra, M. Jin, J. K. Jørgensen, L. E. Kristensen, D. C. Lis, M. R. S. McCoustra, Brett A. McGuire, G. J. Melnick, Karin I. Öberg, M. E. Palumbo, T. Shimonishi, J. A. Sturm, E. F. van Dishoeck, and H. Linnartz. An Ice Age JWST inventory of dense molecular cloud ices. *Nature Astronomy*, 1 2023.

[28] Douglas MacLennan, Karen Trentelman, Yvonne Szafran, Anne T. Woollett, John K. Delaney, Koen Janssens, and Joris Dik. Rembrandt's ¡i¿An Old Man in Military Costume¡/i¿: Combining hyperspectral and MA-XRF imaging to understand how two paintings were painted on a single panel. *Journal of the American Institute for Conservation*, 58(1-2):54–68, 12 2018.

[29] Carmen Quintano, Alfonso Fernández-Manso, Yosio E. Shimabukuro, and Gabriel Pereira. Spectral unmixing, 2012.

[30] Magda K. Raczkowska, Paulina Koziol, Slawka Urbaniak-Wasik, Czeslawa Paluszkiewicz, Wojciech M. Kwiatek, and Tomasz P. Wrobel. Influence of denoising on classification results in the context of hyperspectral data: High Definition FT-IR imaging. *Analytica Chimica Acta*, 1085:39–47, 11 2019.

[31] Amir Bagheri Garmarudi, Mohammadreza Khanmohammadi, Hassan Ghafoori Fard, and Miguel de la Guardia. Origin based classification of crude oils by infrared spectrometry and chemometrics. *Fuel*, 236:1093–1099, 1 2019.

[32] Ya Juan Liu, Michelle Kyne, Cheng Wang, and Xi Yong Yu. Data mining in raman imaging in a cellular biological system, 1 2020.

[33] Anna de Juan, Joaquim Jaumot, and Romà Tauler. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Anal. Methods*, 6(14):4964–4976, 2014.

[34] Anna de Juan and Romà Tauler. Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – A review. *Analytica Chimica Acta*, 1145:59–78, 2 2021.

[35] Eirik Almklov Magnussen, Johanne Heitmann Solheim, Uladzislau Blazhko, Valeria Tafintseva, Kristin Tøndel, Kristian Hovde Liland, Simona Dzurendova, Volha Shapaval, Christophe Sandt, Ferenc Borondics, and Achim Kohler. Deep convolutional neural network recovers pure absorbance spectra from highly scatter-distorted spectra of cells. *Journal of Biophotonics*, 13(12), 9 2020.

[36] Waltraud Kessler, Dieter Oelkrug, and Rudolf Kessler. Using scattering and absorption spectra as mcr-hard model constraints for diffuse reflectance measurements of tablets. *Analytica Chimica Acta*, 642:127–134, 5 2009.

[37] Achim Kohler, Johanne Heitmann Solheim, Valeria Tafintseva, Boris Zimmermann, and Volha Shapaval. Model-based pre-processing in vibrational spectroscopy, 2020.

[38] Willem Windig, Jeremy Shaver, and Rasmus Bro. Loopy MSC: A Simple Way to Improve Multiplicative Scatter Correction. *Applied Spectroscopy*, 62(10):1153–1159, 10 2008.

[39] Paul Bassan and Peter Gardner*. Scattering in Biomedical Infrared Spectroscopy. *Biomedical Applications of Synchrotron Infrared Microspectroscopy*, pages 260–276, 12 2010.

[40] Paul Bassan, Achim Kohler, Harald Martens, Joe Lee, Hugh J. Byrne, Paul Dumas, Ehsan Gazi, Michael Brown, Noel Clarke, and Peter Gardner. Resonant mie scattering (rmies) correction of infrared spectra from highly scattering biological samples. *Analyst*, 135:268–277, 2010.

[41] Manuela Mancini, Giuseppe Toscano, and Åsmund Rinnan. Study of the scattering effects on nir data for the prediction of ash content using emsc correction factors. *Journal of Chemometrics*, 33, 4 2019.

[42] Puneet Mishra, Douglas N. Rutledge, Jean-Michel Roger, Khan Wali, and Haris Ahmad Khan. Chemometric pre-processing can negatively affect the performance of near-infrared spectroscopy models for fruit quality prediction. *Talanta*, 229:122303, 7 2021.

[43] Azeddine Beghdadi and Alain Le Negrate. Contrast enhancement technique based on local detection of edges. *Computer Vision, Graphics, and Image Processing*, 46(2):162–174, 5 1989.

[44] G. Maragatham and S. Mansoor Roomi. A Review of Image Contrast Enhancement Methods and Techniques. *Research Journal of Applied Sciences, Engineering and Technology*, 9(5):309–326, 2 2015.

[45] J. C. Gower. A Comparison of Some Methods of Cluster Analysis. *Biometrics*, 23(4):623, 12 1967.

[46] M. Tabb and N. Ahuja. Multiscale image segmentation by integrated edge and region detection. *IEEE Transactions on Image Processing*, 6(5):642–655, 5 1997.

[47] Shervan Fekri Ershad. Texture image analysis and texture classification methods-a review medical image classification view project surface defect detection view project, 2019.

[48] Sergey Kucheryavski. Extracting useful information from images. *Chemometrics and Intelligent Laboratory Systems*, 108:2–12, 8 2011.

[49] Matti Pietikäinen and Guoying Zhao. Two decades of local binary patterns: A survey, 1 2015.

[50] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC(6):610–621, 11 1973.

[51] Sugata Banerji, Atreyee Sinha, and Chengjun Liu. New image descriptors based on color, texture, shape, and wavelets for object and scene image classification. *Neurocomputing*, 117:173–185, 10 2013.

[52] Manish H. Bharati, J. Jay Liu, and John F. MacGregor. Image texture analysis: Methods and comparisons. *Chemometrics and Intelligent Laboratory Systems*, 72:57–71, 6 2004.

[53] Mryka Hall-Beyer. Practical guidelines for choosing glcm textures to use in landscape classification tasks over a range of moderate spatial scales. *International Journal of Remote Sensing*, 38:1312–1338, 3 2017.

[54] Harmeet Kaur and Satish Kumar. A Review on Decomposition-Reconstruction methods for Fusion of Medical Images. *International Research Journal on Advanced Science Hub*, 2(8):34–40, 8 2020.

[55] Chris Solomon and Toby Breckon. *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab*. Wiley, 1 edition, 1 2011.

[56] M. Heideman, D. Johnson, and C. Burrus. Gauss and the history of the fast fourier transform. *IEEE ASSP Magazine*, 1(4):14–21, 10 1984.

[57] Haris Ahmad Khan, Sofiane Mihoubi, Benjamin Mathon, Jean Baptiste Thomas, and Jon Yngve Hardeberg. Hytexila: High resolution visible and near infrared hyperspectral texture images. *Sensors (Switzerland)*, 18, 7 2018.

[58] Rasmus Bro and Age K. Smilde. Principal component analysis. *Anal. Methods*, 6(9):2812–2831, 2014.

[59] Adhemar Bultheel and Daan Huybrechs. Wavelets with applications in signal and image processing, 2010.

[60] Abdelkader Zitouni, Fatiha Benkouider, Fatima Chouireb, and Mohammed Belkheiri. Comparison Between Gabor Filters and Wavelets Transform for Classification of Textured Images. *Lecture Notes in Electrical Engineering*, pages 1021–1031, 9 2020.

[61] Tal Rapaport, Uri Hochberg, Maxim Shoshany, Arnon Karnieli, and Shimon Rachmilevitch. Combining leaf physiology, hyperspectral imaging and partial least squares-regression (PLS-R) for grapevine water status assessment. *ISPRS Journal of Photogrammetry and Remote Sensing*, 109:88–97, 11 2015.

[62] Frédéric Jamme and Ludovic Duponchel. Neighbouring pixel data augmentation: a simple way to fuse spectral and spatial information for hyperspectral imaging data analysis. *Journal of Chemometrics*, 31, 5 2017.

[63] J. M. Prats-Montalbán, A. de Juan, and A. Ferrer. Multivariate image analysis: A review with applications, 5 2011.

[64] Mario Li Vigni, José Manuel Prats-Montalban, Alberto Ferrer, and Marina Cocchi. Coupling 2d-wavelet decomposition and multivariate image analysis (2d wt-mia). *Journal of Chemometrics*, 32, 1 2018.

[65] Anna De Juan, Marcel Maeder, Thomas Hancewicz, and Romà Tauler. Use of local rank-based spatial information for resolution of spectroscopic images. *Journal of Chemometrics*, 22:291–298, 5 2008.

[66] Siewert Hugelier, Olivier Devos, and Cyril Ruckebusch. On the implementation of spatial constraints in multivariate curve resolution alternating least squares for hyperspectral image analysis. *Journal of Chemometrics*, 29(10):557–561, 8 2015.

[67] Raffaele Vitale, Siewert Hugelier, Dario Cevoli, and Cyril Ruckebusch. A spatial constraint to model and extract texture components in multivariate curve resolution of near-infrared hyperspectral images. *Analytica Chimica Acta*, 1095:30–37, 1 2020.

[68] Jun Li Xu and Aoife A. Gowen. Spatial-spectral analysis method using texture features combined with pca for information extraction in hyperspectral images. *Journal of Chemometrics*, 34, 2 2020.

[69] N. Gorretta, J.M. Roger, G. Rabatel, V. Bellon-Maurel, C. Fiorio, and C. Lelong. Hyperspectral image segmentation: The butterfly approach. *2009 First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 8 2009.

[70] N. Gorretta, G. Rabatel, C. Fiorio, C. Lelong, and J. M. Roger. An iterative hyperspectral image segmentation method using a cross analysis of spectral and spatial information. *Chemometrics and Intelligent Laboratory Systems*, 117:213–223, 8 2012.

[71] J. Jay Liu and John F. MacGregor. On the extraction of spectral and spatial information from images. *Chemometrics and Intelligent Laboratory Systems*, 85:119–130, 1 2007.

[72] Marco S. Reis. An integrated multiscale and multivariate image analysis framework for process monitoring of colour random textures: MSMIA. *Chemometrics and Intelligent Laboratory Systems*, 142:36–48, 3 2015.

[73] Pierre-Marc Juneau, Alain Garnier, and Carl Duchesne. The undecimated wavelet transform–multivariate image analysis (UWT-MIA) for simultaneous extraction of spectral and spatial information. *Chemometrics and Intelligent Laboratory Systems*, 142:304–318, 3 2015.

[74] Alessandro Nardecchia, Raffaele Vitale, and Ludovic Duponchel. Fusing spectral and spatial information with 2-D stationary wavelet transform (SWT 2-D) for a deeper exploration of spectroscopic images. *Talanta*, 224:121835, 3 2021.

[75] Xian Guo, Xin Huang, and Liangpei Zhang. Three-Dimensional Wavelet Texture Feature Extraction and Classification for Multi-Hyperspectral Imagery. *IEEE Geoscience and Remote Sensing Letters*, 11(12):2183–2187, 12 2014.

[76] Mario Beauchemin. Spatial pattern discovery for hyperspectral images based on multiresolution analysis. *International Journal of Image and Data Fusion*, 3(1):93–110, 3 2012.

[77] Seung Chul Yoon and Bosoon Park. Hyperspectral image processing methods, 2015.

[78] Maider Vidal and José Manuel Amigo. Pre-processing of hyperspectral images. essential steps before image analysis. *Chemometrics and Intelligent Laboratory Systems*, 117:138–148, 8 2012.

[79] Yu-hang Li, Xin Tan, Wei Zhang, Qing-bin Jiao, Yu-xing Xu, Hui Li, Yu-bo Zou, Lin Yang, and Yuan-peng Fang. Research and Application of Several Key Techniques in Hyperspectral Image Preprocessing. *Frontiers in Plant Science*, 12, 2 2021.

[80] Beata Walczak. Wavelets in chemistry. *Elsevier eBooks*, 1 2000.

[81] José Manuel Prats-Montalbán, Marina Cocchi, and Alberto Ferrer. N-way modeling for wavelet filter determination in multivariate image analysis. *Journal of Chemometrics*, 29:379–388, 6 2015.

[82] José Manuel Amigo, Arantxa del Olmo Alvarez, Merete Møller Engelsen, Henrik Lundkvist, and Søren Balling Engelsen. Staling of white wheat bread crumb and effect of maltogenic α-amylases. Part 1: Spatial distribution and kinetic modeling of hardness and resilience. *Food Chemistry*, 208:318–325, 10 2016.

[83] José Manuel Amigo, Arantxa del Olmo, Merete Møller Engelsen, Henrik Lundkvist, and Søren Balling Engelsen. Staling of white wheat bread crumb and effect of maltogenic α-amylases. Part 2: Monitoring

the staling process by using near infrared spectroscopy and chemometrics. *Food Chemistry*, 297:124946, 11 2019.

[84] José Manuel Amigo, Arantxa del Olmo, Merete Møller Engelsen, Henrik Lundkvist, and Søren Balling Engelsen. Staling of white wheat bread crumb and effect of maltogenic α-amylases. Part 3: Spatial evolution of bread staling with time by near infrared hyperspectral imaging. *Food Chemistry*, 353:129478, 8 2021.

[85] Emil W. Ciurczak, Benoît Igne, Jerome Workman Jr., and Donald A. Burns. *Handbook of Near-infrared Analysis*. CRC Press, 2021.

[86] I.T. Jolliffe. *Principal Component Analysis*. Springer, 12 2010.

[87] Willem. Windig and Jean. Guilment. Interactive self-modeling mixture analysis. *Analytical Chemistry*, 63(14):1425–1432, 7 1991.

[88] Abraham. Savitzky and M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639, 7 1964.

[89] Paul Eilers and Hans Boelens. Baseline correction with asymmetric least squares smoothing. *Unpubl. Manuscr*, 11 2005.

[90] Åsmund Rinnan, Frans van den Berg, and Søren Balling Engelsen. Review of the most common pre-processing techniques for near-infrared spectra, 11 2009.

[91] Ken-ichi Izutsu, Yukio Hiyama, Chikako Yomota, and Toru Kawanishi. Near-Infrared Analysis of Hydrogen-Bonding in Glass- and Rubber-State Amorphous Saccharide Solids. *AAPS PharmSciTech*, 10(2):524–529, 5 2009.

[92] Heinz Siesler, Yukihiro Ozaki, Satoshi Kawata, and Michael Heise. *Near-Infrared Spectroscopy*. Wiley, Hoboken, NJ, United States, 2002.

[93] Yukihiro Ozaki. *Applications in Chemistry*, chapter 9, pages 179–211. John Wiley and Sons, Ltd, 2001.

[94] Gonzalo Pajares and Jesús Manuel de la Cruz. A wavelet-based image fusion tutorial. *Pattern Recognition*, 37(9):1855–1872, 9 2004.

[95] J.F. Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. *IEE Proceedings F Radar and Signal Processing*, 140(6):362, 1993.

[96] Harpreet Kaur, Deepika Koundal, and Virender Kadyan. Image Fusion Techniques: A Survey. *Archives of Computational Methods in Engineering*, 28(7):4425–4447, 1 2021.

[97] Peter D Wentzell, Tobias K Karakach, Sushmita Roy, M Juanita Martinez, Christopher P Allen, and Margaret Werner-Washburne. Multivariate curve resolution of time course microarray data. *BMC Bioinformatics*, 7(1), 7 2006.

[98] Lionel Blanchet, Julien Réhault, Cyril Ruckebusch, Jean Pierre Huvenne, Romà Tauler, and Anna de Juan. Chemometrics description of measurement error structure: Study of an ultrafast absorption spectroscopy experiment. *Analytica Chimica Acta*, 642(1-2):19–26, 5 2009.

[99] Raffaele Vitale and Cyril Ruckebusch. On a ¡i¿black hole¡/i¿ effect in bilinear curve resolution based on least squares. *Journal of Chemometrics*, 10 2022.

[100] Mahdiyeh Ghaffari, Nematollah Omidikia, and Cyril Ruckebusch. Essential Spectral Pixels for Multivariate Curve Resolution of Chemical Images. *Analytical Chemistry*, 91(17):10943–10948, 7 2019.

[101] C. Ruckebusch, R. Vitale, M. Ghaffari, S. Hugelier, and N. Omidikia. Perspective on essential information in multivariate curve resolution. *TrAC Trends in Analytical Chemistry*, 132:116044, 11 2020.

[102] Laureen Coic, Pierre-Yves Sacré, Amandine Dispas, Charlotte De Bleye, Marianne Fillet, Cyril Ruckebusch, Philippe Hubert, and Éric Ziemons. Selection of essential spectra to improve the multivariate curve resolution of minor compounds in complex pharmaceutical formulations. *Analytica Chimica Acta*, 1198:339532, 3 2022.

[103] Alessandro Nardecchia and Ludovic Duponchel. Randomised SIMPLISMA: Using a dictionary of initial estimates for spectral unmixing in the framework of chemical imaging. *Talanta*, 217:121024, 9 2020.

[104] Mahdiyeh Ghaffari, Nematollah Omidikia, and Cyril Ruckebusch. Joint selection of essential pixels and essential variables across hyperspectral images. *Analytica Chimica Acta*, 1141:36–46, 1 2021.

[105] Jose Maria González-Martínez, Jose Camacho, and Alberto Ferrer. Bilinear modeling of batch processes. Part III: parameter stability. *Journal of Chemometrics*, 28(1):10–27, 11 2013.

[106] Franco Preparata and Michael Shamos. *Computational Geometry*. Springer Publishing, New York, United States, 2012.

[107] Rasmus Bro and Sijmen De Jong. A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics*, 11(5):393–401, 9 1997.

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

[109] Bradley Efron and R Tibshirani. *An Introduction to the Bootstrap (Chapman and Hall CRC Monographs on Statistics and Applied Probability)*. Chapman and Hall CRC, 1 edition, 1 1993.

[110] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 3 1982.