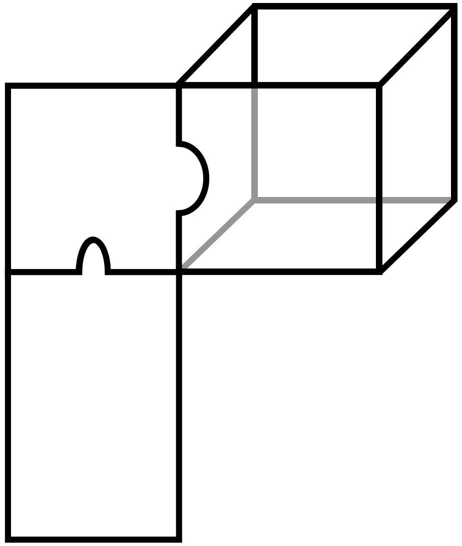
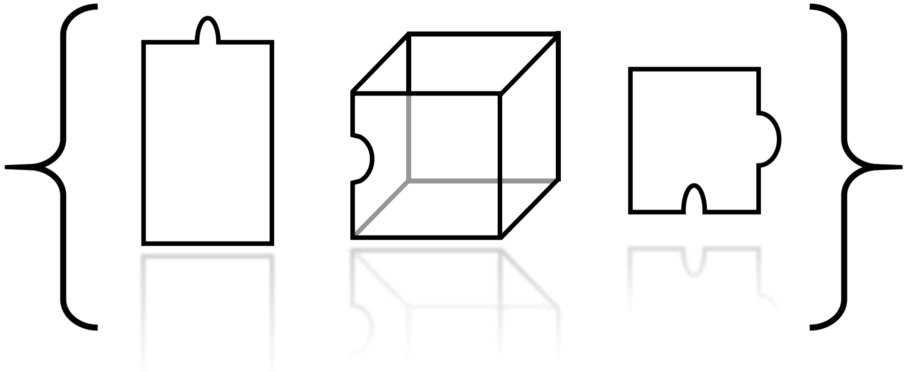


2024 PhD Thesis Adrián Gómez Sánchez

Development and application of new strategies for data fusion of hyperspectral images

Adrián Gómez Sánchez



UNIVERSITAT DE BARCELONA

Université de Lille

FACULTAT DE QUÍMICA
DEPARTAMENT D’ENGINYERIA QUÍMICA I QUÍMICA ANALÍTICA

Programa de doctorat: QUÍMICA ANALÍTICA i MEDI AMBIENT

PhD Thesis A. Gómez Sánchez

Memòria presentada per

Adrián Gómez Sánchez

Per optar al grau de Doctor per la Universitat de Barcelona
en: ***Química Analítica i Medi Ambient***

Directors

Anna de Juan Capdevila

Departament d’Enginyeria Química i Química Analítica

Universitat de Barcelona

Cyril Ruckebusch

Centre National de la Recherche Scientifique (CNRS)

Université de Lille

Université de Lille – Sciences et Technologies

École Doctorale des sciences de la Matière, du Rayonnement et de l'Environnement

LAboratoire de Spectroscopie pour les Interactions, la Réactivité et l'Environnement (LASIRE)

Thèse de Doctorat

En vue de l'obtention du grade de

Docteur de l'Université de Lille

Discipline: Optique et Lasers, Physico-Chimie et Atmosphère

Spécialité du doctorat: Chimie Theorique, Physique, Analytique

Adrián Gómez Sánchez

Development and application of new strategies for data fusion of hyperspectral images

Développement et application de nouvelles stratégies de fusion de données d'images hyperspectrales

Soutenue le 24 avril 2024:

Rapporteurs:

Itziar Ruisánchez (**Président du jury)**	Professeur, Universitat Rovira i Virgili.
Jean-Michel Roger	Professeur, ITAP-INRAE, Institute Agro, Université de Montpellier.

Examinateurs:

José Manuel Amigo	Professeur, Universidad del País Vasco/ Euskal Herriko Unibertsitatea.
Carmen Bedia	Chargé de recherche. Institute of Environmental Assessment and Water Research (IDAEA-CSIC).
Marie-Francoise Devaux	Chargé de recherche. INRAE - National Research Institute for Agriculture, Food and Environment.

Directeurs/trices de Thèse:

Anna de Juan	Professeur, Universitat de Barcelona.
Cyril Ruckebusch	Professeur, Université de Lille.

A mi família y amistades.

"No dejes para mañana lo que puedas hacer hoy,
excepto si es para depositar una tesis. El universo conspirará contra ti".

"He laughed"

Acknowledgements

Me gustaría agradecer, en primer lugar, a mi familia, por su constante apoyo, paciencia y cariño durante todos estos años, aunque todavía no comprendáis muy bien qué es exactamente lo que investigo, y me preguntéis cada poco, a lo que respondo: “Pues más o menos analizo imágenes de microscopio y hago algoritmos nuevos para eso, a ver con qué me encuentro”. Creo que después de presentar la tesis os haré una versión en castellano, así que id cogiendo sitio en el comedor, porque la turra que vendrá será importante.

Quiero que sepáis que, si no hubierais estado ahí siempre que lo he necesitado, dudo que estuvieses leyendo esto ahora mismo. Así que, gracias. Espero poder devolver todo, y un poco más, algún día. Sergio, no sé si podré darte una casa con piscina, pero bueno, arrimaré el hombro.

Quiero agradecer también a todas amistades, de aquí, de allí y de allá, que con el tiempo se han ido esparciendo por el mundo, que han vuelto o que han llegado. Aunque no quiero mencionar nombres (excepto el de Chus, al que voy a seguir agradeciendo cada vez que tenga oportunidad), vosotros sabéis muy bien lo importante que sois para mí, y la de cervezas que quedan por abrir. Algo me dice que ahora viene lo mejor.

También, en especial, quiero mencionar a mis directores de tesis, Anna y Cyril. Quiero que sepáis que nunca hubiese imaginado haber tenido la suerte de teneros como directores. La dedicación y paciencia que habéis tenido conmigo hace que saque pecho de directores cada vez que me preguntan por vosotros. Así que, tomad esta tesis también como vuestra. Al final me llevo, no solo a dos directores excelentes, sino a dos amigos.

A todos aquellos que han contribuido de alguna manera en esto, voy a estar eternamente agradecido.

CONTENTS

ABSTRACT.. **i**

RESUM.. **ii**

RESUMÉ.. **iii**

CHAPTER 1. OBJECTIVES AND STRUCTURE OF THE THESIS.................................. **1**

1.1 Objectives.. 3

1.2 Structure of the thesis.. 5

1.3 List of scientific publications presented in this work.. 6

CHAPTER 2. GENERAL INTRODUCTION.. **9**

2.1 Hyperspectral images: where chemistry reveals the composition and space provides the context.. 11

2.1.1 Raman hyperspectral images.. 15

2.1.2 Infrared hyperspectral images.. 18

2.1.3 Fluorescence hyperspectral images.. 21

2.2 Chemometric tools for the analysis of hyperspectral imaging……….......................... 29

2.2.1 Hyperspectral images: The underlying model of the measurement...............… 29

2.2.2 Unmixing methods. Multivariate Curve Resolution-Alternating Least Squares……...........……...........……...........……...........……...........……..... 33

2.2.3 Data fusion for hyperspectral images... 39

2.2.4 Challenges of image fusion.. 45

CHAPTER 3. RESULTS AND DISCUSSION.. **49**

SECTION I – Addressing challenges of fluorescence images.................................. **51**

3.1 Dealing with missing values on excitation-emission hyperspectral images............. 53

3.2 Improving the analysis of Fluorescence Lifetime Imaging Microscopy data............ 116

SECTION II – Addressing challenges of image fusion... **171**

3.3 Image fusion: A case study applied to vegetal tissues.. 173

3.4 Image fusion. Addressing differences in spectroscopic dimensionality.................... 201

3.5 Fusion of images showing differences in scanned area and spatial resolution......... 233

CONCLUSIONS.. **305**

REFERENCES.. **313**

Abstract

Hyperspectral images (HSIs) are unique analytical measurements that provide spatial and chemical information about samples. Each pixel of a HSI contains a spectroscopic measurement, representing the chemical information of the material present at that specific area. Nowadays, there are extremely diverse hyperspectral imaging platforms in terms of spatial resolution and spectroscopic modalities.

While the individual analysis of HSIs by chemometric methods provides comprehensive and rich chemical information about the nature of samples, the connection and complementary information among images remains too often unused and hidden. The integration and the simultaneous analysis of multiple HSIs in a single data structure or multiset, commonly referred as image fusion, offers a unique chemical multiscale perspective of the sample constituents.

However, image fusion scenarios can be particularly challenging when the images to be merged present differences in scanned areas, spatial resolution or spectral dimensionality. Addressing these problems calls for the development of algorithms extremely flexible and adaptable to the large diversity of data configurations and mathematical models required. Moreover, there is a special interest in improving the analysis of fluorescence images due to their unique chemical and mathematical properties. Image fusion incorporating fluorescence measurements, although challenging, results in a much more accurate characterization of systems.

This thesis proposes first novel algorithms to improve the analysis of excitation-emission fluorescence images and Time-resolved Fluorescence Spectroscopic data, the basic measurement in Fluorescence Lifetime Imaging (FLIM) measurements. These algorithms improve the unmixing processes and facilitate the extraction of crucial information from fluorescence signals.

On the other hand, the thesis provides an open-access protocol for multiplatform image fusion, adapted to handle differences in spatial resolution, scanned sample area and spectroscopic nature across different hyperspectral images. To do so, unmixing methodologies, notably Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS), have been adapted to accommodate simultaneously diverse underlying measurement models and to analyze data structures with missing blocks of information.

The proposed algorithms and methodologies offer a significant progress in the field of hyperspectral imaging analysis, enabling a more comprehensive and insightful understanding of samples across various scales.

Resum

Les imatges hiperespectrals (HSIs) són mesures analítiques que proporcionen informació espacial i química de les mostres. Cada píxel d'una HSI conté una mesura espectroscòpica, que representa la informació química del material present en aquella àrea específica. Avui dia, hi ha plataformes d'imatges hiperespectrals extremadament diverses pel que fa a la seva resolució especial I als modalitats espectroscòpiques que les defineixen.

Tot i que l'anàlisi de HSIs individuals mitjançant mètodes quimiomètrics proporciona informació química sobre la naturalesa de les mostres, la connexió i la informació complementària entre les imatges individuals roman sovint inexplorada. La integració i l'anàlisi simultània de múltiples HSIs en una única estructura de dades, coneguda com a fusió d'imatges, ofereix una perspectiva química multiescala única sobre els constituents de la mostra.

No obstant això, la fusió de dades d’HSIs presenta reptes importants quan les imatges que cal combinar presenten diferències pel que fa a l’àrea de mostra escanejada, a la resolució espacial o a la dimensionalitat espectral. A més, hi ha un interès especial en millorar l'anàlisi d'imatges de fluorescència degut a les seves particulars propietats químiques i matemàtiques. La fusió d’imatges que incorpora mesures de fluorescència és complexa, però proporciona una caracterització molt més acurada dels sistemes.

Aquesta tesi proposa, d'una banda, algorismes innovadors per millorar l'anàlisi d'imatges de fluorescència d’excitació-emissió i de dades de fluorescència procedents d’espectroscòpia resolta en el temps, que són la resposta instrumental associada a les imatges de temps de vida de fluorescència (FLIM). Aquests algorismes milloren els processos d’anàlisi de mescles i faciliten l'extracció d'informació crucial dels senyals de fluorescència.

D'altra banda, la tesi proporciona un protocol d'accés obert per a la fusió multiplataforma d'imatges, adaptat a la gestió de diferències de resolució especial, àrea escanejada i dimensionalitat spectral entre imatges. Per a tal fi, metodologies d’anàlisi de mescles, especialment el mètode de resolució multivariant de corbes per minims quadrats alternats (MCR-ALS), s’ha adaptat per incorporar simultàniament models diversos de descripció de la mesura d’imatge i per a l’anàlisi d’estructures amb blocs d’informació absent.

Els algorismes i metodologies proposats proporcionen un avenç significatiu en el camp de l'anàlisi d'imatges hiperespectrals i permeten una comprensió multiescala més completa i profunda sobre les característiques de les mostres.

Résumé

Les images hyperspectrales (HSI) sont des mesures analytiques qui fournissent des informations spatiales et chimiques sur les échantillons. Chaque pixel d'une HSI contient une mesure spectroscopique, représentant l'information chimique du matériau présent dans cette zone spécifique. De nos jours, il existe une grande diversité de plateformes d'imagerie hyperspectral en ce qui concerne leur résolution spatiale et les modalités spectroscopiques qui les définissent. Bien que l'analyse des HSI individuelles par des méthodes chimiométriques fournisse des informations chimiques sur la nature des échantillons, la connexion et les informations complémentaires entre les images individuelles restent souvent inexplorées. L'intégration et l'analyse simultanée de multiples HSI dans une seule structure de données, connue sous le nom de fusion d'images, offre une perspective chimique multi-échelle unique sur les constituants de l'échantillon. Cependant, la fusion de données HSI présente des défis importants lorsque les images à combiner présentent des différences en termes d'aire scannée, de résolution spatiale ou de dimensionnalité spectrale. De plus, il y a un intérêt particulier à améliorer l'analyse des images de fluorescence en raison de leurs propriétés chimiques et mathématiques particulières. La fusion d'images incorporant des mesures de fluorescence est complexe mais fournit une caractérisation beaucoup plus précise des systèmes. Cette thèse propose, d'une part, des algorithmes innovants pour améliorer l'analyse d'images de fluorescence d'excitation-émission et de données de fluorescence issues de spectroscopie résolue en temps, qui sont la réponse instrumentale associée aux images de temps de vie de fluorescence (FLIM). Ces algorithmes améliorent les méthodes de démélanges et facilitent l'extraction d'informations cruciales des signaux de fluorescence. D'autre part, la thèse propose un protocole d'accès ouvert pour la fusion multiplateforme d'images, adapté à la gestion des différences de résolution spatiale, d'aire scannée et de dimensionnalité spectrale entre images. Pour cela faire, des méthodologies de démélanges, notamment la méthode de résolution multivariée des courbes par moindres carrés alternés (MCR-ALS), ont été adaptées pour incorporer simultanément divers modèles de description de la mesure d'image et pour l'analyse de structures avec des blocs d'information manquante. Les algorithmes et les méthodologies proposés représentent une avancée significative dans le domaine de l'analyse d'images hyperspectrales et permettent une compréhension plus complète et approfondie à plusieurs échelles des caractéristiques des échantillons.

CHAPTER 1. OBJECTIVES AND STRUCTURE OF THE THESIS

1.1 Objectives

Hyperspectral images have emerged as powerful tools for the study and analysis of samples, attracting the attention of very different scientific fields. This interest has led to a great variety of hyperspectral images based on different spectroscopic techniques. The analysis of hyperspectral imaging data has two main scenarios: the analysis of individual images or the simultaneous analysis of several hyperspectral images, also defined as hyperspectral image fusion. The image fusion serves as a bridge for the datasets, connecting the valuable information among the images and providing a comprehensive chemical and spatial description of the samples.

Nowadays, two factors are boosting the image fusion field: the arrival of cutting-edge multimodal imaging platforms, providing different spectroscopic signals of the same sample, and the need to integrate outputs issued from different image platforms to improve the characterization of samples. Therefore, it is essential providing new chemometric tools able to deal with the specificities of imaging data coming from the large diversity of platforms differing in spatial resolution, scanned sample area and in the nature and dimensionality of spectroscopic information.

The main aim of this thesis is to address challenges on the image fusion field, by proposing chemometrics tools that can integrate and analyze data from diverse spectroscopic sources and modalities of hyperspectral images. To reach this goal, two specific objectives will be addressed:

<u>Addressing challenges of fluorescence image analysis</u>

Fluorescence measurements are extensively used due to the fast acquisition, high sensitivity and high spatial resolution associated with this measurement, but they present specific characteristics and problems different from most images. Due to its relevance and the specific challenges that arise from its analysis, an innovation on the analysis of fluorescence images is mandatory to allow the proposal of image fusion protocols that efficiently incorporate this kind of measurement. For this reason, the first section of results in this thesis is devoted to describe the improvements proposed for the fluorescence image analysis.

Within this scope, two main research objectives, related to the investigation of two different typologies of fluorescence images have been proposed, namely:

- ***The proposal of algorithms for the analysis of excitation-emission (EEM) fluorescence images***. This goal has addressed the proposal of flexible modified unmixing algorithms, based on Multivariate Curve

Resolution – Alternating Least Squares (MCR-ALS), to deal with EEM fluorescence images containing systematic patterns of missing entries. To do so, new implementations of the trilinear constraint in the presence of missing entries have been provided. These flexible algorithms adapt to handle individual EEM images, time-series of EEM images and can be used in image fusion scenarios.

- ***The proposal of algorithms for the analysis of Time-resolved Fluorescence Spectroscopic data (TRFS) and Fluorescence Lifetime images (FLIM).*** TRFS fluorescence decay signals are always convolved with the Instrumental Response Function (IRF), which hinders the interpretation of the information. A novel algorithm capable to extract the IRF from the solely fluorescence decay measured is proposed. Another challenge for TRFS and FLIM image analysis is linked to the unmixing analysis of fluorescence decay curves with few time bin channels. The novel kernelizing algorithm, based on a tensorization approach for the analysis of TRFS data is proposed to solve this problem.

Addressing challenges of image fusion

The challenges in image fusion are as diverse as the wide variety of hyperspectral images. Indeed, hyperspectral imaging measurements may present different spatial resolution, sample area scanned, spectral dimensionality and spectroscopic nature. This rich scenario gives rise to a multitude of ways to interconnect the datasets and, consequently, proposes very interesting challenges to be solved. The objectives addressed in this block are:

- ***The proposal of an open-access protocol for classical multiplatform image fusion***. The steps to carry out in the joint unmixing analysis of hyperspectral images with different spectroscopic nature is presented. Specifically, the image fusion of Raman, fluorescence and synchrotron infrared hyperspectral images is studied and its applicability tested on biological samples. This example sets the basic steps to be followed in image fusion and points out some of the limitations solved in the next objectives.

  Coping with differences of spectral dimensionality in image fusion. This is the usual image fusion scenario encountered when excitation-emission fluorescence images, which present a 2D EEM landscape per pixel, are coupled to any other hyperspectral image, which hold a 1D linear spectrum per pixel, e.g., a Raman image. The difference between the linear underlying model among the excitation-emission measurements (trilinear) and the Raman measurements (bilinear) poses a very interesting problem for data fusion. An innovation to improve the flexible implementation of the trilinear constraint in the algorithm MCR-ALS has allowed setting a hybrid bilinear/trilinear model to properly define the underlying behavior of the measurements involved in this image fusion problem. A real case of image fusion with excitation-emission and Raman hyperspectral images has illustrated the algorithm proposed.

- ***Coping with the presence of missing blocks of information in image fusion (fusion of different scanned sample areas and images with different spatial resolution)***. The problematic of missing entries arises very often in the context of image fusion when two or more hyperspectral images contain non-common scanned areas and/or different spatial resolutions among them. This gives datasets with a significant amount of missing entries, designed as incomplete multisets. The exploratory algorithm Principal Component Analysis (PCA) and the unmixing method MCR-ALS have been modified in order to deal with the presence of missing entries. The new PCA and MCR-ALS algorithms proposed adapt to large amounts of missing entries following the challenging missing block pattern, usual in image fusion problems.

1.2 Structure of the thesis

This thesis contains the research performed across nine papers, five of them already published, focusing on the development and implementation of innovative strategies for the analysis of fluorescence images and the fusion of hyperspectral data from different imaging platforms. The manuscript is organized in three chapters and a section of conclusions.

The first chapter presents the objectives and structure of the thesis and the list of derived publications.

The second chapter introduces hyperspectral imaging and provides technical and practical details on the techniques used in this thesis. These imaging techniques include Raman, synchrotron mid-infrared, near-infrared, fluorescence, and TRFS measurements, which will help readers to understand the underlying models discussed.

The section continues presenting the underlying models of the hyperspectral image measurements, well represented by bilinear and trilinear models. Afterwards, it includes the description of the unmixing algorithm Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) and the details related to its application in image analysis. The classical use of MCR-ALS to deal with the fusion of hyperspectral images, related to the concept of multiset analysis, is also presented. The chapter ends with an introduction to the challenges in image fusion, linked to unsolved issues in the analysis of multiset structures.

The third chapter presents the research findings and their discussion and is divided in two sections. The first section is related to the results associated with the study of fluorescence images, i.e., the proposals of algorithms for the analysis of EEM fluorescence images with missing entries and for TRFS data. The second section addresses the challenges of image fusion, containing the results for the traditional image fusion, the algorithm modified to handle image fusion among platforms with different spectral dimensionality and the incomplete multiset structures.

Finally, the conclusions extracted from this thesis are presented.

1.3 List of scientific publications presented in this thesis

The work performed in this thesis resulted in the nine scientific publications below, grouped by topics and following the sequence in the thesis manuscript.

Publication I. The trilinear constraint adapted to solve data with strong patterns of outlying observations or missing values.
Authors: A. Gómez-Sánchez, I. Alburquerque, P. Loza-Álvarez, C. Ruckebusch, A. de Juan.
Citation reference: *Chemometrics and Intelligent Laboratory Systems* (2022), 231:104692.
DOI: 10.1016/j.chemolab.2022.104692

Publication II. The MCR-ALS trilinearity constraint for data with missing values.
Authors: A. Gómez-Sánchez, R. Vitale, P. Loza-Ávarez, R. Tauler, C. Ruckebusch, A. de Juan.
Journal of Chemometrics (2024) (submitted).

Publication III. Study of the photobleaching phenomenon to optimize acquisition of 3D and 4D fluorescence images. A special scenario for trilinear and quadrilinear models.
Authors: A. Gómez-Sánchez, I. Alburquerque, P. Loza-Álvarez, C. Ruckebusch, A. de Juan.
Citation reference: *Microchemical Journal* (2023), 191:108899.
DOI: 10.1016/j.microc.2023.108899

Publication IV. Blind Instrument Response Function Identification (BIRFI) from Fluorescence Decays.
Authors: A. Gómez-Sánchez, O. Devos, R. Vitale, M. Sliwa, D. Sakhapo, J. Enderlein, A. de Juan, C. Ruckebusch.
Biophysical Reports (2024) (submitted).

Publication V. Kernelizing: A way to increase accuracy in trilinear decomposition analysis of multiexponential signals.
Authors: A. Gómez-Sánchez, R.Vitale, O. Devos, A. de Juan, C. Ruckebusch.
Citation reference: *Analytica Chimica Acta* (2023), 1273: 341545.
DOI: 10.1016/j.aca.2023.341545.

Publication VI. Linear unmixing protocol for hyperspectral image fusion analysis applied to a case study of vegetal tissues
Authors: A. Gómez-Sánchez, M. Marro, M. Marsal, S. Zacchetti, R. R. de Oliveira, P. Loza-Álvarez, A. de Juan.
Citation reference: *Scientific Reports* (2021), 11:18665.
DOI: 10.1038/s41598-021-98000-0

Publication VII. 3D and 4D image fusion: coping with differences in spectroscopic modes among hyperspectral images.
Authors: A. Gómez-Sánchez, M. Marro, M. Marsal, S. Zacchetti, R. R. de Oliveira, P. Loza-Álvarez, A. de Juan.
Citation reference: *Analytical Chemistry* (2020), 14:9591–9602
DOI: 10.1021/acs.analchem.0c00780

Publication VIII. Dealing with missing data blocks in Multivariate Curve Resolution. Towards a general framework based on a single factorization model.
Authors: A. Gómez-Sánchez, C. Ruckebusch, R. Tauler, A. de Juan.
Trends in Analytical Chemistry (2024) (submitted).

Publication IX. Solving the missing value problem in PCA by Orthogonalized-Alternating Least Squares (O-ALS)
Authors: A. Gómez-Sánchez, R. Vitale, C. Ruckebusch, A. de Juan.
Chemometrics and Intelligent Laboratory Systems (2024) (submitted).

CHAPTER 2. GENERAL INTRODUCTION

2.1 Hyperspectral images: where chemistry reveals the composition and space provides the context.

Traditional analytical chemistry has often been dominated by univariate measurements, which involve the use of a single parameter or data point, e.g., the use of the absorbance in a specific wavelength to quantify the concentration of a substance or a pH value to define the acidity of a solution. Univariate measurements have served and serve their purpose well, but the complexity of many chemical systems requires the incorporation of more comprehensive information during analyses.

Multivariate measurements have emerged as a solution for this problem, providing richer chemical information. In this context, spectroscopy has become a fundamental tool widely used for expanding analytical capabilities. In a multivariate context, spectroscopic techniques provide a full spectrum, defined as a vector, with each element representing the signal at its respective wavelength (see Fig. 1).

The use of full spectra instead of univariate measurements offers several benefits, such as an enhanced discrimination among chemical species, a reduction of interferences in analytical quantifications and a better description of complex samples, allowing a more advantageously understanding of the nature of chemical systems. Industry and research areas show plenty of applications of multivariate spectroscopic measurements using techniques, such as infrared [Wolfe and Zissis, 1978; Williams and Norris, 2001], Raman [Lewis and Edwards, 2001; Jones et al., 2019], or fluorescence [Lakowicz, 2006; Valeur and Berberan-Santos, 2012] among others.

The analytical power of multivariate spectroscopic techniques has gone a step further with the introduction of spatial information in the measurement. The spatial information has been exploited since long in the microscopy imaging field, which used for decades univariate fluorescence measurements to study biological samples, e.g., labelling specific compounds with fluorophores to locate them and study the structure of a tissue (see Fig. 1). The ability to locate a chemical compound in a sample provides crucial information in some instances. For example, in cancer research, a conventional microscopy image can reveal not only the presence of cancer cells but also their interactions with surrounding healthy tissue and the distribution of blood vessels [Nagy et al., 2009], which are often characteristic properties of cancerous tissues. Without the spatial information, these details remain hind.

From this perspective, both the *spatial* (related to the location of compounds) and the *spectral* (related to the chemistry of the compounds) information have converged into the field of hyperspectral imaging. Thus, a hyperspectral image (HSI) is formed by a large set of spectra associated with spatial areas (pixels) of the sample [Amigo et al., 2015], where the location of each pixel is defined by their spatial coordinates. A HSI can be displayed as a data structure with three dimensions: two spatial dimensions related to the pixel coordinates and sized (x, y) and one chemical dimension, defined by the spectral range covered, sized λ, forming a data cube sized (x, y, λ) (see Fig. 1).

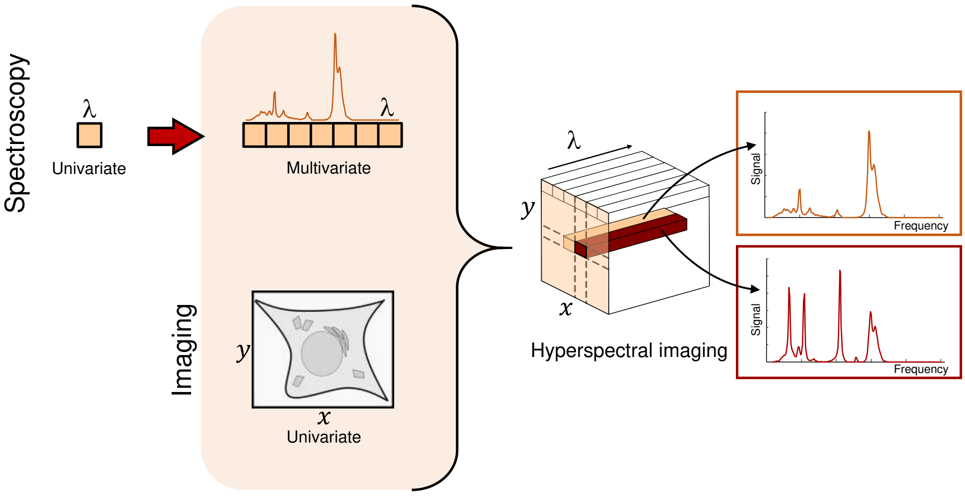


Figure 1. The complexity of the spectroscopic measurement increased from univariate to multivariate. Lately, the imaging and multivariate field converged to the hyperspectral imaging field.

The connection between spatial and chemical information has opened the door to a fascinating and diverse world of hyperspectral images. Nowadays, HSIs can be formed by pixels sized at several nanometers, as in fluorescence imaging techniques [Zavattini et al., 2003; Haaland et al., 2007; Huang et al., 2009], or by pixels covering vast areas spanning several square kilometers, as in remote sensing applications [Landgrebe, 1999]. The spectral dimension in hyperspectral images can be associated with almost all available spectroscopy techniques. It is possible to find HSIs that provide limited chemical information, such as fluorescence HSIs, where the spectra are not very selective, as well as others that provide very specific high-quality chemical information, like mass-spectrometry images [McDonnell and Heerem, 2007; Caprioli, 2015; Buchberger et al., 2018].

HSIs can be obtained using four different acquisition modes [Burger, 2006; Amigo, 2010; Qin et al., 2013; Adão et al., 2017; Amigo and Grassi, 2019], as displayed in Fig. 2.

- ***Point scanning***: The sample grid, formed by individual pixels, is systematically analyzed by directing the source of illumination (usually a beam laser) to each pixel one-by-one and recording the corresponding spectrum (see Fig. 2A). This process continues until the entire area of interest has been scanned. Usually, the sample is placed on a stage that controls the position of the sample on the grid and moves accurately in preset motion steps in the *x*- and *y*- directions. This method is typically used in applications like confocal fluorescence, Raman or infrared imaging. It offers high spatial resolution but the acquisition time can be relatively long [Griffiths, 2002].

- ***Line scanning***: This configuration acquires simultaneously a line of pixel spectra across the sample grid and continues until the entire area is covered (see Fig. 2B). In this acquisition mode, the detector has a grid of sensors (one per each pixel and wavelength) that simultaneously record signals for an entire line of pixels. Pushbroom imaging systems are commonly used for this acquisition mode, where the sample is placed on a stage that moves in the perpendicular direction to the spatial axis of the detector. This acquisition mode is faster than point scanning but presents lower spatial and spectral resolution. Remote sensing devices equipped with Vis-NIR or NIR imaging platforms often use pushbroom systems [Mangold et al., 2008; Gowen et al., 2008].

- ***Plane scanning (or focal plane array)***: In this configuration, the entire sample area is scanned in a single shot, but only the signal of a single spectral channel is acquired (see Fig. 2C), by a detector formed essentially by an array of sensors [Burger, 2006; Primpke et al., 2017]. To obtain the full image, the spectral direction is continuously scanned until all the specified spectral range is covered. This acquisition mode is faster than point line scanning but usually presents lower spatial resolution. This approach is very relevant for some applications, such as monitoring blending processes where the spatial distribution becomes essential [El-Hagrasy et al., 2001; Amigo, 2010], or for the study of living samples, where the chemical dynamics have to be recorded over time and, therefore, the time acquisition must be very fast.

- ***Snapshot acquisition***: It captures the entire hyperspectral data cube in a single shot. In this acquisition mode, there exists several configurations to capture the entire hyperspectral image at once [Hagen et al., 2012]. One of the most interesting advantages, besides the fast time acquisition, is the light collection efficiency of the measurement. While line scanning loses all photons coming from non-scanned lines (although they are illuminated) and plane scanning loses all photons coming from other wavelengths (the sensor scans only one wavelength at a time), the snapshot acquisition can capture efficiently a significant portion of these photons. This technology is still emerging and it seems that there are no available commercial devices yet.

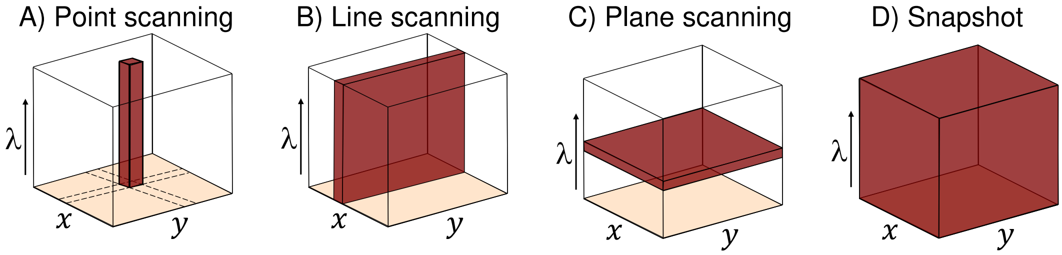


Figure 2. Different hyperspectral image acquisition modes. A) Point scanning. B) Line scanning. C) Plane scanning. D) Snapshot acquisition. The area in dark red represents the information acquired in every acquisition step.

The previous acquisition modes provide most HSIs, which can be displayed as a data cube with dimensions (x, y, λ).

However, HSIs can have more than three dimensions (see Fig. 3). For instance, HSI can have an extra spectral dimension (x, y, Λ, λ), like in excitation-emission fluorescence images, where a 2D excitation-emission fluorescence landscape is obtained per pixel [Appalaneni et al., 2014; Hruska et al., 2014; Omrani et al., 2014; Rodríguez-Vidal et al., 2020]. HSIs can also have an extra spatial dimension $(x, y, \mathrm{z}, \lambda)$, equivalent to a volumetric hyperspectral image [Mertz, 2019; Wen et al., 2020; Gualda et al., 2015]. To add still more diversity to HSIs, the spectral dimension of the image cube can be replaced by a time dimension, such as in the Fluorescence Lifetime Images (FLIM), where the signal collected per pixel is the decay of the emission fluorescence over time. In the following sections, the hyperspectral images used in this thesis will be described in detail.

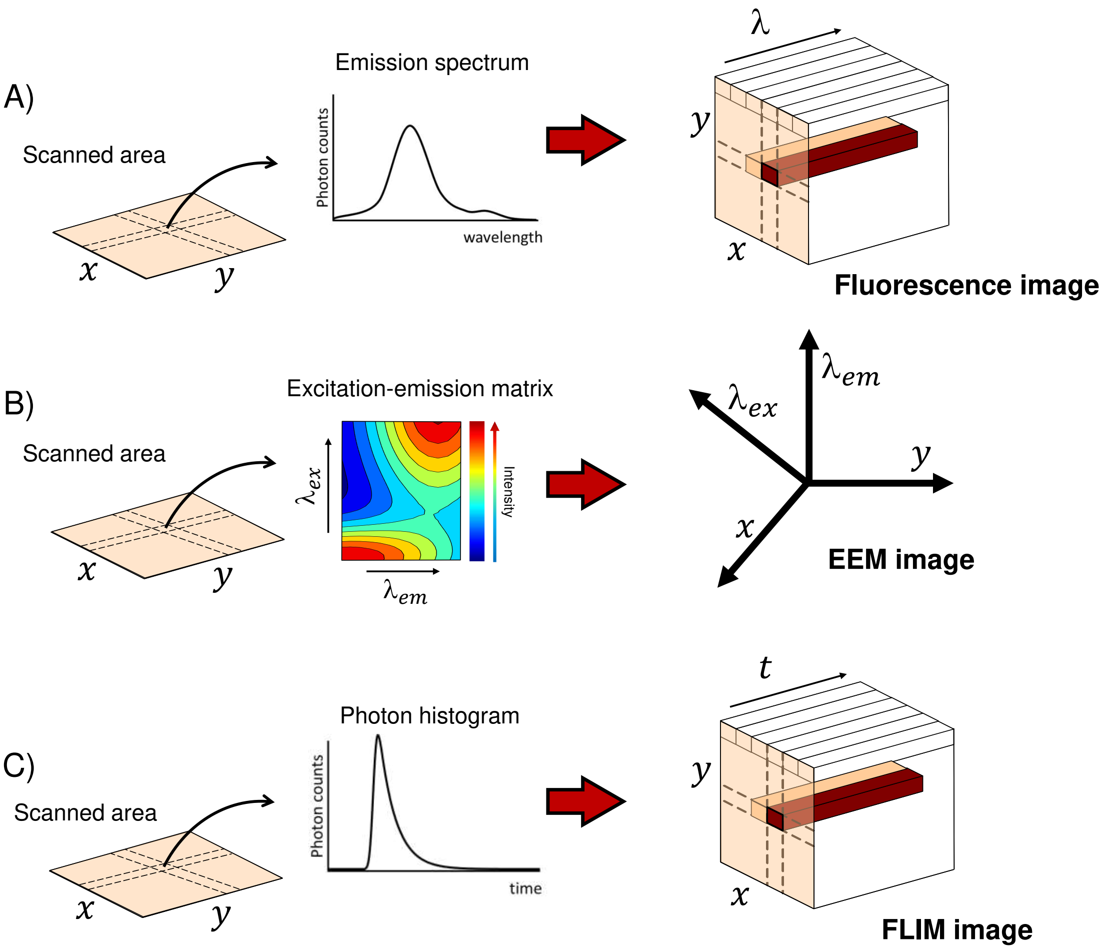


Figure 3. Examples of fluorescence HSIs with different dimensionalities. A) Each pixel contains a single emission spectrum, providing a data cube or 3D image. B) Each pixel contains a full 2D excitation-emission landscape, providing a 4D image. C) Each pixel contains the fluorescence decay over time, providing a data cube where one dimension is *time*.

2.2.1 Raman hyperspectral images

Raman spectroscopy is a powerful analytical technique that has revolutionized the understanding of the molecular composition and structural properties of materials. Named after the Indian physicist Sir C. V. Raman, who discovered the phenomenon in 1928 [Raman, 1928], Raman spectroscopy is a non-destructive and non-invasive technique that measures the scattered light resulting from the interaction between incident photons and a sample.

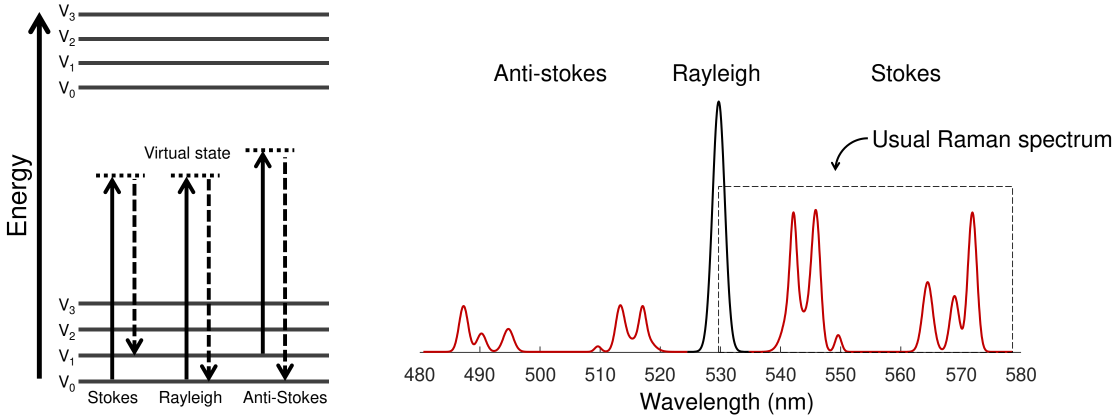


Figure 4. On the left, energy diagram illustrating Stokes, Rayleigh and Anti-Stokes scattering processes. On the right, an example of a Raman spectrum that exhibits characteristic peaks corresponding to molecular vibrational modes. The energy diagram visually represents the interactions between incident photons and molecular vibrations, resulting in Rayleigh scattering (no energy change), Stokes scattering (energy loss), and Anti-Stokes scattering (energy gain).

When a monochromatic laser beam, typically in the visible or near-infrared range, is directed onto a sample, some of the incident photons excite the molecules of the sample to a virtual state, changing their polarizability. During this interaction, some of the incident photons are scattered elastically, as Rayleigh, and inelastically, as stokes or anti-stokes Raman scattering (see Fig. 4) [Jones et al., 2019].

Rayleigh scattering is a fundamental process in Raman spectroscopy, which involves the elastic scattering of photons with no change in energy. Unlike Raman scattering, Rayleigh scattering does not involve a transition between energy levels or vibrational states. In Stokes Raman scattering, the scattered photons have less energy (longer wavelength) than the incident photons. This occurs when the molecules in the sample are in a lower energy state and they are excited to a virtual state. The results when the state decays into lower levels is that the molecule can end in a vibrational level higher than before. This results in the emission of a photon with less energy, causing it to appear to the right of the spectra (see Fig. 4).

In anti-Stokes Raman scattering, the scattered photons have higher energy (shorter wavelength) than the incident photons. This occurs when the molecules in the sample are in a higher vibrational energy state before the excitation. Thus, when molecules return to the base state, the resulting scattered photons have higher frequencies than the incident beam.

The energy difference between the incident and scattered photons in both Stokes and anti-Stokes scattering corresponds to the vibrational and rotational energy levels of the molecules in the sample. This energy difference is represented in the Raman spectrum as peaks at specific wavenumbers, which can be correlated to the molecular bonds and functional groups within the sample [Lewis and Edwards, 2001; Jones et al., 2019]. Analyzing the intensity and position of these Raman peaks provides valuable information about the chemical composition and other properties of the material under investigation. Anti-Stokes Raman scattering is generally less intense than Stokes scattering because the higher vibrational states are used to be less populated at room temperature. Therefore, Raman spectroscopy measures the signal related to Stokes scattering instead of that related to the anti-Stokes scattering.

A clear asset of Raman spectroscopy is related to the rich spectral features offered by this technique and, hence, the amount of chemical information that can be derived. The Raman bands are related to vibrational modes, which can be directly assigned to specific chemical bonds. However, Raman signals are often weak, limiting their ability to detect low-concentration substances. The low Raman intensity signal can be enhanced by increasing the amount of incident photons on the sample using a higher laser power or a longer exposure time, but these options can slow down the measurement and increase the risk of damaging the sample. Another important challenge in Raman spectroscopy is the fluorescence contribution to the measured spectrum, which interferes with the neat Raman signal and can be easily present in the analysis of biological tissues. This emitted fluorescence overlaps with the Raman signal and worsens the quality of the measurement, which requires powerful preprocessing techniques for a proper interpretation.

Researchers and scientists use extensively Raman spectroscopy in fields such as chemistry [Clark and Dines, 1986; McCreery, 2005; Efremov et al., 2008], materials science [Petry et al., 2003; Smith et al., 2016; Shipp et al., 2017] or biology [Patel and Mehta, 2010] to gain insight into the molecular nature of substances and to support a multitude of applications, including quality control [Yang and Ying, 2011].

Raman hyperspectral images are particularly interesting because the high quality of the chemical information goes together with a very good spatial resolution that can reach around 500 nm of pixel size [Turrel and Corset, 1996; Gierlinger and Schwanninger, 2007; Kawata et al., 2017], depending on the laser employed. Nowadays, the Raman imaging systems have evolved a lot and many commercial equipment options available can be found, such as the confocal Raman inVia™ microscope from Renishaw used in this thesis (see Fig. 5), located in the Super-resolution Light Microscopy & Nanoscopy Facility of Dr. Pablo Loza-Alvarez at ICFO - The Institute of Photonic Sciences, the SENTERRA II confocal Raman microscope from Bruker or the LABRAM HR Evolution from Horiba, to mention some of the companies that have been more involved in the design of this kind of instruments. Commercial Raman imaging systems, as the mentioned above, tend to operate in point scanning mode, although line and plane scanning imaging systems can also be found [Schlücker et al., 2003; Bernard et al., 2008].

In a Raman imaging system employing point scanning, a laser beam is directed through an objective lens to focus onto the sample. The scattered light is then collected by the same objective lens and directed towards a detector (see Fig. 5). The most common excitation lasers used in Raman imaging work at the single wavelengths 405, 532, 633 and 785 nm, and the Raman signal collected covers the range 200 to 3500 cm^{-1} (in relative terms to the laser wavelength). The Raman signal is usually measured from the surface of the sample to prevent interferences with the sample itself and due to its lower light penetration.

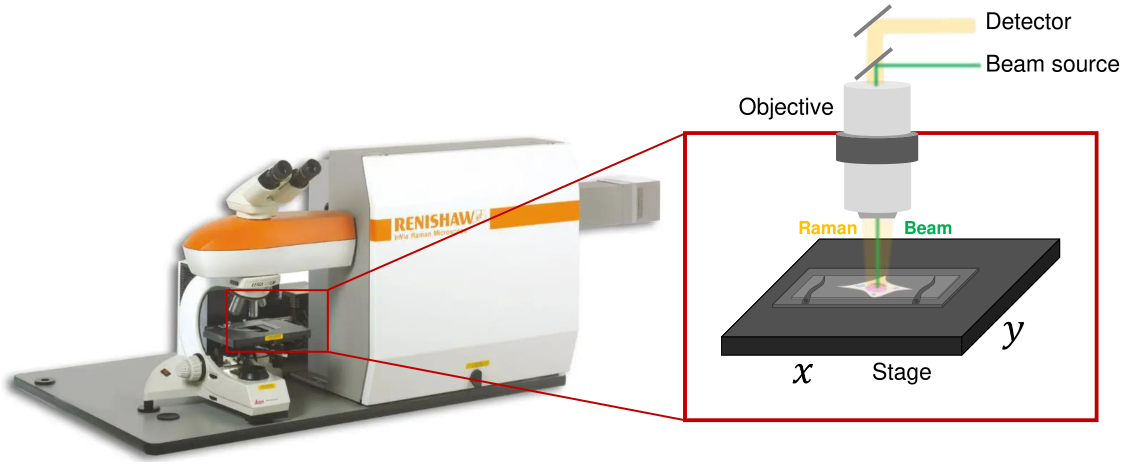


Figure 5. Schematic representation of a Renishaw Raman inVia™ imaging system setup. The system includes a precision stage for sample positioning. Then, the laser light focused on the sample induces molecular vibrations that emit characteristic Raman signals. The scattered light is then collected by the objective and directed to a detector for spectral analysis and spatial mapping of the sample.

2.1.2 Infrared hyperspectral images

Infrared spectroscopy is widely used in various scientific disciplines as a non-destructive and non-invasive tool for characterizing the molecular composition of samples. The infrared radiation was discovered by Sir William Herschel in the early 19th century [Ring, 2000; Herschel, 2013], when he detected the existence of radiation beyond the red end of the visible spectrum. Since then, it has evolved into an invaluable tool for researchers and scientists in fields such as chemistry [Koenig, 1975], biology [Naumann et al., 1991; Soriano-Disla et al., 2014], food control [Cozzolino, 2012; Danezis et al., 2016], and environmental science [Mintenig et al., 2017; Simon et al., 2018].

The infrared absorption is caused by the interaction between the infrared light and the molecules in a sample. Every molecule absorbs energy at specific frequencies and changes the dipole moments of its chemical bonds, leading to vibrational and rotational transitions (see Fig. 6). The energy of these transitions is characteristic of the molecular bonds in the molecules and provide unique infrared absorption patterns, represented by spectra [Wolfe and Zissis, 1978].

Infrared spectroscopy encompasses a broad spectral range, which is divided into three key regions: near-infrared (NIR), mid-infrared (MIR), and far-infrared (FIR). Each of these regions is suitable for different types of analyses. In this thesis, the focus is on two types of infrared imaging systems: NIR and MIR (see Fig. 6).

NIR spectroscopy operates in the near-infrared part of the electromagnetic spectrum, typically from 800 to 2500 nanometers, just beyond the visible spectrum. In this region, the absorption bands primarily result from overtones and combinations of fundamental vibrations, offering information about functional groups and chemical bonds [Wolfe and Zissis, 1978; Williams and Norris, 2001].

On the other hand, MIR spectroscopy operates in the mid-infrared region of the electromagnetic spectrum, ranging from approximately 2500 to 25,000 nanometers. MIR spectroscopy involves fundamental vibrations, including stretching and bending modes of molecular bonds [Wolfe and Zissis, 1978]. The MIR region is sensitive to more specific vibrational transitions, offering a chemical fingerprint linked to the structure and composition of chemical compounds.

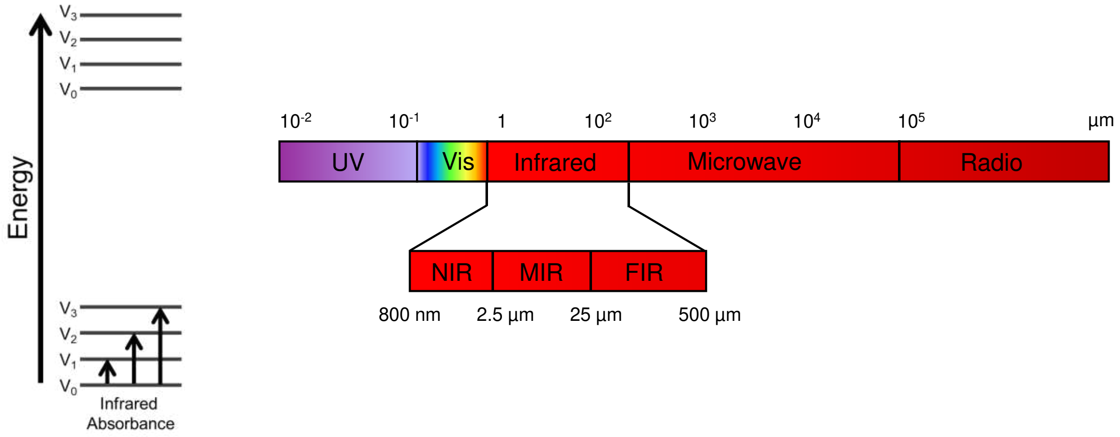


Figure 6. On the left, energy diagram illustrating the infrared absorptions of vibrational levels. On the right, scheme of the electromagnetic spectrum. The infrared region is split in three wavelength ranges: NIR, MIR and FIR.

In infrared imaging systems, the two main modes for spectra acquisition are transmission and reflectance [Burger and Geladi, 2006; Lasch and Naumann, 2006; Chan and Kazarian, 2016; Talari et al., 2017]. In the transmission mode, the infrared radiation goes through the sample (see Fig. 7A). The detector is placed on the opposite side of the sample to measure the intensity of the transmitted radiation. The difference in intensity between the detected radiation and the original beam or *"white"* is used to calculate the absorbance Abs (Eq. 1)

$$Abs = -log\left(\frac{I}{I_0}\right) \qquad \text{Eq. 1}$$

where I is the transmitted intensity and I_0 refers to the *white* or initial intensity. Measurements in transmission mode are suitable for transparent or very thin samples that allow the light beam passing through. If samples are too thick, very little or no light is transmitted and signal saturation appears.

In reflectance mode, the infrared radiation is directed onto the surface of the sample and the detector collects the radiation reflected from the sample surface (see Fig. 7B). Similar to the transmission mode, the detector measures the intensity of a reference beam (white) and a reflected beam. An expression analogous to Eq. 1 is used to obtain the absorbance, where I designs now the reflected intensity. The reflectance mode, in contraposition to the transmission mode, is suitable for thick or opaque samples.

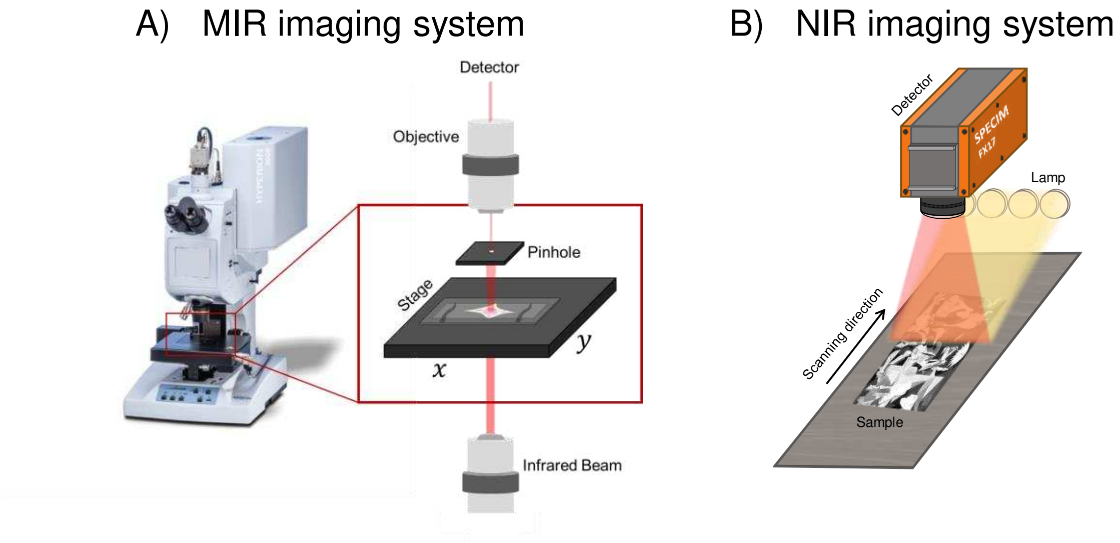


Figure 7. A) Scheme of MIR imaging system working in transmission mode and with point scanning acquisition (Hyperion 3000 FTIR microscope from Brucker company). The sample is placed on a controlled stage for positioning. An infrared beam is focused on the sample, and the transmitted light collected and sent to the detector. B) Scheme of NIR imaging system working in reflectance mode and with line scanning acquisition (Specim FX17 camera from Specim company). The sample is placed in a moving sample bed and scanned along the perpendicular motion direction.

A clear challenge in infrared imaging is related to the pixel size. Compared with other imaging systems, such as Raman, the spatial resolution for conventional infrared imaging systems is usually around 10-50 microns [Lasch and Naumann, 2006] or even lower for focal plane array detectors [Offroy et al., 2010]. Although for some applications this spatial resolution is satisfactory, it is often not enough to measure biological tissues at a cell level. In this thesis, synchrotron mid-infrared images have been acquired because the limitation on the spatial resolution for conventional infrared imaging systems can be bypassed using synchrotron radiation sources [Jamin et al., 1998; Piqueras et al., 2020]. Synchrotrons are specialized particle accelerators that generate extremely bright and highly collimated beams of light, including infrared radiation. The primary reason why a brighter beam improves spatial resolution is that it allows for the use of a smaller pinhole or aperture in the imaging system. The pinhole operates by rejecting out-of-focus light. When using a conventional infrared source with limited brightness, employing a very small pinhole can significantly reduce the intensity of the collected signal. The signal may become too weak to be distinguished from the background noise and this can lead to a loss of signal quality. In this situation, the trade-off between spatial resolution and signal intensity becomes a limiting factor. However, using the very bright source of synchrotron infrared light, an increase of the spatial resolution (up to around two by two microns) can be achieved, enabling the differentiation of structures in biological tissues.

From an instrumental point of view, MIR imaging systems can work using point scanning, line scanning and focal plane array acquisition modes. An example of MIR microscope used in this thesis is the Bruker Hyperion 3000 FTIR microscope from the Brucker company, located at the ALBA synchrotron. It is equipped with a motorized stage and a focal plane array detector, allowing image acquisition in both point and plane scanning modes and with a spectral range around 2.5 to 20 microns (see Fig. 7A).

On the other hand, a NIR camera used in this thesis is the Specim FX17 hyperspectral camera from SPECIM company, which can acquire images in the near-infrared region (around 900 to 1700 nm), operates in a line scanning mode and in reflectance mode (see Fig. 7B) and it has a pixel size around 100 microns. The sample is placed in a belt, while being illuminated by a light source or lamp. Thus, a detector scans the sample line by line, recording the signal, until covers the desired area.

2.1.3 Fluorescence hyperspectral images

Fluorescence spectroscopy is a powerful and non-destructive analytical technique, capable of detecting components at very low concentrations with wide applicability in many different fields. The origins of fluorescence spectroscopy are from the early 19th century, where E.D. Clarke described the phenomenon in fluorite mineral [Clarke, 1819]. The term "fluorescence" itself is derived from the name of the mineral fluorite, known for its fluorescent properties when exposed to ultraviolet light.

Fluorescence is a phenomenon occurring when certain substances excited by light at a specific wavelength (often in the ultraviolet or visible range) rapidly return to their ground state by “relaxation” after some pico or nanoseconds (depending on the electronic states) and emit light at longer wavelengths. Every fluorophore generates a characteristic fluorescence spectrum [Lakowicz, 2006; Valeur and Berberan-Santos, 2012].

Fluorescence imaging is widely used in various fields, including biology, chemistry and materials science. It enables researchers to characterize samples, quantify concentrations of certain fluorophores and study molecular interactions within a sample. It offers a very high spatial resolution, which can range from few nanometers (such as in single molecule fluorescence microscopy [Shashkova and Leake, 2017]) to around 200 nanometers, depending on the instrumental settings. All fluorescence images are usually acquired in point scanning mode.

Several types of images based on the fluorescence phenomenon can be distinguished according to the fluorescence signal measured per pixel. In this thesis, three types have been studied: fluorescence images, where the signal measured per pixel is an emission spectrum obtained at a specific excitation wavelength; excitation-emission fluorescence images, where the signal is an excitation-emission landscape, and Fluorescence Lifetime Imaging Microscopy images (FLIM) [Becker, 2012], where the signal acquired is the decay over time of the emission fluorescence signal obtained at a single emission wavelength. Some additional details associated with these images described above are given in the next subsections.

3D Emission fluorescence images

In this imaging technique, every pixel area of a scanned sample is excited with a specific wavelength to induce fluorescence (see Fig. 8) and the detector captures the fluorescence intensity emitted in a wavelength range, providing a fluorescence hyperspectral image with dimensions (x, y, λ_{em}), where x and y represent spatial dimensions, and λ_{em} represents the emission wavelength range. The spatial resolution of emission fluorescence images can go as low as 100 nm, depending on the instrumental settings. However, despite its high spatial resolution, the fluorescence spectra often present broad emission bands and are not very rich in spectral features when compared to other spectroscopic techniques, such as Raman and infrared. As a consequence, multiple molecules or compounds can emit fluorescence with overlapping emission bands and it is challenging to distinguish between them based solely on their spectral characteristics.

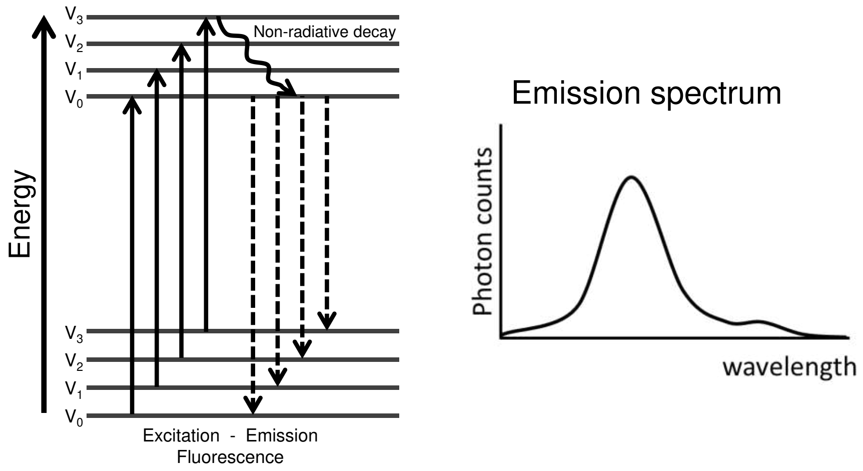


Figure 8. On the left, energy diagram illustrating the excitation phenomenon, typically in the ultraviolet or visible range, that promotes an electron to a higher energy state. The excited electron returns to its lower energy state losing energy by a non-radiative decay, emitting a photon with longer wavelength than the absorbed light. On the right, example of a fluorescence emission signal.

4D Excitation-emission fluorescence images

Excitation-emission imaging is an advanced fluorescence imaging technique that extends the concept of fluorescence imaging by providing more detailed information about the fluorophores at different excitation and emission wavelengths. Instead of just capturing a fluorescence emission spectrum per pixel, it records a 2D excitation-emission fluorescence landscape per pixel, with one dimension representing the excitation wavelength range and the other dimension representing the emission wavelength range (see Fig. 9). This kind of image provides a 4D array with dimensions $(x, y, \lambda_{ex}, \lambda_{em})$.

Excitation-emission hyperspectral imaging is a powerful tool to characterize samples with complex fluorescence behavior, enabling researchers to gain deeper insights into the nature of fluorophores and their interactions within a sample. It provides valuable data for quantitative analysis and can be a crucial tool for understanding environmental or biological processes. However, this imaging technique is not widely used in contrast to other techniques in the microscopy field.

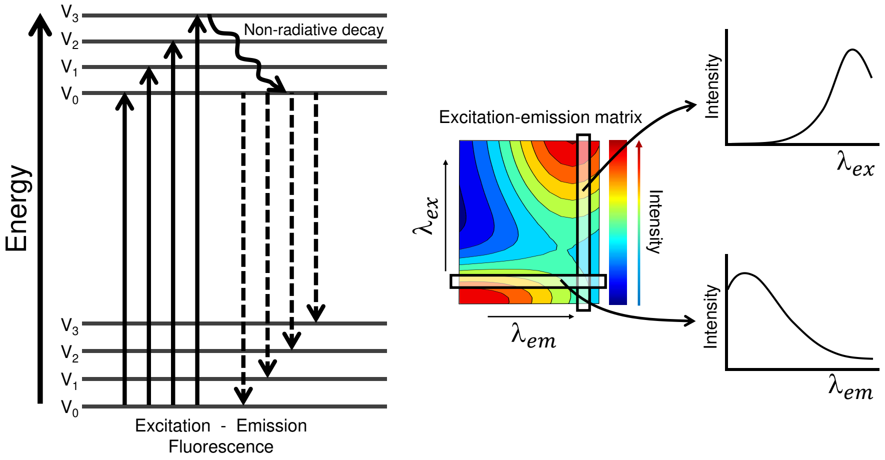


Figure. 9. On the left, an energy diagram illustrating different excitations and promoting an electron to a higher energy state. The excited electron returns to its lower energy state through losing energy by a non-radiative decay, emitting a photon with longer wavelength than the absorbed light. On the right, an example of a fluorescence excitation-emission landscape. Here, each column represents the excitation spectrum, while each row represents the emission spectrum.

During the excitation-emission fluorescence acquisition, several unwanted spectroscopic contributions can be observed in addition to the fluorescence signals. Thus, it is common to observe Rayleigh and Raman scattering signals crossing the EEM landscape, since these phenomena occur simultaneously with the fluorescence (see Fig. 10). The scattering contributions hinder the analysis of EEM spectra and preprocessing is required to mitigate this problem. In addition, due to the nature of the measurement, no fluorescence signal is observed below the first order Rayleigh scattering since the emission signal always appears at wavelengths longer than the excitation used.

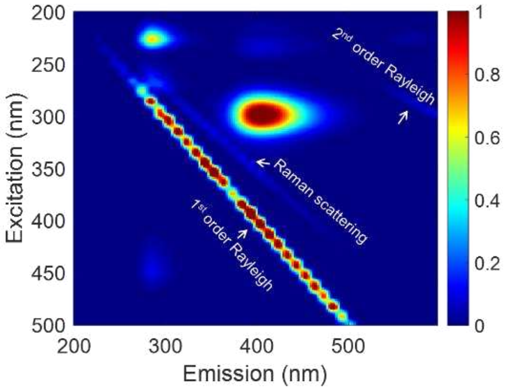


Figure 10. Example of excitation-emission measurement of a mixture of pharmaceutical compounds (ibuprofen and acetylsalicylic acid). Rayleigh and Raman scattering can be observed through the EEM landscape.

Usually, fluorescence microscopes operate in point scanning mode. An example of a fluorescence microscope used in this thesis is the Leica TCS SP8 STED 3× microscope, manufactured by Leica (see Fig. 11), located in the

Super-resolution Light Microscopy & Nanoscopy Facility of Dr. Pablo Loza-Alvarez at ICFO - The Institute of Photonic Sciences. It is equipped with a motorized stage, allowing image acquisition in point scanning mode. It allows the acquisition of excitation-emission images thanks to the white light excitation laser that covers a spectral range from 470 to 670nm.

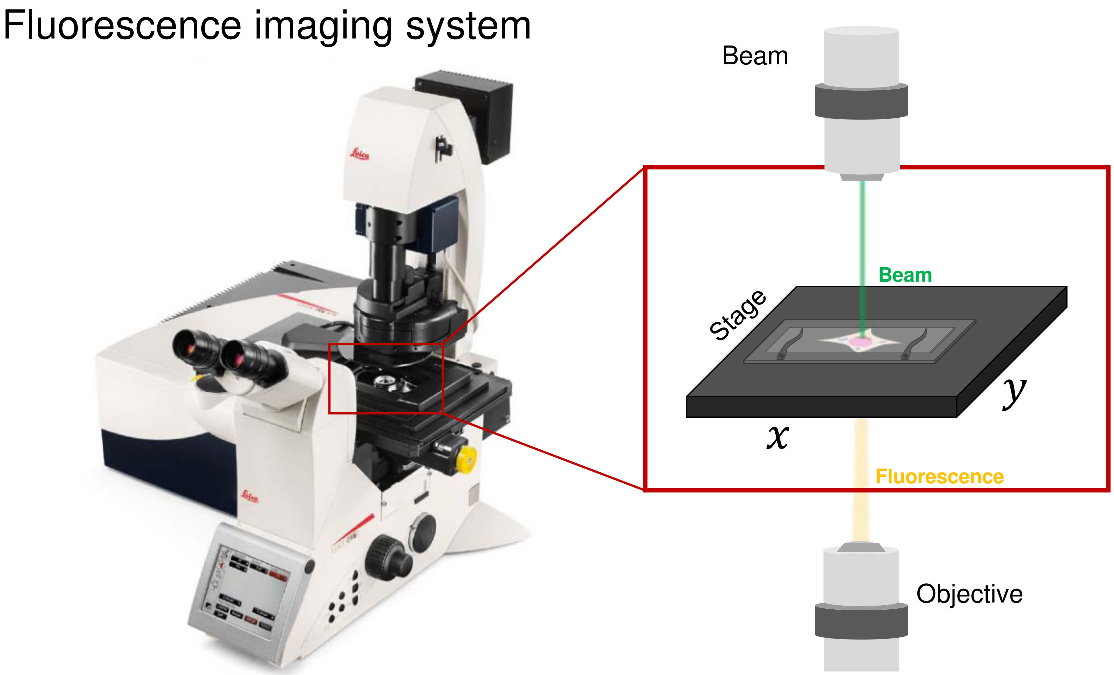


Figure 11. On left) Example of fluorescence imaging system (Leica TCS SP8 STED 3× microscope from Leica company). On right) Scheme of the acquisition. The sample is placed on a controlled stage for positioning. A laser beam is focused on the sample, and the emitted light is sent to the detector.

Fluorescence Lifetime Imaging Microscopy (FLIM) images

Fluorescence Lifetime Imaging Microscopy (FLIM) provides valuable insights into the fluorescence properties of samples. Unlike traditional fluorescence imaging, which captures the fluorescence intensity at different emission wavelengths, FLIM focuses on the fluorescence decay kinetics of fluorophores, measured by Time Resolved Fluorescence Spectroscopy (TRFS) [Lakowicz, 2006; Becker, 2012; Lemmetyinen et al., 2014; Liput et al., 2020].

In FLIM, a pixel is illuminated with a pulsed light at a specific wavelength. Then, instead of measuring the emitted intensity at various wavelengths, FLIM measures how many photons stay in the excited state in every pulse before returning to their ground state for a particular time. When a molecule is excited to a different electronic state, it remains a certain time on that state before returning to the ground state (usually pico or nanoseconds) by a non-radiative loss of energy (see Fig. 8). The time that a specific molecule remains in the excited state changes every time that a new excitation takes place. Indeed, all these times follow an exponential decay distribution (see Fig. 12A and Eq. 2).

$$x(t) = Ae^{-\frac{t}{\tau}} \qquad \text{Eq. 2}$$

Where $x(t)$ represents the intensity or number of photons expected to be detected at a time t, A represents the preexponential factor, and τ is the decay ratio (called lifetime) and is a distinctive characteristic of each fluorophore.

The FLIM images can be represented as a data cube with dimensions (x, y, t), where x and y represent spatial dimensions, and t represents the time axis of the decay curve. The spatial resolution of FLIM images can reach easily sub-micrometer levels, making it a powerful tool for studying biological and cellular processes [Suhling et al., 2015].

However, when the fluorescence decay is measured in each pixel, the signal is affected by the so-called Instrumental Response Function (IRF) [Luchowski et al., 2009]. The IRF refers to the time profile of the response of the instrument to an instantaneous signal, such as a short pulse of light. In other words, it describes how the instrument *sees* a perfect, short-event signal. Ideally, the IRF shape should be infinitely narrow, following a Dirac delta function, but this is never the case. The IRF shape depends on several factors such as the instrument optics, electronics and the laser characteristics and it usually presents an approximate Gaussian shape.

FLIM measurements can be accurately characterized as the convolution of the IRF with the inherent fluorescence decay signal $\boldsymbol{x}$, i.e., the recorded signal in FLIM, denoted as $\boldsymbol{y}$, can be expressed analytically as in Eq. 3 (see Fig. 12B).

$$\boldsymbol{y} = \mathbf{IRF} * \boldsymbol{x} \qquad \text{Eq. 3}$$

where * denotes convolution.

Since the real IRF is not infinitely narrow, it affects the fluorescence decay measured and the subsequent analysis. On the one hand, the correct extraction of the lifetime of the fluorophores gets compromised due to the difficulty to perform an exponential fitting of the data in the initial region of the signal and, on the other hand, a loss of available signal usually happens since the non-exponential part of the signal is often left out of the analysis, increasing the error in the estimated preexponential factors.

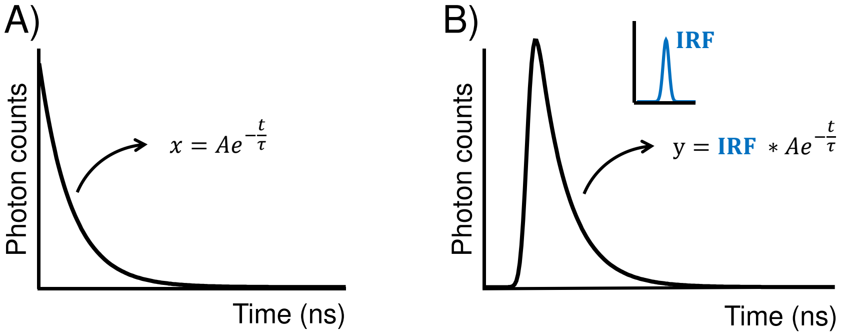


Figure 12. A) The fluorescence decay can be defined as a monoexponential decay function. B) The measured fluorescence decay is affected by the IRF (in blue). The measurement follows the convolution process, where both signals, the IRF and the exponential decay x are convolved. Therefore, the observed fluorescence decay behaves as B) instead as A), being A) the ideal scenario.

To deal with this problem, especially when fluorophores present short lifetimes around or below 1 ns, deconvolution is a standard operation. This process requires knowledge or experimental estimation of the IRF. The experimental methods for the estimation of the IRF, which involve measuring the emission of a fluorophore with a very short fluorescence lifetime or the elastic scattering of the excitation laser pulse, are the most used. Thus, once the IRF is estimated, deconvolution of the measured signal (y) can be performed to estimate the true signal (x) according to Eq. 3.

However, the experimental measurement of the IRF has a significant limitation. The IRF depends on the experimental conditions, which may change across measurements due to chemical and instrumental factors. For example, the use of different solutions, different wavelength filters, different emission windows, etc. As a consequence, if the IRF is not well characterized, a bias in the extraction of x through deconvolution will occur.

The fluorescence lifetime imaging microscopes used follow the same scheme of Fig. 11. For example, the same LEICA microscope shown before can acquire fluorescence lifetime images if equipped with the commercial FALCON detection system, which is a specific detector suitable for FLIM measurements.

The main difference between the 3D and 4D fluorescence images and a FLIM image can be found in the laser and the detector used for their acquisition. In FLIM, the laser emits a short burst of light at a specific wavelength, typically in the ultraviolet, visible, or near-infrared range. The laser pulse is often very brief, in the order of picoseconds, to ensure precise temporal control over the excitation process, while the detector can capture the emitted fluorescence from the sample in picosecond scale.

As a main conclusion of this section, there exists a wide range of hyperspectral images that offer numerous possibilities to the researchers for exploring and comprehending the nature of the samples, each one from a unique perspective and allowing a more comprehensive understanding of their chemical and spatial features.

In the following section, the underlying models for the analysis of hyperspectral images will be explained in detail, along with the challenges associated with image fusion of hyperspectral images.

2.2 Chemometric tools for the analysis of hyperspectral imaging

Hyperspectral images contain a vast amount of complex and challenging data to be analyzed. For instance, Raman hyperspectral images can be easily built by several thousand of spectra with thousands of variables and the number of spectra can even go up to millions in fluorescence hyperspectral images. The massive number of pixels, variables and the complexity of the chemistry present in many samples requires chemometric tools able to transform the raw data into reliable and interpretable information. This section starts presenting the fundamental linear models for hyperspectral images, i.e., bilinear and high order multilinear models). Afterwards, unmixing methods, with a specific focus on Multivariate Curve Resolution Alternating Least Squares, are discussed in detail as the idoneous tools to recover the underlying model of the image measurement. Finally, the section introduces the data fusion structures, called multisets, for hyperspectral images and the challenges associated with the image fusion scenario.

2.2.1 Hyperspectral images. The underlying model of the measurement

There exists a wide variety of chemometrics methods to address different aspects related to the analysis of hyperspectral images and the selection of the chemometric tool depends on the objective of the analysis and data characteristics. In this thesis, the main focus is on solving the mixture analysis problem for hyperspectral images by bilinear or multilinear decomposition methods, i.e., to find the pure constituents present in one or several acquired images [Lawton and Sylvestre, 1971; Hamilton and Gemperline, 1990; de Juan and Tauler, 2021]. The justification of this choice responds to several reasons. First and most important, the great match between the underlying nature of spectroscopic measurements and the linear models. Second, the simplicity offered by linear models to accurately describe the information of hyperspectral images. Third, the meaningful data compression of the raw information into a small number of chemically recognizable image constituents (components) that facilitates enormously the interpretation of the results.

In the following subsections, a comprehensive explanation of both bilinear and higher order linear models is provided, describing their theoretical principles and their relevance for the analysis of hyperspectral images.

Bilinear models

To understand the bilinear model and its intrinsic relation with hyperspectral images, it is needed to consider the fundamental principle in spectroscopy, i.e., the Beer-Lambert-Bouguer law [Swinehart, 1962]. This law provides a straightforward relationship between the concentration of a substance in a sample and the amount of light absorbed at specific wavelengths.

Specifically, the Beer-Lambert-Bouguer law correlates the concentration of a given analyte n with its absorbance. This relationship can be expressed as a bilinear model (Eq. 4):

$$d_{i,j} = \sum_{n=1}^{N} c_{i,n}\varepsilon_{j,n} \qquad \text{Eq. 4}$$

where $d_{i,j}$ designs the absorbance for a given sample i at wavelength j, $c_{i,n}$ is the concentration value of the n compound in the sample i and $\varepsilon_{j,n}$ the absorptivity of the n compound at wavelength j (the constant pathlength l is skipped from the expression for simplicity). Thus, the total absorbance measured on a sample is the sum of the individual absorbances of the N compounds forming it.

When $d_{i,j}$ is displayed for all J wavelengths, the absorbance spectrum $\boldsymbol{d}_i$ of the sample i is obtained. In addition, when $\varepsilon_{j,n}$ is displayed for all J wavelengths, the pure absorbance spectrum $\boldsymbol{s}_n^{\mathrm{T}} = [\varepsilon_{1,n}, \varepsilon_{2,n}, \varepsilon_{2,n} \ldots\ \varepsilon_{J,n}]$ of the compound n is obtained. Eq. 5 represents the absorbance spectrum of sample i as the sum of the pure absorbance spectra of the N compounds weighted by their respective concentrations.

$$\boldsymbol{d}_i = \sum_{n=1}^{N} c_{i,n}\boldsymbol{s}_n^{\mathrm{T}} \qquad \text{Eq. 5}$$

When a real set of I spectra from samples formed by mixtures of N compounds is considered, the bilinear model in Eq. 5 can be expressed in matrix form as in Eq. 6.

$$\mathbf{D} = \mathbf{CS}^{\mathrm{T}} + \mathbf{E} \qquad \text{Eq. 6}$$

where **D** is sized (I,J) and contains the spectra of the I samples formed by mixtures of N components and $\mathbf{C}$ (I,N) and $\boldsymbol{S}$ (J,N) represent the matrices of column concentration profiles and pure spectra profiles for each of the N compounds in the sample, respectively. The term $\mathbf{E}$, sized (I,J), corresponds to the unexplained variance or residuals, usually related to noise. The bilinear model based on the Beer-Lambert-Bouguer law also applies to many other spectroscopies, such as emission fluorescence and Raman, for which the total signal measured can be expressed as well as a weighted sum of the pure spectral signal of the components.

Bilinear models are particularly valuable for many data that can be expressed in two modes of variation. In a bilinear model, the information of every compound or component is always defined by a dyad of vectors $\boldsymbol{c}_i$ and $\boldsymbol{s}_i^{\mathrm{T}}$ and the contribution of a component to the total signal measured is represented by the product $\boldsymbol{c}_i\boldsymbol{s}_i^{\mathrm{T}}$.

Hyperspectral images can be very well represented by bilinear models because of the bilinear nature of the spectroscopic techniques employed [Gemperline, 1989; Hamilton and Gemperline, 1990; Ruckebusch and Blanchet, 2013; de Juan, 2018]. In order to express the hyperspectral image measurements as a bilinear model, the image cube is unfolded into a data matrix **D,** sized $(x \times y, \lambda)$ where the pixel spectra are placed one below the other (see Fig. 13). With this matrix configuration, all pixel spectra can be described as a combination of the spectra of the pure components (**S)** in different proportions (**C**), as in Eq. 6.

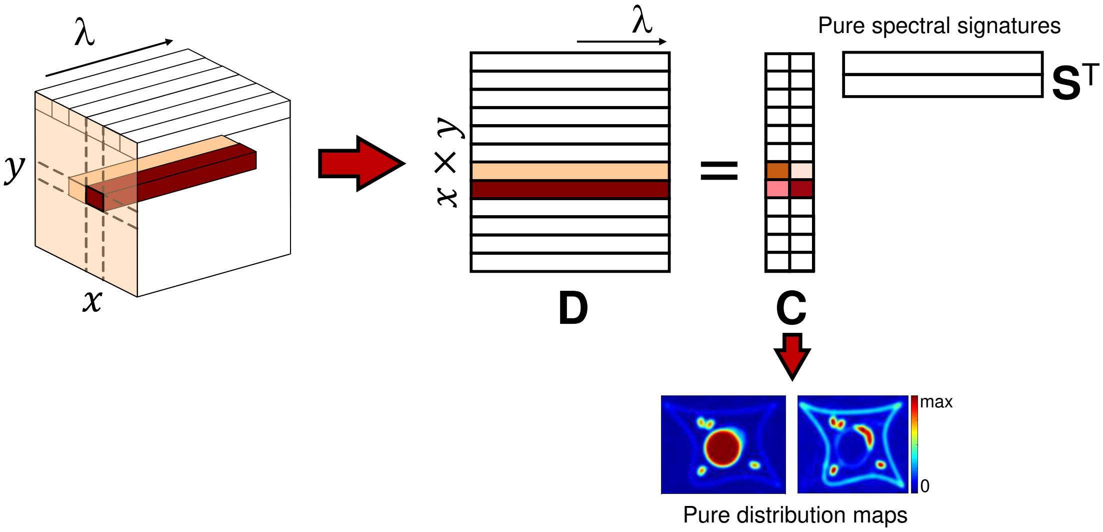


Figure 13. The hyperspectral image cube is unfolded by concatenating the spectra one below the other one, forming the matrix **D**. The matrix **D** is then expressed as a bilinear model. When the pure concentration profiles (**C**) are refolded, the distribution maps of every image constituent (component) are recovered.

The concentration profiles **C** can be refolded into distribution maps that represent the initial image spatial configuration. The distribution maps contain information about the concentration and spatial distribution of every sample component across the hyperspectral image. The underlying bilinear model of a hyperspectral image inherently encloses information about the spatial distribution and identity of the components present in the sample, allowing a full characterization of the system under study.

Multilinear models

Certain spectroscopic measurements present a complexity that goes beyond the bilinear structure. While bilinear models define adequately the two modes of variation (**C** and **S**) present in Raman, infrared and emission fluorescence images, there are other spectroscopies that require multilinear models adapted to describe additional modes of variation [Hirschfeld, 1980; Olivieri et al., 2004; Malik and Tauler, 2013; Alcaraz et al., 2019]. This is the case of 4D excitation-emission fluorescence images, where every pixel has associated a full 2D excitation-emission landscape and the image measurement is defined by a four-dimensional data set $(x,\ y, \lambda_{ex},\ \lambda_{em})$. This additional spectral dimension requires a trilinear model to be properly described [Andersen and Bro, 2003; Tauler et al., 1998; Marín-García and Tauler, 2020]. To understand the connection between a 4D image and a trilinear model, the initial four-way array needs to be unfolded into a cube or tensor **D̲**, sized $(x \times y, \lambda_{ex}, \lambda_{em})$, by placing every excitation-emission landscape one below the other. In a trilinear model, each component is defined by a tryad of pure profiles (see Fig. 14).

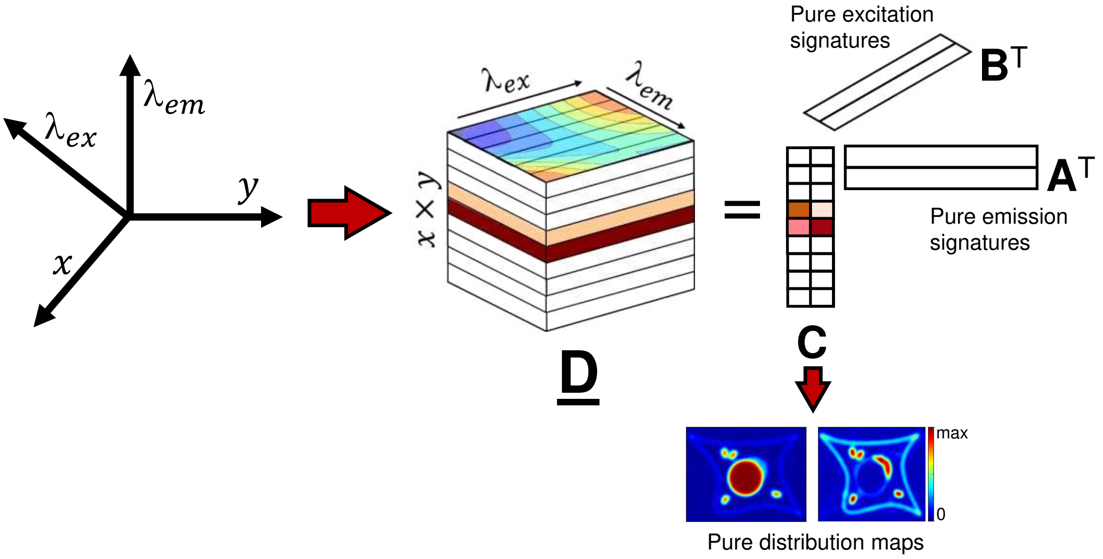


Figure 14. The excitation-emission image is unfolded by concatenating each excitation-emission landscape one below the other one, forming the data cube **D̲**. The information in the data cube **D̲** is described by a trilinear model, where the three modes of pure profiles are the pure concentration profiles (**C**), the pure excitation profiles (**B**) and the pure emission profiles (**A**). As in bilinear models, the pure concentration profiles can be refolded and the distribution maps are recovered.

The trilinear model extends the bilinear concept for data with three modes of variation (**C**, related to the concentration profiles, **A**, related to the pure emission profiles and **B**, related to the pure excitation profiles). As described above, the concentration profiles **C** can be refolded into distribution maps that represent the information on the spatial distribution of the related components. In a 4D image, the additional spectral dimension provides richer chemical information and new opportunities to find characteristic spectral features to differentiate compounds.

In addition, the existence of even more complex scenarios that require higher order linear models, as the quadrilinear model or higher to be described needs to be taken into consideration [Malik and Tauler, 2014; Alcaraz et al., 2019]. While spectroscopic measurements that directly provide quadrilinear or higher order linear data are not very common, the introduction of additional modes to the measurements, such as time, may demand the use of this multilinear models, as will also be shown in this thesis.

2.2.2 Unmixing methods. Multivariate Curve Resolution Alternating Least Squares (MCR-ALS)

As seen in subsection 2.2.1, the bilinear model is the natural, simple and compact way to describe the information of a hyperspectral image measurement through the signal contributions of their components. However, in most of the cases only the raw image measurement, i.e., the **D** matrix of Eq. 6, is available. To retrieve the underlying bilinear model formed by matrices **C** and **S** from the sole information in matrix **D**, unmixing methods are needed. In this thesis, the focus is on the unmixing method called Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) [de Juan and Tauler, 2021], an iterative method meant to address the mixture analysis problem, i.e., to obtain **C** and **S** via an alternating least-squares optimization under constraints. MCR-ALS is widely used in many research fields and it is particularly suited for hyperspectral imaging [Duponchel et al., 2003; Amigo et al., 2008; Ruckebusch and Blanchet, 2013; de Juan, 2018].

The use of MCR-ALS to analyze HSI data passes through several steps:

1) Initially, the optimal number of components to describe the information in the data set **D** is determined.
2) Then, an initial estimate of **C** or **S** is selected.
3) An iterative process through an alternating least squares optimization of the matrices **C** and **S** is performed under constraints.
4) Lastly, the algorithm stops when convergence criteria are fulfilled.

A detailed explanation of each step is presented below, addressing the critical aspects of each one.

1) *Determination of the number of components to describe the data set* ***D****.*

The analysis of the data by MCR-ALS starts by establishing the suitable number of components to describe the dataset **D**. The number of components can be determined in several ways, such as estimating the chemical rank[1] by Principal Component Analysis (PCA) [Joliffe and Morgan, 1992]. PCA provides the bilinear decomposition of the data **D** into the uncorrelated sources of variation that most contribute to explain the total variance of the data, called principal components. The explained variance of each principal component, linked to the related eigenvalue, can be plotted using a *scree plot* that shows a break-off point that separates the principal components related to relevant information

[1] C*hemical rank* is defined as the mathematical rank of the data set in absence of noise and perfect bilinearity.

from those associated with noise. This point helps to estimate how many components are needed to properly fit the dataset **D** using a bilinear model and is a good indication for the number of components required by an MCR model.

However, relying solely on statistical methods such as PCA to determine the number of components may not always capture all the underlying components of the data. For example, the detection of minor chemical species or the presence of very correlated pure spectra can be unclear when PCA is used, particularly in image data sets with a high noise level. Thus, when in doubt, it is advisable to run several MCR-ALS analyses with different number of components in order to assess whether the resolved pure components have chemical sense. The knowledge about the samples and the underlying chemistry of the data is the best ally, in conjunction with the use of chemometric tools, to properly determine the number of components required by the MCR model.

2) *Proposal of initial estimates of **C** or **S**.*

Once the number of components is selected, the choice of the initial estimates (an initial **C** or **S** matrix) is performed. The choice of initial estimates is an important step in MCR-ALS, since it can have an impact on the convergence of the model. There are two usual ways to obtain the initial estimates, i.e., from prior knowledge, e.g., when the pure spectra of some components are known, or from the selection of the purest columns (concentration profiles) or rows (spectral profiles) of the initial data set **D**[2]. Since the aim is optimizing these estimates to obtain chemically meaningful profiles, it is not recommended starting with initial estimates formed by random numbers.

As a general rule, it is desired to use the pure profiles of some or all components if they are available. In absence of this information, several chemometric methods can be used to choose good initial estimates based solely on the available data **D**. Among them, there are purest variable selection methods, such as Simple-to-use Interactive Self-modelling Mixture Analysis (SIMPLISMA) [Windig and Guilment, 1991], essential spectra selection methods [Sawall et al., 2022; Coic et al., 2022] and many others [Cocchi et al., 2018]. These methods aim to identify the most dissimilar rows or columns in a data matrix **D**, i.e., the purest ones, and provide very good initial estimates for the MCR-ALS analysis.

[2] Selecting the purest rows involves choosing samples where the presence of a particular component is the highest, i.e., the presence of the rest of the components is the lesser. Similarly, selecting the purest columns consist of choosing those wavelengths where the signal is primarily attributed to a specific component, while the presence of the other ones is as minimum as possible.

3) *Alternating least squares optimization of **C** and **S** matrices*

When the initial estimates are selected, the next step in MCR-ALS is the alternating least squares (ALS) optimization to provide **C** and **S** matrices of pure profiles according to Eqs. 7 and 8.

$$\mathbf{C} = \mathbf{D}(\mathbf{S}^{\mathrm{T}})^{+} \qquad \text{Eq. 7}$$

$$\mathbf{S}^{\mathrm{T}} = \mathbf{C}^{+}\mathbf{D} \qquad \text{Eq. 8}$$

where $\mathbf{C}^{+}$ and $(\mathbf{S}^{\mathrm{T}})^{+}$ correspond to the Moore–Penrose inverse, hereafter referred as pseudoinverse of **C** and **S**, respectively.

In hyperspectral image unmixing, the start of the iterative process most often consists of the calculation of **C** given **D** and an initial estimate of **S**[3], (Eq. 7), followed by the calculation of $\mathbf{S}^{\mathrm{T}}$ given **D** and the previous calculated **C** (Eq. 8). After completing the first iteration, the process is repeated recalculating **C** and **S** while updating the pure matrices with the newly calculated ones each until an optimal reproduction of the initial data set **D** is reached according to a preset convergence criterion.

One of the most interesting and useful characteristics of MCR-ALS is the use of constraints during the iterative optimization process. A constraint is a specific characteristic that the pure profiles in matrices **C** or **S** must fulfil. The role of constraints is helping in the chemical interpretation of the components and improving the accuracy of the retrieved profiles of the bilinear model.

The versality of MCR-ALS allows applying constraints in very different ways [Tauler et al., 2020; de Juan and Tauler, 2021]. The constraints can be imposed on the entire pure matrices **C** or **S**, to specific elements in the profiles and to specific components. This gives to MCR-ALS the flexibility to cover a wide range of analytical problems and the ability to adapt to very diverse data. There are a wealth of constraints adapted to many application scenarios, some based on mathematical properties and some others that rely on chemical knowledge [Tauler et al., 1995; Bro and Sidiropoulos, 1998; Blanchet et al., 2007; Tauler et al., 2020]. However, in this section, the attention will be focused on the constraints commonly used and specifically proposed for hyperspectral image analysis.

The first constraint proposed for MCR methods was non-negativity, which enforces the values of a pure profile to be positive. This can be applied in different ways, such as replacing negative values by zeros after the calculation of matrices **C** or **S** or employing softer algorithms, such as non-negative least squares [Lawson and Hanson, 1995] or fast non-negative least squares [Bro and De Jong, 1997]. This constraint is always applied to concentrations and is

[3] Note that **C** can be chosen as well as initial estimates, if available. If this is the case, the order of the calculation of **C** and **S** is inverted, i.e., Eq. 8 is performed first, followed by Eq. 7.

widely employed in spectroscopic data, since the majority of spectra, including fluorescence, Raman, Infrared, etc., provide non-negative signals.

The local rank constraint [Tauler et al., 1995] is also essential in many applications. This constraint is based on the identification of regions (often called *windows*), where certain components are absent in the profiles **C** and **S**. For example, if it is known that a specific component is not present in certain regions of **C** or **S**, then these parts can be forced to have zero values during the iteration process. This allows the active use of purity or selectivity information and increases significantly the accuracy of the retrieved profiles. The local rank constraint has been specifically adapted to be applied on the distribution maps of images [de Juan et al., 2005; de Juan et al., 2008; Zhang et al., 2016]. The application of this constraint is based on the determination of the number of components in subareas of the image formed by windows of few neighboring pixels, followed by a subsequent identification of the absent components in regions with rank lower than the total. Thus, during the iterative process, the matrix **C** is forced to have null signal for the components known to be absent in certain pixels due to the local rank information.

In addition, there have been new image-specific constraints that use the spatial information of the components to improve the definition of the distribution maps in hyperspectral image analysis. Examples of image-specific constraints are the segmentation [Hugelier et al., 2015] or the smoothness constraint [Hugelier et al., 2015b]. The segmentation constraint forces the pure profiles of **C** to fulfill segmented patterns; therefore, it is an indirect way to introduce local rank information on the pure distribution maps. On the other hand, the smoothness constraint is the opposite to the segmentation constraint, being useful for those components who exhibit a smooth gradient of concentration on the distribution maps. The application of image-specific constraints during the iteration process is performed by refolding the matrix **C** after its calculation in Eq. 7 into the distribution maps. Then, the selected constraint is applied to the specific components. Once it is applied, the distribution map is again unfolded, and the iterative process continues.

A special attention must be given to the trilinear constraint that allows MCR-ALS to analyze data that require high-order linear models. Although MCR-ALS is a bilinear decomposition method, trilinear models can be accommodated through the application of the trilinear constraint during the iterative alternating least squares optimization [Tauler and Barceló, 1993; Tauler et al., 1998; Marín-García and Tauler, 2020; Alier et al., 2011]. In this thesis, the trilinear constraint plays a central role, since many fluorescence images follow this underlying model. Besides, the possibility to perform trilinear decompositions of datasets has the advantage of providing unique solutions, ensuring the correct retrieval of component profiles. The implementation of the trilinearity constraint in the context of image analysis is explained in detail in subsection 4.1.

4) Establishing the convergence criterion in the ALS iterative optimization.

Determining when to stop the iterative process is also a necessary aspect in MCR-ALS analysis. The convergence criterion defines the conditions under which the algorithm should stop iterating, ensuring that the results are sufficiently accurate while avoiding unnecessary computation, i.e., too few iterations may result in an insufficiently optimized solution, while excessive iterations can lead to spend computation time without a clear improvement of the results.

Usual convergence criteria include reaching a specified number of iterations or observing minimal changes in the fit between successive iterations. Generally, the convergence criterion based on the change on the lack of fit of the model across the iterations is more desired since it is less arbitrary and reflects more clearly the quality evolution of the sought solutions. The lack of fit of the model is defined as in Eq. 9, while the relative change on the lack fit of the model is defined in Eq. 10.

$$LOF\ (\%) = 100 \times \sqrt{\frac{\sum_{i,j} {e_{ij}}^2}{\sum_{ij} {d_{ij}}^2}} \qquad \text{Eq. 9}$$

$$Conv\ \ (\%) = 100 \times \frac{LOF_{it-1} - LOF_{it}}{LOF_{it-1}} \qquad \text{Eq. 10}$$

Where d_{ij} is an element of the original matrix **D** and e_{ij} is element of the residual matrix **E**. Finally, LOF_{it} indicates the lack of fit at iteration it.

However, at this point it is worth noticing that small changes on the fit do not necessarily imply irrelevant changes on the profiles in **C** and **S**. This is especially important when the pure profiles are very correlated (such as in TRFS data), since changes on the **C** and **S** profiles do not always go together with a significant variation of the model fit. In these specific instances, it is advisable to use a low convergence criterion (10^{-6}% or lower) to prevent stopping the iterative process when the profiles are still changing. For this reason, it is recommended that the researcher actively supervises the convergence process and monitors the variation of profile shapes as an additional way to decide on the convergence of the optimization.

Rotational, scale and permutation ambiguity.

Despite bilinear decomposition methods, such as MCR-ALS, are powerful and flexible modeling approaches, the resulting bilinear model may suffer from ambiguity. Thus, rotational, scale and permutation ambiguities can impact the interpretability and reliability of the obtained solutions. In this context, comprehensive explanations are provided to address each ambiguity, pointing to strategies to reduce or completely avoid them.

Rotational ambiguity describes the possibility to obtain bilinear models of components with different profile shape while still effectively fitting the **D** matrix in an identical and optimal way [Borgen and Kowalski, 1985]. As shown in Eq. 11, matrices **C** and **S** can be multiplied by a $\mathbf{TT^{-1}}$ term, equal to the identity matrix, where **T** (N,N) is a transformation matrix containing real values. The inclusion of this term does not affect the model fit and allows for an optimal reproduction of matrix **D**.

$$\mathbf{D} = \mathbf{CTT}^{-1}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \qquad \text{Eq. 11}$$

However, Eq. 11 can now be rearranged such that **C**' = **CT**, and $\mathbf{S}'^{\mathrm{T}} = \mathbf{T}^{-1}\mathbf{S}^{\mathrm{T}}$ (Eq. 12).

$$\mathbf{D} = \mathbf{C}'\mathbf{S}'^{\mathrm{T}} + \mathbf{E} \qquad \text{Eq. 12}$$

This implies that the matrix **D** can be described using components with shapes formed by linear combinations of those from the true chemical species contained in the data. Thus, there may be multiple solutions associated with a bilinear decomposition and the reduction or the elimination of rotational ambiguity in MCR methods is essential.

The only way to reduce or completely avoid the rotational ambiguity is the application of constrains on the pure profiles of **C** and $\mathbf{S^T}$. Not all constraints are equally powerful to do this task and selectivity is key in this respect [Manne, 1995; Tauler et al., 1995]. A pure profile contains selective information if there is a region where the contribution of a single component is present. For example, in a hyperspectral Raman image, a concentration profile $\boldsymbol{c}_n$ of the component *n* has selective information if it contains one or more pixels where only the component *n* is present. When this is the case, the related $\boldsymbol{s}_n$ profile can be retrieved without ambiguity. Similarly, a pure spectral profile $\boldsymbol{s}_n$ contains selective information if there are one or more wavelengths where the signal recorded is only attributed to component *n*. The spectral selective information would ensure the correxct retrieval of the related concentration profile $\boldsymbol{c}_n$. Rolf Manne [Manne, 1995], extended the results previously obtained by Maeder and Malinowski about the general local rank conditions required to obtain bilinear models in absence of rotational ambiguity [Maeder, 1987; Malinowski, 1992]. As a general rule, the more selective the profiles are, the higher the probability to reduce or suppress the rotational ambiguity.

Scale and permutation ambiguity are less relevant from the data analysis point of view since they do not modify the profile shapes and preserve the interpretability of the results, but they should be addressed to further stabilize the bilinear model [Tauler and Maeder, 2009; Golshan et al., 2016; Sawall et al., 2019].

The scale ambiguity describes the possibility to obtain bilinear models with pure profiles different in scale (not the shape) without changing the fit of the model. This can be clearly seen in Eq. 13 and 14, where the pure matrix **C** is multiplied by a scalar a, while simultaneously the pure matrix **S** is divided by the same scalar. This operation can be done individually per component using different scaling factors and reaching the same conclusions.

$$\mathbf{D} = \mathbf{C}aa^{-1}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \qquad \text{Eq. 13}$$

$$\mathbf{D} = \mathbf{C}'\mathbf{S}'^{\mathrm{T}} + \mathbf{E} \qquad \text{Eq. 14}$$

Here, the new pure profiles of **C**’ and **S**’ have different scale than those of **C** and **S,** but their shapes remain the same. To address the scale ambiguity, normalization procedures are commonly employed. Normalization involves dividing the pure profiles of one pure matrix by a scalar, often based on metrics like their Euclidean norm, their maximum value, or their total area. This normalization ensures that the scale of the pure profiles is fixed and consistent among resolution and among components, facilitating the interpretation and analysis of the data.

Finally, the last ambiguity is the permutation ambiguity. This ambiguity refers to the ability of rearranging or permuting the order of the components in the model without changing the fit of the model to the observed data. In other words, different permutations of the components lead to equivalent models that explain the data equally well. The permutation ambiguity is often ignored since it does not modify neither the shape or scale of the profiles retrieved. It only needs to be taken into account when the order of components is helpful for interpretation or comparison of different models.

2.2.3 Data fusion of hyperspectral images

In this section, the core concept of this thesis, the image fusion, is addressed. In the MCR framework, this concept is linked to multiset analysis [Tauler, 1995; Tauler et al., 1995; de Juan, 2019; de Juan et al., 2019; Tauler et al., 2020; de Juan et al., 2024].

When exploring scenarios where two or more HSIs are present, the individual analysis of each image appears as the most used approach. However, data fusion provides a much more advantageous strategy. Data fusion of hyperspectral images involves the integration and analysis of information from multiple datasets, such as multiple HSIs from different samples or/and acquired with different spectroscopic techniques (modality). All datasets can be joined in the so-called *multisets* [Tauler et al., 2020]. The use of multiple sources of information is particularly valuable when dealing with HSIs, since multiset analysis provides a more comprehensive and accurate representation of the system under investigation. Generally, data fusion approaches are categorized into three levels, each one using different information from the data sets to be combined [Borràs et al., 2015; Brereton et al., 2017; Smolinska et al., 2019].

- **Low-level** data fusion involves the direct concatenation of signals. For example, a multiset where a block of Raman spectra is connected with fluorescence spectra of analogous samples.
- **Medium-level** data fusion is based on the connection of data blocks represented by their signal features, such as the fusion of the scores of principal components of each individual dataset.
- **High-level** data fusion involves often the combination of model outputs that come from the analysis of individual data blocks. This kind of fusion is more often encountered in certain kinds of data analysis, such as classification methods.

The selection of the data fusion level is based on the nature of the data and the objectives of the analysis. For instance, in chemical systems that often follow simple models as the Beer-Lambert law, the application of low-level fusion models is usually performed since it preserves the information in the original form. Conversely, when considering other scientific domains, such as industrial process control, where the response of multiple and very diverse sensors, such as air flow or temperature and spectroscopic signals, needs to be combined, the use of features extracted from the data allows a better use of the relevant information.

In this thesis, only low fusion level is considered for several reasons. First, most hyperspectral images follow a bilinear model and a concatenation of signals can be treated without modifying the underlying data analysis model. Second, the direct interpretation of the model is very relevant, i.e., preserving the identity of real pure distribution maps and pure spectral profiles is required. Third, the low fusion level strategy adapts to handle directly the output provided by multimodal hyperspectral imaging platforms, able to acquire hyperspectral images using different spectroscopic techniques.

Multiset structures

In the context of low-level image fusion, multisets are defined as single data structures formed by a concatenation of different data blocks. Image multisets can be formed by data from several hyperspectral images collected on different samples with the same spectroscopic technique (Fig. 15A), data acquired on the same sample but with different spectroscopic imaging platforms (Fig. 15B), or by combinations of images from different samples and platforms at the same time (Fig. 15C) [de Juan and Tauler, 2016; Tauler et al., 2020]. The analysis of multisets can be performed by MCR-ALS in the same manner as the analysis of a single HSI described in subsection 2.2.2. However, the pure matrices **C** and **S** will contain information of multiple samples or multiple spectroscopic techniques, depending on the multiset structure.

Figure 15A shows an example of image fusion formed by the concatenation of HSIs from different samples, structured as a column-wise multiset. In this case, the multiset **D_{aug}** is built by multiple HSIs acquired with the same spectroscopic technique and covering the same spectral range. It is important to know that only the spectral direction needs to be common among images. Instead, images connected can have different number of pixels, spatial geometry and spatial resolution. The bilinear model obtained can be expressed as in Eq. 15:

$$\mathbf{D_{aug}} = \begin{pmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \vdots \\ \mathbf{D}_m \end{pmatrix} = \begin{pmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \vdots \\ \mathbf{C}_m \end{pmatrix} \mathbf{S}^{\mathrm{T}} + \begin{pmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \\ \vdots \\ \mathbf{E}_m \end{pmatrix} = \mathbf{C_{aug}}\mathbf{S}^{\mathrm{T}} + \mathbf{E_{aug}} \qquad \text{Eq. 15}$$

where $\mathbf{D_{aug}}$ contains the spectra of the fused *m* images, $\mathbf{C_{aug}}$ the concentration profiles related to every image that can be refolded into distribution maps and $\mathbf{S^T}$ is the single matrix of pure spectra, valid for all images analyzed.

The simultaneous analysis of HSIs of several samples has several benefits, namely:

- The distribution maps of the components in all samples are more easily retrieved because of the complementary information in the multiset, e.g., the map of a minor compound in a sample can be more easily retrieved because this compound may be very well represented in a different sample.
- The pure spectral profiles **S** are better defined since more pixels with more diverse spectral information are being used to estimate **S**.
- And, in general, the rotational ambiguity of the model decreases due to the major diversity of information and the potential inclusion of a higher number of pure pixels in the multiset.

Figure 15B shows an example of image fusion obtained by analyzing the same sample by *n* different spectroscopic image platforms. This is the main type of image fusion studied in this thesis. Here, the multiset is built by the concatenation of HSIs blocks in the row-wise direction. i.e., each pixel is defined by the concatenated spectral responses of the employed imaging techniques.

The bilinear model obtained can be expressed as in Eq. 16:

$$\mathbf{D_{aug}} = (\mathbf{D}_1\,\mathbf{D}_2\ \ldots \mathbf{D}_n) = \mathbf{C}\big(\mathbf{S}_1{}^{\mathrm{T}}\mathbf{S}_2{}^{\mathrm{T}} \ldots \mathbf{S}_n{}^{\mathrm{T}}\big) + (\mathbf{E}_1\,\mathbf{E}_2\ \ldots\ \mathbf{E}_n) = \mathbf{C}\mathbf{S_{aug}}{}^{\mathrm{T}} + \mathbf{E_{aug}} \qquad \text{Eq. 16}$$

This means that the response of the matrix $\mathbf{D_{aug}}$ can be expressed by linear combinations of the pure matrix **S** which contains simultaneously the pure spectra of each spectroscopic technique. The simultaneous analysis of images coming from different platforms in a single multiset structure has several advantages.

- First, the more complete vision of the identity of components in the data set. In this case, the pure profiles on **S** contain chemical complementary information from different spectroscopic techniques and it becomes easier to characterize and differentiate the components present in the sample. For example, if a fluorescence HSI is fused with a Raman HSI, the richer chemical information of Raman (more interpretable and less overlapped bands than fluorescence) facilitates enormously the retrieval and identification of components.

- Secondly, the pure concentration maps **C** are better defined since more diverse spectral information are being used to calculate **C**[5].

  And, in general, the rotational ambiguity of the model decreases due to the major diversity of information and the potential inclusion of a higher number of pure spectral channels in the multiset.

Finally, Fig. 15C shows the multiset $\mathbf{D_{aug}}$ formed by the measurement of different samples scanned by several spectroscopic platforms.

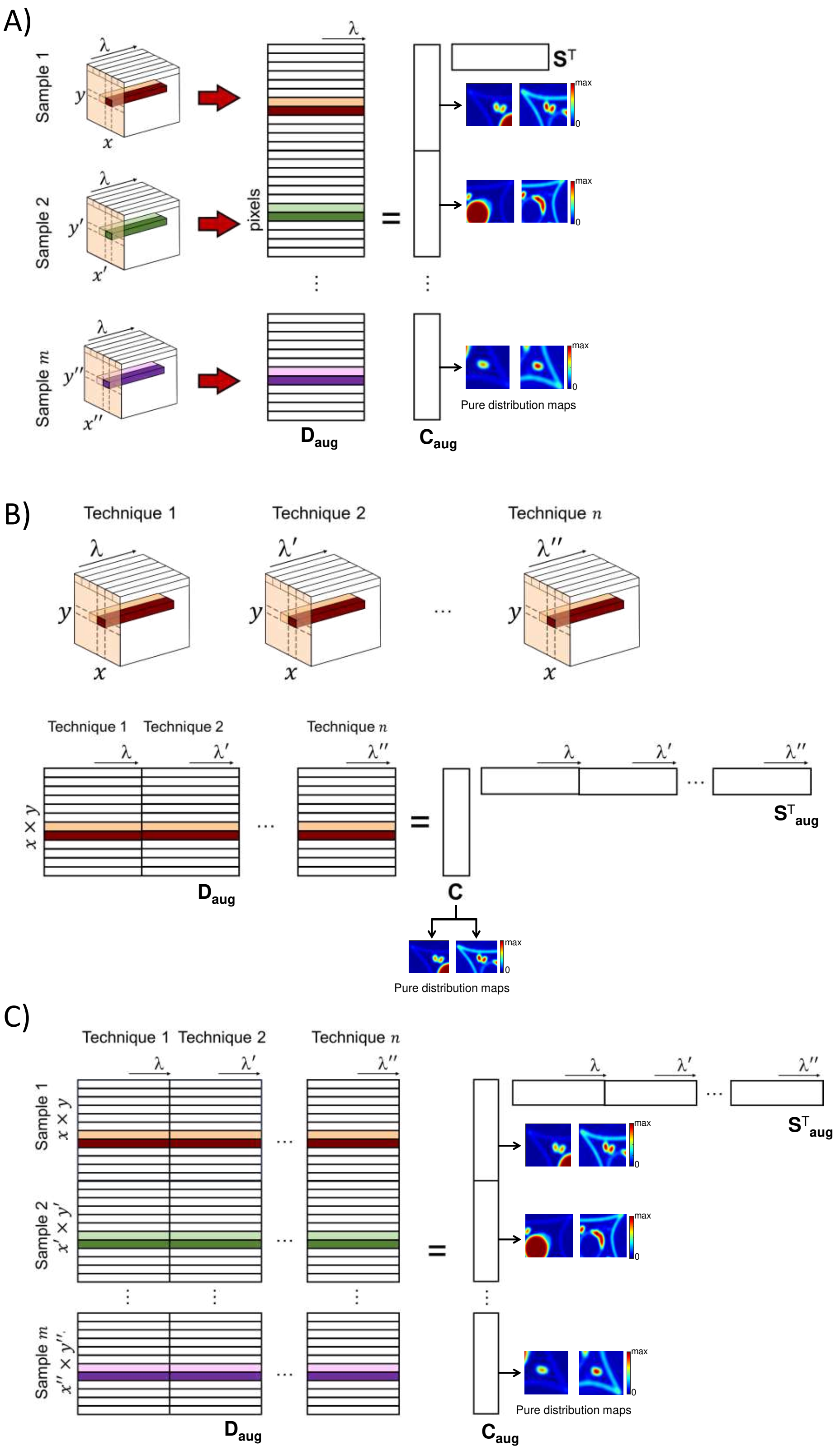


Figure 15. Multiset configurations and related bilinear models. A) Multiset augmented column-wise formed by m hyperspectral images acquired with the same technique. B) Multiset augmented in row-wise direction formed by one sample monitored by *n* spectroscopic image platforms. C) Multiset augmented in row and column-wise direction formed by hyperspectral images acquired on *m* samples by *n* different techniques.

The bilinear model obtained can be expressed as in Eq. 17

$$\mathbf{D_{aug}} = \begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \dots \mathbf{D}_{1n} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \dots \mathbf{D}_{2n} \\ \vdots & \vdots \ddots \vdots \\ \mathbf{D}_{m1} & \mathbf{D}_{m2} \dots \mathbf{D}_{mn} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \vdots \\ \mathbf{C}_m \end{pmatrix} (\mathbf{S}_1{}^{\mathrm{T}} \mathbf{S}_2{}^{\mathrm{T}} \dots \mathbf{S}_n{}^{\mathrm{T}}) + \begin{pmatrix} \mathbf{E}_{11} & \mathbf{E}_{12} \dots \mathbf{E}_{1n} \\ \mathbf{E}_{21} & \mathbf{E}_{22} \dots \mathbf{E}_{2n} \\ \vdots & \vdots \ddots \vdots \\ \mathbf{E}_{m1} & \mathbf{E}_{m2} \dots \mathbf{E}_{mn} \end{pmatrix} \qquad \text{Eq. 17}$$

$$= \mathbf{C_{aug}} \mathbf{S_{aug}}^{\mathrm{T}} + \mathbf{E_{aug}}$$

Here, the benefits from both row-wise and column-wise augmentation can be obtained.

Generally, the multisets based on column-wise augmentation (Fig. 15A) are instrumentally and mathematically accessible and very simple to be built and allow the fusion of HSI with different area scanned, different pixel size and different geometries, since only the spectral mode has to be common among images. However, building multisets based on the row-wise augmentation is not trivial. In this case, the pixel mode has to be common among images, i.e., there should be spatial congruence among the HSIs acquired. This means that the coordinates of the pixels of one HSI should spatially correspond to the same coordinates on the others HSI when they are concatenated [Piqueras et al., 2017]. Thus, the HSIs often need to be spatially transformed to match the pixels among HSIs, e.g., by balancing the pixel size of the images and/or performing translations and rotations. The complexity to perform the data analysis and the often unavailability of two or more spectroscopic imaging techniques to perform measurements makes that this kind of image fusion be insufficiently exploited.

However, although multiplatform image fusion is not yet very common, there are starting efforts of institutions and companies oriented to develop multimodal imaging platforms [Wang et al., 2014; Dochow et al., 2015; Selci, 2019; Wightman et al., 2019; Bec et al., 2020; Tuck et al., 2020; Schie et al., 2021; Neal et al., 2023; Occhipinti et al., 2023]. The shift towards multimodality in imaging is driven by the advantages that it provides, since it significantly enhances the chemical information provided with the need for spatial congruence among signals solved. An example of this technology is the partnership of Renishaw and Becker & Hickl companies, which have recently produced a multimodal FLIM-Raman microscope. This instrument allows measuring simultaneously FLIM and Raman images of the same sample. Another example is the spectrally-resolved FLIM systems (or spectral FLIM systems), where the excitation, emission and fluorescence decay can be recorded simultaneously using the same microscope.

Thus, the increasing interest on multimodal platforms and, in general, on the joint use of different spectroscopic signals in imaging has been the main motivation of this thesis, oriented to provide data analysis solutions for a broad diversity of image fusion scenarios, making them accessible to the scientific community. In the following section, the main studied challenges of image fusion are presented.

2.2.4 Challenges of image fusion

Multisets have been presented as the way to integrate information from different HSIs acquired on different samples by different spectroscopic techniques. However, the explanations provided in the previous section are valid for datasets with images obeying bilinear models and showing similar spatial characteristics. New challenges arise when images to be fused present a higher spatial and spectroscopic diversity. In this instance, working with all available information and proposing joint models that may preserve the underlying natural behavior of the individual images is not yet solved.

Spatial diversity among images is specifically linked to differences in the scanned area and/or in the spatial resolution attained by the different platforms, as displayed in Fig. 16 and 17 and described in the next paragraphs.

Figure 16 illustrates the challenge associated with the fusion of HSIs covering partially different scanned areas. This situation may occur for several reasons, such as the impossibility of measuring the full area to avoid the sample photodamage caused by one of the spectroscopic techniques, or because the acquisition by one of the image platforms is very slow and this measurement is reduced to a small sample area. Fig. 16A shows an example where an emission fluorescence image is acquired scanning all the sample area, while the Raman image restricts to scan a smaller part. The simple solution to address this issue is to discard the non-common scanned area and perform conventional image fusion, as shown in Fig. 15B. To avoid any loss of information, a potential solution is to incorporate the fluorescence pixel spectra related to the non-common scanned area into the analysis by placing them below the rest of the pixel spectra of the same technique (Fig. 16B). Since these pixels do not have equivalent Raman measurements, an empty block is generated in the multiset structure. This type of multiset structure is called incomplete multiset [Alier and Tauler, 2013] because one or more data blocks are missing.

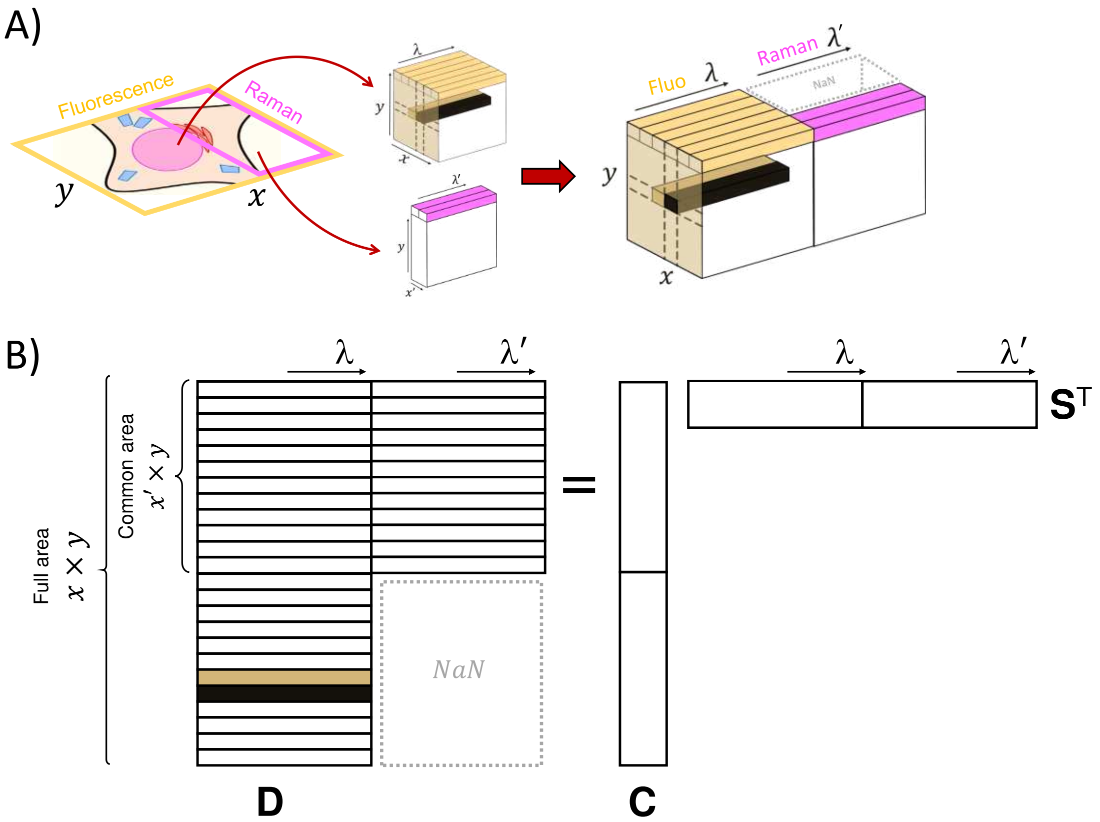


Figure 16. A) The full area of a sample is measured using fluorescence imaging, while only a small region of the scanned area is measured by Raman imaging. The concatenation of all available information in the two HSIs generates an empty block, designing conditions where no Raman data are recorded (set as Not A Number, *NaN*), and a complete region where both signals are available for each pixel. B) The concatenated images are unfolded to form an incomplete multiset. The incomplete multiset contains a full row-wise augmented region, where the fusion of both fluorescence and Raman signals are available per pixel (common scanned area) and a block of fluorescence pixel spectra with a neighboring missing block of information, linked to the area where no Raman signal is recorded.

A similar scenario is encountered when fusion of hyperspectral images acquired by imaging techniques with different spatial resolution is required. This is also a common situation since the pixel size of hyperspectral images depends on the optical characteristics of the spectroscopic technique used. To fuse images with different pixel size, a spatial preprocessing is needed to equal the pixel size among images and allow the image coregistration. Fig. 17A illustrates this image fusion challenge with an example where a fluorescence image with a pixel size of 1 × 1 μm^2 needs to be fused with a Raman image with a pixel size of 4 × 4 μm^2. To be able to fuse both images in the conventional way presented in Figure 15B, the fluorescence image must be binned, i.e., increasing the pixel size four times by summing up pixel spectra of 4 × 4 pixel windows to obtain a new bigger pixel, sized 4 × 4 μm^2, as in the Raman image.

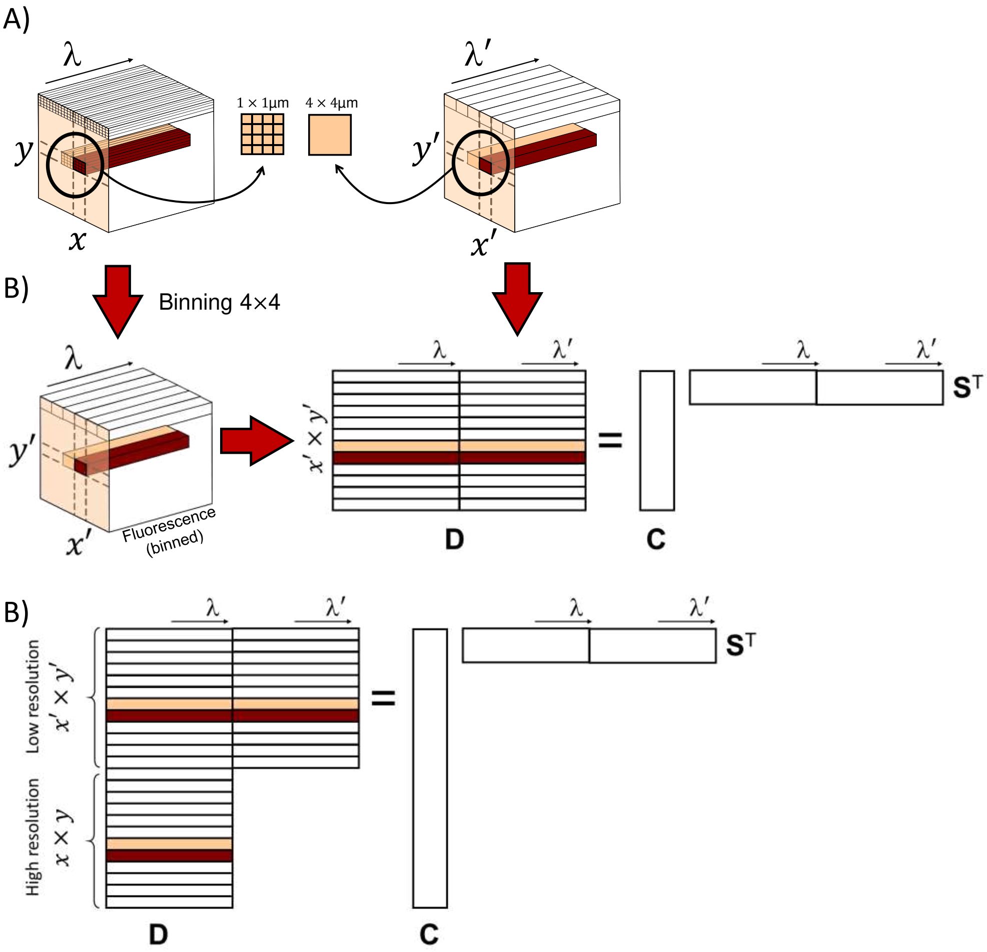


Figure 17. A) Illustrative example of a fluorescence and Raman image with a pixel size of 1x1 µm and 4x4 µm, respectively. The fluorescence image is binned in a factor of 4x4 to match the pixel size of Raman. Thus, a conventional image fusion concatenating both images can be performed. B) The original pixel fluorescence spectra with high spatial resolution can be concatenated below the binned fluorescence spectra. However, this generates an incomplete multiset since Raman spectra with high spatial resolution are not available.

This allows for the concatenation of both (binned) fluorescence and Raman images (Fig. 17A) in a complete multiset. However, binning the spectra of an image has two clear drawbacks. First, if the spatial resolution is decreased, the observation of certain detail in spatial structures of the components can be lost. Second, binning involves summing groups of pixel spectra to obtain bigger pixels with more mixed information. Therefore, pure pixels that help in the MCR-ALS resolution are swallowed in the binned spectrum and the rotational ambiguity of the analysis increases. A way to circumvent the loss of spatial resolution is concatenating the pixel spectra of the original high spatial resolution fluorescence image below their binned version (Fig. 17B). The incomplete multiset obtained preserves all the information and allows retrieving pure resolved maps with high spatial resolution.

Differences in spatial resolution lead to incomplete multisets, as those shown in Fig. 17, if all available image information needs to be taken into account. Therefore, solving these image fusion problems calls for the proposal of new and fast adaptations of the MCR-ALS algorithm to handle multiset structures with one or more missing blocks of information.

The last challenge tackled in this thesis involves the fusion of 3D and 4D images, issued from techniques providing signals with different spectroscopic dimensionalities. Fig. 18 shows an example where a 3D Raman image is fused to a 4D excitation-emission fluorescence image. The challenge is not only related to the construction of a single data structure formed by an unfolded matrix and an unfolded tensor, but to the fact that these image measurements require different underlying linear models, bilinear and trilinear, to be properly described. Integrating information from datasets with distinct dimensional structures poses a challenge in image fusion that can only be solved by proposing algorithms able to accommodate hybrid models, which consider both the bilinear and trilinear behavior of the 3D and 4D images fused.

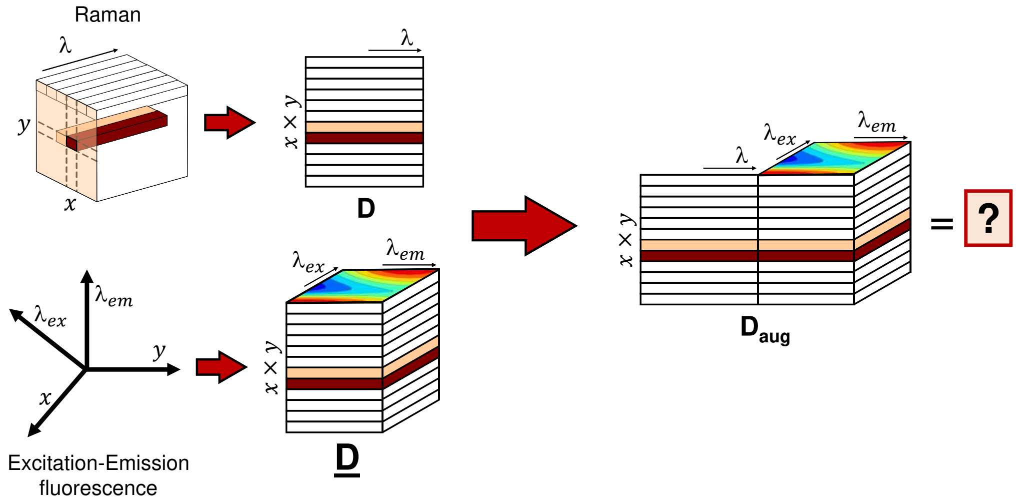


Figure 18. Scheme of image fusion among a Raman HSI and an Excitation-Emission fluorescence HSI. Both images are unfolded and concatenated forming a combined matrix-tensor structure. The block coming from the Raman HSI on $\mathbf{D_{aug}}$ follows a bilinear model, while the block coming from the Excitation-emission HSI follows a trilinear model.

Summarizing, the challenges related to the fusion of images with different dimensionalities and spatial characteristics requires the proposal of new flexible algorithms able to handle incomplete multisets and to perform combined tensor and matrix factorization. Only in this way, a general framework for image fusion will be obtained.

CHAPTER 3. RESULTS AND DISCUSSION

SECTION I – Addressing challenges of fluorescence image analysis

This section contains five scientific publications related to the improvement of the analysis of fluorescence hyperspectral images, specifically excitation-emission fluorescence images and time-resolved fluorescence spectroscopy data, the basic measurement of Fluorescence Lifetime Images (FLIM). The improvement of the analysis of the fluorescence images has a significant impact when image fusion using this kind of measurements is required. New implementations of the trilinear constraint for the analysis of excitation-emission fluorescence images containing missing values and new approaches for the analysis of time-resolved fluorescence spectroscopy data are provided.

3.1 Dealing with missing values on excitation-emission hyperspectral images

Data sets formed by Excitation-emission fluorescence measurements (EEM) are the best-known paradigm of chemical measurements providing a trilinear model, with triads of profiles related to the excitation, emission and sample modes [Stedmon et al., 2008; Rodríguez-Vidal et al., 2020; Alcaraz et al., 2019; Marín-García and Tauler, 2020].

Trilinear models have the interesting property of uniqueness (absence of rotational ambiguity) [Kruskal, 1977] and this makes trilinear decomposition methods very attractive to solve the mixture analysis problem. However, as stated in Chapter 2, the excitation-emission matrices are challenging to analyze due to the presence of non-linear Raman and Rayleigh scattering or the absence of emission signal below the excitation wavelength when EEM measurements are acquired with overlapped excitation and emission wavelength ranges. The scattering phenomena introduce signals that break the inherent trilinear behavior of excitation-emission matrices and must be addressed before analyzing data with trilinear models [Bahram et al., 2006; Eilers and Kroonenberg, 2014]. On the other hand, most algorithms are envisioned to deal with complete EEM landscapes and do not provide clear solutions when a systematic pattern of missing values associated with the lack of signal at emission wavelengths lower than the excitation wavelengths arise. There exist several proposals to deal with the absence of signal below the excitation wavelength and the presence of Raman and Rayleigh scattering in EEM data with benefits and drawbacks. One approach is directly selecting a rectangular region of interest (ROI) on the EEM landscape to avoid Raman and Rayleigh signals. Another is removing Raman and Rayleigh signals and estimate the missing values, or simply removing Raman and Rayleigh signals and set them as missing values [Andersen and Bro, 2003; Bahram et al., 2006; Elcoroaristizabal et al., 2015], as in the right plot of Fig. 19.

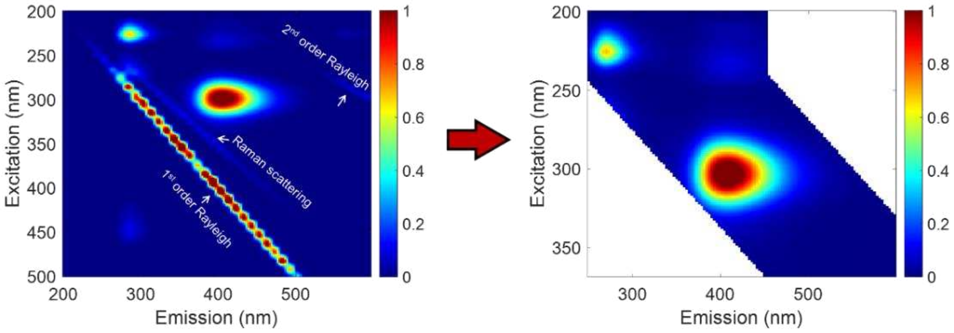


Figure 19. Example of excitation-emission measurement of a mixture of pharmaceutical compounds (ibuprofen and acetylsalicylic acid). On the left, Rayleigh and Raman scattering can be observed in the EEM landscape. On the right, a specific region of the full landscape that contains only fluorescence signal is considered. White areas indicate missing entries.

The algorithms meant to solve the unmixing problem for trilinear data, such as Parallel Factor Analysis-Alternating Least Squares (PARAFAC-ALS) [Bro, 1997] or MCR-ALS with trilinear constraint, cannot be straightforwardly applied in the presence of missing values. For this reason, one of the main efforts in this work has been the adaptation of the trilinearity constraint to handle missing data in MCR-ALS. This results subsection includes the **Publications I**, **II** and **III** and the related discussion.

Publication I is a new implementation of the trilinearity constraint for MCR-ALS to deal with strongly patterned missing data. This initial adaptation is complex and requires a sequential implementation with several steps, such as a submatrix selection, to be implemented.

Publication II offers an improved, simple and easy way to implement the trilinear constraint based on an adapted use of the NIPALS algorithm. This implementation does not require sequential steps, adapts to any pattern of missing values and it can be easily implemented in the MCR-ALS framework.

Publication III shows the application of the adapted trilinearity constraint for the presence of missing values in a photobleaching study of the natural fluorophores of a vegetal tissue. The flexibility of application of the multilinear constraints is tested in a challenging scenario requiring a quadrilinear model.

Publication I. The trilinear constraint adapted to solve data with strong patterns of outlying observations or missing values.
Authors: A. Gómez-Sánchez, I. Alburquerque, P. Loza-Álvarez, C. Ruckebusch, A. de Juan.
Citation reference: *Chemometrics and Intelligent Laboratory Systems* (2022), 231:104692.
DOI: 10.1016/j.chemolab.2022.104692

Publication II. The MCR-ALS trilinearity constraint for data with missing values.
Authors: A. Gómez-Sánchez, R. Vitale, P. Loza-Ávarez, C. Ruckebusch, A. de Juan.
Journal of Chemometrics (2024) (submitted).

Publication III. Study of the photobleaching phenomenon to optimize acquisition of 3D and 4D fluorescence images. A special scenario for trilinear and quadrilinear models.
Authors: A. Gómez-Sánchez, I. Alburquerque, P. Loza-Álvarez, C. Ruckebusch, A. de Juan.
Citation reference: *Microchemical Journal* (2023), 191:108899.
DOI: 10.1016/j.microc.2023.108899

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

The trilinear constraint adapted to solve data with strong patterns of outlying observations or missing values

Adrián Gómez-Sánchez [a,c,*], Iker Alburquerque [a], Pablo Loza-Álvarez [b], Cyril Ruckebusch [c], Anna de Juan [a,**]

[a] Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028, Barcelona, Spain
[b] ICFO-Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology, 08860, Castelldefels, Barcelona, Spain
[c] LASIRE - Laboratory of Advanced Spectroscopy, Interactions, Reactivity and Environment Université Lille, CNRS, UMR 8516, Cité Scientifique, Bâtiment C5, 59000, Lille, France

ARTICLE INFO

Keywords:
Trilinearity
Missing values
Multivariate curve resolution
Constraints
Excitation-emission fluorescence

ABSTRACT

The possibility to perform trilinear decompositions of data sets has the clear advantage of providing unique solutions. Excitation-emission fluorescence matrices (EEM) are the best known paradigm of chemical measurements providing a trilinear structure associated with the configuration of excitation, emission and sample modes. Chemometric tools, such as Parallel Factor Analysis (PARAFAC) and Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) with trilinear constraint, assist in solving the mixture analysis problem by exploiting the trilinear behavior of the EEM measurements. However, the spectroscopic nature of EEM measurements makes that no emission signal can be recorded below the current excitation wavelength, generating a strong and systematic pattern of outlier (zero observations) in EEM data that challenges the classical analysis by MCR-ALS or PARAFAC. Several approaches have been proposed to deal with this problem, such as the identification of outlying values below the excitation wavelength and, thus, the use of data imputation in PARAFAC, but they show severe limitations when systematic outlying data patterns occur. In this paper, we propose a new implementation of the trilinear constraint in MCR-ALS algorithm to cope with EEM measurements where a strongly patterned of outlying data is present. This approach preserves the trilinear property and does not require any data imputation step to replace the outlying observations. Its performance is tested on simulated data, controlled pharmaceutical mixtures and hyperspectral images of a plant tissue (HSI). It should be noted that the approach proposed is applicable to EEM data, where a systematic pattern of outlying observations exist, but can be generalized to the treatment of any trilinear data set with a strong pattern of missing values.

1. Introduction

Excitation-emission fluorescence (EEM) spectroscopy allows characterizing and quantifying fluorophores taking advantage of differences in their excitation and emission profiles [1–4]. EEM spectroscopy provides a full 2D excitation emission matrix or landscape per sample. When EEM from different samples are organized in a single 3D structure, the three dimensions refer to the sample direction (s), excitation direction (ex) and emission direction (em), forming a data cube of size $s \times ex \times em$. In the microscopy field, excitation-emission hyperspectral imaging (EEM-HSI) associates an excitation-emission fluorescence measurement with every pixel and provides 4D images, where two dimensions x- and y- are the pixel coordinates, and the remaining ones correspond to the 2D EEM landscapes, forming a hypercube of size $x \times y \times ex \times em$ [5,6].

In absence of Rayleigh and Raman scatter and for emission ranges higher than the excitation ranges used in the measurement, EEM measurements follow naturally a trilinear model, i.e., every component coming from a set of EEM matrices (i.e. a set of samples) can be expressed as a combination of three different profiles: a concentration profile, which describes the relative abundance of a fluorophore in the different samples, and the associated excitation and emission spectra. When a set of samples is analyzed, a concentration profile describes the relative abundance of a fluorophore in the different samples. When EEM

* Corresponding author. Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028, Barcelona, Spain.
** Corresponding author.
E-mail addresses: agomezsa29@alumnes.ub.edu (A. Gómez-Sánchez), anna.dejuan@ub.edu (A. de Juan).

https://doi.org/10.1016/j.chemolab.2022.104692
Received 14 June 2022; Received in revised form 6 October 2022; Accepted 10 October 2022
Available online 13 October 2022

fluorescence images are studied, the values in a concentration profile refer to abundance of a fluorophore in every pixel. In this context, concentration profiles are refolded to recover the 2D structure of the original image and display distribution maps (Fig. 1).

Characterizing samples or image fluorophores from raw EEM measurements needs suitable chemometric methods that take advantage of the underlying trilinear model of the method. In this scenario, PARAFAC [7,8] and Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) [9,10] become an excellent alley to deal with the mixture analysis problem. Whereas PARAFAC naturally provides trilinear data decomposition as shown in Fig. 1, the underlying MCR-ALS model is bilinear. However, trilinearity can also be applied in the MCR-ALS framework as a constraint [11–13]. Thus, different works have reported on the flexibility to impose the trilinear condition per component or per block in a multiset configuration [5,11,14], offering very versatile scenarios of hybrid bilinear/trilinear models. Hence, MCR-ALS becomes a very good and flexible chemometric tool adapted to the characteristics of the EEM measurement, where the applied trilinear constraint relies on the fact that the fluorescence emission shape of the components remains constant across all the excitation wavelengths [5,12,15]. At this point, it is important to remind that all trilinear decomposition methods provide unique solutions in absence of degeneracies in all modes of the tensor analyzed [16]. This property is an excellent asset when compared with methodologies relying on bilinear decompositions, such as MCR-ALS when the trilinear constraint is not imposed [8,11,13,16].

However, some limitations can be observed in all trilinear decomposition algorithms when dealing with missing data. In this sense, fluorescence measurements have the particularity that no emission signal is produced at wavelengths shorter than the excitation wavelength used. This fact may cause a systematic pattern of zero observations in EEM measurements, linked to the natural fluorescence phenomenon, as can be seen in Fig. 2. In this scenario, an option is selecting a rectangular region of interest (ROI) in the EEM landscape to avoid the regions with absent data. However, data selection may discard relevant information for the characterization of some sample compounds, as shown in Fig. 2, where there is no possible rectangular ROI including information of all sample compounds simultaneously. Another alternative is replacing the outlying observations using data imputation methods, which is the same treatment given to data sets with missing values. During the analysis, the EEM outlying observations (or missing values in a wider context) are replaced by predictions coming from the model itself to avoid algorithm incompatibilities. However, it is difficult to perform a reliable data imputation when the outlying (or missing) values show a strongly patterned structure, such as the one in Fig. 2 [17].

The classical trilinear decomposition methods, Incomplete Data PARAFAC (INDAFAC) or PARAFAC-ALS are well suited to handle missing values when their spatial distribution is random, but not for systematic patterns of missing values, such as the one in Fig. 2. In contrast, MCR-ALS can analyze a complete multiset, obtained by unfolding adequately the data in Fig. 2. However, since the current implementation of the trilinearity constraint in MCR-ALS is not prepared to handle missing values, only a bilinear decomposition would be possible.

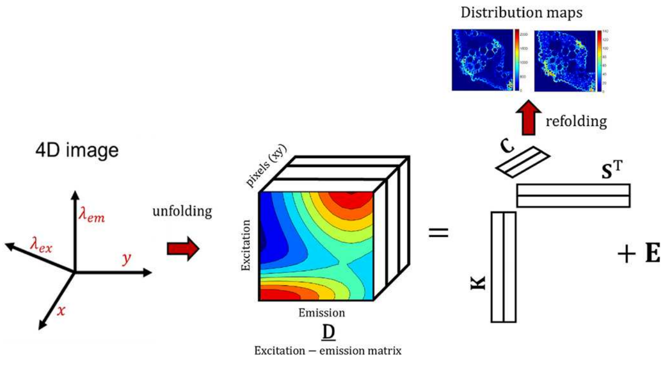


Fig. 1. Underlying model of EEM fluorescence measurements. In a trilinear model, each component has a pure concentration profile (**C**), a pure excitation spectrum (**K**) and a pure emission spectrum (**S**).

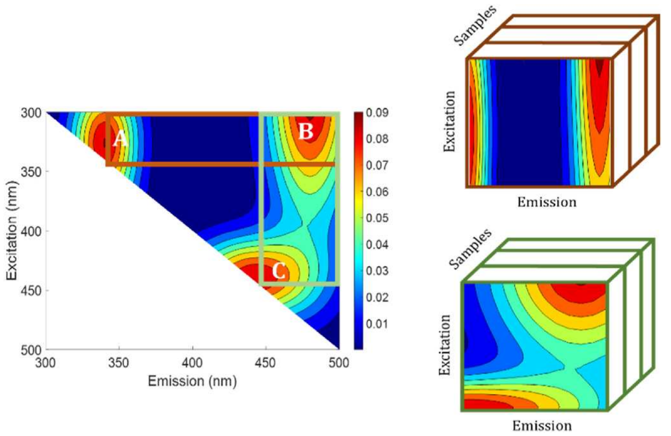


Fig. 2. Possible scenario in a mixture of three components (A, B, C). It can be observed that there is no rectangular ROI that includes the signal from the three components.

In this work, we propose a new implementation of the trilinear constraint in MCR-ALS capable to deal with the presence of systematic-patterned missing values. Our approach can be optionally applied to individual components and does not require any data imputation step. To proof the potential of this new approach, the new constraint has been tested in simulated data, in EEM from controlled pharmaceutical samples and in EEM-HSI from cross-sections of rice roots as examples. It is worth noting that the outlying systematic pattern of EEM fluorescence data will be handled in the same way as a systematic pattern of missing values would be. Hence, on the theoretical description of the proposed approach, the expression missing values will be used because the approach proposed can be applicable to both scenarios.

2. Data sets

This section includes simulated and real examples of EEM measurements. The simulations have been performed mimicking the maps and spectral fingerprints of an EEM-HSI of vegetal tissue and introducing variations related to different noise level and noise structures and structures and to diverse spectral overlap conditions. Examples of real EEM-HSIs and EEM measurements of solution samples of pharmaceutical mixtures are studied.

2.1. Excitation-emission hyperspectral images of plant tissue

2.1.1. Simulated excitation-emission hyperspectral image

The simulated data set is an EEM-hyperspectral image where the shape of the distribution maps is taken from the analysis of a similar real EEM leaf sample image done by the authors on a rice leaf sample [5]. The maps show a considerable overlap among components. In total, the EEM-HSI simulated sample surface has a size of 119 × 119 pixels. The simulated range is from 200 nm to 500 nm with a step size of 6 nm for the excitation wavelengths (51 channels) and from 270 nm to 570 nm with a step size of 6 nm for the emission wavelengths (51 channels), giving a hypercube sized 119× 119× 51× 51. The distribution maps and the different fluorescence excitation and emission fluorescence spectra used for the simulation are shown in Fig. S1 (Supporting Information). Once the image has been obtained, different levels of white or Poisson noise representing 16 and the 30% approximately of the total signal were added, mimicking the usual noise level found in these measurements in normal or harsh conditions, respectively. For more detail in the generation of the simulated data, see the Supporting Information.

2.1.2. Excitation-emission hyperspectral image of plant tissue

Rice plants were grown as in Ref. [5]. After harvest, small pieces of plant roots were collected and embedded in agarose (5% w/w). 50 μm thickness microsections were prepared and put on a 1 mm-thickness CaF_2 slide with a drop of Phosphate-buffered saline solution, covered with a 0.5 mm-thickness CaF_2 coverslip and sealed with nail polish, to avoid water evaporation during the experiment.

EEM-HSIs were acquired by a confocal microscope (Leica TCS SP8 STED 3X, Leica Microsystems, Mannheim, Germany) with an HC PL APO CS2 10 × /0.40 DRY objective. Several excitation wavelengths were selected: 405, 470, 520 and 570 nm. For the 405 nm laser beam, a power approximately of the 70% (89 μW at the sample plane) was used. For the 470, 520 and 570 nm excitations, a supercontinuum white light laser (WLL) with a power approximately of 70% (146 μW at the sample planned) was used.

The emission range for each excitation was 435–663 nm, 495–663 nm, 543–663 nm and 591–663 nm, respectively. The fluorescence spectra were collected using a hybrid photodetector (HYD SMD) with 12 nm sampling interval and a bandwidth of 12 nm. This provides a 4D hyperspectral image with x and y as the spatial directions, and λ_{exc} and λ_{em} as the spectral dimensions. Spectra were collected by point mapping with dwell times of 32 μs in all excitation wavelengths, except for 405 nm, where 15 μs were used. Each of the three images acquired has 1024 × 512 pixels, a pixel size of 450 × 450 nm^2 and a field of view of 460 × 230 $μm^2$.

2.2. Excitation-emission matrices of pharmaceutical mixtures

Nine mixtures of ibuprofen (IP) and acetylsalicylic acid (ASA) (a.r., Sigma Aldrich) were prepared in an ammonia-ammonium chloride buffer solution (pH 10) and measured by an AB2 Aminco-Bowman spectrofluorometer. A common fluorescence linear range was found for the two compounds from 0.25 to 5.00 mg/L ($R^2 = 0.998$). Excitation and emission slits were set to 5 and 10 nm respectively and the voltage of the photomultiplier was set to 560 V. A Hellma quartz cell (4 × 10 mm optical pathlength, and 400 μL volume) was used. The excitation range was 200–500 nm and the emission range was 200–600 nm. Table 1 shows the concentrations of the pharmaceutical compounds in each mixture. The dataset formed by the pharmaceutical mixtures was a data cube formed by 9 samples, 61 excitation channels and 42 emission channels, sized 9 × 61× 42.

3. Data analysis

3.1. Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)

MCR-ALS is an algorithm meant to solve the mixture analysis problem via a bilinear decomposition, which has been applied successfully in many different fields [9,10]. For spectroscopic data, the bilinear model can be expressed by (Eq. (1))

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \qquad \text{Eq. 1}$$

where **D** is the matrix containing all spectra and **C** and $\mathbf{S}^{\mathrm{T}}$ are matrices of concentration profiles and spectral signatures of the sample constituents, respectively. **E** is the matrix of the residual variation unexplained by the MCR model. MCR-ALS is an algorithm that optimizes matrices **C** and $\mathbf{S}^{\mathrm{T}}$ by an alternating least squares iterative procedure under constraints. The end of the optimization procedure is defined by the convergence criterium, often expressed as a threshold based on the relative difference of the lack of fit (LOF) during consecutive iterations. The parameters used to estimate the quality of the MCR model fit are the LOF and the explained variance, as expressed in Eq. (2) and Eq. (3).

$$\mathrm{LOF}\ (\%) = 100 \times \sqrt{\frac{\sum_{i,j} e_{i,j}^2}{\sum_{i,j} d_{i,j}^2}} \qquad \text{Eq. 2}$$

$$\mathrm{Var}\ (\%) = 100 \times \left(1 - \frac{\sum_{i,j} e_{i,j}^2}{\sum_{i,j} d_{i,j}^2}\right) \qquad \text{Eq. 3}$$

Where $d_{i,j}$ is the ij^{th} element of **D** and $e_{i,j}$ is the residual associated with the reproduction of $d_{i,j}$ by the MCR model.

During the iterative process, several constraints can be applied to **C** and $\mathbf{S}^{\mathrm{T}}$ according to the natural behavior of the profiles or responding to mathematical conditions. Constraints can be applied optionally per mode (**C** or $\mathbf{S}^{\mathrm{T}}$), per block in a multiset arrangement and per profile (component) within **C** or $\mathbf{S}^{\mathrm{T}}$. A group of dedicated constraints for multiset data are the model constraints, which incorporate multi-way models, such as multilinear or factor interaction models, in the MCR-ALS decomposition [12,18]. A detailed explanation on the current implementation of trilinearity and the new proposal for strongly patterned missing data sets can be found in the next subsections.

3.2. The trilinearity constraint in MCR-ALS. Implementation for complete data sets and for data sets with strongly patterned missing values

The first step for the implementation of trilinearity in MCR-ALS is transforming the original data cube into a multiset configuration. In complete EEM measurements, where for each excitation wavelength the emission spectrum has the same wavelength range, the tensor $\underline{\mathbf{D}}$ can be unfolded as a data matrix by transforming two dimensions in a single extended one (Fig. 3A). Thus, every row of the multiset contains a vectorized 2D EEM landscape, where the emission spectra of the different excitation wavelengths are concatenated.

In this case, the trilinear model can be implemented as a constraint during the iterations. As shown in previous work [12], in every iteration, each row profile in $\mathbf{S}^{\mathrm{T}}$, related to a specific component, is folded into the excitation-emission matrix $\mathbf{S}_{fi}$, where i refers to the component (Fig. 3B). This new EEM matrix $\mathbf{S}_{fi}$ is decomposed by singular value decomposition (SVD) and it is reconstructed using the first SVD-component. This gives a new matrix $\widehat{\mathbf{S}}_{fi}$ where all the emission profiles have the same shape and only differ in scale, depending on the excitation wavelength they are associated with. The new matrix $\widehat{\mathbf{S}}_{fi}$ is unfolded again and is used to replace the row profile related to component i in $\mathbf{S}^{\mathrm{T}}$. It is important to note that the $\mathbf{S}_{fi}$ matrix needs that every emission spectrum has the same emission range and Raman or Rayleigh scattering must be either removed or corrected to keep the trilinear behavior in the data. As mentioned before, the *per component* implementation of the trilinear

Table 1
Pharmaceutical mixtures.

	Mixture								
Pharmaceutical compound	1	2	3	4	5	6	7	8	9
IP (mg/L)	0.25	1.00	0.25	2.50	0.25	1.00	1.50	1.50	1.50
ASA (mg/L)	1.50	0.50	1.00	0.25	2.50	2.50	0.5	0.25	1.50

All data sets used are available on request.

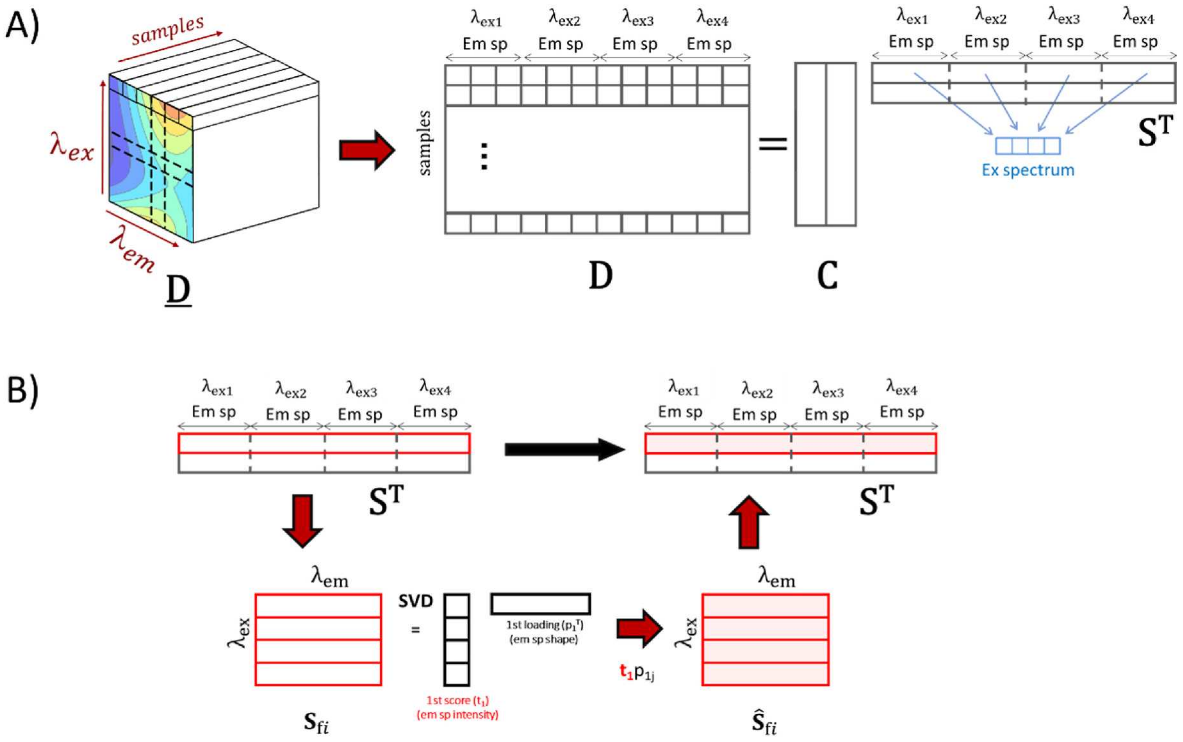


Fig. 3. A) Structure of a three-way complete EEM data set. The cube $\underline{\mathbf{D}}$ is unfolded by concatenating emission spectra at different excitations in matrix **D**. **D** is decomposed into the product of matrix **C**, related to the concentration profiles, and $\mathbf{S}^{\mathrm{T}}$, related to the spectral signatures. B) Classical application of the trilinearity constraint *per component* during MCR-ALS iteration. The spectral profile $\mathbf{S}^{\mathrm{T}}$ of one component *i* is folded as an EEM matrix ($\mathbf{S}_{fi}$) and decomposed by SVD. Then, it is reconstructed by the first component of the SVD analysis ($\widehat{\mathbf{S}}_{fi}$) and the suitable values of $\mathbf{S}^{\mathrm{T}}$ are replaced.

constraint allows obtaining full trilinear models (when all components are constrained) or hybrid bilinear/trilinear models when only some of them obey this model condition.

The excitation spectrum is recovered using the area of the pure fluorescence emission for each excitation wavelength (Fig. 3A). Note that for each component, every emission spectrum has the same shape. This gives us a trilinear model, where for each component there is a concentration, an excitation and an emission profile.

3.2.1. *Trilinearity constraint for data with strongly patterned missing values in Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)*

When using EEM measurements with overlapping excitation and emission wavelength ranges, a strongly patterned missing data set, as shown in Fig. 2, is obtained. However, transforming this cube into a multiset configuration can be done as in Fig. 3A. The main difference in this case is that the length of the emission spectra concatenated per every pixel changes depending on the related excitation wavelength (Fig. 4A). The multiset **D** obtained does not contain missing data and could be easily analyzed using a bilinear model decomposition. However, the uniqueness in solutions provided by the trilinear constraint would be lost. In the classical application of the trilinear constraint shown in Fig. 3B, the matrix $\mathbf{S}_{fi}$ must be complete and cannot have missing values. In EEM measurements where the excitation and the emission range overlap, every emission spectrum has a different length and this prevents applying the trilinear constraint, as shown in Fig. 4A, because $\mathbf{S}_{fi}$ would be a ragged matrix.

To solve the problem described in the previous section, the following procedure is proposed (Fig. 4B). In each iteration, after the $\mathbf{S}^{\mathrm{T}}$ matrix is calculated, each profile of $\mathbf{S}^{\mathrm{T}}$ is folded as a ragged matrix, filling the empty spaces with NaN and refolding the data as in the original structure. The idea is applying the trilinear constraint to a suitable number of complete $\mathbf{S}_{fi}$ submatrices until all elements in the original $\mathbf{S}^{\mathrm{T}}$ matrix are used. As a result, the trilinear profiles are reconstructed sequentially without any imputation step. In the example of Fig. 4B, the criterion chosen was selecting the rectangular submatrices according to the number of rows covered in decreasing order. Thus, the green submatrix is the first detected and is decomposed by SVD and reconstructed using the first component. The corresponding values of $\mathbf{S}_{fi}$ are replaced by the reconstructed submatrix (green colour in Fig. 4B). Then, the second submatrix, in purple, is detected and the same decomposition is applied, replacing only the values in the $\mathbf{S}_{fi}$ matrix that were not modelled by the previous submatrix analysis. This is repeated sequentially with all possible additional submatrices until all the area of $\mathbf{S}_{fi}$ considered for trilinearity is covered ($\widehat{\mathbf{S}}_{fi}$). The matrix $\widehat{\mathbf{S}}_{fi}$ is then vectorized by concatenating the emission spectra at the different excitation wavelengths to replace the *i*th suitable profile of the $\mathbf{S}^{\mathrm{T}}$ matrix. There are several ways to sort the submatrices used to describe $\widehat{\mathbf{S}}_{fi}$. Each correspond to different criteria (bigger area, bigger number of row or columns …). The criterium to sort the submatrices in an optimal way will be discussed later.

3.2.1.1. *Optimal submatrix selection.*

When running the MCR-ALS algorithm, only under non-negativity, every spectral profile in $\mathbf{S}^{\mathrm{T}}$ is likely to contain slightly mixed contributions. In this scenario, the trilinearity constraint should aim first at removing this initial mixed profile nature and afterwards to provide a common emission shape associated with all excitation wavelengths. Hence, the selection of the submatrices $\mathbf{S}_{fi}$ on which to apply sequentially trilinearity will consider this double goal in two steps.

Step 1 (removal of signal mixing in $\mathbf{S}_{fi}$)

An automated algorithm was designed to detect all possible rectangular submatrices in the ragged matrix $\mathbf{S}_{fi}$. These submatrices are afterwards sorted in decreasing order according to their mixture level

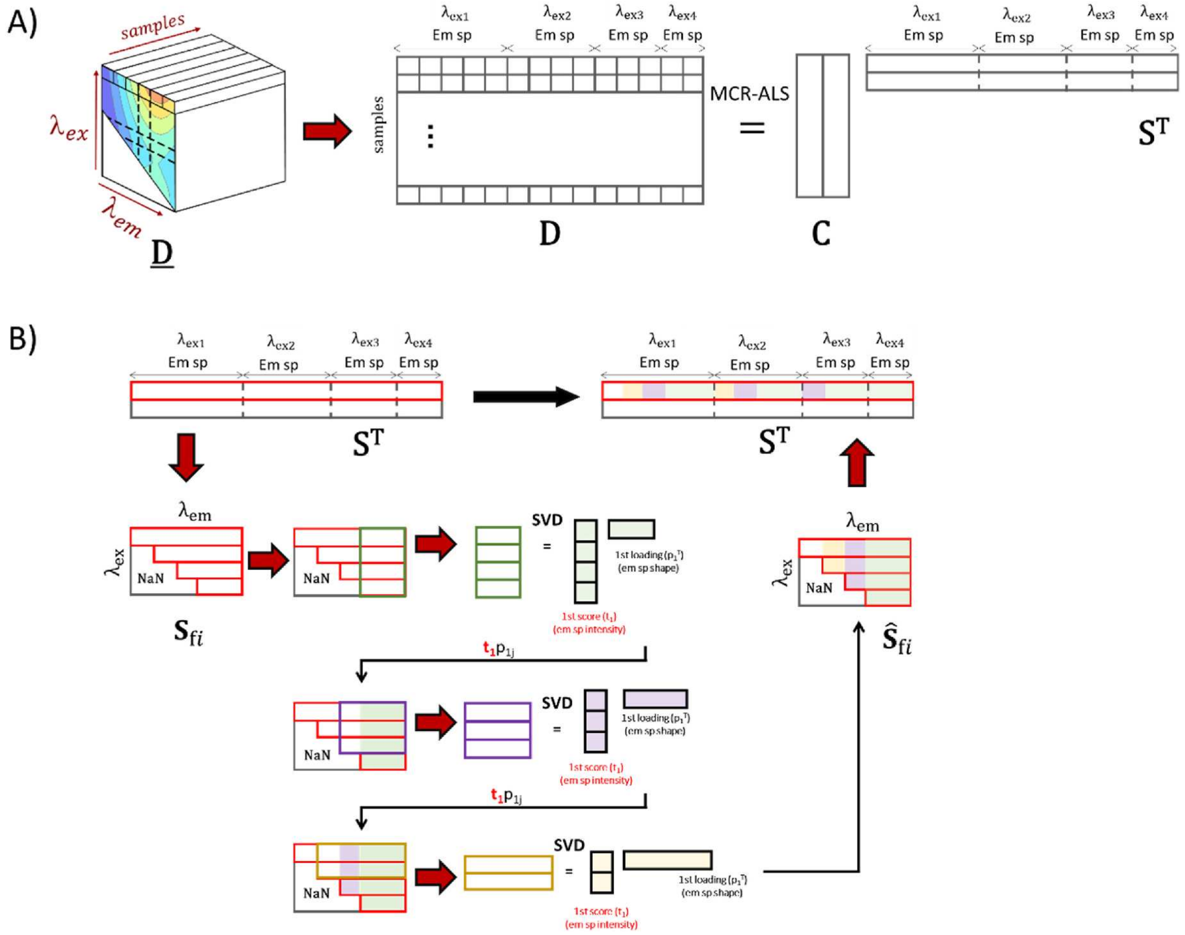


Fig. 4. A) Structure of a three-way EEM cube with systematic missing value pattern. The cube $\underline{\mathbf{D}}$ is unfolded by concatenating different excitations in the matrix **D**. B) Trilinearity constraint for irregular EEM measurements. The spectral profile $\mathbf{S}^T$ of the ith component is folded as an EEM ragged matrix ($\mathbf{S}_{fi}$). In this example, the algorithm sorts in decreasing order the rectangular submatrices according to the maximum number of rows included until the full $\mathbf{S}_{fi}$ matrix is covered. Sequentially, the submatrices are submitted to SVD and the first component is used for reconstruction ($\hat{\mathbf{S}}_{fi}$) and replacement of the suitable elements in the matrix $\mathbf{S}_{fi}$. Any new submatrix analysis only replaces values that were not modified by previous submatrix analyses. Finally, when all the ragged $\mathbf{S}_{fi}$ matrix has been covered by the different submatrix analyses, the matrix $\hat{\mathbf{S}}_{fi}$ is vectorized by concatenating the excitation dimension to replace the *i*th profile of the $\mathbf{S}^T$ matrix.

(ML), estimated as the trace of the submatrix $\widetilde{\boldsymbol{\Sigma}}$, defined as the diagonal matrix containing the eigenvalues $\boldsymbol{\Sigma}$ divided by $\boldsymbol{\Sigma}_{11}$ and with *N* as the number of components (Eq. (4)).

$$\mathrm{ML} = \frac{trace(\widetilde{\boldsymbol{\Sigma}})}{N} \qquad \text{Eq. 4}$$

ML can move from 1/N for a perfect rank one matrix (i.e. in a noiseless case, when only a pure component exists) to one, when the variance is evenly spread in all calculated components. The closer ML is to 1, the higher the mixture level in the analyzed submatrix.

SVD is applied first to the most mixed submatrix of $\mathbf{S}_{fi}$, framed in green color. The reconstructed submatrix only using the first component helps to remove the non-common signal features that could come from residual contributions of other compounds. The procedure continues gradually, every time taking the most mixed remaining submatrix (following the purple and yellow sequence in Fig. 5), doing the SVD analysis and incorporating only the reconstructed part of the submatrix absent in previous steps, until the full area of the $\mathbf{S}_{fi}$ ragged matrix has been covered. This algorithm is fast and automatic since it does not require to set any parameter.

Step 2 (ensuring trilinear profiles)

The first step described above helps to 'clean' the original mixed contributions in $\mathbf{S}_{fi}$; however, the emission profiles associated with every excitation step may be slightly different because the reconstructed values in each emission channel may come from different submatrix reconstructions. To obtain perfect trilinear profiles, a second step of sequential application of trilinearity is done taking now submatrices sorted as in Fig. 4B.

It is important to note that the procedure presented in section 3.2.1 is useful to apply the trilinearity constraint to ragged matrices with any kind of pattern of missing values without any step of value imputation. As all other constraints in MCR-ALS, this constraint can be applied to all or to specific components of the data set. The current implementation proposed does not show limitations neither in terms of number of components of the system nor in profile overlap. However, it needs to be noted that the step of computation of the submatrices covering the EEM landscape increases in computation time when the landscapes treated have a high number of excitation and emission channels. In any case, though, even with hundreds of channels in each direction, a desktop computer would be sufficient to perform this task. In this situation, a previous binning in the spectral direction can alleviate problems of computation time.

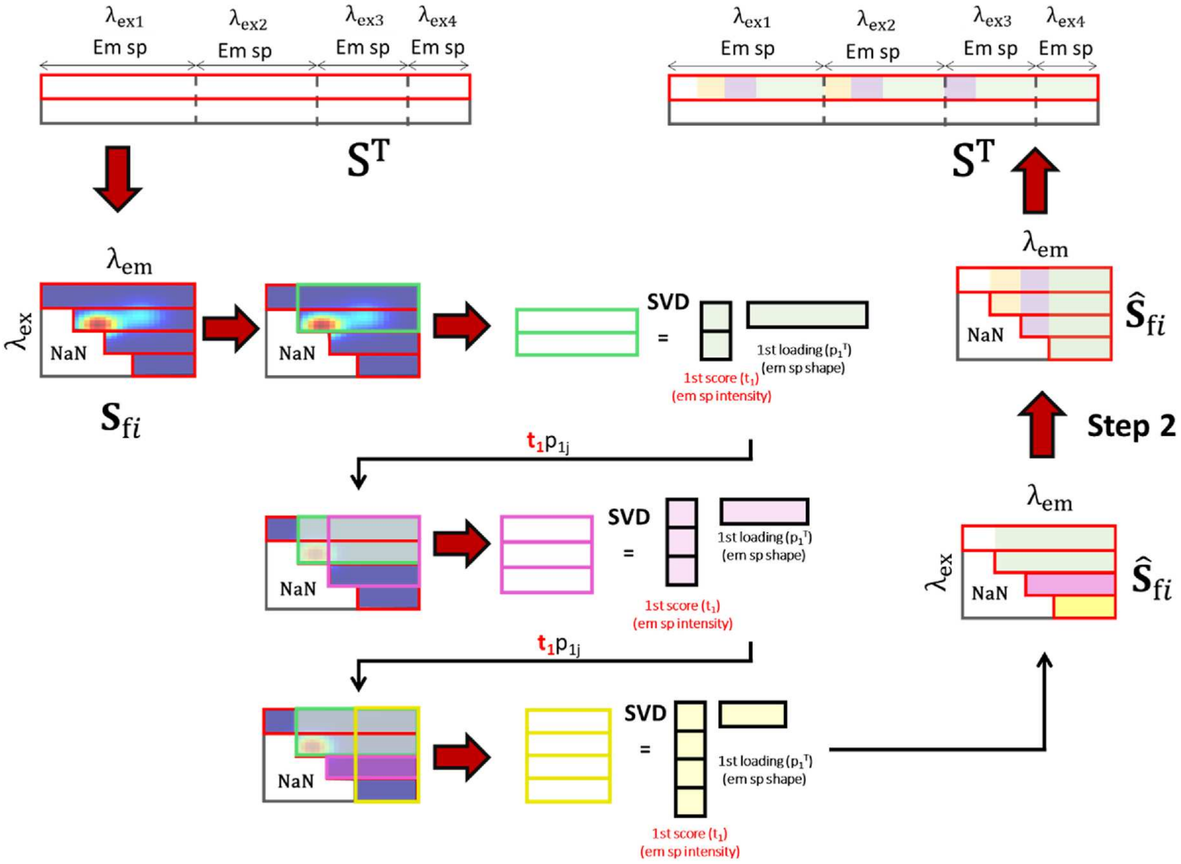


Fig. 5. Application of trilinearity constraint on irregular EEM measurements. The spectral profile $\mathbf{S}^T$ of the ith component is folded as an EEM ragged matrix ($\mathbf{S}_{fi}$). In this example it is possible to see a minor signal contribution from another component in this pure profile. The algorithm sorts in decreasing order the rectangular submatrices according to the mixture degree ML. Sequentially, the submatrices are submitted to SVD and the first component is used for reconstruction ($\widehat{\mathbf{S}}_{fi}$) and replacement of the suitable elements in the matrix $\mathbf{S}_{fi}$. Any new submatrix analysis only replaces values that were not modified by previous submatrix analyses. Finally, when all the ragged $\mathbf{S}_{fi}$ matrix has been covered by the different submatrix analyses, the matrix $\widehat{\mathbf{S}}_{fi}$ is vectorized by concatenating the excitation dimension to replace the *i*th profile of the $\mathbf{S}^T$ matrix and the algorithm continues to step 2.

3.3. Software

The PARAFAC method was used as implemented in the N-way Toolbox for MATLAB. Version 3.31 [19]. MCR-ALS was applied using in-house coded routines that incorporated the new trilinearity implementation.

4. Results and discussion

4.1. EEM HSI data sets. Simulated and real images

In both simulated and real image data sets, non-informative background pixels were removed to reduce the data set size. In the real EEM-HSIs, images were spatially binned using a 3 × 3 factor to increase the signal-to-noise ratio and the emission channels 585–597 nm were removed due to the presence of an instrumental artefact.

The potential of the methodology presented was first tested on the simulated 4D images. The simulated data were analyzed in three different ways: using MCR-ALS applying a bilinear model, using MCR-ALS with the adapted trilinearity constraint for strongly patterned missing data and using a PARAFAC-ALS model. To analyze the simulated dataset by MCR-ALS, the 4D image was unfolded according to Fig. 4A. During the iterative optimization, non-negativity constraint was applied to both modes. In all MCR-ALS analyses, initial spectral estimates were obtained by a SIMPLISMA-based algorithm [20]. For the application of the PARAFAC-ALS model, only the pixel spatial dimensions were unfolded, as shown in Fig. 1B. The PARAFAC-ALS algorithm was applied using SVD as the method to provide the initial estimates. Non-negativity was applied in the three modes. The imputation proposed by the algorithm (based on an expectation-maximization approach) was used to estimate missing data [17]. In all PARAFAC-ALS and MCR-ALS models, the maximum number of iterations was set to 5000 and the convergence criterion based on differences in error among consecutive iterations was 10^{-6}%. Results are summarized in Table 2.

A first observation is that all bilinear and trilinear models for this data set provided a very similar lack of fit for all cases, which is in good agreement with the amount of noise added in the simulation. This means that the noise is well separated from the signal and no local minima are reached in any of the analyses presented. Actually, when trilinearity is an appropriate constraint, it is not expected a strong variation in the variance explained between bilinear and trilinear models [13,21].

The assessment of the quality of the models was also checked by observing the correlation coefficients between the concentration profiles and pure EEM landscapes recovered by the applied algorithm and the corresponding true solutions, for the different models. It should be noted that for the comparison of EEM landscapes, only non-imputed values of PARAFAC-ALS model were considered. Fig. 6 displays the excitation and emission profiles recovered by the models tested (black lines) overlaid with the profiles used for simulation (red dotted lines). All excitation and emission profiles obtained by MCR-ALS without trilinearity were plotted. The excitation and emission profiles for MCR-ALS analysis with trilinearity were extracted plotting the longest spectrum associated with

Table 2
Lack of fit (LOF) and correlation coefficients among recovered solutions and true solutions for the different data sets and models tested.

System	Profile overlap	Noise (%) (structure)	Component	MCR-ALS (bilinear model)			MCR-ALS (trilinearity for missing data)			PARAFAC-ALS		
				C profile(+)	S profile(+)	LOF (%)	C profile (+)	S profile(+)	LOF (%)	C profile(*)	S profile(*)	LOF(*) (%)
1	Low	16.1 (White)	1	0.99	1.00	16.1	1.00	1.00	16.1	1.00	1.00	16.1
			2	1.00	0.99		1.00	1.00		1.00	1.00	
			3	0.99	1.00		1.00	1.00		1.00	1.00	
2	High	16.5 (White)	1	0.97	0.99	16.5	0.99	1.00	16.5	0.89	0.25	16.5
			2	1.00	0.99		1.00	1.00		0.99	0.99	
			3	0.96	0.82		0.98	0.99		0.96	0.67	
3	Low	31.0 (White)	1	1.00	1.00	31.0	1.00	1.00	31.0	1.00	1.00	31.0
			2	0.99	0.99		1.00	1.00		1.00	1.00	
			3	1.00	1.00		1.00	1.00		1.00	1.00	
4	High	31.8 (White)	1	0.93	0.71	31.8	0.98	1.00	31.8	0.87	−0.07	31.8
			2	0.98	0.95		0.99	0.99		0.95	0.96	
			3	0.90	0.21		0.95	0.98		0.96	0.81	
5	High	28.5 (Poisson)	1	0.96	0.90	28.4	0.98	1.00	28.5	0.89	0.45	28.5
			2	0.99	0.97		0.98	1.00		0.95	−0.30	
			3	0.93	0.74		0.96	0.99		0.64	0.36	

* Missing values in PARAFAC-ALS are estimated using the imputation of the algorithm. The imputed values are not considered neither for the calculation of the correlation coefficients in the pure EEM landscape nor in the lack of fit.

+ Correlation coefficients between recovered profile by MCR-ALS and simulated profiles.

emission and excitation wavelengths from the resolved EEM landscape, respectively. Only the results associated with system 5, the worst case in terms of noise level and profiles overlap, are shown for illustration purposes. As can be seen, the recovered profiles by the bilinear MCR-ALS model and the PARAFAC-ALS model are not satisfactory, especially for component 3.

The variability in the emission and excitation profiles recovered by the bilinear MCR-ALS model with only non-negativity constraints can be explained by the strong profile overlap existing among components and the associated rotational ambiguity (see Supporting material for systems 1 and 5). Instead, the cause of the poor recovery of some profiles by PARAFAC-ALS is due to the data imputation required when trilinearity is imposed, more prone to fail when a systematic pattern of missing values is present [10]. In contrast to the two approaches mentioned, MCR-ALS with the modified trilinear constraint retrieves very accurately the true profiles. The improvement in the solutions is due to both the trilinear property, which suppresses the rotational ambiguity [11–13, 21,22], and to the fact that no data imputation is required. As a consequence, the strong pattern of missing data does not affect the quality of the final results. These results confirm that, even if a very huge number of patterned missing values is present, the true solutions can be correctly reached with the presented novel approach.

In the following real example of EEM-HSI image, described in section 2.1, only MCR-ALS will be used either using a bilinear model or the modified implementation of the trilinear constraint. In this case, the benefit of the trilinear constraint is obtaining more accurate results and, hence, improving the interpretability of the components obtained.

The real data set consists of three hyperspectral EEM images from rice root cross sections. Fig. 7 shows the global intensity map (the total fluorescence counts in each pixel) of one of the samples and its global intensity EEM (the total fluorescence counts in each spectral channel).

Each 4D image was unfolded following Fig. 4A scheme. The blocks of pixel spectra of every image were put one on top of each other to form a multiset. As a result, after MCR analysis, the matrix **C** provides concentration profiles for every component in the different samples, which can be refolded into distribution maps. The matrix $\mathbf{S}^\mathbf{T}$ contains their related stretched emission spectra, which can be refolded into the pure 2D EEM landscapes (see scheme of the multiset configuration in the support information). The multiset described was analyzed by MCR-ALS using only non-negativity constraints and a bilinear model and by MCR-ALS using non-negativity and the modified trilinear constraint. In all analyses initial spectral estimates were obtained by SIMPLISMA algorithm and the maximum number of iterations was 500 (a convergence criterion was 10^{-8}%). Four different components were detected. Fig. 8 shows the pure excitation-emission matrices of the root compounds and the distribution maps for one of the three root samples found by MCR-ALS using a bilinear and a trilinear model. The complete MCR-ALS results of the multiset analysis are shown in the Supporting Information.

The explained variance of the bilinear and the trilinear model were 99.0% and 98.9%, respectively, confirming that the trilinear model is suitable to analyze this kind of data.

When comparing the results obtained by both approaches, components 1 and 3 are well resolved in both models since no differences are present in the pure EEM landscapes and maps. However, components 2 and 4 show clear changes in the emission spectra shapes associated with the different excitation wavelengths when bilinear models are used, a clear sign that rotational ambiguity is affecting the results. This ambiguity is known to affect not only the EEM landscapes but also the structure of the distribution maps. Therefore, interpretation of the biological information extracted will be performed from the results shown in Fig. 8B.

The components recovered by the trilinear model have a clear biological meaning. The first component is strongly related to the root cortex and the stele. The emission maximum can be observed at 441–453 nm and the excitation providing the highest signal is 405 nm. Based on the location and spectral characteristics of this component, it can be assigned to a type of non-specific lignin or phenolic compounds, normally observed in all the vegetal tissue [23]. The third component is related to the sclerenchyma layer of the epidermis and the inner part of the stele. The emission maximum is at 489–501 nm and the maximum excitation is at 405 nm. This component could be strongly related with lignin since both root zones are highly lignified cells and the emission maximum matches with the maximum reported in literature [24]. To the knowledge of the authors, the second and fourth components have never been reported based on autofluorescence measurements, probably due to the difficulty to extract a clear signal from the raw EEM measurement. Fig. 8B shows that the second component is closely related to the inner cortical and sclerenchyma layer of the root exodermis. The emission maximum is found at 573–585 nm and the maximum excitation signal is observed at 570 nm. To the best of our knowledge, identification of the inner cortex was only reported once, by inmunoprofiling [25]. Likewise, the fourth component could be related to the Casparian strip, and the sclerenchyma layer of the epidermis and it is also present in the phloem.

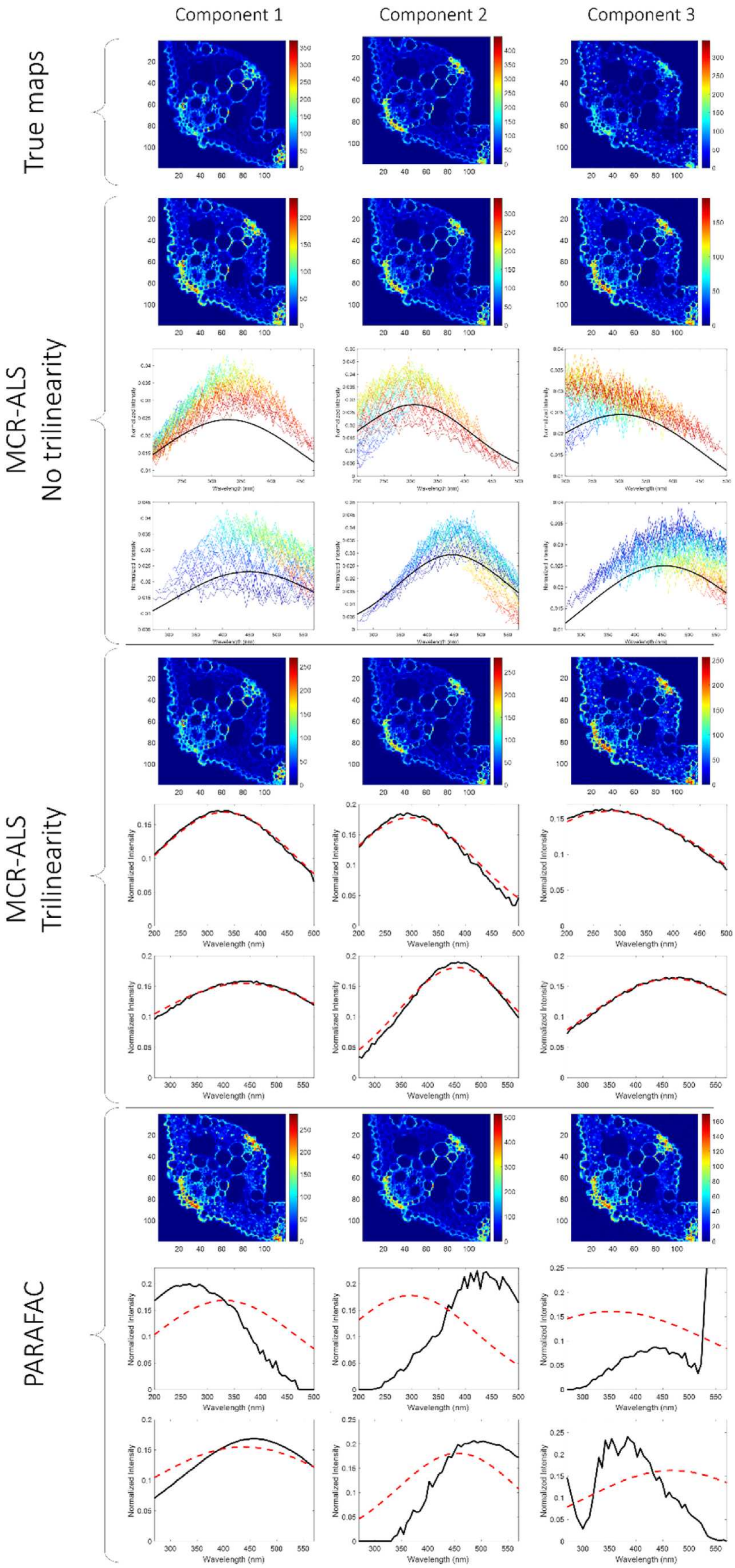


Fig. 6. Results of the analysis for system 5. A) True distribution maps. B) Distribution maps (first row), excitation (second row) and emission profiles (third row) for each excitation provided by MCR-ALS without applying trilinearity constraint. The excitation profiles were extracted following the scheme of Fig. 3A. C) Distribution maps (first row), excitation (second row) and emission profile (third row) solutions provided by MCR-ALS applying trilinearity constraint. D) Distribution maps (first row), excitation (second row) and emission profile (third row) solutions provided by PARAFAC-ALS. Black lines are solutions provided by the respective models. Red dotted lines are the true solutions. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

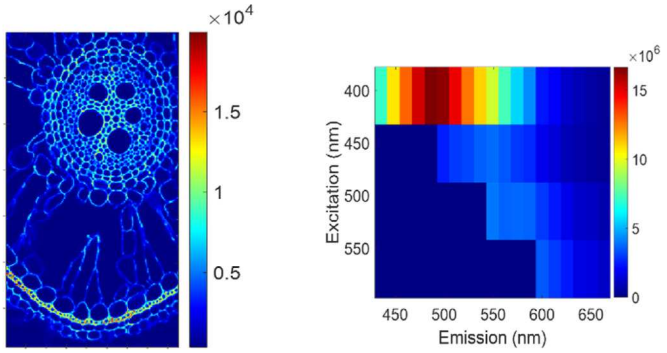


Fig. 7. Global intensity map (left) and global intensity EEM (right) of the hyperspectral image. Scale units refer to fluorescence counts.

The maximum emission can be observed at 537–549 nm and the highest excitation signal is at 470 nm. The location of this component is related to the presence of the Casparian strip and matches well with the results reported in literature [26].

4.2. Real pharmaceutical mixtures

The new implementation of trilinearity was tested to analyze controlled mixtures of ibuprofen and acetylsalicylic acid in harsh conditions for the MCR-ALS algorithm. In this case, the signal contributions of the two compounds have more than one order of magnitude of difference between them and no pure sample is present in the dataset. Several mixtures were prepared using IP and ASA as described in Section 2.1.

As a previous step to the analysis, a ROI was selected from the 2D EEM landscape of each sample so that the useful fluorescence signal of the two compounds was included and the zones with Rayleigh and Raman scattering were discarded, as it can be seen in Fig. 9A. It is important to note that the ROI selected does not have a rectangular shape and that this does not preclude the application of the trilinearity constraint, as described in section 3.2.1. The dataset is formed by equally shaped ROIs from nine mixture samples, covering 30 excitations and 33 emission channels. Pure EEM of ibuprofen and acetylsalicylic acid are shown in Fig. 9B.

The dataset was analyzed by MCR-ALS using a bilinear model and non-negativity constraint and with non-negativity and the modified

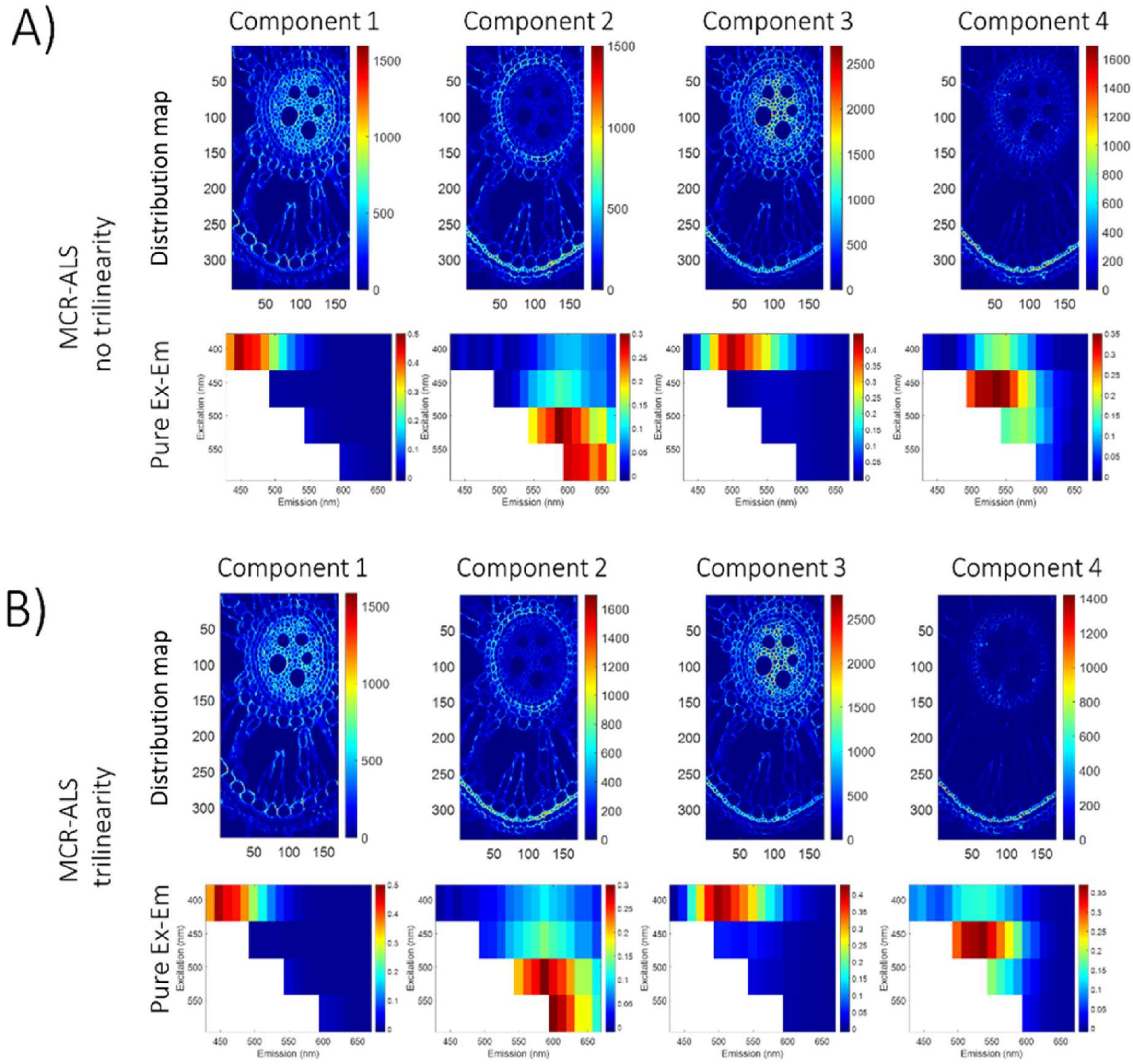


Fig. 8. A) Predicted concentration maps (of sample 1) and pure EEM profiles found by MCR-ALS without trilinearity constraint. B) Predicted concentration maps (of sample 1) and pure EEM profiles found by MCR-ALS with trilinearity constraint. Scales in distribution maps and pure 2D EEM landscapes are concentrations and fluorescence intensities in arbitrary units.

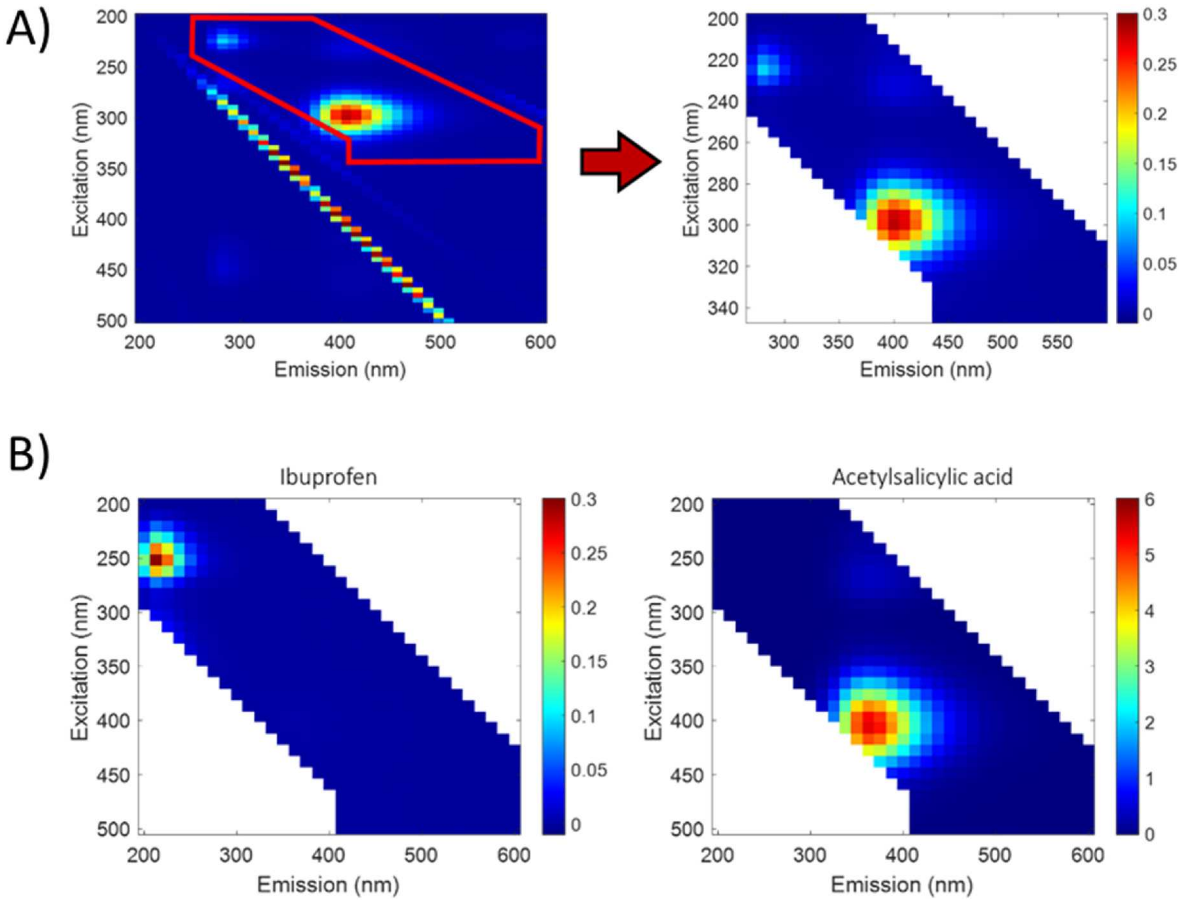


Fig. 9. A) ROI selected from a mixture of IP and ASA, discarding the Raman and Rayleigh dispersion. B) Pure excitation-emission matrices of IP and ASA at the same concentration (5 mg/L) (note the high difference in the fluorescence signal magnitude).

trilinearity constraint. Initial estimates and the convergence criterion were set as in previous examples.

Table 3 shows the lack of fit of the different models tested and the correlation coefficients between the recovered concentration profiles and pure EEM landscapes obtained with the models and the true profiles. As in previous examples, the lack of fit is similar between the bilinear and the trilinear model.

Fig. 10 shows the pure 2D EEM landscapes and the comparison between true and recovered concentration profiles for the two compounds of the samples. If only non-negativity constraint and a bilinear model is applied, the pure EEM recovered for IP is not correct and the effect of ambiguity is also perceived in the concentration profiles of the two compounds. Indeed, Fig. 10A shows that high contributions of ASA are present in the pure spectral landscape of IP, seen also in the low correlation coefficient, equal to 0.2, between the true solution and the MCR-ALS profiles for this component. Although the correlation coefficients for the concentration profiles of IP and ASA are 0.99 and 1.00, respectively, a certain bias between the real and the recovered concentrations can also be seen. Instead, the use of the MCR-ALS method with the adapted trilinear constraint provides excellent solutions for the concentration profiles and pure fluorescence EEM landscapes (see Fig. 10B), due to the uniqueness associated with this kind of data decomposition.

Table 3
Correlation coefficients between MCR-ALS profiles and true profiles for the different models tested.

Model	Component	Concentration profile	Pure landscape profile	Lack of fit (%)
MCR-ALS without trilinearity constraint	ASA IP	1.00 0.99	1.00 0.20	0.69
MCR-ALS with trilinearity constraint for incomplete data	IP ASA	1.00 1.00	1.00 0.98	0.73

5. Conclusions

The new implementation of the trilinear constraint in MCR-ALS for EEM data sets with strongly patterned outlying data surmounts the limitations linked to data imputation when natural trilinear decomposition methods are applied, and the ones related to the rotational ambiguity associated with the multiset analysis carried out on the unfolded three-way data cube, when a classical MCR-ALS bilinear model is applied.

The trilinear profiles obtained with this method are issued from SVD analyses performed in a sequential way on complete submatrices issued from the ragged 2D matrix that contains the emission profiles forced to show the same shape. This sequential approach allows obtaining trilinear profiles without requiring any data imputation step that are subsequently submitted to the MCR-ALS optimization. In this manner, the ambiguity associated with MCR bilinear decompositions is also suppressed. An additional advantage of the implementation of this constraint is that it is not restricted to the triangular pattern related to the nature of EEM measurements data, but to any other kind of systematic pattern of missing values that may be encountered in the initial ragged matrix to be constrained.

The value of this constraint implementation has been validated on

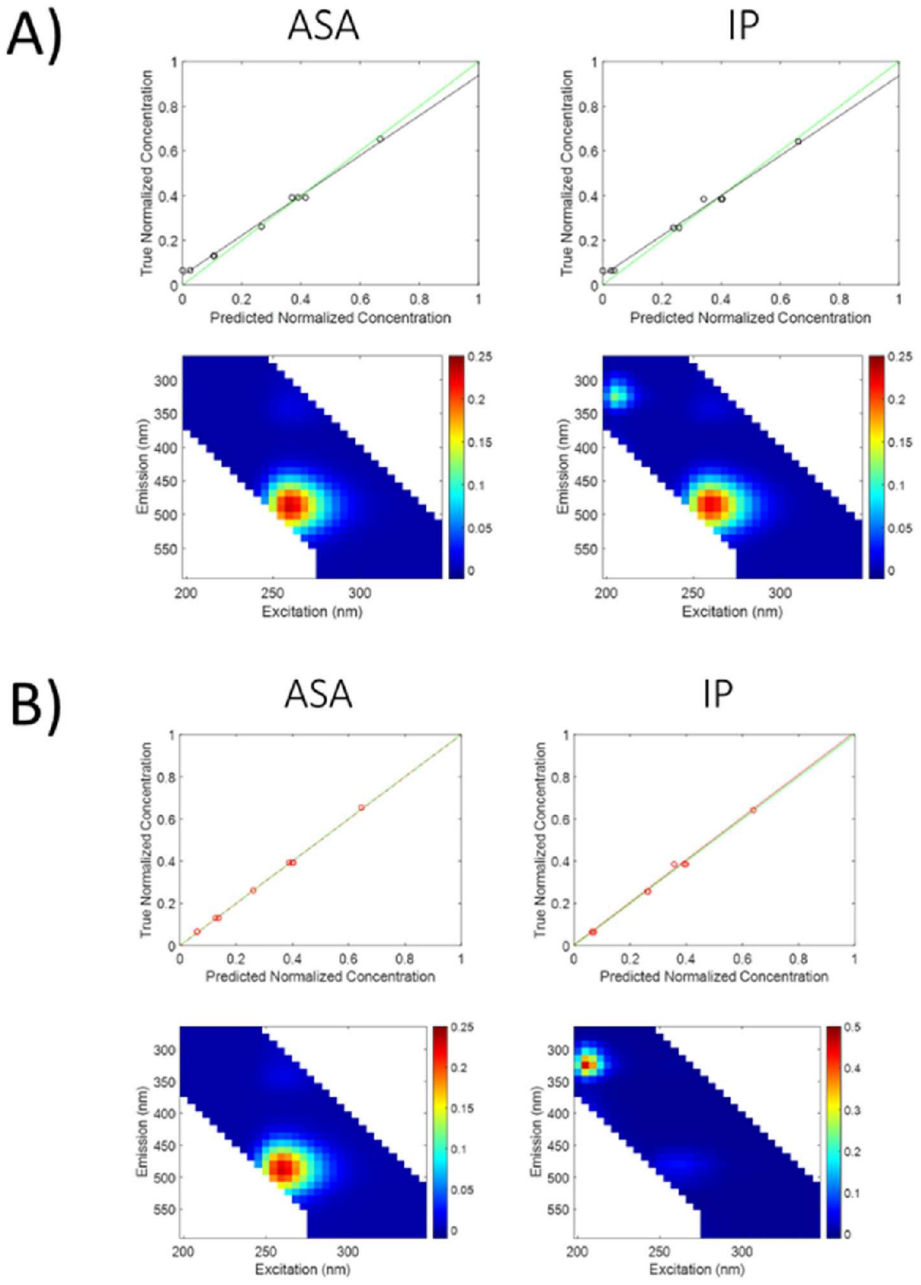


Fig. 10. A) Predicted concentration (top) and pure EEM profiles (bottom) by MCR-ALS applying only non-negativity constraint. Green line indicates the perfect prediction ($R^2 = 1$). B) Predicted concentration (top) and pure EEM profiles (bot) by MCR-ALS applying trilinearity constraint for incomplete datasets. Green line indicates the perfect prediction ($R^2 = 1$). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

simulated data sets and has been also applied to real EEM data sets with different systematic patterns of outlying values. Although EEM data are a natural context of application of this implementation of the trilinear constraint, it could also be applied onto any other kind of trilinear data set with a systematic pattern of missing data.

Author statement

Adrián Gómez-Sánchez: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization. Iker Albuquerque: Validation, Formal analysis, Investigation, Data Curation, Writing - Review & Editing. Pablo Loza-Álvarez: Resources, Writing - Review & Editing, Funding acquisition. Cyril Ruckebusch: Resources, Writing - Review & Editing, Discussion, Supervision, Funding acquisition. Anna de Juan: Conceptualization, Methodology, Formal analysis, Resources, Writing - Original Draft, Discussion, Supervision, 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.

Data availability

Data will be made available on request.

Acknowledgements

A. G.-S., I.A. and A.J. acknowledge financial support from the Spanish government Project PID 2019-1071586B-IOO and the Catalan government (2017 SGR 753). A. G.-S. acknowledges scholarship from the MOBLILEX U Lille program. ICFO authors acknowledge financial support from the Spanish Ministerio de Economía y Competitividad (MINECO) through the "Severo Ochoa" program for Centres of Excellence in R&D (CEX2019-000910-S [MCIN/AEI/10.13039/501100011033]), Fundació Privada Cellex, Fundació Mir-Puig, and Generalitat de Catalunya through CERCA program; LaserlabEurope EU-H2020 GA no. 871124. The authors thank to Raffaele Vitale from the University of Lille for fruitful discussion in the frame of the optimal matrix selection.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2022.104692.

References

[1] F.J. Rodríguez-Vidal, M. García-Valverde, B. Ortega-Azabache, Á. González-Martínez, A. Bellido-Fernández, Characterization of urban and industrial wastewaters using excitation-emission matrix (EEM) fluorescence: searching for specific fingerprints, J. Environ. Manag. 263 (2020), 110396.

[2] M.R. Alcaraz, O. Monago-Maraña, H.C. Goicoechea, A.M. de la Peña, Four-and five-way excitation-emission luminescence-based data acquisition and modeling for analytical applications. A review, Anal. Chim. Acta 1083 (2019) 41–57.

[3] M. Marín-García, R. Tauler, Chemometrics characterization of the Llobregat river dissolved organic matter, Chemometr. Intell. Lab. Syst. 201 (2020), 104018.

[4] C. Stedmon, R. Bro, Characterizing dissolved organic matter fluorescence with parallel factor analysis: a tutorial, Limnol Oceanogr. Methods (6) (2008) 572–579.

[5] A. Gómez-Sánchez, M. Marro, M. Marsal, P. Loza-Alvarez, A. de Juan, 3D and 4D image fusion: coping with differences in spectroscopic modes among hyperspectral images, Anal. Chem. 92 (14) (2020) 9591–9602.

[6] C.F. Kaminski, R.S. Watt, A.D. Elder, J.H. Frank, J. Hult, Supercontinuum radiation for applications in chemical sensing and microscopy, Appl. Phys. B 92 (3) (2008) 367–378.

[7] R.A. Harshman, M.E. Lundy, PARAFAC: parallel factor analysis, Comput. Stat. Data Anal. 18 (1994) 39–72.

[8] R. Bro, PARAFAC. Tutorial and applications, Chemometr. Intell. Lab. Syst. 38 (2) (1996) 149–171.

[9] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods 6 (14) (2014) 4964–4976.

[10] A. de Juan, R. Tauler, Multivariate Curve Resolution: 50 years addressing the mixture analysis problem-A review, Anal. Chim. Acta 1145 (8) (2021) 59–78.

[11] R. Tauler, Multivariate curve resolution applied to second order data, Chemometr. Intell. Lab. Syst. 30 (1) (1995) 133–146.

[12] R. Tauler, I. Marques, E. Casassas, Multivariate curve resolution applied to three-way trilinear data: study of a spectrofluorimetric acid-base titration of salicylic acid at three excitation wavelengths, J. Chemometr. 12 (1998) 55–75.

[13] R. Tauler, A. Smilde, B.R. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, J. Chemometr. 9 (1995) 31–58.

[14] O. Devos, M. Ghaffari, R. Vitale, A. de Juan, M. Sliwa, C. Ruckebusch, Multivariate curve resolution slicing of multiexponential time-resolved spectroscopy fluorescence data, Anal. Chem. 93 (37) (2021) 12504–12513.

[15] S. Elcoroaristizabal, A. de Juan, J.A. García, N. Durana, L. Alonso, Comparison of second-order multivariate methods for screening and determination of PAHs by total fluorescence spectroscopy, Chemometr. Intell. Lab. Syst. 132 (2014) 63–74.

[16] J.B. Kruskal, Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistic, Linear Algebra Appl 18 (2) (1977) 95–138.

[17] G. Tomasi, R. Bro, PARAFAC and missing values, Chemometr. Intell. Lab. Syst. 75 (2) (2005) 163–180.

[18] A. Malik, R. Tauler, Extension and application of multivariate curve resolution-alternating least squares to four-way quadrilinear data-obtained in the investigation of pollution patterns on Yamuna River, Indiada casestudy, Anal. Chim. Acta 794 (2013) 20–28.

[19] The N-way Toolbox for MATLAB, Version 3.31, 16/05/2022, http://www.models.life.ku.dk/nwaytoolbox/download.

[20] W. Windig, Guilment, Interactive self-modeling mixture analysis, J. Anal. Chem. 63 (1991) 1425–1432.

[21] A. De Juan, R. Tauler, Comparison of three-way resolution methods for non-trilinear chemical data sets, J. Chemom. 15 (10) (2001) 749–771.

[22] M. Ghaffari, H. Abdollahi, Duality based interpretation of uniqueness in the trilinear decompositions, Chemometr. Intell. Lab. Syst. 177 (2018) 17–25.

[23] L. Donaldson, Autofluorescence in plants, Molecules 25 (10) (2020) 2393.

[24] M. Hazman, K.M. Brown, Progressive drought alters architectural and anatomical traits of rice roots, Rice 11 (1) (2018) 1–16.

[25] S. Henry, F. Divol, M. Bettembourg, C. Bureau, E. Guiderdoni, C. Périn, A. Diévart, Immunoprofiling of rice root cortex reveals two cortical subdomains, Front. Plant Sci. 6 (2016) 1139.

[26] T. Kreszies, L. Schreiber, K. Ranathunge, Suberized transport barriers in Arabidopsis, barley and rice roots: from the model plant to crop species, J. Plant Physiol. 227 (2018) 75–83.

Supplementary material

The trilinear constraint adapted to solve data with strong patterns of outlying observations or missing values.

Adrián Gómez-Sánchez, Iker Alburquerque, Pablo Loza-Álvarez, Cyril Ruckebusch, Anna de Juan.

Abstract

The supporting information includes the description of the simulated data sets with the related Table S1, Figure S1 and S2. Figure S3 contains the borgen plots of simulated datasets Case 1 and 5. The pseudocode for the implementation of the constraint has been added. Figure S4 contains the results of the analysis of the simulated dataset.

Simulated data set

The simulated data is formed by a 4D EEM fluorescence image of a single sample. The sample has three components. Figure S1 and S2 show the concentration-related maps used in the simulation for each component. These concentration-related maps have been extracted from a previous MCR-ALS analysis. The three EEM landscapes were generated by sum of different Gaussian shapes to mimic fluorescence excitation and emission spectra. Two scenarios (low and high spectral overlap) were simulated. For each scenario, two cases were studied: low or high amount of noise. In order to stress even more the system, a final scenario where the spectral overlap is high, the amount of noise is high, and the signal follows a Poisson distribution was studied as well (Table S1).

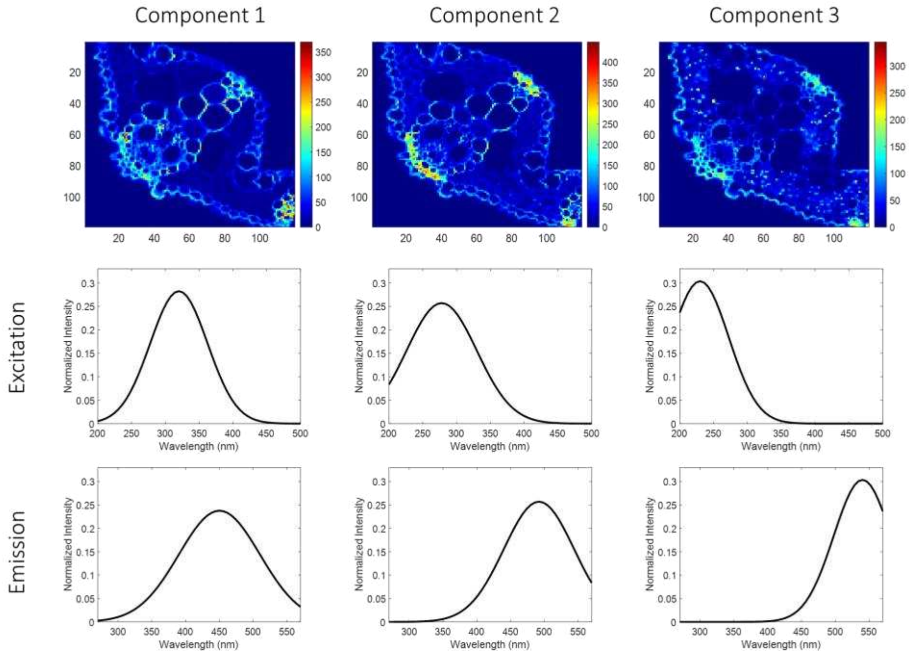


Figure S1. Pure components used to simulate the dataset with low spectral overlap. Top plots, pure concentration-related maps of the components. Bottom plots, pure incomplete EEM of the components simulating the systematic missing value-pattern in EEM measurements.

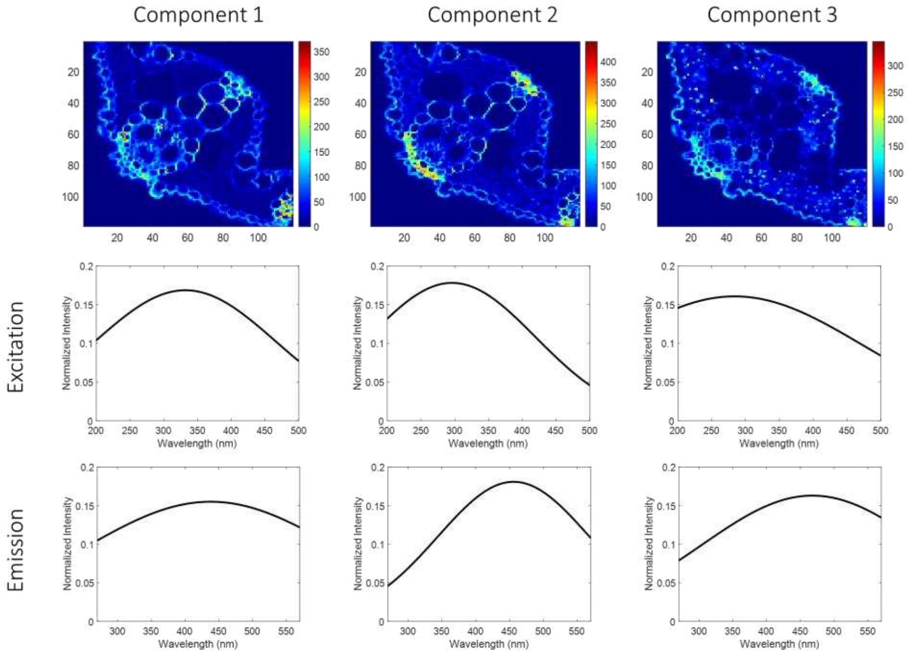


Figure S2. Pure components used to simulate the dataset with high spectral overlap. Top plots, pure concentration-related maps of the components. Bottom plots, pure incomplete EEM of the components simulating the systematic missing value-pattern in EEM measurements.

Using these profiles, an incomplete EEM HSI ($\underline{\mathbf{D}}_{\text{in}}$) where missing values are present was simulated. To generate the data matrix $\mathbf{D}_{\mathbf{c}}$ and $\mathbf{D}_{\mathbf{in}}$, the related 2D EEM landscape, sized (λ_{ex}, λ_{em}), is vectorized concatenating in a single row vector the emission spectra related to the different excitation wavelengths. The vectorized EEM spectra of the three components were put together to form the matrix, $\mathbf{S}^{\mathbf{T}}$. Note that $\mathbf{S}^{\text{T}}_{\text{c}}$ contains the complete landscape, while $\mathbf{S}^{\text{T}}_{\text{in}}$ contains missing values due to the fact that the set is incomplete. The distribution maps of the three components were unfolded into linear concentration-related profiles to form the matrix $\mathbf{C}_.$ Both data matrix $\mathbf{D}_{\text{c}}$ and $\mathbf{D}_{\text{in}}$ were obtained through the product $\mathbf{D} = \mathbf{CS}^{\mathrm{T}}$. White noise or poisson were added to $\mathbf{D}$ to simulate the natural noise in the fluorescence hyperspectral images in soft or hard conditions.

The amount of noise added to the multiset has been calculated according to Eq S1.

$$E(\%) = \sqrt{\frac{\sum_i \sum_j (d_{ij,n} - d_{s_{ij,nf}})^2}{\sum_i \sum_j \left(d_{s_{ij,n}}\right)^2}} \cdot 100 \qquad \text{(S1)}$$

Where $d_{s_{ij,n}}$ and $d_{ij,nf}$ denote the ij$^{\text{th}}$ element of the raw $\mathbf{D}$ matrix with noise added and the ij$^{\text{th}}$ element of the noise-free matrix. The $\mathbf{E}$ (%) added is shown in Table S1. for $\mathbf{D}_{\text{in}}$. When this matrix is analysed by MCR-ALS, a lack of fit very similar to $\mathbf{E}$ (%) should be obtained. Missing values of $\mathbf{D}_{\text{in}}$ were removed according to the Figure 6 of the main manuscript.

Table S1. Scenarios simulated.

Case	Noise structure	Noise added (%)	Spectral overlap
1	White	16.1	Low
2	White	16.5	High
3	White	31.0	Low
4	White	31.8	High
5	Poisson	28.5	High

Rotational ambiguity of the simulated data set

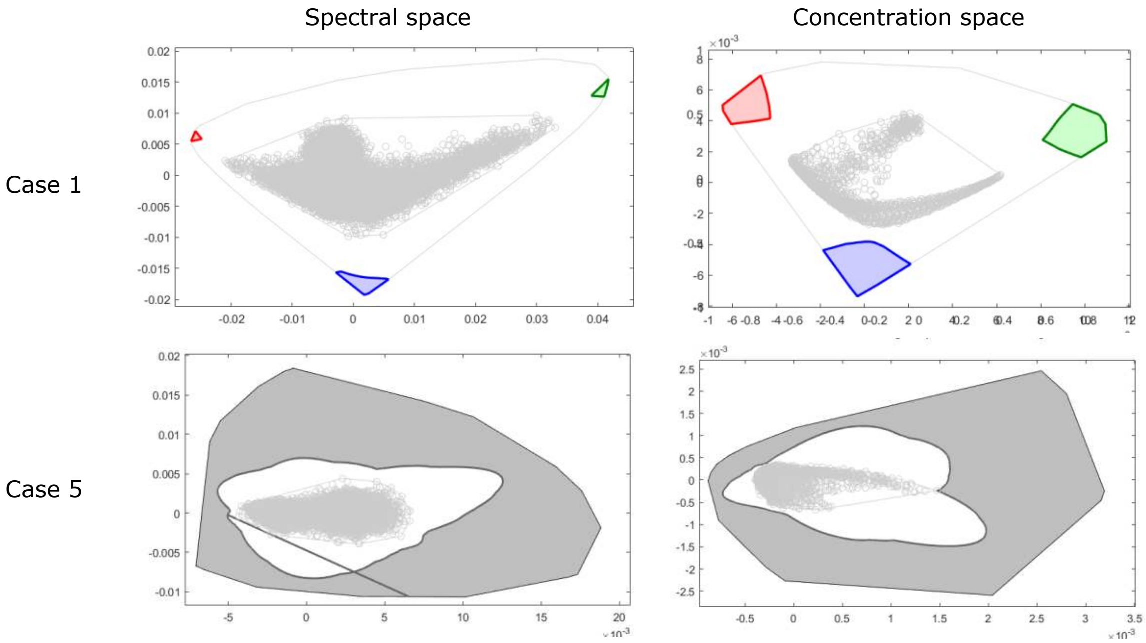
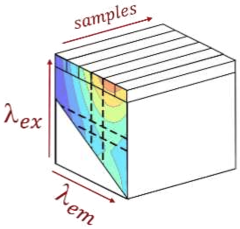


Figure S3. Borgen plots (updating low negative values in the resolved to zero) by FACPACK for simulated examples 1 (system with the lowest noise level and lowest spectral overlap) and 5 (system with the highest noise level with Poisson structure and most severe spectral overlap). It can be observed that in both cases there is rotational ambiguity, but when increasing overlap and noise level, the system becomes extremely ambiguous. The figures have been generated with *FACPACK: a software for the computation of multi-component factorizations and the area of feasible solutions.* https://swmath.org/software/27898 (Last access 8/9/2022).

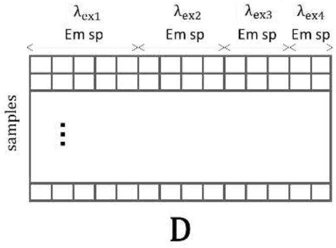
Pseudocode of the trilinear constraint adapted to solve data with strongly patterned missing values.

Syntaxis is inspired in MATLAB notation.

```
%Initialization
% Cube is the data sized (s,ex,em) (Fig 4A on manuscript)
```



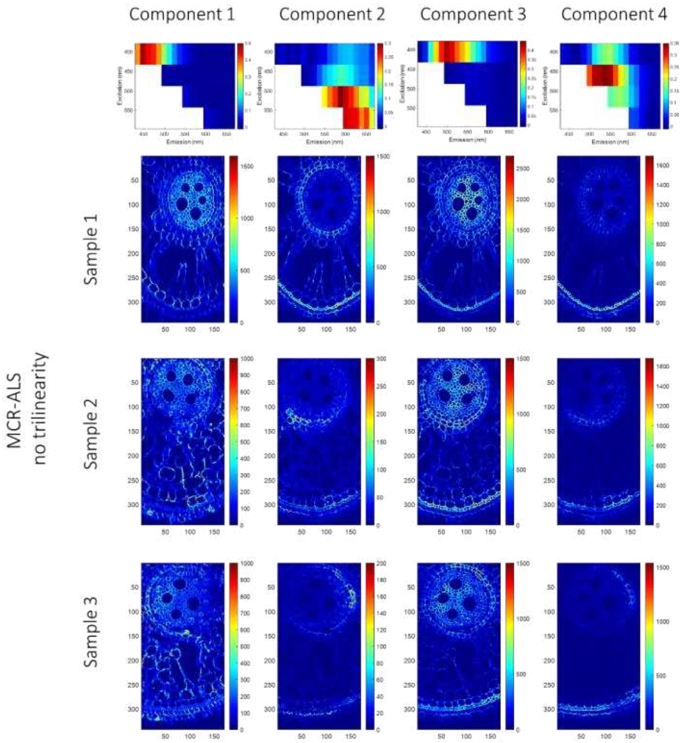
```
% D is the matrix formed by concatenation of all excitation channels of Cube without
missing values. (Fig 4A on manuscript)
```



```
% n is number of components.
% nc is the number of columns of D.
% crit is the convergence criterion.
% conv is the difference of LOF between two consecutive iterations.
% C is the pure concentration profile matrix.
% S is the pure spectral profile matrix.
% find_submatrices is the script which provides all possible submatrices in a
landscape.
% ml_index is the script which provides the ML index of each submatrix and sort them
from high to low value until the landscape is covered.
% ml_evaluation contains in each row the indices that define each submatrix sorted by
ml_index.
% ns_ml is the number of rows of ml_evaluation.
% row_index is the script which sorts the submatrices from higher to lower number of
rows until the landscape is covered.
% row_evaluation contains in each row the indices that define each submatrix sorted by
row_index.
% ns_row is the number of rows of ns_evaluation.
% reshaping allows to pass from unfolded mode to folded mode or vice versa.


%Spectral initial estimates by a based SIMPLISMA algorithm.
sp=pure(D); %sp is sized [n,nc].


%Reconstruction of the landscape ExEm of sp.
ExEm=reshaping(sp,’fold’); %ExEm is sized [n,ex,em].
```

```
%Calculation of the optimal submatrices

for i=1:n
       submatrix_list=find_submatrices(ExEm(i,:,:)); %Finds the submatrices on the
component i.
        ml_evaluation=ml_index(submatrix_list); %Calculates the ML index and sorts the
submatrix_list according to decreasing order of ML.
end


%Initialize MCR-ALS

while conv > crit
S=C\D;

%%Trilinearity constraint%%

%Step 1 (removal of signal mixing in Sfi)
       for i=1:n
              S_folded=reshaping(S(i,:),’fold’);
                     for j=1:ns_ml
                            submatrix= S_folded (ml_evaluation(j,:)));
                            [u,s,v]=svd(submatrix);
                            u=u(:,1);
                            s=s(1,1);
                            v=v(:,1);
                            submatrix=u*s*v';
                            S_folded(ml_evaluation(j,:))=submatrix;
                     end
              S(i,:)= reshaping(S_folded,’unfold’); %Unfold the S_folded matrix in
       vectorized form.
       end

row_evaluation=row_index(submatrix_list); %Sorts the submatrix_list according to the
number of rows.

%Step 2 (ensuring trilinear profiles)
       for i=1:n
              S_folded=reshaping(S(i,:),’fold’);
                     for j=1:ns_ml
                            submatrix= S_folded (row_evaluation(j,:)));
                            [u,s,v]=svd(submatrix);
                            u=u(:,1);
                            s=s(1,1);
                            v=v(:,1);
                            submatrix=u*s*v';
                            S_folded(row_evaluation(j,:))=submatrix;
                     end
              S(i,:)= reshaping(S_folded,’unfold’); %Unfold the S_folded matrix in
       vectorized form.
       end

S=norm(S); %Normalize each profile by the 2-norm.
C=D/S;

end
```

Complete results of HSI multiset analysis



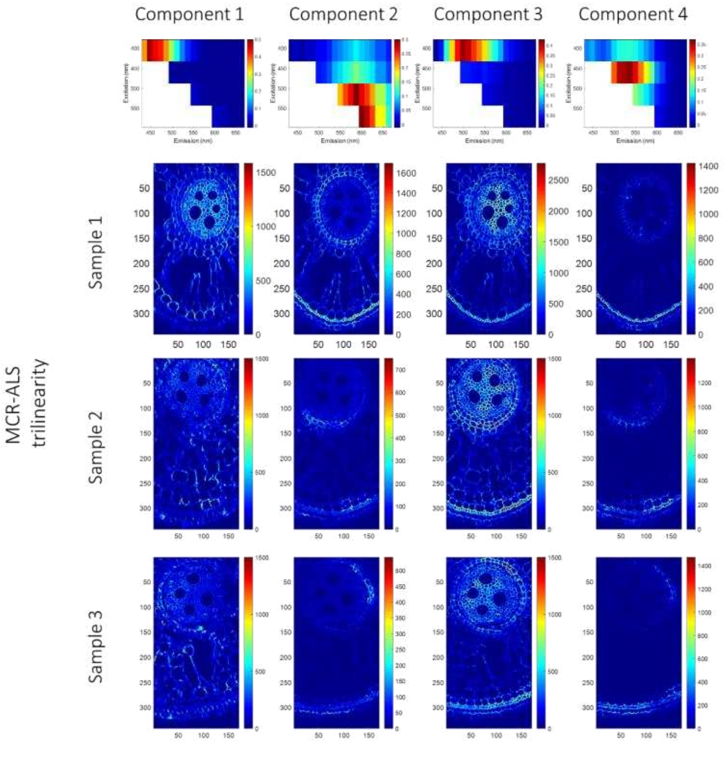


Figure S4. Pure concentration-related maps and pure EEM found by MCR-ALS without and with trilinear constraint applied.

The MCR-ALS trilinearity constraint for data with missing values

Adrián Gómez-Sánchez[1,2], Raffaele Vitale[2], Pablo Loza-Alvarez[3], Cyril Ruckebusch[2], Anna de Juan[1]

[1]Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028, Barcelona, Spain

[2] Univ. Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, F-59000 Lille, France

[3]ICFO-Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology, 08860 Castelldefels, Barcelona, Spain

*Corresponding authors:

Adrián Gómez-Sánchez (agomezsa29@alumnes.ub.edu)

Anna de Juan (anna.dejuan@ub.edu)

Abstract

Trilinearity is a property of some chemical data that lead to unique decompositions when curve resolution or multi-way factorization methods are used. Curve resolution algorithms, such as Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS), can provide trilinear models by implementing the trilinearity condition as a constraint. However, some trilinear analytical measurements, such as Excitation-Emission Matrices (EEM) measurements, usually exhibit systematic patterns of missing data due to the nature of the technique, which imply a challenge of the classical implementation of the trilinearity constraint. In this instance, extrapolation or imputation methodologies may not provide optimal results.

Recently, a novel algorithmic strategy to constrain trilinearity in MCR-ALS in the presence of missing data was developed. This strategy relies on the sequential imposition of a classical trilinearity restriction on different submatrices of the original investigated data set, but, although effective, was found to be particularly slow and requires a proper submatrix selection criterion.

In this paper, a much simpler implementation of the trilinearity constraint in MCR-ALS capable of handling systematic patterns of missing data and based on the principles of the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm is proposed. This novel approach preserves the trilinearity of the retrieved component profiles without requiring data imputation or subset selection steps and, as for all other constraints designed for MCR-ALS, offers the flexibility to be applied component-wise or data block-wise providing hybrid bilinear/trilinear factorization models. Furthermore, it can be easily extended for coping with any trilinear or higher-order datasets with whatever pattern of missing values.

Keywords: trilinearity, missing data, Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS), Nonlinear Iterative Partial Least Squares (NIPALS), constraints.

Introduction

Trilinear models are mathematical representations of the decomposition of three-way data arrays into the product of three pure matrices, each connected to one of the modes or dimensions of such arrays and containing the underlying components or factors of the three-way data (Figure 1A).

Data decomposition approaches providing trilinear models are particularly relevant in scientific fields such as chemistry, spectroscopy and environmental science due to the fact the solutions they return are unique, i.e., the extracted component profiles do not exhibit rotational ambiguity under mild conditions [1]. Furthermore, when multiblock or multiset data are handled, trilinear decompositions also yield the so-called second-order advantage and enable the quantification of analytes in the presence of unknown interferents [2]. Parallel Factor Analysis-Alternating Least Squares (PARAFAC-ALS) [3,4], Direct Trilinear Decomposition (DTD) [5] and Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) [6,7] with trilinearity constraint [2,8] are three highly effective algorithms widely employed to obtain trilinear models. Whereas PARAFAC-ALS and DTD hold to the decomposition in Figure 1A, when MCR-ALS is used, the initial data cube is unfolded in one direction and the trilinearity constraint is applied by-component to the blocks of the unfolded mode to ensure a common profile shape among them (Figure 1B) [2,8].

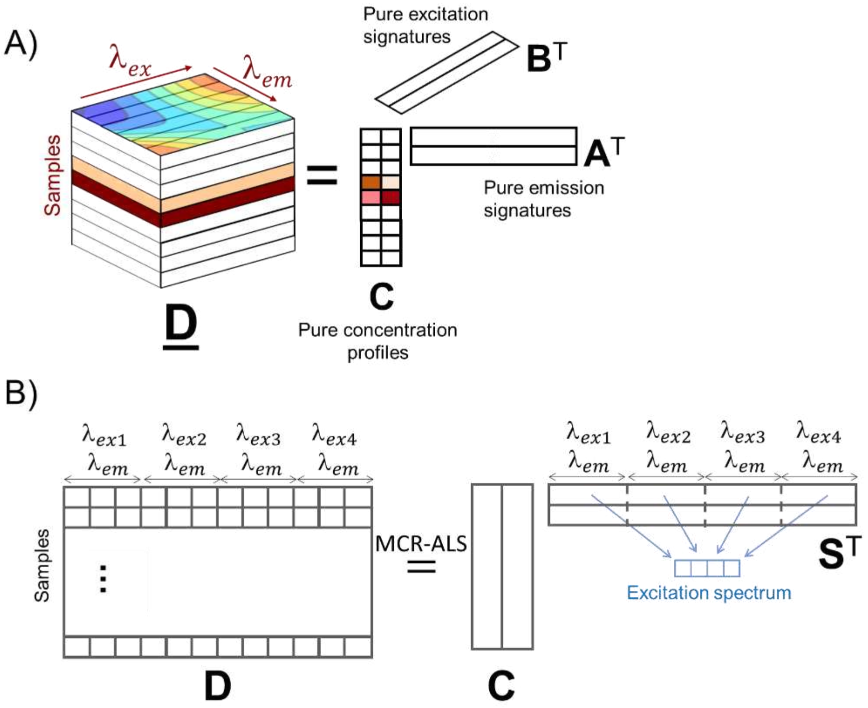

A)
λ_{ex}
λ_{em}
Samples
D
=
Pure excitation signatures
B^T
A^T
Pure emission signatures
C
Pure concentration profiles
B)
λ_{ex1} λ_{em}
λ_{ex2} λ_{em}
λ_{ex3} λ_{em}
λ_{ex4} λ_{em}
Samples
D
MCR-ALS =
C
S^T
Excitation spectrum

Figure 1. A) Schematic representation of the trilinear decomposition of a EEM data cube (**D**) enabling the retrieval of the pure concentration profiles (**C**), the pure excitation spectral profiles (**B**) and, the pure emission spectral profiles (**A**) of the underlying components. B) Schematic representation of the trilinearity-constrained MCR-ALS decomposition of **D**. Here, **D** is unfolded by concatenating in a row-wise augmented multiset all the emission spectra collected at the different excitation wavelengths. In this case, MCR-ALS provides the pure concentration profiles (**C**) and the augmented pure spectral fingerprint (**S**) of every component, containing the excitation and emission pure profiles. Notice that every pure emission spectral profile extracted at the different excitation wavelengths are forced to have the same shape.

Although the practical applications of trilinear data analysis can be substantially diverse [9-11], the most emblematic example to illustrate the relevance of trilinear models in science and, more specifically in analytical chemistry, relates to excitation-emission matrix (EEM) fluorescence measurements. EEM measurements provide a 2D excitation-emission landscape per sample analyzed and represent an excellent tool to characterize fluorophores due to variations in their excitation and emission spectra [12-14]. If several samples are considered, a 3D data structure can be built, (see Figure 1A), where three dimensions correspond to the number of samples (s), the number of excitation wavelength channels (λ_{ex}) and the number of emission wavelength channels (λ_{em}), the resulting in a data cube sized ($s, \lambda_{ex}, \lambda_{em}$). A trilinear decomposition of this 3D data structure can then be carried out to obtain the pure excitation and emission spectra and the sample profile of components.

Similar decompositions can be achieved when dealing with excitation-emission hyperspectral images (EEM-HSI). In EEM-HSI, every image pixel is associated to a 2D EEM landscape. EEM-HSI can be therefore looked at as 4D data arrays with dimensions equal to the number of image pixels along the x-direction times the number of image pixels along the y-direction times the number of excitation wavelength channels times the number of emission wavelength channels ($x, y, \lambda_{ex}, \lambda_{em}$). In this scenario, trilinear decompositions are achieved after unfolding pixel-wise these 4D data arrays.

Although EEM constitute an ideal example to illustrate how trilinear factorization methodologies operate and work, their actual analysis may sometimes be extremely challenging due to the fact that the collected measurements might be perturbed by signals such as Rayleigh and Raman scattering. In such cases, the signal of these scattering contributions does not follow a trilinear model and, consequently, has to be corrected or removed from the initial data set. In addition, when the emission range and the excitation range in the EEM overlap, no emission signal is detected below the excitation range. These facts cause a systematic pattern of missing data in EEM measurements linked to the natural fluorescence phenomenon and the instrumental settings used that need somehow to be dealt with.

Dealing with missing data poses a substantial challenge when employing trilinear modelling approaches since conventional algorithms are not designed to directly handle them. Different strategies have been proposed to overcome this limitation, such as missing data interpolation or extrapolation based on neighbouring values or missing data imputation [15,16]. However, it is well-established that imputation algorithms may converge very slowly in the presence of large amounts of missing data following systematic patterns of absence in the data structure [17]. In image fusion analysis, the percentage of missing data can easily exceed 50% [18], making imputation not a viable option in such cases.

In trilinear MCR-ALS models, dealing with missing data implies modifying the way the trilinear constraint is implemented. Indeed, the forced common shape in the blocks of the extended mode in Figure 1B is based on performing a Singular Value Decomposition (SVD) analysis of a matrix formed by all profiles linked to a single component and taking the profile of the first principal component calculated as the common reference [8]. If the profiles do not have the same number of entries (because of missing emission observations values are missing), the classical implementation cannot be applied.

As an alternative to data extrapolation and imputation, Gómez-Sánchez et al. [19] have lately proposed an innovative algorithmic procedure to constrain trilinearity when modelling three-way data with missing values by MCR-ALS. This approach allows to skip missing entries by imposing the trilinearity restriction only on local subsets of the original data at hand. Unfortunately, the selection of the submatrices is data set-dependent and the algorithm gets complex and difficult to implement.

In this work, we present a much simpler and computationally efficient implementation of the MCR-ALS trilinearity constraint capable of handling missing data and based on an adapted use of the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm [20]. As detailed in the next sections, the valuable characteristic of NIPALS is that it can be adapted to handle datasets with missing values by skipping the missing entries during the rank-one approximation calculation [21]. This is possible because calculation of the scores and loadings of is performed the row-by-row and column-by-column calculation of the scores and loadings, respectively.

The adaptability of NIPALS to work only with the available data information values generalizes the use of this trilinearity implementation to analyze data with a large diversity of percentage and pattern of missing data without the need to perform any step of data imputation. As for the classical implementation of the trilinear constraint in MCR-ALS environment, the new implementation can be optionally applied per component or per block [22], ensuring the possibility to work with hybrid bilinear-trilinear models. It is important to note that the approach would also apply when the multilinear constraint is applied to higher-order data sets.

To proof the potential of this approach, the new trilinearity constraint based on NIPALS has been tested in simulated data, in EEM from controlled pharmaceutical samples and in EEM-HSI from cross-sections of rice roots as examples.

Datasets

This section includes the details of both simulated and EEM measurements. Simulations were conducted to replicate the spatial structures and EEM fingerprints naturally found in a plant tissue, while introducing variations related to varying noise levels and diverse spectral overlap conditions. Real EEM-HSIs and EEM measurements of pharmaceutical mixture solutions are also analyzed to show the performance of the trilinear constraint under experimental controlled conditions and for exploratory analysis.

Excitation-emission hyperspectral images of plant tissue

Simulated excitation-emission hyperspectral images

The simulated dataset is based on an EEM-hyperspectral image, with distribution maps inspired by the components of a real EEM leaf sample. These maps exhibit a significant overlap among components. In total, the EEM-HSI simulated sample surface encompasses 119×119 pixels. The emission range goes from 200 nm to 500 nm, with a step size of 6 nm (51 channels). The excitation range goes from 200 nm to 500 nm, with a step size of 6 nm (51 channels), resulting in a hypercube sized 119×119×51×51. Since in EEM measurements there is no emission signal below the excitation wavelength, we set as Not a Number (*NaN)* all emission values which are below the excitation wavelength to mimic the missing value pattern naturally found in EEM. Thus, the dataset presents approximately 50% of missing data. The pure distribution maps and pure fluorescence EEM landscapes are presented in Figures S1-3 of Supporting Information.

The pattern of missing data used for the simulations can be seen in Figure 2A. In order to test the algorithm, two scenarios of low and high overlap of pure component EEM profiles, respectively, were explored. In both cases, different levels of Poisson noise (0.5%, 5%, 15% and 30% of the total data variance) were accounted for. These noise levels mimic typical conditions encountered when conducting EEM measurements under excellent, good, standard and severe experimental conditions. Additional information on the generation of these simulated data is provided in the Supporting Information.

Excitation-emission hyperspectral image of a plant tissue sample

A sample of plant tissue was imaged under a fluorescence confocal microscope (Leica TCS SP8 STED 3X, Leica Microsystems, Mannheim, Germany) at five different excitation wavelengths (405, 470, 520, 570, and 620 nm). Emission spectra were recorded within 5 specific ranges (435–663 nm, 495–663 nm, 543–663 nm, 591–663 nm and 647–663) with a sampling interval and a bandwidth of 12 nm to avoid Rayleigh scattering due to the sensor sensibility. Pixel size was set at 450 × 450 nm², which resulted in a final 4D data structure of dimensions 1024×512×5×20 covering a global field of view of 460 × 230 µm² and featuring approximately 47% of missing values in each EEM landscape (see Figure 2B) For additional details on the data collection procedure, please refer to Ref. [19].

Excitation-emission matrices of pharmaceutical mixtures

The EEM of nine mixtures of ibuprofen (IBU) and acetylsalicylic acid (ASA) were measured using an AB2 Aminco-Bowman spectrofluorometer within the excitation wavelength range 200–500 nm and emission wavelength range 200–600 nm. Table 1 shows the concentrations of the two pharmaceutical compounds in each investigated mixture. The final 3D dataset formed by the pharmaceutical mixtures was a data cube formed by 9 samples, 61 excitation channels and 42 emission channels, sized 9×61×42. For additional details, please refer to Ref. [19]. Spectral regions clearly exhibiting Rayleigh and Raman scattering were removed from the initial data which resulted in approximately 46% of missing values in every EEM landscape recorded (see Figure 2C).

Table 1. Concentration of ibuprofen (IBU) and acetylsalicylic acid (ASA) in the nine pharmaceutical mixtures under study

	Mixture								
Pharmaceutical compound	1	2	3	4	5	6	7	8	9
IBU (mg/L)	0.25	1.00	0.25	2.50	0.25	1.00	1.50	1.50	1.50
ASA (mg/L)	1.50	0.50	1.00	0.25	2.50	2.50	0.5	0.25	1.50

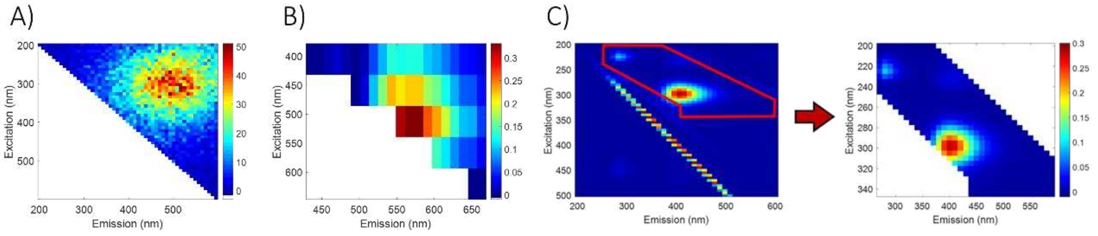


Figure 2. Missing value patterns in A) the simulated EEM data, B) the real EEM-HSI data and C) the real EEM data collected on the pharmaceutical mixtures of ibuprofen and acetylsalicylic acid.

Software

Data analysis was performed by means of in-house-coded Matlab scripts and routines.

Data analysis

Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)

MCR-ALS is an algorithm meant to solve the mixture analysis problem and it has been widely applied in many different fields [6,7]. MCR-ALS decomposes the data into the pure signatures weighted by their contributions or concentrations, following a bilinear model (Eq. 1). This model matches the nature of the spectroscopic measurements, where the data can be generally expressed as a bilinear model following the Beer-Lambert law.

$$\mathbf{D} = \mathbf{CS}^{\mathrm{T}} + \mathbf{E} \qquad \text{Eq. 1}$$

where $\mathbf{D}$ is the matrix sized (I, J) (usually, samples and wavelengths, respectively) which contains all the spectra and $\mathbf{C}$ and $\mathbf{S}^{\mathrm{T}}$ are the matrices of concentration profiles, sized (I, N) (samples and components) and spectral signatures of the image constituents, sized (N, J) (components and wavelengths), respectively. $\mathbf{E}$, sized (I, J) , is the matrix of residual variation unexplained by the MCR model. In MCR-ALS, the matrices $\mathbf{C}$ and $\mathbf{S}^{\mathrm{T}}$ are estimated through an iterative optimization process based on alternating least squares and during which constraints, such as non-negativity or trilinearity, can be optionally imposed per mode ($\mathbf{C}$ or $\mathbf{S}^{\mathrm{T}}$), per block in a multiset arrangement and per profile (component) within $\mathbf{C}$ or $\mathbf{S}^{\mathrm{T}}$. Calculations are stopped when the relative difference in the values of the model lack of fit expressed as:

$$LOF(\%) = 100 \times \sqrt{\frac{\sum_{i,j} e_{i,j}^2}{\sum_{i,j} d_{i,j}^2}} \qquad \text{Eq. 2}$$

becomes lower than a user-defined threshold. In Eq. 2, $d_{i,j}$ represents the i,j^{th} element of **D** and $e_{i,j}$ is the residual associated with the reproduction of $d_{i,j}$ through the MCR-ALS model.

Standard implementation of the trilinearity constraint in MCR-ALS

The standard algorithmic scheme by which trilinearity constraint applied during the MCR-ALS optimization procedure is represented in Figure 3. The cube $\underline{\mathbf{D}}$, formed by the EEM measurement of several samples, need to be first unfolded into a data matrix with size $s \times \lambda_{ex}\lambda_{em}$ (see Figure 1B).

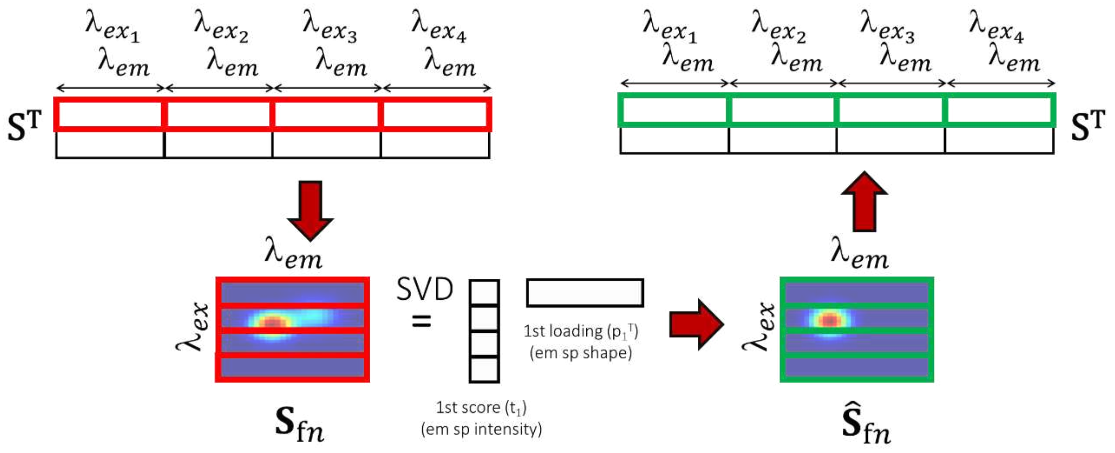

λ_{ex_1} λ_{em}
λ_{ex_2} λ_{em}
λ_{ex_3} λ_{em}
λ_{ex_4} λ_{em}
$\mathbf{S}^{\mathrm{T}}$
λ_{ex_1} λ_{em}
λ_{ex_2} λ_{em}
λ_{ex_3} λ_{em}
λ_{ex_4} λ_{em}
$\mathbf{S}^{\mathrm{T}}$
λ_{em}
λ_{ex}
SVD
=
1st loading (p_1^T)
(em sp shape)
1st score (t_1)
(em sp intensity)
$\mathbf{S}_{fn}$
λ_{em}
λ_{ex}
$\hat{\mathbf{S}}_{fn}$

Figure 3. Schematic representation of the classical SVD-based implementation of the MCR-ALS trilinearity constraint. At each MCR-ALS iteration, trilinearity is imposed on each row of $\mathbf{S}^{\mathrm{T}}$ as illustrated. Notice that the decomposition and reconstruction of $\mathbf{S}_{\mathrm{f}n}$ forces all its row profiles to have identical shape, while weighted by the corresponding score.

Afterwards, at each individual MCR-ALS iteration, when $\mathbf{S}^{\mathrm{T}}$ $(N \times \lambda_{ex}\lambda_{em})$ is estimated, each one of its rows is refolded into a two-dimensional data array, $\mathbf{S}_{\mathrm{f}n}$, with dimensions $\lambda_{ex} \times \lambda_{em}$ which is subjected to an SVD. The first principal component, expressed by the scores and the loadings, serves to build a new matrix $\hat{\mathbf{S}}_{\mathrm{f}n}$, where all emission profiles have the same shape thanks to the rank-one SVD reconstruction. Note that, here, the score vector is defined as the left singular vector multiplied by its singular value for the sake of simplification. $\hat{\mathbf{S}}_{\mathrm{f}n}$ is finally unfolded again and used to replace the corresponding row of $\mathbf{S}^{\mathrm{T}}$ before the following MCR-ALS iterative process. Once convergence is achieved, the pure component excitation spectra are retrieved by computing the area of the respective pure emission profiles at each excitation wavelength.

This implementation of the trilinearity constraint in MCR-ALS, being it based on the principles of SVD, cannot readily handle datasets containing missing values.

A NIPALS-based implementation of the trilinearity constraint in MCR-ALS

As stated before, it is very common to find situations where no emission signal is recorded below certain excitations or where specific scattering or Raman bands need to be removed from the data yielding to EEM landscapes to contain patterned missing data (see Figure 4A). In these cases, the MCR-ALS trilinearity constraint can be adapted to address the missing values following the scheme illustrated in Figure 4B. This algorithmic scheme basically encompasses the same computational steps as the one represented in Figure 3, but when it comes to decomposing the $\mathbf{S}_{\mathrm{f}n}$ matrices resulting from the refolding of the individual rows of $\mathbf{S}^{\mathrm{T}}$, the factorization is conducted by means of the NIPALS algorithm and not through SVD.

NIPALS is an iterative algorithm used in multivariate analysis to extract principal components, as SVD does. However, a significant advantage over SVD regards the fact that NIPALS can converge in the presence of missing data to the same solution as SVD for rank-one matrix approximations [21]. NIPALS calculates sequentially the scores and loadings of every component so that they capture the maximum variance in the data. After the calculation of every component, the initial data are deflated and the remaining information is used to estimate the following component until all data variance is explained [20]. When $\mathbf{S}_{\mathrm{f}n}$ contains missing values, the rank-one approximation is done by performing the least squares estimation of the score and loading vector row by row and column by column, respectively, as displayed in Figure 5.

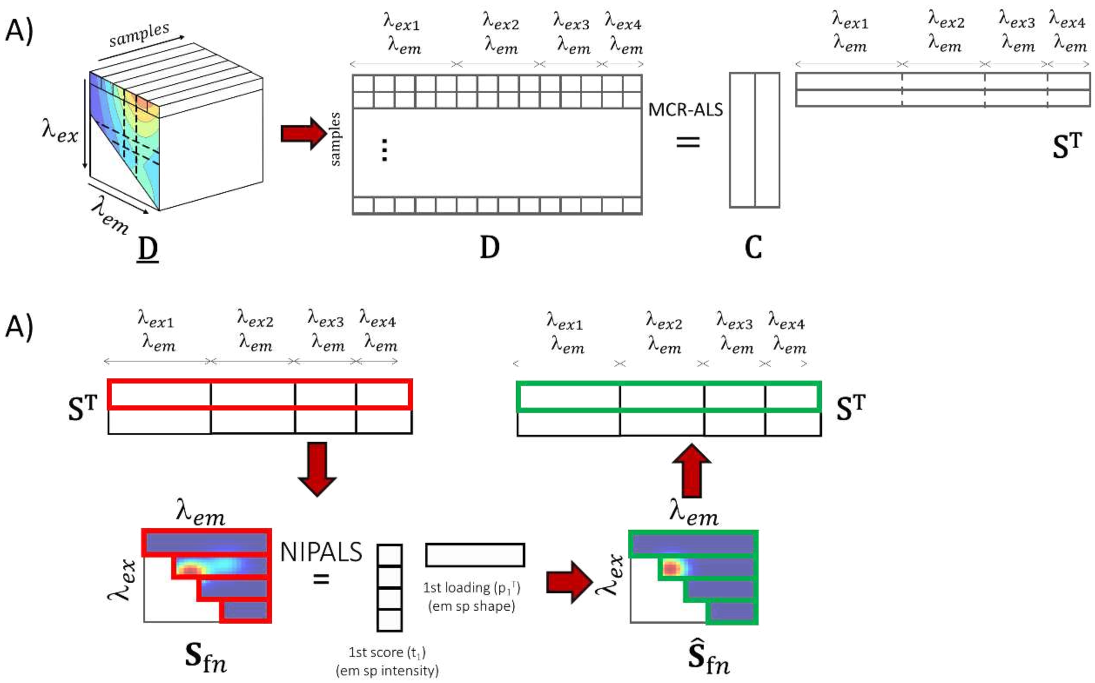


Figure 4. Schematic representation of the NIPALS-based trilinear MCR-ALS factorization of an EEM data array $\underline{\mathbf{D}}$. A) $\underline{\mathbf{D}}$ is unfolded into the matrix **D** by concatenating the different excitations. Then, **D** is decomposed as the product of a matrix $\mathbf{C}$, containing the pure component concentration profiles, and $\mathbf{S}^{\mathrm{T}}$, containing augmented pure component spectral signatures. B) During iterations, the trilinearity constraint is applied on each row of $\mathbf{S}^{\mathrm{T}}$, forcing $\mathbf{S}_{fn}$ to have the same emission shape across the excitation using NIPALS.

As shown in Figure 5A, NIPALS is initialized with an estimate of the first-component loading vector **p,** obtained, for instance, as the column-wise average of the available entries of $\mathbf{S}_{\mathrm{f}n}$. Then, the first-component score vector **t** is calculated using **p** and $\mathbf{S}_{\mathrm{f}n}$. More specifically, every element of **t** is calculated independently, using only the respective row of $\mathbf{S}_{\mathrm{f}n}$ and the loading vector **p**, as in Eq. 3.

$$\mathbf{t}(i,1) = \mathbf{S}_{\mathrm{f}n}(i,:)(\mathbf{p}^{\mathrm{T}})^{+} \qquad \text{Eq. 3}$$

If missing values appear along $\mathbf{S}_{\mathrm{f}n}(i,:)$, only its available entries and the corresponding portion of the loading vector **p** are considered. Once all the elements of **t** have been calculated, **p** is reestimated by using the score vector **t** and $\mathbf{S}_{\mathrm{f}n}$, as shown in Figure 5B. In this case, every column of **p** is calculated independently, using **t** and the related column of $\mathbf{S}_{\mathrm{f}n}$, as:

$$\mathbf{p}^{\mathrm{T}}(1,j) = \mathbf{t}^{+}\mathbf{S}_{\mathrm{f}n}(:,j) \qquad \text{Eq. 4}$$

If the column of $\mathbf{S}_{\mathrm{f}n}$ contains missing values, only its available entries and the corresponding elements of **t** are taken into account. This procedure is repeated for all columns of $\mathbf{S}_{\mathrm{f}n}$. Both calculations of **t** and **p** are repeated until convergence. When convergence is achieved, the algorithm stops providing two refined vectors **t** and **p** which are finally used to obtain $\hat{\mathbf{S}}_{\mathrm{f}n}$.

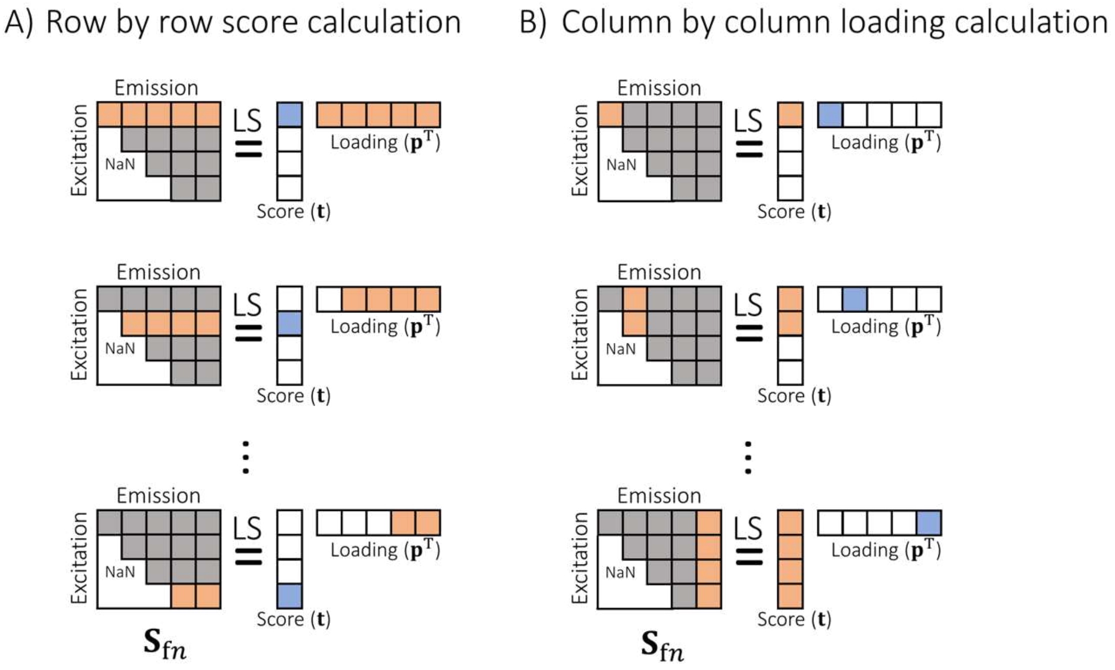

A) Row by row score calculation
B) Column by column loading calculation
Emission
Excitation
NaN
LS
Loading ($\mathbf{p}^{\mathrm{T}}$)
Score ($\mathbf{t}$)
$\mathbf{S}_{fn}$

Figure 5. A) Schematic representation of the row-by-row calculations underlying the NIPALS algorithm. The loading vector $\mathbf{p}^{\mathrm{T}}$ and a single row of $\mathbf{S}_{\mathrm{f}n}$ (orange) are used to calculate the corresponding score value $\mathbf{t}(j,1)$ (blue). In case this single row of $\mathbf{S}_{\mathrm{f}n}$ contains missing values, the size of $\mathbf{p}^{\mathrm{T}}$ is adapted accordingly. B) Schematic representation of the column-by-column calculations underlying the NIPALS algorithm. The score vector $\mathbf{t}$ and a column of $\mathbf{S}_{\mathrm{f}n}$ (orange) are used to calculate the corresponding loading value $\mathbf{p}^{\mathrm{T}}(1,k)$ (blue). Once again, in case this single column of $\mathbf{S}_{\mathrm{f}n}$ contains missing values, the size of $\mathbf{t}$ is adapted accordingly.

This NIPALS-based implementation of the trilinearity constraint in MCR-ALS allows skipping missing values present in the investigated data, thereby bypassing the need for imputation methods and preserving the integrity of MCR-ALS decompositions. Due to its simplicity, it can constitute a valuable addition to all publicly available MCR-ALS interfaces [23], and can be easily adapted to be applied to higher-order multi-way data arrays.

Results and discussion

Simulated excitation-emission hyperspectral images

The new implementation of the MCR-ALS trilinearity constraint was first tested on the simulated 4D images described before and containing around 50% of missing values. All the generated datasets were analyzed using two different approaches: each 4D image was first unfolded as in Figure 3A and then subjected to two different MCR-ALS factorization procedures, one during which only non-negativity constraints were imposed on both $\mathbf{C}$ and $\mathbf{S}^{\mathrm{T}}$ (bilinear model) and the other during which also the adapted trilinearity constraint was applied (trilinear model). In all cases, initial spectral estimates were obtained through a SIMPLISMA-based algorithm [24]. For all MCR-ALS models, the maximum number of iterations was set at 2000 while convergence was considered achieved if the difference between the LOF values resulting from two consecutive iterations was found to be lower than 10^{-11}%. In order to evaluate the quality of these models, the final LOF percentages and the pair-wise correlation coefficients between the pure profiles in $\mathbf{C}$ and $\mathbf{S}^{\mathrm{T}}$ and the corresponding ground-truth ones were estimated and assessed.

Results are summarized in Table 2.

Table 2. LOF values and pair-wise correlation coefficients between recovered and ground-truth profiles yielded by the bilinear and trilinear MCR-ALS factorization of the simulated EEM datasets.

Noise level (%)	Profile overlap	Component	MCR-ALS (bilinear model)			MCR-ALS (trilinearity for missing data)		
			C profile$^{(+)}$	**S** profile$^{(+)}$	LOF (%)	**C** profile$^{(+)}$	**S** profile$^{(+)}$	LOF (%)
0.5	Low	1	1.000	1.000	0.5	1.000	1.000	0.5
		2	1.000	0.997		1.000	1.000	
		3	0.996	0.999		1.000	1.000	
	High	1	0.994	0.999	0.5	1.000	1.000	0.5
		2	0.993	0.995		1.000	1.000	
		3	0.997	0.951		1.000	1.000	
5	Low	1	1.000	1.000	5	1.000	1.000	5
		2	1.000	0.998		1.000	1.000	
		3	0.998	0.998		1.000	1.000	
	High	1	0.999	1.000	5	1.000	1.000	5
		2	0.993	1.000		1.000	1.000	
		3	1.000	0.942		1.000	1.000	
15	Low	1	1.000	1.000	15	1.000	1.000	15
		2	0.999	1.000		1.000	1.000	
		3	0.998	1.000		1.000	1.000	
	High	1	0.999	0.996	15	1.000	1.000	15
		2	0.989	1.000		1.000	1.000	
		3	0.990	0.897		1.000	1.000	
30	Low	1	0.999	0.999	30	0.999	1.000	30
		2	0.999	0.999		0.999	1.000	
		3	0.998	0.998		0.999	1.000	
	High	1	0.999	0.986	30	0.999	1.000	30
		2	0.991	0.999		0.998	1.000	
		3	0.994	0.924		0.998	1.000	

+Correlation coefficients between profiles recovered by MCR-ALS and simulated profiles.

In general, for low noise levels (0.5 and 5% of the total data variation) and low spectral overlap, both types of MCR-ALS factorizations yielded satisfactory outcomes. However, when the spectral overlap among pure components becomes more pronounced, the profiles recovered by the purely bilinear MCR-ALS decomposition show a significant degradation due to the increase of rotational ambiguity. Conversely, when the NIPALS-based trilinearity constraint is also imposed, stable and accurate MCR-ALS decompositions are obtained for both noise levels and all degrees of component overlap.

On the other hand, as the noise level increases (15 and 30%), the performance of the bilinear model degrades, as it is observed in Table 2. This effect is inherent to least squares problems, since the model tries to explain as much variance as possible, no matter if the variance comes from components or noise. However, it is worth noting that, even at high noise levels, the trilinear MCR-ALS model performs very well compared to the bilinear model, since the profiles are meant to be trilinear, and thus, more robust to the noise.

In summary, these results highlight that i) the new NIPALS-based implementation of the MCR-ALS trilinearity constraint is actually effective when it comes to extracting trilinear component profiles from trilinear data containing missing values and ii) trilinear MCR-ALS models obtained through the application of this novel constraint provide more accurate representations of trilinear data (compared to their purely bilinear counterparts) even when the noise level and the amount missing values are relatively high.

Excitation-emission hyperspectral image of a plant tissue sample

The conclusions drawn after the analysis of the simulated datasets are strongly corroborated by the results obtained for the real EEM hyperspectral image of a root tissue (see Figure 6). It is worth noticing that here, prior to their MCR-ALS modelling, the investigated data were preprocessed as described in Ref. [19].

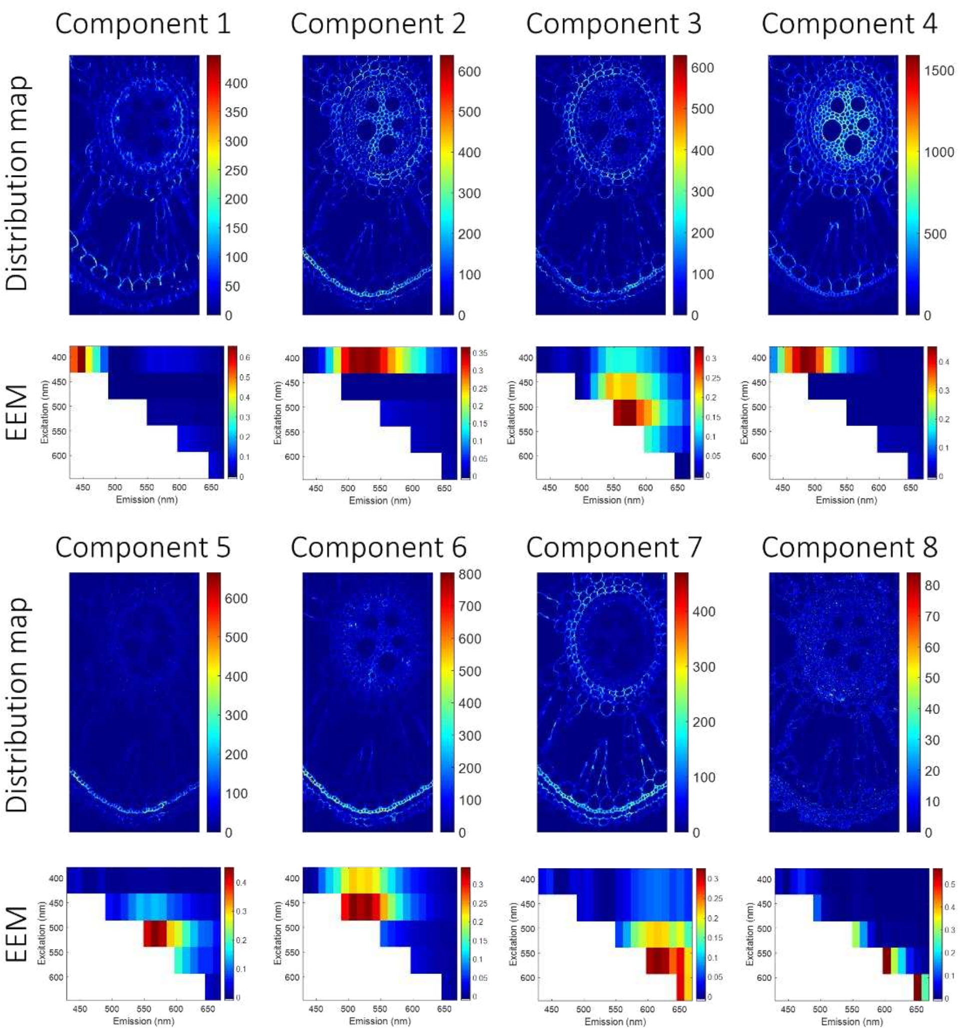


Figure 6. Pure component spatial distribution maps and pure component EEM landscapes resulting from the trilinear MCR-ALS factorization of the real EEM hyperspectral image.

Figure 6 shows the pure distribution maps and pure EEM landscapes achieved by MCR-ALS applying the NIPALS-based trilinearity constraint. The components obtained match very well those described in Ref [19]. Component 1 is present in the surrounded tissues of the center vessel (pericycle). This component boasts an excitation peak at 405 nm and an emission peak at approximately 440 nm. It is specific of specialized cells in the center vessel (phloem companion cells) and the inner part of the epidermis. Component 2 appears across the root tissue and is likely representative of non-specific lignin tissue. It is characterized by an excitation peak at 405 nm and an emission peak at around 500 nm. Component 3 is mainly associated with the outer part of the center vessel (endodermis), although it can be observed throughout the root, making it a common feature across the root tissue. Component 3 exhibits an excitation peak at around 520 nm and an emission peak at around 570 nm. Component 4 appears in the center vessel (pith), in particular in highly lignified regions. It exhibits an excitation peak at around 405 nm and an emission peak at approximately 480 nm. This particular component is likely associated with the lignified cells of the pith. Component 5 relates to specialized lignified cells of the epidermis (sclerenchyma layer of the exodermis) and is characterized by an excitation peak at around 520 nm and an emission peak at approximately 560 nm. Similarly, component 6 is prevalent in the sclerenchyma layer of the exodermis, as well as in the plant tissue regions where one would expect to find the Casparian strip (outer ring of the center vessel). The excitation spectral interval of this component ranges from approximately 405 to 470 nm, while its emission occurs at around 500 and 550 nm. Its spatial distribution across the root cross sections is in agreement with the findings reported by Vishal, B. et al. [25], which may indicate the presence of suberin. Component 7 appears in the outer ring of the center vessel and the root (endodermis and exodermis-epidermis). Interestingly, small vesicles within certain vessels are specifically associated with this component which could evidence the existence of silica bodies over the surface of the plant tissue section. Component 7 exhibits an excitation peak at around 570 nm and an emission peak at around 620 nm. Finally, Component 8 explains an artifact attributed to residual Rayleigh scattering with a non-relevant signal on the model.

As mentioned above, the results reported are in a very good agreement with those reported in Ref. [19], where a trilinearity constraint for MCR-ALS based on sequential use of calculations using submatrices was proposed. Such a fact confirms the goodness of the new implementation of the constraint, which provides comparable results to those previously obtained with the correct, but more complex and data–dependent implementation of trilinearity described in Ref. [19].

EEM of a pharmaceutical mixture

The EEM mixture data described before were also analyzed by means of MCR-ALS imposing uniquely non-negativity constraints on $\mathbf{C}$ and $\mathbf{S}^{\mathrm{T}}$ (bilinear model) and forcing at the same time non-negativity and trilinearity (trilinear model). In all models, we set the maximum number of iterations to 2000 and employed a convergence criterion of 10^{-11}%.

The summarized results are shown in Table 3.

Table 3. LOF values and pair-wise correlation coefficients between recovered and ground-truth profiles yielded by the bilinear and trilinear MCR-ALS factorization of the EEM pharmaceutical data.

Component	MCR-ALS (bilinear model)			MCR-ALS (trilinearity for missing data)		
	C profile$^{(+)}$	**S** profile$^{(+)}$	LOF (%)	**C** profile$^{(+)}$	**S** profile$^{(+)}$	LOF (%)
ASA	1.000	1.000	0.7	1.000	1.000	0.8
IBU	0.993	0.744		0.998	0.997	

+ Correlation coefficients between profiles recovered by MCR-ALS and ground-truth profiles.

Table 3 clearly shows that both the bilinear and trilinear MCR-ALS model shows perfect correlations (1.000) in concentration profiles for ASA when they are compared to the true concentration profile (Figure 7). However, when the trilinearity constraint is applied, the recovered concentration profile for IBU is slightly better for the trilinear model (0.993 vs 0.998). This bias is observed when the true concentration profile is plotted against the recovered profile.

On the other hand, while the pure spectral profile of ASA is perfectly recovered in both models (1.000), a significant difference is observed in the recovered spectral profile of IBU for the bilinear model (0.744) (Figure 7). This result is expected since the dataset does not contain enough selectivity on the concentration profile and this causes the presence of rotational ambiguity in the related pure spectrum. In addition, the huge difference among the signal of ASA (major) and IBU (minor) can result in a degradation of the solution for the minor compound IBU.

The LOF values yielded by the two different models are very similar (0.7 and 0.8% for the bilinear and trilinear model, respectively). This is an indicator of the fact that trilinearity holds in this case, since in similar situations the model residuals should not vary significantly for bilinear and trilinear decompositions.

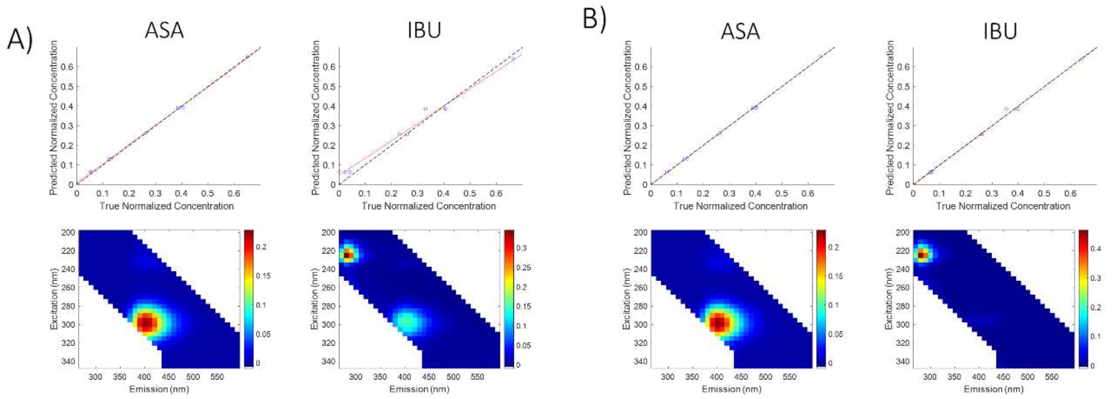


Figure 7. A) Predicted concentration (top) and pure EEM profiles (bottom) by MCR-ALS applying only non-negativity constraint. B) Predicted concentration (top) and pure EEM profiles (bot) by MCR-ALS applying trilinearity constraint. Notice that concentration profiles are graphed in an expected *vs* predicted value plot where actual and ideal fit lines are represented as red solid line and dashed black lines, respectively.

Conclusions

A novel implementation of the trilinearity constraint in MCR-ALS capable of handling data containing missing values was presented. This implementation is based on the application of the NIPALS algorithm to force the common shape required for trilinear component profiles. NIPALS allows skipping missing values during computations through a sequence of row-by-row and column-by-column least-squares estimation operations involving only the available entries of the data set. For this reason, it bypasses the use of imputation methods and its mathematical simplicity constitutes a considerable improvement over existing approaches based, for example, on the principles of SVD. Besides, it is suited to cope with any kind of missing data pattern and even with data exhibiting high amounts of missing elements. The idea behind this implementation can easily be extended for imposing multilinearity constraints when higher-order multi-way data are handled and incorporated in all publicly available MCR-ALS interfaces.

References

1. Kruskal, Joseph B. "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics." *Linear algebra and its applications* 18.2 (**1977**): 95-138.
2. Tauler, Roma. "Multivariate curve resolution applied to second order data." Chemometrics and intelligent laboratory systems 30.1 (**1995**): 133-146.
3. Harshman, Richard A., and Margaret E. Lundy. "PARAFAC: Parallel factor analysis." Computational Statistics & Data Analysis 18.1 (**1994**): 39-72.
4. Bro, Rasmus. "PARAFAC. Tutorial and applications." Chemometrics and intelligent laboratory systems 38.2 (**1997**): 149-171.
5. Sanchez, Eugenio, and Bruce R. Kowalski. "Tensorial resolution: a direct trilinear decomposition." Journal of Chemometrics 4.1 (**1990**): 29-45.
6. De Juan, Anna, Joaquim Jaumot, and Romà Tauler. "Multivariate Curve Resolution (MCR). Solving the mixture analysis problem." Analytical Methods 6.14 (**2014**): 4964-4976.
7. de Juan, Anna, and Roma Tauler. "Multivariate Curve Resolution: 50 years addressing the mixture analysis problem–A review." *Analytica Chimica Acta* 1145 (**2021**): 59-78.

8. Tauler, R., I. Marqués, and E. Casassas. "Multivariate curve resolution applied to three-way trilinear data: Study of a spectrofluorimetric acid–base titration of salicylic acid at three excitation wavelengths." *Journal of Chemometrics: A Journal of the Chemometrics Society* 12.1 (**1998**): 55-75.
9. Engelsen, Søren Balling, and Rasmus Bro. "PowerSlicing." *Journal of Magnetic Resonance* 163.1 (**2003**): 192-197.
10. Devos, Olivier, et al. "Multivariate curve resolution slicing of multiexponential time-resolved spectroscopy fluorescence data." *Analytical Chemistry* 93.37 (**2021**): 12504-12513.
11. Bech Risum, Anne, Jesper Løve Hinrich, and Åsmund Rinnan. "Multiway Decomposition Followed by Reconvolution of Fluorescence Time Decay Data." *Analytical Chemistry* 95.51 (**2023**): 18697-18708.
12. Câmara, Anne BF, et al. "Excitation-emission fluorescence spectroscopy coupled with PARAFAC and MCR-ALS with area correlation for investigation of jet fuel contamination." *Talanta* 266 (**2024**): 125126.
13. Marin-Garcia, Marc, and Romà Tauler. "Chemometrics characterization of the Llobregat river dissolved organic matter." *Chemometrics and Intelligent Laboratory Systems* 201 (**2020**): 104018
14. Gómez-Sánchez, Adrián, et al. "Study of the photobleaching phenomenon to optimize acquisition of 3D and 4D fluorescence images. A special scenario for trilinear and quadrilinear models." *Microchemical Journal* 191 (**2023**): 108899.
15. Tomasi, Giorgio, and Rasmus Bro. "PARAFAC and missing values." *Chemometrics and Intelligent Laboratory Systems* 75.2 (**2005**): 163-180.
16. Andersen, Charlotte Møller, and Rasmus Bro. "Practical aspects of PARAFAC modeling of fluorescence excitation-emission data." *Journal of Chemometrics: A Journal of the Chemometrics Society* 17.4 (**2003**): 200-215.
17. Elcoroaristizabal, Saioa, et al. "PARAFAC models of fluorescence data with scattering: A comparative study." *Chemometrics and Intelligent Laboratory Systems* 142 (**2015**): 124-130.
18. Gomez-Sanchez, Adrian, et al. "Dealing with missing data blocks in Multivariate Curve Resolution. Towards a general framework based on a single factorization model." *Submitted to: Trends in analytical Chemistry* (**2024**).
19. Gomez-Sanchez, Adrian, et al. "The trilinear constraint adapted to solve data with strong patterns of outlying observations or missing values." *Chemometrics and Intelligent Laboratory Systems* 231 (**2022**): 104692.
20. Wold, Herman. "Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach." *Journal of Applied Probability* 12.S1 (**1975**): 117-142.
21. Grung, Bjørn, and Rolf Manne. "Missing values in principal component analysis." *Chemometrics and Intelligent Laboratory Systems* 42.1-2 (**1998**): 125-139.
22. Gómez-Sánchez, Adrián, et al. "3D and 4D image fusion: coping with differences in spectroscopic modes among hyperspectral images." *Analytical Chemistry* 92.14 (**2020**): 9591-9602.
23. Jaumot, Joaquim, et al. "A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB." *Chemometrics and intelligent laboratory systems* 76.1 (**2005**): 101-110.
24. Windig, Willem, and Jean Guilment. "Interactive self-modeling mixture analysis." Analytical chemistry 63.14 (**1991**): 1425-1432.25.
25. Vishal, Bhushan, et al. "Os TPS 8 controls yield-related traits and confers salt stress tolerance in rice by enhancing suberin deposition." *New Phytologist* 221.3 (**2019**): 1369-1386.

Supplementary material

The MCR-ALS trilinearity constraint for data with missing values

Adrián Gómez-Sánchez[1,2], Raffaele Vitale[2], Pablo Loza-Alvarez[3], Cyril Ruckebusch[2], Anna de Juan[1]

[1]Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028, Barcelona, Spain

[2]Univ. Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, F-59000 Lille, France

[3]ICFO-Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology, 08860 Castelldefels, Barcelona, Spain

*Corresponding authors:

Adrián Gómez-Sánchez (agomezsa29@alumnes.ub.edu)

Anna de Juan (anna.dejuan@ub.edu)

Abstract

This document contains a more comprehensive description of the simulated data sets investigated in this study.

Simulated data sets

A series of three-component 4D EEM fluorescence images was here simulated based on the computational procedure proposed by Gómez-Sánchez et al in [19]. Figures S1 and S2 display the spatial distribution maps generated for the individual components concerned as well as their corresponding excitation and emission spectra (calculated as combinations of Gaussian functions).

Two scenarios of low and high spectral overlap among components, respectively, were accounted for. In each scenario, images were contaminated with Poisson noise at different levels: 0.5%, 5%, 15%, and 30% of the total data variation as per the following equation:

$$E(\%) = \sqrt{\frac{\sum_i \sum_j (d_{i,l} - d_{n_{i,l}})^2}{\sum_i \sum_j \left(d_{n_{i,l}}\right)^2}} \cdot 100 \qquad \text{(S1)}$$

where $d_{i,l}$ and $d_{n_{i,l}}$ denote the i,l^{th} element of the noisy data and the i,l^{th} element of the noise-free data, respectively. Notice that MCR-ALS should ideally yield a lack-of-fit value very close to $E(\%)$.

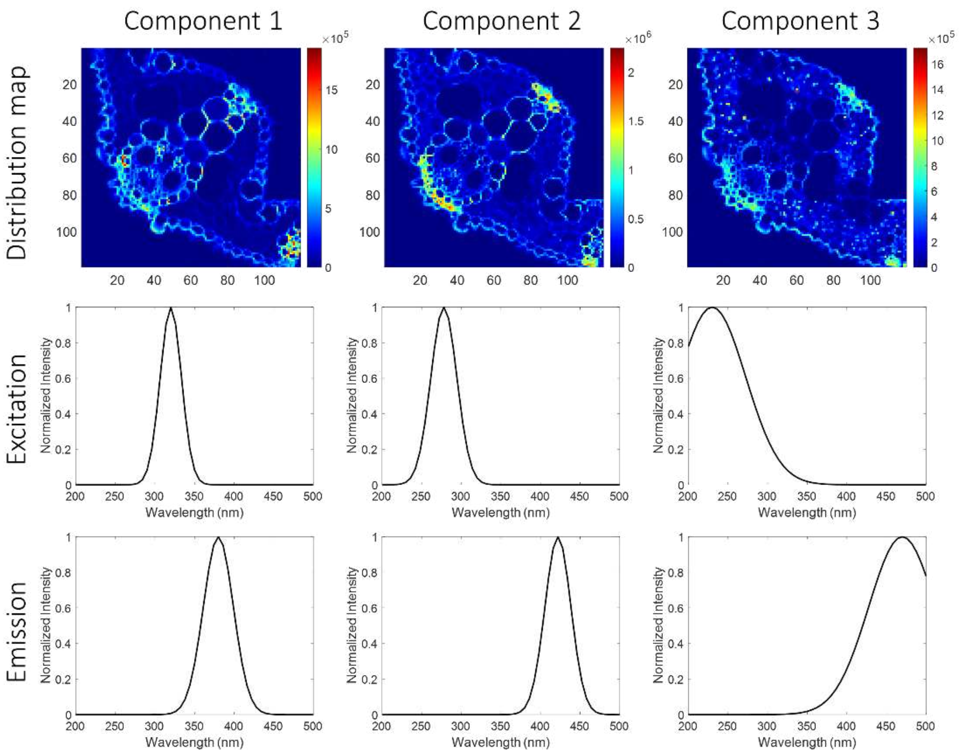


Figure S1. Pure component profiles used to simulate data with low spectral overlap.

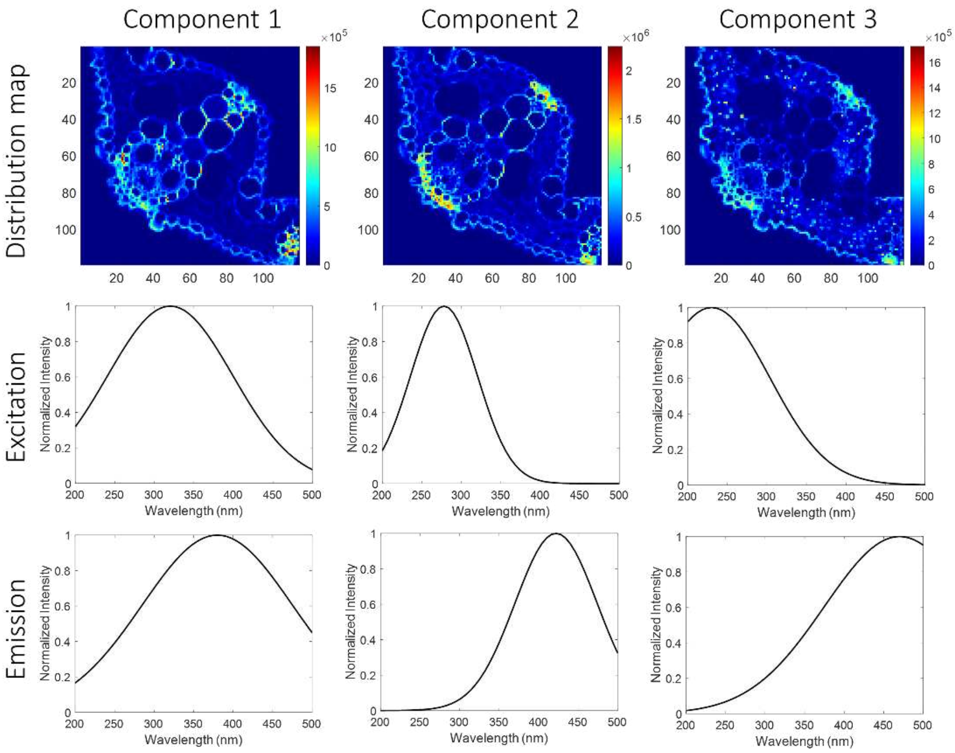


Figure S2. Pure component profiles used to simulate data with high spectral overlap.

In total, eight EEM hyperspectral images, were simulated per each simulation condition. A systematic pattern of missing data was finally imposed on every one of these images.

Contents lists available at ScienceDirect

Microchemical Journal

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

Study of the photobleaching phenomenon to optimize acquisition of 3D and 4D fluorescence images. A special scenario for trilinear and quadrilinear models

Adrián Gómez-Sánchez [a,b,*], Iker Alburquerque Alvarez [a], Pablo Loza-Alvarez [c], Cyril Ruckebusch [b], Anna de Juan [a,*]

[a] Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028 Barcelona, Spain
[b] LASIRE – Laboratory of Advanced Spectroscopy, Interactions, Reactivity and Environment Université Lille, CNRS, UMR 8516, Cité Scientifique, Bâtiment C5, 59000 Lille, France
[c] ICFO-Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology, 08860 Castelldefels, Barcelona, Spain

ARTICLE INFO

Keywords:
Hyperspectral imaging
Fluorescence imaging
Multivariate Curve Resolution
Trilinear and quadrilinear models
Vegetal tissues
Rice root

ABSTRACT

Emission (3D) and excitation-emission (4D) fluorescence images allow covering wide excitation and emission spectral ranges and, hence, provide very complete information for a good characterization and location of fluorophores in samples. However, when the acquisition time of the image is too long, degradation of the fluorescence signal of compounds and sample photodamage can occur due to photobleaching. This phenomenon is due to the long exposure time of the sample to the light source and can hinder the detection and the proper characterization of the fluorophores in samples.

The main purpose of this research is providing a methodology to obtain and interpret the information of fluorescence images for the characterization of samples without suffering the consequences of photobleaching. Such a goal implies a first thorough knowledge of the photobleaching phenomenon to adapt the fluorescence imaging measurement for an optimal characterization of the fluorophores present in samples.

The proposed approach relies first on a study of time-series of 3D or 4D fluorescence images to characterize spatially and spectroscopically the fluorophores present in the samples and their photobleaching behaviour. Since photobleaching is fluorophore-dependent, the unmixing algorithm Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) is applied to the set of fluorescence images acquired as a function of time to understand the specific behaviour of every fluorophore. The characteristics of the photobleaching phenomenon and the nature of the fluorescence measurement offer a challenging scenario to look for adapted implementations of trilinear and quadrilinear models within the MCR framework. From the results obtained, appropriate instrumental settings are adopted for an image acquisition that allows the correct spatial and spectroscopic characterization of fluorophores in samples.

To test the potential of this methodology, the characterization of thin cross-sections of the *Oryza sativa* (commonly called rice) root have been studied due to the co-occurrence of several natural fluorophores in vegetal tissues.

1. Introduction

Hyperspectral imaging is a powerful analytical technique that provides spectral and spatial information of the sample surface. Each pixel of a hyperspectral image (HSI) is related to spectral information and imaging platforms are adapted to work with many spectrophotometric techniques. Hyperspectral imaging is used in many research fields, such as food quality [1,2], medicine [3–5], pharmacy [6–8] por biology [9–11]. Within the spectroscopic techniques adopted in imaging systems, fluorescence offers a high sensitivity to detect low concentrations of fluorophore compounds and the capability to provide a very detailed information on the distribution of these compounds on the sample surface due to its high spatial resolution, which can go down to several tens of nm if super resolution imaging techniques are used [12–14].

[*] Corresponding authors.
E-mail addresses: aderegomez@gmail.com (A. Gómez-Sánchez), anna.dejuan@ub.edu (A. Juan).

https://doi.org/10.1016/j.microc.2023.108899
Received 5 April 2023; Received in revised form 19 May 2023; Accepted 21 May 2023
Available online 22 May 2023

Confocal fluorescence microscopy can provide 3D or 4D fluorescence HSI depending on the spectroscopic information recorded [15]. Thus, a 3D fluorescence image acquires a full emission spectrum per each pixel of the sample surface at a fixed excitation wavelength, providing three-dimensional data $(x, , \lambda)$. Instead, a 4D fluorescence image associates a full 2D excitation/emission landscape to every pixel, covering a range of excitation and emission wavelengths, providing four-dimensional data $(x, y, \lambda_{ex}, \lambda_{em})$. The combined use of 3D and 4D fluorescence images provides information on a wide spectral range of excitation and emission wavelengths and, therefore, allows an accurate sample characterization [15]. However, issues related to the instrumental image settings and to the big size of image data need to be addressed, as explained below.

The acquisition time of a 4D fluorescence HSI is relatively long. As a consequence, the degradation of the fluorescent signal of the compounds by photobleaching can occur [16]. Thus, photobleaching can hinder the detection and proper characterization of fluorophores in samples. In the worst scenario, a sample can be damaged due to laser exposition, making impossible the acquisition of images of living tissues. Hence, this phenomenon needs to be studied by collecting consecutive images as a function of time and studying the intensity decay of each fluorophore in the sample, instead of the global intensity decay. This kind of preliminary study is necessary to obtain suitable settings that enable a proper fluorescence image acquisition for sample characterization.

Understanding the information offered by fluorescence images requires chemometric tools due to the large size and complexity of the data sets acquired. Both for the study of the photobleaching phenomenon and the subsequent characterization of samples, it is relevant unmixing the raw signal into the contributions of the pure fluorophores in the samples analyzed. A solution to this problem is provided by the Multivariate Curve Resolution – Alternating Least Squares (MCR–ALS) method [17] that works iteratively decomposing the raw HSI into the pure spectral signatures and concentration maps of the image constituents. Such a procedure can work analyzing a single image or an ensemble of related images in a multiset fashion. Besides, the flexibility in data configuration allows handling 3D and 4D images [15]. The profiles issued from MCR-ALS provide a complete chemical, semiquantitative and distributional characterization of the fluorophores present in the samples and an additional description of the individual decay behavior of every fluorophore when photobleaching is investigated.

In the present work, the image acquisition protocol assisted by MCR-ALS, adapted to study photobleaching and to handle the specificities of 3D and 4D fluorescence HSI, is tested on samples of cross sections of *Oryza sativa* (rice) root. This example is a perfect testing scenario since vegetal tissues contain many natural fluorophores colocalized across the sample surface analyzed. In the following sections, the protocol to acquire and interpret images obtained in photobleaching and characterization studies is described, together with the most important results related to the specific study of *Oryza Sativa* root cross-sections.

2. Experimental work

2.1. Plant growth and sample preparation

Rice plants were grown as in Ref. [15]. After harvest, thin cross sections of roots were manually cut and placed on a 1 mm-thickness CaF2 slide with a drop of phosphate-buffered saline solution, covered with a 0.5 mm-thickness CaF2 coverslip and sealed with nail polish, to avoid water evaporation during the experiment. Fig. 1 shows the structure of a cross section of *Oryza sativa* root with the different parts identified [18].

2.2. Image acquisition

All images were collected using a Leica TCS SP8 STED 3X microscope (Leica Microsystems, Mannheim, Germany) with an HC PL APO CS2 10 × /0.40 DRY objective. The instrumental parameters were set, as explained below, depending on the experiment (photobleaching on 3D images, photobleaching on 4D images and characterization images).

For studying the photobleaching on 3D images, a 405 nm excitation laser (power of 89 μW at the sample plane) was selected with an emission range from 432.5 to 597.5 nm. The sampling interval and bandwidth were 5.69 nm with a dwell time set 3.8 μs. Six consecutive images of the same root cross-section were acquired to study the photobleaching phenomenon covering an interval of time going from 0 to 40 min. The total acquisition time of every 3D image is eight minutes. The 3D images were split in two regions of interest (ROIs): the epidermis and the stele, sized 428 × 315 μm^2 and 270 × 315 μm^2 respectively.

On the other hand, for studying the photobleaching on 4D images, a supercontinuum white light laser (WLL) was used. The excitation range covered 470 to 582 nm with a sampling interval of 8 nm (power of 146 μW at the sample planned). The emission range covered 504–624 nm with a sampling interval and bandwidth of 6 nm. The dwell time was set to 7.7 μs. Three consecutive images of the same root cross-section were acquired to study the photobleaching across the images covering an interval of time going from 0 to 24 min. The total acquisition time of every 4D image is 12 min.

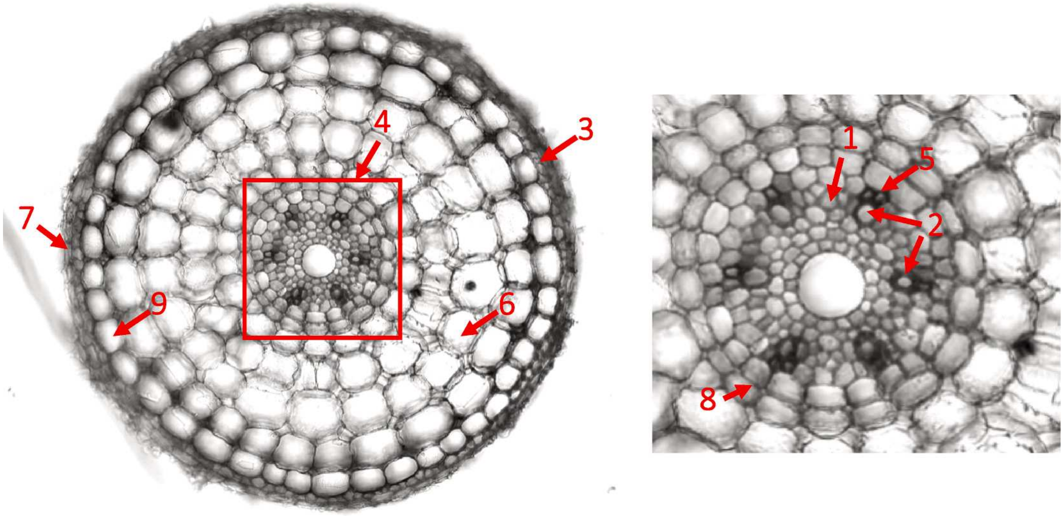


Fig. 1. Main anatomy of a cross-section of Oryza sativa Japonica root cross-section. 1) Phloem. 2) Xylems. 3) Sclerenchyma layer. 4) Stele. 5) Xylem-pole pericycle. 6) Cortex. 7) Epidermis. 8) Endodermis. 9) Exodermis.

To study the characterization of the natural fluorophores present on the rice root, a final image of a new root cross-section was acquired with only five excitation wavelengths (405, 470, 520, 570, 620 nm) and covering an emission range of 435–663 nm with 12 nm sampling interval and a bandwidth of 12 nm. The dwell time was 32 µs in all excitation wavelengths, except for 405 nm, where 15 µs were used.

All images had a pixel resolution of 450 × 450 nm^2.

3. Data treatment

This section covers the preprocessing applied to improve the signal-to-noise ratio of the fluorescence images and the description of the protocol based on the use of Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) for photobleaching and characterization studies using 3D and 4D fluorescence images.

3.1. Preprocessing

The images obtained have a high spatial resolution, with a pixel size of 450 × 450 nm^2, but the signal to noise ratio is too low for multivariate analysis. To address this issue, a binning of adjacent pixels (10 × 10) was done to improve the spectroscopic signal quality. Finally, the background was cropped.

In EEM measurements, the scattering produced by the Rayleigh emission has to be removed from the data, since it does not follow either bilinear nor trilinear responses. Thus, few emission channels close to the laser wavelength were set as not a number (NaN).

For the characterization study, the channel at excitation 405 nm and emission 591 nm showed a loss of signal, probably due to a problem with the detector. The channel was interpolated using the nearest emission wavelengths.

3.2. Multivariate Curve Resolution-Alternating least Squares (MCR-ALS)

Multivariate Curve Resolution-Alternating Least Squares [17] is an unmixing algorithm that decomposes the raw mixed information contained in an initial data set into a bilinear model of profiles related to their pure components, according to equation (1):

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

In a spectroscopic context, **D** is a table of raw mixture spectra and **S** and **C** are the matrices that contain the pure spectra and the related concentration profiles of the pure compounds, which can reproduce appropriately the information contained in the initial data set **D**. **E** is the variance unexplained by the bilinear model.

The MCR-ALS method works optimizing iteratively the **C** and **S** matrices via an alternating least squares procedure under constraints. An initial estimate of the matrix **C** or **S**, often obtained with a purest variable selection method, is required to start the optimization procedure. The constraints used are based on general mathematical or chemical properties that the profiles in **C** and **S** matrices naturally obey. The role of constraints is providing chemically meaningful profiles and decreasing the rotational ambiguity associated with the bilinear decomposition. The choice of the constraints is adapted to the nature of the profiles to be resolved and can be optionally applied per component and per mode (**C** and **S**). The iterative optimization is finished when a convergence criterion is reached, usually related to the fulfillment of a preset threshold value linked to the relative difference in lack of fit (see eq. 2) among consecutive iterations.

$$LOF(\%) = 100 \bullet \sqrt{\frac{\sum_{i,j} e_{i,j}^2}{\sum_{i,j} d_{i,j}^2}}$$

In the lack of fit expression, $d_{i,j}$ is an element of matrix **D** and $e_{i,j}$ from matrix **E**.

MCR-ALS has been applied in many diverse research fields and is particularly suitable for hyperspectral image analysis. The flexible data configurations and use of constraints of this algorithm allows the analysis of 3D and 4D HSIs [15,17,19–22]. The multiset modality of the method adapts to the simultaneous analysis of several related HSIs acquired with the same spectroscopic platform and to challenging image fusion scenarios where HSIs come from platforms differing in spectral dimensionality and spatial resolution.

The application of MCR-ALS to a single 3D or 4D fluorescence image implies transforming the initial image array into a data table and applying the suitable constraints. 3D images can be displayed as a cube where a full emission spectrum acquired at a fixed excitation wavelength is associated with each pixel of the sample surface. The image cube consists of two spatial pixel coordinates, x and y, and one spectral dimension, λ (see Fig. 2A). The cube can be easily unfolded into a data table by putting one pixel emission spectrum under the other. The **D** matrix obtained has a number of rows equal to the number of pixels ($x \times y$) and a number of columns equal to the number of emission wavelengths (λ). The bilinear model provided by MCR consists of a matrix with the pure emission spectra of the fluorophores in the sample (**S** matrix) and the related pixel concentration arrays (**C** matrix), which refolded according to the structure of the 2D sample surface provide the fluorophore distribution maps. The basic constraints that can be used in this resolution are non-negativity for both the concentration profiles and the pure emission spectra and spectra normalization in the **S** matrix.

4D images instead associate a 2D excitation-emission (EEM) landscape per every pixel. Therefore, the data table **D** is obtained by putting the vectorized 2D EEM landscape of every pixel, i.e., the concatenated emission spectra λ_{em} obtained at the different excitation wavelengths (λ_{ex_1} to λ_{ex_N}), one under the other (see Fig. 2B). The **D** matrix obtained has a number of rows equal to the number of pixels ($x \times y$) and a number of columns equal to the total number of emission channels in the vectorized EEM spectrum. The bilinear model provided by MCR consists of a matrix with a set of pure pixel concentration arrays, which are turned into fluorophore distribution maps, and a matrix with pure vectorized EEM landscapes related to each fluorophore (**S** matrix), which can be refolded into 2D EEM landscapes as well. All constraints mentioned for 3D images can be applied to 4D images. However, the nature of 2D EEM spectra allows for the application of the trilinearity constraint to matrix **S.** In plain words, the action of this constraint is forcing that all emission spectra within the EEM vectorized profile of a pure compound show the same shape across all the excitation wavelengths covered. Since the 2D EEM landscapes of fluorescence images show a systematic pattern of missing values when the excitation and emission ranges overlap, i.e., no emission fluorescence values are obtained if the emission wavelength is lower than the excitation wavelength, a dedicated implementation of the trilinearity constraint able to handle structures with missing values has been applied [23]. The three modes of the trilinear model linked to a 4D single image analysis by MCR-ALS would be the concentration and the emission spectra (explicit modes in the MCR decomposition) and the excitation mode (embedded in the vectorized EEM landscape).

An asset of MCR-ALS is the possibility to work with several related HSIs into a single multiset structure. For both 3D and 4D images, this can be easily accomplished by appending the blocks of spectral information of the different sample images (either single emission spectra in 3D images or vectorized 2D EEM landscapes in 4D images) one under the other. Such a multiset configuration is used in fluorophore characterization studies by analyzing multisets formed by images of different samples and the constraints used would be the same as for the analysis of single 3D and 4D images.

Multisets related to photobleaching studies deserve a special comment. As described in the introduction, the study of the photobleaching phenomenon can be easily carried out by analyzing simultaneously a set of fluorescence HSIs obtained on the same sample over time (see Fig. 3).

The nature of the photobleaching phenomenon allows taking advantage of the singular behavior of the fluorophores and the related

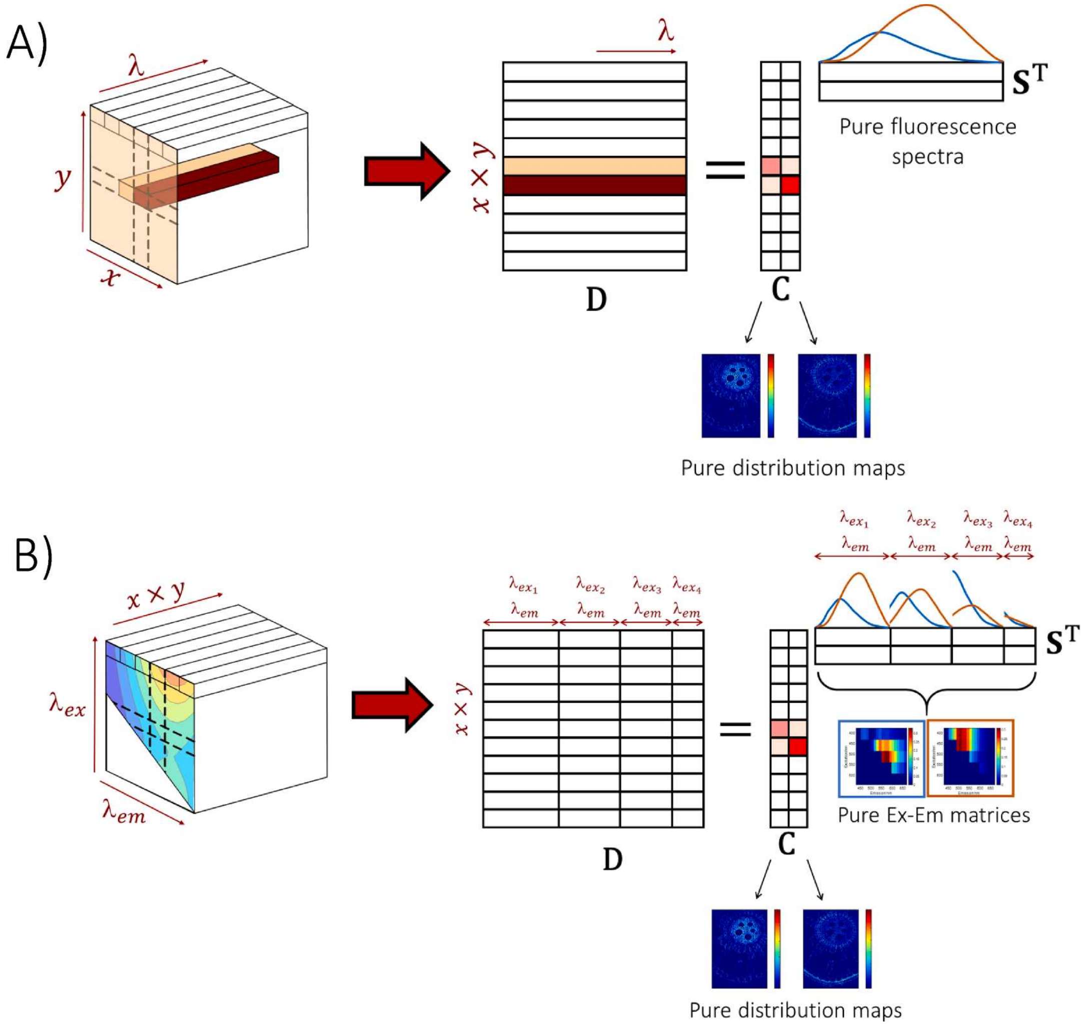

A)
λ
y
x
λ
x × y
D
C
S^T
Pure fluorescence spectra
Pure distribution maps
B)
x × y
λ_{ex}
λ_{em}
λ_{ex_1} λ_{em}
λ_{ex_2} λ_{em}
λ_{ex_3} λ_{em}
λ_{ex_4} λ_{em}
x × y
D
C
S^T
Pure Ex-Em matrices
Pure distribution maps

Fig. 2. A) MCR model for a 3D fluorescence image. B) MCR model for a 4D fluorescence image.

samples in the experiments performed. On the one hand, since the measurement is based on fluorescence spectroscopy, the shape of the emission spectra for every pure fluorophore remains invariant across the wavelength excitation range and/or across time and only the global signal intensity varies; on the other hand, since HSIs are collected on the same sample as a function of time, the shape of the concentration map of every fluorophore remains invariant across time and only the fluorescence intensity is modified because of photobleaching. These two facts allow a high flexibility in terms of data configuration and implementation of model constraints, as will be described below for studies carried out by using 3D and 4D images.

Fig. 3A shows the data arrangement and MCR model for a photobleaching study based on a time-series of 3D images. In this case, since the spatial structure of the concentration maps of every fluorophore is invariant over time and, hence, the shape of the related concentration profile, the blocks of spectral information from the different images are appended one beside the other forming a row-wise augmented multiset. As a consequence, the resulting MCR model obtained is formed by a single **C** matrix, which reflects the invariant shape of concentration profiles and, hence, of the related refolded maps of every fluorophore, and an augmented $\mathbf{S}^\mathbf{T}$ matrix, which contains in every row the concatenated emission spectra of a single fluorophore at the different photobleaching times studied. Other than non-negativity in the concentration and spectral profiles, trilinearity is applied to the $\mathbf{S}^\mathbf{T}$ matrix, forcing the shape of the emission spectrum of every pure fluorophore to remain constant at the different photobleaching times monitored. Finally, the pure decay signal of every fluorophore is obtained representing the area under the profile of each concatenated pure emission spectrum as a function of the related photobleaching time. In the case of photobleaching based on 3D images, the direction extended in time has been the spectral one because this mode was the least selective. Due to this fact, trilinearity could be applied and the pure spectra were resolved without ambiguity [17].

Photobleaching based on collecting 4D images over time required a different data configuration, shown in Fig. 3B). Every 4D acquired was unfolded as shown in Fig. 2B) and the blocks of information related to every photobleaching time were organized one under the other. The result is a row- and column-wise augmented multiset and the resulting MCR model will be formed by an augmented **C** matrix formed by the pixel concentration arrays (turned into concentration maps) linked to each photobleaching time studied and an augmented **S** matrix that has the same information as when a single 4D image is analyzed (see Fig. 2B). Adding to the non-negativity constraint, the trilinearity constraint has been used in the two directions of the MCR model. In the spectral direction, the adaptation of this constraint to handle missing values has been used. In the concentration direction, the classical implementation of trilinearity was applied forcing all concentration profiles (hence, maps) of the same fluorophore to have the same shape along time. The fluorescence decay profiles of every fluorophore can be subsequently obtained by integrating (summing) the values of the elements of the concentration profiles at every photobleaching time.

Looking carefully at the way photobleaching studies are done and

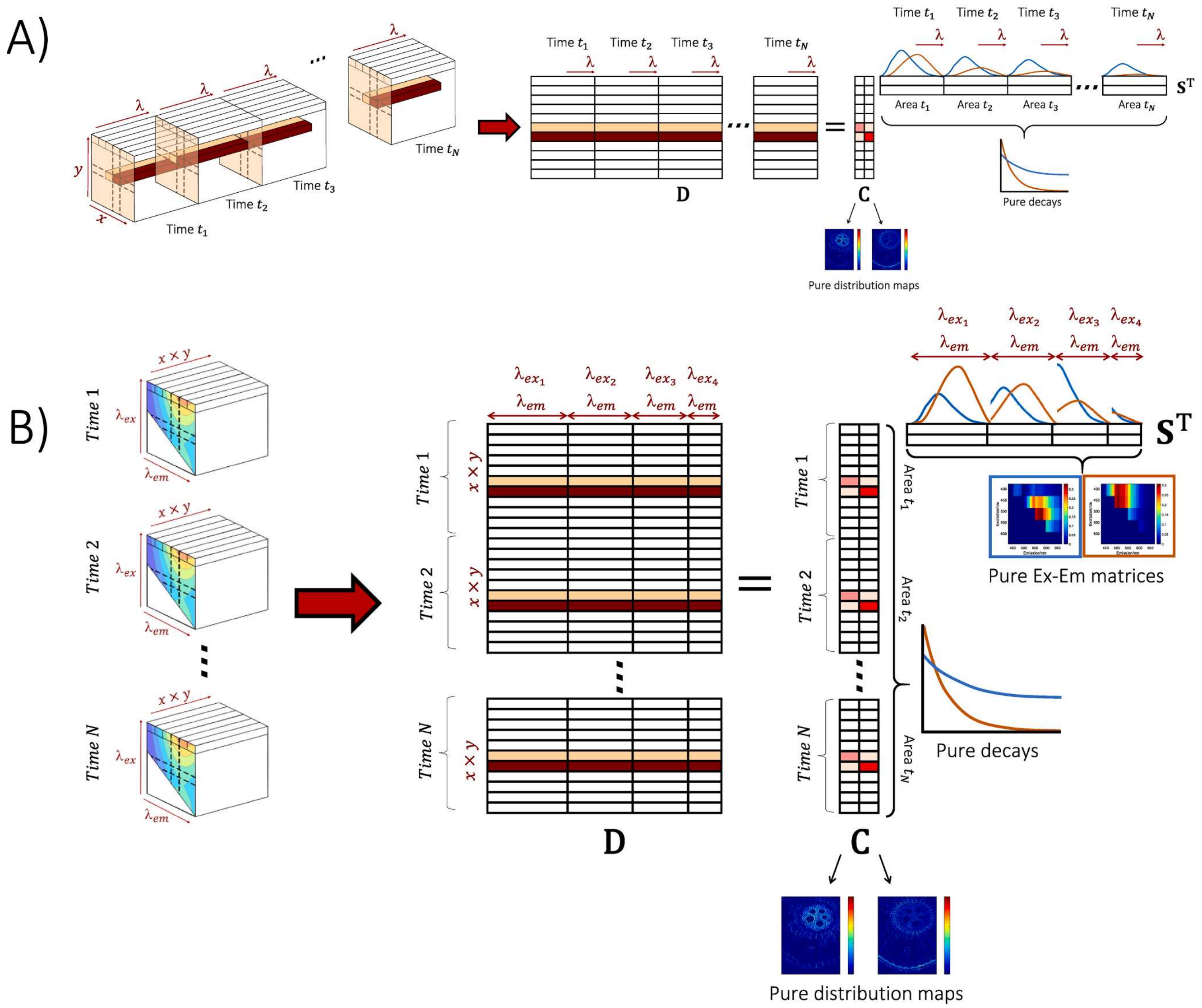


Fig. 3. A) Multiset configuration and related MCR model for a photobleaching experiment based on: A) 3D fluorescence images and B) 4D fluorescence images.

the way MCR analysis is performed, photobleaching studies based on 3D fluorescence images obey a trilinear model, where the three modes are the concentration profiles and the emission spectra (explicit MCR modes), and the decay profile mode (embedded in the set of emission spectra augmented over time). Analogously, photobleaching based on 4D fluorescence images obeys a quadrilinear model (see Fig. 3B), where the four modes are the concentration profiles and the emission spectra (explicit MCR modes), the excitation mode (embedded in the vectorized EEM landscape) and the decay profile mode (embedded in the set of concentration profiles). Thus, MCR provides a flexible framework for multilinear model implementation and, as a consequence of the use of these higher-order multilinear models, the fluorophore information can be recovered in a unique way.

3.3. Software

All in-house routines, scripts and analyses generated to preprocess and the analyze the data were performed using MATLAB 2021 (The Mathworks, Inc., Natick, MA).

4. Results and discussion

The photobleaching studies were performed by taking consecutive images (3D or 4D) on a single root cross-section over time under the same instrumental conditions and analyzing them by MCR-ALS. The results obtained in the study of the photobleaching effect helped to set the optimal image acquisition settings for the final characterization study of the fluorophores present in rice root samples. Table 1 shows a summary of the models obtained in all studies carried out. The nature, location and photobleaching decay of every component will be discussed in the following sections.

4.1. Photobleaching studies

For the results reported below, initial spectral estimates were calculated by a SIMPLISMA-based method [24]. Non-negativity was applied to **C** and **S**. Trilinearity constraint is often used and adapted depending on the context of the study, as will be described below. Convergence criterion was set to 10^{-6} % difference in lack of fit among consecutive iterations. After the MCR-ALS analysis, to recover the distribution maps at the initial high resolution, an extra least squares step

Table 1
Instrumental conditions and MCR-ALS results of each of the characterization experiments.

Dataset	Excitation range (nm)	Emission range (nm)	No. images (time step/min)	NC (*)	LOF (%)	Explained variance (%)
Photobleaching						
3D	405	432.5–597.5	6 (8)	3	10.7	98.9
4D	470–582	504–624	3 (12)	5	18.5	96.6
Characterization						
4D	405,470,520, 570,620	435–663	–	8	3.6	99.9

(*) NC: Number of components.

was performed using the optimal spectral profiles and the image data without binning.

4.1.1. 3D fluorescence images

The six 3D fluorescence images were concatenated as Fig. 3A. Trilinearity constraint was applied to force the emission shape to be the same for each component over the time. Three components were resolved (Fig. 4). The variance of the original data, i.e., the binned image (see section 3.1), is well explained by the MCR – ALS model (98.9%).

The first component found by MCR-ALS is significantly affected by photobleaching, losing 41% of the signal after 40 min of image acquisition. This component seems to be related to small vesicles (red color in Fig. S1), the cortex and the external part of the epidermis, with an emission maximum at 450 nm. The second component is present in the endodermis and the sclerenchyma layer of the epidermis (green color in Fig. S1). This component is less affected than the first one by photobleaching, having a loss of signal only of 6% after 40 min of image acquisition, with an emission maximum at 500 nm. Finally, the last component is specific of the inner part of the stele, where the cells are used to be more lignified (blue color in Fig. S1), with an emission maximum around 470 nm. The photobleaching affects even less this component than the previous ones, having a loss of signal of 3% after 40 min of imaging.

4.1.2. 4D fluorescence images

The three 4D fluorescence images were concatenated as in Fig. 3B into a single multiset. Trilinearity constraint was applied to force the concentration profile shape to be the same for each component over the time. On the other hand, the pure fluorescence profiles were forced to have the same shape across the excitations [23], resulting in a quadrilinear model. Convergence criterion was set to 10^{-6} % difference in lack of fit among consecutive iterations.

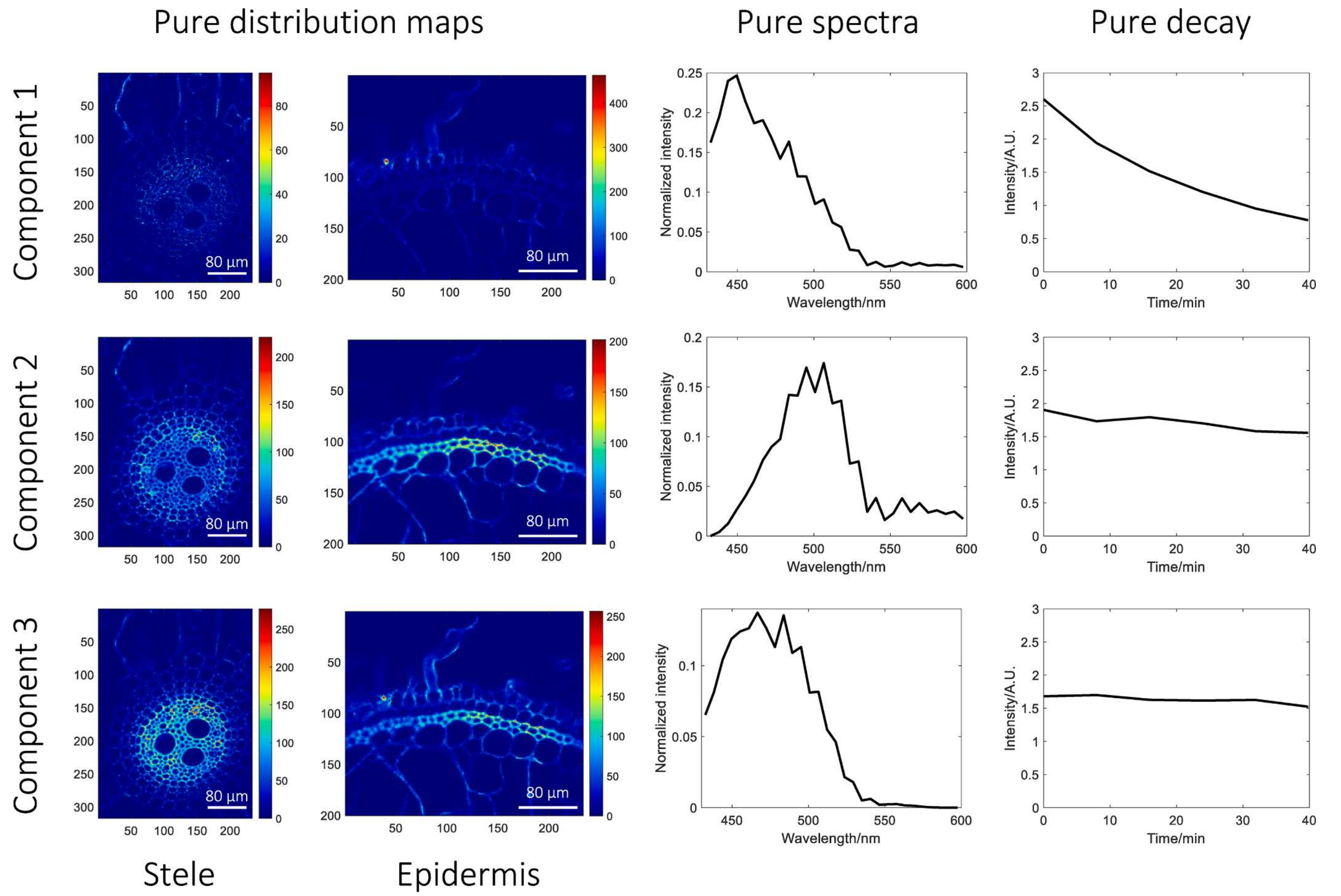


Fig. 4. MCR results obtained from the 3D photobleaching data. Plots from left to right: distribution maps (stele and epidermis ROIs), pure fluorescence emission spectra and pure photobleaching decay. Note that time 0 in the pure decay plots means the final time of acquisition of the first image acquired in the photobleaching experiment, i.e., 8 min.

Using MCR-ALS, five components were detected (Fig. 5), which explained 96.6% of the variance. In order to recover the distribution maps at the initial high resolution, an extra least squares step was performed using the optimal spectral profiles and the image data without binning.

Component 1 is related to an accumulation in the stele. The excitation maximum is around 480 nm, while the emission maximum is around 520 nm. There is no photobleaching phenomenon associated with this component. Component 2 is located in the endodermis and the sclerenchyma layer of the exodermis. This component has a similar excitation maximum than component 1, around 480 nm. However, the pure emission spectrum has a maximum around 530 nm. The photobleaching effect is significant for this component, with a loss of 55% of its signal in 24 min. Component 3 is specific of the sclerenchyma layer of the exodermis. The excitation maximum is around 530 nm and its emission around 570 nm. This component is affected by photobleaching as well, losing 11% of its intensity after 24 min of imaging. Component 4 appears on the cortex, but it has significant signal in the rest of the root, being a general component across the tissue. It has two different emission regions. The first one is located at an excitation of 500 nm and an emission of 540 nm. The second region is located at an excitation of 570 nm and an emission 620 nm, being a yellow–red fluorescence, in contraposition of the rest of components, located in the blue-green emission region. Finally, component 5 is the most shifted to the blue emission. Its excitation is 470 nm and the emission around 500 nm, and it is located in the phloem vessels and the sclerenchyma layer of the exodermis. This component seems to be almost unaffected by photobleaching, having a loss of signal around 4%.

The study of photobleaching in data sets formed by 3D and 4D fluorescence images using MCR-ALS has revealed that the photobleaching phenomenon is fluorophore-dependent and it is mandatory to perform an unmixing task of the image to properly characterize this behavior for the individual components of the sample. In the context of the rice root study, the results have shown that certain fluorophores present in the root tissue are sensitive to laser exposure at certain wavelengths, and their decay can significantly impact the quality of the measurement. Therefore, in order to accurately characterize the natural fluorophores present in the root tissue, it is important to minimize photobleaching effects. To achieve this, a strategy has been developed to reduce the number of excitation and emission channels to decrease the exposure time, while simultaneously increasing the bandwidth of the detector to increase the signal-to-noise ratio. This approach has allowed the acquisition of complete 4D images with a high signal-to-noise ratio in just 10 min, effectively minimizing the impact of photobleaching. It is clear that the effect of photobleaching must be considered on a per-component basis, as different fluorophores react differently to laser exposure.

4.2. Characterization study

Once the photobleaching phenomenon was confirmed in some components, the instrumental settings for the characterization of fluorescence components in root sections were tuned according to Section 2.2. The 4D hyperspectral image was unfolded as in Fig. 2B. Trilinearity constraint was applied to force the fluorescence emission profiles to have the same shape across the excitations. Using MCR-ALS, eight components were detected (Fig. 6), which explained 99.9% of the variance.

Fig. 7 displays the distribution maps of most of the biological components in false color, to highlight the differences among them. Component 1 (in red in Fig. 7A and B) appears on the pith of the root where the tissue is highly lignified, being more intense in the xylems and early xylems. It has an excitation maximum around 405 nm, while the emission maximum is around 480 nm. This compound could be attributed to lignin [25], specifically that which is related with lignified cells conforming the pith. Component 2 (in red in Fig. 7C) is related to the endodermis, but it has significant signal in the rest of the root, being a component located across all the tissues. Its excitation maximum is

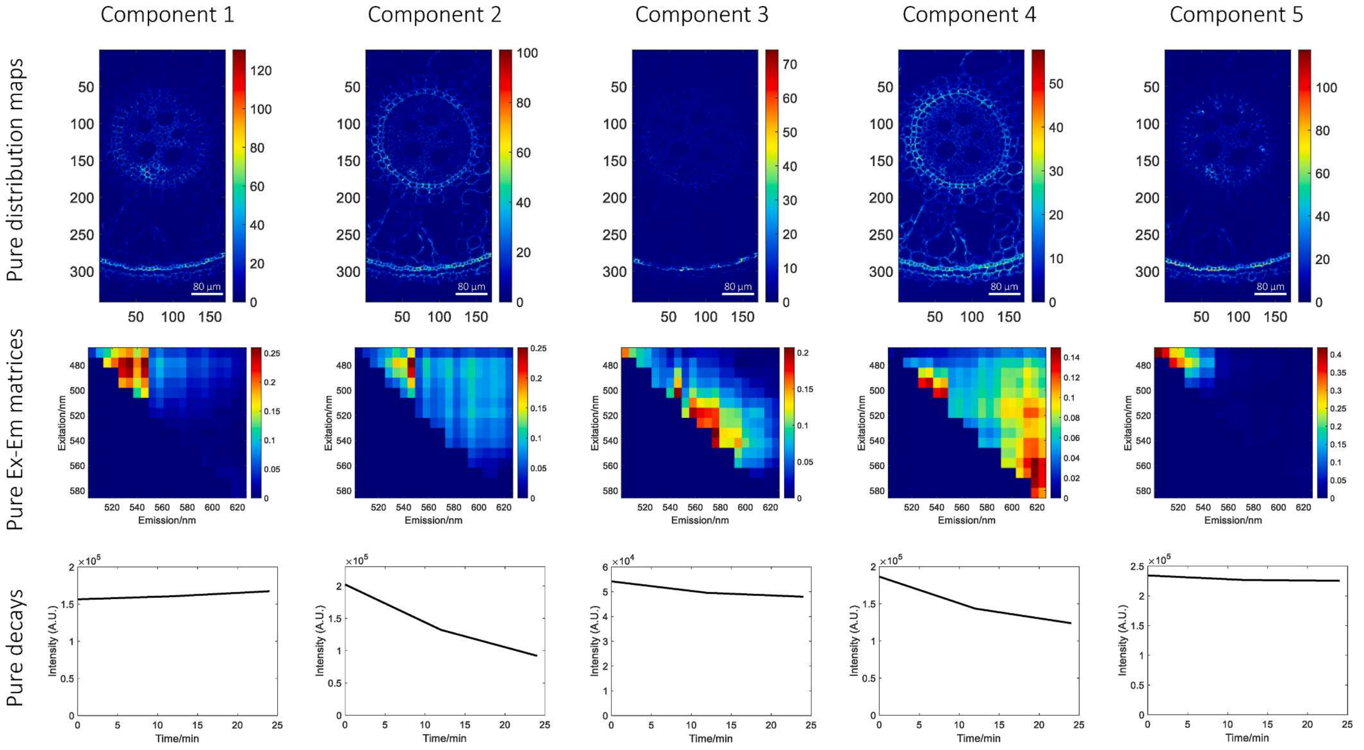


Fig. 5. MCR results obtained from the 4D photobleaching data. Plots from top to bottom: distribution maps, pure excitation-emission matrices and pure photobleaching decay. Note that time 0 in the pure decay plots means the final time of acquisition of the first image acquired in the photobleaching experiment, i.e., 12 min.

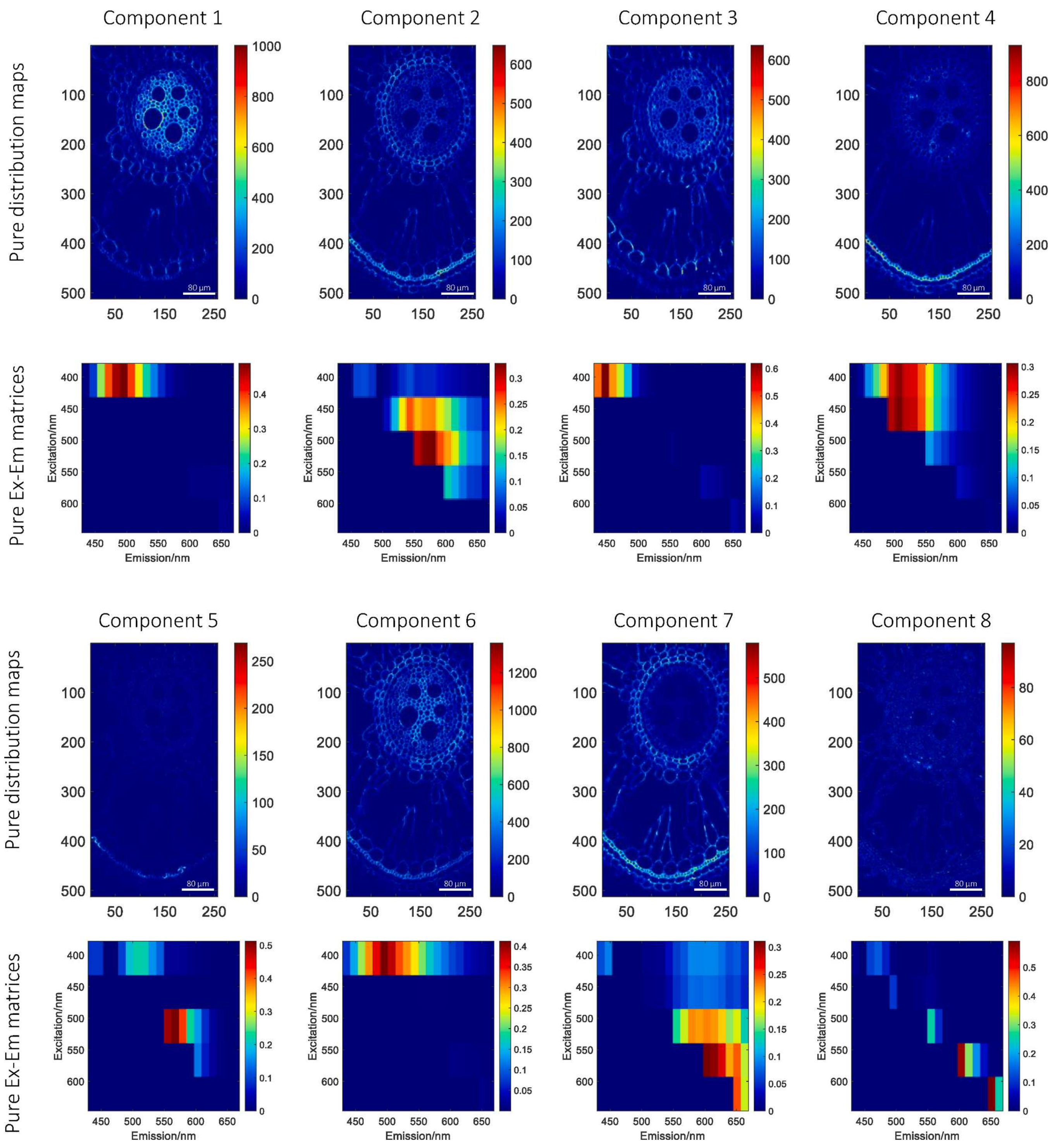


Fig. 6. MCR results for the characterization study performed using the 4D image. Distribution maps and pure EEM landscapes of the resolved components.

around 520 nm, while the emission is around 570 nm. Further investigation is needed to correctly characterize this component. Component 3 (in green in Fig. 7A and B) has an excitation of 405 nm and an emission around 440 nm, and it is located in the pericycle. It can be observed specifically in the phloem companion cells. To the knowledge of the authors, the fluorescence of the phloem companion cells has not been reported yet in rice roots. Component 4 (in blue in Fig. 7) is present in the sclerenchyma layer of the exodermis, but also in the xylem-pole pericycle. It can be observed as well where the Casparian strip should be located. This is a similar distribution found by Vishal, B. [26], which may indicate the presence of suberin. It has an excitation around 405–470 nm, while the emission is around 500 and 550 nm. Component 5 is specific of the sclerenchyma layer of the exodermis, having an excitation around 520 and an emission around 560 nm. Component 6 is located in all the root tissues, being characterized probably as a type of lignin not specific of any particular tissue. The excitation maximum is 405 while the emission is around 500 nm. Component 7 (in green in Fig. 7C) appears in the endodermis, the exodermis and the epidermis. In addition, small vesicles inside some vessels are specifically related to this component. A reasonable hypothesis is that these vesicles can be

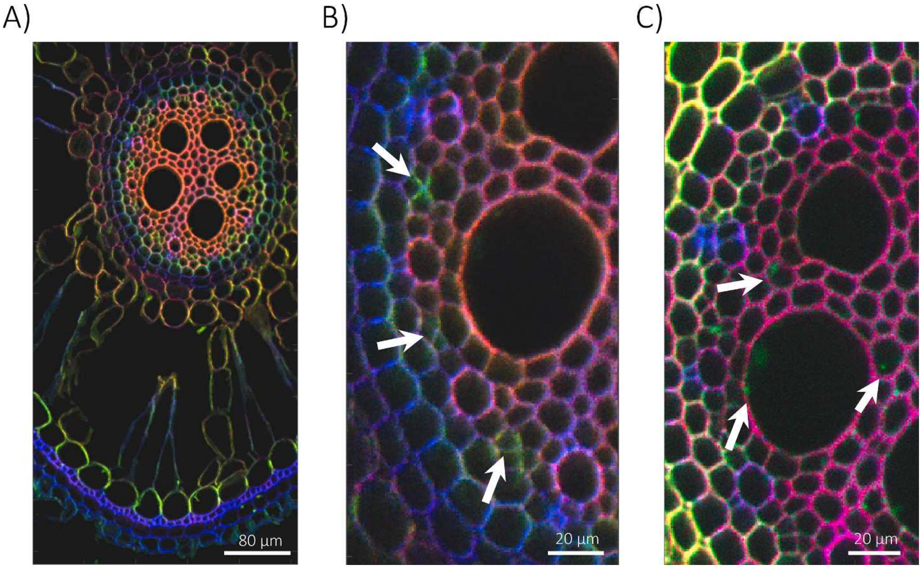


Fig. 7. Overlap of normalized distribution and color-saturated maps from the MCR-ALS analysis shown in Fig. 6 in false color. A) Red, green and blue are components 1, 3 and 4, respectively. B) Detailed zoom of the pith. C) Red, green and blue are components 2, 7 and 4, respectively.

related to silica bodies [27]. Its excitation is around 570 nm and its emission around 620 nm. Finally, the Component 8 is an artifact due to Rayleigh scattering.

Thus, thanks to the application of MCR-ALS to 3D and 4D images and to the previous study of the photobleaching phenomenon, the optimal image acquisition parameters to minimize photobleaching and improve the overall quality of data obtained from fluorescence imaging of root tissue could be obtained. Improving the signal-to-noise ratio (SNR) of fluorescence data, MCR-ALS could improve the detection and unmixing of natural fluorescence compounds present in root tissue, as it is reflected in Table 1, i.e., the number of components detected and the explained variance by the MCR model increased significantly even if the characterization experiment took only 10 min. In this context, the use of MCR-ALS also enabled to identify previously undetected compounds, which can provide valuable insights into the biology and physiology knowledge of plants.

5. Conclusions

The photobleaching phenomenon needs to be adequately described to guide a proper fluorescence image acquisition that allows detecting and characterizing all fluorophores present in samples while minimizing sample photodamage.

To do so, sets of 3D or 4D fluorescence images need to be acquired over time and be analyzed simultaneously. Characterization of the photobleaching phenomenon of a sample requires considering that this phenomenon is fluorophore-specific and some compounds may show a high signal decay while others may be almost invariant along time. To assist in this individual characterization, the unmixing methodology MCR-ALS is particularly suitable. Owing to the flexibility in the use of model constraints, pure fluorophore specificities like the invariance of the shape of maps and of pure excitation-emission spectra during photobleaching can be appropriately considered. To do so, trilinearity constraints applied in the maps and/or spectral directions and inclusion of model constraint variants that can handle the presence of systematic patterns of absent values in 2D excitation-emission landscapes is exploited. Thus, photobleaching in 3D and 4D images can be adequately described with dedicated trilinear and quadrilinear models, respectively, always providing concentration and spectral profiles and, most important, photobleaching decay curves, for every resolved fluorophore. Additionally, the multilinear nature of the models applied ensures unique solutions.

Once the photobleaching phenomenon is characterized, strategies like selecting few excitation channels to build sufficiently informative 4D fluorescence images with a limited sample exposure time ensure the characterization of all fluorescent compounds despite their different sensitivity to photobleaching.

The presented approach has been shown to be particularly suitable for challenging biological samples, with a high number of spectrally and spatially overlapped fluorescent compounds with very different photobleaching behavior, and can be extended to any kind of samples containing natural fluorescent compounds or fluorophores used for staining purposes.

CRediT authorship contribution statement

Adrián Gómez-Sánchez: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing - original draft, Writing - review & editing. **Iker Alburquerque Alvarez:** Formal analysis, Investigation, Data curation, Visualization, Writing - review & editing. **Pablo Loza-Alvarez:** Resources, Funding acquisition, Writing - review & editing. **Cyril Ruckebusch:** Resources, Funding acquisition, Writing - review & editing. **Anna de Juan:** Conceptualization, Methodology, Formal analysis, Resources, Supervision, Project administration, Funding acquisition, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

Acknowledgements

A. G.-S. and A.J. acknowledge financial support from the Spanish government Project PID 2019-1071586B-IOO and the Catalan government (2021 SGR 00449). A. G.-S. acknowledges scholarships from the MOBLILEX U Lille program and from the Santander bank.

PL-A acknowledge the Spanish Ministerio de Economía y

Competitividad (MINECO) through the "Severo Ochoa" program for Centres of Excellence in R&D (CEX2019-000910-S [MCIN/AEI/10.13039/501100011033]), Fundació Privada Cellex, Fundació Mir-Puig. Generalitat de Catalunya through CERCA program; Excellent recognized group 2021 SGR 01456 by the Departament de Recerca i Universitats de la Generalitat de Catalunya"and Laserlab-Europe EU-H2020 GA no. 871124.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.microc.2023.108899.

References

[1] B. Park, R. Lu (Eds.), Hyperspectral imaging technology in food and agriculture, Springer, Berlin, Germany, 2015.

[2] A. Gowen, C. Odonnell, P. Cullen, G. Downey, J. Frias, Hyperspectral imaging–an emerging process analytical tool for food quality and safety control, Trends Food Sci. Technol. 18 (12) (2007) 590–598.

[3] G. Lu, B. Fei, Medical hyperspectral imaging: a review, J. Biomed. Opt. 19 (1) (2014) 010901.

[4] M.A. Calin, S.V. Parasca, D. Savastru, D. Manea, Hyperspectral imaging in the medical field: Present and future, Appl. Spectrosc. Rev. 49 (6) (2014) 435–447.

[5] B. Fei, Hyperspectral imaging in medical applications. En Data Handling in Science and Technology, Elsevier, 2019, pp. 523–565.

[6] J.M. Amigo, J. Cruz, M. Bautista, S. Maspoch, J. Coello, M. Blanco, Study of pharmaceutical samples by NIR chemical-image and multivariate analysis, TrAC Trends Anal. Chem. 27 (8) (2008) 696–713.

[7] C. Gendrin, Y. Roggo, C. Collet, Pharmaceutical applications of vibrational chemical imaging and chemometrics: a review, J. Pharm. Biomed. Anal. 48 (3) (2008) 533–553.

[8] H.C. Goicoechea, A.C. Olivieri, R. Tauler, Application of the correlation constrained multivariate curve resolution alternating least-squares method for analyte quantitation in the presence of unexpected interferences using first-order instrumental data, Analyst 135 (3) (2010) 636–642.

[9] V. Olmos, L. Benítez, M. Marro, P. Loza-Alvarez, B. Piña, R. Tauler, A. de Juan, Relevant aspects of unmixing/resolution analysis for the interpretation of biological vibrational hyperspectral images, TrAC Trends Anal. Chem. 94 (2017) 130–140.

[10] K. Fackler, L.G. Thygesen, Microspectroscopy as applied to the study of wood molecular structure, Wood Sci. Technol. 47 (1) (2013) 203–222.

[11] N. Gierlinger, M. Schwanninger, Chemical imaging of poplar wood cell walls by confocal Raman microscopy, Plant Physiol. 140 (4) (2006) 1246–1254.

[12] M.J. Rust, M. Bates, X. Zhuang, Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM), Nat. Methods 3 (10) (2006) 793–796.

[13] S.W. Hell, J. Wichmann, Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy, Opt. Lett. 19 (11) (1994) 780–782.

[14] S. Hugelier, J.J. de Rooi, R. Bernex, S. Duwé, O. Devos, M. Sliwa, P. Dedecker, P.H.C. Eilers, C. Ruckebusch, Sparse deconvolution of high-density super-resolution images, Sci. Rep. 6 (1) (2016).

[15] A. Gómez-Sánchez, M. Marro, M. Marsal, P. Loza-Alvarez, A. de Juan, 3D and 4D image fusion: coping with differences in spectroscopic modes among hyperspectral images, Anal. Chem. 92 (14) (2020) 9591–9602.

[16] R.A. Hoebe, C.H. Van Oven, T.W.J. Gadella, P.B. Dhonukshe, C.J.F. Van Noorden, E.M.M. Manders, Controlled light-exposure microscopy reduces photobleaching and phototoxicity in fluorescence live-cell imaging, Nat. Biotechnol. 25 (2) (2007) 249–253.

[17] A. De Juan, R. Tauler, Multivariate Curve Resolution: 50 years addressing the mixture analysis problem–A review, Anal. Chim. Acta 1145 (2021) 59–78.

[18] J. Rebouillat, A. Dievart, J.L. Verdeil, J. Escoute, G. Giese, J.C. Breitler, P. Gantet, S. Espeout, E. Guiderdoni, C. Périn, Molecular genetics of rice root development, Rice 2 (1) (2009) 15–34.

[19] Anna De Juan, Multivariate curve resolution for hyperspectral image analysis, in: En Data Handling in Science and Technology, Elsevier, 2019, pp. 115–150.

[20] S. Hugelier, O. Devos, C. Ruckebusch, On the implementation of spatial constraints in multivariate curve resolution alternating least squares for hyperspectral image analysis, J. Chemom. 29 (10) (2015) 557–561.

[21] S. Hugelier, et al., Application of a sparseness constraint in multivariate curve resolution–alternating least squares, Anal. Chim. Acta 1000 (2018) 100–108.

[22] A. de Juan, M. Maeder, T. Hancewicz, R. Tauler, Use of local rank-based spatial information for resolution of spectroscopic images, J. Chemometrics: J. Chemometrics Soc. 22 (5) (2008) 291–298.

[23] A. Gómez-Sánchez, I. Alburquerque, P. Loza-Álvarez, C. Ruckebusch, A. de Juan, The trilinear constraint adapted to solve data with strong patterns of outlying observations or missing values, Chemom. Intel. Lab. Syst. 231 (2022) 104692.

[24] W. Windig, D.A. Stephensom, Self-modeling mixture analysis of second-derivative near-infrared spectral data using the SIMPLISMA approach, Anal. Chem. 64 (22) (1992) 2735–2742.

[25] L. Donaldson, N. Williams, Imaging and spectroscopy of natural fluorophores in pine needles, Plants 7 (1) (2018) 10.

[26] B. Vishal, et al., Os TPS 8 controls yield-related traits and confers salt stress tolerance in rice by enhancing suberin deposition, New Phytol. 221 (3) (2019) 1369–1386.

[27] N. Zexer, R. Elbaum, A. Lux, Unique lignin modifications pattern the nucleation of silica in sorghum endodermis, J. Exp. Bot. 71 (21) (2020) 6818–6829.

Supplementary material

STUDY OF THE PHOTOBLEACHING PHENOMENON TO OPTIMIZE ACQUISITION OF 3D AND 4D FLUORESCENCE IMAGES. A SPECIAL SCENARIO FOR TRILINEAR AND QUADRILINEAR MODELS.

Adrián Gómez-Sánchez[1,2], Iker Alburquerque Alvarez[1], Pablo Loza-Alvarez[3], Cyril Ruckebusch[2] & Anna de Juan[1]

[1]Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028, Barcelona, Spain

[2]LASIRE - Laboratory of Advanced Spectroscopy, Interactions, Reactivity and Environment Université Lille, CNRS, UMR 8516, Cité Scientifique, Bâtiment C5, 59000, Lille, France

[3]ICFO-Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology, 08860 Castelldefels, Barcelona, Spain

3D fluorescence images

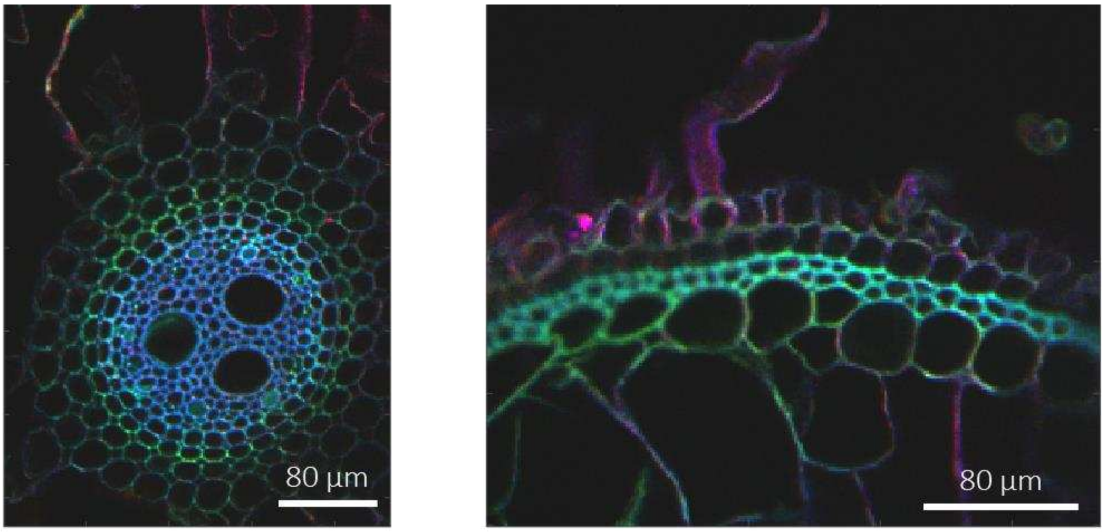


Figure S1. Overlap of normalized distribution and color-saturated maps from the MCR-ALS analysis shown in Figure 5 in false color. Red, green and blue are components 1, 2 and 3, respectively.

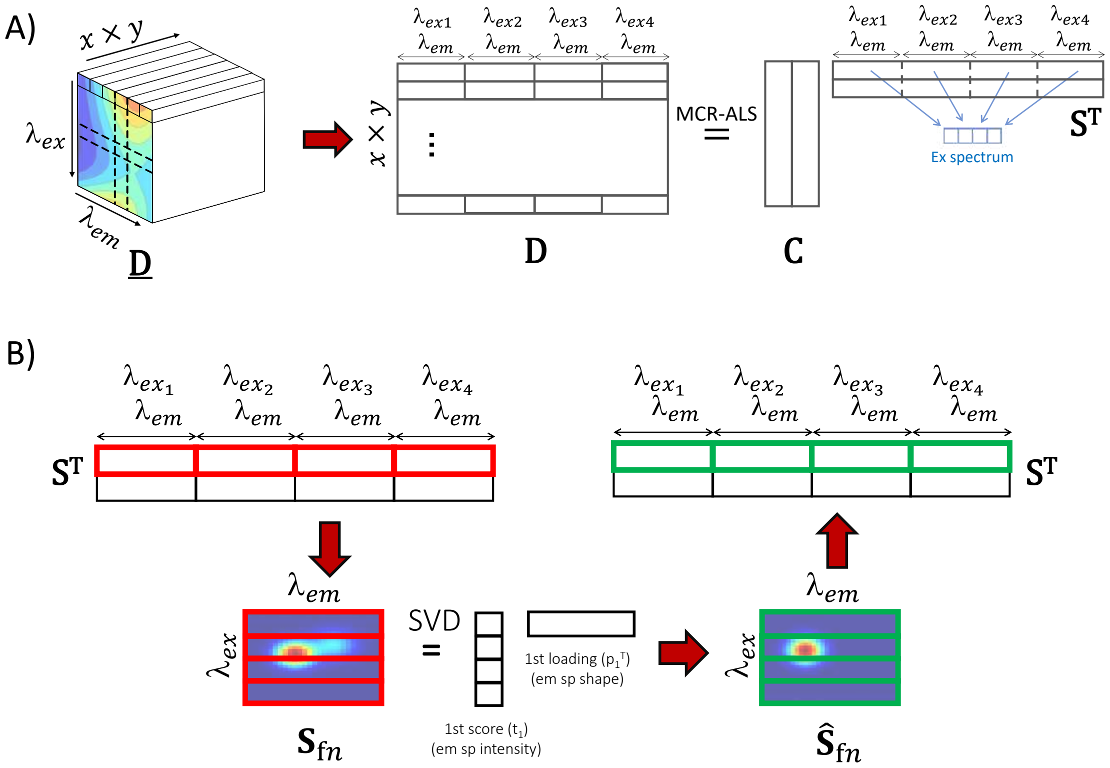
Publications I and **II** propose two different approaches to adapt the trilinear constraint in MCR-ALS for EEM data sets with strongly patterned missing values. The classical implementation of the trilinear constraint for data with full EEM landscapes is described to understand better the modifications introduced to deal with the presence of missing values.

Classical implementation of the trilinear constraint in MCR-ALS

Trilinear models can be achieved through the application of the trilinear constraint during the iterative alternating least squares optimization in MCR-ALS. The implementation of the trilinearity constraint was proposed by Romà Tauler in 1993, where it was successfully applied in the analysis of chromatographic data [Tauler and Barceló, 1993].

The scheme of Fig. 20 illustrates the steps followed in the implementation of the trilinearity constraint for a 4D EEM image. To use MCR-ALS on this data, the data cube of EEM measurements $\underline{\mathbf{D}}$ is unfolded into a data matrix **D** concatenating the emission spectra obtained at the different excitation wavelengths for each pixel (Fig. 20A). The implementation of the trilinearity constraint is based on the fact that the shape of the pure emission spectra in $\mathbf{S^T}$ for a specific component is the same for all excitation wavelengths and the only modification observed among them is in scale intensity. Thus, the trilinear constraint works forcing this invariant shape on the pure emission spectra profiles, providing in this way a trilinear model [Tauler, 1995; Tauler et al., 1998; Alier et al., 2011; Tauler, 2021].

Figure 20B shows the algorithmic steps to implement the trilinearity constraint, which is applied individually to each component of matrix $\mathbf{S^T}$. To impose the trilinear constraint, the extended profile of concatenated emission spectra for a particular component *n* (in red in Fig. 20B) is folded to form the matrix $\mathbf{S}_{fn}$, which contains the emission spectra related to the different excitation wavelengths one on top of each other. Subsequently, $\mathbf{S}_{fn}$ is decomposed by Singular Value Decomposition (SVD) to find the first principal component, which will contain the information related to the common shape that all emission spectra of the component *n* should present. Here, the loading vector $\mathbf{p}_1^T$ contains the shape of the emission spectrum, while the score vector $\mathbf{t}_1$ have the scaling weighting factors to define the variations of the emission intensity across the different excitation wavelengths. In other words, $\mathbf{t}_1$ represents the shape of the pure excitation spectrum. Afterwards, the reconstruction of $\mathbf{S}_{fn}$ is achieved using the first score and the first loading vectors, providing $\hat{\mathbf{S}}_{fn}$. In the new matrix $\hat{\mathbf{S}}_{fn}$, all emission profiles share the same shape and are only differently scaled, thanks to the rank-one principal component reconstruction. To finish the application of the trilinearity constraint, the $\hat{\mathbf{S}}_{fn}$ is unfolded and the extended profile (in green) is adopted as the pure emission spectra profile for component *n* in $\mathbf{S^T}$. Once the iterative process finishes, the pure excitation spectrum is derived by considering the area of pure fluorescence emission for each excitation. This procedure allows MCR-ALS to provide trilinear models and obtain unique solutions [Tauler, 2021].



A)
$x \times y$
λ_{ex}
λ_{em}
D
λ_{ex1} λ_{em}
λ_{ex2} λ_{em}
λ_{ex3} λ_{em}
λ_{ex4} λ_{em}
$x \times y$
D
MCR-ALS
=
C
λ_{ex1} λ_{em}
λ_{ex2} λ_{em}
λ_{ex3} λ_{em}
λ_{ex4} λ_{em}
Ex spectrum
S^T
B)
λ_{ex_1} λ_{em}
λ_{ex_2} λ_{em}
λ_{ex_3} λ_{em}
λ_{ex_4} λ_{em}
S^T
λ_{em}
λ_{ex}
S_{fn}
SVD
=
1st loading (p_1^T)
(em sp shape)
1st score (t_1)
(em sp intensity)
λ_{em}
λ_{ex}
$\hat{S}_{fn}$
λ_{ex_1} λ_{em}
λ_{ex_2} λ_{em}
λ_{ex_3} λ_{em}
λ_{ex_4} λ_{em}
S^T

Figure 20. A) Structure of EEM data. The data cube is unfolded by concatenating the emission spectra at different excitations, resulting in matrix **D**. This matrix is decomposed into matrix **C**, associated with concentration profiles, and **S**^{T}, linked to the unfolded pure excitation-emission signatures. The third mode is achieved by integrating the emission signal for each excitation. B) During iterations, the trilinearity constraint is applied to each pure spectral profile, forcing **S** to exhibit the same emission shape across excitations through SVD.

However, the trilinear constraint cannot be applied in presence of Raman and Rayleigh scattering, since these phenomena introduce signals that break the inherent trilinear behavior of excitation-emission matrices. Therefore, it becomes mandatory to suppress these interferences before the application of trilinear models. As mentioned in the introduction of this section, several preprocessing approaches exist to address Raman and Rayleigh scatterings in EEM data, but the one that avoids the introduction of artifacts and preserves all the available trilinear fluorescence signal is setting the Rayleigh and Raman as missing entries. The same option can be adopted for EEM measurements where emission and excitation wavelength ranges overlap, since emission values below excitation wavelengths can also be defined as missing entries. In this situation, the traditional implementation of the trilinearity constraint cannot be applied due to the presence of missing values and different strategies need to be considered. For this reason, one of the main objectives of this thesis has been the adaptation of the trilinearity constraint to handle missing data in MCR-ALS.

Sequential application of the trilinearity constraint for data sets with missing values.

As seen in Fig. 20B, the conventional implementation of the trilinearity constraint requires that the matrix $\mathbf{S}_{fn}$ does not contain missing values, i.e., all emission spectra should cover the same wavelength range. This does not happen in EEM structures such as the one shown in Fig. 21A, where the emission spectra get shorter as the related excitation wavelength increases. If $\mathbf{S}_{fn}$ contains missing values, as shown in Fig. 21B, SVD cannot be straightforwardly applied.

Publication I proposes a solution based on the sequential application of SVD to complete submatrices of $\mathbf{S}_{fn}$ instead of doing it on the whole matrix with missing values.

The starting step of the MCR-ALS analysis also consists of unfolding the data cube $\underline{\mathbf{D}}$ by concatenating the emission spectra acquired at the different excitation wavelengths for each pixel (Fig. 21A). Note that now, the concatenated emission spectra do not cover the same spectral range. Then, MCR-ALS is applied and the profiles of the $\mathbf{S}^{\mathbf{T}}$ matrix subject to the trilinear constraint. As in the conventional implementation of the trilinearity constraint, each pure profile of $\mathbf{S}^{\mathbf{T}}$ (in red) is folded to recover the original structure of the excitation-emission landscape. Here, the refolded matrix $\mathbf{S}_{fn}$ contains missing values, set as not-a-number (NaN) (Fig. 21B). The adaptation of the trilinear constraint consists of applying the SVD decomposition to a certain number of $\mathbf{S}_{fn}$ submatrices until all elements in the original $\mathbf{S}_{fn}$ matrix are covered. However, this operation takes place in two steps. Step 1 is devoted to obtaining an $\mathbf{S}_{\mathbf{fn}}$ matrix with as unmixed information as possible to start step 2, oriented to obtain the common emission spectrum shape required for trilinear models.

The choice of the submatrices submitted to SVD in step 1 is not arbitrary and starts by the submatrices with a highest mixture level, estimated with ML, defined as in Eq. 18:

$$\mathrm{ML} = \frac{trace(\widetilde{\mathbf{\Sigma}})}{N} \qquad \text{Eq. 18}$$

Where $\widetilde{\mathbf{\Sigma}}$ is defined as the diagonal matrix containing the eigenvalues Σ divided by Σ_{11} and with N as the number of components.

ML can move from 1/*N* for a perfect rank one matrix (i.e. in a noiseless case, when only a pure component exists) to one, when the variance is evenly spread in all calculated components underlying this specific submatrix. The closer ML is to 1, the higher the mixture level in the analyzed submatrix.

In the example of Fig. 21B, the initial submatrix with highest ML is identified in green. This submatrix is decomposed by SVD and reconstructed using the first component. The values in $\mathbf{S}_{fn}$ that correspond to the reconstructed submatrix are replaced by the new ones. This process is repeated for the next submatrix identified (in purple). However, only the values in the $\mathbf{S}_{fn}$ matrix that were not defined by the preceding submatrix analysis are replaced during this step. Next, the same procedure is applied to the third identified submatrix, in yellow. This sequential process continues until all necessary submatrices to cover the area of $\mathbf{S}_{fn}$ considered for trilinearity are forced. Step 1 helps to 'clean' the original mixed contributions in $\mathbf{S}_{fn}$, but the emission profiles associated with every excitation step may be slightly different because they may come from different submatrix reconstructions.

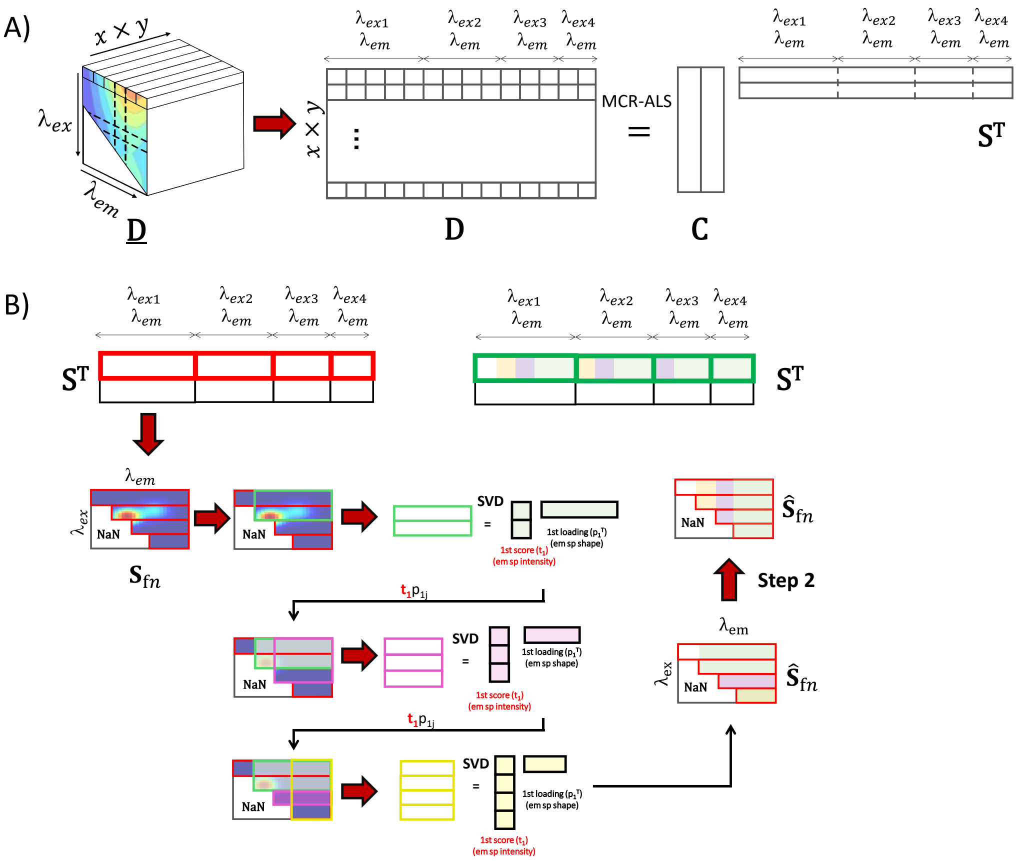


Figure 21. A) The excitation-emission dataset is unfolded by concatenating the emission spectra at different excitation for each pixel in row-wise direction. The dataset **D** is then analyzed by MCR-ALS. B) During the iterative process, the pure profiles **S** are refolded, and the trilinearity constraint is applied to the different submatrices. Then, the $\hat{\mathbf{S}}_{fn}$ is then vectorized by concatenating the emission spectra at different excitation wavelengths and replacing the corresponding pure profile.

Step 2 is similar to Step 1, but the submatrices submitted to SVD are sorted in decreasing order of covered rows. In this way, the reconstructed part by the first submatrix covers the spectral range related to the submatrix with maximum number of rows (in light green). The reproduction by the next submatrix updates the spectral range with the second highest number of rows not covered by the first submatrix (in purple) until all the $\mathbf{S_{fn}}$ matrix is reconstructed. Step 2 guarantees obtaining a common emission spectral shape of emission spectra due to the ordered design of the SVD analyses on submatrices sorted in decreasing order of sorted rows. The resulting matrix $\mathbf{\hat{S}}_{\mathrm{fn}}$ is then vectorized by concatenating the emission spectra at different excitation wavelengths and used to replace the corresponding pure profile of the $\mathbf{S}^{\mathbf{T}}$ matrix and the iterative process continues until convergence.

This new adaptation of the trilinear constraint in MCR-ALS for EEM data effectively overcomes challenges linked to trilinear data with systematic patterns of missing values and does not require the data imputation step used in conventional trilinear decomposition methods.

<u>Use of adapted Non-linear Iterative Partial Least Squares (NIPALS) to implement the trilinearity constraint for data with missing values.</u>

Despite the efficiency of the previous adaptation of the trilinear constraint for EEM data with missing values, the implementation of the constraint is complex and it is formed by two independent steps. In addition, for large data sets, the exhaustive search of the suitable submatrices for step 1 can significantly slow down the analysis.

Publication II offers an improved implementation of the trilinearity constraint for data sets with missing values (Fig. 22A), based on an adapted use of the Non-linear Partial Least Squares (NIPALS) [Wold, 1975] algorithm instead of SVD. This approach is much easier to implement, does not require two steps, is completed in a short computation time and provides comparable results to the proposed implementation in **Publication I**.

The NIPALS algorithm is, in essence, a simple alternating least squares procedure that extracts sequentially the principal components of a data set. It starts by providing an initial estimate of the first score or loading vector to initiate the alternating least squares optimization of the score and loading profile until convergence is achieved. Once the first principal component is calculated, a deflation of the matrix by subtracting the variance described by the component calculated is carried out and a new component is estimated. The component extraction-deflation sequence is continued until all components required are calculated.

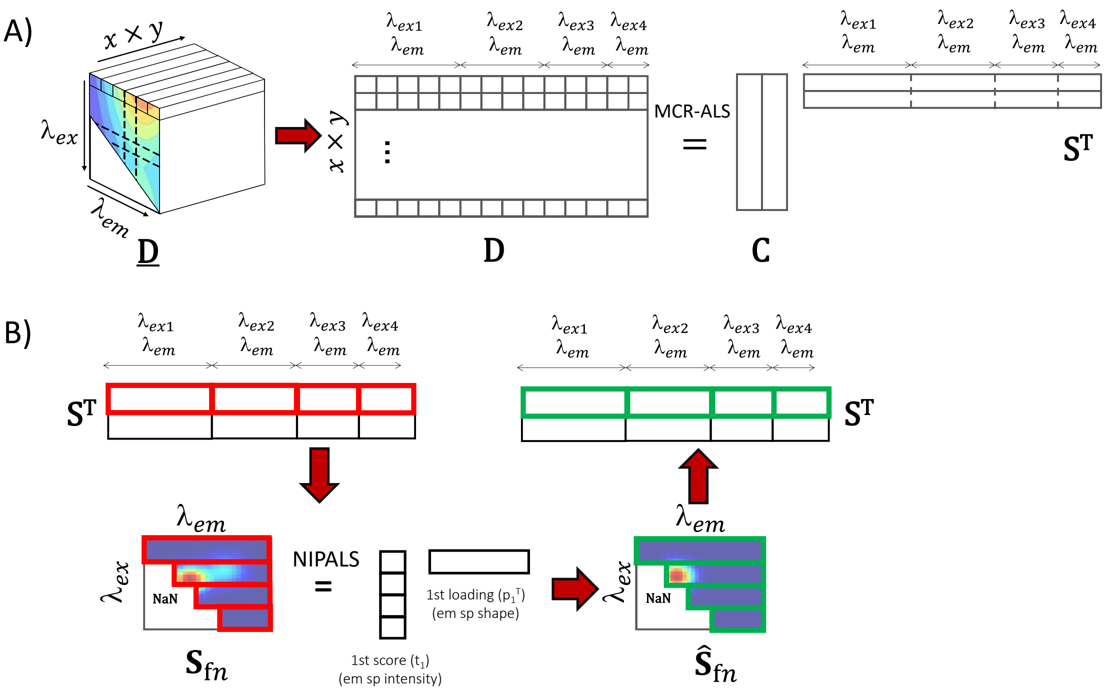


Figure 22. Implementation of the trilinear constraint based on an adapted NIPALS algorithm. A) The excitation-emission dataset is unfolded by concatenating the emission spectra at different excitation for each pixel in row-wise direction. The dataset **D** is then analyzed by MCR-ALS. B) During iterations, the trilinearity constraint is applied to each pure spectral profile, forcing **S** to exhibit the same emission shape across excitations by using an adapted version of the NIPALS algorithm.

Figure 22B shows that the NIPALS algorithm is used here to obtain the first principal component that defines the common emission profile shape required for the implementation of the trilinear constraint.

A good asset of the NIPALS algorithm is that it can be adapted to skip the missing values during the alternating least squares optimization of the score and loading profiles [Grung and Manne, 1998]. To deal with missing entries, the least squares calculations of the score and loading profile are performed row by row (Eq. 19) and column by column (Eq. 20), respectively, adapting these steps to the available information in $\mathbf{S}_{\mathrm{fn}}$, sized (I,J). The information used in these calculations is visually shown in Fig. 23.

$$\mathbf{t}(i,\mathbf{1}) = \mathbf{S}_{\mathrm{fn}}(i,:)\left(\mathbf{p}^{\mathrm{T}}\right)^{+} \qquad \text{Eq. 19}$$

$$\mathbf{p}^{\mathrm{T}}(\mathbf{1},j) = \mathbf{t}^{+}\mathbf{S}_{\mathrm{fn}}(:,j) \qquad \text{Eq. 20}$$

where i and j go from 1 to I and 1 to J, respectively. When missing values are encountered in the row $\mathbf{S}_{\mathrm{fn}}(i,:)$**,** the calculation of the related score value, $\mathbf{t}(i,1)$, according to Eq. 19 is done by using only the available entries of $\mathbf{S}_{\mathrm{fn}}(i,:)$ and the analogous values of $\mathbf{p}^{\mathrm{T}}$, as shown in Figure 23A. The same approach is applied when the calculation of every element in the loading vector, $\mathbf{p}^{\mathrm{T}}(1,j)$, is carried out, as proposed in Eq. 20 and displayed in Figure 23 B). The calculations are carried out using the available entries in the column $\mathbf{S}_{\text{fn}}(:,j)$ and the analogous values of $\mathbf{t}$.

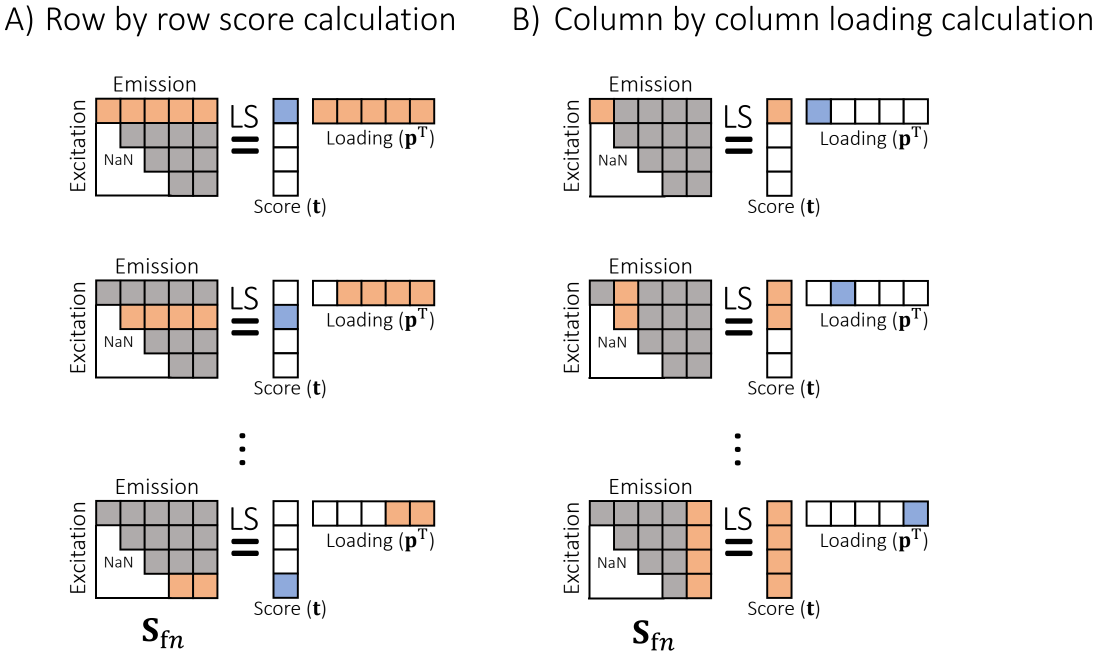

A) Row by row score calculation
B) Column by column loading calculation
Emission
Excitation
NaN
LS
=
Loading ($\mathbf{p}^{\mathrm{T}}$)
Score ($\mathbf{t}$)
⋮
$\mathbf{S}_{fn}$

Figure 23. A) Row-by-row calculation of scores by NIPALS. The loading $\mathbf{p}^{\mathrm{T}}$ and a row $\mathbf{S}_{\text{fn}}(i,:)$ (orange) are used to calculate the corresponding score value $\mathbf{t}(i,1)$ (blue). The loading is adapted to match the corresponding available entries in $\mathbf{S}_{\text{fn}}(i,:)$. B) Column-by-column calculation of loadings. The score $\mathbf{t}$ and a column $\mathbf{S}_{\text{fn}}(:,j)$ (orange) are used to calculate the corresponding loading value $\mathbf{p}^{\mathrm{T}}(1,j)$ (blue). Again, the score is adapted to match the available entries.

After imposing trilinearity through NIPALS to $\mathbf{S}_{\text{fn}}$, the resulting matrix $\hat{\mathbf{S}}_{\text{fn}}$ is unfolded and used to replace the corresponding pure profile of the $\mathbf{S}^{\mathrm{T}}$ matrix. This procedure is repeated for all components desired to be trilinear. Once the trilinearity constraint has been applied to the components, the MCR-ALS iterative process continues until convergence.

This simple approach allows MCR-ALS to deal with systematic patterns of missing data when applying the trilinear constraint without the need of any data imputation procedure. Another significant advantage of this constraint implementation is that it can be adapted to any missing value pattern encountered in the EEM landscapes or in other kinds of trilinear data.

The results obtained from the approaches presented in **Publication I** and **Publication II** demonstrate equivalence in terms of the quality of the solutions obtained, as shown in Fig. 24 for a real data set of binary mixtures of pharmaceutical compounds.

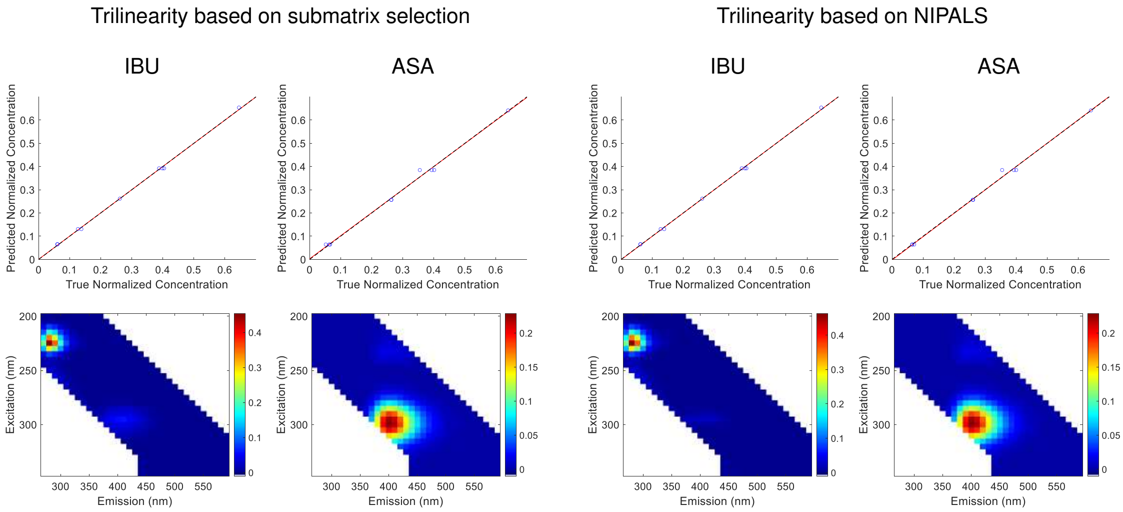


Figure 24. Comparison of one of the datasets (mixtures of ibuprofen, IBU, and acetylsalicylic acid, ASA) analyzed by, left, trilinearity based on submatrix selection (**Publication I**) and right, the NIPALS-based algorithm (**Publication II**). Both approaches provide nearly identical results.

Summarizing, **Publication I** introduces a two-step process involving sequential SVD analyses to handle missing values. While effective, the drawback of this approach is the complexity and time-consuming execution, particularly in large datasets.

In contrast, **Publication II** proposes a straightforward approach based on the use of the adapted NIPALS algorithm. This implementation is simple and adapts to any missing pattern, offers computational advantages and is easily scalable to higher-order linear models.

Photobleaching study in vegetal tissues. A practical application of the adapted trilinearity constraint for excitation-emission fluorescence images.

Publication III applies the adapted trilinearity constraint of **Publication I** to a real case, where the fluorophores of a natural vegetal tissue are studied using 3D and 4D fluorescence images. This real example is a challenging case, where the potential of trilinear and quadrilinear models can be extensively tested.

3D and 4D fluorescence images provide comprehensive spatial and chemical information of the natural fluorophores present in samples. However, 3D and 4D fluorescence can require prolonged image acquisition times that may cause a degradation of the fluorescence signals, the so-called photobleaching, and potential sample damage [Hoebe et al., 2007]. Photobleaching occurs due to reactions between the fluorophore and surrounding molecules when excited by the laser. The study of the photobleaching phenomenon during sample measurements is particularly crucial for researchers, since it allows them to establish optimal instrumental parameters to avoid sample damage and to perform the correct characterization of the fluorophores.

This research focuses on developing a methodology to mitigate the impact of photobleaching by analyzing time-series of 3D and 4D fluorescence images through MCR-ALS. The results obtained will help to understand the photobleaching behavior of the fluorophores for a posterior instrumental parameter tuning. The study of the evolution of the pure profiles over time provides useful information related to how sensitive each specific fluorophore is to photobleaching. Therefore, this information can be used to choose if the instrumental parameters must be modified to avoid photodamage.

According to the nature of the fluorescence measurement, this study assumes that all fluorophores preserve the shape of their pure emission spectra (Fig. 25A) or pure excitation-emission response (Fig. 25B) when photobleaching occurs. Observing the fluorophore-specific photobleaching curves, strategic approaches, such as decreasing the acquisition time by selecting a few excitation channels or increasing the detected bandwidth to improve the signal-to-noise ratio, will contribute to improve the characterization of all sample fluorophores.

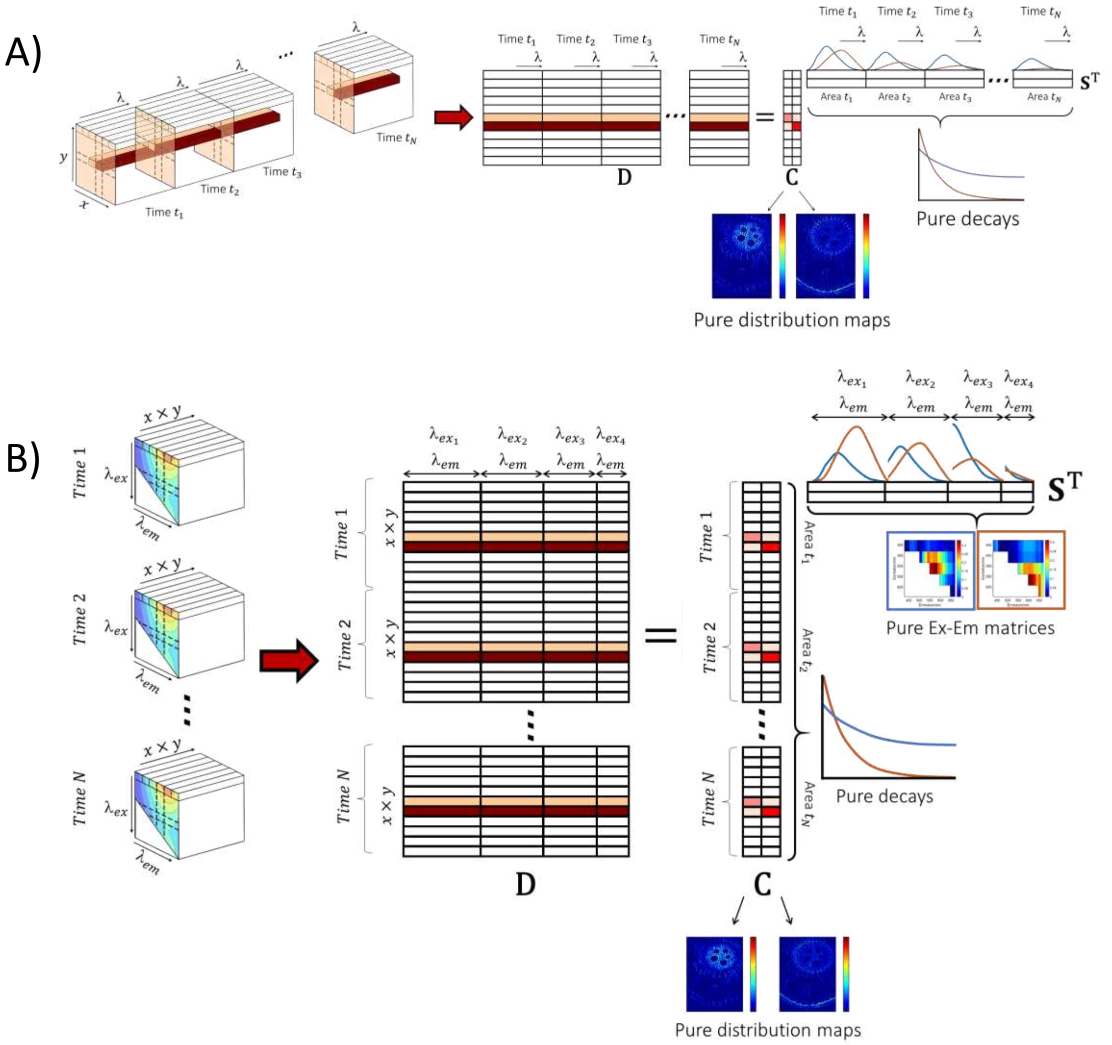


Figure 25. A) Multiset configuration and related MCR model for a photobleaching experiment based on: A) time series of 3D fluorescence images and B) time series of 4D fluorescence images. For time series of 3D fluorescence images, the traditional implementation of the trilinearity constraint has been applied. For the time series of 4D fluorescence images, the adapted trilinearity constraint has been applied to the pure matrix **S**, while the traditional trilinear constraint has been applied to **C**, leading to a quadrilinear model, where distribution maps, excitation, emission and decay modes can be studied.

Study of 3D images over time

The multiset formed by time series of 3D fluorescence images was built by the concatenation in row-wise direction of 3D fluorescence images acquired consecutively on the same sample over time (Fig. 25A). The analysis of this multiset by MCR-ALS was performed under trilinearity constraint on the pure matrix **S**, i.e., keeping constant the shape of the emission pure profiles over the time. MCR-ALS provided a set of pure distribution maps, the pure emission spectrum and the pure decay of each component. The pure photobleaching curves were obtaining by integration of each pure emission spectra over time per each component. The results are shown in Fig. 26.

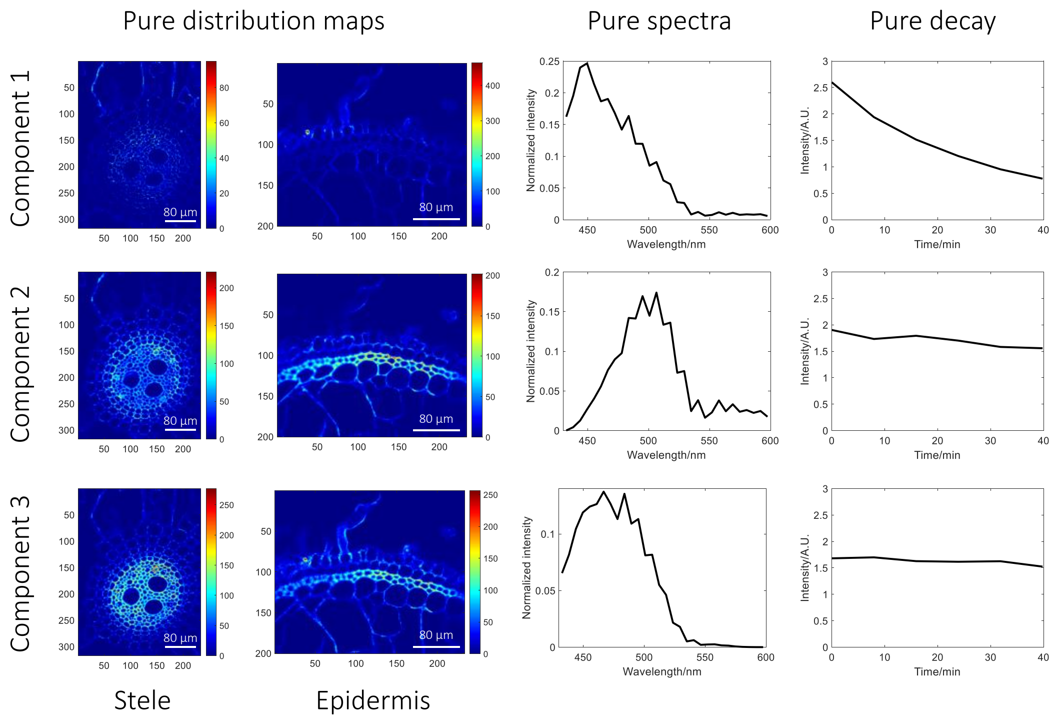


Figure 26. MCR results obtained from the 3D photobleaching data. Plots from left to right: distribution maps (stele and epidermis ROIs), pure fluorescence emission spectra and pure photobleaching decay. Note that time 0 in the pure decay plots means the final time of acquisition of the first image acquired in the photobleaching experiment, i.e., 8 min.

The first MCR-ALS component, associated with small vesicles and the cortex, is significantly affected by photobleaching, with a 41% signal loss after 40 minutes of image acquisition. The second component, found in the endodermis and the sclerenchyma layer, is less affected, with only a 6% signal loss. The third component, specific to the inner part of the stele, is even less affected, with a 3% signal loss after 40 minutes of imaging. It is important to note that the fluorophore-specific photobleaching behavior can only be obtained previous unmixing analysis of the fluorescence images.

Study of 4D images over time

Three 4D fluorescence images were acquired consecutively. Rayleigh dispersion entries were suppressed and emission values below the excitation wavelength, providing a systematic pattern of missing values.

The images were concatenated into a single multiset as in Fig. 25B. Then, the unfolded images acquired at different times were concatenated in a column-wise direction. The trilinearity constraint was applied to force the concentration profile shape to be the same for each component over time (**C**). This responds to the fact that the sample does not move during the time-series acquisition and all fluorophores remain located in the same position, showing the same distribution map with different intensity. On the other hand, the pure emission fluorescence profiles were forced to have the same shape across the excitation wavelengths (**S**). The double application of the trilinear constraint in the augmented concentration and spectral mode results into a quadrilinear model. Note that the trilinearity constraint applied to the excitation-emission unfolded spectra follows the implementation proposed in **Publication I**, since the EEM landscapes collected per pixel show strongly patterned missing values. The results of the analysis of 4D time-series using a quadrilinear model are shown in Fig. 27.

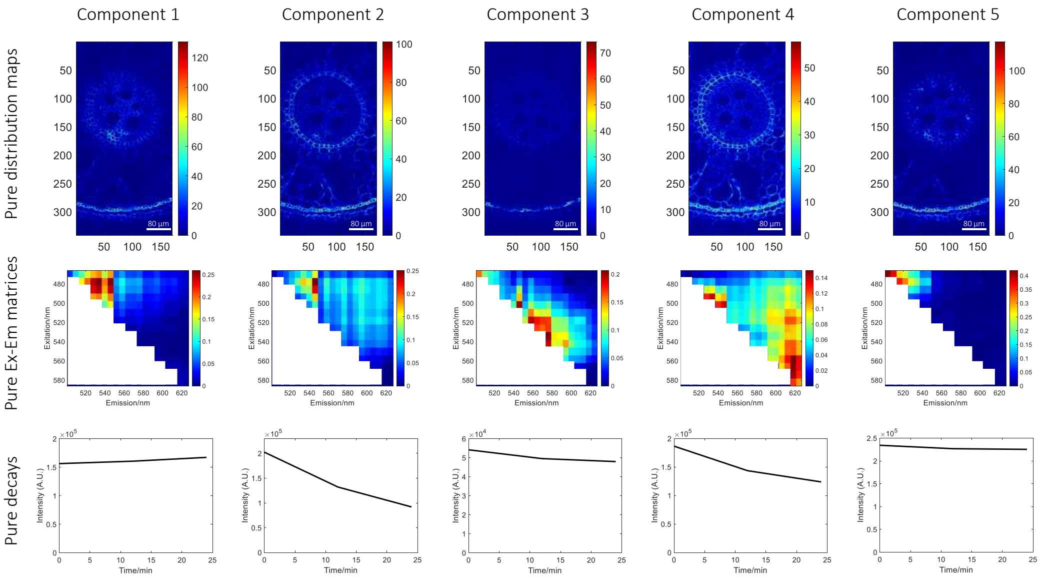


Figure 27. MCR results obtained from the 4D photobleaching data. Plots from top to bottom: distribution maps, pure excitation-emission matrices and pure photobleaching decay. Note that time 0 in the pure decay plots means the final time of acquisition of the first image acquired in the photobleaching experiment, i.e., 12 min.

According to the results, Component 1 is unaffected by photobleaching. Component 2 experiences significant photobleaching, losing 55% of its signal in 24 minutes. Component 3 also undergoes photobleaching, with an 11% loss in intensity after 24 minutes. Component 4 is significantly affected by photobleaching, losing approximately a 30% of the signal. In contraposition, Component 5 is minimally affected by photobleaching, with only a 4% signal loss.

The investigation of photobleaching in 3D and 4D fluorescence image datasets using MCR-ALS indicated the fluorophore-dependent nature of this phenomenon. In the case of the rice root study, certain fluorophores are sensitive to specific laser wavelengths and their decay significantly influences the overall quality of the measurement.

To characterize natural fluorophores accurately, minimizing photobleaching is crucial. A strategy devised from the results obtained was reducing the number of excitation and emission channels to decrease exposure time while widening the detector bandwidth to enhance signal-to-noise ratio. This approach facilitated the acquisition of complete 4D images with minimal photobleaching impact in just 10 minutes, emphasizing the need to consider photobleaching effects on a per-component basis due to varying fluorophore responses to laser exposure.

4D images for fluorophore characterization

Once the photobleaching phenomenon was confirmed for some components, the instrumental settings for the characterization of fluorescence components in root sections were tuned reducing the wavelength channels while keeping the same spectral range and increasing the detector bandwidth. The 4D hyperspectral image was unfolded as in Fig. 21A. The adapted trilinearity constraint of **Publication I** was applied to the pure matrix **S** (Fig. 21B). Using MCR-ALS, eight components were detected (Fig. 28).

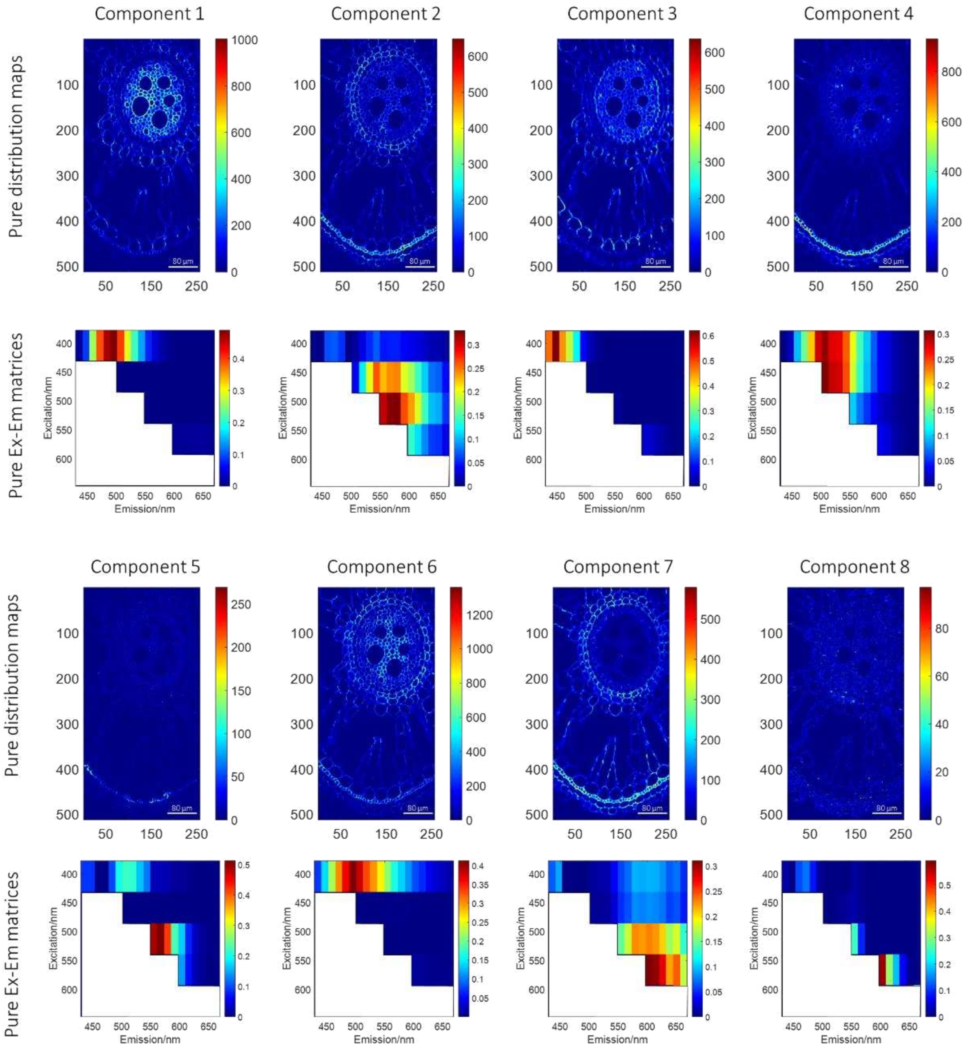


Figure 28. MCR results for the characterization study performed using the 4D image. The distribution maps and pure EEM landscapes of the resolved components are provided. Note that white regions on pure EEM landscapes contain NaN.

The components identified in Fig. 28 showed biological relevance and were assigned to specific regions of the root [Donaldson and Williams, 2018; Vishal et al., 2019; Zexer et al., 2020]. Component 1 was associated with lignified cells in the central pith, possibly related to lignin. Component 2 was found in the endodermis, but is also spread across all tissues, requiring further characterization. Component 3 is located in the pericycle, specifically in the phloem companion cells. Component 4 is present in the sclerenchyma layer, potentially indicating the presence of suberin. Component 5 is specific to the sclerenchyma layer. Component 6 is a type of lignin found throughout the root tissues. Component 7 appears in the endodermis, exodermis, and epidermis, possibly related to silica bodies. Finally, Component 8 is identified as an artifact due to a residual signal from Rayleigh scattering.

The application of MCR-ALS to time series of 3D and 4D fluorescence images was useful for the detection of fluorophore-specific photobleaching phenomena. Since photobleaching was detected in some components, an optimization of the instrumental parameters was required. This optimization minimized photobleaching, enhancing the quality of the data. By improving the signal-to-noise ratio, MCR-ALS enhances the detection and unmixing power of natural fluorescence compounds in root tissue, as evidenced in Fig. 28, even with a short 10-minute acquisition experiment. Furthermore, MCR-ALS facilitated the identification of previously undetected compounds, offering valuable insights into plant biology and physiology.

As a final conclusion, the outcomes of this study show the effectiveness of the adapted trilinearity constraint to adequately model the information on fluorescence images and provide a more reliable characterization of fluorophores in complex biological samples.

3.2 Improving the analysis of Fluorescence Lifetime Imaging Microscopy data

Time-resolved fluorescence spectroscopy (TRFS) is a well-known fluorescence spectroscopic technique that provides valuable information about the fluorophores present on a sample via the measurement of fluorescence decay curves [Lakowicz, 2006].

However, as stated in Chapter 2, analyzing fluorescence decay curves is challenging due to the impact of the convolution of the instrument response function (IRF) with the true TRFS fluorescence signal [Luchowski et al., 2009]. To solve this problem, deconvolution is commonly used, but it requires the knowledge or estimation of the IRF.

In addition, analyzing TRFS multiexponential decay curves to obtain the underlying monoexponential signals of the pure fluorophores is particularly challenging due to high correlation of these signal and the lack of selectivity in the pure monoexponential profiles. For this reason the **Publications IV** and **V** in this subsection are devoted to facilitate and improve the quality of the analysis of TRFS data.

Publication IV is a new algorithm aimed to derive the IRF function from the decay curves measured and without the need of any additional experimental measurement. This approach is based on the mathematical properties of the exponential decays and the convolution process.

Publication V proposes the kernelizing approach, based on the convolution of different kernels to the TRFS data, as an approach to generate trilinear data from the bilinear data measured with TRFS. The trilinear data formed improve significantly the unmixing of the multiexponential decays into their pure monoexponential contributions.

Publication IV. Blind Instrument Response Function Identification (BIRFI) from Fluorescence Decays.
Authors: A. Gómez-Sánchez, O. Devos, R. Vitale, M. Sliwa, D. Sakhapo, J. Enderlein, A. de Juan, C. Ruckebusch.
Biophysical Reports (2024) (submitted).

Publication V. Kernelizing: A way to increase accuracy in trilinear decomposition analysis of multiexponential signals.
Authors: A. Gómez-Sánchez, R.Vitale, O. Devos, A. de Juan, C. Ruckebusch.
Citation reference: *Analytica Chimica Acta* (2023), 1273: 341545.
DOI: 10.1016/j.aca.2023.341545

Blind Instrument Response Function Identification from Fluorescence Decays

Adrián Gómez-Sánchez[1,2]*, Olivier Devos[2], Raffaele Vitale[2], Michel Sliwa[2], Damir Sakhapo[3], Jörg Enderlein[3], Anna de Juan[1], Cyril Ruckebusch[2]

[1]Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028, Barcelona, Spain

[2]Univ. Lille, CNRS, UMR 8516 - LASIRe - Laboratoire Avancé de Spectroscopie pourles Intéractions la Réactivité et l'Environnement, F-59000 Lille, France

[3]III. Institute of Physics – Biophysics, Georg-August Universität, Göttingen, Germany.

*Corresponding authors: adrian.gomezsanchez.etu@univ-lille.fr

Abstract

Time-resolved fluorescence spectroscopy (TRFS) plays a crucial role when studying dynamic properties of complex photochemical systems. Nevertheless, the analysis of measured time decays and the extraction of exponential lifetimes often requires either the experimental assessment or the modeling of the instrument response function (IRF). However, the intrinsic nature of the IRF in the measurement process, which may vary across measurements due to chemical and instrumental factors, jeopardizes the results obtained by reconvolution approaches. In this paper, we introduce a novel methodology, called Blind Instrument Response Function Identification (BIRFI), that enables the direct estimation of the IRF from the collected data. It capitalizes on the properties of single exponential signals to transform a deconvolution problem into a well-posed system identification problem. To delve into the specifics, we provide a step-by-step description of the BIRFI method and a protocol for its application to fluorescence decays. The performance of BIRFI is evaluated using simulated and Time-Correlated Single-Photon Counting (TCSPC) data. Our results demonstrate that the BIRFI methodology allows an accurate recovery of the IRF, yielding comparable or even superior results compared to those obtained with experimental IRFs when they are used for reconvolution by parametric model fitting.

Key words: Time-Resolved Fluorescence Spectroscopy (TRFS), Blind Instrument Response Function Identification (BIRFI), Deconvolution.

Why it matters

Time-Resolved Fluorescence Spectroscopy is a crucial tool for understanding dynamic processes in photochemical systems. However, the interference of the Instrument Response Function (IRF) complicates accurate analysis. The novel Blind Instrument Response Function Identification (BIRFI) methodology proposed in this paper addresses this challenge. BIRFI transforms the deconvolution problem into a well-posed system identification problem, allowing the direct estimation of the IRF from measurements on fluorophores exhibiting single exponential behavior. This innovation is particularly significant since it suppresses the need to estimate the IRF through specific experimental procedures involving fluorophores with very short lifetimes or detectors that may suffer from color effects.

1. Introduction

Time-Resolved Fluorescence Spectroscopy (TRFS) is a widely used tool to investigate the dynamic properties of complex photochemical systems [1,2]. A short laser pulse is employed to initiate the excitation of fluorescent molecules within a complex sample. The subsequent return of these molecules to the ground state generates a fluorescence signal, denoted as $\boldsymbol{x}$, which diminishes over time. Ideally, this signal can be fitted by a mono- or multiexponential model, enabling the extraction of the lifetime(s) and amplitude(s) of the fluorescent species involved.

However, the detection of rapidly decaying responses that might be observable in a particular experiment is hindered by the instrument response function (IRF). This complication arises due to the interference of the IRF with the fluorescence signal, especially in scenarios where fast dynamics are at play [3]. TRFS measurements can be accurately characterized as the convolution of the IRF with the inherent fluorescence signal $\boldsymbol{x}$. In other words, the recorded signal, denoted as $\boldsymbol{y}$, can be expressed analytically as detailed in Equation 1.

$$\boldsymbol{y} = \boldsymbol{x} * \mathrm{IRF} \qquad \text{Eq. 1}$$

where * denotes functional convolution.

In the realm of discrete-time signals, operators and practitioners may encounter the need to: i) compute $\boldsymbol{x}$ based on $\boldsymbol{y}$ and the IRF through deconvolution [4], ii) determine the IRF based on $\boldsymbol{y}$ and $\boldsymbol{x}$ using system identification [5], or iii) simultaneously estimate both $\boldsymbol{x}$ and the IRF given $\boldsymbol{y}$ through blind deconvolution [4]. Each of these can pose challenges, since it is often ill-posed and not easily solved. To address these complexities, additional constraints, assumptions, or prior knowledge about the signals and systems under investigation must frequently be considered.

In TRFS, deconvolution is a standard operation since, when dealing with short lifetimes, typically around or below 1 ns, it requires knowledge or experimental estimation of the IRF. In literature, experimental estimation of the IRF often involves measuring the emission of a fluorophore with a very short fluorescence lifetime, such as Erythrosine B or Rose Bengal in a potassium iodide solution [6,7], or the elastic scattering of the excitation laser pulse using a LUDOX solution [3]. Once the IRF is estimated, deconvolution of the measured signal $\boldsymbol{y}$ can be performed, enabling the estimation of the true signal $\boldsymbol{x}$. Various approaches can be employed for this purpose, including polynomial long division [8], least squares [9], Fourier deconvolution [9], and reconvolution [10]. Reconvolution is a parametric fitting approach, where, given a specific IRF, Eq.1 is least squares-fitted to estimate the parameters of a mono- or multiexponential model that describes the behavior of $\boldsymbol{x}$. This method is widely used due to its numerical stability, but it does have several limitations [11,12].

One significant limitation is the intrinsic variability of the IRF, which may change across measurements due to chemical and instrumental factors. For example, TRFS measurements are often 'emission-dependent', in part because of the wavelength-dependent timing response of certain detectors (such as Micro Photon Devices - Single Photon Avalanche Photodiodes, MPD - SPADs). In this case, the measured IRF is only strictly valid for wavelength ranges that are close to those used for the IRF estimation, specifically the laser wavelength for scattering or dye emission. This dependence is particularly noticeable in the red region of the electromagnetic spectrum, crucial for biological imaging [13]. This leads to potential biases in the extraction of $\boldsymbol{x}$ when the measured IRF is used for reconvolution and parameter estimation. Various solutions have been suggested to address this issue, including the implementation of detectors robust to color effects or the utilization of a single exponential decay for correcting the phasor plot domain [14,15].

Additional limitations stem from the signal processing methods employed for deconvolution. Optimization algorithms like the Levenberg-Marquardt algorithm, commonly used in reconvolution, can be sensitive to local minima [9]. Furthermore, approaches like polynomial long division or Fourier transform are known to be highly sensitive to noise [9].

In contrast, system identification is a relatively straightforward scenario when $\boldsymbol{x}$ is known and the IRF signal comprises fewer sampling points than $\boldsymbol{y}$. In such instances, convolution can be expressed as an overdetermined system of linear equations (more equations than unknowns) by utilizing a Hankel matrix composed of corresponding shifted values of $\boldsymbol{x}$ [16]. Consequently, an ordinary least squares solution to this system of equations can always be found, as long as the Hankel matrix is invertible.

Blind deconvolution, on the other hand, often proves to be a challenging and critical process. It involves the simultaneous estimation of both the IRF and $\boldsymbol{x}$, presenting an ill-posed and underdetermined problem [4]. This essentially implies that the solution for $\boldsymbol{x}$ and IRF is not unique.

In TRFS data analysis, although the IRF is typically much shorter than the measured signal, practical application of system identification is rare. This is primarily due to the requirement of the knowledge of the underlying signal $\boldsymbol{x}$, which is generally unavailable. However, if $\boldsymbol{x}$ exhibit a single exponential decay behavior, the IRF can be extracted solely from the analysis of the measured signal $\boldsymbol{y}$, since its tail behaves as $\boldsymbol{x}$, up to scale variations. This is what Blind Instrument Response Function Identification (BIRFI), the new algorithm proposed in this paper, can achieve by leveraging the properties of exponential signals. BIRFI overcomes the main limitations of current methods, such as those related to potential color effects of detectors or the necessity to use specific fluorophores with very short lifetimes. Therefore, BIRFI provides the researchers with the possibility to extract the IRF from measurements carried out on affordable commercial fluorophores under experimental conditions that match the characteristics of their samples. Once the IRF is extracted, it can also be exploited for the analysis of unknown samples.

We provide a comprehensive, step-by-step description of BIRFI and its specific application to fluorescence decays. To gauge its efficacy, we rigorously assess its performance using simulated data and real Time-Correlated Single-Photon Counting (TCSPC) datasets representative of diverse analytical scenarios, specifically in the red emission range and using a common picosecond diode laser with IRF-dependent power.

2. Material and methods

We examined various simulated and real datasets to evaluate the effectiveness of the proposed approach and the extraction of the IRF in TCSPC measurements. The simulated datasets were utilized to assess the algorithm performance under tightly controlled conditions, while the real datasets were employed to evaluate its potential in practical experimental scenarios.

2.1 Simulated datasets

Three distinct IRFs were generated with shapes resembling those commonly observed in TRFS experiments: the first IRF is a Gaussian function, the second is a sum of two overlapping Gaussian functions, and the third is a sum of three overlapping Gaussian functions (refer to Fig. S3). The underlying signal $\boldsymbol{x}$ was simulated as a monoexponential decay with a lifetime of 1 ns, utilizing 1500 time-bins and a time resolution of 25 ps. Subsequently, $\boldsymbol{x}$ was convolved with the respective IRFs, resulting in three datasets: dataset 1, dataset 2, and dataset 3 for the first, second, and third IRF, respectively.

To emulate real conditions, different levels of Poisson noise (no noise, a low level of noise representing 0.5% of the total variance, and a high level of noise representing 2% of the total variance) were introduced to the convolved signal. This comprehensive approach allowed us to thoroughly assess the performance of the proposed method under various conditions.

2.2 Experimental datasets

A solution of the commercial dye ALEXA 647 (Thermo Fisher Scientific, Invitrogen) in PBS at pH 7.4 with a concentration of $5 \cdot 10^{-7}$ M was prepared. Measurements was performed using a PicoQuant TCSPC system,equipped with a FluoTime 200 spectrometer (bandpass 4 nm, 90° and magic angle configuration) and a picosecond laser diode emitting at 640 nm, with a repetition rate of 8 MHz. Detection was performed using a microchannel plate photomultiplier tube (MCP-PMT, Hamamatsu) (PicoHARP300) with a bin time of 4 ps.

Measurements were conducted at three different power intensity (40, 50, and 80% for dataset 1, 2, and 3, respectively), with ten replicates per power. The IRF was determined at the laser emission wavelength using the scattering of the laser from a nonfluorescent scattering solution (LUDOX colloidal silica solution). The full width at half maximum (FWHM) of the IRFs was approximately 100 ps but very different shapes were observed for the different power intensities. The measurements were carried out for an emission wavelength of 660 nm (4 nm bandpass) and stopped when the number of counts reached a maximum value of 10,000.

3. Blind IRF Identification (BIRFI)

BIRFI is designed to estimate the IRF solely from the analysis of the measured signal $\boldsymbol{y}$. The methodology builds upon the properties of monoexponentially decaying functions when convolved, which guarantee that the lifetime of $\boldsymbol{x}$ is the same as the one determining the behavior of the tail of $\boldsymbol{y}$ (Eq. 1). By leveraging this property, the IRF can be readily extracted from the sole measurement of $\boldsymbol{y}$ as the solution of a system identification problem.

To elucidate how we can detect the monoexponential tail of $\boldsymbol{y}$, consider the example depicted in Fig. 1. In this illustration, the signal $\boldsymbol{y}$ results from the convolution of a given IRF with a monoexponential $\boldsymbol{x}$ (refer to Fig. 1A). Calculating the derivative of $\boldsymbol{y}$, for instance, using the Savitzky-Golay algorithm [17], produces a profile with two distinctive trends in two different time intervals (see Supplementary Material for additional details about the properties of this derivative profile): the first (interval 1 in Fig. 1B) is associated with the region where the IRF significantly influences the shape of $\boldsymbol{y}$, and the second (interval 2 in Fig. 1B) corresponds to the region where the derivative of $\boldsymbol{y}$ exhibits monoexponential behavior. The time point that separates these two regions is referred to as the "cutting point," and it can be easily observed through visual inspection in Fig. 1B.

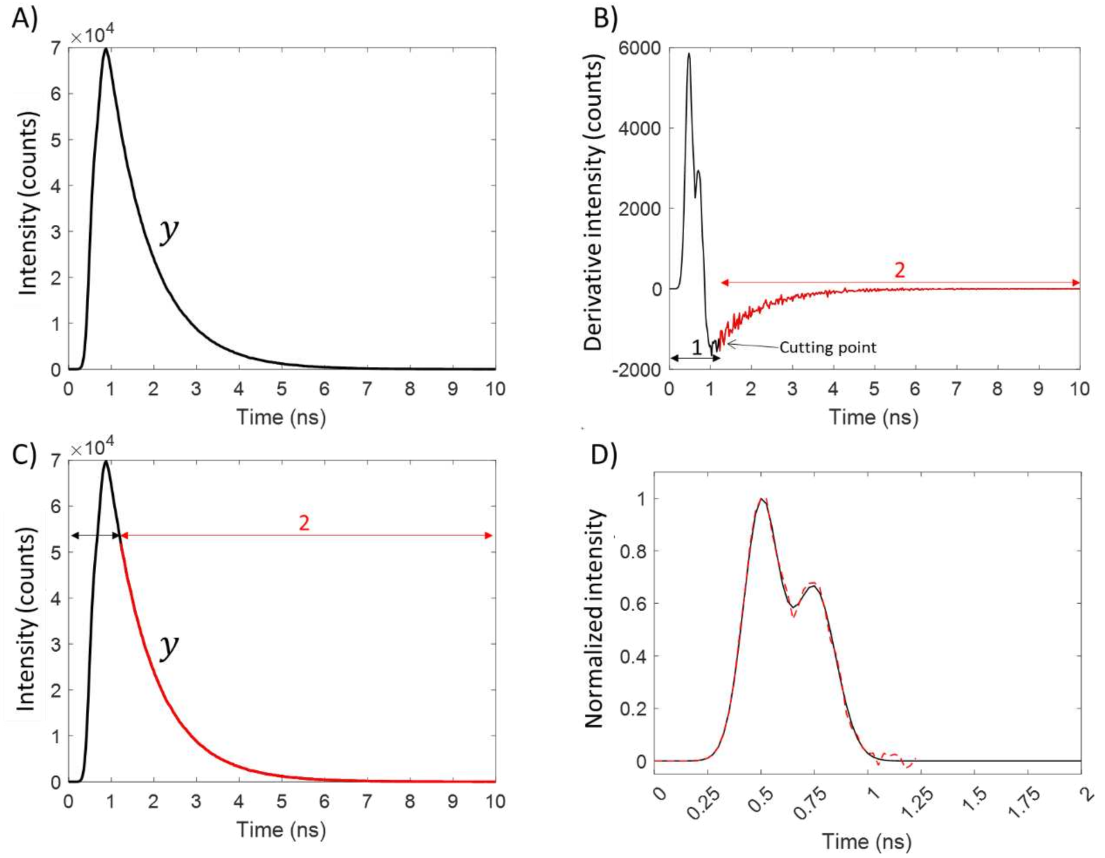


Figure 1. A) Measured convolved decay $\boldsymbol{y}$. B) Derivative of the measured convolved decay featuring an initial nonexponential trend (black, 1) and a tail with a characteristic exponential behaviour (red, 2). C) The exponential tail (red) is detected based on the recognition of the cutting point in Fig. 1B. D) The tail and the convolved decay are finally exploited to perform an ordinary deconvolution and estimate the IRF (red dashed line). The solid black line in D) represents the true IRF.

Here, three crucial points arise: first, computing the derivative of the measured signal facilitates the identification of the tail of $\boldsymbol{y}$. Second, beyond the cutting point, $\boldsymbol{y}$ and $\boldsymbol{x}$ exhibit identical single exponential shapes but differ in amplitude (as illustrated by the red line in Fig. 1C) and, third, the non-exponential part of $\boldsymbol{y}$ (Fig. 1C, black arrows) has the same time-span of the IRF— for a rigorous mathematical explanation of this property, please refer to the Supplementary Material.

At this juncture, with both $\boldsymbol{y}$ and $\boldsymbol{x}$ available ($\boldsymbol{x}$ being the tail of $\boldsymbol{y}$), the blind identification problem and Eq. 1 transforms into a well-posed problem and estimating the IRF can be accomplished through ordinary long polynomial division, least squares, or Fourier transform division.

Several noteworthy advantages afforded by the proposed methodology warrant emphasis. Foremost is the ability to retrieve the IRF by analyzing conventional fluorophores characterized by single exponential behaviors. Deconvolution is thus no longer impeded by emission wavelength dependencies since the deconvolved IRF is extracted directly from the measured signal and there is no need to use complex fluorophores with very short lifetimes which operate at specific wavelength, or to employ detectors robust to color effects. Additionally, no assumption is made about the shape of the IRF, and no parameter optimization operations are required.

A pivotal aspect of BIRFI lies in its capability to reliably, automatically identify and extract the tail of $\boldsymbol{y}$ that decays exponentiality as the inherent fluorescence signal $\boldsymbol{x}$. Therefore, it is valuable not only for applying system identification procedures but also for tail analysis or global fitting, both of which necessitate isolating the tail of the measured signal.

However, it is essential to acknowledge some inherent limitations of the BIRFI method. Firstly, BIRFI assumes that the signal $\boldsymbol{x}$ is well-approximated by a monoexponential decay function. In the rare scenarios where this requirement cannot be met by using very conventional fluorophores, blind unmixing tools should be applied to decompose $\boldsymbol{y}$ into its latent monoexponential decays before utilizing BIRFI to extract the IRF from such individual single exponential decays. To illustrate this, we provide as Supplementary Information results obtained coupling a novel multivariate curve resolution approach named MCR-Slicing [18,19] to BIRFI in a case concerning multiexponential decaying signals.

In addition, it is crucial to highlight that since BIRFI extracts the IRF from the experimental measurement acquired, a sufficiently high signal-to-noise ratio in $\boldsymbol{y}$ is essential for accurate deconvolution. To mitigate noise effects, one may consider increasing photon accumulation during TRFS experiments or applying noise-filtering approaches like Savitzky-Golay [17] or Whittaker smoothing [20], as demonstrated in this work.

Software

Data simulation and analysis were performed by means of in-house routines and scripts coded in MATLAB 2021 (The Mathworks, Inc., Natick, MA). The BIRFI algorithm is explained in the Supplementary Material. The DecayFit - Time-Resolved Emission Decay Analysis Software for MATLAB (version 1.4, Søren Preus, Ph.D) was used to perform the reconvolution analysis.

4. Results and discussion

4.1 Simulated datasets

To evaluate the performance of BIRFI, we delved into nine distinct data analysis scenarios, each distinguished by varying complexities in the shape of the RF and different levels of Poisson noise (refer to the Supplementary Material for comprehensive details). Whittaker smoothing was systematically applied to all analyzed decays to mitigate the influence of noise. Decay curves for each scenario were replicated 100 times to scrutinize the variability of the solutions obtained due to noise. Notably, in all instances, the single exponential tail of the decays was successfully retrieved, as elaborated in Section 3. The conclusive results of these analyses are presented in Fig. 2.

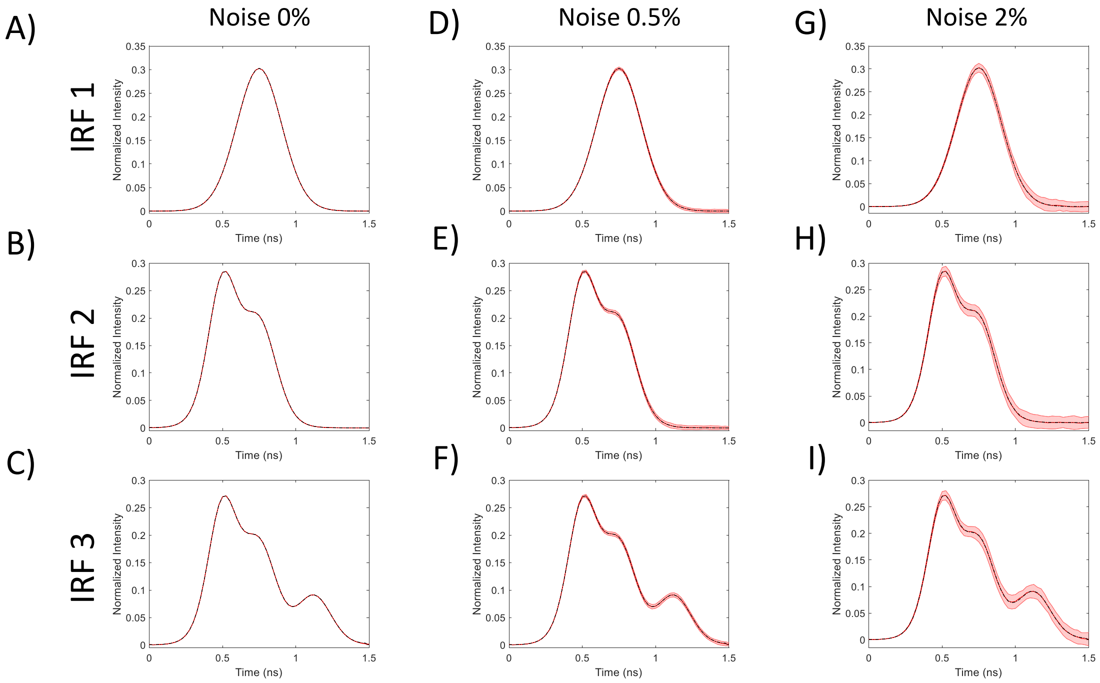


Figure 2. Predicted IRFs for three different levels of noise and three different expected IRF shapes. A-C) Results of BIRFI in the absence of noise for three expected IRFs of different shapes. D-F) Results of BIRFI in the presence of noise with intensity equal to the 0.5% of the total signal intensity for three expected IRFs of different shapes. G-I) Results of BIRFI in the presence of noise with intensity equal to the 2% of the total signal intensity for three expected IRFs of different shapes. Each analysis round was repeated 100 times: the average predicted IRF is represented as a red solid line, the red shaded area denotes the 95% confidence interval associated to this average estimation, while the expected IRF is represented as a black dashed line.

Initially, in the absence of noise (refer to Fig. 2A-C), as anticipated, the recovery of all IRFs is flawless, indicative of a well-posed signal identification problem.

Introducing noise into the scenario (see Fig. 2D-I) reveals a minor dispersion in the recovered IRFs across the 100 replicated analysis rounds. Notably, the mean representation across these runs (depicted by the red solid line) impeccably aligns with the ground truth (indicated by the black dashed line) in all instances. This alignment underscores that the observed dispersion is solely a result of the introduced noise and emphatically establishes the method accuracy.

These findings affirm that the methodology adeptly derives reliable estimates of the IRF from measured signals, operating without any prior assumptions about its shape.

4.2 Real datasets

To assess the performance of BIRFI on real experimental data, we conducted a series of TCSPC experiments. We analyzed ten replicates of TCSPC measurements conducted at three laser powers using a commercial ALEXA647 solution with a lifetime of 3.55 ns. Once again, Whittaker smoothing was applied as a preliminary step to each recorded decay, and the cutting point was set at 2.00 ns.

Figure 3 illustrates the preprocessed data and the IRFs obtained through the application of BIRFI (depicted by the red solid lines). These results showcase a highly satisfactory agreement between the estimated IRFs and the measured ones (indicated by the black dashed lines). Minor discrepancies, possibly stemming from solvent effects, are the only observable differences.

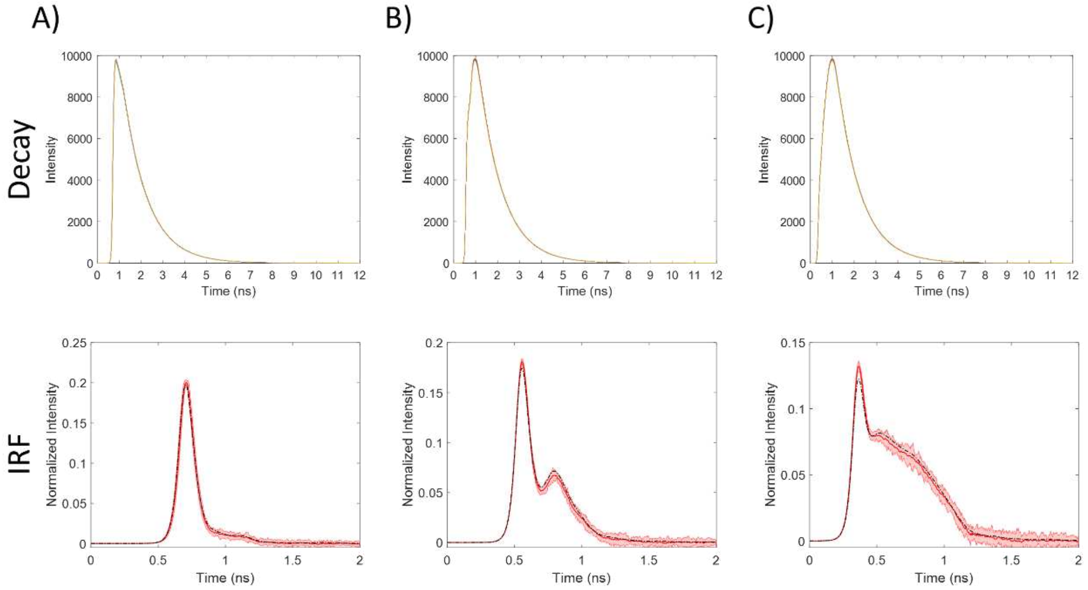


Figure 3. Results yielded by BIRFI for the first (A), the second (B) and the third (C) pure fluorescent dye dataset. Top panel: smoothed fluorescence decays (10 replicates per case). Bottom panel: average predicted IRFs (red solid lines) three standard deviation intervals (red shaded areas), and measured IRFs (black dashed lines).

To further validate the BIRFI approach and underscore its utility, we employed a parametric model (reconvolution) for each measured signal. This fitting process involved utilizing both i) the actual measured IRF and ii) the IRF estimated with BIRFI. The residuals of the parametric model are depicted in Fig. 4.

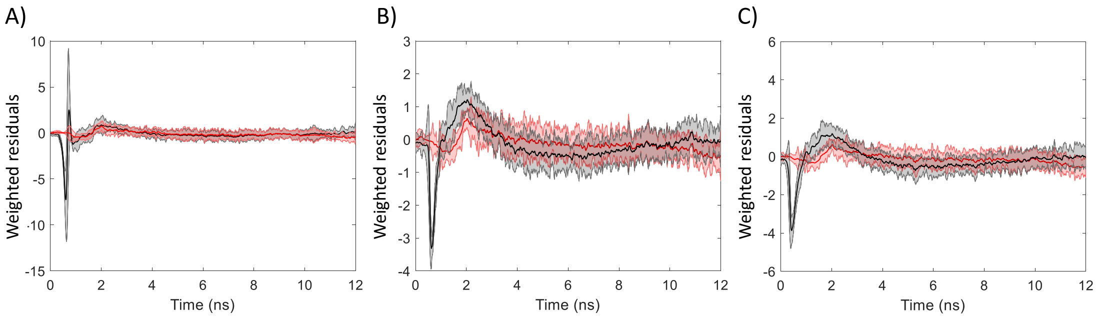


Figure 4. Weighted residuals resulting from the monoexponential fitting of the fluorescence decays in the first (A), the second (B) and the third (C) pure fluorescent dye dataset. The fitting was performed with reconvolution using the IRF estimated by BIRFI (red) and the measured IRF (black). Each measurement replicate was fitted individually: the solid lines represent the average residual profiles, while the red and black shaded areas denote the corresponding three standard deviation intervals.

Moreover, the estimates of lifetimes were examined. In dataset 1, a mean lifetime of 1.09(0.01) ns was derived using the predicted IRF, while a slightly lower estimate of 1.07(0.01) ns was obtained using the measured IRF. For comprehensive analysis, considering only the tail of the measured signal and fitting it with a monoexponential function yielded a mean lifetime of 1.09(0.01) ns. Although the lifetimes obtained in both cases are quite similar, the fitting residuals are notably less structured when reconvolution is performed with the IRF estimated through BIRFI compared to using the measured IRF (refer to Fig. 4A).

Similar conclusions can be drawn for datasets 2 and 3. In dataset 2, the average lifetimes are 1.07(0.01) ns and 1.04(0.01) ns when reconvolution is conducted using the IRF estimated by BIRFI and the measured IRF, respectively, while tail fitting provides a lifetime estimate of 1.08(0.01) ns. For dataset 3, the average lifetimes are 1.08(0.02) ns and 1.05(0.02) ns, respectively, while tail fitting provides a lifetime estimate of 1.08(0.01) ns. In both cases, the fitting residuals obtained when reconvolution is carried out with the IRF estimated by BIRFI are significantly less structured (refer to Fig. 4B and C).

Overall, these results indicate that utilizing IRFs estimated by BIRFI for fitting TRFS data collected on individual fluorophore solutions (via reconvolution) results in lower and less structured residuals compared to when measured IRFs are employed. This implies that better estimates of the fluorescence lifetimes of the species underlying the investigated systems may be achieved by employing BIRFI prior to TRFS data processing. Furthermore, considering residual analysis as a tool for assessing the validity of statistical models in the realm of TRFS, another critical point emerges from the presented outcomes: high residuals may arise not only from inadequately formulated models but also from poorly estimated or inaccurately measured IRFs. TRFS users should thus be attentive when attempting to address high residuals by introducing more complex models with a higher number of parameters.

5. Conclusions

In this communication, we introduced a novel method, called Blind Instrument Response Function Identification (BIRFI) which enables the estimation of the IRF inherent in TRFS measurements, relying solely on the measurement of an exponential decay. This straightforward approach is grounded in the theoretical principles of convolution/deconvolution applied to monoexponential decays. As demonstrated, it possesses broad applicability, considering additional practical aspects that were also explored. Overall, BIRFI consistently delivered accurate estimates of the IRF across scenarios of varying complexity in terms of IRF shape, thus improving the accuracy of lifetime determination. Its robustness is now open to scrutiny and validation by potential users.

Supporting material

The Supporting material contains the mathematical foundations BIRFI, emphasizing key properties of exponential decays essential for understanding its principles. It includes detailed information on simulated datasets and provides the complete MATLAB code for the BIRFI algorithm.

Acknowledgments

A. G.-S. and A.J. acknowledge financial support from the Spanish government Project PID 2019-1071586 B-IOO and the Catalan government (2021 SGR 00449). A. G.-S. acknowledges scholarships from the MOBLILEX U Lille program, from the Santander bank and from Fundació Montcelimar.

Author contributions

Adrián Gómez-Sánchez: Conceptualization, Methodology, Data acquisition, Software, Formal analysis, Investigation, Discussion, Data curation, Writing – original draft, Visualization.

Olivier Devos: Data acquisition, Discussion, Writing – review & editing.

Raffaele Vitale: Conceptualization, Discussion, Writing – review & editing.

Michel Sliwa: Data acquisition, Discussion, Writing – review & editing.

Damir Sakhapo: Discussion, Writing – review & editing.

Jörg Enderlein: Discussion, Writing – review & editing.

Anna de Juan: Resources, Writing – original draft, Discussion, Supervision, Funding acquisition.

Cyril Ruckebusch: Conceptualization, Methodology, Formal analysis, Resources, Writing – original draft, Discussion, Supervision, Project administration, Funding acquisition.

Declaration of interests

The authors declare no competing interests.

References

1. Liput, D. J.; Nguyen, T. A.; Augustin, S. M.; Lee, J. O.; Voge, S. S. (**2020**). A guide to fluorescence lifetime microscopy and Förster's Resonance Energy Transfer in Neuroscience. *Current Protocols in Neuroscience*. 94(1): e108.
2. J. R. Lakowicz. *Principles of Fluorescence Spectroscopy*, 3rd ed.; Springer, USA, **2006**.

3. Luchowski, R.; Gryczynski, Z.; Sarkar, P.; Borejdo, J.; Szabelski, M.; Kapusta, P.; Gryczynski, I. (**2009**). Instrument response standard in time-resolved fluorescence. *Review of Scientific Instruments*. 80(3), 033109.
4. Riad, S. M. The deconvolution problem: An overview. Proceedings of the IEEE (**1986**). 74(1), 82-85.
5. Abed-Meraim, K.; Qiu, W.; Hua, Y. (**1997**) Blind system identification. *Proceedings of the IEEE*. 85(8) 1310-1322.
6. Szabelski, M.; Ilijev, D.; Sarkar, P.; Luchowski, R.; Gryczynski, Z.; Kapusta, P.; Erdmann, R.; Gryczynski, I. (**2009**). Collisional quenching of erythrosine B as a potential reference dye for impulse response function evaluation. Applied Spectroscopy 63(3): 363-368.
7. Szabelski, M.; Luchowski, R.; Gryczynski, Z.; Kapusta, P.; Ortmann, U.; Gryczynski, I. (**2009**). Evaluation of instrument response functions for lifetime imaging detectors using quenched Rose Bengal solutions. *Chemical Physics Letters*, 471(1-3), 153-159.
8. Bini, D.; Pan, V. (**1986**). Polynomial division and its computational complexity. Journal of Complexity 2(3): 179-203.
9. Vetterling, W. T.; Teukolsky, S. A.; Press, W. H.; Flannery, B. P. (**1999**). Numerical Recipes in C: The Art of Scientific Computing. *Cambridge University Press.*
10. O'Connor, D. V.; Ware, W. R.; Andre, J. C. (**1979**). Deconvolution of fluorescence decay curves. A critical comparison of techniques. Journal of *Physical Chemistry* 83(10): 1333-1343.
11. Xiao, D.; Sapermsap, N.; Safar, M.; Cunningham, M. R.; Chen, Y.; Li, D. D. U. (**2021**). On synthetic instrument response functions of time-correlated single-photon counting-based fluorescence lifetime imaging analysis. *Frontiers in Physics* 9: 635645.
12. Wahl, Ph.; Auchet, J. C.; Donzel, B. (**1974**). The wavelength dependence of the response of a pulse fluorometer using the single photoelectron counting method. *Review of Scientific Instruments* 45(1): 28-32.
13. Reja, S. I.; Minoshima, M.; Hori, Y.; Kikuchi, K. (**2021**). Near-infrared fluorescent probes: a next-generation tool for protein-labeling applications. *Chemical Science*, 12(10), 3437-3447
14. Ranjit, S.; Malacrida, L.; Jameson, D. M.; Gratton, E. (**2018**). Fit-free analysis of fluorescence lifetime imaging data using the phasor approach. *Nature protocols*, 13(9), 1979-2004.
15. Štefl, M.; James N. G.; Ross, J. A.; Jameson, D. M.; (**2011**). Applications of phasors to in vitro time-resolved fluorescence measurements. *Analytical biochemistry*, 410(1), 62-69.
16. Eilers, Paul HC; Ruckebusch, Cyril. (**2022**). Fast and simple super-resolution with single images. *Scientific Reports*, 12(1), 11241.
17. Savitzky, A.; Golay, M. J. E. (**1964**). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8), 1627-1639.
18. Devos, O.; Ghaffari, M.; Vitale, R.; de Juan, A.; Sliwa, M.; Ruckebusch, C. (**2021**). Multivariate curve resolution slicing of multiexponential time-resolved spectroscopy fluorescence data. *Analytical Chemistry*, 93(37), 12504-12513.

19. Gómez-Sánchez, A.; Marro, M.; Marsal, M.; Loza-Alvarez, P.; de Juan, A. (**2020**). 3D and 4D image fusion: Coping with differences in spectroscopic modes among hyperspectral images. *Analytical Chemistry*, 92(14), 9591-9602.
20. Eilers, P. H. C. (**2003**). A perfect smoother. *Analytical Chemistry*, 75(14), 3631-3636.

Supporting material

Blind Instrumental Response Function Identification (BIRFI) from Fluorescence Decays

Adrián Gómez-Sánchez[1,2], Olivier Devos[2], Raffaele Vitale[2], Michel Sliwa[2], Damir Sakhapo[3], Jörg Enderlein[3], Anna de Juan[1], Cyril Ruckebusch[2]

[1]Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028, Barcelona, Spain

[2]Univ. Lille, CNRS, UMR 8516 - LASIRe - Laboratoire Avancé de Spectroscopie pourles Intéractions la Réactivité et l'Environnement, F-59000 Lille, France

[3]III. Institute of Physics – Biophysics, Georg-August Universität, Göttingen, Germany.

Abstract

This supporting material provides additional insights into the mathematical rationale behind the Blind Instrumental Response Function Identification (BIRFI) approach and describes key mathematical properties of exponential decays that are essential for a comprehensive understanding of its operating principles. Detailed information on the simulated datasets is also provided, as well as on the potential use multivariate curve resolution approaches to unmix multiexponential decays before applying BIRFI. Finally, the full MATLAB code of the BIRFI algorithm is made available.

Definition and properties of convolved decays

Here, the convolution of a single exponential decay $\boldsymbol{x}$ with a known IRF is defined. We demonstrate that the tail of the convolved exponential decay $\boldsymbol{y}$ is characterized by the same time constant as the single exponential decay $\boldsymbol{x}$, the one would ideally observe in the absence of convolution, which constitutes the basis of the BIRFI method.

Consider the fluorescence monoexponential decay $x(t)$ in Eq. S1,

$$x(t) = Ae^{\frac{-t}{\tau}} \qquad \text{Eq. S1}$$

where A denotes the preexponential factor, τ denotes the lifetime, t denotes the time variable and $x(t)$ has a domain $\{t \in \mathbb{R} : t \geq 0\}$. Given the fact that the IRF can be defined as a finite function, $IRF(t)$, with domain $\{t \in \mathbb{R} : 0 \leq t \leq m\}$, then the measured signal $y(t)$ is the product of the convolution of $x(t)$ with $IRF(t)$ (see Eq. S2).

$$y(t) = \int_{-\infty}^{+\infty} Ae^{\frac{-(t-t_o)}{\tau}} IRF(t)\, dt_o \qquad \text{Eq. S2}$$

The full convolution in Eq. S2 can be expressed as a combination of three different convolution operations in three distinct domains of t.

Domain 1)

$$t < 0 \ \rightarrow y(t) = \int_{-\infty}^{0} Ae^{\frac{-(t-t_o)}{\tau}} IRF(t_o)\, dt_o = 0 \qquad \text{Eq. S3}$$

Domain 2)

$$0 \leq t < m \rightarrow y(t) = \int_{0}^{t} Ae^{\frac{-(t-t_o)}{\tau}} IRF(t_o)\, dt_o = Ae^{\frac{-t}{\tau}} \int_{0}^{t} e^{\frac{t_o}{\tau}} IRF(t_o) dt_o \qquad \text{Eq. S4}$$

Domain 3)

$$t \geq m \rightarrow y(t) = \int_{0}^{m} Ae^{\frac{-(t-t_o)}{\tau}} IRF(t_o)\, dt_o = Ae^{\frac{-t}{\tau}} \int_{0}^{m} e^{\frac{t_o}{\tau}} IRF(t_o)\, dt_o \qquad \text{Eq. S5}$$

In Domain 3, (i.e., for $t \geq m$), it can be observed that $y(t)$ is equal to the scalar $\int_0^m e^{\frac{t_o}{\tau}} IRF(t_o)\, dt_o$ multiplied by the original fluorescence decay $Ae^{\frac{-t}{\tau}}$ (see also Fig. S1). This means that, when an exponential signal is convolved, irrespectively of the nature of the IRF, the underlying exponential $\boldsymbol{x}$ is recovered for $t \geq m$, which defines the tail of this signal.

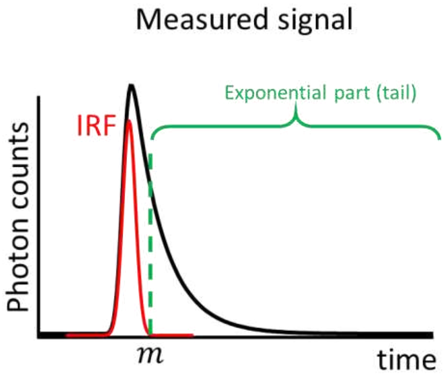


Figure S1. Measured signal $y(t)$ (black solid line) resulting from the convolution of a monoexponential decay and the IRF represented in red. For $t > m$, with m being the size of the IRF, a pure monoexponential behaviour is observed.

Optimal data-cutting point

One of the problems that TCSPC users most commonly encounter and face regards the identification of the so-called cutting point, i.e., the time instant after which the measured fluorescence decays assume an exponential behaviour. Identifying correctly this cutting point permits to increase the signal-to-noise ratio and, thus, enhance the quality of the final results, since more sampling points can be taken into account during the analysis of the data. Detecting when the aforementioned exponential behaviour begins becomes, therefore, particularly relevant in many practical scenarios. To do this, as already highlighted before, based on Eq. S5, it can be said that such an exponential behaviour is observed and measured without any distortion at $t > m$ (Domain 3). Calculating the first derivatives of Eqs. S3, S4 and S5 as in Eq. S9, the difference on the frontier between the case 2 and the case 3 is significative enhanced (Fig. 3).

$$\frac{\partial y(t)}{\partial t} = \begin{cases} 0, & \text{Domain 1: } t \leq 0 \\ \dfrac{\partial B(t) \cdot Ae^{\frac{-t}{\tau}}}{\partial t}, & \text{Domain 2: } 0 < t < m \\ \dfrac{-B}{\tau} Ae^{\frac{-t}{\tau}}, & \text{Domain 3: } m \leq t \end{cases} \qquad \text{Eq. S6}$$

One can observe that this derivation affects the shape of the original decay within domain 2, but not within domain 3 – in this latter domain, the original exponential decay is only reflected and scaled by $\frac{B}{\tau}$. This property can be exploited to enhance the visualization of the frontier between domains 2 and 3 as shown in Fig. S2. Derivatives of higher order can also be employed in order to enhance even more the visualization of this frontier, but it has to be borne in mind that higher order derivatives gradually amplify noise. For this reason, an optimal compromise should always be found.

In addition, it is worth noticing that the derivative minimum does not exactly correspond to the size of the IRF, but approximates very accurately the frontier separating domains 2 and 3. For this reason, it should be resorted to as a graphical rather than absolute mathematical criterion for the identification of the value of m.

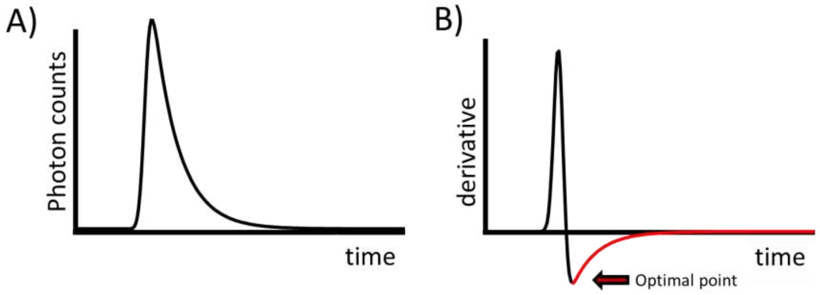


Figure S2. A) A fluorescence measured decay $y(t)$. B) First derivative of $y(t)$. It is possible to observe the visual enhancement of the frontier between domains 2 and 3 in B) with respect to A), which provides an easier way to separate and isolate (for data analysis purposes) the non-exponential and exponential intervals of $y(t)$.

This optimal cutting point determination approach is also valid for mixtures of fluorescence decays, as could be easily demonstrated by calculating the first derivative of Eq. S8 (not shown). Therefore, the cutting-point approach can be used to properly detect the tail of a multiexponential signal, being useful for tail-fitting analysis.

Simulated datasets

Three IRFs with different shapes were generated. The first is gaussian (see Fig. S3A), while the second and the third result from the sum of two and three distinct gaussian functions, respectively (see Figs. S3B and S3C). In addition, a single exponential decay was simulated with tau equal to 1 ns and 1498 sampling points. The time bin resolution was 25 ps. The single exponential decay was convolved with each IRF. After the convolution, Poisson noise was added in different proportions (see the manuscript for further details). The amount of noise added to the convolved decay has been calculated as in Eq. S10,

$$\mathbf{E}(\%) = \sqrt{\frac{\Sigma_i\left(y_i - y_{i,f}\right)^2}{\Sigma_i\left(y_{i,f}\right)^2}} \cdot 100 \qquad \text{(Eq. S7)}$$

where y_i and $y_{i,f}$ denote the i-th element of the noisy convolved decay and the i-th element of the noise-free convolved decay, respectively.

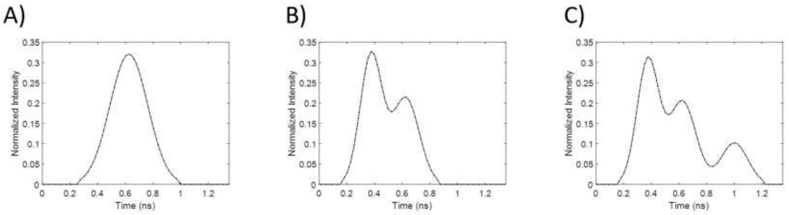


Figure S3. A) Gaussian IRF. B) IRF resulting from the sum of two different gaussian functions. C) IRF resulting from the sum of three different gaussian functions.

Multiexponential decay case

The ideal scenario for applying BIRFI to extract the IRF is working with a fluorophore whose decay can be characterized by a single exponential model. However, if required, for more complex fluorophores, the signal can be unmixed into a sum of individual monoexponential components to which can BIRFI can be applied. For this purpose, we use here MCR-Slicing (see Ref. 15 of the main manuscript), but other tools could be considered. As an illustrative example, seven mixtures of ATTO 655, ATTO 665, and ATTO 647N in different proportions were measured using the same TCSPC system employed for the acquisition of the ALEXA measurement in the main manuscript. The convolved decays are provided in Fig. S4A (left) and the individual component recovered are shown in Fig. S4A (right). Note that the only input of MCR-Slicing is the number of components. In Fig. S4B are provided the results obtained applying BIRFI to each extracted individual decay. In Fig. S4B (right), the mean of the three IRFs extracted and the measured IRF is shown. The predicted IRF align exceptionally well with the IRF measured using a LUDOX solution. This fact proves the quality of the unmixing task provided by MCR-slicing and the reliability of the BIRFI approach, which provides very consistent IRFs with the pure resolved profiles of the three dyes in the mixture.

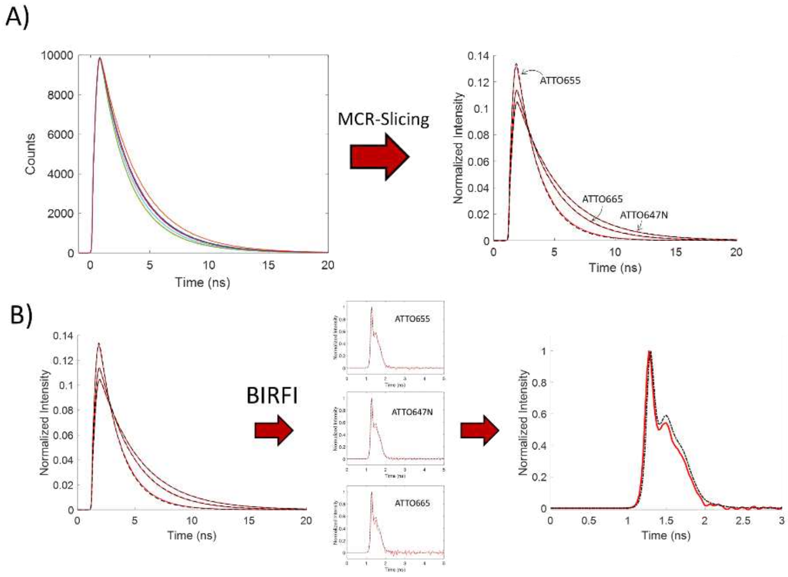


Figure S4. A) Left plot: The data set containing fluorescence decay of the seven mixtures is shown. Right plot: Results of the MCR-Slicing applied to the data set. In red solid line, output of MCR-Slicing. In black dashed line, measured pure decays. B) The single exponential decays obtained by MCR-Slicing are used by BIRFI to achieve the IRF. Right plot: mean predicted IRF in solid red line and measured IRF (black dashed line).

BIRFI code

The function `birfi_ls` is based on least squares. The function `birfi_dec` is based on the MATLAB function `deconv`. Both functions yield equivalent outputs.

```
function [irf]=birfi_ls(decay,irf_size)
%"decay" is the full decay sized 1xD.
%"irf_size" is a scalar indicating the size of the irf or
cutting point.
decaysize=size(decay,2);
decay_cut=decay(1,irf_size:end);
f2=[zeros(1,irf_size-1),decay_cut,zeros(1,irf_size-1)];
for i=1:decaysize
HankelM(i,:)=fliplr(f2(i:i+decaysize-size(decay_cut,2)));
end
irf=decay/HankelM';
end
```

```
function [irf]=birfi_dec(decay,irf_size)
%"decay" is the full decay sized 1xD.
%"irf_size" is a scalar indicating the size of the irf or
cutting point.
decay_cut=decay(1,irf_size:end);
irf=deconv(decay,decay_cut);
end
```

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Kernelizing: A way to increase accuracy in trilinear decomposition analysis of multiexponential signals

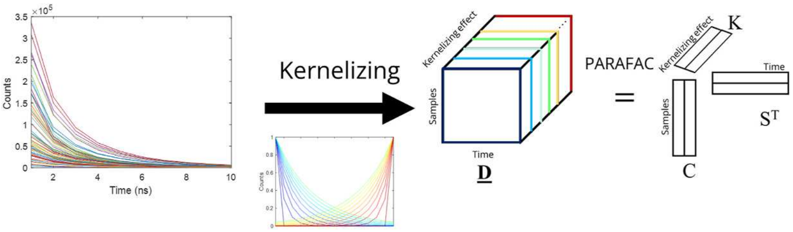
Adrián Gómez-Sánchez [a,b,*], Raffaele Vitale [b], Olivier Devos [b], Anna de Juan [a], Cyril Ruckebusch [b]

[a] *Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028, Barcelona, Spain*
[b] *Univ. Lille, CNRS, UMR 8516, LASIRe, Laboratoire Avancé de Spectroscopie pour Les Intéractions La Réactivité et L'Environnement, F-59000, Lille, France*

HIGHLIGHTS

- A tensorization approach based on convolution of exponential decays is proposed.
- The new Kernelizing approach is able to generate trilinear from bilinear data.
- This new unmixing approach provides the pure monoexponential components.
- Kernelizing can be applied even if a low number of sampling points is only available.
- Kernelizing has been tested in simulations, experiments and FLIM images.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Prof. L. Buydens

Keywords:
Kernelizing
Multiexponential decays
Tensorization
Trilinear models
Parallel factor analysis-alternating least squares (PARAFAC-ALS)

ABSTRACT

The unmixing of multiexponential decay signals into monoexponential components using soft modelling approaches is a challenging task due to the strong correlation and complete window overlap of the profiles. To solve this problem, slicing methodologies, such as PowerSlicing, tensorize the original data matrix into a three-way data array that can be decomposed based on trilinear models providing unique solutions. Satisfactory results have been reported for different types of data, e.g., nuclear magnetic resonance or time-resolved fluorescence spectra. However, when decay signals are described by only a few sampling (time) points, a significant degradation of the results can be observed in terms of accuracy and precision of the recovered profiles.

In this work, we propose a methodology called Kernelizing that provides a more efficient way to tensorize data matrices of multiexponential decays. Kernelizing relies on the invariance of exponential decays, i.e., when convolving a monoexponential decaying function with any positive function of finite width (hereafter called "kernel"), the shape of the decay (determined by the characteristic decay constant) remains unchanged and only the preexponential factor varies. The way preexponential factors are affected across the sample and time modes is linear, and it only depends on the kernel used. Thus, using kernels of different shapes, a set of convolved curves can be obtained for every sample, and a three-way data array generated, for which the modes are sample, time and kernelizing effect. This three-way array can be afterwards analyzed by a trilinear decomposition method, such as PARAFAC-ALS, to resolve the underlying monoexponential profiles. To validate this new approach and assess its performance, we applied Kernelizing to simulated datasets, real time-resolved fluorescence spectra

[*] Corresponding author. Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028, Barcelona, Spain.
E-mail address: agomezsa29@alumnes.ub.edu (A. Gómez-Sánchez).

https://doi.org/10.1016/j.aca.2023.341545
Received 14 February 2023; Received in revised form 9 June 2023; Accepted 18 June 2023
Available online 19 June 2023

collected on mixtures of fluorophores and fluorescence-lifetime imaging microscopy data. When the measured multiexponential decays feature few sampling points (down to fifteen), more accurate trilinear model estimates are obtained than when using slicing methodologies.

1. Introduction

The analysis of exponential decay signals is usually performed by multiexponential fitting approaches that allow extracting the characteristic decay constants and preexponential factors of the different monoexponential components [1]. However, multiexponential fitting remains difficult due to the high natural correlation among the monoexponential decays of these individual components and becomes even more complex for signals exhibiting a low signal-to-noise ratio [2]. Besides, the results obtained can be very user-dependent since selecting the correct number of monoexponential components and setting appropriate initial parameters for the fitting are tasks that require expertise and are often based on a trial-and-error approach.

In this context, factor analysis can constitute a good alternative to multiexponential fitting. Indeed, specific chemometric approaches are available to solve the unmixing (curve resolution) problem for exponential mixtures and have been successfully applied in e.g., Nuclear Magnetic Resonance (NMR) [3,4] and Time-Resolved Fluorescence Spectroscopy (TRFS) [5–7]. Among these approaches, PowerSlicing [8] is a method aimed at resolving mixtures of monoexponential decays, based on (i) the reorganization of the collected dataset into a three-way data array by a so-called slicing approach and (ii) the subsequent application of Parallel Factor Analysis-Alternating Least Squares (PARAFAC-ALS) [9]. From a broad perspective, data slicing can be considered a tensorization approach [10] which consists of splitting the exponential decays of a data matrix, say **D**, into several equally sized slabs or “slices” covering different signal time ranges separated by a certain lag. The slices obtained are afterwards rearranged into a three-way data array $\underline{\mathbf{D}}$, to which a trilinear decomposition method is applied (see Fig. 1A). Trilinearity offers the advantage of uniqueness, being trilinear models more robust to noise and less affected by the choice of the initial estimates as long as the datasets analyzed have full rank [9,11]. These properties enhance significantly the capacity for unmixing multiexponential signals even in conditions of complete window overlap and high correlation among profiles. However, methodologies like PowerSlicing usually require that the measured signals encompass many sampling points (hundreds), so that the slicing procedure can be efficiently performed returning accurate results. In real scenarios where the exponential signals would consist of a few tenths of sampling points, results could be much less accurate or even incorrect.

In this work, we propose an alternative approach, called Kernelizing, to tensorize multiexponential signals characterized by only few sampling points. Kernelizing exploits the following invariance property of exponential functions, i.e., when convolving a monoexponential decaying function with any positive function of finite width (hereafter called “kernel”), the shape of the monoexponential decay (determined by the characteristic decay constant) remains unchanged and only the preexponential factor varies. This approach provides a new way to build three-way data arrays from the measured two-way data matrices of decay curves. To do this, each measured exponential decay (each row of the data matrix) is convolved with a set of different kernels, yielding new signals for which the decay constants of the individual monoexponential components are unchanged and only the corresponding preexponential factors are modified, but preserving the relative proportion of the components across the samples analyzed. Such an operation yields the “slices” needed to build a three-way trilinear data array from the original bilinear data to which a trilinear data decomposition, such as PARAFAC-ALS, can be applied to extract decay constants and concentration profiles of individual components (see Fig. 1B).

Both PowerSlicing and Kernelizing generate slices that are arranged

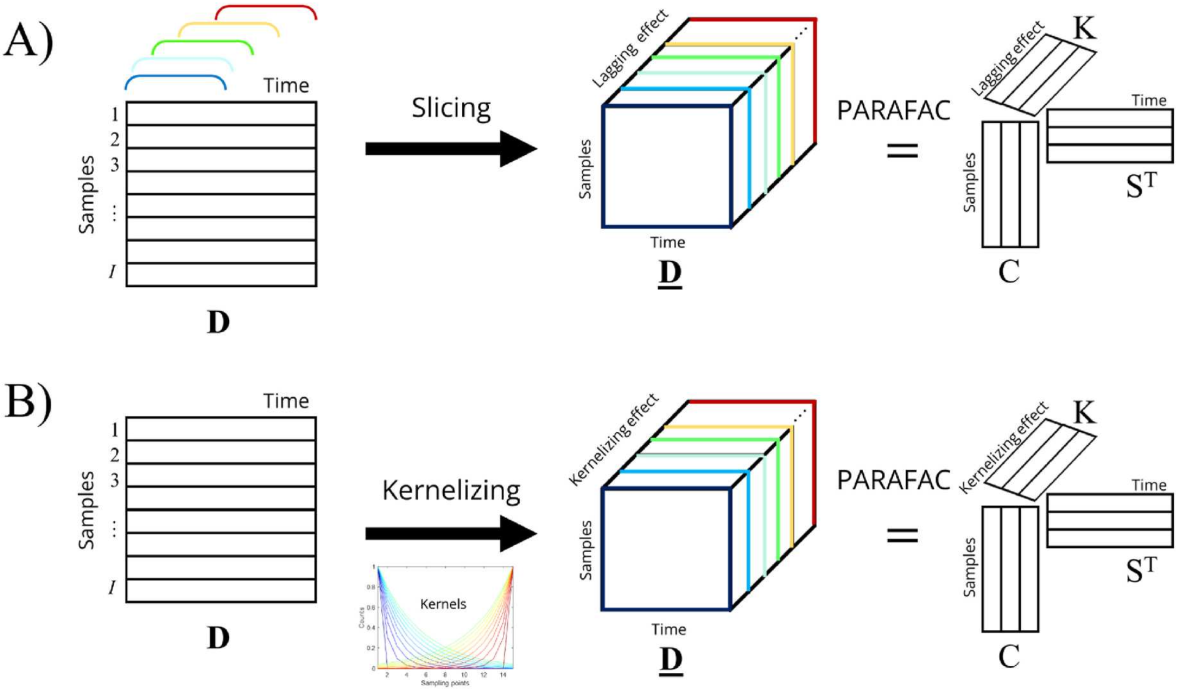


Fig. 1. Multilinear decomposition of exponential signals A) Slicing allows tensorizing a bilinear dataset into a trilinear dataset obtained by the selection of slices of the original decay separated by certain lags. B) Kernelizing allows tensorizing a bilinear dataset into a trilinear dataset convolving the original signals with a set of kernels.

Table 1
Composition of each analyzed mixture (expressed as volume ratios of standard solutions of pure dyes).

Mixture	ATTO 647 N (v/v)	ATTO 655 (v/v)	ATTO 665 (v/v)
1	1/3	1/3	1/3
2	4/6	1/6	1/6
3	1/6	4/6	1/6
4	1/6	1/6	4/6
5	5/12	5/12	2/12
6	2/12	5/12	5/12
7	5/12	2/12	5/12

in a three-way data array. However, the substantial difference between PowerSlicing (or other methodologies based on data lagging) and the Kernelizing approach is that the latter can provide a three-way array with an unlimited number of slices (as many slices as the number of kernels used), whereas PowerSlicing can only work with a limited number of slices, defined as a function of the total number of data points resulting from the sampling of the measured exponential signals. Such a difference becomes crucial when working with exponential signals with a low number of sampling points, a scenario in which Kernelizing improves significantly the accuracy and precision of the monoexponential profiles resolved.

To assess the performance of the proposed approach, several datasets simulated considering mixtures of monoexponential decays with different numbers of sampling points were studied. To complement these simulations, TRFS datasets were also investigated. TRFS is a well-established spectroscopic technique aiming at measuring the emission decay of a fluorophore in the picosecond to nanosecond timescale for the characterization of its lifetime (decay constant) [5]. Fluorescence lifetime is fluorophore-specific, dependent on the physicochemical environment probed, and provides valuable information about the sample. A real dataset was obtained gathering measurements performed on mixtures of known composition of ATTO fluorophores in solution using Time Correlated Single Photon Counting (TCSPC), a TRFS-based technique. Both PowerSlicing and Kernelizing were applied to the simulated and the TCSPC datasets (for which the ground truth solutions were known) to illustrate the benefits of the new approach proposed. Finally, the potential of Kernelizing was tested on a challenging real example of Fluorescence-Lifetime Imaging Microscopy (FLIM) [12], another TRFS-based technique widely used in the bioimaging field [13].

Although all the case-studies presented involve TRFS measurements, the results and conclusions inferred can be generalized to any analytical signal following a multi-exponential decaying behavior.

2. Datasets and software

The Kernelizing methodology was tested on simulated and experimental TRFS datasets. All datasets are representative of challenging scenarios for curve resolution approaches (high correlation among decay profiles, no selectivity, i.e., absence of subwindows with pure channels in either the concentration or the decay direction, and significant amount of Poisson-structured noise). The simulated dataset 1 consists of a set of multiexponential decay curve signals sampled at 500 time points (columns of the data matrix). The simulated dataset 2 consists of the same signals sampled at 50 points. The simulated dataset 3 consists of the same signals, with only 15 sampling points. The ATTO experimental dataset contains exponential decay curves of known mixtures of fluorescence dyes, originally sampled at 1500 time points. Finally, the ConvM dataset corresponds to a real FLIM image of a *Convallaria Majali*, for which pixel decay curves were sampled at 16 time points only. Further details about each dataset are given below. It should be noted that all datasets are full-rank.

2.1. Simulated dataset 1

Three pure monoexponential decays were simulated with decay constants of 0.8 ns, 1.4 ns and 2.4 ns, respectively, and with 500 equispaced sampling points. The three pure exponential profiles were organized in a matrix sized 3 × 500. These profiles show strong correlation with complete window overlap, i.e., all the pure monoexponential decays have a correlation coefficient $\geq$0.9 among them and cover the full time range concerned. A total of 200 mixtures with different composition were simulated to build a concentration matrix, sized 200 × 3. The two-way dataset 1 (200 × 500) of multiexponential decay curves was generated by matrix multiplication of the concentration and pure monoexponential decay matrices. All samples have a significant contribution of the three components, resulting in a non-trivial unmixing problem, since no pure sample for any component exists. Poisson noise was added representing approximately 10.0% of the global signal.

2.2. Simulated dataset 2

As for dataset 1, three pure monoexponential decays with constants of 0.8 ns, 1.4 ns and 2.4 ns were used. However, in this case, they were downsampled (the total time range covered was unchanged) and only 50 sampling points were considered , which resulted in a pure profile matrix sized 3 × 50. The concentration matrix, sized 200 × 3, was identical to the one exploited for dataset 1. The dataset 2, sized 200 × 50, was generated by matrix multiplication of the concentration and pure monoexponential decay matrices. Poisson noise was added representing approximately 5.0% of the total signal.

2.3. Simulated dataset 3

The same pure monoexponential decays as for dataset 1 and 2, with constants of 0.8 ns, 1.4 ns and 2.4 ns, were used; however, only 15 sampling points were now considered. The concentration matrix was left unchanged. The resulting dataset 3, sized 200 × 15, was obtained as described before and Poisson noise was added representing approximately the 4.5% of the total signal.

2.4. Time Correlated Single Photon Counting (TCSPC) data of ATTO fluorophores

Seven mixtures were prepared using solutions of three commercial dyes (ATTO 647, ATTO 655 and ATTO 665 from ATTO-TEC GmbH, a. r.) at the volume ratios listed in Table 1. Standard solutions of pure dyes were prepared in phosphate buffer solution (PBS) at pH 7.4 with a concentration of $5{\cdot}10^{-7}$ M.

All TCSPC measurements were performed using a PicoQuant TCSPC system with a FluoTime 200 spectrometer equipped with a picosecond laser diode emitting at 640 nm with a pulse width <90 ps full width at half-maximum (FWHM) and a repetition rate of 8 MHz.

A microchannel plate photomultiplier tube (MCP-PMT) connected to a TCSPC system (TimeHarp260, time precision 20 ps, dead time 25 ns) with a bin time of 25 ps was used for detection. The instrumental response function (IRF) of system (75 ps FWHM) was measured using a nonfluorescent scattering solution (LUDOX colloidal silica solution). Measurements stopped when the maximum reached 10 000 counts. The signals were recorded at 700 nm using a band pass filter with a 4 nm band pass, which resulted, after cropping the non-exponential signal portion originated from the IRF, in a TRFS data matrix composed of seven fluorescence decay curves in the range 0–20 ns with 1500 time points each. Thus, the experimental ATTO dataset is sized 7 × 1500 and does not contain any pure sample. In addition, a second dataset with fewer sampling points was built by sampling the ATTO dataset decay curves once every 100 points. The reduced dataset obtained is sized 7 × 15.

To assess the accuracy of the results obtained from the analysis of the mixtures, the fluorescence decays of the pure solutions of ATTO 655, ATTO 665 and ATTO 647N were measured and fitted by a monoexponential decay to retrieve an estimate of the ground-truth lifetimes (1.9 ns, 2.9 ns and 3.5 ns, respectively).

2.5. ConvM dataset

The ConvM dataset was generated from a FLIM image of *Convallaria Majali* acquired by G. Williams et al. [12]. Instrumental details are available in the original reference. In FLIM, for a given spectral channel, a multiexponential fluorescence decay is provided for each pixel. The resulting dataset is, therefore, a hypercube of four dimensions, sized 256 × 256 pixels × 512 spectral channels × 16 time points. For every pixel, the spectral dimension of the FLIM image was integrated by summing the intensity values of all the spectral channels at each time point in order to analyze only the time dimension and increase the signal-to-noise ratio. This operation provided a FLIM image of 256 × 256 pixels and 16 time-sampling points, with a decay curve associated to every pixel. The first 30 columns of pixels were cropped because the related signal was saturated. In addition, time channels 1 to 3 and 14 to 16 were removed because they presented a non-exponential behavior. After cropping, the final size of the image was 256 × 227 pixels × 11 sampling points. The related unfolded convallaria dataset had size 58 112 × 11.

2.6. Software

All in-house routines, scripts and analyses generated for the Kernelizing approach were performed using MATLAB 2021 (The Mathworks, Inc., Natick, MA). PowerSlicing was adapted from Engelsen et al. [8]. The N-Way331 toolbox [14] was used for PARAFAC-ALS analysis.

For all analyses, the PARAFAC-ALS convergence criterium was set as 10^{-10}%, this small value being justified by the high correlation of the monoexponential profiles to be retrieved. In this scenario, when pure profiles are very correlated, a clear change in the decay constants of the resolved profiles may result in a very small change in the model residuals. Therefore, the algorithm needs more iterations to refine the resolved profiles. The analysis was repeated 1000 times adding the same amount of Poisson-structured noise in each run for each simulated dataset. The results obtained across these multiple runs help in the assessment of the accuracy and precision of the solutions obtained.

3. Data analysis

Let us consider the dataset **D** shown in Fig. 1, constituted by a set of multiexponential decay curves from different samples, sized $I\times J$ (samples and number of time-sampling points, respectively). Using curve resolution methods and under certain constraints, such as non-negativity, the bilinear decomposition of **D** provides component profiles (i.e., decay profiles and related sample concentration profiles) with direct chemical meaning, identifiable as those of the pure chemical compounds present in the samples [15,16]. However, due to the high correlation and time overlap among the monoexponential signals of the pure components, the bilinear decomposition of the data, even under constraints, seldom results in unique solutions, which hinders the interpretation of the results [11].

As shown in Fig. 1A and B, a two-way dataset of decay curves can be transformed into a three-way data array through different tensorization approaches. These data arrays or tensors (say, generically, $\underline{\mathbf{D}}$), follow a trilinear model as defined in Eq. (1), [9].

$$\underline{\mathbf{D}} = \mathbf{C}(\mathbf{S}\odot\mathbf{K})^{\mathrm{T}} + \underline{\mathbf{E}} \qquad \text{Eq. 1}$$

where $\underline{\mathbf{D}}$ is sized $I\times J\times K$ and the N-components model is built from the data matrices **C** ($I\times N$), **K** ($K\times N$), and **S** ($J\times N$), with $\odot$ denoting the Khatri-Rao product. The matrix **C** contains the pure concentration profiles, **S** the related pure monoexponential decays and the matrix **K** is related to the way the initial two-way dataset of full decay curves is transformed into $\underline{\mathbf{D}}$. $\underline{\mathbf{E}}$ is the data array of residual variation unexplained by the model, sized $I\times J\times K$.

One of the most useful properties of trilinear data factorization is the uniqueness of the solutions obtained [9,11]. This property is especially useful to resolve datasets formed by mixtures of exponential decay curves, with high correlation among them and showing no selectivity. To perform the trilinear decomposition of the data array $\underline{\mathbf{D}}$, Parallel Factor Analysis-Alternating Least Squares (PARAFAC-ALS) [9] will be used to obtain the trilinear model given by the matrices **C**, **K** and **S**, which contain chemically meaningful profiles for the components in the system studied.

A parameter used to assess the fit quality of the PARAFAC-ALS model is the lack of fit (LOF) expressed as in Eq. (2).

$$\mathrm{LOF}\,(\%) = 100\times\sqrt{\frac{\Sigma_{i,j,k}e^2_{i,j,k}}{\Sigma_{i,j,k}d^2_{i,j,k}}} \qquad \text{Eq. 2}$$

where $d_{i,j,k}$ is the ijk $^{\mathrm{th}}$ element of $\underline{\mathbf{D}}$ and $e_{i,j,k}$ is the residual associated with the reproduction of $d_{i,j,k}$ by the trilinear model.

To assess the quality of the final trilinear model, the core consistency diagnostic (CORCONDIA) [17] can be used as an indicator parameter. CONCORDIA takes values from 100 (perfect fit by a trilinear model) to 0 or even negative (when the fitted trilinear model is less appropriate).

Despite the fact that PARAFAC-ALS has been the chosen algorithm in this work, it is important to note that the Kernelizing approach would increase the accuracy of the solutions provided by any other trilinear decomposition method.

3.1. Tensorization methods: slicing

Trilinear data arrays can be built from bilinear matrices by adequately reorganizing the original data, an example of so-called Data Tensorizing [10]. In this context, Pedersen et al. [4] proposed to use a slicing methodology to generate trilinear data from bilinear data based on the mathematical properties of exponential decays. Later, PowerSlicing [8], which consists of an optimal way to perform such slicing, was introduced. PowerSlicing and other slicing techniques exploit the invariance of exponential functions taken at different lags (see Eq. (3)):

$$Ae^{-\frac{t+\Delta t}{\tau}} = e^{-\frac{\Delta t}{\tau}}Ae^{-\frac{t}{\tau}} \qquad \text{Eq. 3}$$

When the independent variable t is lagged ($t+\Delta t$), the characteristic decay time τ of the exponential function (i.e., its shape, $e^{-\frac{t}{\tau}}$) remains unchanged, and only the value of the preexponential factor (A) gets modified.

If several slices, lagged Δt from one another, are extracted from the same decay (Fig. 1A), they will share the same τ but their respective preexponential factors will be different. Thus, linking Eq. (3) with Eq. (1), the preexponential factor A is related to the concentration mode (**C**), the $e^{-\frac{t}{\tau}}$ term is related to the shape of the monoexponential decay (**S**) and the new preexponential term $e^{-\frac{\Delta t}{\tau}}$ is related to the lag, which is in turn linearly related to the signal and becomes the new dimension **K**. The same occurs if the decay curve results from a mixture of exponential decays: individual τ are unchanged, but the preexponential factors for all the components get modified as previously described. Thanks to this mathematical property, slices can be arranged as a three-way data array that can be readily decomposed (Eq. (1)). For slicing methodologies, the profiles in **K** describe how the preexponential factors vary over the K slices, i.e., they define the lagging effect per component.

When applying PowerSlicing, we define the number of slices K to consider for the tensorization step according to the relation $\frac{J}{2}\geq 2^{K-1}$, being J the number of sampling points of the decays. The size of every slice is then $J-2^{K-1}+1$. For instance, for a matrix built from expo-

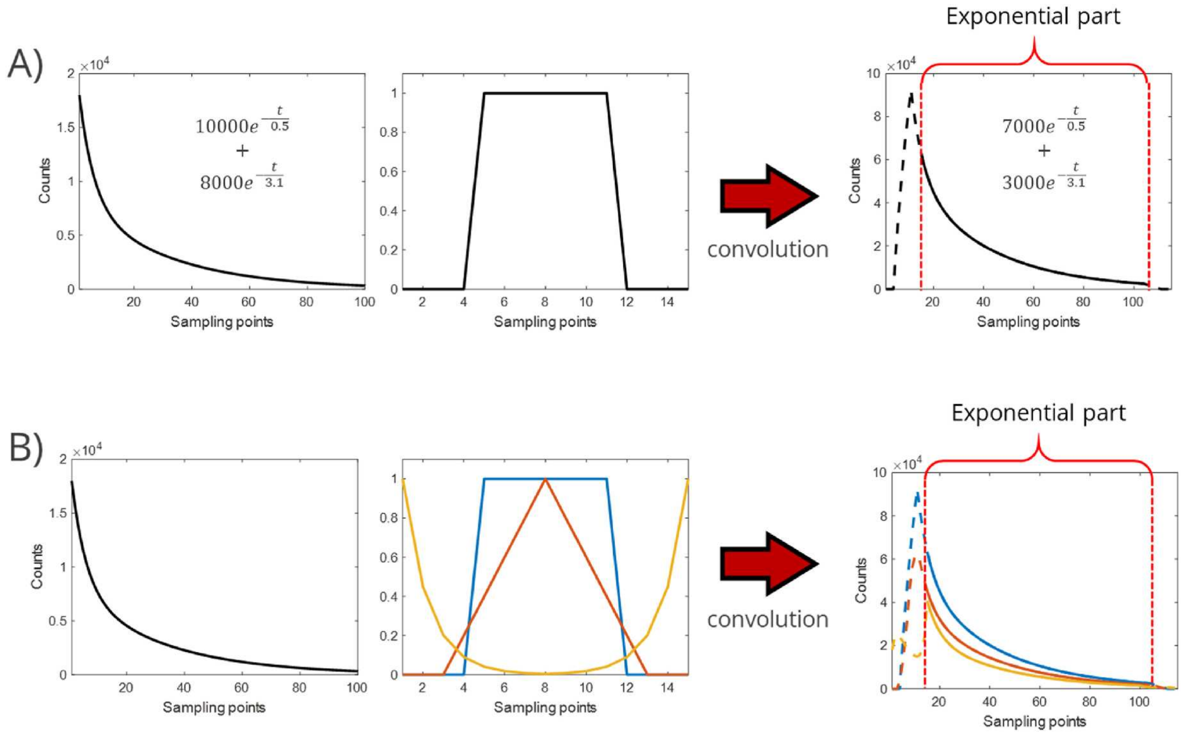


Fig. 2. Representation of a multiexponential decay resulting from the contributions of two monoexponential components (preexponential factors of 10 000 and 8000 and decay constant of 0.5 and 3.1, respectively) convolved with A) a single kernel (the preexponential factors change but the decay constants remain unchanged) and B) three kernels, yielding three new signal profiles, each of them characterized by different preexponential factors but the same decay constants. The dotted lines indicate the extremes of the convolved signals exhibiting non-exponential behaviors. Only the exponential part of the convolved signals is used for further trilinear decomposition analysis.

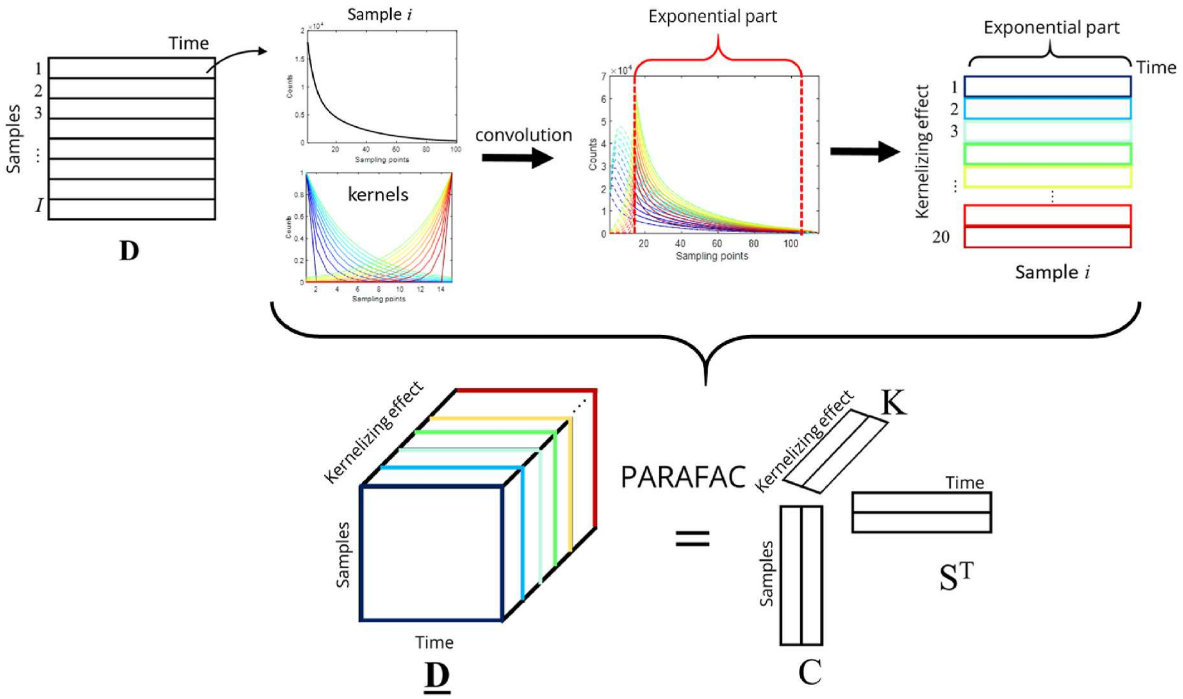


Fig. 3. Schematic representation of the Kernelizing approach. Each decay of the matrix **D** is convolved individually with a set of different kernels. Then, only the exponential parts of the resulting curves are kept and gathered into a matrix. After convolving every sample with the aforementioned kernels, a trilinear data array **D̲** can be built and subsequently analyzed by PARAFAC-ALS.

Table 2
Median of the correlation coefficients between simulated and recovered concentration profiles and of the decay constants recovered by PowerSlicing and Kernelizing in the 1000 analyzed runs. Values between brackets show the interval between the 2.5th and 97.5th percentiles of the corresponding metric. Median of CONCORDIA parameter across all runs is shown.

	True decay constant	Simulated dataset 1			
		PowerSlicing		Kernelizing	
		Concentration profile	Decay constant	Concentration profile	Decay constant
Component 1	0.80	0.98 [0.98, 0.99]	0.78 [0.76, 0.80]	0.99 [0.99, 0.99]	0.79 [0.78, 0.80]
Component 2	1.40	0.95 [0.91, 0.97]	1.39 [1.35, 1.43]	0.98 [0.97, 0.98]	1.40 [1.37, 1.43]
Component 3	2.4	0.96 [0.95, 0.97]	2.26 [2.13, 2.41]	0.98 [0.98, 0.99]	2.34 [2.26, 2.44]
CONCORDIA (%)		55		97	
	True decay constant	Simulated dataset 2			
		PowerSlicing		Kernelizing	
		Concentration profile	Decay constant	Concentration profile	Decay constant
Component 1	0.80	0.95 [0.94, 0.97]	0.76 [0.73, 0.78]	0.97 [0.96, 0.98]	0.77 [0.75, 0.79]
Component 2	1.40	0.86 [0.74, 0.92]	1.38 [1.30, 1.42]	0.94 [0.91, 0.96]	1.42 [1.37, 1.46]
Component 3	2.4	0.91 [0.87, 0.93]	2.16 [2.01, 2.35]	0.95 [0.93, 0.96]	2.27 [2.16, 2.41]
CONCORDIA (%)		24		77	
	True decay constant	Simulated dataset 3			
		PowerSlicing		Kernelizing	
		Concentration profile	Decay constant	Concentration profile	Decay constant
Component 1	0.80	0.92 [0.86, 0.94]	0.71 [0.60, 0.75]	0.94 [0.93, 0.95]	0.73 [0.69, 0.76]
Component 2	1.40	0.64 [0.56, 0.81]	1.29 [1.12, 1.40]	0.83 [0.72, 0.89]	1.41 [1.33, 1.47]
Component 3	2.4	0.83 [0.76, 0.88]	2.00 [1.86, 2.18]	0.89 [0.85, 0.91]	2.15 [2.04, 2.29]
CONCORDIA (%)		−2		24	

nential signals consisting of 1000 sampling points, nine slices can be generated, with 745 points each. If now only 15 sampling points are available, only three slices can be obtained, each of 12 sampling points. The effect of reducing the number of slices and the number of sampling points per slice on the stability of the trilinear model can be dramatic when few sampling points are available.

3.2. Tensorization methods: kernelizing

Kernelizing provides an alternative approach to generate trilinear data arrays from mixtures of monoexponential decay signals. It exploits the invariance of exponential functions resulting from the fact that by convolving a monoexponential decay function with any positive function (kernel), the preexponential factor gets modified but the characteristic decay time remains unchanged.

An example is shown in Fig. 2A, which illustrates this property. From left to right, an exponential signal of 100 sampling points generated from the sum of two monoexponential decays is convolved by a given kernel, with a width of 15 points. Once cut at the beginning and at the end (removing as many points as the size of the kernel window used), the resulting signal corresponds to a decay signal formed by the two same monoexponential components, but with different preexponential factors (for a mathematical proof and additional explanations see Eq. S1-7 in the Supplementary Material).

If a single decay curve is convolved with a set of K kernels (see Fig. 2B, $K = 3$), K new decays are obtained, characterized by different preexponential factors but the same decay constants. Then, if each decay curve (rows of $\mathbf{D}$) is convolved with the same set of K kernels, the K convolved decays obtained can be arranged into a three-way data array ($\underline{\mathbf{D}}$) and analyzed by PARAFAC-ALS to obtain the underlying monoexponential contributions associated with the individual components.

A schematic representation of the Kernelizing approach is shown in Fig. 3. I samples, characterized by different mixtures of two monoexponential decays, and a set of 20 kernels of different shapes, providing 20 new convolved decays per sample, are here considered. Since the preexponential factors of the components change along the kernel direction, a third linearly independent dimension (mode) is obtained, and, thus, $\underline{\mathbf{D}}$ results to be a trilinear array. Differently from slicing, this procedure remains applicable for signals characterized by very few sampling points.

It should be noted that kernel normalization is recommended to obtain convolved sets of decay curves with similar signal intensity for the subsequent PARAFAC-ALS analysis. Fig. 3 shows an example where kernels were normalized to a maximum value of 1.

Several features can be expected for the Kernelizing approach. First, the number of kernels used to build $\underline{\mathbf{D}}$ does not depend on the number of sampling points of the decay curves and, hence, a rich trilinear dataset, with no limited number of slabs can always be generated. The second feature is the denoising action associated with the Kernelizing approach, since signal convolution can be interpreted as a weighted moving average methodology, where every point in the convolved signal results from the sum of the signal points covered by the kernel window, weighted by the related kernel coefficients. The third feature is linked to the wide diversity of kernel shapes that can be chosen to emphasize different parts of the original decay signal. For example, by choosing a kernel corresponding to a monotonically decreasing function, the emphasis would be on the components with longer decay times, since the decay points at longer times will be weighted with a larger coefficient in the convolved signal, (see the darkest blue kernel and the related

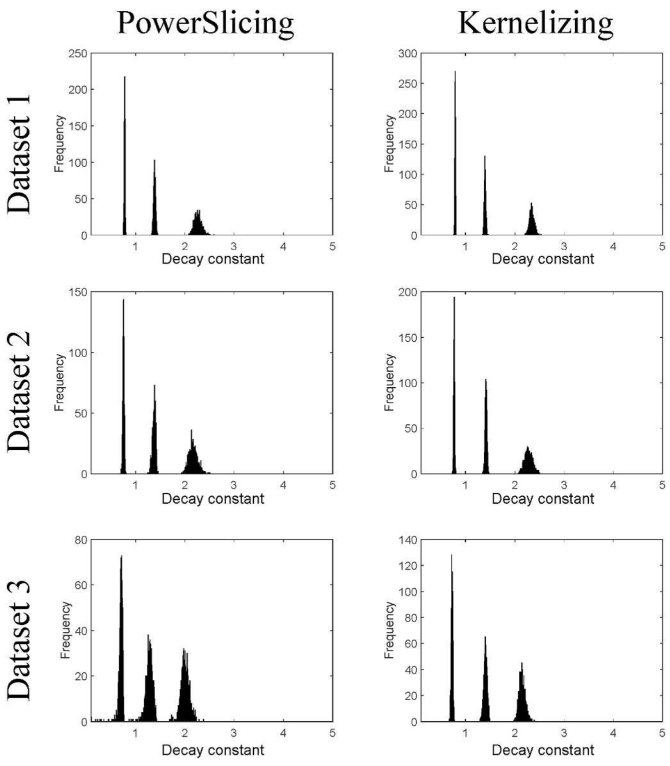


Fig. 4. Decay constants returned after the application of PowerSlicing and Kernelizing to the simulated datasets 1, 2 and 3 over the 1000 calculation runs. A significant reduction of the estimate scatter is observed for Kernelizing with respect to PowerSlicing when the number of sampling points is decreased.

convolved blue signal in Fig. 3). In our case, we have selected a set of kernels based on symmetric exponential functions to weight equally the short and long decay time components, as it is shown in Fig. 3. In general, it would be advisable to use a set of kernels that allows weighting in a similar way the entire range of points of the decay curves to be convolved.

4. Results and discussion

Kernelizing has been tested on simulated and time-resolved fluorescence experimental data acquired under controlled conditions. It has also been applied to a FLIM image for exploratory analysis. The results are discussed in the following sections.

4.1. Simulated examples

This subsection is divided in two parts. The first part aims at showing and discussing the performance of Kernelizing on simulated datasets with different numbers of sampling points (datasets 1, 2 and 3). The results are compared with those yielded by PowerSlicing. The second part aims at assessing the effect of key parameters of the Kernelizing approach on the results obtained.

4.2. Performance of the kernelizing approach

Datasets 1, 2 and 3 were simulated as explained in Section 2 and tensorized by both the Kernelizing and PowerSlicing approaches. The data arrays provided by Kernelizing were built by tensorizing the datasets 1, 2 and 3 using a set of 20 kernels, sized 20 × 250, 20× 25 and 20 × 8, respectively (in all cases, the kernel size is half the number of sampling points of the corresponding dataset). The typology and diversity of kernel functions were chosen according to the guidelines provided in Section 3 (see Fig. 3 and S3). The dimensions of the corresponding three-way data arrays are 200 × 20 × 250, 200 × 20 × 25 and 200 × 20× 7 for datasets 1, 2 and 3, respectively. In the case of PowerSlicing, considering the rules to define the number of slices and the sampling points therein, the dimensions of the three-way arrays generated are 200 × 8 × 373, 200 × 5 × 43 and 200 × 3 × 12, respectively.

PARAFAC-ALS was used to analyze the three-way data arrays resulting from both tensorization procedures. Initial estimates were set as the profiles yielded by the best-fitting models obtained after several PARAFAC-ALS runs started with a variety of initial estimates and fitted using only a few iterations [14].

To compare the results obtained by both the Kernelizing and PowerSlicing approaches, the individual pure component decay profiles in **S** were fitted by a monoexponential model to extract their respective characteristic decay constants and the correlation coefficients between the recovered concentration profiles in **C** and their corresponding ground truth profiles were calculated (see Table 2). The analysis was repeated 1000 times adding the same amount of Poisson-structured noise in each run for each dataset. To assess the accuracy and spread of the final solutions, the median and the 2.5th-97.5th percentile interval of these metrics were considered for every component. For each run, the CONCORDIA was calculated. The lack of fit of all models was found in agreement with the quantity of noise added to the simulated dataset (data not shown).

For the tensorized dataset 1 (500 sampling points), the values of the three decay constants are well recovered by both Powerslicing and Kernelizing. The full distribution of the decay constants is shown in Fig. 4. The exponential profiles and related decay constants are very well recovered for components 1 and 2, whereas a higher scatter, slightly more pronounced for Powerslicing, can be observed for component 3.

The correlation coefficients obtained for the concentration profiles are satisfactory for both approaches (equal or higher than 0.9 for all components), despite a slightly lower value for the concentration profile of component 2 returned by Powerslicing. Finally, it should be noted that a clear difference exists in the CONCORDIA values, 55% for PowerSlicing *vs* 97% for Kernelizing, meaning that the PARAFAC-ALS model yielded by Kernelizing is closer to an ideal trilinear model than the one yielded by PowerSlicing.

Overall, although Kernelizing provides slightly better results, both approaches guarantee a satisfactory performance when dealing with exponential decay signals for which a sufficiently number of sampling points is available.

When inspecting the results for datasets 2 and 3 (featuring 50 and 15 sampling points, respectively) the differences between PowerSlicing and Kernelizing become more pronounced (see Table 2). For the concentration profiles, Kernelizing provides correlation coefficients very close or higher than 0.9 in all components and datasets whereas the accuracy of Powerslicing decreases, as can be seen from the outcomes obtained for component 2 in dataset 2 and components 2 and 3 in dataset 3. Looking at the decay constants, correct decay constant values are recovered by Kernelizing for datasets 2 and 3 (with only a small bias for component 3 in the latter case). An additional relevant fact is that all Kernelizing models are very stable, the spread of the calculated decay constants does not change across datasets (see Fig. 4 and Table 2). Conversely, biased results are obtained for the decay constants of component 3 in dataset 2 and components 2 and 3 in dataset 3 when applying PowerSlicing. Additionally, the spread of the solutions is significantly larger when the number of sampling points decreases, which derives from the instability PARAFAC-ALS exhibits when taking into account reduced numbers of slices. Another indicator of the quality of the modelling approach is the CONCORDIA value. It can be observed that this parameter decreases for both approaches from dataset 1 to dataset 3, but the effect is more pronounced for PowerSlicing. All the differences mentioned above stem from the fact that reducing the number of sampling points heavily affects the amount of information that can be encoded in the three-way data array generated by Powerslicing since less slices can be built (five and

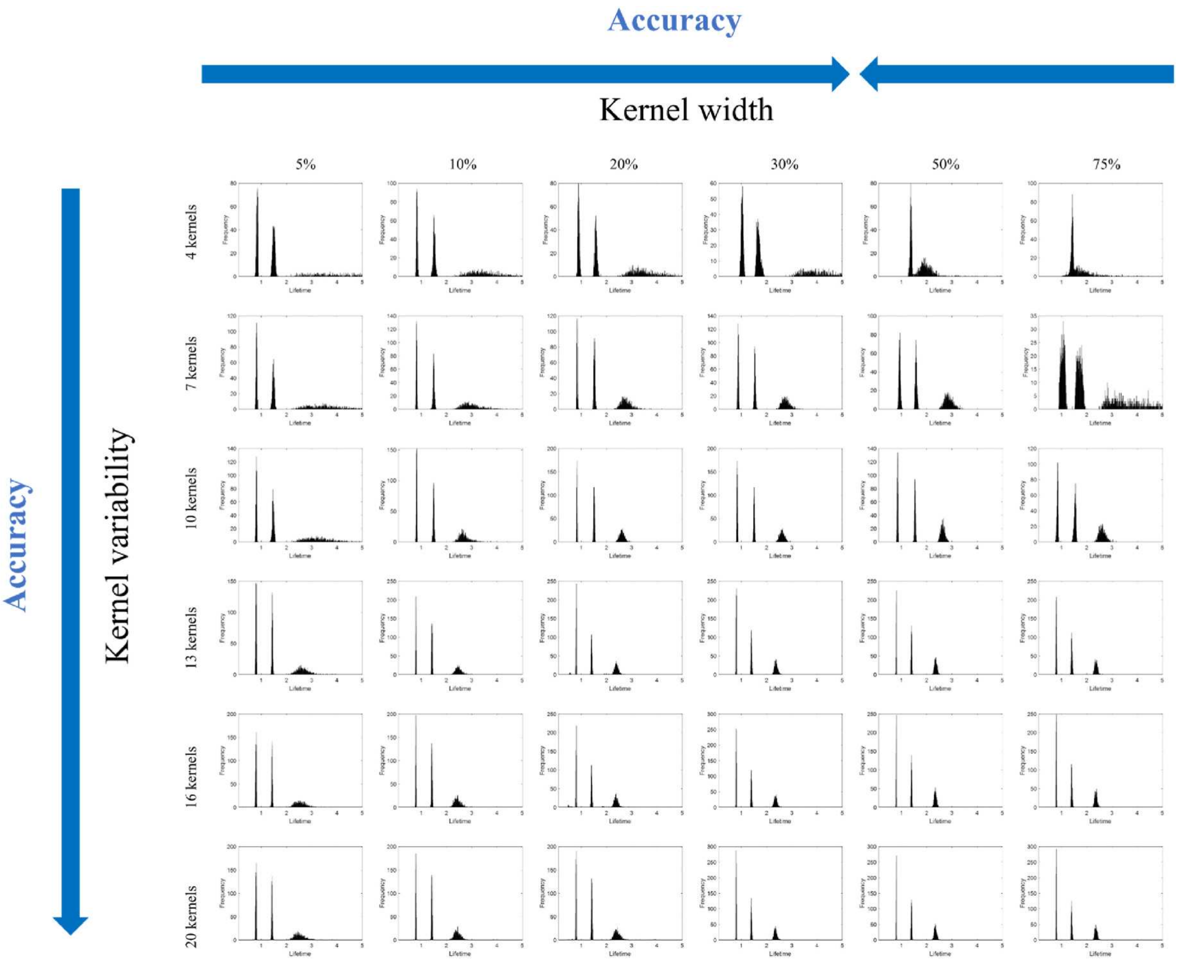


Fig. 5. Dispersion of the values of the decay constants recovered using Kernelizing on dataset 1. From top to bottom, the number and diversity of the kernel functions increase. From left to right, the width of the kernel functions (expressed as percentage of points over the total number of sampling points of the original decay curves) increases.

Table 3
Summary of results yielded by PowerSlicing and Kernelizing for the ATTO dataset in the two tested scenarios.

	True lifetime (ns)	ATTO data (1500 sampling points)				ATTO data (15 sampling points)			
		PowerSlicing		Kernelizing		PowerSlicing		Kernelizing	
		Concentration profile*	Lifetimes (ns) +	Concentration profile*	Lifetimes (ns)+	Concentration profile*	Lifetimes (ns) +	Concentration profile*	Lifetimes (ns) +
ATTO 655	1.9	0.75	1.786 [1.783–1.789]	0.76	1.887 [1.886–1.887]	0.70	1.86 [1.85–1.88]	0.72	1.93 [1.92–1.94]
ATTO 665	2.9	0.86	2.435 [2.432–2.438]	0.87	2.486 [2.486–2.486]	0.83	2.64 [2.62–2.66]	0.88	2.8 [2.79–2.83]
ATTO 647 N	3.5	0.90	3.419 [3.416–3.421]	0.99	3.494 [3.494–3.495]	0.77	3.65 [3.61–3.68]	0.86	3.74 [3.72–3.77]

+ Monoexponentially fitted lifetime and 95% confidence interval associated with the fitting error.

three for datasets 2 and 3, respectively). This, together with the decreasing number of sampling points per slice, hinders the correct recovery of monoexponential decays. For the Kernelizing approach, the number of sampling points per slice is also reduced, but the number of slabs remains unchanged (20 for all datasets). The shape diversity of the kernels is also preserved and, as a consequence of both facts, a three-way data array with large variability of information can still be generated in the worst-case scenario and satisfactory results can be obtained.

4.3. Effect of the kernel size and shape variation on the final solutions

In this section, we investigate several key features of the Kernelizing approach, such as the number of kernels used, their widths and shapes.

Six different sets of kernels were generated, enclosing from four to 20 exponential functions with different shapes, chosen so that the set of 20 kernels weights similarly short and long decay time components. The kernel sets used are shown in Fig. S3. In addition, for each set, six different kernel widths were considered (covering 5, 10, 20, 30, 50 and 75% of the total number of sampling points of the data). This results in a total of 36 different sets of kernels that were used to build the three-way data arrays corresponding to datasets 1, 2 and 3. For each combination of kernel shape and kernel width, PARAFAC-ALS analysis was repeated 1000 times adding to the data the same amount of Poisson-structured noise in each run.

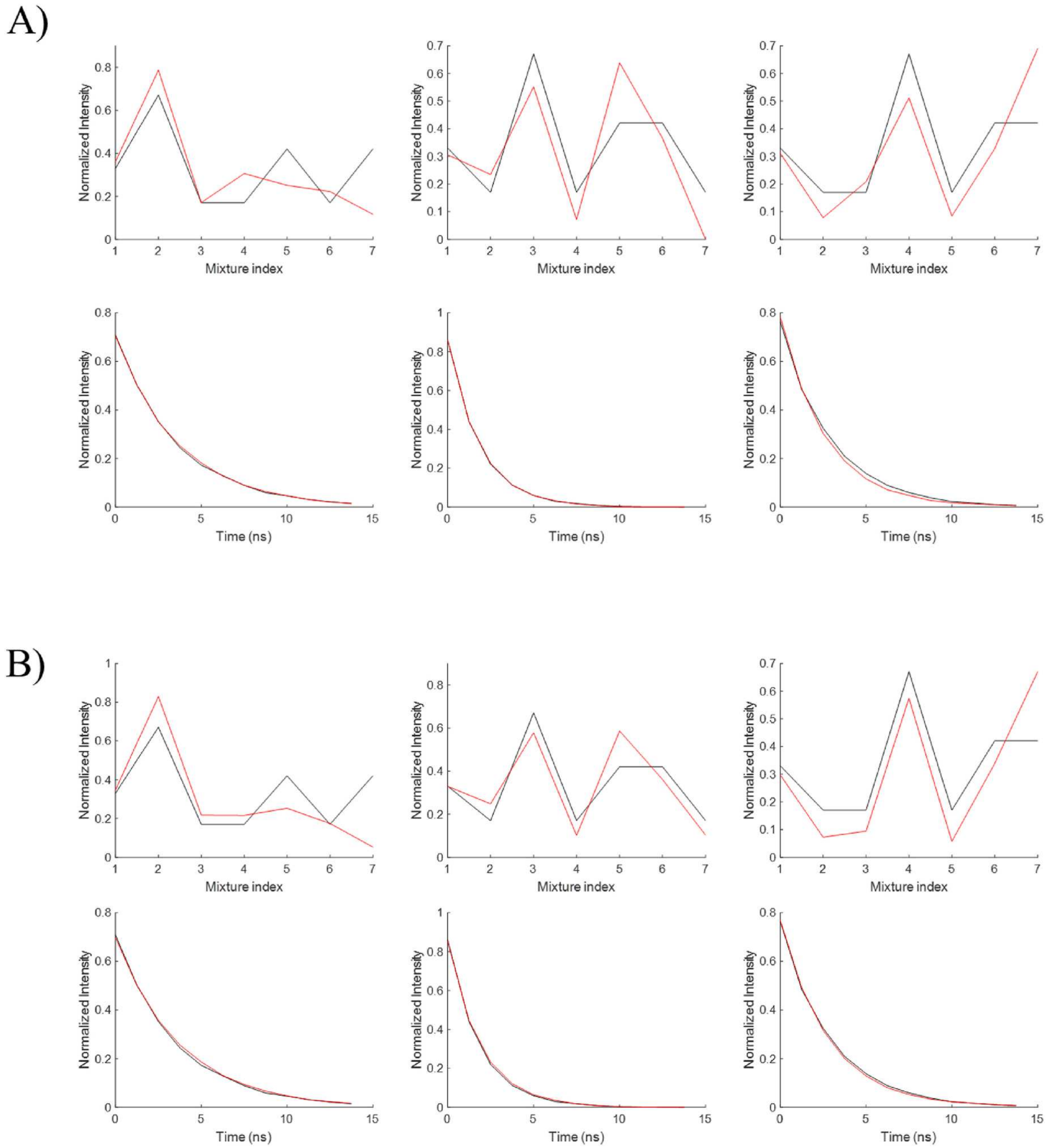


Fig. 6. ATTO dataset (15 sampling points). A) Top panel, concentration profiles recovered by PARAFAC-ALS for PowerSlicing (red) and expected concentration profiles (black). Bottom panel, pure fluorescence decays recovered by PARAFAC-ALS for PowerSlicing (red) and expected pure fluorescence decays (black). B) Top panel, concentration profiles recovered by PARAFAC-ALS for Kernelizing (red) and expected concentration profiles (black). Bottom panel, pure fluorescence decays recovered by PARAFAC-ALS for Kernelizing (red) and expected pure fluorescence decays (black). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Fig. 5 shows the decay constants obtained from the analysis of dataset 1. Looking at the results from top to bottom, it can be clearly observed that the higher the number of kernels and the wider the diversity of their shape, the closer to the ground truth the solutions are. For this reason, a key point is to always use a high number of kernels with very different shapes, since more variance is induced in the preexponential factors of all components, irrespective of their characteristic decay time (short or long).

Looking at the results in Fig. 5 from left to right, the effect of the kernel width can be assessed. In general, the larger the kernel, the more signal needs to be removed (see Fig. 3) and, thus, the lower the amount of available information exploitable for the resolution of strongly overlapping monoexponential components. On the other hand, when the kernel is too narrow, the variability induced in the preexponential factors is too small, and so is the denoising action, both effects resulting in a larger dispersion of the results. In practice, a compromise should be found and for the cases explored here, choosing a kernel width in the range between 20% and 50% of the total number of sampling points provided good results. However, different approaches may be advisable when coping with real-world datasets for which the ground truth is unknown, e.g., generating replicates by the noise addition method [18] and looking at the spread of the final results.

As a final conclusion, it is also important to understand the interaction between the effects of the key factors described above. When the number of kernels increases, the kernel width can be significantly reduced and, hence, the part of the convolved signals with non-

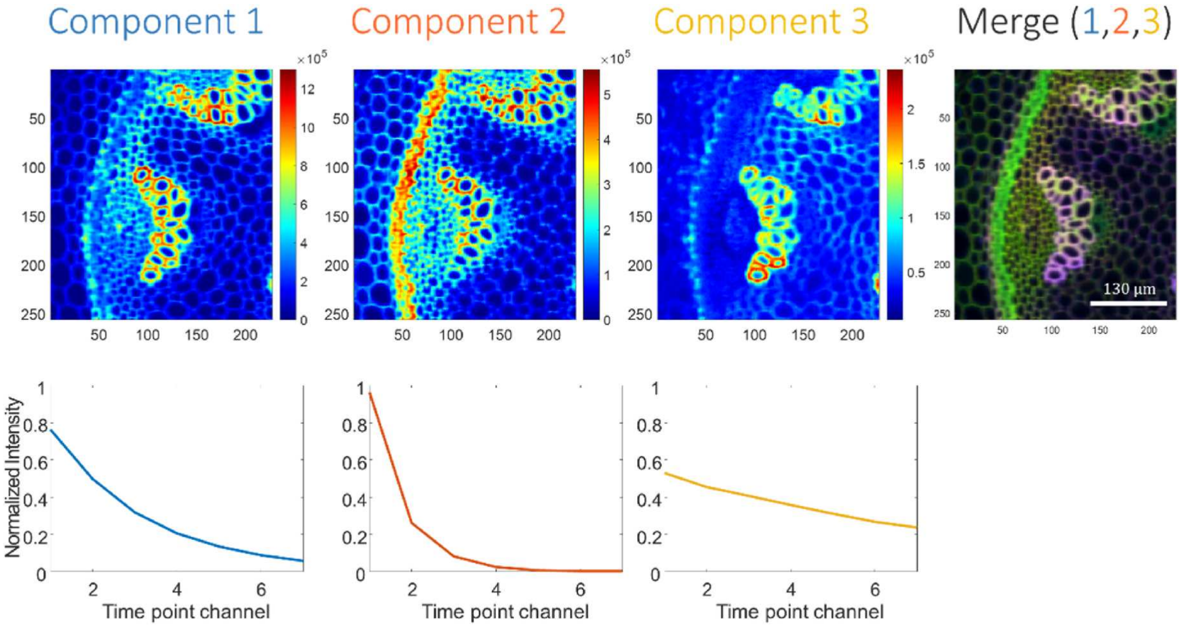


Fig. 7. FLIM dataset. Pure distribution maps (top) and pure fluorescence decays (bottom) recovered by Kernelizing-PARAFAC-ALS.

exponential behavior minimized.

The results obtained for datasets 2 and 3 provide the same general conclusions as for dataset 1 and are presented in the Supplementary Material (Figs. S4 and 5).

4.4. ATTO fluorescence data

The ATTO dataset, composed of fluorescence decays of seven ternary mixtures of three ATTO dyes sampled at 1500 points, was used to compare the results obtained by both Kernelizing and Powerslicing on real data. Kernelizing was applied using a set of 20 kernels of 300 point width (20%), resulting in a data array sized 7 × 20 × 1200. PowerSlicing was applied using 10 slices and the corresponding data size was 7 × 10 × 989. The tensorized data arrays were then decomposed by means of PARAFAC-ALS. After this, the resolved time profiles were fitted and the respective lifetimes extracted with their corresponding fitting error (95% confidence). The CONCORDIA parameter was also calculated.

Table 3 shows a summary of the results obtained. For full length signals, the concentration profiles and decay constants are, generally, well recovered by both approaches when they are compared with the ground truth (Fig. S6), despite the high correlation among the undelying monoexponential profiles and the very low number of samples handled. The decay constants of the components found for PowerSlicing and Kernelizing match well the expected ones. However, a small deviation for the component ATTO 665 can be observed for both approaches. The concentration profiles are also in good agreement with the true ones. Thus, in this scenario, both approaches are found to perform equivalently from a practical point of view, which is in line with the results obtained on simulated data for a large number of sampling points (dataset 1).

In a second step, the sampling points in the ATTO dataset were reduced to 15 (see Section 2). Kernelizing (20 kernels, 3 points width) was then applied to the resulting signals and a 7 × 20× 12 data array was obtained. PowerSlicing was applied to the same data using three slices which yielded a 7× 3× 12 data array. PARAFAC-ALS was used to analyze both datasets. Results are shown in Fig. 6. It can be observed that the concentration profiles and the decay constants retrieved by PARAFAC-ALS are more consistent with the ground truth when employing the Kernelizing approach rather than PowerSlicing, especially for components 2 and 3. This can be explained by the fact that only three slices are here available for Powerslicing as opposed to the 20 kernelized versions of the initial dataset, as previously observed for dataset 3. It is also important to notice that the error associated with the fitted lifetimes is higher for Powerslicing, indicating that the pure decay curves extracted may be a bit further from the pure monoexponential shapes expected.

4.5. Analysis of FLIM data

A FLIM image of *Convallaria majali* has also been investigated for illustrative purposes. The FLIM data was made available by Williams et al. [12]. The FLIM image has size 256 × 227 × 11 after preprocessing (see Section 2), where the first two dimensions are the *x*- and *y*-spatial directions and the third represents the 11 sampling points of the decay curve of every pixel, respectively. A set of 20 kernels of 3 time points each was used to tensorize the data, resulting in a dataset sized 256 × 227 × 20 × 8. The dataset was decomposed (after unfolding along the pixel direction) by a three-component PARAFAC-ALS model. The concentration profiles obtained were refolded to recover the 2D concentration maps of every component. Fig. 7 shows the pure component concentration maps and the related pure monoexponential decays obtained.

As can be observed, despite no ground truth is available, the three resolved concentration maps highlight quite specific biological zones of the vegetal tissue. A tentative assignment would be the following: component 1 relates to the xylem and could correspond to the safranin dye linked to lignin. Besides, the endodermis is highlighted. The safranin dye stains lignin, which is generally located on the xylem and the endodermis [19]. Component 2 (fast decay) generally appears in the mesophyll cells as well as in some lignified cells and might be related to the fast-green stain. It has been reported that fast-green stains well the phloem and cellulosic cell walls in the pith [19]. On the other hand, component 3 (slow decay) appears specifically on the xylem, differentiating multiple environments for lignin. These results are in agreement with Kaminski et al. [19], who characterized a similar sample by means of excitation-emission spectroscopy.

As all the investigated components have biological sense and were satisfactorily recovered in spite of the high complexity of the case-study dealt with, this last example shows the potential of the Kernelizing approach for the analysis of FLIM images of vegetal tissues featuring a low number of sampling points.

5. Conclusions

In this work, Kernelizing is proposed as a very efficient way to obtain trilinear data arrays with a high degree of variability from bilinear data matrices of multiexponential decays. The richness of information encoded in these kernelized three-way arrays allows PARAFAC-ALS resolving chemical mixtures into individual components characterized by monoexponential decays with an equal or higher accuracy than well-established slicing approaches (such as PowerSlicing).

Thus, although PowerSlicing is a fast and robust method that can serve the same purpose as Kernelizing in most practical situations, we have identified specific scenarios for which the robustness of the PowerSlicing solutions can be questioned, i.e., situations for which very few slices can be obtained because the number of sampling points in the original multiexponential curves is low. Such a problem does not affect the proposed Kernelizing approach since the number of convolved decays that can be generated for each sample signal is not limited by the number of sampling points. As has been proven, the possibility to choose the number and shapes of the kernels used is an excellent asset to increase the variability in the three-way arrays to be analyzed and, consequently, the accuracy and precision of the solutions obtained.

Kernelizing has been found very useful to handle multiexponential measurements for which binning is required to increase the signal-to-noise ratio or whose number of sampling points is low due to instrumental limitations. Fluorophores characterized by very similar decaying behaviors, for example, could be unmixed in a FLIM imaging case-study, where the number of sampling points was limited due to specific features of the instrument resorted to.

At this point, the main aspect that needs further exploration is the choice of the kernel width. In this study, a trial-and-error approach was utilized, but we acknowledge that a more systematic strategy would be useful, e.g., defining the width of the kernels based on the characteristics of the dataset (number of sampling points, noise, etc.) would further simplify the generalized use of the Kernelizing approach.

Finally, it is worth pointing out that the results and conclusions drawn in this article mainly relate to the PARAFAC analysis of TRFS datasets, but can be generalized to any measurement that can be expressed by multiexponential decay curves and to any algorithm devoted to perform trilinear decomposition analysis.

CRediT authorship contribution statement

Adrián Gómez-Sánchez: Conceptualization, Methodology, Software, Formal analysis, Investigation, Discussion, Data curation, Writing – original draft, Visualization. **Raffaele Vitale:** Conceptualization, Discussion, Writing – review & editing. **Olivier Devos:** Data acquisition, Discussion, Writing – review & editing. **Anna de Juan:** Resources, Writing – original draft, Discussion, Supervision, Funding acquisition. **Cyril Ruckebusch:** Conceptualization, Methodology, Formal analysis, Resources, Writing – original draft, Discussion, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare no competing interests.

Data availability

Data will be made available on request.

Acknowledgements

A. G.-S. and A.J. acknowledge financial support from the Spanish government Project PID 2019-1071586 B-IOO and the Catalan government (2021 SGR 00449). A. G.-S. acknowledges scholarships from the MOBLILEX U Lille program and from the Santander bank.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2023.341545.

References

[1] H. Lemmetyinen, N.V. Tkachenko, B. Valeur, J.I. Hotta, M. Ameloot, N.P. Ernsting, T. Gustavsson, N. Boens, Time-resolved fluorescence methods (IUPAC technical report), Pure Appl. Chem. 86 (12) (2014) 1969–1998.
[2] A.A. Istratov, O.F. Vyvenko, Exponential analysis in physical phenomena, Rev. Sci. Instrum. 70 (2) (1999) 1233–1257.
[3] W. Windig, B. Antalek, Direct exponential curve resolution algorithm (DECRA): a novel application of the generalized rank annihilation method for a single spectral mixture data set with exponentially decaying contribution profiles, Chemometr. Intell. Lab. Syst. 37 (2) (1997) 241–254.
[4] H.T. Pedersen, R. Bro, S.B. Engelsen, Towards rapid and unique curve resolution of low-field NMR relaxation data: trilinear SLICING versus two-dimensional curve fitting, J. Magn. Reson. 157 (1) (2002) 141–155.
[5] T. Ohno, Z. Wang, R. Bro, PowerSlicing to determine fluorescence lifetimes of water-soluble organic matter derived from soils, plant biomass, and animal manures, Anal. Bioanal. Chem. 390 (8) (2008) 2189–2194.
[6] O. Devos, M. Ghaffari, R. Vitale, A. de Juan, M. Sliwa, Ruckebusch, Multivariate curve resolution slicing of multiexponential time-resolved spectroscopy fluorescence data, Anal. Chem. 93 (37) (2021) 12504–12513.
[7] D. Cevoli, S. Hugelier, R. Van den Eynde, O. Devos, P. Dedecker, C. Ruckebusch, Multilinear Slicing for curve resolution of fluorescence imaging with sequential illumination, Talanta 241 (2022), 123231.
[8] S.B. Engelsen, R. Bro, PowerSlicing, J. Magn. Reson. 163 (1) (2003) 192–197.
[9] R. Bro, PARAFAC. Tutorial and applications, Chemometr. Intell. Lab. Syst. 38 (2) (1997) 149–171.
[10] O. Debals, L.D. Lathauwer, August. Stochastic and deterministic tensorization for blind signal separation, in: International Conference on Latent Variable Analysis and Signal Separation, Springer, Cham, 2015, pp. 3–13.
[11] R. Tauler, A. Smilde, B. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, J. Chemometr. 9 (1) (1995) 31–58.
[12] G.O. Williams, E. Williams, N. Finlayson, A.T. Erdogan, Q. Wang, S. Fernandes, A. R. Akram, K. Dhaliwal, R.K. Henderson, J.M. Girkin, M. Bradley, Full spectrum fluorescence lifetime imaging with 0.5 nm spectral and 50 ps temporal resolution, Nat. Commun. 12 (1) (2021) 1–9.
[13] W. Becker, Fluorescence lifetime imaging–techniques and applications, J. Microsc. 247 (2) (2012) 119–136.
[14] The N-Way Toolbox for MATLAB Version 3.31, 2022, 18/11, http://www.models.life.ku.dk/nwaytoolbox/download. Last access: 9/6/2023.
[15] A. de Juan, R. Tauler, Multivariate curve resolution-alternating least squares for spectroscopic data, Data Handling Sci. Technol. 30 (2016) 5–51 (Elsevier).
[16] S. Benabou, C. Ruckebusch, M. Sliwa, A. Avino, R. Eritja, R. Gargallo, A. de Juan, Study of conformational transitions of i-motif DNA using time-resolved fluorescence and multivariate analysis methods, Nucleic Acids Res. 47 (13) (2019) 6590–6605.
[17] R. Bro, H.A. Kiers, A new efficient method for determining the number of components in PARAFAC models, J. Chemometr.: A Journal of the Chemometrics Society 17 (5) (2003) 274–286.
[18] J. Jaumot, R. Gargallo, R. Tauler, Noise propagation and error estimations in multivariate curve resolution alternating least squares using resampling methods, J. Chemometr.: A Journal of the Chemometrics Society 18 (7-8) (2004) 327–340.
[19] C.F. Kaminski, R.S. Watt, A.D. Elder, J.H. Frank, J. Hult, Supercontinuum radiation for applications in chemical sensing and microscopy, Appl. Phys. B 92 (3) (2008) 367–378.

Supplementary material

Kernelizing: a way to increase accuracy in trilinear decomposition analysis of multiexponential signals

Adrián Gómez-Sánchez[1,2,*], Raffaele Vitale[2], Olivier Devos[2], Anna de Juan[1], Cyril Ruckebusch[2]

[1]Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028, Barcelona, Spain

[2] Univ. Lille, CNRS, UMR 8516 - LASIRe - Laboratoire Avancé de Spectroscopie pour les Intéractions la Réactivité et l'Environnement, F-59000 Lille, France

[*]Corresponding author

Abstract

This supporting material includes the mathematical proof that the tensorization behind the Kernelizing approach yields trilinear data and the results of the study of the impact that the number of kernels and their width have on the final solutions obtained for the simulated datasets 2 and 3.

1. The Kernelizing tensorization yields trilinear data

Let us consider the fluorescence monoexponential decay in Eq. S1

$$f(t) = Ae^{\frac{-t}{\tau}}, \qquad \text{Eq. S1}$$

where A is the preexponential factor, τ is the lifetime, t is the time variable and $f(t) = 0$ for $t < 0$. The Kernelizing tensorization is based on the invariability of the decay constant τ when convolving an exponential decay with a function $g(t)$. The convolution of an exponential decay is defined as in Eq. S2

$$h(t) = \int_{-\infty}^{+\infty} Ae^{\frac{-(t-t_o)}{\tau}} g(t_o)\, dt_o. \qquad \text{Eq. S2}$$

where $h(t)$ is the convolved function and $g(t)$ is a non-negative finite function with positive entries in the interval $[0, m]$, here called kernel.

For calculating the convolution in Eq. S2, three cases should be considered

Case 1)

$$t < 0 \rightarrow \int_{-\infty}^{0} Ae^{\frac{-(t-t_o)}{\tau}} g(t_o)\, dt_o = 0. \qquad \text{Eq. S3}$$

Case 2)

$$0 \leq t < m \rightarrow \int_{0}^{t} Ae^{\frac{-(t-t_o)}{\tau}} g(t_o)\, dt_o = Ae^{\frac{-t}{\tau}} \int_{0}^{t} e^{\frac{t_o}{\tau}} g(t_o) dt_o. \qquad \text{Eq. S4}$$

Case 3)

$$t \geq m \rightarrow \int_{0}^{m} Ae^{\frac{-(t-t_o)}{\tau}} g(t_o)\, dt_o = Ae^{\frac{-t}{\tau}} \int_{0}^{m} e^{\frac{t_o}{\tau}} g(t_o)\, dt_o. \qquad \text{Eq. S5}$$

For the objectives of our work, Case 3 is the interesting one, where it can be observed that, after $t \geq m$, the recovered function is a scalar $\int_{0}^{m} e^{\frac{t_o}{\tau}} g(t_o)\, dt_o$ multiplied by the original fluorescence decay $Ae^{\frac{-t}{\tau}}$. This means that, when an exponential signal is convolved by any kernel, the exponential behavior is recovered for $t \geq m$. This property is what Kernelizing exploits for generating trilinear data.

The equations described above can be generalized for datasets formed by multiple samples and multiple monoexponential decays (components). In this case, Eq. S5 becomes Eq. S6.

$$\int_{-\infty}^{+\infty} \left(A_{i_1} e^{\frac{-(t-t_o)}{\tau_1}} + A_{i_2} e^{\frac{-(t-t_o)}{\tau_2}} + \ldots + A_{i_N} e^{\frac{-(t-t_o)}{\tau_N}} \right) g_k(t_o)\, dt_o = \int_{-\infty}^{+\infty} \left(\sum_{n=1}^{N} A_{i_n} e^{\frac{-(t-t_o)}{\tau_n}} \right) g_k(t_o)\, dt_o. \qquad \text{Eq. S6}$$

where A_{in} is the preexponential of the sample i of component n , k the number of kernel and N the number of components, then:
The integral in Eq. S6 can be split in N parts by linearity of integration

$$\int_{-\infty}^{+\infty} \left(A_{i_1} e^{\frac{-(t-t_o)}{\tau_1}} \right) g_k(t_o)\, dt_o + \int_{-\infty}^{+\infty} \left(A_{i_2} e^{\frac{-(t-t_o)}{\tau_2}} \right) g_k(t_o)\, dt_o + \ldots + \int_{-\infty}^{+\infty} \left(A_{i_N} e^{\frac{-(t-t_o)}{\tau_N}} \right) g_k(t_o)\, dt_o. \qquad \text{Eq. S7}$$

It can be noted that, each integral term in Eq. S7, being of the same form as Eq. S2, can be rewritten as in Eq. S5 for $t \geq m$, resulting in

$$\sum_{n=1}^{N} \int_{0}^{m} \left(A_{i_n} e^{\frac{-(t-t_o)}{\tau_n}} \right) g_k(t_o)\, dt_o = \sum_{n=1}^{N} A_{i_n} e^{\frac{-t}{\tau_n}} \int_{0}^{m} e^{\frac{t_o}{\tau_n}} g_k(t_o)\, dt_o. \qquad \text{Eq. S8}$$

Considering each of the N component individually, the argument of the summation in Eq. S8 can be written as a triad of vectors corresponding to sample, time and kernel, respectively, as illustrated in Figure S1.

$$A_{i_n} e^{\frac{-t}{\tau_n}} \int_0^m e^{\frac{t_o}{\tau_n}} g_k(t_o)\, dt_o$$

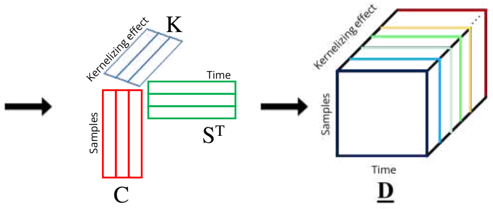


Figure S1. Triadic representation of Eq. S8. Each colored term contributes linearly to the data. Therefore, the combination of the matrices **C**, **K** and $\mathbf{S}^{\mathrm{T}}$ yields trilinear data.

2. Uniqueness of the PARAFAC-ALS solutions yielded after Kernelizing

A test was performed in order to detect if the high correlation inherent to multiexponential data can jeopardize the uniqueness of a PARAFAC-ALS solution yielded after the application of Kernelizing. If this solution is unique, PARAFAC-ALS, independently of how its initialization parameters are set, should always converge to the same solution. To show this, a single version of dataset 1 was analyzed by PARAFAC-ALS 100 times changing iteratively the initial estimates input to its algorithm (orthogonalized random profiles). All runs converged to an identical solution, meaning that data are full-rank and uniqueness is not lost (Fig. S2).

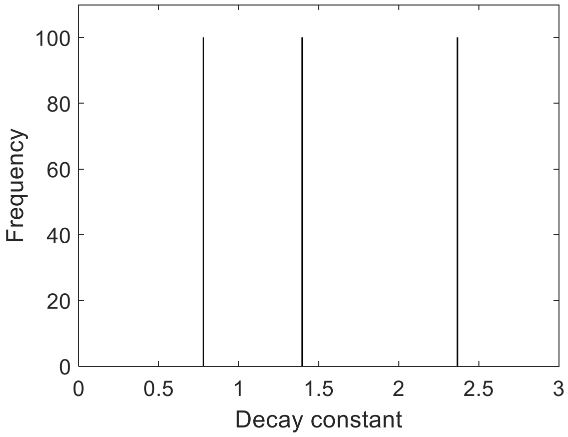


Figure S2. Decay constants estimated over 100 calculation runs performed applying Kernelizing and PARAFAC-ALS to an individual version of the simulated dataset 1 and inputting different random initial estimates to the PARAFAC-ALS algorithm. No dispersion was found.

3. Example of kernel sets for Kernelizing the Dataset 1

Fig. S3 shows the shape of some of the kernel functions used for this particular assessment.

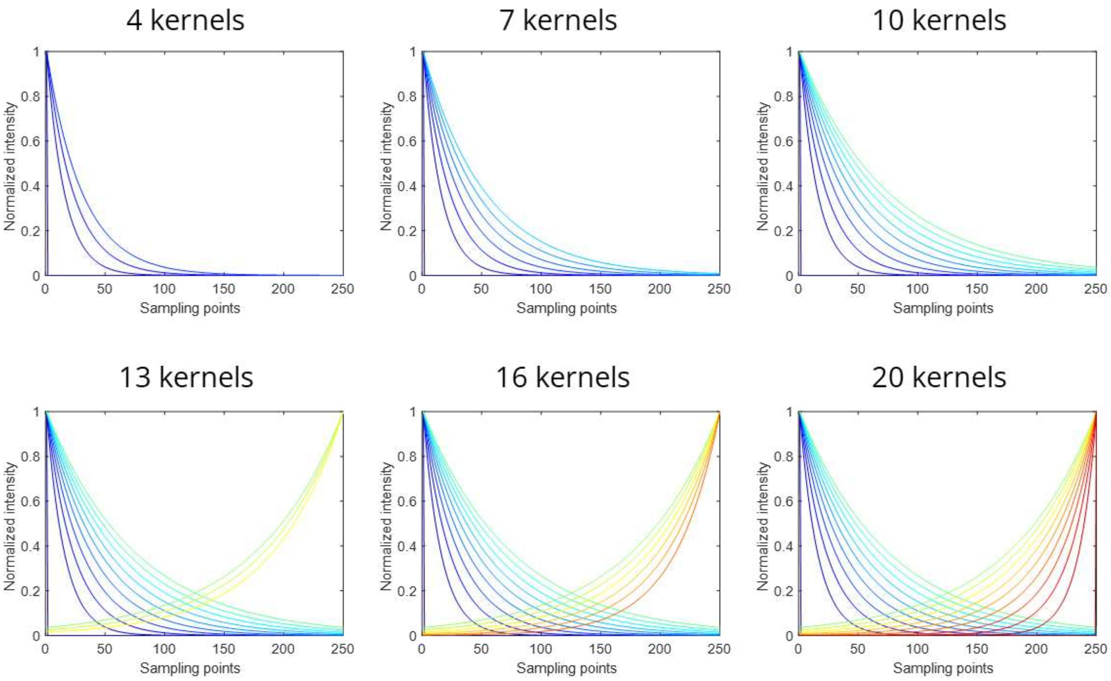


Figure S3. Example of kernel functions characterized by a width equal to the 50% of the total number of sampling points of dataset 1 (250 out of 500 sampling points).

4. Effect of the size and the variation of the kernel set on the solutions for Datasets 2 and 3

Fig. S4 and S5 provide the results obtained from the application of Kernelizing to datasets 2 and 3 containing exponential signals sampled over 50 and 15 sampling points, respectively. For these cases, the kernel sets used for the analysis of Dataset 1 (Fig. S3) were downsampled in the same proportion to preserve their shape. For dataset 3, 5% width kernels were not considered because this would have implied using kernels of less than two points. The kernel set containing only four functions was not considered either because it would have yielded a too low accuracy.

Focusing first on the results obtained when increasing the number of kernels, outcomes very similar to the ones obtained for dataset 1 are observed. On the other hand, the effect of increasing the diversity and the number of kernels is more significant for dataset 2 and 3 than for dataset 1. As can be seen in Fig. S4, for dataset 2, a reduction of the bias as well as an increase of the precision of the solutions can be clearly for all tested kernel widths when the number of kernels increased from 4 to 20 (column direction). Analogous results are also obtained for dataset 3 (Fig. S5). This is due to the fact that, increasing the number of kernels with different shapes also increases the variability in the convolved curves constituting the analyzed cube. When the decay curves under study feature a large number of sampling points (as in dataset 1), though, the information in the three-way array tends to be sufficiently discriminating even if a small number of kernels (or slices) is used; instead, the poorly defined decay curves in datasets 2 and 3 do need a larger number of kernels (or slices) with varied information to describe sufficiently well the underlying monoexponential decay profiles to be retrieved.

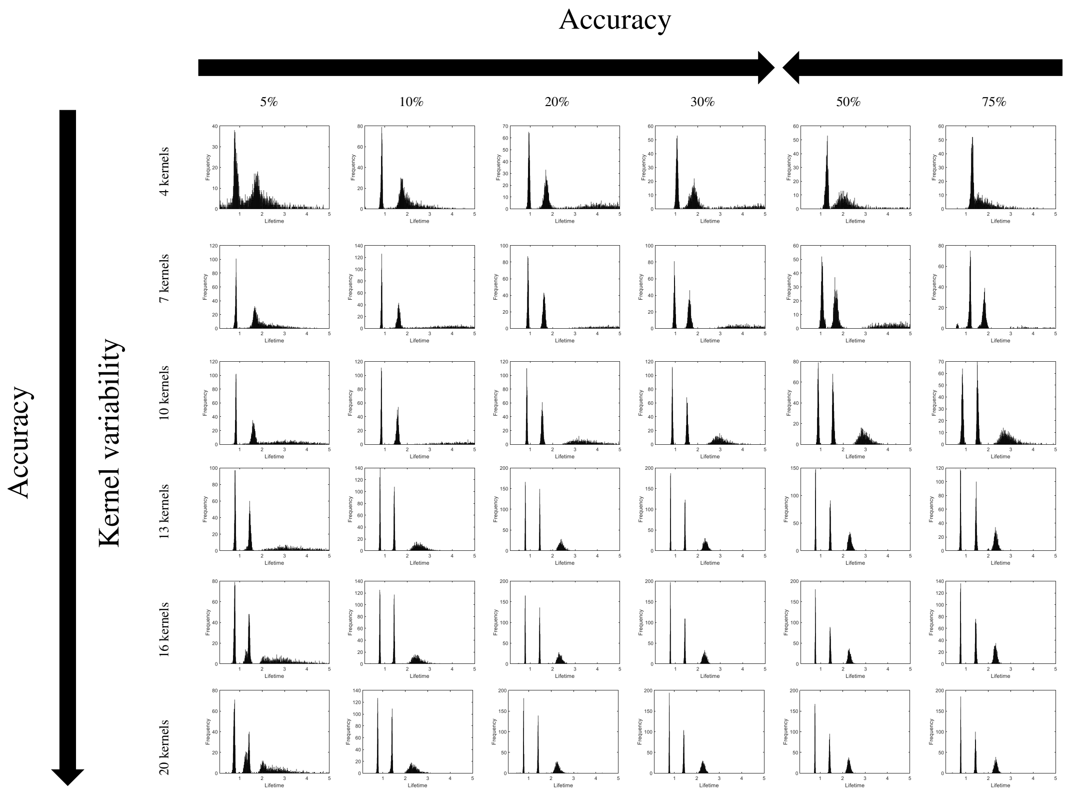


Figure S4. Dispersion of the decay constants estimated using Kernelizing on dataset 2.

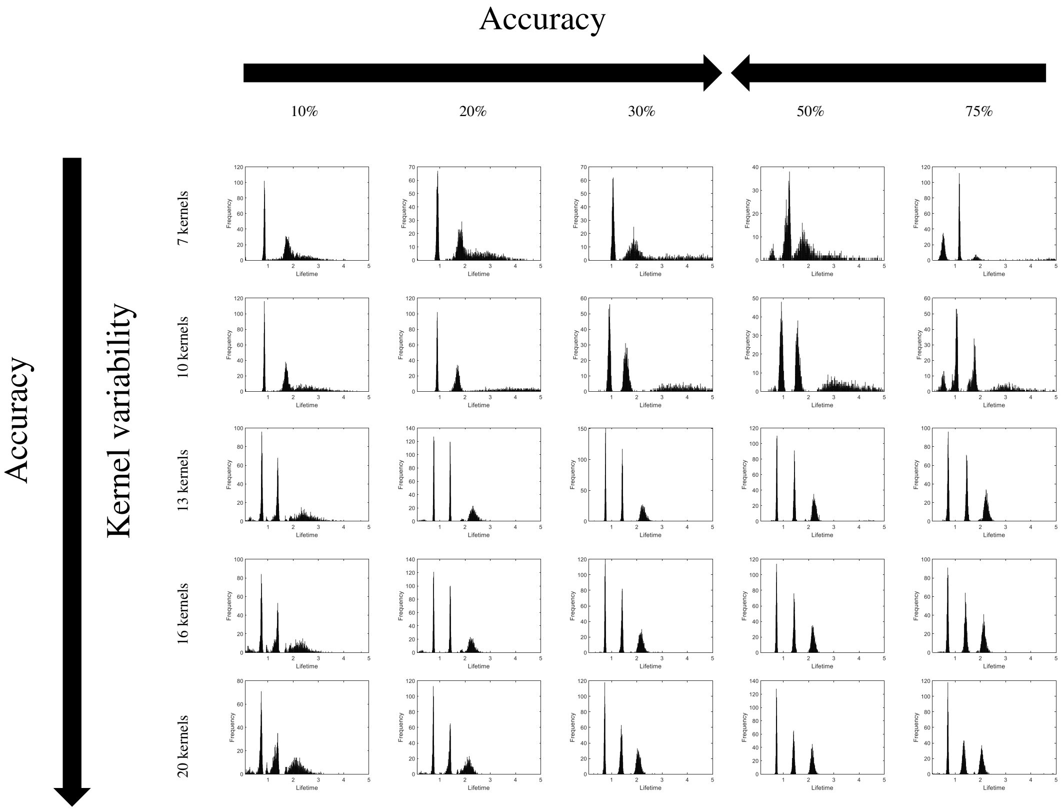


Figure S5. Dispersion of the decay constants estimated using Kernelizing on dataset 3.

5. ATTO fluorescence data - High number of sampling points

Fig. S6 provide the results obtained from the application of PowerSlicing and Kernelizing to the ATTO fluorescence data with a high number of sampling points.

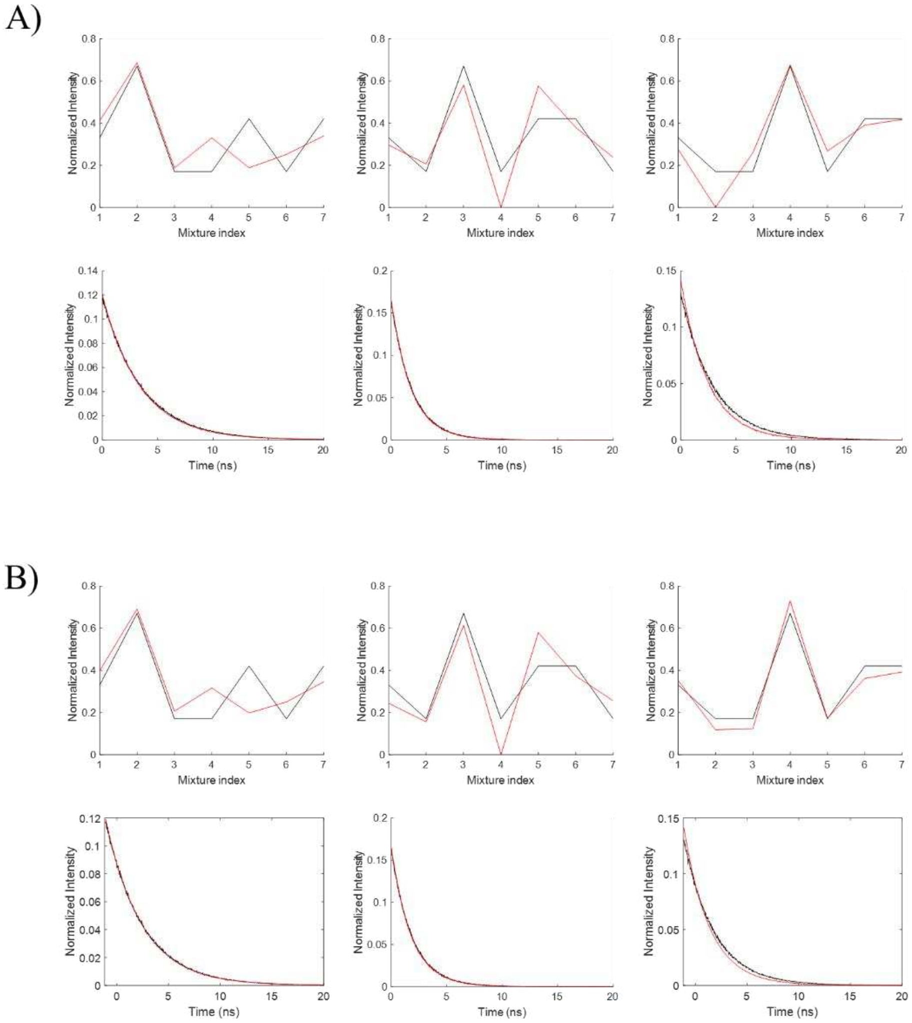


Figure S6. ATTO dataset (1500 sampling points). A) Top panel, concentration profiles recovered by PARAFAC-ALS for PowerSlicing (red) and expected concentration profiles (black). Bottom panel, pure fluorescence decays recovered by PARAFAC-ALS for PowerSlicing (red) and expected pure fluorescence decays (black). B) Top panel, concentration profiles recovered by PARAFAC-ALS for Kernelizing (red) and expected concentration profiles (black). Bottom panel, pure fluorescence decays recovered by PARAFAC-ALS for Kernelizing (red) and expected pure fluorescence decays (black).

Both **Publications IV** and **V** are based on the idea of the invariability of the shape of exponential signals through the convolution process. This shape invariability is key in the proposal of a mathematical tool to extract the IRF from an exponential decay or in the unmixing of TRFS multiexponential data into their pure monoexponential profiles. The details of this mathematical property are discussed here to show how it provides the basis for the Blind Instrument Response Function Identification (BIRFI) and Kernelizing methodologies.

The invariability of the shape of exponential signals across the convolution: the core of BIRFI and Kernelizing approaches

Studying mathematically the nature of the fluorescence decay signals helps to understand some limitations in the analysis of this measurement and to exploit mathematical properties to improve the results obtained. Assuming that the fluorescence decay follows an exponential function (Eq. 21), two important properties of this kind of signal need to be taken into account.

The first property is, as stated in Chapter 2 (see Eq. 3), that the measured fluorescence decay can be expressed as the convolution of the pure exponential decay by the IRF. The second property is that if an exponential decay is convolved with a function (called kernel), the resulting signal keeps the original decay constants, while only changing the preexponential factors.

Both convolution cases are mathematically represented as the integral of the product of the exponential decay and the convolving function after one is reversed and shifted δ (Eq. 22). In this context, the function that convolves the exponential decay $x(t)$ is named $g(t)$. Here, $g(t)$ must be understood as the IRF in **Publication IV** and as any kernel in **Publication V**.

Equation 23 shows the analytical solution $y(t)$ for any exponential decay $x(t)$ convolved by any kernel $g(t)$. To solve analytically the convolution process, the problem is split in three domains: for $t \leq 0$, for $0 < t < m$ and finally, for $m \leq t$ where m is the time range covered by the convolving function, e.g., if a kernel function covers five time points, $m = 5$.

$$x(t) = Ae^{\frac{-t}{\tau}} \qquad \text{Eq. 21}$$

$$y(t) = IRF * x(t) = \int_{-\infty}^{+\infty} x(t-\delta)g(\delta)\, d\delta \qquad \text{Eq. 22}$$

$$y(t) = \begin{cases} 0, & \text{Domain 1: } t \leq 0 \\ B(t) \cdot Ae^{\frac{-t}{\tau}}, & \text{Domain 2: } 0 < t < m \\ B \cdot Ae^{\frac{-t}{\tau}}, & \text{Domain 3: } m \leq t \end{cases} \qquad \text{Eq. 23}$$

Figure 29 illustrates the different regions defined in Eq. 23 for the convolved signal $y(t)$.

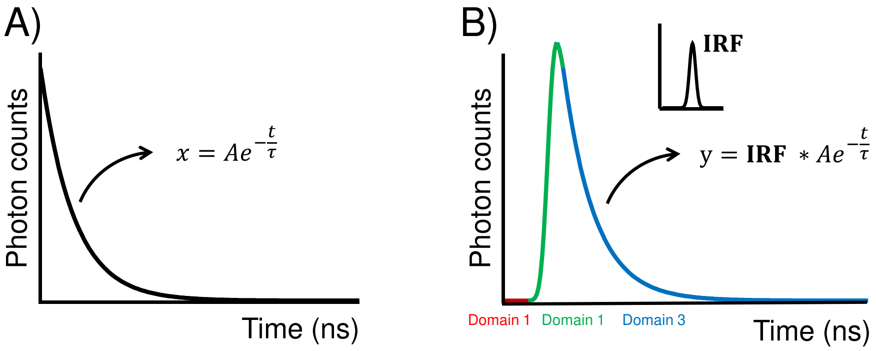


Figure 29. A) An exponential decay $x(t)$ and B) the signal $y(t)$, issued from the convolution of $x(t)$ with the IRF, $g(t)$. To solve the convolution, the problem is split in three domains, for which the shape is defined in Eq. 23. The blue shape (domain 3) is equal to $x(t)$, but scaled in intensity.

Domain 1 determines the shape before the signal starts to grow. Domain 2 defines the shape of the non-exponential part of the convolved exponential decay, where $B(t)$ changes as a function of time. Domain 3 follows a pure exponential shape and it is the most interesting part of the signal both for the estimation of IRF in the BIRFI algorithm of **Publication IV** and for the kernelizing approach of **Publication V**.

Domain 3 explicitly shows that after the point m, i.e., after the time range covered by $g(t)$, $y(t)$ recovers the exponential behavior of $x(t)$ albeit there is a scale effect in the signal shape, represented by *B.* That is, the shape of $x(t)$ remains invariant when it is convolved by any function beyond time m. Based on this invariability, the BIRFI algorithm extracts the IRF from $y(t)$ by combining the exponential tail, which plays the role of $x(t)$ (domain 3), and the full convolved signal $y(t)$ through ordinary deconvolution (see Eq. 3 of Chapter 2). On the other hand, Kernelizing uses the fact that the convolution of any function with a multiexponential decay gives as a result a convolved multiexponential decay with the same decay constants as the original signal, but showing different proportions of the monoexponential contributions. Therefore, if the multiexponential decay signal of a sample is convolved with many diverse kernels, the result is a set of convolved signals with the same multiexponential behavior as the original one but each one with different proportions of the underlying pure monoexponential decays. When this kernelizing operation is done on a bilinear data matrix formed by decay curves of different samples, the sets of convolved signals obtained *per sample* can be structured as a trilinear data set.

BIRFI algorithm: predicting the IRF from the measured fluorescence decay

The measured fluorescence decay ($\boldsymbol{y}$) is the result of the convolution of the IRF and the intrinsic exponential decay ($\boldsymbol{x}$), as expressed in the following equation, $\boldsymbol{y} = \mathbf{IRF} * \boldsymbol{x}$. To recover $\boldsymbol{x}$ given $\boldsymbol{y}$, deconvolution is the most used operation. This procedure needs first the estimation of the IRF, to be subsequently used to obtain $\boldsymbol{x}$. Typically, the estimation of the IRF can be done by measuring the emission of a fluorophore with a very short fluorescence lifetime or the elastic scattering of the excitation laser pulse [Luchowski et al., 2009; Szabelski et al., 2009; Szabelski et al., 2009b]. Then, once the IRF is estimated, deconvolution of the measured signal ($\boldsymbol{y}$) can be performed to estimate the true signal ($\boldsymbol{x}$).

However, the experimental measurement of the IRF has a limitation. The IRF shape is inherent to specific experimental conditions and may vary among measurements due to chemical and instrumental factors. If the measurement of $\boldsymbol{y}$ is done under different experimental conditions to those used to measure IRF e.g., employing distinct wavelength filters, or different emission windows than those for IRF measurement, this difference may lead to biases in the extraction of $\boldsymbol{x}$ through deconvolution.

Publication IV aims to propose a solution for this issue, i.e., to estimate the IRF from the measured fluorescence decay instead of using the emission of a fluorophore with a very short fluorescence lifetime or the elastic scattering of the laser pulse. The algorithm was tested on simulated and experimental datasets. However, in this discussion, only experimental datasets are considered to focus on real-case scenarios and as a direct validation of the accuracy of the approach.

The BIRFI algorithm operates on the assumption that the monoexponential signal ($\boldsymbol{x}$), is equal in shape to the signal found in the tail of the measured decay signal ($\boldsymbol{y}$), except for scale differences. Therefore, using the complete convolved ($\boldsymbol{y}$) signal and the monoexponential part of ($\boldsymbol{y}$) playing the role of ($\boldsymbol{x}$), the IRF can be readily extracted since two of the three unknows of equation $\boldsymbol{y} = \mathbf{IRF} * \boldsymbol{x}$ are available.

To exploit this idea, the derivative of the convolved signal ($\boldsymbol{y}$) helps to visualize the border between the non-exponential and exponential signal domains (tail) and the rightmost minimum of the derivative serves as a practical approximation for the cutting-point of these signal regions (Fig. 30). Once the monoexponential part of ($\boldsymbol{y}$) is detected, IRF is extracted by direct deconvolution using both the full convolved ($\boldsymbol{y}$) and its monoexponential part, playing the role of ($\boldsymbol{x}$).

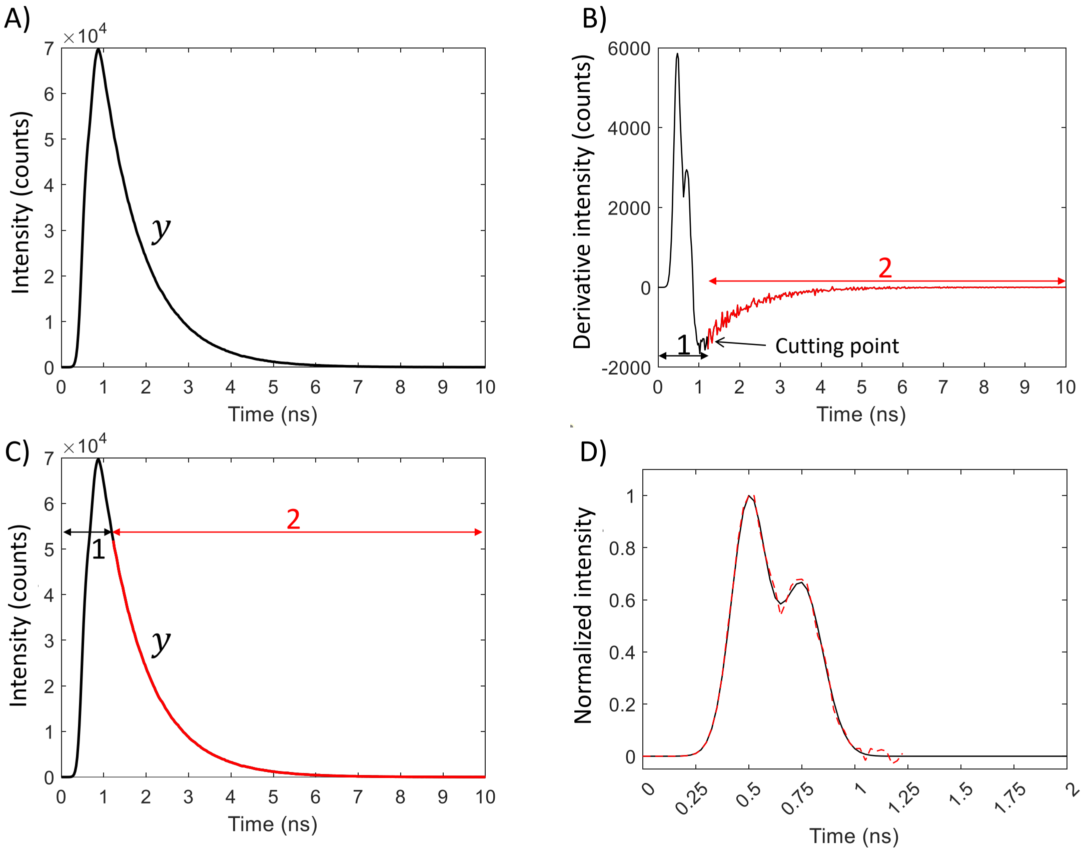


Figure 30. A) Measured convolved decay y. B) Derivative of the measured convolved decay featuring an initial nonexponential trend (black, 1) and a tail with a characteristic exponential behavior (red, 2). C) The exponential tail (red) is identified based on the recognition of the cutting point in Fig. 30B. D) The tail and the convolved decay are finally exploited to perform an ordinary deconvolution and estimate the IRF (red dashed line). The solid black line in D) represents the true IRF.

BIRFI was tested on simulated data and on controlled experimental conditions, where a fluorophore that exhibits a monoexponential decay behavior was measured in three different laser power conditions. This measurement provided a set of fluorescence decays convolved with three different IRFs. The results are displayed in Fig. 31. For the sake of simplicity and interest, only real cases are shown in this subsection. For results on simulated datasets, where BIRFI was tested on examples of IRFs with different shapes and data sets with different noise levels, please refer to **Publication IV**.

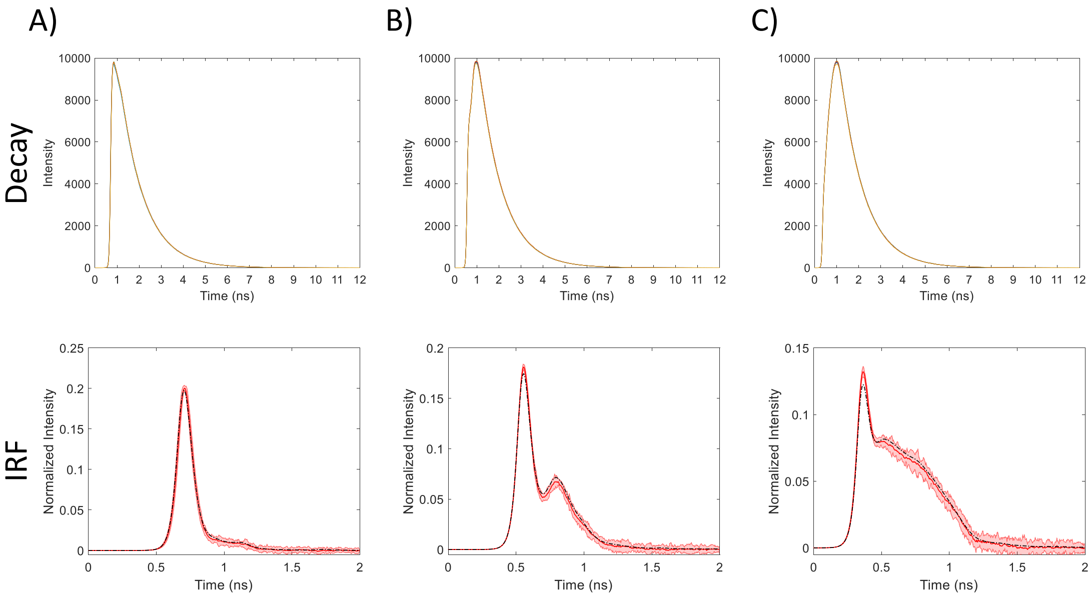


Figure 31. Results provided by BIRFI for the first (A), the second (B) and the third (C) pure fluorescence decay datasets, obtained using different laser powers on the same dye solution. Top panel: smoothed fluorescence decays (10 replicates per case). Bottom panel: average predicted IRFs (red solid lines) +/- three standard deviation intervals (red shaded areas), and measured IRFs (black dashed lines).

Figure 31 shows the three different IRF predicted by the sole use of the fluorescence decays measured. The results of the application of BIRFI indicate a very good ability to predict the IRF irrespective of its shape. To complete the study, the IRF predicted by BIRFI and the IRF measured experimentally were used to fit the fluorescence decays through reconvolution. The fitting residuals are shown in Fig. 32.

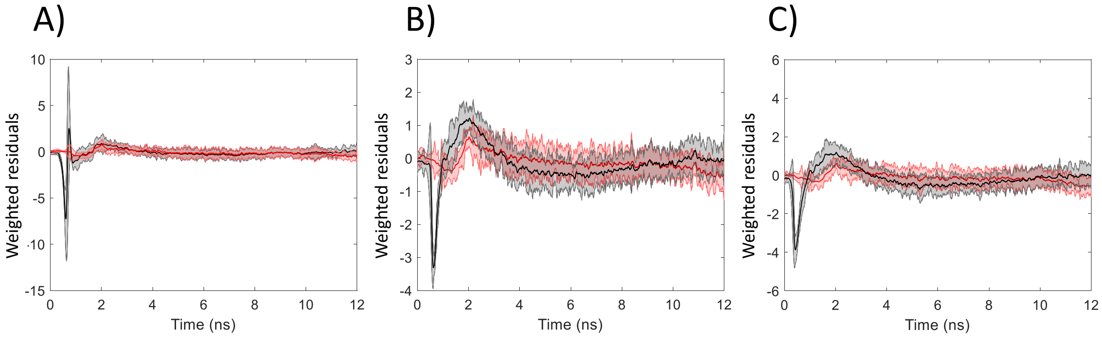


Figure 32. Weighted residuals resulting from the monoexponential fitting of the fluorescence decays in the first (A), the second (B) and the third (C) pure fluorescent dye dataset. The fitting was performed with reconvolution using the IRF estimated by BIRFI (red) and the measured IRF (black). Each measurement replicate was fitted individually: the solid lines represent the average residual profiles, while the red and black shaded areas denote the corresponding three standard deviation intervals.

In the three datasets studied, the fitting residuals showed less structure when reconvolution was performed with the IRF estimated through BIRFI than if the experimentally measured IRF was used (see Fig. 32). This reduction in structured residuals indicates a promising approach, able to improve as well the accuracy of the lifetime estimates.

In summary, the BIRFI algorithm has been introduced to estimate the IRF directly from measured decay data. The potential implications of this algorithm lie in its ability to mitigate biases introduced by experimental variations in IRF measurements, helping to the researchers to find a robust estimation of the IRF by using very conventional fluorophores, instead of those with a very short lifetime.

Although the study of the IRF in FLIM has not been explored by BIRFI, the algorithm holds the potential for providing interesting results. This aspect is considered as a future direction for investigation since especially in systems where spatial resolution is crucial, the characteristics of the IRF may vary across pixels. This variability is influenced by factors such as optical aberrations, scattering and differences in detector response across the field of view. Therefore, a single IRF may not accurately fit the data.

To address this issue, BIRFI may be applied to each pixel or group of pixels to map the IRF across the image. This approach has the potential to enable the deconvolution of the measured fluorescence decay curve at each pixel by accounting for the corresponding local IRF, opening a direction for future investigations.

Generating multiexponential trilinear data from bilinear data through the Kernelizing approach

Unmixing multiexponential fluorescence decay signals poses a challenging task for bilinear decomposition approaches due to the strong correlation and complete window overlap of the pure monoexponential profiles sought. To address this issue, *slicing* methodologies like PowerSlicing [Engelsen et al., 2003] have been employed, which elegantly transform the original data matrix into a three-way data array for trilinear decomposition. The tensorization of the matrix of multiexponential decays is done by building equally sized slices of time windows of the initial data set separated by a certain lag until the full time interval is covered. Due to the properties of the exponential functions, the multiexponential decays in the slices will have the same behavior as the complete decay curves in the initial matrix. The tensor formed in this way can be submitted to a trilinear decomposition method to obtain the profiles of the pure monoexponential contributions a unique way.

While slicing methods have shown satisfactory results for the analysis of multiexponential data, such as nuclear magnetic resonance decays [Engelsen et al., 2003] or TRFS [Devos et al., 2021], they suffer from degradation in accuracy and precision when the decay signals to be analyzed have only a few sampling points. This is due to the low number of slices derived and the small number of time points in each, which hinders the adequate definition of the decay behavior.

For example, in the context of multimodal microscopes, the spectral-FLIM provides a full emission spectrum, time-resolved at each wavelength for every pixel. In this kind of measurement, to obtain a sufficient signal-to-noise ratio, a binning in the time decay direction must be applied, reducing the number of time-channels to few sampling points.

The study presented in **Publication V** introduces a new methodology called Kernelizing, designed to efficiently tensorize data matrices of multiexponential decays, e.g., FLIM data, even if they contain few sampling points.

Kernelizing is based on the property that if an exponential decay is convolved with a function (called kernel), the resulting signal keeps the original decay constants, while only changing the preexponential factors (see Eq. 23, Domain 3). This idea is applied to build trilinear tensors from the initial matrices of measured decay curves, as shown in Fig. 33.

Thus, to build the trilinear data, each measured multiexponential decay (each row of the data matrix **D**) is convolved with a set of different kernels, yielding a set of new multiexponential decays for which the decay constants of the individual monoexponential components are unchanged and only the corresponding preexponential factors are modified. This operation is done for every decay curve in the original matrix and the result is a data cube, where every "slice" comes from the use of a particular kernel to convolve all initial decay curves (note that the cube is formed only by the exponential part of the convolved signals obtained by kernelizing). The cube can be afterwards analyzed by a trilinear decomposition method, such as PARAFAC-ALS, to uniquely extract the pure monoexponential components and, hence, the related lifetimes.

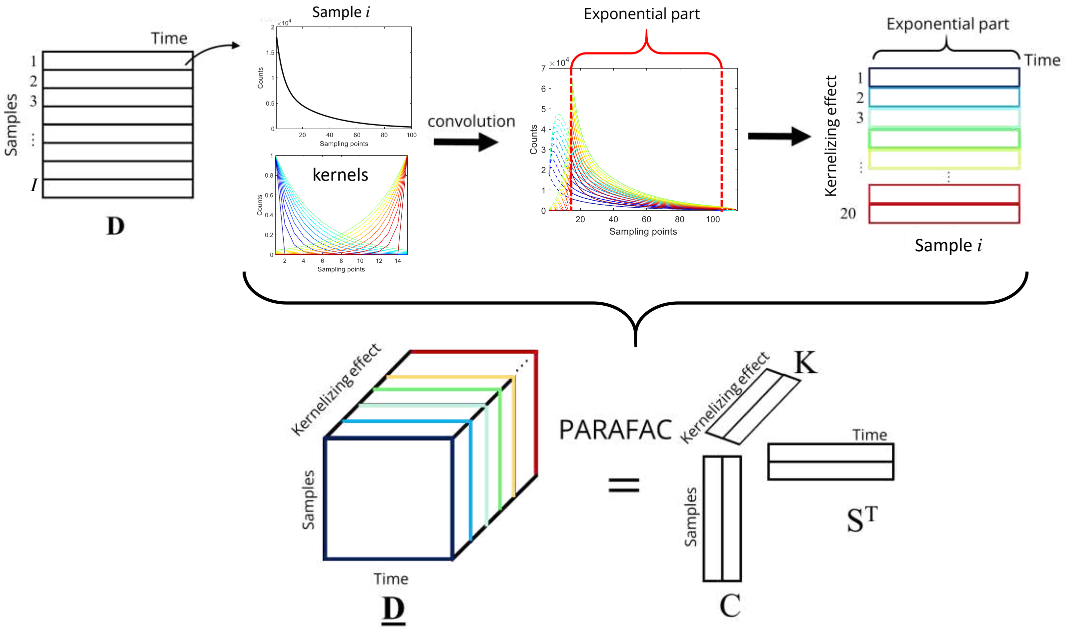


Figure 33. Schematic representation of the Kernelizing approach. Each decay of the matrix **D** is convolved individually with a set of different kernels. Then, only the exponential parts of the resulting curves are kept and gathered into a data cube **D̲**. After convolving every sample with the aforementioned kernels, a trilinear data array can be built and subsequently analyzed by PARAFAC-ALS.

To test the performance of the algorithm, Kernelizing was applied to simulated and real datasets of TRFS data and compared with the PowerSlicing method.

Simulated datasets

Simulated data set 1 was generated as the result of 200 mixtures of three pure monoexponential decays in different proportions, each with decay constants of 0.8 ns, 1.4 ns, and 2.4 ns, with 500 sampling points. The pure monoexponential profiles used exhibit strong correlation and complete window overlap, making them challenging for bilinear unmixing methods. Simulated dataset 2 retained similar characteristics to dataset 1 but involved downsampling the decay curves to 50 points, resulting in a matrix sized (200,50). Dataset 3 further reduced the sampling points to 15 giving a matrix sized (200,15). The matrices of the three data sets were tensorized using both the Kernelizing and PowerSlicing methods and submitted to PARAFAC-ALS for analysis. For the sake of simplification, the details of the tensorization procedure are only shown on **Publication V**.

To compare the PARAFAC-ALS results of Kernelizing and PowerSlicing, the resolved pure monoexponential decay profiles in **S** were fitted by monoexponential functions to extract the decay constants. PARAFAC analyses were performed 1000 times on each simulated data set, adding Poisson-structured noise. The median of the results and 2.5th-97.5th percentile intervals were considered to assess accuracy and precision in the recovery of the decay constants.

For the tensorized dataset 1 (500 sampling points), the values of the three decay constants are well recovered by both Powerslicing and Kernelizing (Fig. 34). The exponential profiles and related decay constants are very well recovered for components 1 and 2, whereas a higher scatter, slightly more pronounced for Powerslicing, can be observed for component 3. When analyzing datasets 2 and 3, characterized by 50 and 15 sampling points, respectively, significant differences between PowerSlicing and Kernelizing can be observed (Fig. 34). The outputs of Kernelizing consistently produced correlation coefficients close to or above 0.9 for all components and datasets when compared with the ground truth profiles used for the simulation. Conversely, PowerSlicing showed a decreasing accuracy, particularly for component 2 in dataset 2 and components 2 and 3 in dataset 3. Regarding the decay constants, Kernelizing yielded correct values for datasets 2 and 3, showing stability across datasets. In contrast, PowerSlicing introduces bias in the decay constants of component 3 in dataset 2 and components 2 and 3 in dataset 3 and shows a larger scatter in the estimation of the decay constants as the number of sampling points decreases.

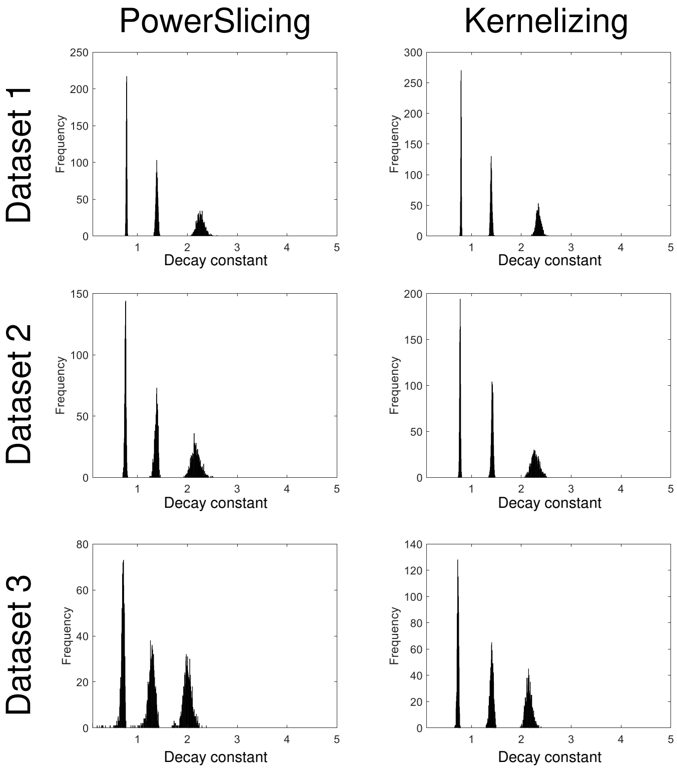


Figure 34. Decay constants returned after the application of PowerSlicing and Kernelizing to the simulated datasets 1, 2 and 3 over the 1000 calculation runs. A significant reduction of the estimated scatter is observed for Kernelizing with respect to PowerSlicing when the number of sampling points is decreased.

The results on simulated datasets demonstrate the ability of Kernelizing to obtain more accurate trilinear model estimates, especially when dealing with decay signals featuring few sampling points (as low as fifteen), compared to traditional slicing methodologies.

There are two key explanations for the results obtained. First, the time-lagging procedure used by PowerSlicing for tensorization restricts dramatically the number of possible slices built, being only three for data set 3, where the initial decay curves had only 15 sampling points. Instead, there is no limit for the number of “slices” provided by the kernelizing approach, the number being as large as the number of kernels used to perform the convolution. In the examples presented, 20 kernels were always used for tensorization, providing three-way arrays with 20 slices each. A second explanation for the higher precision in the Kernelizing results comes from the fact that the convolution operation involves a certain filtering in the noise of the data, which decreases the scatter of the estimated decay constants.

Effect of the kernel size and shape variation on the final solutions

Kernelizing easily generates trilinear data using the convolution of an unlimited set of kernels. However, how to select the length and the shape of the different kernels used is not a trivial question. To shed light on this issue, several simulations have been carried out by varying the kernel width and the kernel shape.

To do so, six sets of kernels were generated, each containing from four to 20 functions with different shapes. Additionally, six different kernel widths were considered for each set. This resulted in 36 sets of kernels used to construct three-way data arrays for three datasets. Each of these 36 sets of kernels was used to analyze dataset 1, 2 and 3. The analysis was repeated 1000 times for each combination of kernel shape and width, with Poisson-structured noise added each time.

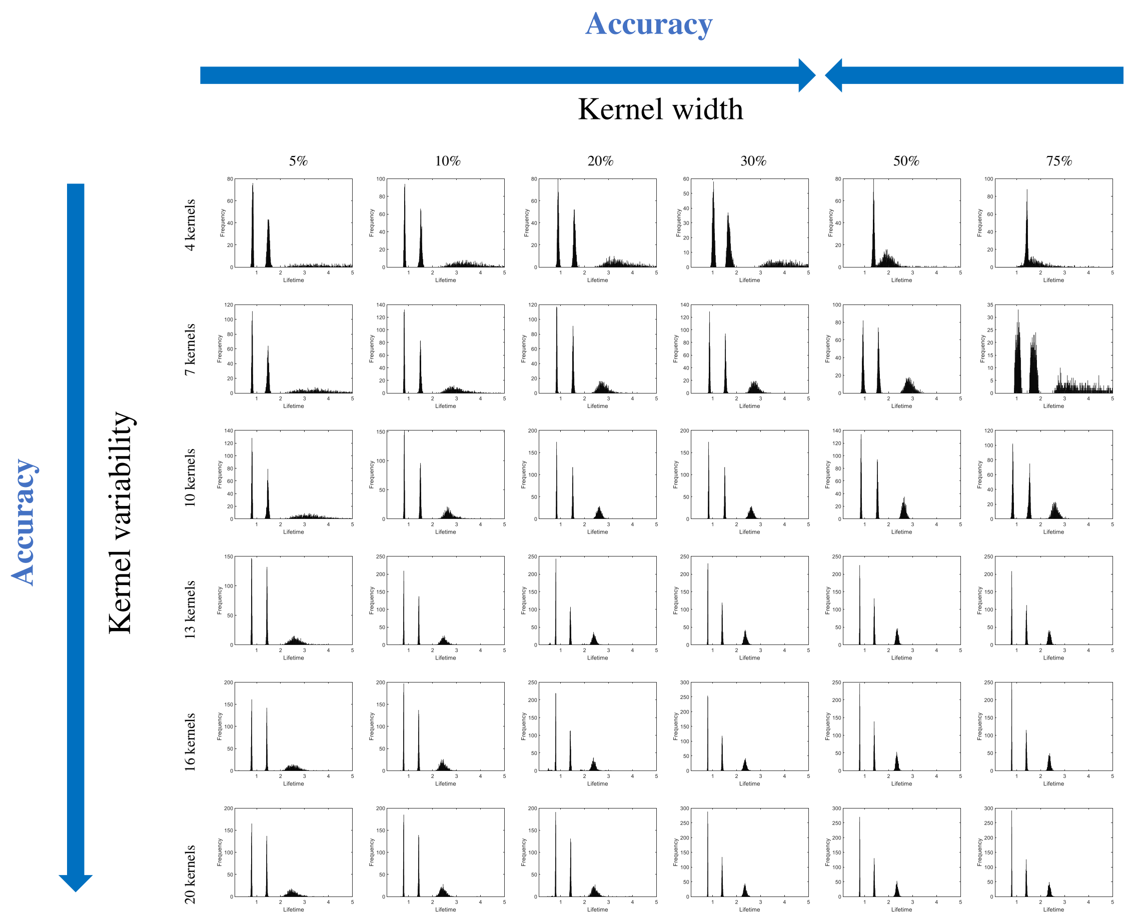


Figure 35. Decay constants recovered using Kernelizing on dataset 1. From top to bottom, the number and diversity of the kernel functions increase. From left to right, the width of the kernel functions (expressed as percentage of points over the total number of sampling points of the original decay curves) increases.

The results depicted in Fig. 35 for dataset 1 show that a higher number of kernels with diverse shapes leads to solutions closer to the ground truth. This can be explained because the amount of information in the tensor increases and the diversity of kernel shapes produces signals with a large variability in the preexponential factors, facilitating the unmixing task.

The effect of the kernel width requires a more complex interpretation. If the kernels used are excessively narrow, they induce a very low variability in the signal preexponential factors, the unmixing task becomes more complex and there is a broader dispersion of the decay constants obtained. However, the use of too broad kernels requires discarding a large part of the convolved signal because it does not show an exponential behavior. Therefore, the small amount of information used results in a broader dispersion of the decay constants again. In practice, it seems necessary to have a balance and use a kernel width that induces sufficient variability in the convolved signals without sacrificing too many sampling points.

The results obtained for datasets 2 and 3 provide the same conclusions as for dataset 1 and are presented in the Supplementary Material section of **Publication V.**

At this stage, while it seems to be clear that the more the variability of the kernel shapes in a set, the better, more investigation is needed to determine the optimal kernel width.

Real datasets

For illustrative purposes, a FLIM image of *Convallaria majali* was analyzed with FLIM data provided by Williams et al. (2021). The preprocessed FLIM image has dimensions 256×227×11, where the first two dimensions represent x- and y-spatial directions, and the third dimension corresponds to the 11 sampling points of the decay curves. Using a set of 20 kernels, each with 3 time points, the data was tensorized, resulting in a dataset sized (256×227,20,8). A three-component PARAFAC-ALS model was applied to decompose the dataset (after unfolding along the pixel direction). Figure 36 shows the pure component concentration maps and the corresponding pure monoexponential decays derived from the PARAFAC analysis.

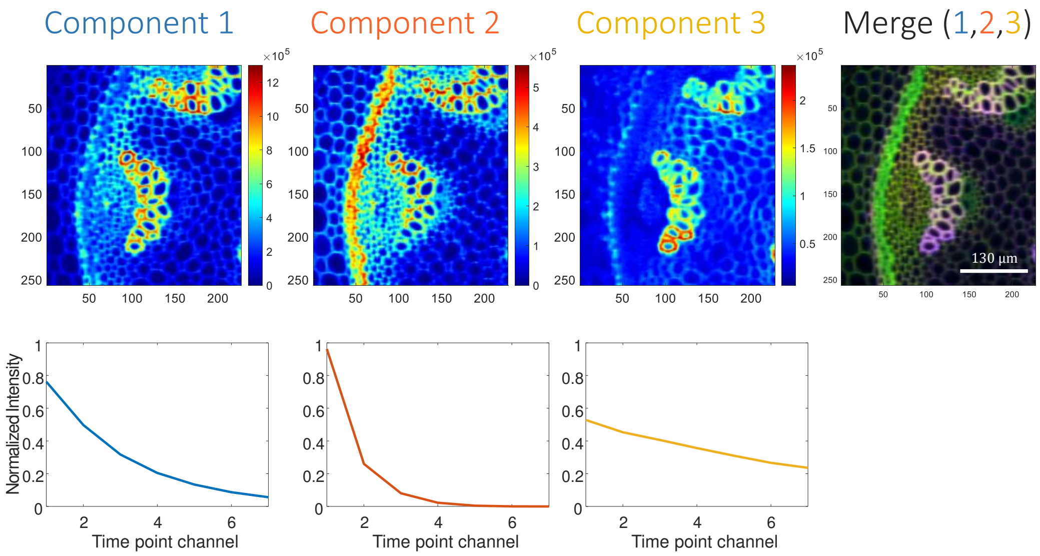


Figure 36. Analysis of FLIM of a sample of *Convallaria majali*. Pure distribution maps (top) and pure fluorescence decays (bottom) recovered by Kernelizing-PARAFAC-ALS.

The analysis revealed specific biological zones in the vegetal tissue. While there is no ground truth available, the concentration maps suggest that component 1 corresponds to the xylem and may relate to safranin dye linked to lignin, highlighting the endodermis as well. Component 2, with a fast decay, appears in mesophyll cells and lignified cells, potentially related to fast-green stain, known to highlight the phloem and cellulosic cell walls. Component 3, with a slow decay, specifically appears in the xylem, differentiating multiple lignin environments.

In conclusion, Kernelizing has been found very useful to handle multiexponential measurements for which binning is required to increase the signal-to-noise ratio or with number of sampling points low due to instrumental limitations.

SECTION II – Addressing challenges of image fusion

This section contains four scientific publications addressing specific challenges of the image fusion field. The methodologies proposed allows for the fusion of a widely variety of hyperspectral images, such as different scanned areas, different spatial resolution and different dimensionality.

3.3 Image Fusion: A case study applied to vegetal tissues

In this work, a case study involving the fusion of hyperspectral images of different spectroscopic imaging platforms, i.e., synchrotron infrared, Raman, and fluorescence, is presented. The fusion challenge lies in merging images with different spatial resolutions, sample orientations and scanned areas. To show how the classical image fusion protocol works, a case study about the investigation of cross-sections from rice leaves, where Raman, synchrotron infrared, and fluorescence images have been independently measured, is described. The images are fused in a single multiset and analyzed by MCR-ALS. This first example of image fusion describes the preprocessing protocols oriented to build a suitable complete multiset. The benefits of image fusion vs. individual image analysis become clear, but the drawbacks of classical fusion, i.e., loss of spatial resolution and scanned areas to allow obtaining a regular multiset structure, also affect the quality of the results obtained.

Publication VI. Linear unmixing protocol for hyperspectral image fusion analysis applied to a case study of vegetal tissues

Authors: A. Gómez-Sánchez, M. Marro, M. Marsal, S. Zacchetti, R. R. de Oliveira, P. Loza-Álvarez, A. de Juan.

Citation reference: *Scientific Reports* (2021), 11:18665.

DOI: 10.1038/s41598-021-98000-0

scientific reports

Check for updates

OPEN

Linear unmixing protocol for hyperspectral image fusion analysis applied to a case study of vegetal tissues

Adrián Gómez-Sánchez[1]✉, Mónica Marro[2], Maria Marsal[2], Sara Zacchetti[1,3], Rodrigo Rocha de Oliveira[1], Pablo Loza-Alvarez[2] & Anna de Juan[1]✉

Hyperspectral imaging (HSI) is a useful non-invasive technique that offers spatial and chemical information of samples. Often, different HSI techniques are used to obtain complementary information from the sample by combining different image modalities (Image Fusion). However, issues related to the different spatial resolution, sample orientation or area scanned among platforms need to be properly addressed. Unmixing methods are helpful to analyze and interpret the information of HSI related to each of the components contributing to the signal. Among those, Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) offers very suitable features for image fusion, since it can easily cope with multiset structures formed by blocks of images coming from different samples and platforms and allows the use of optional and diverse constraints to adapt to the specific features of each HSI employed. In this work, a case study based on the investigation of cross-sections from rice leaves by Raman, synchrotron infrared and fluorescence imaging techniques is presented. HSI of these three different techniques are fused for the first time in a single data structure and analyzed by MCR-ALS. This example is challenging in nature and is particularly suitable to describe clearly the necessary steps required to perform unmixing in an image fusion context. Although this protocol is presented and applied to a study of vegetal tissues, it can be generally used in many other samples and combinations of imaging platforms.

Hyperspectral imaging (HSI) is a useful non-invasive analytical technique that allows preserving the morphological and chemical information associated with samples. This technique consists of collecting spectroscopic information associated with different points (pixels) of a scanned area in a sample. In this way, spatial and chemical information about the samples is provided and limitations linked to traditional single point spectroscopic techniques, such as the lack of spatial information, are clearly overcome. Nowadays, imaging platforms offer a wealth of spatial resolution scales and are adapted to the specificities of many spectroscopic (and spectrometric) modalities[1,2]. Despite the clear value of the complementary information provided by the currently available imaging platforms, image fusion is still a challenge that does not have a generalized solution[3].

The size and complexity of the information provided by hyperspectral images need powerful chemometric techniques for their adequate interpretation. Very often, the goal of HSI is providing information about the nature and location of sample constituents. In the beginning of hyperspectral imaging, the compound location was described displaying maps at selected spectral channels and the compound spectral fingerprint was associated with spectra of pixels located in specific sample regions. However, such an approach is clearly insufficient for complex multicomponent samples, where often no selective spectral channels exist and the extraction of clear compound fingerprints is hindered by the colocation of components in the pixels of the image. Unmixing methods come then into play to provide pure spectral signatures and pure concentration maps of the image constituents and, hence, a global chemical, quantitative and morphological information of the samples studied.

The unmixing task can be tackled by linear and non-linear methods depending on the underlying model assumed to define the spectroscopic measurement, i.e., the spectroscopic signal in every pixel is defined as a

[1]Department of Chemical Engineering and Analytical Chemistry, Universitat de Barcelona, 08028 Barcelona, Spain. [2]ICFO- Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology, 08860 Castelldefels, Spain. [3]Department of Chemistry, Università Di Roma "La Sapienza", 00185 Rome, Italy. ✉email: agomezsa29@alumnes.ub.edu; anna.dejuan@ub.edu

concentration-weighted linear combination of the pure spectra of the image constituents in linear models or the signal definition obeys more complex models in non-linear approximations. Linear unmixing methods reflect exactly the basic form of the spectroscopic Beer-Lambert law, where every component is defined by an invariant spectral fingerprint that contributes to the signal measured proportionally to its concentration. Non-linear methods, instead, take into consideration that there may be variability in the spectral fingerprint of a particular component in certain instances. The non-linear unmixing problem is often solved by using deep learning methods based on the use of neural network autoencoders[4,5]. Such an approach has found applicability basically in remote sensing scenarios, where the definition of component, e.g., soil, vegetation… and the conditions of the image acquisition may sometimes justify the assumption of a certain variability in the spectral signatures of components. However, in a very large number of cases, particularly when image platforms located in the laboratory are used, the results provided by linear unmixing methods are a very good approximation of the real behavior of the spectroscopic measurement, need a lower computation time and allow a simpler implementation of external available information under a variety of constraints[6]. As in any other data analysis context, the parsimony principle stating that the simplest model that provides a satisfactory description of the phenomenon studied has to be chosen prevails in this case and the protocol proposed in this work is based on linear unmixing methodologies.

Within the family of linear unmixing methods, Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)[7] is a chemometric method that has been widely used for image analysis due to the flexible characteristics offered in terms of the data structures that can be potentially studied and the diversity of information that the algorithm incorporates under the form of constraints to help in the modelling of concentration maps and spectral signatures of the pure components[7–9]. Other linear unmixing methods, often used in the remote sensing area, tend to work forcing necessarily non-negativity and normalization constraints or using libraries of previously known spectral signatures and do not change the modus operandi when dealing with an individual image or with an image fusion scenario[10]. Hence, the choice of the MCR-ALS method in this work.

MCR-ALS has been successfully applied to the analysis of single images or sets of related images acquired with the same spectroscopic platform[9]. However, the use of this methodology for fusion of images from different platforms is not extended yet, although few examples are reported to address problems such as the fusion of images with different spatial resolutions[11,12], with spectroscopic modalities showing different dimensions[13] or with the combination of spectroscopic and color information[12,14].

In a general multiplatform fusion context, images from each platform could be analyzed separately by MCR-ALS, but the fusion of the information from the different image techniques provides a complete description of the sample constituents and more accurate solutions[3,9]. Multiplatform image fusion allows exploiting the complementary information provided by different spectroscopic techniques and obtaining simpler models including all the gathered information in a single complete, reliable, and robust model to answer to the scientific question of interest. Image fusion has started to emerge as an excellent methodology to analyze data of different chemical systems.

An area where image fusion can be particularly relevant is the research of the structure and composition of tissues in living organisms. Indeed, the natural complexity of the biological tissues looks as a problem that can be adequately addressed with the use of imaging platforms sensitive to different information and components. The results obtained can hopefully provide the necessary link between chemical structure and function required for the understanding of these systems.

An interesting case study of biological interest to apply image fusion strategies is the characterization of vegetal tissues. Indeed, different hyperspectral imaging techniques can be used simultaneously to obtain complementary information for this particular biological system. In this respect, the combination of fluorescence, Raman and infrared imaging techniques is a good option. Fluorescence images collect the emission fluorescence spectra from natural fluorophores in plants, such as lignin, chlorophylls[15], while Raman images provide information about cellulose, lignin, carotenes and other components[16]. Finally, infrared images provide information about molecular components, such as proteins, lipids and carbohydrates[17]. However, spatial resolution from conventional infrared HSI is insufficient for a good definition of micron vegetal tissue substructures. Synchrotron Radiation Fourier Transform Infrared (SR-FTIR) imaging, instead, has the necessary spatial resolution and helps to reveal the microstructures at tissue level[18].

In this work, SR-FTIR, Raman and fluorescence HSI from rice leaf cross-sections were acquired to study thoroughly the different constituents and structures found in the tissues of this plant. For the first time, SR-FTIR, Raman and fluorescence HSI are fused and analysed by MCR-ALS. To do so, images from the different platforms had to be balanced in terms of spatial resolution, orientation and area scanned before being analysed. As a result, spectral signatures of plant components showing the relevant features of all different spectroscopic techniques used and the related distribution maps defining accurately the spatial structure of the biological elements identified were obtained. The steps followed and the gain obtained when using image fusion as compared with the analysis of images coming from individual platforms is clearly proven. Despite the intrinsic interest of the characterisation of components in vegetal tissues, the main goal of the work is providing a general framework that can be generally adopted to address linear unmixing in any multiplatform image fusion problem.

Experimental

Plant growth and sample preparation.

Rice plants were obtained from *Oryza Sativa Japonica Nipponbare* seeds provided by the Center for Research in Agricultural Genomics (CRAG) at Autonomous University of Barcelona. This public university center complies with all necessary legislative regulations on the treatment of plant seeds and living organisms and the seeds used do not present any kind of hazardous risk for their growth and use. Seeds were germinated for two days at 30 °C in a wet environment. After germination, seeds were planted in small individual pots with a universal substrate BATLLE, composed by coconut fiber, peat moss,

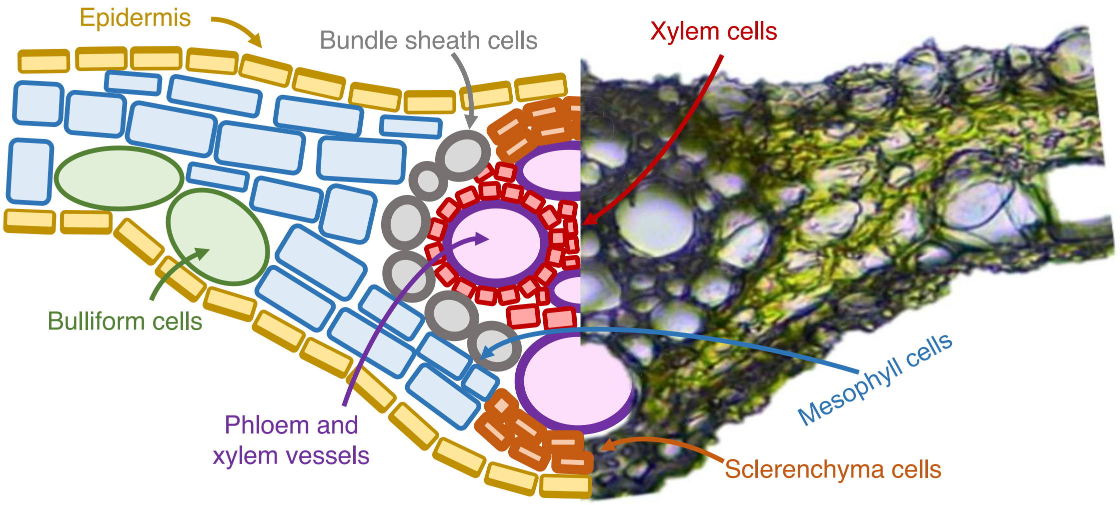


Figure 1. Schematic and optic cross-section image of a rice leaf.

composted vegetal material and perlite with pH 7.25. Rice plants were watered two times per week with 400 mL of Milli-Q water for 33 days under controlled conditions of temperature, light and humidity in an Environmental Test Chamber MLR-352H (PANASONIC) in the Institute of Environmental Assessment and Water Research-Spanish National Research Council (IDAEA-CSIC).

Once the plants were grown, small pieces of plant leaves of different plants were collected and embedded in agarose. Straightaway, three cryosections of seven µm-thickness were obtained using a cryostat at the Parc Científic of Barcelona at − 20 ± 5 °C. Sections were placed on a calcium fluoride slide of 1 mm-thickness, covered with a calcium fluoride coverslide of 0.5 mm-thickness and sealed with nail polish. In every cross-section, different regions can be observed (Fig. 1). Mesophyll cells, where photosynthetic activity is located and chlorophyll or carotenoids can be found. The characteristic green color of the plants comes from this type of cells. Also, the epidermis can be observed. The function of the epidermis is to protect the plant tissue from external damages. Several compounds can be found there, such as resins that cover the epidermis to avoid water loss. In the vascular system, two parts are differentiated: xylem and phloem. Xylem is a type of lignified tissue that transports water and minerals and it is formed by a conglomerate of the bigger channels. Phloem is another type of lignified tissue that transports nutrients as sugar or other biomolecules. It is located on top of the xylem in a cross-section view. Finally, sclerenchyma cells can be located on the top and the bottom of the vascular system. Sclerenchyma cells are strongly lignified cells and give hardness to the plant.

Image acquisition

Synchrotron infrared image acquisition.

All SR-FTIR HSI were collected at the SYNCHROTRON ALBA (Cerdanyola del Vallès, Catalunya, Spain, MIRAS beamline). The Fourier transform infrared spectrometer used was equipped with a TE Cooled DLaTGS Detector Vertex coupled to a HYPERION 3000 Microscope. The detector of the IR microscope was a liquid-nitrogen-cooled 50 µm HgCdTe detector, covering the range of 10000–600 cm^{-1}. The microscope was operating using a 36 × objective. IR spectra were acquired in transmission mode by point mapping and every spectrum was associated with a pixel sized 3 × 3 µm^2. Spectra were collected in the infrared region covering the range of 4000–1000 cm^{-1} with 4 cm^{-1} resolution and 64 accumulations. Background was collected every 25 spectra with 128 accumulations.

Fluorescence image acquisition.

Fluorescence HSI were collected using a LEICA TCS SP8 STED 3X microscope (LEICA MICROSYSTEMS, Mannheim, Germany). A 405 nm laser beam with a power approximately of 160 µW focused through a 10 × objective LEICA HC Pl Apo was used as a light source. A Gated HyD hybrid detector in photon counting mode was used for the spectra collection. Spectra were collected by laser point scanning with an exposure of 0.825 µs/pixel with 70% of total laser power. Every line is formed approximately by 800 pixels and each line was accumulated two times to improve signal to noise ratio. The studied spectral emission range goes from 420 to 750 nm, with a spectral resolution of 5 nm and the pixel size of 0.25 × 0.25 µm^2.

Raman image acquisition.

Raman HSI were collected using an INVIA RAMAN Microscope spectrometer (RENISHAW, Gloucestershire, UK). A 532 nm laser beam focused through a 20 × objective Leica (NA = 0.4) with a power of 25 mW was used as a light source. Spectra were collected by point mapping with 0.25 s exposure time and 10 % of total laser power per pixel. The studied spectral range goes from 270 to 2015 cm^{-1}, with a spectral resolution of 1.55–1.95 cm^{-1} depending on the Raman shift scanned. Pixel size was of 2 × 2 µm^2. The Raman spectrum is recorded on a deep depletion charge coupled device (CCD) detector (RENISHAW RenCam).

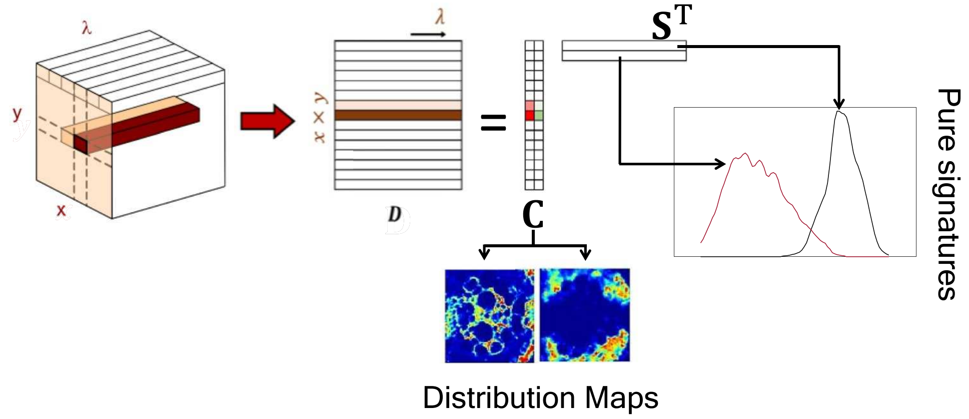


Figure 2. Bilinear model of an HSI.

Data analysis

HSI preprocessing. In the case of SR-FTIR images, due to the opacity of the sample in several regions, some IR spectra were saturated. These pixels were removed and not used in further analysis. Also, for all samples, infrared spectra were first cropped within 3000–1200 cm^{-1} spectral range. Wavenumbers out of this range were not used because of the signal saturation observed. Strong baseline artifacts were also detected. The second derivative was applied to the infrared spectra to remove offsets and linear baselines and to enhance the separation of overlapping peaks through the Savitzky-Golay algorithm[19]. From the derivative spectra, only spectral regions with useful information were selected for further analysis (3000 to 2800 cm^{-1} and 1800 to 1360 cm^{-1}).

In fluorescence images, emission fluorescence per pixel was low due to the high spatial resolution (pixel was sized 0.25 × 0.25 µm^2) and to the low quantum yield of natural fluorophores for this instrumental set. To improve the spectroscopic signal quality, spectra of adjacent pixels were binned to create a pixel with a single spectrum (binning). The binning chosen was (3 × 3) pixels, which provided a final pixel size of 0.75 × 0.75 µm^2. Despite the pixel binning, spatial resolution was still high and the spectral quality was improved.

In Raman images, all spectra showed a high fluorescence baseline contribution due to the natural fluorophores in the rice tissue. Fluorescence baseline interferes with the Raman peaks hiding them and hardening the interpretation. Raman spectra were corrected by Asymmetric Least Squares[20] to remove fluorescence baselines. Also, cosmic peaks were corrected by interpolation of Raman intensities of nearest channels. In addition, the range 1100 to 1800 cm^{-1} was used for the analyses. An example of raw and preprocessed SR-FTIR, fluorescence and Raman spectra can be found in Fig. S1 in the Supporting Information.

Image linear unmixing: multivariate curve resolution-alternating least squares (MCR-ALS). An HSI can be visualized as a data cube where x and y are the pixel coordinates and λ the spectral dimension. In this cube, a full spectrum is associated with each pixel coordinate. If the HSI cube is unfolded, the data acquires a matrix structure **D** (I×J) (Fig. 2) that contains all pixel spectra of the image one under the other. Unmixing methods are able to describe the mixed information in the original pixel spectra in **D** through a bilinear model analogous to the Lambert-Beer law, where the total spectroscopic signal collected can be expressed as the sum of the signal contributions of each individual image constituent. Following the linear Beer-Lambert law, the contribution of each image constituent to the total signal collected can be mathematically expressed by the pure spectrum of the compound $\mathbf{s}_i^{\mathrm{T}}$ weighted by its concentration in the different pixels, $\mathbf{c}_i$, defined by the term $\mathbf{c}_i\,\mathbf{s}_i^{\mathrm{T}}$ (Eq. 1). Finally, the typical bilinear model associated with unmixing methods is expressed in compact format as shown in Eq. (2), where the matrix $\mathbf{S}^{\mathrm{T}}$ contains the profiles of the pure spectra of the image constituents and the matrix **C** the related concentration profiles. The residuals of the model are expressed by **E** (I×J). The spectroscopic bilinear model of an image allows expressing the information of every sample constituent with a spectral signature $\mathbf{s}_i^{\mathrm{T}}$ and a concentration profile $\boldsymbol{c}_i$ that conveniently refolded provides the related distribution map.

$$\mathbf{D} = \sum_i c_i s_i^{\mathrm{T}} + \mathbf{E} \tag{1}$$

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \tag{2}$$

Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) is a multivariate least-squares based iterative resolution (or unmixing) method that alternatingly optimizes matrices **C** and $\mathbf{S}^{\mathrm{T}}$ under the action of constraints that help to provide chemically meaningful spectral and concentration profiles. MCR-ALS is used in many fields of application and is especially suitable for hyperspectral image analysis[7–9].

The method starts doing an estimation of the number of components present in the original data set **D** by Principal Component Analysis (PCA)[21] or taking advantage of previous knowledge of the sample (note that when

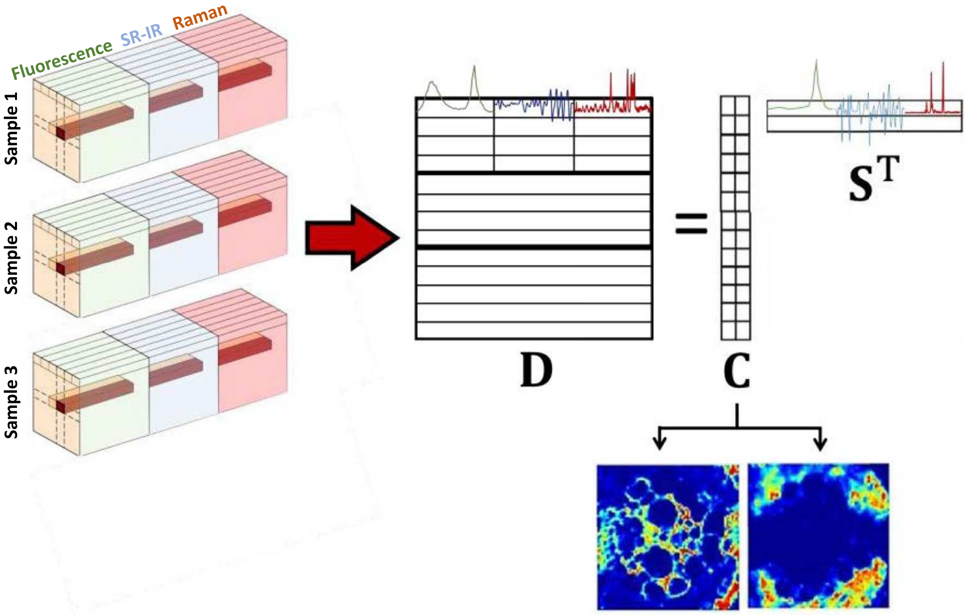


Figure 3. Scheme of the multiset structure issued from image fusion and related bilinear model. In this case, nine data blocks form the multiset: three samples imaged by three spectroscopic platforms.

a high number of components is detected in this step, the potential need for a non-linear unmixing method can be considered). Afterwards, an initial estimate of matrix **C** or $\mathbf{S}^T$ (most often spectral estimates in HSI analysis) is built by a pure variable selection method based on Simple-to-use Interactive Self-modelling Mixture Analysis (SIMPLISMA)[22] or on similar algorithms. Such an estimate and the matrix **D** are used to start the least-squares alternating optimization of the profiles in matrices **C** and $\mathbf{S}^T$ of the bilinear model under the action of constraints until convergence is achieved. The convergence criterion can be a maximum number of iterations or a value related to the difference in fit improvement between consecutive iterations.

The quality of the MCR model fit is described by the lack of fit *LOF* (%), defined by

$$LOF\ (\%\ ) = 100 \times \sqrt{\frac{\sum_{i,j} e_{ij}^2}{\sum_{i,j} d_{ij}^2}} \tag{3}$$

and the percent of variance explained, defined by

$$r^2 = \left(1 - \frac{\sum_{i,j} e_{ij}^2}{\sum_{i,j} d_{ij}^2}\right) \tag{4}$$

where d_{ij} is an element of **D** and e_{ij} is related to **E**. In a HSI context, when the final bilinear model is obtained, the pure spectral signatures of the image constituents are the profiles in the $\mathbf{S}^T$ matrix and their pure distribution maps can be recovered refolding the related concentration profiles into the original spatial geometry of the image.

Often, several images may contain related information. When this is the case, it is possible to build multiset structures that contain several connected images. Multisets can be formed appending blocks of spectra from related images obtained with the same spectroscopic technique on under the other in a column-wise augmented fashion. In this case, the spectral dimension needs to be common for all images. A multiset can also be built appending spectra from images of the same sample obtained with different imaging platforms in a row-wise augmented fashion. In this case, the pixel dimension needs to be common for all images. The most complex and complete multisets can be built connecting images from different samples obtained with different platforms (see Fig. 3) in a row- and column-wise augmented fashion. MCR-ALS can also be used to analyze these multiset structures and a bilinear model is also obtained, where the matrix **C** and/or $\mathbf{S}^T$ can also be formed by small blocks (submatrices) related to concentration profiles of the different images and/or to pure spectral signatures of the different platforms used.

The MCR-ALS analysis of a single image or an image multiset takes the benefit of the use of constraints on **C** or/and $\mathbf{S}^T$ to obtain chemically meaningful and more accurate spectral signatures and distribution maps. Classical constraints, such as non-negativity, are often applied to concentration maps and to some spectroscopic measurements[23]. Another useful constraint is the selectivity/local rank[24,25]. In this context, this constraint may force particular pixels to show null concentration values or some spectral ranges to show null signal for particular image constituents. Such an information can come from previous knowledge or from image-adapted local rank analysis methods[26]. Recently, a new generation of constraints has appeared that takes into account characteristics of the spatial distribution of components as well[27,28]. An asset of the use of constraints in MCR-ALS analysis is that they can be optionally set per component, per mode (**C** or $\mathbf{S}^T$) and per block in each of the **C** or $\mathbf{S}^T$ submatrices in a multiset context. This flexibility allows the preservation of the specific characteristics of the spatial distribution of components in the different samples and the properties of the spectral signatures of the different spectroscopic techniques.

Studying adequately a multiplatform image multiset passes through the solution of problems linked to the multiset configuration and the multiset analysis. Thus, a proper configuration of the multiset needs a common pixel dimension among the images combined, i.e., having congruent pixels, and an efficient multiset analysis demands a proper balance of the information linked to the data blocks of the different platforms. A proper description of the sequence needed to solve these problems is carried out taking as example the data linked to the case study presented. In this case, the final multiset will have the same structure as shown in Fig. 3 and will be the result of fusing images from three different samples analyzed with Raman, fluorescence and SR-FTIR platforms. This multiplatform image fusion should follow the steps displayed in Fig. 4, which are:

Building a multiset with congruent pixels. Such a goal requires matching the pixel size of all images and the image area scanned. Afterwards, a spatial transformation (shift and rotation) of images is required to ensure pixel congruency. In the combination of fluorescence, SR-FTIR and Raman images of our case study, this will happen as follows:

Matching pixel size of all images and image area scanned. SR-FTIR images are those with largest pixel size, 3 × 3 μm^2 and smallest area scanned. The rest of imaging techniques are binned to achieve this pixel size. Thus, the pixels of the fluorescence HSI, sized 0.25 × 0.25 μm^2, were binned by a factor of 12 × 12 to achieve the pixel size 3 × 3 μm^2. Raman HSI had a pixel size of 2 × 2 μm^2. An inhouse developed MATLAB script was used to bin and interpolate the pixel values to achieve a 3 × 3 μm^2-pixel size. Finally, the fluorescence and Raman HSI were also cropped until covering approximately the same area than SR-FTIR HSI.

Spatial transformations (shift and rotation of images) for pixel congruency. This step is oriented to compensate the pixel shift and/or rotation among the images to be combined. For this reason, HSI need to be moved in x and y directions and/or rotated until pixels are congruent among fluorescence, SR-IR and Raman HSI.

To obtain the transforming parameters, binarized maps issued from global intensity maps from each HSI technique can be used. The global intensity maps are 2D representations displaying the sum of all spectral intensities of the channels of each pixel spectrum in the image. Global intensity maps are binarized i.e., pixels are assigned a value equal to one (when signal is significant) or zero (when there is no detectable signal). When images have a clear contour, pixels on the sample have much higher intensity than pixels on the background sample support. This contour shape information can be used for the alignment because all images of the same sample must have the same contour, independently on the spectroscopic techniques used for imaging.

The SR-FTIR binarized map is always taken as reference for the alignment (Ar). Sequentially, shifts in in x and y and rotation angle θ were computed for fluorescence and Raman images (As) with the SR-FTIR reference image. To do that, initial estimates for shifts in x and y (dx, dy) and rotation α (Θ) are defined and the map of the image to be aligned is modified accordingly. An error function (Eq. 3), defined as:

$$ssq(\Theta, dx, dy) = \sum\sum\left(A^r_{i,j(x,y,\alpha)} - A^s_{i,j(x+dx,y+dy,\alpha+\Theta)\,(i,j)}\right)^2 \quad (5)$$

is calculated among the binarized values of common pixels in the image to be aligned and the reference image. This is an iterative process that uses a SIMPLEX optimization algorithm and stops when the error defined in Eq. (3) gets sufficiently small[29]. When shift and rotation parameters are found, the whole HSI is spatially transformed to match the reference image. Only pixels from common sample areas scanned by all techniques are used to create the multiset used for further analysis.

Balancing the importance of the data blocks related to each platform in the multiset. The pixel spectra provided by the different imaging platforms can show significant differences related to the scale of the signal recorded and to the number of spectral channels in each measurement. If blocks of the raw pixel spectra are appended in the multiset, platforms that provide spectra with higher signal intensity and formed by a large number of spectral channels will have a major influence in the results obtained. A good quantitative representation of the overall signal contribution of an image is provided by the 2-norm of the related unfolded matrix **D**. Hence, to balance the importance of images coming from different platforms on the same sample, the data block of each image will be divided by its 2-norm before the multiset is built. This is a clear mathematical procedure to keep similar the weights of the different blocks of the multiset, less biased than trying to find suitable scaling factors by visual inspection.

Once a balanced and pixel-congruent multiset is built, MCR-ALS can be properly applied setting the appropriate constrains to the profiles in each block of the **C** and/or $\mathbf{S}^T$ matrices. The application of MCR-ALS to analyze single or fused images has been done using a freely downloadable graphical user interface under MATLAB environment that follows the steps described above and provides the possibility to incorporate in a flexible way the suitable constraints[30].

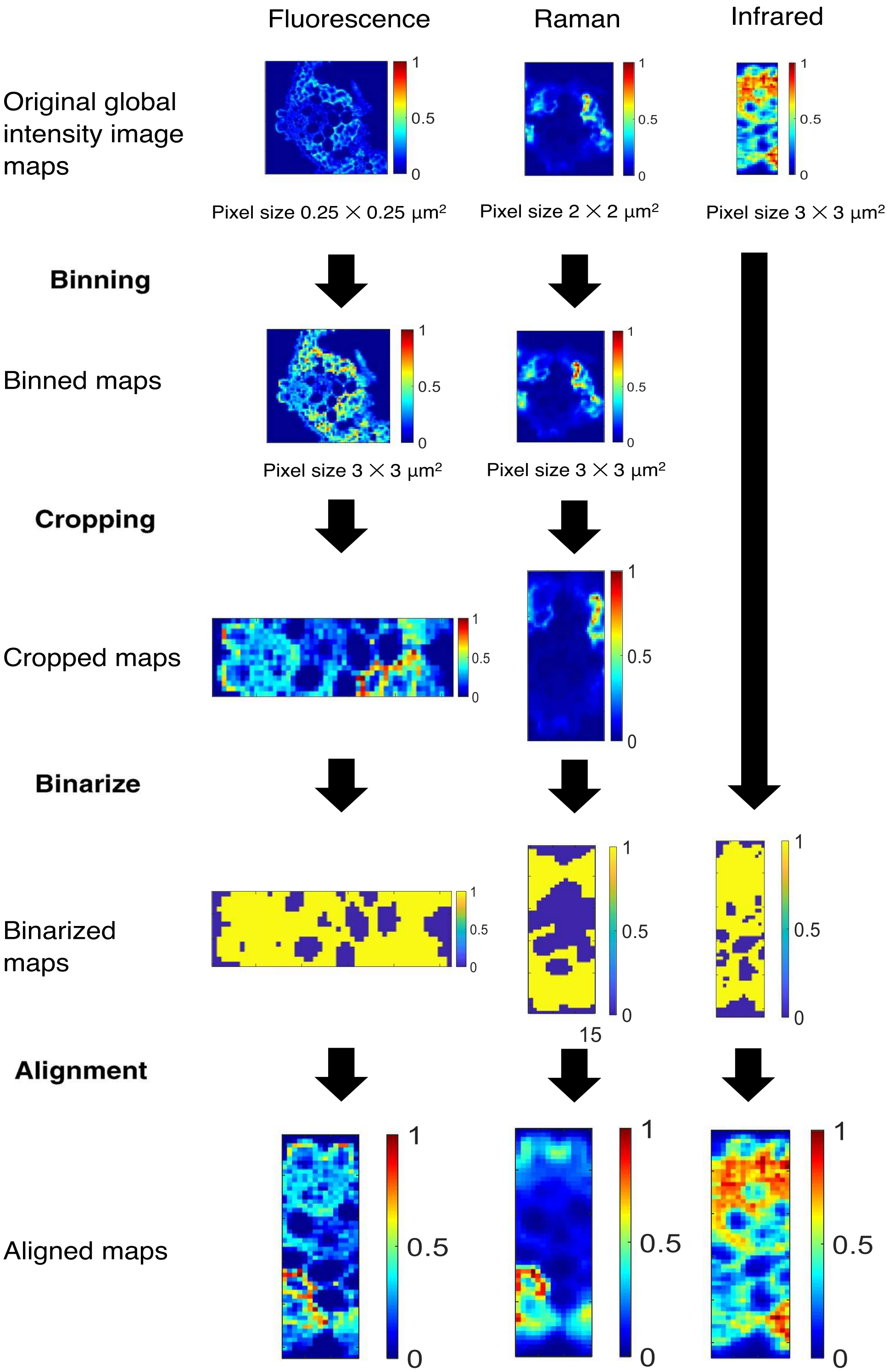


Figure 4. General scheme of the image matching procedure. In order, images are resized until match pixel size. Then, they are cropped until have approximately the same area covered. Next, the images are binarized and aligned among them. Once optimal translational and rotational parameters are achieved, hyperspectral images are aligned.

Results and discussion

MCR-ALS was used to elucidate the sample constituents present in the cross-sections of rice leaves analyzed by SR-IR, fluorescence and Raman HSI. To show the gain of global information obtained by fusing images

Multiset	Techniques	Nr. of components	LOF (%)	Explained variance (%)
1	Fluorescence	5	13	98
2	SR-FTIR	3	57	67
3	Raman	4	32	90
4	Fluorescence, SR-FTIR, Raman	6	31	91

Table 1. Summary of MCR-ALS results from the image multisets analyzed.

from different platforms, two different analyses were performed. On the one hand, three separate multisets (for fluorescence, for SR-IR and for Raman) containing three images each collected with the same platform were structured in a multiset extended in the column-wise direction and were subsequently analyzed by MCR-ALS. This per platform analysis gives a vision of the information that can be obtained without using image fusion. On the other hand, an MCR analysis of a multiset incorporating the images from all platforms (Fig. 3) was carried out to illustrate the gain of information linked to image fusion. To build the fused multiset, the alignment of the images previously described to achieve the congruence of pixels among platforms and the suitable balance between data blocks was carried out.

For all MCR-ALS analyses, the convergence criterion was 0.1% difference among lack of fit between consecutive iterations. The main results of MCR-ALS applied to the different multisets analyzed are summarized in Table 1.

As can be seen, the variance explained is satisfactory in all multisets taking into consideration the quality of the spectra analyzed. Thus, SR-FTIR provides the lowest variance explained due to the enhancement of noise when derivative spectra are used. Raman and fluorescence multisets show higher variance explained due to the quality of the original spectra. The analysis of the multiset using all platforms provides a good description of all images analyzed.

As it was expected, the results provided by the imaging platforms used in this work show differences in the number of components modelled due to the complementary information of the related spectroscopic techniques and the differences in the detectable response for the different biological tissues and molecules. These differences suggest the need of a multiplatform fusion to exploit the complementary information and achieve a complete description of the sample.

In the next subsections, a description of the components found by each spectroscopic technique and by the fused multiset containing all platforms is provided.

Fluorescence HSI multiset analysis.

Initial spectral estimates found by a SIMPLISMA-based method pointed out to the presence of components similar to those identified in a previous work[13]. Thus, some components were identified as chlorophylls, which emit in wavelengths higher than 625 nm, lignins in lower wavelengths and an additional component linked to small vesicles, probably oil or silica bodies, inside the mesophyll cells and epidermis in leaves, emitting between these two families of compounds.

The multiset could be described by five components, as suggested by PCA. In the MCR optimization, the non-negativity constraint was applied to the concentration profiles and spectral signatures of all components because emission spectra are not negative. Considering the prior information mentioned above, selectivity/local rank was also applied to the spectral signature of chlorophylls, set to have null emission in wavelengths lower than 625 nm, and to the component presumably linked to vesicles, set to have null emission below 505 nm and above 680 nm. Figure 5 shows the resolved spectral signatures and the distribution maps of the three samples used in the fusion of images of all platforms.

Five components were identified as natural fluorophores in leaves. Component yellow and blue are identified as chlorophylls, showing a maximum at 682 nm. The maps show clearly that chlorophylls are located at the mesophyll cells, where there are chloroplasts and biochemical activity, such as the photosynthesis[15]. Orange and green components could be lignins, since the emission range observed goes from 430 to 550 nm. As it can be observed in the distribution maps, lignins are present in plant cell walls, plant vascular system and in the epidermis of the leaves[15]. Finally, the purple component, based on its location and its shape as a droplet or vesicle, is presumably identified as a type of body-lipid or body-silica. Yellow fluorescence with a long range can be observed. The characterization of these vesicles was not possible using only the fluorescence emission.

SR-FTIR HSI multiset analysis.

Three components were suggested by PCA in all SR-FTIR HSI to describe the multiset. During the iterations, non-negativity was applied only to the concentration profiles of all components because pure signatures have negative values due to the second derivative preprocessing. Figure 6 shows the resolved spectral signatures and the distribution maps of the three samples analyzed. Three components could be identified with distinct IR spectral signatures. The blue component shows mainly protein bands (Amide I (1655 cm^{-1}) and Amide II (1543 cm^{-1})[31]. It is possible to observe the location of the proteins mainly in the mesophyll cells, where there are proteins as enzymes related to the biochemical activity. The orange component shows bands associated with lignin (carbonyl (1732 cm^{-1}))[32] and it is possible to observe its presence in the vascular system. The vascular system has cells fortified with lignin to give robustness to the plant. The yellow component is related to lipid bands (methylene groups (2916 and 2846 cm^{-1}))[32]. This yellow component is present in the epidermis. Often, leaves show a small layer of resins in their epidermis to avoid water loss. These results are in agreement with the biological compounds naturally present in leaves.

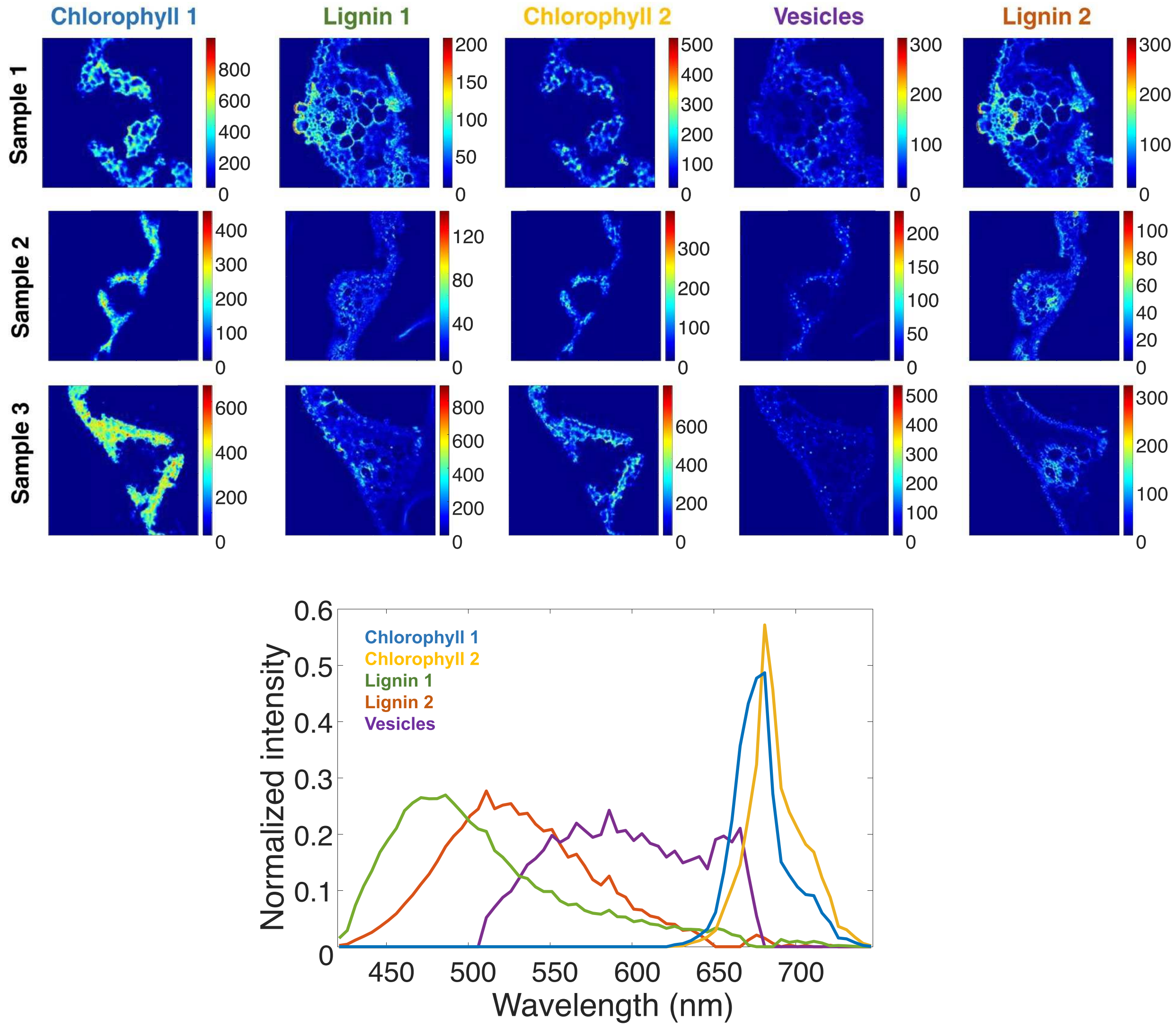


Figure 5. Bottom plot, pure fluorescence spectra profiles. It can be observed five different components. Lignins have an emission in blue and green, while. Vesicles in yellow, and chlorophylls in orange. Top plot, pure distribution maps are showed.

Raman HSI multiset analysis. Four components were suggested by PCA in all Raman HSI to describe the multiset. During the iterations, non-negativity constraint was applied to the concentration profiles and the pure spectral signatures of all components. In Fig. 7 it is possible to observe two components associated with biological contributions (in blue and orange) were identified, whereas two additional components (in gray) were attributed to instrumental noise detected in previous works[33]. The blue component was characterized as β-carotene (with typical Raman features at 1155 and 1525 cm^{-1})[34]. β-carotene is a strongly colored red–orange pigment and during photosynthesis β-carotene normally serves as antenna pigments, transferring singlet excitation energy to chlorophyll. Therefore β-carotene can be found at mesophyll cells, where chlorophyll is. The orange characterized component is lignin (with typical Raman features at 1598 and 1631 cm^{-1})[35]. Lignin can be found at the vascular system and sclerenchyma cells, which are strongly lignified.

Image fusion of fluorescence, SR-FTIR and Raman HSI. MCR-ALS was applied to identify in a complete way the constituents present in the rice leaves. Several MCR-ALS models were tested with different number of components. Six components were needed to explain the relevant variation in images. Adding more components did not provide additional interpretable information. Several initial estimates based on SIMPLISMA or on the connection of resolved signatures coming from multisets of individual techniques were tested.

In a data fusion, constraints can be applied in $\mathbf{C}_i$ and $\mathbf{S}_i^{\mathrm{T}}$ submatrices in different ways. In all analyses, non-negativity was applied to concentration profiles. Non-negativity was applied to fluorescence and Raman $\mathbf{S}_i^{\mathrm{T}}$ profiles, whereas SR-FTIR profiles were left unconstrained. Table 2 shows the identification of the six components resolved in the definitive MCR-ALS model. This identification was useful to set local rank constraints.

Several components are not detected by all techniques with the instrumental parameters used in this work. Thus, chlorophyll and lipids do not have Raman signal according to previous measurements. The contribution

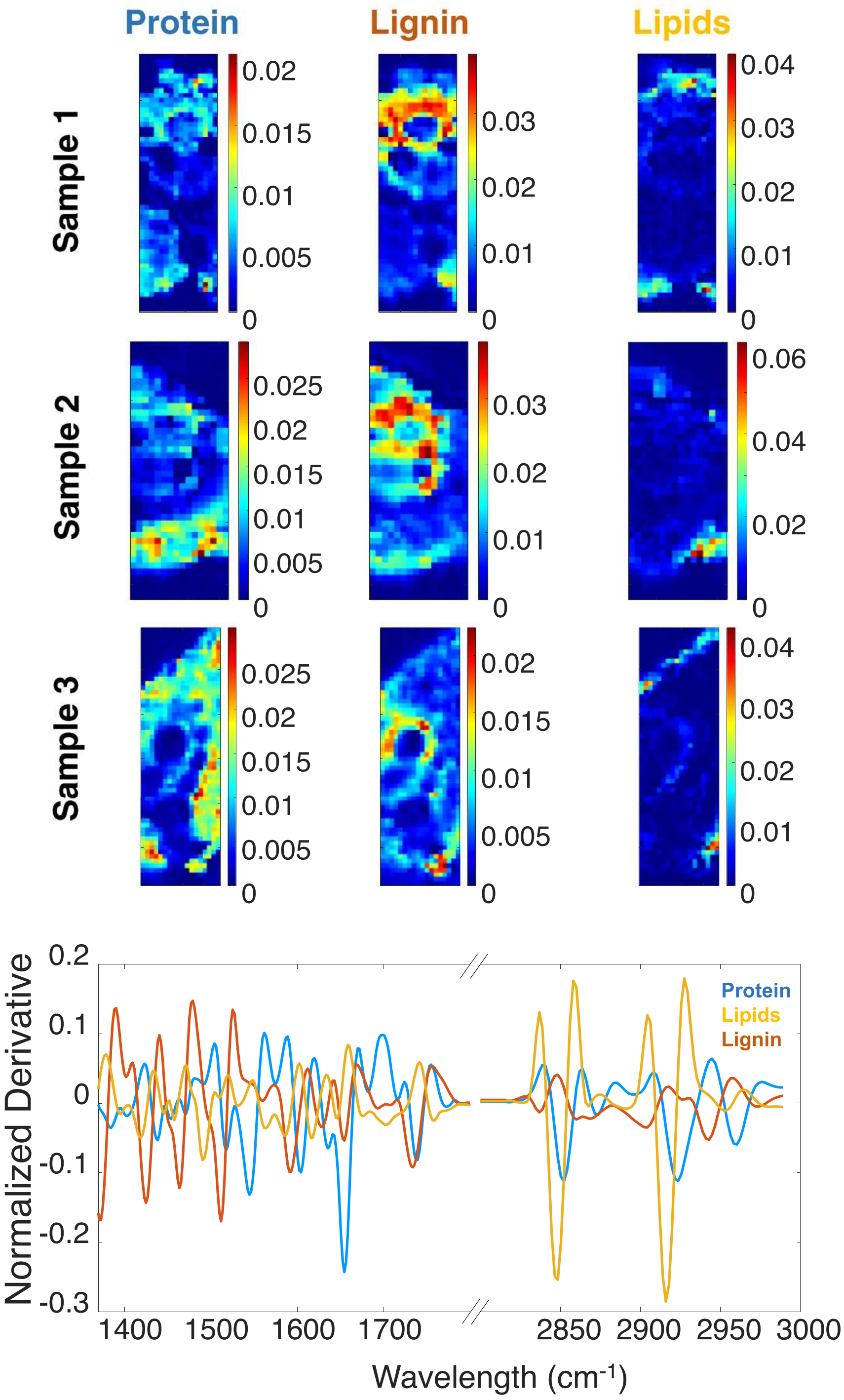


Figure 6. Top plot, pure distribution maps for the three components. Bottom plot, pure SR-FTIR spectral signatures of the component related to lipids (in yellow), to proteins (in blue) and lignin (orange).

most linked to proteins in SR-FTIR does not have either fluorescence or Raman signal. Furthermore, fluorescence from β-carotene was not detected. For these reasons, selectivity/local rank was applied to force null signals in all components that are not detected in the suitable technique. Other local rank constraints related to spectral regions with null fluorescence were also applied in the analysis of the multiset of all fused techniques.

The results of the MCR-ALS analysis are shown in Fig. 8. Six components were identified. The blue component was characterized as chlorophyll. It is possible to observe in the distribution maps that chlorophyll is located at mesophyll cells on all samples. The fluorescence pure signature exhibits a typical emission spectrum of chlorophyll with a maximum of 682 nm[15]. The infrared spectrum has bands that could be related to the chlorophyll structure (alkanes (2930 to 2840 cm^{-1}), ester (1741 cm^{-1}) and alkenes (1660 cm^{-1}). The green component was characterized as β-carotene. The component was located at mesophyll cells in distribution maps and Raman

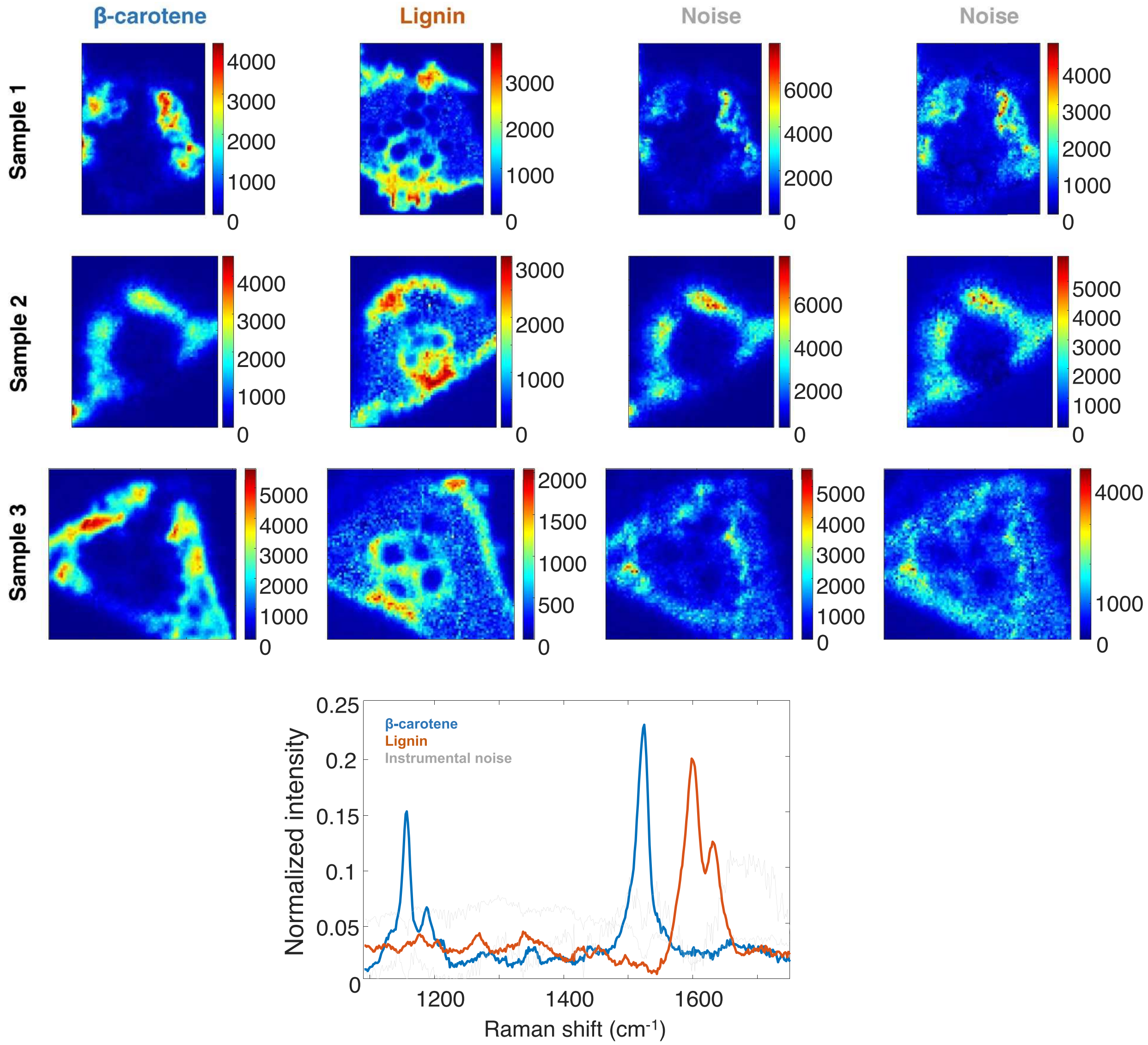


Figure 7. Top plot, pure distribution maps of the resolved components in the three samples. β-carotene, first column and lignin in the second column, the last two columns show the distribution maps related to the noise components. Bottom plot, MCR-ALS resolved pure Raman spectra profiles related to carotenes and lignin. In gray, the two spectral profiles related to noise.

		Local rank constraint in S^T spectral range (forced to be zero)		
Component	Identified as	Fluorescence	SR-IR	Raman
1	Chlorophyll	420–625	–	All*
2	Lignin 1	–	–	–
3	Lipids	505–680	–	All*
4	Protein	All*	–	All*
5	β-carotene	All*	–	–
6	Lignin 2	–	–	–

Table 2. Summary of the components identified and the imposed selectivity/local rank. *No signal was detected for the technique in the related component. Note that for infrared interval, no selectivity constraint was imposed. Several components were forced to be zero (labelled ‘All’) in certain techniques. Finally, for lipids and chlorophyll, some spectral regions of the fluorescence spectra were forced to be zero.

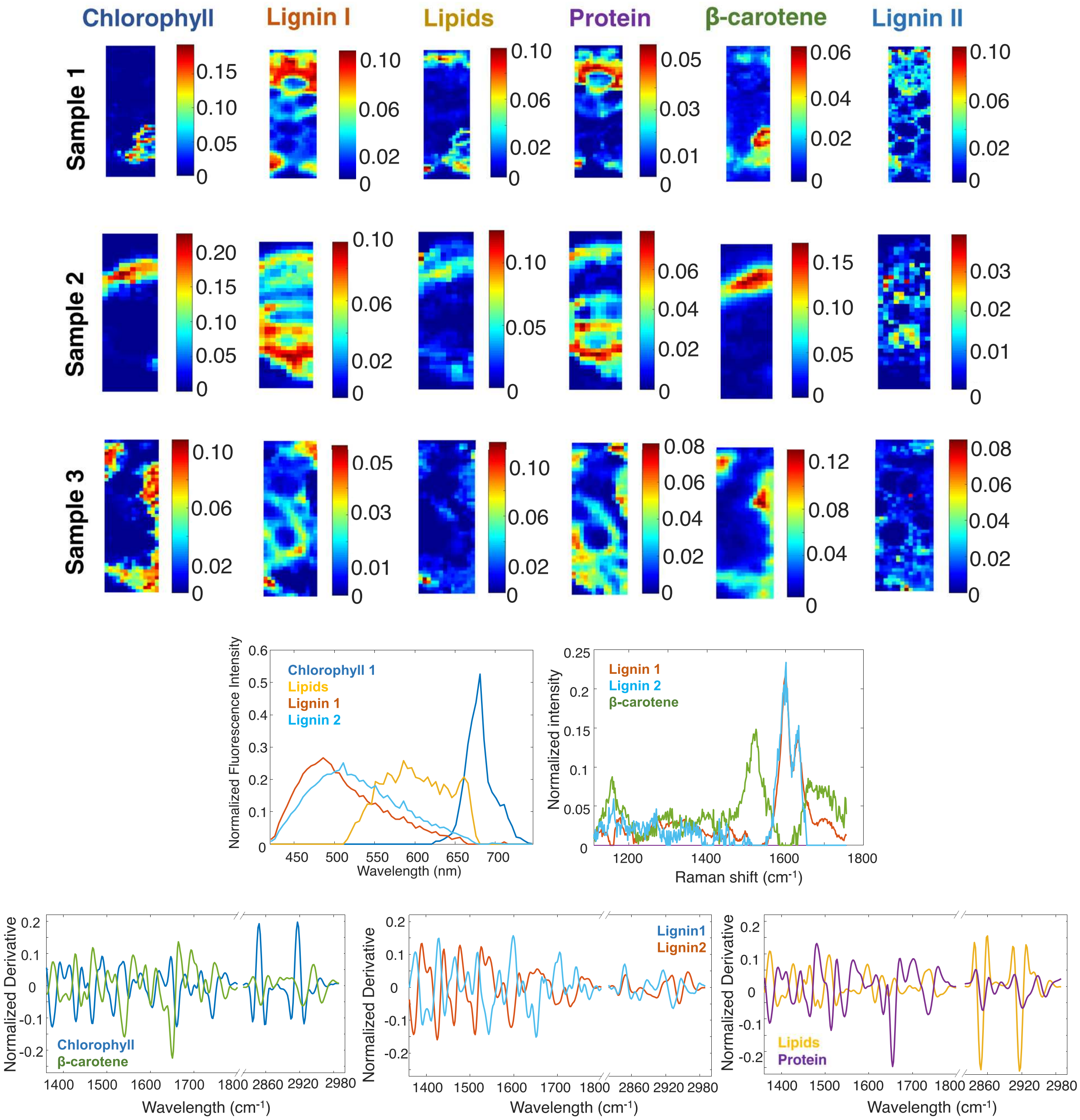


Figure 8. Top, pure distribution maps of the six components identified with the fused images for the three samples. Bottom, pure spectral signatures of fluorescence, Raman and SR-FTIR. SR-FTIR spectra were broken down in plots of two components for better visualization.

pure signature shows the typical Raman peaks at 1525 and 1157 cm^{-1}[34]. As is expected, β-carotene appears in the same leaf zone as chlorophyll. This component also shows a relevant protein band in SR-FTIR. This may be the consequence of β-carotenes binding specifically to some protein receptors. Orange and cyan components were characterized as types of lignin. Lignin can be observed on concentration map at vascular tissues. The presence of lignin was high in the sclerenchyma cells as well. For the orange component, the pure fluorescence spectrum has a maximum at 487 nm, while a maximum at 512 nm is observed for the cyan lignin. The pure infrared spectrum of the orange component has a band with maximum at 1730 cm^{-1} (carbonyl), but it was not observed for the cyan component. The pure Raman signatures of both components have two typical Raman features from lignin (1632 cm^{-1} and 1601 cm^{-1})[35]. The yellow component was identified as rich in lipids. The pure fluorescence spectral signature coincides with the pure signature presumably attributed to the epidermis and the vesicles. The SR-FTIR pure signature confirms the identity of this component, with two strong peaks at the lipid region (2916 and 2848 cm^{-1})[32]. Vesicles were not possible to be clearly observed in the concentration maps, probably due to the spatial binning.

The purple component was identified as rich in proteins. The pure infrared spectrum shows a strong peak at 1655 cm^{-1}, typically of the group amide of the proteins[31]. This component was located in structural regions in distribution map. This result can indicate that these proteins could interact with plant structural elements.

As it can be observed, the fusion gives richer information than the individual analysis since the distinction of components becomes easier due to their now extended multitechnique spectroscopic signature.

There is a clear synergic effect linked to the compensation of weak properties of a technique by stronger points in another one. For instance, fluorescence images have a good spatial resolution but poor spectroscopic features to identify the nature of many compounds, i.e. lipids were not possible to be identified using only fluorescence images since the fluorescence shape is not selective of functional groups. Instead, the fusion with SR-FTIR images allows characterizing unequivocally this component through characteristic bands in the infrared region, with much richer in spectral features. Along this line, the infrared spectra associated with Raman or fluorescence signatures help to identify molecular compounds (lipids and proteins) linked to typical constituents found in plant tissues (carotenes, lignin, chlorophylls, …).

The image fusion also allows distinguishing components with very similar signatures in a technique taking advantage of the clear distinction of the same components in another fused technique. The lignin components depict this situation. On the one hand, the difference in spectral signatures of the two lignin contributions in fluorescence helps in the distinction of the variants of the same compound in Raman spectroscopy, impossible to achieve when Raman images were analysed alone. On the other hand, the very characteristic Raman features for lignin help to confirm the identity of these components, a task more difficult to do only based on the fluorescence information. The increase in discriminating power is also seen in the six resolved infrared signatures in image fusion, which were reduced to three components when this technique was analyzed alone.

Conclusions

The operating procedure related to image fusion in a multiplatform scenario has been clearly described and the steps detailed, from the data preprocessing and image matching to the unmixing with MCR-ALS multiset analysis and interpretation of information can be generally applied to perform a complete characterization of the components in any imaged sample. The great benefits of joining different kinds of spectroscopic information for a better morphological and chemical characterization of components has been clearly proven in a case study linked to a vegetal tissue.

The fusion strategy presented is the basic pipeline for many image fusion situations that can be encountered in practice. However, it is relevant to know that fusion approaches are being developed recently to compensate for drawbacks, such as the possible presence of relevant components located in image areas not common to all images or the loss of spatial resolution of some techniques to achieve pixel congruence with techniques that provide a lower level of spatial detail. Additionally, the adaption of algorithms that can combine images providing a linear spectrum per pixel, e.g., Raman, infrared, with others yielding a 2D spectroscopic landscape per pixel, e.g., excitation-emission spectra, has also been proposed. Although these approaches are not generally used yet, they open a new direction to make image fusion even more powerful.

Received: 28 May 2021; Accepted: 24 August 2021
Published online: 20 September 2021

References

1. Salzer, R. & Siesler, H. W. (eds) *Infrared and Raman Spectroscopic Imaging* (Wiley, 2014).
2. Amigo, J. M. Hyperspectral and multispectral imaging: Setting the scene. In *Data Handling in Science and Technology*, Vol. 32 (ed. Amigo, J. M.) 3–16 (Elsevier, 2020).
3. de Juan, A., Gowen, A., Duponchel, L. & Ruckebusch, C. Image fusion. In *Data Handling in Science and Technology*, Vol. 31 (ed. Cocchi, M.) 311–344 (Elsevier, 2019).
4. Borsoi, R. A., Imbiriba, T. & Bermudez, J. C. M. Deep generative endmember modeling: An application to unsupervised spectral unmixing. *IEEE Trans. Comput. Imaging* **6**, 374–384 (2019).
5. Palsson, B., Sigurdsson, J., Sveinsson, J. R. & Ulfarsson, M. O. Hyperspectral unmixing using a neural network autoencoder. *IEEE Access* **6**, 25646–25656 (2018).
6. Bioucas-Dias, J. M. & Plaza, A. An overview on hyperspectral unmixing: Geometrical, statistical, and sparse regression based approaches. In *2011 IEEE International Geoscience and Remote Sensing Symposium*, 1135–1138 (2011).
7. de Juan, A. & Tauler, R. Multivariate curve resolution: 50 years addressing the mixture analysis problem - a review. *Anal. Chim. Acta* **1145**, 59–78 (2021).
8. de Juan, A., Maeder, M. & Tauler, R. Multiset data analysis: Extended multivariate curve resolution. In *Comprehensive Chemometrics* Vol. 2 (eds Brown, S. *et al.*) 305–336 (Elsevier, 2020).
9. de Juan, A. Multivariate curve resolution for hyperspectral image analysis. In *Data Handling in Science and Technology*, Vol. 32 (ed. Amigo, J. M.) 115–150 (Elsevier, 2020)
10. Dobigeon, N., Altmann, Y., Brun, N. & Moussaoui, S. Linear and nonlinear unmixing in hyperspectral imaging. In *Data Handling in Science and Technology*, Vol. 30 (ed. Ruckebusch, C.) 185–224 (Elsevier, 2016).
11. Piqueras, S. *et al.* Handling different spatial resolutions in image fusion by multivariate curve resolution-alternating least squares for incomplete image multisets. *Anal. Chem.* **90**(11), 6757–6765 (2018).
12. Bedia, C., Sierra, À. & Tauler, R. Application of chemometric methods to the analysis of multimodal chemical images of biological tissues. *Anal. Bioanal. Chem.* **412**(21), 5179–5190 (2020).
13. Gómez-Sánchez, A., Marro, M., Marsal, M., Loza-Alvarez, P. & de Juan, A. 3D and 4D image fusion: Coping with differences in spectroscopic modes among hyperspectral images. *Anal. Chem.* **92**(14), 9591–9602 (2020).
14. Mas, S. *et al.* Use of physiological information based on grayscale images to improve mass spectrometry imaging data analysis from biological tissues. *Anal. Chim. Acta* **1074**, 69–79 (2019).
15. Donaldson, L. Autofluorescence in plants. *Molecules* **25**(10), 2393 (2020).
16. Gierlinger, N., Keplinger, T. & Harrington, M. Imaging of plant cell walls by confocal Raman microscopy. *Nat. Protoc.* **7**(9), 1694–1708 (2012).

17. Movasaghi, Z., Rehman, S. & Rehman, D. I. Fourier transform infrared (FTIR) spectroscopy of biological tissues. *Appl. Spectrosc. Rev.* **43**(2), 134–179 (2008).
18. Yu, P. *et al.* Chemical imaging of microstructures of plant tissues within cellular dimension using synchrotron infrared microspectroscopy. *J. Agric. Food Chem.* **51**(20), 6062–6067 (2003).
19. Savitzky, A. & Golay, M. J. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **36**(8), 1627–1639 (1964).
20. Eilers, P. H. Parametric time warping. *Anal. Chem.* **76**(2), 404–411 (2004).
21. Jolliffe, I. T. Principal components in regression analysis. In *Principal Component Analysis* 129–155 (Springer, 1986).
22. Windig, W. & Guilment, J. Interactive self-modeling mixture analysis. *Anal. Chem.* **63**(14), 1425–1432 (1991).
23. Bro, R. & De Jong, S. A fast non-negativity-constrained least squares algorithm. *J. Chemom. J. Chemom. Soc.* **11**(5), 393–401 (1997).
24. Tauler, R., Smilde, A. & Kowalski, B. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J. Chemom.* **9**(1), 31–58 (1995).
25. de Juan, A., Maeder, M., Hancewicz, T. & Tauler, R. Use of local rank-based spatial information for resolution of spectroscopic images. *J. Chemom. J. Chemom. Soc.* **22**(5), 291–298 (2008).
26. de Juan, A., Maeder, M., Hancewicz, T. & Tauler, R. Local rank analysis for exploratory spectroscopic image analysis. Fixed size image window-evolving factor analysis. *Chemom. Intell. Lab. Syst.* **77**(1–2), 64–74 (2005).
27. Hugelier, S., Devos, O. & Ruckebusch, C. On the implementation of spatial constraints in multivariate curve resolution alternating least squares for hyperspectral image analysis. *J. Chemom.* **29**(10), 557–561 (2015).
28. Ghaffari, M., Hugelier, S., Duponchel, L., Abdollahi, H. & Ruckebusch, C. Effect of image processing constraints on the extent of rotational ambiguity in MCR-ALS of hyperspectral images. *Anal. Chim. Acta* **1052**, 27–36 (2019).
29. Piqueras, S., Maeder, M., Tauler, R. & de Juan, A. A new matching image preprocessing for image data fusion. *Chemom. Intell. Lab. Syst.* **164**, 32–42 (2017).
30. Jaumot, J., de Juan, A. & Tauler, R. MCR-ALS GUI 2.0: New features and applications. *Chemom. Intell. Lab. Syst.* **140**, 1–12 (2015).
31. Krimm, S. & Bandekar, J. Vibrational spectroscopy and conformation of peptides, polypeptides, and proteins. *Adv. Protein Chem.* **38**, 181–364 (1986).
32. Heredia-Guerrero, J. A. *et al.* Infrared and Raman spectroscopic features of plant cuticles: a review. *Front. Plant Sci.* **5**, 305 (2014).
33. Olmos, V. *et al.* Combining hyperspectral imaging and chemometrics to assess and interpret the effects of environmental stressors on zebrafish eye images at tissue level. *J. Biophotonics* **11**(3), e201700089. https://doi.org/10.1002/jbio.201700089 (2018).
34. Tschirner, N. *et al.* Resonance Raman spectra of β-carotene in solution and in photosystems revisited: an experimental and theoretical study. *Phys. Chem. Chem. Phys.* **11**(48), 11471–11478 (2009).
35. Zhang, X., Chen, S. & Xu, F. Combining Raman imaging and multivariate analysis to visualize lignin, cellulose, and hemicellulose in the plant cell wall. *J. Vis. Exp. JoVE* **124**, 55910. https://doi.org/10.3791/55910 (2017).

Acknowledgements

We are grateful to ALBA synchrotron facilities for enabling the SR-FTIR measurements, especially to Imma Martínez Rovira and Ibraheem Yousef for helping in the SR-FTIR measurements. A.G.-S., R.R.O. and A.J. acknowledge financial support from the Spanish government Project PID 2019-1071586B-IOO and the Catalan government (2017 SGR 753). ICFO acknowledges financial support from the Spanish Ministry of Economy and Competitiveness through the "Severo Ochoa" program for Centres of Excellence in R&D (SEV-2015-0522), from Fundació Privada Cellex, Fundacion Mig-Puig, Generalitat de Catalunya through the CERCA program, and "Laser lab Europe" (No. 654148).

Author contributions

A.G.-S. and A.J. are responsible for the conception of the work, performed the experiments, analyzed the data and wrote the manuscript. M.M., M.M., R.R.O., S.Z., P.L.-A. performed the experiments and made a substantial revision of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-98000-0.

Correspondence and requests for materials should be addressed to A.G.-S. or A.d.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary material

LINEAR UNMIXING PROTOCOL FOR HYPERSPECTRAL IMAGE FUSION ANALYSIS APPLIED TO A CASE STUDY OF VEGETAL TISSUES

Adrián Gómez-Sánchez[1],*, Mónica Marro[2], Maria Marsal[2], Sara Zacchetti[1,3], Rodrigo Rocha de Oliveira[1], Pablo Loza-Alvarez[2], Anna de Juan[1]*

Abstract

The supporting information includes S1 illustrate image related to the preprocessing of each spectroscopic technique.

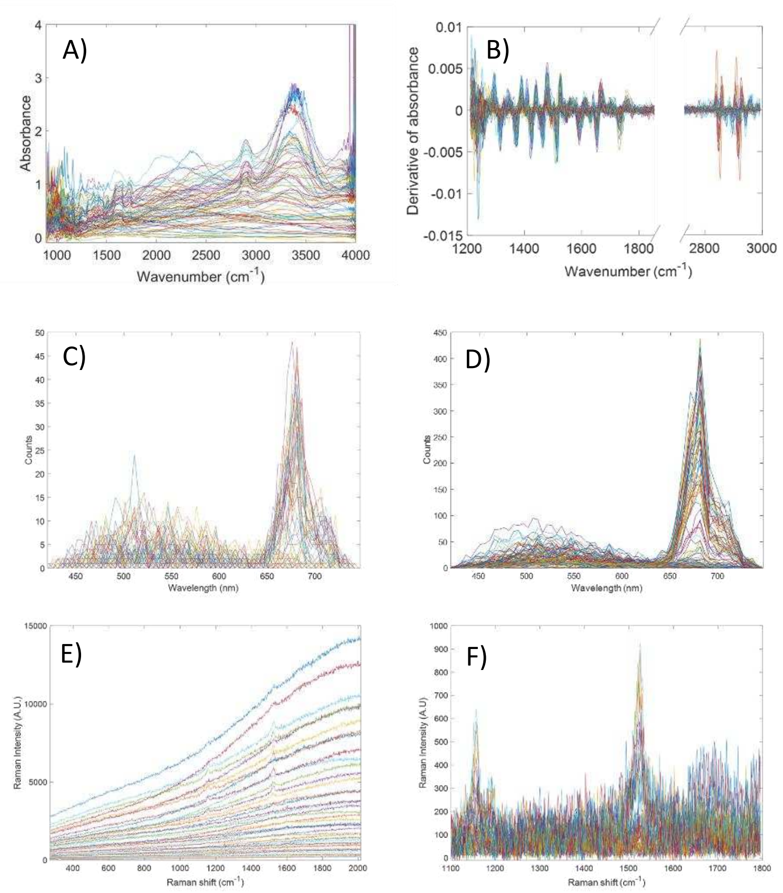


Figure S1. Image preprocessing. A) Raw synchrotron infrared spectra. B) Spectra after preprocessing. C) Original fluorescence spectra. D) Fluorescence spectra after binning. E) Raw Raman spectra. F) Preprocessed Raman spectra.

Publication VI shows an example of image fusion by combining fluorescence, Raman and synchrotron radiation infrared (SR-IR) hyperspectral images to investigate a case study devoted to characterize the natural compounds of cross-sections of rice leaves.

The aim of this work was to establish an open-access protocol for classical image fusion analysis based on MCR-ALS and to show the benefits and drawbacks of this procedure. To do so, the results of the independent analysis of each typology of HSIs considering one spectroscopic technique at a time are first presented. Afterwards, the systematic protocol followed for image fusion is described and, finally, the results of the analysis of the single fused multiset by MCR-ALS are shown. The improvement with respect to the individual technique analysis can be noticed as well as the limitations associated with classical fusion, which are the focus of the developments linked to **Publications VIII** and **IX**.

<u>**Unmixing analysis of HSIs from single platforms**</u>

For all HSIs techniques, MCR-ALS has been applied to analyze a column-wise augmented data set (see Fig. 15A of Chapter 2) formed by the pixel spectra of three rice leaf cross-section images. The number of components of the MCR model was estimated by PCA and the initial estimates based on a SIMPLISMA-based method. Non-negativity constraints were used in all concentration maps and the suitable constraints for the pure spectra of each particular technique can be found in **Publication VI.** The MCR results consist of a pure $\mathbf{S}^T$ matrix and concentration maps for all components in the three different samples are obtained.

Analysis of fluorescence hyperspectral images.

Figure 37 shows the results of the multiset analysis by MCR-ALS. The analysis identified five potential natural fluorophores in leaves. The distribution maps visually demonstrated the spatial consistency of these components with the leaf structures across all three samples. According to the pure spectra, yellow and blue components were identified as chlorophyll contributions, with a maximum at 682 nm, in agreement with reported spectra in literature [Donaldson, 2020]. The distribution maps of chlorophylls are uniquely located on mesophyll cells, where photosynthesis occurs. Orange and green components are likely lignin contributions, found in plant cell walls, vascular systems and leaf epidermis and present pure emission spectra similar to those found in literature [Donaldson, 2020]. Lignins are complex polymers found in the plant cell walls, as can be clearly seen in the related pure distribution maps. The purple component seems to appear as a droplet or vesicle, and was tentatively identified as a body-lipid or body-silica. It shows a pure fluorescence spectrum with an extensive emission range. However, the characterization of these vesicles was challenging using only fluorescence emission data.

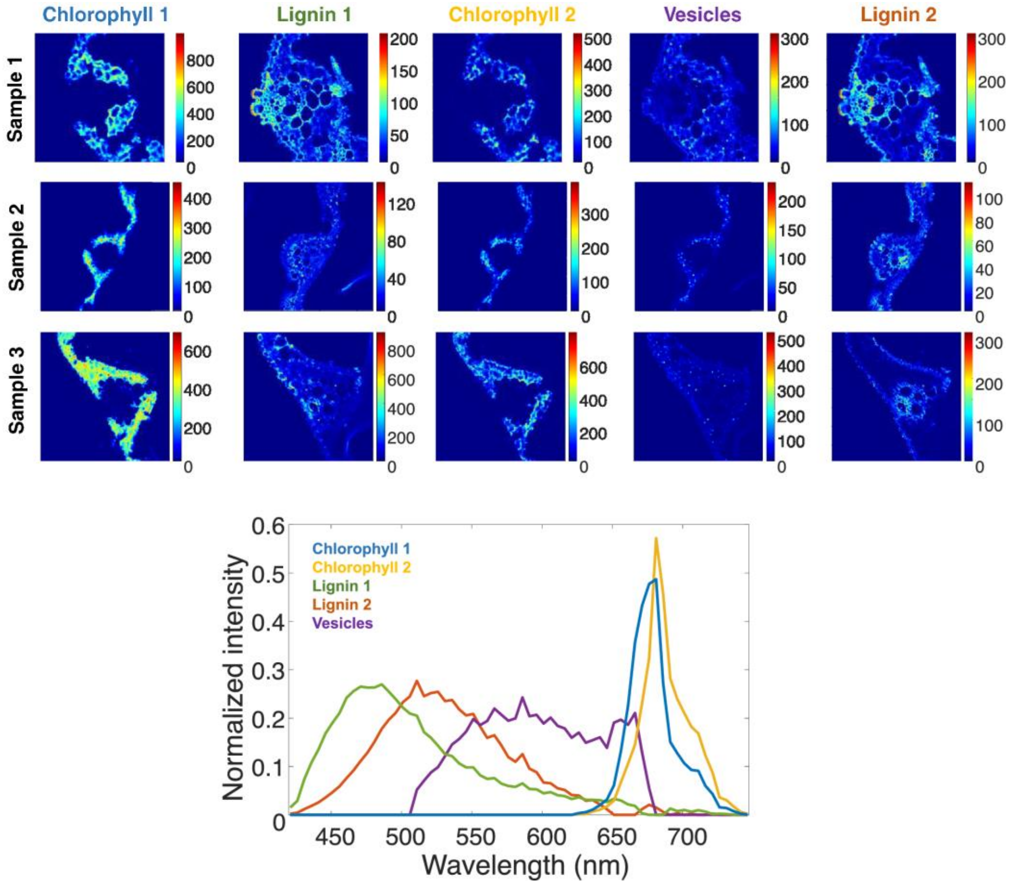


Figure 37. Results of the fluorescence multiset analysis by MCR-ALS. Top plot, pure distribution maps. Bottom plot, pure fluorescence spectra profiles. The analysis shown five components, characterized as two lignin contributions (presence in the vascular system of the leaf and a blue and green emission), two chlorophyll contributions (presence in the mesophyll cells of the leaf and showing a red fluorescence) and an uncharacterized component, related to vesicles spread around the vegetal tissue, mainly in the mesophyll cells and showing a yellow fluorescence.

Generally, fluorescence images yield interesting results due to the presence of many natural fluorophores in plant tissues. The spatial resolution of fluorescence provides very detailed information across distribution maps of various components, but the fluorescence spectra do not present very specific features, required for an accurate characterization of components.

Analysis of SR-IR hyperspectral images.

Multiset analysis of the three SR-IR hyperspectral images of the three rice leaf cross-sections was performed on second derivative spectra to remove offsets and linear baselines and to enhance the separation of overlapping peaks through the Savitzky-Golay algorithm [Savitzky and Golay, 1964].

Figure 38 shows the resolved spectral signatures and the distribution maps of the three analyzed samples by MCR-ALS. Three components were identified, each with characteristic IR spectral signatures of biological compounds and consistent distribution maps across samples. The blue component primarily shows protein bands (Amide I at 1655 cm^{-1} and Amide II at 1543 cm^{-1}). The spatial distribution could reveal the presence of proteins, such as enzymes associated with biochemical activity, concentrated in the mesophyll cells. The orange component shows bands that may be associated with lignin (carbonyl at 1732 cm^{-1}), and is spatially distributed in the vascular system. Finally, the yellow component is linked to lipid bands (groups at 2916 and 2846 cm^{-1}) and is predominantly found in the epidermis. It is common for leaves to exhibit a thin layer of resins in the epidermis to mitigate water loss. Therefore, this component could be related with the presence of a lipidic component.

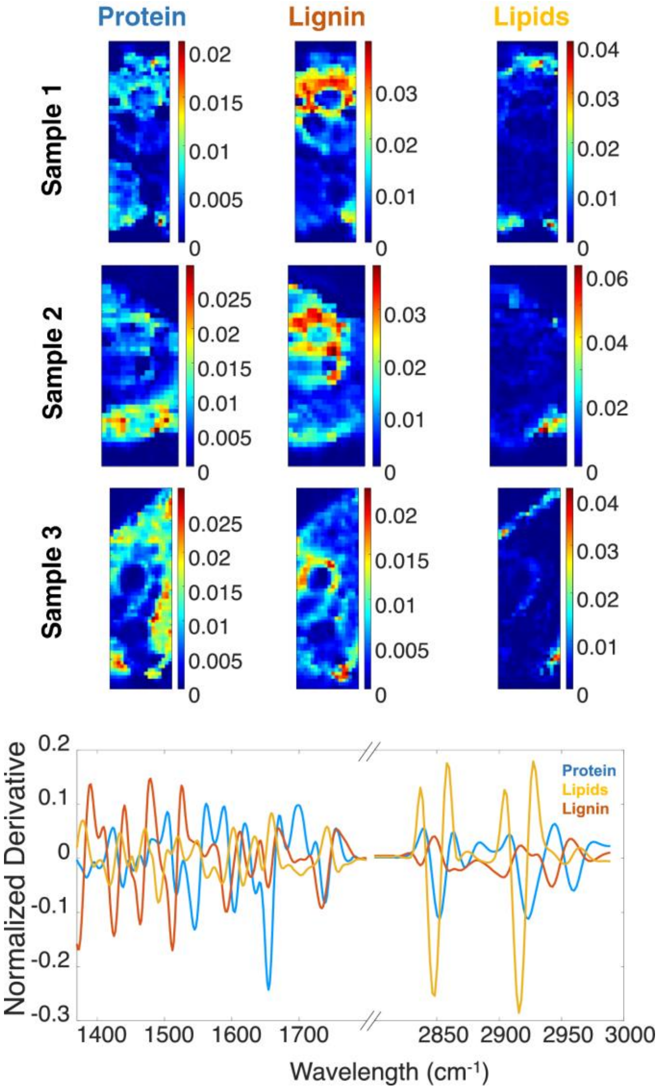


Figure 38. Results of the infrared multiset analysis by MCR-ALS. Top plot, pure distribution maps for the three components. Bottom plot, pure SR-IR spectral signatures of the component related to lipids (in yellow), to proteins (in blue) and lignin (orange).

It is important to note that the spatial resolution of infrared imaging may not be sufficient to resolve some components at a cellular level. For instance, while fluorescence imaging can capture the component related to small vesicles, the limited spatial resolution of infrared imaging prevents their observation. Nevertheless, this technique provides much richer spectroscopic features, useful to chemically characterize various components across the distribution maps.

Analysis of Raman hyperspectral images.

Figure 39 shows the results of the MCR-ALS analysis of Raman images. Two components associated with biological contributions (in blue and orange) were observed, while two additional components (in gray) are attributed to instrumental noise or artifacts due to the residual fluorescence after baseline correction, as identified in previous studies [Olmos et al., 2018].

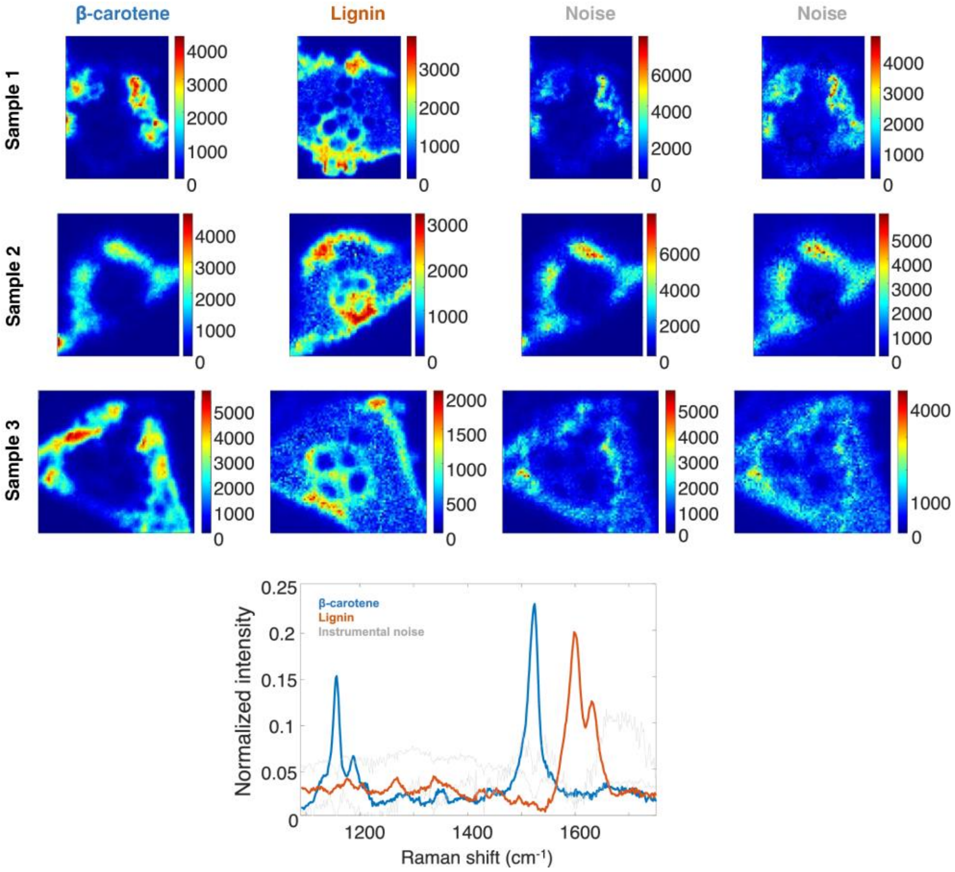


Figure. 39. Top plot, pure distribution maps for the four Raman components. Bottom plot, pure Raman spectral signatures of the component related to lignin (orange), to β-carotene (in blue) and instrumental artefacts (light-gray).

The blue component was identified as β-carotene, exhibiting typical Raman features at 1155 and 1525 cm^{-1} [Tschirner et al., 2009]. β-carotene is a pigment molecule in photosynthetic organisms that captures light energy and transfers it to the chlorophyll molecules in the reaction center of the photosystem. As a result, β-carotene is predominantly located in mesophyll cells, where chlorophyll is concentrated. On the other hand, the orange-characterized component corresponds to lignin, featuring typical Raman peaks at 1598 and 1631 cm^{-1} [Zhang et al., 2017].

Raman imaging provides an effective combination of high spatial resolution and spectral features. However, the image acquisition process can be relatively slow. To address this, Raman images were obtained using a pixel size of 2×2 µm, losing partially the possibility to acquire exploit the potential high spatial resolution that this technique can offer. Additionally, the presence of fluorescence in certain plant tissues, as observed in this study, poses a challenge to the analysis, lowering data quality due to the significant presence of Poisson noise after baseline correction.

Building a multiplatform multiset. Aspects to be taken into account.

Ensuring consistent sample positioning for all imaging techniques and accurately measuring the same area across images is very challenging, if not impossible, because variations in pixel coordinates and rotations are common among images. However, to build a multiplatform multiset, the HSIs must be spatially aligned to allow the row-wise augmentation, i.e., the concatenated pixel spectra in each row should refer to the sample pixel area. Achieving pixel congruency among hyperspectral images from fluorescence, Raman, and synchrotron infrared spectroscopy platforms requires a spatial preprocessing procedure, involving several steps. First, equalizing the pixel size among imaging platforms. Second, selecting a common scanned area. Third, binarizing the global intensity map of the different techniques to facilitate the spatial transformation of the images by translation and rotation until pixel congruence among them is achieved [Piqueras et al., 2017]. Fig. 40 shows graphically these preprocessing steps.

It is interesting to compare the initial information present in the individual images acquired and the structure of the final information to be fused in the multiplatform multiset to understand some of the results that will be commented afterwards. The original pixel size of fluorescence images was 0.25×0.25 $µm^2$, for Raman images was 2×2 $µm^2$ and for infrared images 3×3 $µm^2$. In terms of area scanned, fluorescence and Raman images span a much larger area than SR-IR images. The final information fused in the multiset needs to be spatially equivalent and congruent among images. In this context, it means that the pixel size will be 3×3 $µm^2$, matching the size of the SR-IR image with worse spatial resolution, and the sample surface analyzed will be restricted to the common scanned area by the three imaging techniques, defined by the SR-IR image as well. Although working with complementary spectroscopic signals is the main asset of image fusion, it is important to realize that a lot of useful and available information in terms of scanned areas and spatial detail is lost on this process.

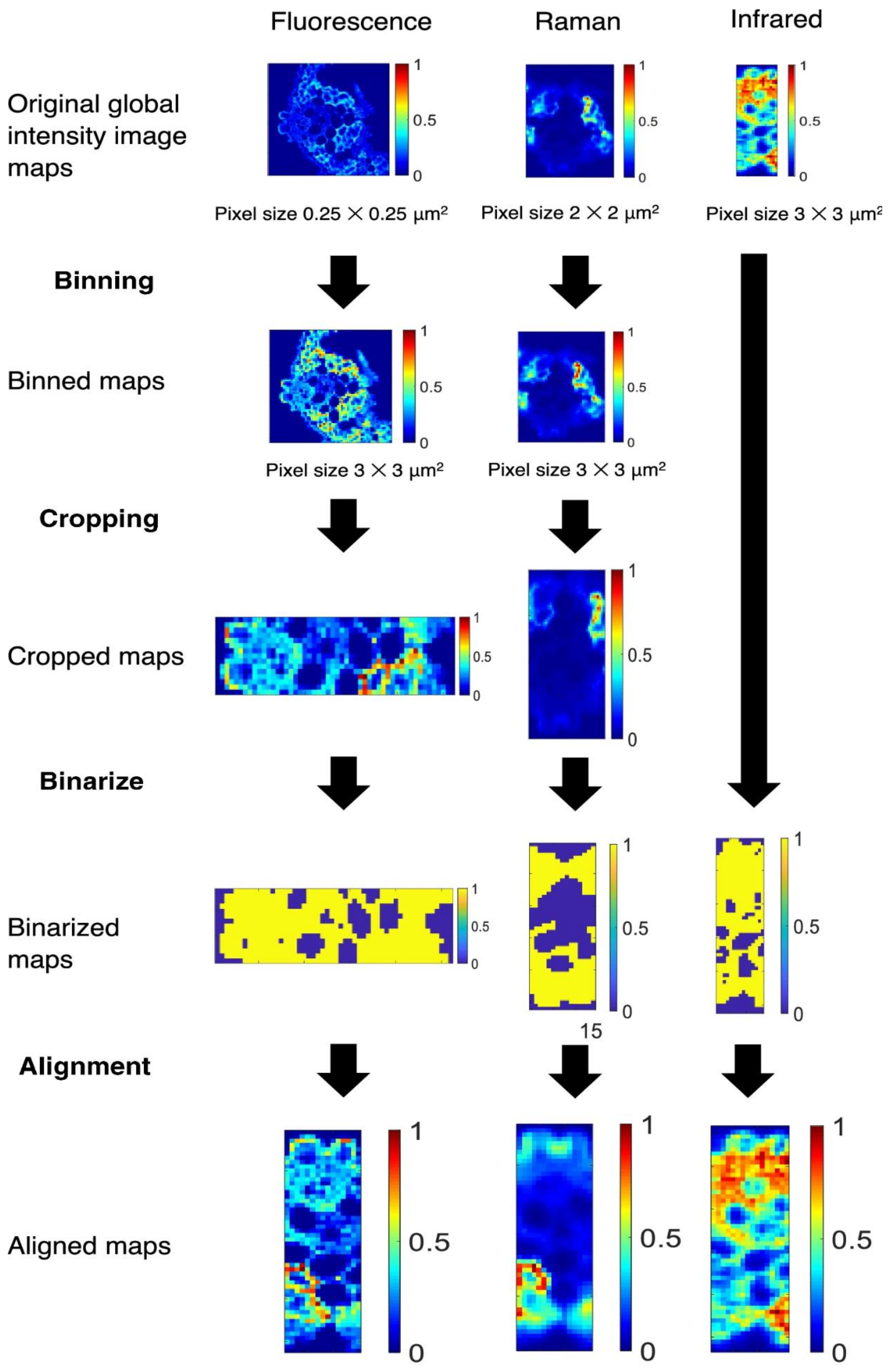


Figure 40. Schematic representation of the spatial transformations needed to align the HSIs. Initially, the images are resized until they share a common pixel size. Subsequently, they are cropped to cover approximately the same area. Afterwards, the corresponding binarized maps (based on the global intensity map or others approximations) are aligned to a reference. After determining the optimal translation and rotation parameters, the hyperspectral images are aligned accordingly.

An additional aspect to consider when building multiplatform multisets is the relative importance that each image block has in the fused structure to be analyzed. The signal intensity and the number of spectral channels can differ a lot among platforms and this causes severe differences in terms of the variance attributed to each imaging technique. To ensure a balanced representation of all imaging platforms, the data blocks of each technique can be normalized, usually dividing each element by the suitable 2-norm. After normalization, the blocks of the aligned images are concatenated to form a single multiset, as shown in Fig. 41.

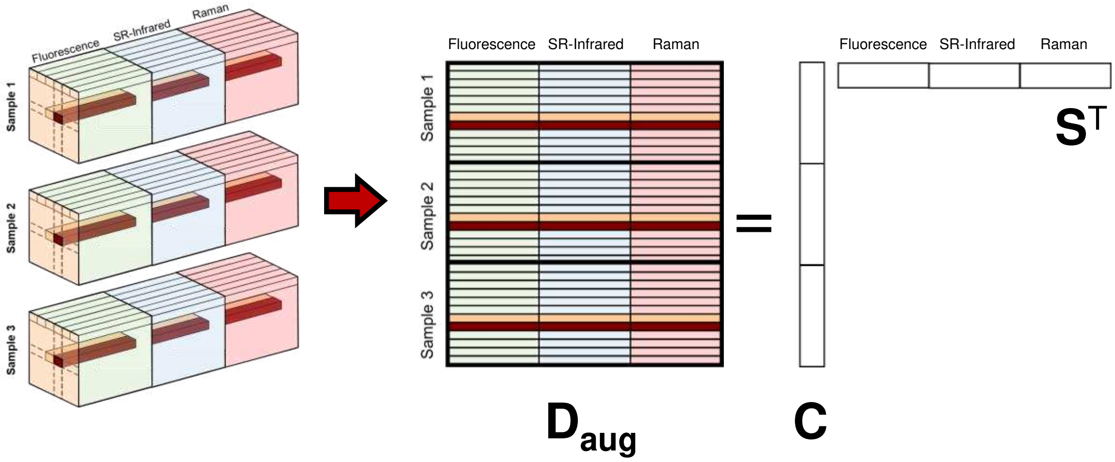


Figure 41. Structure of the augmented multiset **Daug** and the associated bilinear model for the three rice leaf cross-sections monitored by the fluorescence, SR-IR and Raman imaging techniques. The multiset has nine data blocks (three samples imaged by three spectroscopic platforms).

Unmixing analysis of the multiplatform multiset (image fusion results)

MCR-ALS is applied to the multiset displayed in Fig. 41 using the appropriate constraints for the profiles in each block of matrices **C** and/or **S**^{T} (see **Publication VI** for more detail). The bilinear model provides extended spectral signatures concatenating the fluorescence, SR-IR and Raman signal, very helpful for the characterization of every component and distribution maps for all components in the three samples. Before describing the results obtained, Table 1 shows the comparison between the number of components retrieved in the individual analysis of the different imaging techniques and in the image fusion approach.

Table 1 Summary of MCR-ALS results from the image multisets analyzed.

Multiset	Techniques	Nr. of components	LOF (%)	Explained variance (%)
1	Fluorescence	5	13	98
2	SR-FTIR	3	57	67
3	Raman	4	32	90
4	Fluorescence, SR-FTIR, Raman	6	31	91

A first observation shows that six components can be modelled with image fusion, a number higher than that provided by any individual technique. This fact confirms the major differentiation ability provided by the joint use of the complementary spectroscopic information provided by the different techniques.

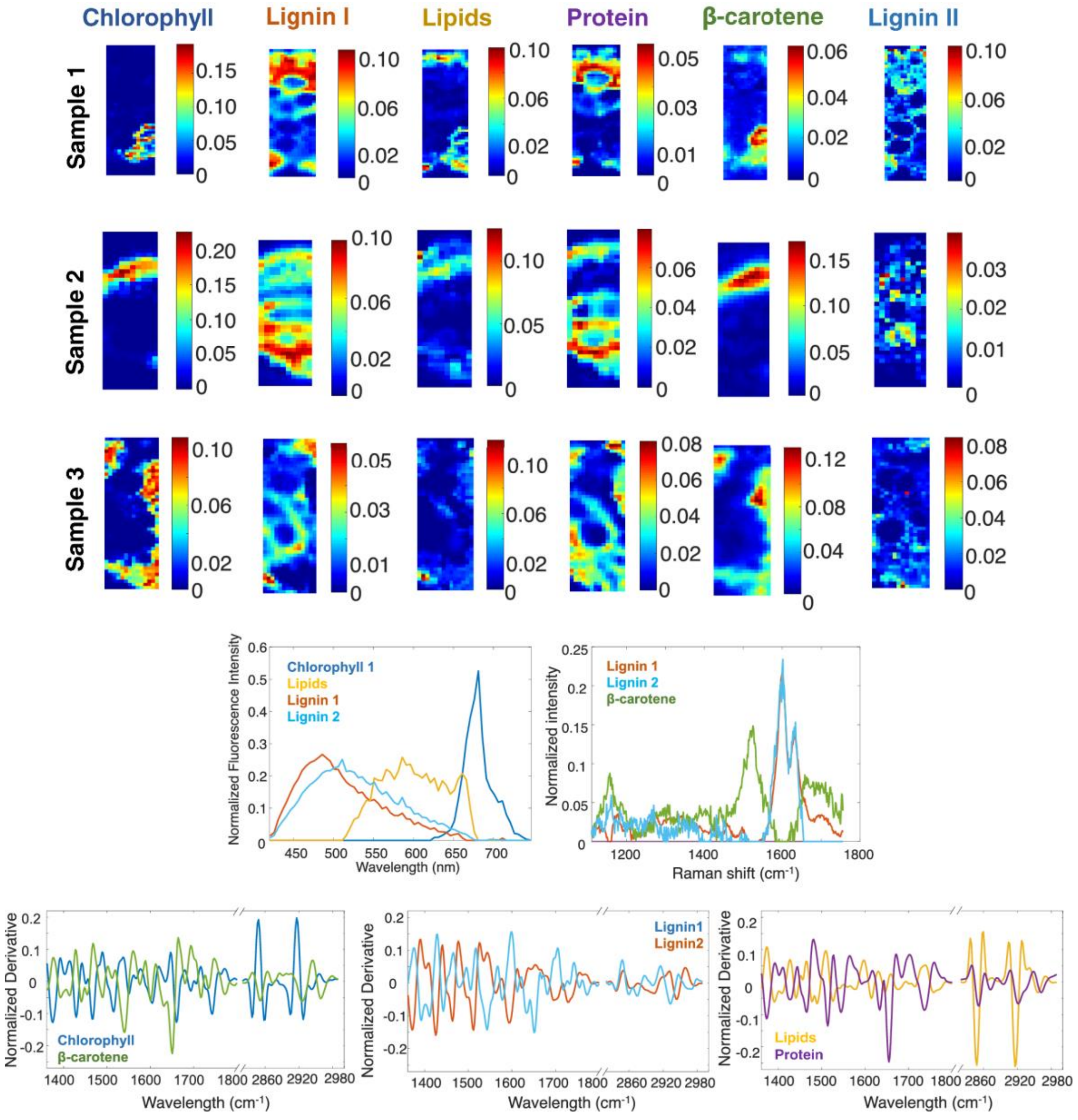


Figure 42. Top: Pure distribution maps of the six identified components for the three samples. Bottom: Pure spectral signatures of fluorescence, Raman, and SR-IR. The SR-IR spectra were segregated into plots featuring two components each to enhance visualization.

Figure 42 displays the characteristics of the six components modelled, which are characterized as follows.

Blue Component: identified as chlorophyll. It is observed in mesophyll cells across all samples in distribution maps. The pure fluorescence signature displays a typical emission spectrum of chlorophyll with a maximum at 682 nm.

Additionally, the infrared spectrum exhibits several bands that may be related to the chlorophyll structure, including alkanes, ester and alkenes, but not showing clearly signatures. No Raman signal was found.

Green Component: characterized as β-carotene, is located in mesophyll cells according to distribution maps. The Raman pure signature reveals typical Raman peaks at 1525 and 1157 cm^{-1}. As expected, β-carotene appears in the same leaf zone as chlorophyll. Notably, this component also shows a relevant protein band in SR-IR, suggesting possible co-location with protein-rich tissues. No fluorescence signal was found.

Orange and Cyan Components: identified as types of lignin, both components are observed at vascular tissues and sclerenchyma cells. The pure fluorescence spectrum of the orange component has a maximum at 487 nm, while the cyan lignin at 512 nm. The orange compounded includes a carbonyl band at 1730 cm^{-1}, absent in the cyan lignin. The pure Raman signatures of both components feature typical lignin peaks at 1632 cm^{-1} and 1601 cm^{-1}.

Yellow Component: Identified as a lipid-rich component, has a pure fluorescence spectral signature similar to the found in the epidermis and vesicles, according to the individual MCR-ALS analysis. The SR-IR pure signature confirms its lipid-rich identity, with two strong peaks at the lipid region (2916 and 2848 cm^{-1}). However, vesicles were challenging to observe in distribution maps due to the poor spatial resolution.

Benefits and drawbacks of image fusion

As already commented, the analysis of the fused multiset enables the characterization of a higher number of components than the individual analyses due to the joint use of complementary spectroscopic information. As a consequence, the characterization of components becomes easier thanks to the extended pure multitechnique spectroscopic signature. For instance, fluorescence images, with unspecific spectroscopic features did not allow identifying the chemical nature of the lipidic vesicles detected. However, when the fusion is performed, the connected SR-IR spectra enables unequivocal characterization of lipids through characteristic bands in the infrared region. Another example is the lignin identification in Raman images. In this case, the distinct spectral signatures in fluorescence assist in separating lignin contributions in Raman spectroscopy, which were not apparent when analyzing Raman images alone. Conversely, the characteristic features in Raman for lignin confirm the identity of these components, a task more challenging with only fluorescence information. Additionally, the combination of diverse spectroscopic information provides resolved profiles with a lower rotational ambiguity.

However, there are two clear drawbacks when classical fusion of hyperspectral images is carried out. First, for several imaging techniques, there is a clear loss of spatial resolution when the pixel size is balanced. For instance, the pixel size of the fluorescence images is increased by a factor of 144, i.e., the spectral information of (12×12) =144 fluorescence pixels is summed up to obtain a single pixel with the same size as a SR-IR pixel. As a result, the spectral information of this larger pixel is much more mixed and the capability to differentiate components located very closely decreases dramatically. This is evident in the case of the two chlorophylls, which are distinguishable in the individual fluorescence analysis, but not after fusion because the spectral selectivity and spatial detail of the initial measurement is lost. The same issue occurs with lipidic/silica bodies; although the spectra are observed in the pure component, the spatial location is lost, and only the epidermis is clearly observable. The second drawback is the discard of non-common scanned areas during the analyses. This step is necessary if a complete multiset is to be constructed. Without it, there would be pixels lacking corresponding spectra from other techniques and the application of classical MCR-ALS multiset analysis would not be possible. The exclusion of non-common scanned areas provides only a partial vision of the system under study and may also cause a loss of valuable information for the analysis, such as pixels presenting high purity, which could be beneficial in the unmixing procedure.

The possibility to work with all available information provided by the different imaging platforms in terms of sample area scanned and spatial detail is addressed in **Publication VIII**, where an adaptation of MCR-ALS is proposed to deal with incomplete multisets, with missing information.

3.4 Image fusion. Addressing differences in spectroscopic dimensionality.

Image fusion often encounters differences in spatial resolution among the different spectroscopic platforms, as it was shown in the previous work of **Publication VI,** where a solution for this problem was proposed. However, the image fusion challenge becomes even more complex when there are differences in the dimensionality of the spectroscopic data. Most imaging systems like Raman or infrared capture 3D images with a spectrum per pixel and the measurement follows a bilinear model. Instead, imaging platforms such as excitation-emission fluorescence imaging provide 4D images with a 2D spectral landscape per pixel and a trilinear model is required to define their signal behavior. The fusion of 3D and 4D images obeying different bilinear and trilinear models, respectively, is not trivial.

This subsection introduces a dedicated variant of the MCR-ALS algorithm meant to address differences in spectral dimensionality in image fusion. In this variant, the trilinearity constraint can be optionally applied *per block*, providing hybrid bilinear-trilinear models that help to preserve the natural linear behavior of the fused techniques. The proposed solution has been tested on the fusion of real 3D Raman and 4D fluorescence images of cross sections of rice leaf samples.

Publication VII. 3D and 4D image fusion: coping with differences in spectroscopic modes among hyperspectral images.
Authors: A. Gómez-Sánchez, M. Marro, M. Marsal, S. Zacchetti, R. R. de Oliveira, P. Loza-Álvarez, A. de Juan.
Citation reference: *Analytical Chemistry* (2020), 14:9591–9602
DOI: 10.1021/acs.analchem.0c00780

analytical chemistry

pubs.acs.org/ac Article

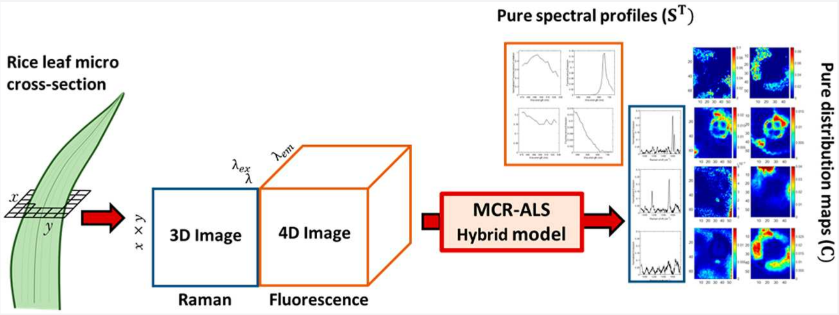
3D and 4D Image Fusion: Coping with Differences in Spectroscopic Modes among Hyperspectral Images

Adrián Gómez-Sánchez,* Mónica Marro, Maria Marsal, Pablo Loza-Alvarez, and Anna de Juan*

Cite This: *Anal. Chem.* 2020, 92, 9591−9602

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information



ABSTRACT: Image fusion is often oriented to solve differences in spatial scale and orientation among different spectroscopic platforms. However, an additional problem arises when the nature of the spectroscopic information differs in dimensionality as well. Indeed, most imaging systems, e.g., Raman, IR, MS, etc., allow acquisition of 3D images, with a linear spectrum per pixel, but new platforms have emerged, such as the recent excitation−emission fluorescence imaging platforms that provide 4D images, with a 2D spectral landscape per pixel. A proper 3D/4D image fusion needs to take into account the difference in the dimension of the spectral information and in the underlying models of both measurements (bilinear for 3D images and trilinear for 4D images). This work solves this image fusion problem through a new dedicated variant of the multivariate curve resolution-alternating least squares (MCR-ALS) algorithm for multiset analysis based on the incorporation of a hybrid bilinear/trilinear model that can handle the image fused structure preserving the natural behavior of the 3D and 4D imaging techniques coupled. The example is illustrated on the fusion of real 3D Raman and 4D fluorescence images recorded on cross sections of rice leaf samples.

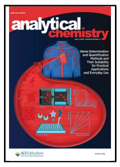
Hyperspectral images (HSIs) provide spatial and chemical information on samples and have become essential to solve many biological, industrial, and environmental problems. Nowadays, imaging platforms can differ extremely in spatial resolutions and can incorporate many spectroscopic and spectrometric measurements.[1,2] Image fusion then becomes an excellent approach to achieve a comprehensive description of the complexity of many chemical and biological systems.

Most image fusion works are oriented to solve spatial differences among imaging platforms, i.e., differences in spatial resolution and/or spatial orientation. Along this line, there are interesting works that solve the problem of spatial coregistration,[3] whereas other approaches are more oriented to handle spatial-resolution differences among imaging platforms through regression models between high- and low-spatial-resolution images or using multiblock or adapted multiset methodologies.[4,5]

The image fusion proposed in this work tackles a different aspect, the combination of image platforms having different spectroscopic modes and underlying models. Thus, 3D images are measurements defined as a data cube, where two dimensions *x*- and *y*- are the pixel coordinates, and the third dimension is spectral. This definition includes all platforms providing a 1D (vector) spectrum per pixel, such as Raman, IR, or MS imaging. Instead, a 4D image is defined as a hypercube of four dimensions, where two dimensions *x*- and *y*- are the pixel coordinates, and every pixel is associated with a 2D spectroscopic measurement. This definition adapts to excitation−emission (EEM) fluorescence platforms, where every pixel is associated with a 2D EEM landscape. When trying to fuse 3D and 4D images, the dimensionality of the two data structures is an obvious problem to handle but not the only one, since spectra in 3D and 4D images obey different underlying models.

Thus, 3D images adequately preprocessed are in general well approximated by the Lambert−Beer bilinear model (see Figure 1a). In this case, once the pixel spectra in a 3D cube are unfolded into a data table **D**, each pixel spectrum can be

Received: February 21, 2020
Accepted: June 10, 2020
Published: June 10, 2020



https://dx.doi.org/10.1021/acs.analchem.0c00780

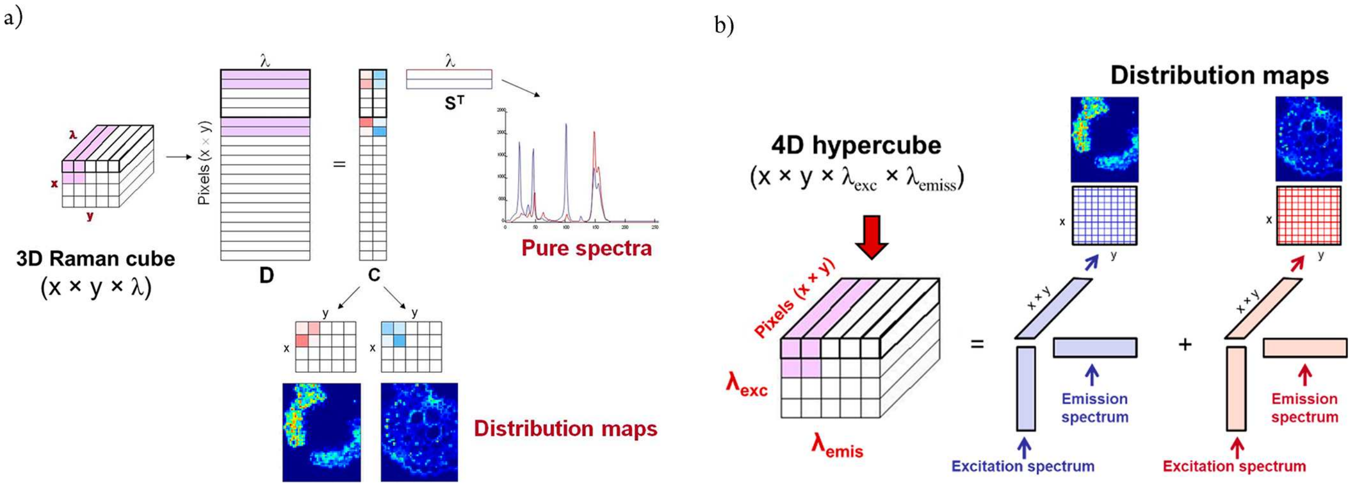


Figure 1. Underlying models of spectroscopic measurements. (a) Bilinear model (3D images) and (b) trilinear model (4D images).

defined as the concentration-weighted sum of the pure spectral signatures of the constituents present in the image. Thus, the full HSI can be expressed as a bilinear model, and every image constituent can be expressed by a dyad of profiles (i.e., concentration profile and spectrum). Instead, 4D images obey a trilinear model (see Figure 1b). In this case, once the pixel spectra in the original 4D hypercube are unfolded into a data cube $\underline{\mathbf{D}}$, where every pixel has an associated EEM slice, the full HSI can be expressed by a trilinear model and every image constituent by a tryad of profiles (i.e., concentration profile, excitation spectrum, and emission spectrum).

Although HSIs have information on the spatial composition of samples, chemical compounds often overlap on the same area, and unmixing or multivariate resolution methods are needed to extract the identity and spatial distribution of each particular image constituent, i.e., to recover the underlying model in Figures 1a,b from the raw image measurement. Most unmixing algorithms are associated with the interpretation of 3D images. There are different linear and nonlinear unmixing algorithms linked to the remote sensing area,[6] but within the chemometric field, multivariate resolution methods, such as multivariate curve resolution-alternating least squares (MCR-ALS), have been widely applied to hyperspectral image analysis.[7,8] Indeed, MCR-ALS allows the incorporation of constraints linked to spatial characteristics of the image and is well adapted to work with image multiset structures,[8−10] formed by several 3D images coming from different samples and/or from different imaging platforms. Unmixing 4D images, where every pixel has a 2D spectroscopic measurement associated, can be done with trilinear decomposition methods, such as parallel factor analysis (PARAFAC),[11] used since long to handle EEM data associated with sets of samples. In this respect, it is worth to mention that MCR-ALS, which allows incorporating trilinearity as a constraint, would be equally applicable. At this point, we can say that unmixing of 3D or 4D images separately is a problem already solved.[8,11]

As mentioned above, fusing 3D and 4D images involves the solution of two problems, the different dimensionalities of the images and the difference in underlying spectroscopic models. To do so, none of the algorithms previously described can be straightforwardly applied. Going to the simplest scenario, i.e., that the pixel mode among 3D and 4D platforms was common, the resulting data structure would need to address the combination of pixels associated with 1D spectra and 2D spectra.

Out of the field of application of image analysis, there are very few approaches that can work with this kind of problem. Acar et al. proposed a combined tensor and matrix factorization (CTMF) algorithm that allows a separate factorization of data structures involving data matrices (as a 3D image would give) and data cubes (as a 4D image would be).[12] In this algorithm, the factorizations of data matrices and cubes are separate and obey their respective bilinear and trilinear models, but there is a single error function related to the reproduction of all data sets used and a constraint that forces a common factor matrix (the concentration profiles) for both factorizations. Such an approach has been used in other application domains but requires a first diagnostic on which components are common or specific to the different data sets fused. Besides, the constraints used in the factor retrieval are scarce compared with other unmixing algorithms, more devoted to hyperspectral image analysis.

The approach proposed in this work supposes a step beyond the current possibilities offered by image unmixing algorithms and provides a new framework to solve the simultaneous analysis of 2D and 3D arrays based on the use of a single factorization on a fused 2D/3D data structure that can preserve the natural underlying model of the parent arrays. Such an approach is based on a modification of the MCR-ALS algorithm and is presented in the context of image fusion but has potential application for any 2D/3D array fusion where one of the modes of the two arrays is common.

Going back to the image application, a clear option to cope with the different data dimensionalities of 3D and 4D images would imply unfolding even more of the 4D image by vectorizing the 2D spectrum of a pixel, i.e., in an EEM context, concatenating in the same row all emission spectra related to the different excitation wavelengths. In this way, the originally 4D array would turn into a data table, where every row would show the vectorized EEM spectrum of a pixel, which could be connected with the table of a natural 3D image to form a multiset. This multiset could be analyzed by the current implementation of the MCR-ALS algorithm, which provides a bilinear model (as shown in Figure 1a). This methodology would circumvent the problem of defining common and specific components for the data sets combined, since a single factorization, valid for the full multiset structure, is used.

However, a bilinear model would be imposed to a measurement (the 4D image), which is known to obey a trilinear model.

Therefore, the novel solution proposed in this work to fuse 3D and 4D images involves the development of a new variant of MCR-ALS that allows handling the multiset formed by 3D and 4D images with a hybrid bilinear/trilinear model that allows preserving the natural underlying model of the images fused. To do so, a *per block* partial trilinearity constraint is designed that enables applying the trilinear model only to the spectral blocks linked to 4D images, whereas the rest of the fused structure, linked to 3D images, keeps obeying the bilinear model.

To show the development and application of this fusion strategy, biological samples consisting of cross sections of rice (*Oriza sativa* Japonica Nipponbare) leaves will be analyzed by Raman imaging (3D image platform) and EEM fluorescence imaging (4D image). A simulated example is also included to clarify and validate the methodology used. The satisfactory performance and the advantages of the use of MCR-ALS with the hybrid bilinear/trilinear model as opposed to fusion based on the use of the classical pure bilinear MCR model will be described.

■ EXPERIMENTAL SECTION

Plant Growth and Sample Preparation. Plant growth of *Oryza sativa* Japonica Nipponbare seeds was performed using a procedure previously described in ref 13. Briefly, seeds were obtained from the Center for Research in Agricultural Genomics (CRAG) at the Autonomous University of Barcelona and were germinated for 2 days at 30 °C in a wet environment. After germination, seeds were transferred to flowerpots of 3.5 cm diameter and watered three times per week with 150 mL of Milli-Q water for 22 days under controlled conditions of light, humidity, and temperature.

After harvest, small pieces of plant leaves were collected and embedded in agarose (5% w/w). Straightaway, one microsection of 50 μm thickness was prepared using a vibratome in a quartz slide with a drop of water, covered with a quartz coverslip (20 μm) and sealed with nail polish, to avoid water evaporation during the experiment. The final samples analyzed were two cross sections of the main and the secondary vessels of rice leaf samples.

3D Raman Images. Two Raman images of midrib and the secondary vein of rice leaves were acquired using an inVia confocal Raman microscope (Renishaw, Gloucestershire, United Kingdom) with a 20× DRY objective. The midrib is the central vein, which crosses the leaf longitudinally, while the secondary veins are smaller and parallel to the midrib. The samples were excited with a 532 nm green laser. The scattered Raman signal was collected with a spectrometer (1200 g mm^{-1} grating, spectral resolution about 1.79 cm^{-1} in a range of λ = 270 to 2015 cm^{-1}) and detected by the CCD camera (Andor DU401 BV, Belfast, North Ireland). For both images, the laser power was set at 10% of the 36 mW total power laser with an integration time of 0.50 s. The pixel motion step in the x- and y- directions was 2 μm, and the pixel size Raman images were recorded in point scanning mode. Images were recorded with a pixel size of 2 × 2 μm^2. The Raman image of the midrib has a field of view of 158 × 130 μm^2, and the related data set will be designated $\mathbf{D}_{\mathbf{R1}}$. The Raman image of the secondary vein has a field of view of 98 × 120 μm^2, and the data set will be designated $\mathbf{D}_{\mathbf{R2}}$.

4D Fluorescence Images. EEM images from the rice leaf cross sections were acquired using a Leica TCS SP8 STED 3× microscope (Leica, Mannheim, Germany) with an HC PL APO CS2 10×/0.40 DRY objective. The sample was excited with a supercontinuum white light laser (WLL) in a range of λ_{exc} = 470 to 526 nm with a 4 nm sampling interval. The fluorescence spectra were collected using a detector HYD SMD in a range of λ_{em} = 532.5 to 727.5 nm with 5.74 nm sampling interval and a bandwidth of 5 nm, providing a hyperspectral image with four dimensions: x and y as spatial directions and λ_{exc} and λ_{em} as spectral directions (4D). Fluorescence images were recorded by spectral scan mode.

Two images were collected; one of the midrib, giving a data set designated $\mathbf{D}_{\text{F1}}$, and another of the secondary vein, designated $\mathbf{D}_{\text{F2}}$. Each hyperspectral image has a field of view of 182.63 × 182.63 μm^2 and 188 × 188 nm^2 of pixel size. To be on the safe side and to avoid the effect of scattered light in the analysis of fluorescence images, the images analyzed cover the excitation range from 470 to 514 nm and the emission range from 555 to 727.5 nm.

■ DATA ANALYSIS

Data Sets. The methodology proposed in this work will be tested on a simulated data set and on the real images previously described in the Experimental Section.

The simulated example consists of the Raman and EEM fluorescence images of a sample. The different components are detected (present) or not according to the spectroscopic techniques used as described in Table 1.

Table 1. Components in Simulated Raman and EEM Images

component	Raman	EEM
1	present	present
2	present	present
3	absent	present

The shape of the distribution maps used for the simulation comes from the analysis of a similar image done by the authors on another rice leaf sample. The shapes of the two pure Raman spectra and the three excitation and emission spectra present a considerable overlap (all these profiles can be seen in Figure S1 of the Supporting Information). 3D Raman and 4D fluorescence images were reconstructed using a bilinear and a trilinear model, respectively, based on the profiles described, as shown in Figure 1. Once the images are obtained, some noise is added, mimicking the usual noise level found in these measurements. For more detail about the simulated images, see the Supporting Information.

Image Preprocessing. Raman spectra of real images showed a high-fluorescence baseline contribution due to the natural fluorophores in leaf tissue. The fluorescence baseline contribution was corrected by asymmetric least squares (AsLs).[14] Cosmic peaks, generated by cosmic rays hitting the detector, produced spurious needle-type features, and were corrected by interpolation of Raman intensities of nearest channels. Finally, the range 915 to 1755 cm^{-1} was chosen for the analysis, since it was the region containing useful Raman information.

The original pixels in the fluorescence images were binned to achieve the same pixel size as the Raman image. The binning allowed the signal-to-noise ratio of the original fluorescence spectra to be improved and facilitated the fusion

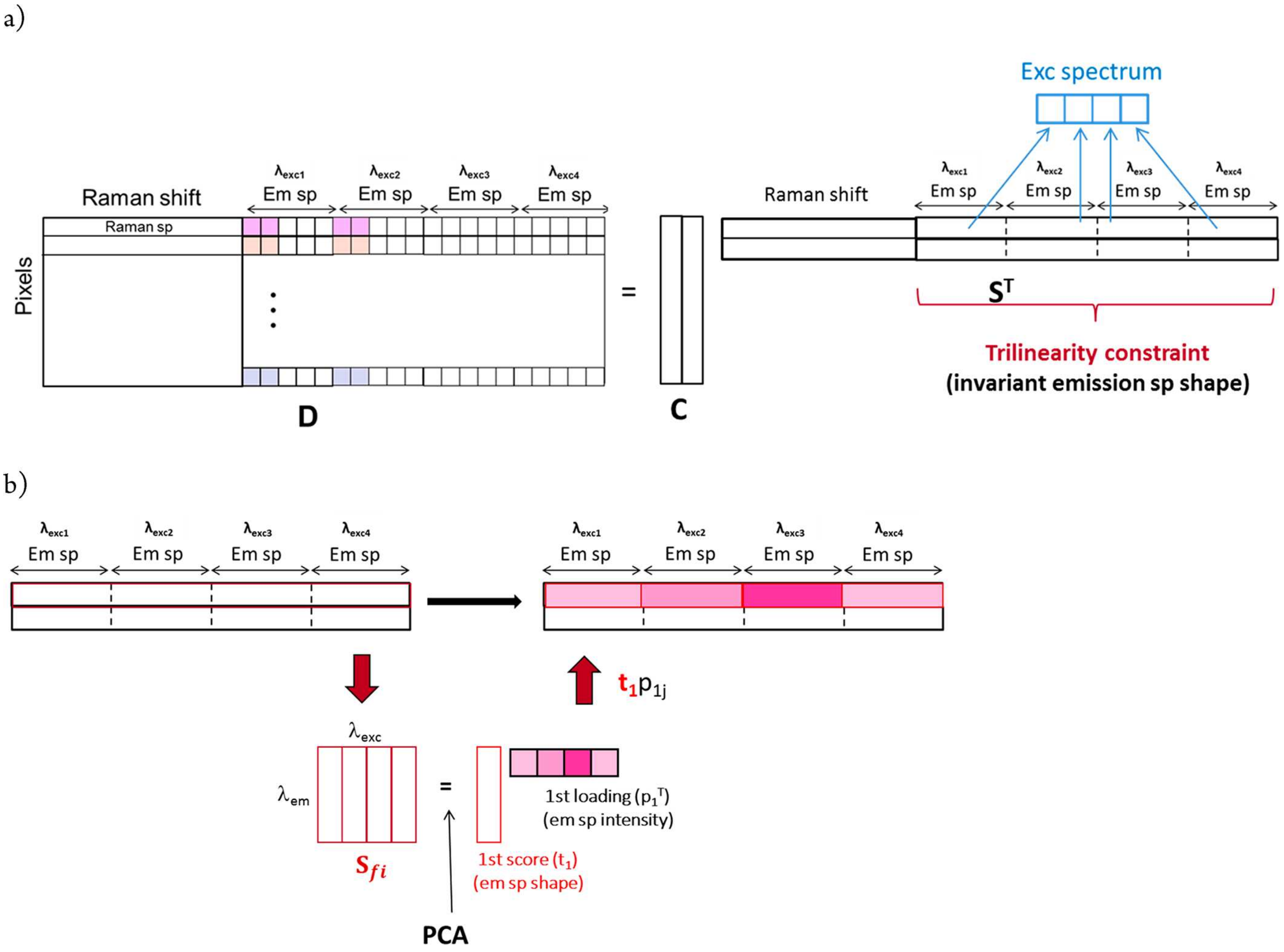


Figure 2. (a) Multiset formed by a 3D Raman image and a 4D fluorescence image and related MCR model. (b) Implementation of partial trilinearity in the emission spectra blocks.

with the Raman image. An example of raw and preprocessed Raman and fluorescence spectra can be found in Figure S2 in the Supporting Information.

Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) for Hyperspectral Image Analysis. MCR-ALS is a multivariate iterative resolution method meant to solve the mixture analysis problem. MCR-ALS is used in several fields and is especially suitable for hyperspectral image analysis.[2,8] For the resolution of 3D images, the original image cube is unfolded into the data table shown in Figure 1a. Then, MCR-ALS decomposes the original **D** matrix into the bilinear model expressed by eq 1

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

where **D** is the matrix containing all pixel spectra, and **C** and $\mathbf{S}^{\mathbf{T}}$ are the matrices of meaningful concentration profiles and spectral signatures of the image constituents, respectively. **E** is the matrix of residual variation unexplained by the MCR model. As can be seen in Figure 1a, concentration maps are easily recovered, refolding each concentration profile into a 2D map showing the original image spatial structure.

MCR-ALS involves the iterative alternating optimization of **C** and $\mathbf{S}^{\mathbf{T}}$ matrices under constraints. Available constraints for hyperspectral image analysis are non-negativity in **C** and/or $\mathbf{S}^{\mathbf{T}}$,[8] adapted local rank constraints[15] and spatial constraints.[9,10] Constraints are optionally applied *per mode* (**C** and $\mathbf{S}^{\mathbf{T}}$) and *per component*, providing chemical meaning to the profiles retrieved and reducing the ambiguity of the MCR solutions.[7,8]

The convergence criterion can be defined by a maximum number of iterations or by a value related to the difference in fit improvement between consecutive iterations, lack of fit (LOF). The parameters used to estimate the quality of the MCR model fit are the LOF and the explained, as expressed in eq 2 and eq 3.

$$\mathrm{LOF}(\%) = 100^{*}\sqrt{\frac{\sum_{i,j} e_{i,j}^2}{\sum_{i,j} d_{i,j}^2}} \tag{2}$$

$$\mathrm{var}(\%) = 100^{*}\left(1 - \frac{\sum_{i,j} e_{i,j}^2}{\sum_{i,j} d_{i,j}^2}\right) \tag{3}$$

where $d_{i,j}$ is the ij^{th} element of **D**, and $e_{i,j}$ is the residual associated with the reproduction of $d_{i,j}$ by the MCR model.

Image Fusion of 3D and 4D Images by MCR-ALS. MCR-ALS can also be applied to multiset structures formed by several images.[8,16] Multisets follow the bilinear model and can be built with several images obtained with different platforms, as shown in eq 4, where $\mathbf{D}_{ij}$ refers to image of sample *i* and platform *j*.

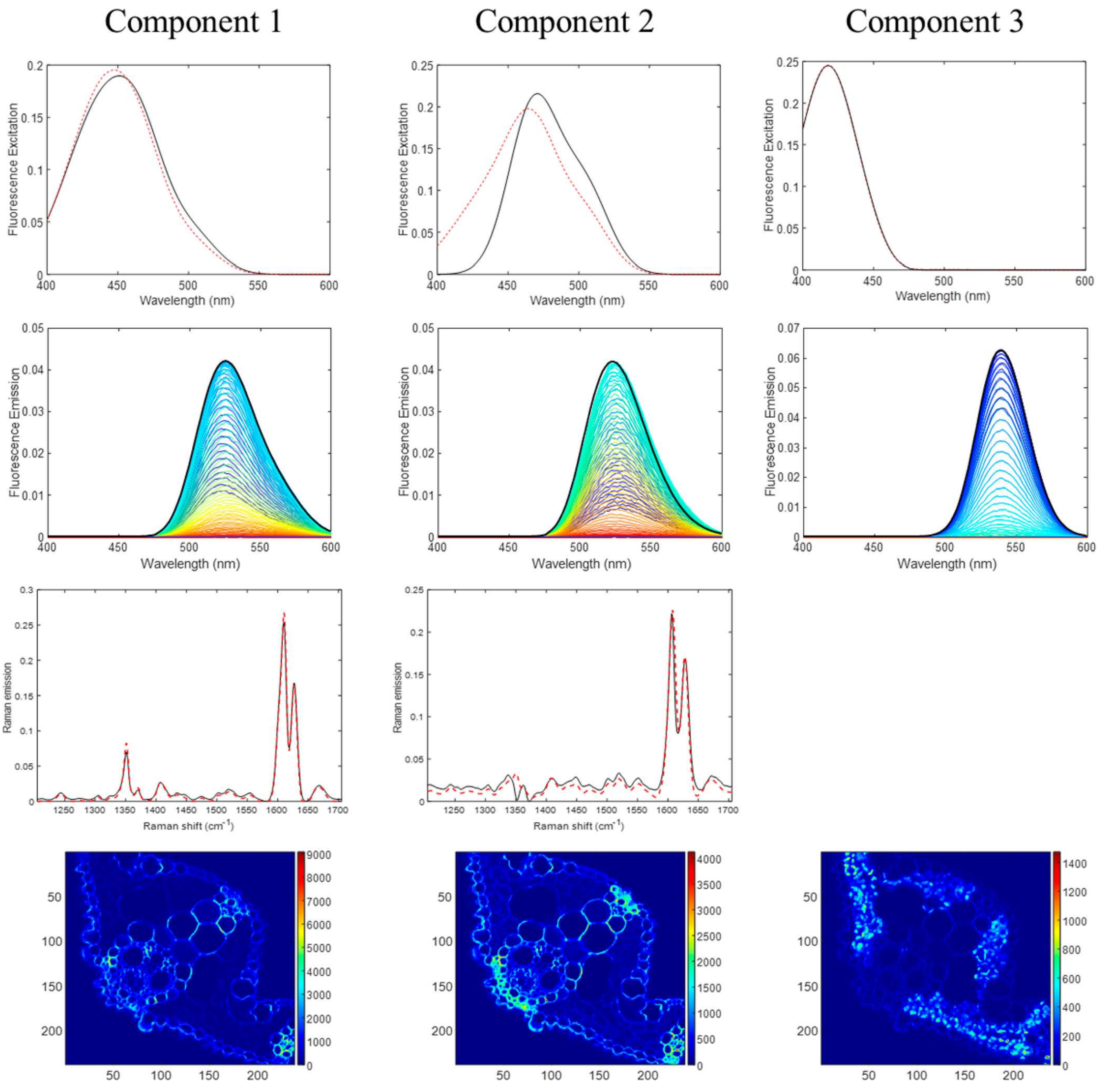


Figure 3. MCR results on the simulated data set fusing a 3D Raman image and a 4D EEM fluorescence image from the same sample. Top plots: excitation spectra (black line: simulated profiles; red dashed lines: MCR spectra). Second row plots: emission spectra (black line: simulated profile shape; colored spectra from blue to red, MCR emission spectra from lowest to highest excitation wavelength). Third row plots: Raman spectra (black line: simulated profiles; red dashed lines: MCR spectra). Bottom plots: MCR distribution maps.

$$\begin{pmatrix} \mathbf{D}_{11}\mathbf{D}_{12}\mathbf{D}_{13}\ldots\mathbf{D}_{1L} \\ \mathbf{D}_{21}\mathbf{D}_{22}\mathbf{D}_{23}\ldots\mathbf{D}_{2L} \\ \mathbf{D}_{31}\mathbf{D}_{32}\mathbf{D}_{33}\ldots\mathbf{D}_{3L} \\ \ldots \\ \mathbf{D}_{K1}\mathbf{D}_{K2}\mathbf{D}_{K3}\ldots\mathbf{D}_{KL} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \mathbf{C}_3 \\ \ldots \\ \mathbf{C}_K \end{pmatrix} (\mathbf{S}_1^{\mathbf{T}}\mathbf{S}_2^{\mathbf{T}}\mathbf{S}_3^{\mathbf{T}}\ldots\mathbf{S}_L^{\mathbf{T}}) + \mathbf{E}_{\mathbf{aug}} = \mathbf{C}_{\mathbf{aug}}\mathbf{S}_{\mathbf{aug}}^{\mathbf{T}} + \mathbf{E}_{\mathbf{aug}} \tag{4}$$

When connecting images monitored with the same technique, the spectral range needs to be common among images. When fusing different imaging techniques, the pixel mode has to be common among images, and both requirements need to be addressed in multisets responding to eq 4. Ensuring that the pixel mode is common among images from different platforms means that the pixel size needs to be the same (spatial binning can solve this aspect) and that the area scanned and the spatial orientation are also common. Matching the spatial orientation of images can be carried out using a procedure based on comparing binarized maps of the images to be fused proposed by Piqueras et al.[3]

3D and 4D images with the same pixel size and spatially matched can be organized into a multiset, where every pixel

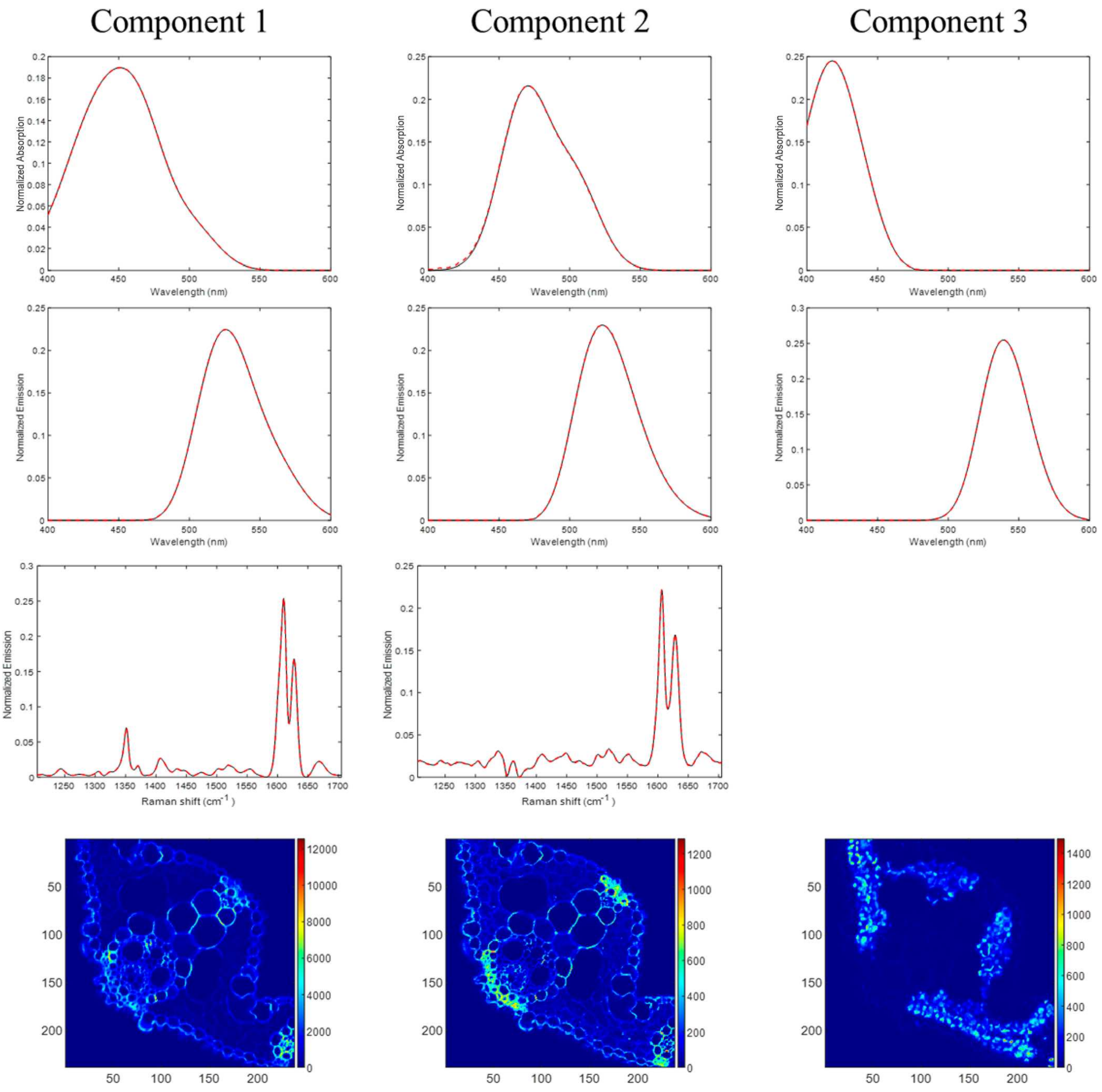


Figure 4. MCR results on the simulated data set fusing a 3D Raman image and a 4D EEM fluorescence image from the same sample. Top plots: excitation spectra (black line: simulated profiles; red dashed lines: MCR spectra). Second row plots: emission spectra (black line: simulated profiles; red dashed lines: MCR spectra) Third row plots: Raman spectra (black line: simulated profiles; red dashed lines: MCR spectra). Bottom plots: MCR distribution maps.

will be described by the 1D spectrum coming from the 3D image and the vectorized 2D EEM spectral information from the 4D image (see Figure 2a). For a single sample, the resulting MCR-ALS model would consist of a common matrix of concentration profiles and an augmented spectral matrix ($\mathbf{S}^{\mathbf{T}}_{\mathbf{aug}}$) that would contain a composed spectral signature per each image constituent, formed by the 1D pure spectrum and the vectorized 2D EEM spectrum. Excitation spectra of each component can be afterward retrieved by integrating the area under the emission spectra per each of the excitation wavelengths analyzed.

In order to preserve the natural underlying spectroscopic model of 3D and 4D images, a modification of the MCR-ALS algorithm is proposed that allows solving the multiset in Figure 2a using a hybrid bilinear/trilinear model. In order to deal adequately with 3D/4D image multisets, we propose here for the first time to extend the flexibility of application of the trilinearity constraint *per block*. To do so, in every iteration of the algorithm, only the profiles from the 2D EEM blocks in $\mathbf{S}^{\mathbf{T}}_{\mathbf{aug}}$ will be submitted to obey the trilinearity constraint, whereas the spectral blocks related to 3D images will not be subject to this condition (see Figure 2a). Note that this constraint is used only on the EEM blocks of $\mathbf{S}^{\mathbf{T}}_{\mathbf{aug}}$, and it is applied every time to a single row of these blocks, assigned to the emission spectra of a single pure fluorophore at the different excitation wavelengths. Doing it this way, it can be clearly assumed that the emission spectra of that single fluorophore should have the same shape in all excitation wavelengths and that they will only vary in signal intensity. Bearing in mind this assumption, the trilinearity constraint for a single fluorophore (i) would be applied as shown in Figure 2b. Thus, in each MCR-ALS iteration and for the i^{th} component, the trilinearity constraint is applied following the next steps.

(a) The i^{th} row of the emission spectra blocks at the different excitation wavelengths of the $\mathbf{S_{aug}}^{T}$ matrix, sized $(1, \lambda_{em} \times \lambda_{exc})$ is refolded into a matrix sized $\boldsymbol{S_{fi}}$ $(\lambda_{em}, \lambda_{exc})$, the columns of which contain all emission spectra at the different excitation wavelengths of the i^{th} component.

(b) $\boldsymbol{S_{fi}}$ is decomposed by principal component analysis (PCA) according to the equation $\boldsymbol{S_{fi}} = \boldsymbol{TP^{T}}$, where $\boldsymbol{T}$ is the score matrix, and $\boldsymbol{P^{T}}$ is the loading matrix. Since the emission spectra in $\boldsymbol{S_{fi}}$ belong to the same fluorophore and should have the same shape, the first score in $\boldsymbol{T}$ ($\mathbf{t_1}$) represents appropriately the common shape that all emission spectra should have. The elements in the first loading in $\boldsymbol{P^{T}}$ ($\mathbf{p_1}^{T}$) account for the different signal intensities that the emission spectra have, depending on the excitation wavelength.

(c) The constrained emission spectra resubmitted to the MCR optimization all have the same shape and are obtained by multiplying the first score of $\boldsymbol{S_{fi}}$ (describing the common shape for the emission spectra of the fluorophore) by the loading related to the suitable excitation wavelength (accounting for the signal intensity linked to that excitation wavelength), i.e., to reconstruct the emission spectrum related to the block of the j^{th} excitation wavelength, we would do $\mathbf{t_1}\mathrm{p_{1j}}^{T}$.

This simple and flexible way to implement trilinearity in MCR-ALS adapts to select the application of this constraint *per block* and also *per component* in a multiset structure.

RESULTS AND DISCUSSION

Simulated Example. The potential of the methodology presented is first tested on the fusion of the simulated 3D Raman and 4D fluorescence images belonging to the same sample. A multiset, as the one shown in Figure 2a, is formed and is designated as $[\mathbf{D_{RS}D_{FS}}]$, with $\mathbf{D_{RS}}$ being the simulated data matrix coming from the Raman image and $\mathbf{D_{FS}}$ being the simulated data matrix coming from the 4D image. Every row of the multiset contains the Raman spectrum and the concatenated emission spectra at the different excitation wavelengths associated with a particular pixel. The MCR model of this multiset is described as $[\mathbf{D_{RS}D_{FS}}] = \mathbf{C_S}[\mathbf{S_{RS}}^{T}\mathbf{S_{FS}}^{T}]$, where $\mathbf{C_S}$ contains the concentration profiles of the three components of the sample, which can be refolded into the related maps, and $[\mathbf{S_{RS}}^{T}\mathbf{S_{FS}}^{T}]$ is the augmented pure spectra matrix with $\mathbf{S_{RS}}^{T}$ being the block containing the pure Raman signatures and $\mathbf{S_{FS}}^{T}$ being the block of the pure concatenated emission spectra.

This multiset is analyzed by MCR-ALS in two different ways: (a) using a pure bilinear model and equality and non-negativity constraints in the pure spectra, matrix $[\mathbf{S_{RS}}^{T}\mathbf{S_{FS}}^{T}]$, and in the concentration profiles in $\mathbf{C_S}$ and (b) using a hybrid bilinear/trilinear model and equality and non-negativity in the spectral and concentration direction and partial trilinearity in the blocks related to emission spectra in $\mathbf{S_{FS}}^{T}$. In both MCR-ALS analyses, the same initial spectral estimates obtained a method based on SIMPLISMA were used.[17]

Figures 3 and 4 show the pure spectra and related maps resolved by MCR-ALS in the analyses performed using a pure bilinear model and a hybrid bilinear/trilinear model, respectively. A first comment about the results obtained is that both analyses give the same lack of fit, 14.49%, in agreement with the amount of noise added in the simulation (see Supporting Information).

Figure 3 shows that not all distribution maps are well recovered; especially, the map of component 3 shows patterns of the cell wall structure in the leaf that should be absent (see Figures S1 and 4). The emission spectra of components 1 and 2 differ slightly in shape for the different excitation wavelengths, shifting their maxima with respect to the real spectrum used in the simulation (in black). The shift may not be very apparent visually, because the three simulated emission spectra were very similar among them. However, the intensity of the emission spectra at the different excitation wavelengths is not correct, as it is clearly reflected by the incorrect recovery of the shape of the related excitation spectra (especially for component 2). The pure Raman spectra are recovered reasonably well; although this is not surprising if we consider that the original shapes of the spectra used in the simulation are very similar to each other.

Figure 4 instead, where the partial trilinearity constraint is applied, manages to perfectly recover the shapes of the simulated distribution maps, excitation and emission spectra, and Raman spectra, validating the proposed methodology and showing that the use of a hybrid bilinear/trilinear model in 3D/4D image fusion makes a substantial difference in the quality of the results compared with the mere use of a bilinear model.

In the Supporting Information, Table S1 shows the correlation coefficients between the simulated and recovered MCR profiles to complement this information. As can be appreciated, profile recovery in the hybrid bilinear/trilinear resolution always has a correlation coefficient $r > 0.998$ for all concentration and pure spectra profiles in the three components, whereas the values of the correlation coefficient for the pure bilinear model are clearly lower, with values that can get until $r < 0.85$. For a better differentiation, if the angle between simulated and recovered MCR profiles, defined as $\alpha(\text{deg}) = a(\cos(r))$ is used, the hybrid bilinear/trilinear model always present α values lower than 3.5° (0° would be the value for $r = 1$) for all profiles, whereas the pure bilinear model presents higher angles, sometimes reaching α values around 30°.

Real Images. Before performing the 3D/4D image fusion of the real 3D Raman and 4D fluorescence images of rice leaves cross sections, a preliminary analysis using MCR-ALS was performed on the images of each technique separately to have a first insight on the components present in the samples. Thus, column-wise augmented fluorescence and Raman multisets formed by the images of the two different samples studied were analyzed separately by MCR-ALS. The number of components used in each MCR analysis was first explored using singular value decomposition, SVD. If SVD was not giving conclusive results, a few models with variable numbers of components were tested, and other criteria, such as the chemical meaningfulness of the maps and spectral signatures resolved and a lack of fit in agreement with the noise level of the data were taken into consideration for the final decision on the size of the MCR model.show the pure spectra

A method based on SIMPLISMA was used to obtain the initial spectral estimates in both multisets.[17] The trilinearity constraint was used in the fluorescence multiset to respect the natural behavior of the EEM data. In addition, non-negativity was applied in both multisets, since neither fluorescence nor Raman spectra have negative values

Results are shown in Table 2. For both multisets, the lack of fit and variance explained are satisfactory considering the noise level of the initial spectra analyzed. Distribution maps and pure spectral signatures derived from the fluorescence and the Raman MCR-ALS multiset analysis are attached in Figures S3 and S4 of the Supporting Information, respectively.

Table 2. Summary of Multiset Structures and MCR-ALS Main Results

data set	NC[a]	initial estimates	constraints	% lack of fit	% expl. variance
$[D_{F1};D_{F2}]$	2	SIMPLISMA-based	non-negativity, trilinearity	8.10	99.34
$[D_{R1};D_{R2}]$	3	SIMPLISMA-based	non-negativity	11.13	98.76
$[D_{R1}D_{F1};D_{R2}D_{F2}]$	4	SIMPLISMA-based	non-negativity	9.58	99.08
	4	supervised[b]	non-negativity, equality	9.70	99.06
	4	SIMPLISMA-based	non-negativity, trilinearity	9.71	99.06
	4	supervised[b]	non-negativity, equality, trilinearity	10.09	98.98

[a] NC: number of components. [b] Supervised describes combining pure spectra from multiset analysis on individual imaging techniques.

Two components were detected in fluorescence images, characterized as chlorophyll and lignin. In Raman images, three components were detected: lignin, β-carotene, and a non-biological contribution due to an instrumental artifact. A brief description of the spectral features and location of these components in the rice leaf samples is provided to facilitate interpretation of the output of the fused 3D/4D image analyses.

Chlorophyll has a strong fluorescence emission around 675 nm[18] but no associated Raman emission. It is located on chloroplast in mesophyll cells, where the photosynthesis is carried out. As it can be observed in Figure S3, the chlorophyll fluorescence emission spectrum agrees with the literature, and its distribution map coincides with the mesophyll cell location.

Another component was characterized as lignin. Lignin is usually located in the vascular tissue of leaves.[19] The resolved fluorescence emission spectrum in Figure S3 agrees with the literature showing a band with a maximum around 500−550 nm.[20] The Raman lignin spectrum in Figure S4 presents two characteristic strong Raman bands at 1600 and 1635 cm^{-1}, In our resolution, lignin is located in the outer part of the leaf and, as reported in the literature, in the vascular system of the leaf.[20]

Another component was characterized as β-carotene. β-carotene can be found at mesophyll cells, where chlorophyll is, forming protein complexes in thylakoid membranes. The resolved Raman spectrum shown in Figure S4 shows typical Raman bands at 1004, 1155, 1186, and 1523 cm^{-1}, which agree with the literature.[21] The fluorescence signal has not been detected. The distribution map of β-carotene seems to be affected by photobleaching, but it is still possible to observe that it is located in mesophyll cells.

Finally, a last component present in Raman is related to an instrumental artifact, and it has a noisy sinusoidal behavior, see Figure S3. This kind of component has been previously found in Raman images performed on other samples with the same instrument.[22] As can be seen in Figure S5, the pattern in the resolved component in Figure S4 matches the pattern found in the raw Raman spectra before any baseline correction is performed. Since this artifact is more present in spectra with a high-fluorescence contribution, the related distribution maps show higher intensity in the regions where carotene is present.

3D/4D Image Fusion. *3D/4D Image Fusion (Effect of the Application of the Partial Trilinear Constraint).* To perform the 3D/4D image fusion, a multiset formed by the four images coming from the Raman and fluorescence images from the midrib and the secondary vein of the cross sections of rice leaves was built. To concatenate the Raman and fluorescence images, as in Figure 2a, fluorescence images were down-sampled by binning (188 × 188 nm^2 to 2 × 2 μm^2 pixel size) to obtain an identical pixel size to Raman images. Once the same pixel size was achieved, fluorescence and Raman images were cropped until having the same scanned area approximately. Several ways to achieve good initial maps to align images were proposed in Piqueras et al.[3] In this case, binarized distribution maps of lignin, based on maps provided by MCR-ALS analysis on separate fluorescence and Raman multisets, were used for the alignment of the images of each sample analyzed. The choice was due to the clear morphological structure of the lignin maps, more easily comparable among techniques than the morphology of binarized global intensity maps, often used for image alignment purposes.

The 3D(Raman)/4D(fluorescence) image multiset obtained was analyzed by MCR-ALS and provided an augmented matrix $\mathbf{C}_{\mathbf{aug}}$, formed by the concentration profiles that will give the distribution maps of the samples of the main and secondary vein of rice leaves, and an augmented matrix $\mathbf{S}_{\mathbf{aug}}{}^{\mathbf{T}}$, formed by the extended Raman and emission fluorescence signatures of each component. This multiset, coded $[\mathbf{D}_{\mathbf{F1}},\mathbf{D}_{\mathbf{R1}};\mathbf{D}_{\mathbf{F2}},\mathbf{D}_{\mathbf{R2}}]$ was analyzed by applying different initial estimates and constraints according to Table 1. As initial estimates, two options were tested: an initial estimate obtained by the use of a SIMPLISMA-based method on the multiset and a second kind of initial estimate based on the combination of pure spectra resolved in the MCR-ALS analyses performed on each imaging technique separately. The constraints used in different combinations were non-negativity in the concentration and spectral direction, equality constraints in the spectral direction forcing null spectroscopic signal in components not detectable by one of the spectroscopic techniques used, i.e., null signal for chlorophyll in Raman spectra and null signal for β-carotene in fluorescence signatures, and trilinearity in the blocks related to fluorescence emission spectra.Table 1 summarizes relevant MCR results for all possible combinations of initial estimates and constraints tested. In all cases, the number of resolved components is the same, and the lack of fit is satisfactory and very similar among analyses, which indicates that all constraints tested are really obeyed by the multiset of interest.

3D/4D Image Fusion (Bilinear Model). Two multiset analyses based on the use of a pure bilinear model were tried, as shown in Table 1. Figure 5 shows the distribution

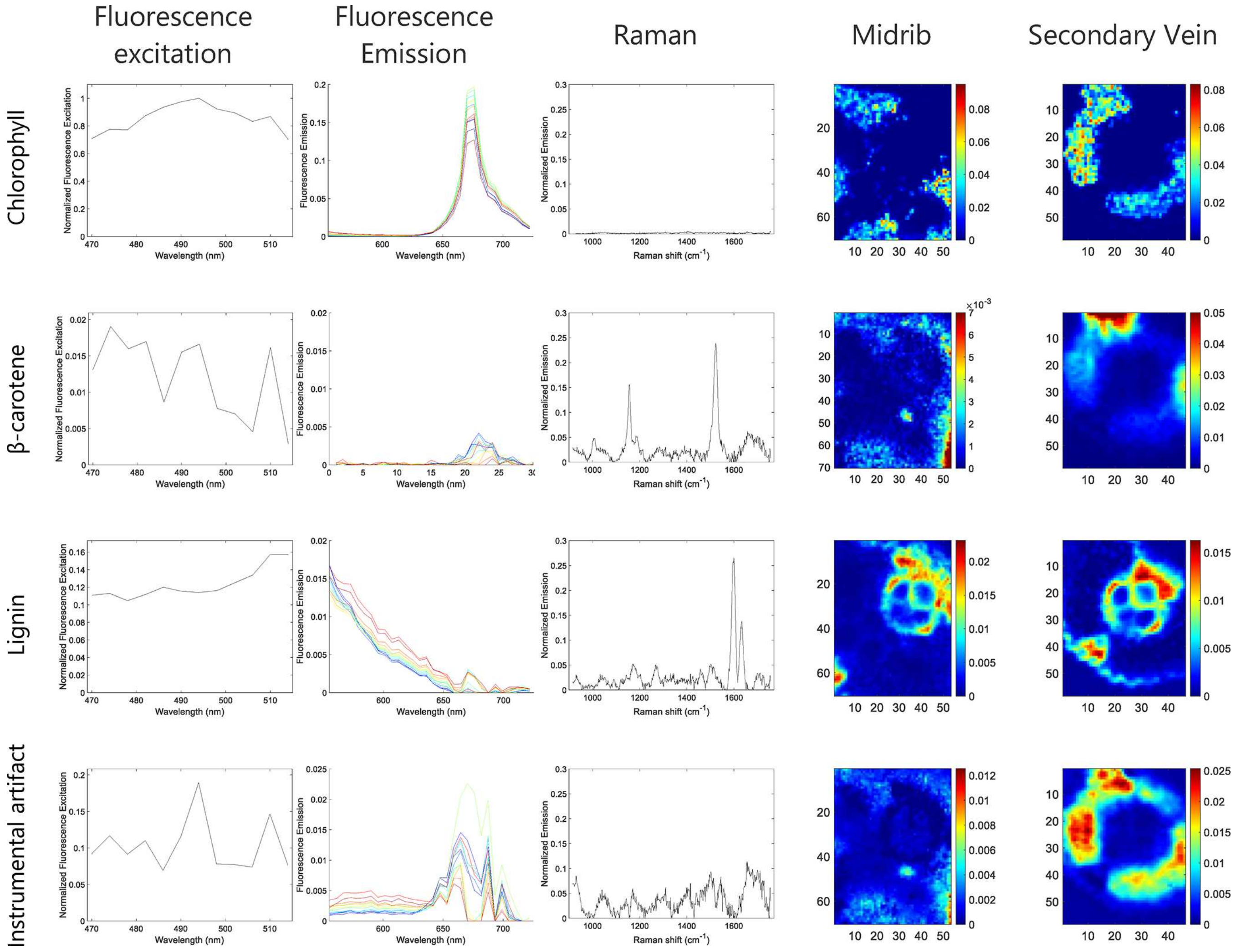


Figure 5. MCR results obtained on the Raman/fluorescence image multiset using a bilinear model. Plots in columns indicated from left to right: resolved excitation spectra, overlapped emission spectra for all excitation wavelengths, Raman spectra, maps from midrib and from secondary vein of rice leaves.

maps and related Raman, excitation, and emission spectra from the MCR analysis with less prior assumptions (initial estimates are done with a SIMPLISMA-based method) and the mildest application of constraints (only non-negativity in the concentration and spectral direction).

As positive outcomes of this resolution, we can mention that the four components expected according to the analysis of the multisets on each separate imaging technique were obtained. Thus, distribution maps related to chlorophyll, carotene, lignin, and the instrumental artifact seen by Raman show the morphology already observed in the separate analysis of these two imaging techniques (see Figures S3 and S4). The spectral signatures for the Raman spectra of carotene, lignin, and the instrumental artifact and the spectral excitation spectra of the chlorophyll and lignin show correct shapes. Looking at the relationship between connected Raman/fluorescence signatures for the same component, consistent observations with previous analyses can be found, such as the fact that the fluorescence chlorophyll signature connects with noisy Raman patterns, in agreement with the absence of signal for these components in the Raman technique. Likewise, the carotene signature and the artifact previously detected in Raman connect with low and unpatterned fluorescence signatures, as expected for nonfluorescent components or noise-related contributions. The lignin component, common to Raman and fluorescence images, shows the expected features in the Raman and excitation spectrum.

However, the recovered emission spectra per each component show clearly unacceptable features. The only component well recovered is the chlorophyll, showing the expected dominant band around 675 nm and, most importantly, a consistent shape of emission spectra in all excitation wavelengths analyzed, seen in the plot with the resolved overlapped emission spectra at different excitation wavelengths in Figure 5.

The emission spectra recovered for the lignin show a clear variation in shape as a function of the excitation wavelength, notably, a shift of the main emission band toward higher wavelengths in the lignin. In fluorescence spectroscopy, the emission spectrum shape of a pure component should be invariant in all excitation wavelength ranges scanned. The reason why this requirement is not obeyed by some components is clearly associated with the ambiguity of the MCR results obtained using a bilinear decomposition with mild constraints, such as non-negativity.

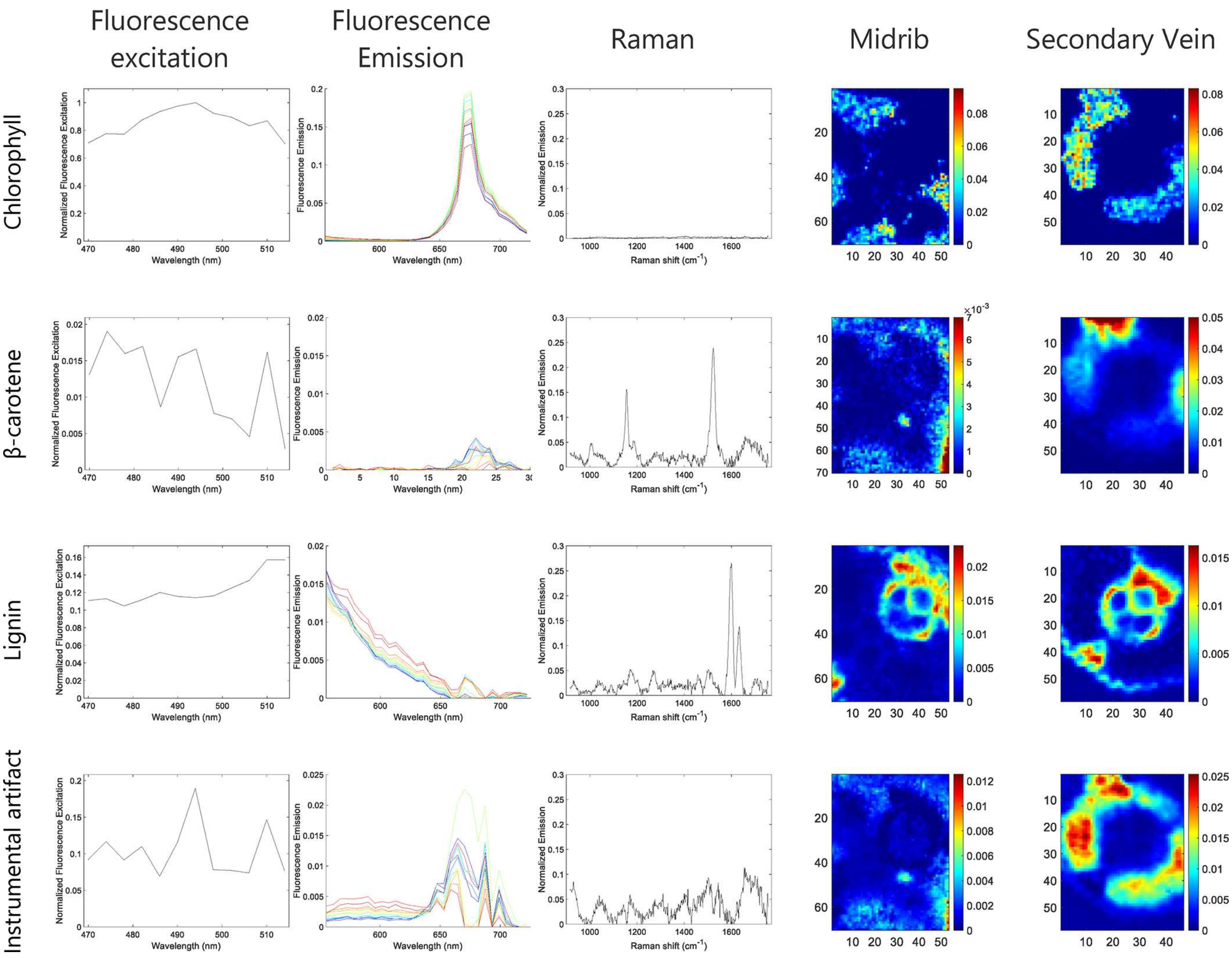


Figure 6. MCR results obtained on the Raman/fluorescence image multiset using a hybrid bilinear/trilinear model. Plots in columns indicated from left to right: resolved excitation spectra, emission spectra, Raman spectra, map from midrib and from secondary vein of rice leaves.

Additional improvements on this simplest analysis, based on the initialization with better spectral estimates and the inclusion of spectral equality constraints, as expressed on Table 1 (see Figure S6) do not solve the problem of the intracomponent shape variation in the resolved emission spectra These analyses, performed using more information and constraints, confirm the fact that MCR-ALS based on a bilinear decomposition is insufficient to solve adequately the analysis of multisets based on 3D/4D image fusion.

3D/4D Image Fusion (Hybrid Bilinear/Trilinear Model). MCR-ALS was applied also to the same multiset using a hybrid bilinear/trilinear model. It is important to note the extreme flexibility of the trilinear constraint in this context, which takes advantage of the new *per block* implementation and the existing *per component* optional application. Thus, trilinearity is applied to all blocks of fluorescence emission spectra, leaving the Raman block outside of the constraint. And, within the fluorescence blocks, trilinearity is applied to all components, except to the fluorescence part related to the Raman artifact, because a noise contribution is not expected to follow a trilinear behavior.

Using this hybrid bilinear/trilinear model, two MCR-ALS analyses were carried out varying initial estimates and constraints. As shown in Table 1, the lack of fit and variance explained in all MCR-ALS runs using a hybrid bilinear/trilinear model are similar among them and to analogous MCR-ALS analyses assuming a complete bilinear behavior. This similarity in fit with previous bilinear decompositions confirms, from a mathematical point of view, that the fluorescence part of the data set behaves according to a trilinear model.[23]

Figure 6 shows the MCR results obtained when working in the same conditions as in the analysis of Figure 5, i.e., SIMPLISMA-based initial estimates and non-negativity in the concentration and spectral direction, but incorporating the partial trilinearity constraint in the emission spectra blocks of $\mathbf{S}_{\mathbf{aug}}^{\mathbf{T}}$. The expected four components are retrieved, with similar distribution maps and spectral shapes in the Raman and excitation spectra. Now, the main difference is that, since partial trilinearity is applied, the correct single shape of the emission spectrum per each component is retrieved, and the unmixing task produces a better definition of all sample components.

Figure S7 in the Supporting Information, obtained using better estimates and incorporating additional equality constraints, does not imply substantial improvement in the results, since the action of the trilinearity constraint drives the system to the correct solution without the need of further information.

As a summary, the quality of the results obtained from this hybrid bilinear/trilinear variant of MCR-ALS is significantly better than the results recovered from multiset MCR-ALS image analysis based on pure bilinear decomposition. The modification of the MCR-ALS algorithm with the partial trilinearity constraint is a very valid alternative to solve both the differences in spectral dimensionality and in underlying measurement models associated with the problem of 3D/4D image fusion.

CONCLUSIONS

3D/4D image fusion is a recent challenge linked to the emergence of new powerful 4D imaging platforms providing 2D spectroscopic landscapes per pixel. The difficulty in 3D/4D image fusion lies both in the different dimensionalities of the spectroscopic information among platforms and the different specific underlying models required to explain the spectroscopic measurements, bilinear for 3D images and trilinear for 4D images.

Differences in spectroscopic dimensionality are easily solved by forming a multiset that connects the 1D pixel spectra from 3D images with the spatially congruent vectorized version of the related 2D spectral landscapes coming from 4D images. In addition, the difference in spectroscopic measurement models is addressed with the new block-wise optional implementation of the trilinearity constraint within the MCR model.

The hybrid bilinear/trilinear model provided by this new MCR-ALS variant has many benefits, namely (a) the complementary spectral information of the different image platforms is taken simultaneously into consideration in a multiset analysis to provide a better description of image constituents, (b) every hyperspectral image in the multiset is modeled respecting the suitable spectroscopic underlying model, bilinear for the 3D image and trilinear for the 4D image, (c) the use of the trilinear constraint in the resolution dramatically decreases the uncertainty linked to the resolved concentration and spectral profiles, and (d) since the 3D and 4D images are in the same multiset structure, there is no need to know which components may be common or specific to each of the images, as other methods based on separate factorizations of 3D and 4D images require.

This simple, yet powerful solution to the combination and analysis of 3D/4D image measurements is a step forward in the necessary image fusion tools demanded by the large variety of spatial and spectroscopic properties emerging from the modern hyperspectral image instrumentation. Although for simplicity, this work has only focused on coping with the spectroscopic differences among imaging platforms (1D spectroscopy vs 2D spectroscopy), more sophisticated approaches combining this solution with existing approaches to tackle differences in spatial resolution, such as those in reference 5, can be envisioned.

To wrap up this work, it is necessary to comment that the potential of the solution presented goes well beyond the field of image analysis and could be successfully applied to deal with any 2D/3D array fusion as long as one direction among arrays was common, such as in simultaneous process monitoring, where 1D and 2D spectroscopic sensors were used, environmental data, where the same sampling sites could be monitored with a set of physicochemical parameters (1D measurement) and a 2D spectroscopic sensor, or in 3D/3D array fusion if only the nonspectroscopic mode is common.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.0c00780.

> Description of the simulated data sets with a related figure. Additional figures illustrate image preprocessing and the pattern of instrumental artifacts in Raman images. Figures with MCR-ALS results of multiset analyses described in Table 2 related to individual imaging techniques and to fused multisets with equality constraints are also enclosed (PDF)

AUTHOR INFORMATION

Corresponding Authors

Adrián Gómez-Sánchez – *Chemometrics Group, Universitat de Barcelona, 08028 Barcelona, Spain*;
Email: adrian.gomez.sanchez@ub.edu

Anna de Juan – *Chemometrics Group, Universitat de Barcelona, 08028 Barcelona, Spain*; orcid.org/0000-0002-6662-2019;
Email: anna.dejuan@ub.edu

Authors

Mónica Marro – *ICFO-Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology, 08860 Barcelona, Spain*

Maria Marsal – *ICFO-Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology, 08860 Barcelona, Spain*

Pablo Loza-Alvarez – *ICFO-Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology, 08860 Barcelona, Spain*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.0c00780

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

A.G.-S. and A.J. acknowledge financial support from the Spanish government project PID 2019-1071586B-IOO and the Catalan government (2017 SGR 753). ICFO acknowledges financial support from the Spanish Ministry of Economy and Competitiveness through the "Severo Ochoa" program for Centres of Excellence in R&D (SEV-2015-0522), from Fundació Privada Cellex, Fundación Mig-Puig, Generalitat de Catalunya through the CERCA program, and "Laser lab Europe" (No. 654148).

REFERENCES

(1) *Infrared and Raman Spectroscopic Imaging*, 2nd ed.; Salzer, R., Siesler, H., Eds.; Wiley-VCH: Weinheim, Germany, 2014.

(2) *Hyperspectral imaging*, 1st ed.; Amigo, J. M., Ed.; Data Handling in Science and Technology; Elsevier: Oxford, 2020; Vol. 32.

(3) Piqueras Solsona, S.; Maeder, M.; Tauler, R.; de Juan, A. *Chemom. Intell. Lab. Syst.* **2017**, *164*, 32–42.

(4) De Juan, A.; Gowen, A.; Duponchel, L.; Ruckebusch, C. *Data Handl. Sci. Technol.* **2019**, *31*, 311–344.

(5) Piqueras, S.; Bedia, C.; Beleites, C.; Krafft, C.; Popp, J.; Maeder, M.; Tauler, R.; De Juan, A. *Anal. Chem.* **2018**, *90* (11), 6757–6765.

(6) Bioucas-Dias, J. M.; Plaza, A.; Dobigeon, N.; Parente, M.; Du, Q.; Gader, P.; Chanussot, J. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE J-STARS* **2012**, *5* (2), 354–379.

(7) Rutan, S. C.; de Juan, A.; Tauler, R. *Comprehensive Chemometrics* **2020**, *2*, 85−94.

(8) De Juan, A.; Tauler, R. *Data Handl. Sci. Technol.* **2016**, *30*, 5−51.

(9) Hugelier, S.; Devos, O.; Ruckebusch, C. *J. Chemom.* **2015**, *29*, 557−561.

(10) Hugelier, S.; Piqueras, S.; Bedia, C.; De Juan, A.; Ruckebusch, C. *Anal. Chim. Acta* **2018**, *1000*, 100−108.

(11) Bro, R. *Chemom. Intell. Lab. Syst.* **1997**, *38* (2), 149−171.

(12) Acar, E.; Bro, R.; Smilde, A. K. *Proc. IEEE* **2015**, *103* (9), 1602−1620.

(13) Navarro-Reig, M.; Jaumot, J.; García-Reiriz, A.; Tauler, R. *Anal. Bioanal. Chem.* **2015**, *407* (29), 8835−8847.

(14) Eilers, P. *Anal. Chem.* **2004**, *76*, 404−411.

(15) De Juan, A.; Maeder, M.; Hancewicz, T.; Tauler, R. *J. Chemom.* **2008**, *22* (5), 291−298.

(16) Tauler, R. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 133−146.

(17) Windig, W.; Guilment, J. *Anal. Chem.* **1991**, *63*, 1425−1432.

(18) Ndao, A. S.; Konté, A.; Biaye, M.; Faye, M. E.; Faye, N. A. B.; Wagué, A. *J. Fluoresc.* **2005**, *15* (2), 123−129.

(19) Donaldson, L. *IAWA J.* **2013**, *34* (1), 3−19.

(20) Zhang, X.; Chen, S.; Xu, F. *J. Visualized Exp.* **2017**, No. 124, e55910.

(21) Tschirner, N.; Schenderlein, M.; Brose, K.; Schlodder, E.; Mroginski, M. A.; Thomsen, C.; Hildebrandt, P. *Phys. Chem. Chem. Phys.* **2009**, *11*, 11471−11478.

(22) Olmos, V.; Marro, M.; Loza-Alvarez, P.; Raldúa, D.; Prats, E.; Padrós, F.; Piña, B.; Tauler, R.; de Juan, A. *J. Biophotonics* **2018**, *11*, No. e201700089.

(23) De Juan, A.; Tauler, R. *J. Chemom.* **2001**, *15* (10), 749−771.

https://dx.doi.org/10.1021/acs.analchem.0c00780
Anal. Chem. 2020, 92, 9591−9602

Supplementary material

3D and 4D image fusion: coping with differences in spectroscopic modes among hyperspectral images

Adrián Gómez-Sánchez[†*], Mónica Marro[‡], Maria Marsal[‡], Pablo Loza-Alvarez[‡], Anna de Juan[†*].

Abstract

The supporting information includes the description of the simulated data sets with the related figure S1. Figures S2 and S5 illustrate image preprocessing and the pattern of instrumental artifacts in Raman images, respectively. Figures with MCR-ALS results mentioned in Table 2 are related to multisets of fluorescence images (Figure S3), of Raman images (Figure S4) and to fused multisets of 3D Raman and 4D fluorescence images with equality constraints (Figures S6 and S7).

Simulated data set

The simulated data consist of the 3D Raman and 4D EEM fluorescence images of a single sample. The sample has three components, two of them detectable by EEM fluorescence and only two of them detectable by Raman. Figure S1 shows the distribution maps used in the simulation for the three components, extracted from an MCR-ALS analysis on another rice leaf sample, the three EEM spectra (red for excitation spectra and black for emission spectra) and the two Raman spectra. As it can be seen, there is a significant overlap in the spectra used, especially noticeable in the emission and Raman spectra. This situation is common in spectra of biological compounds from similar tissues.

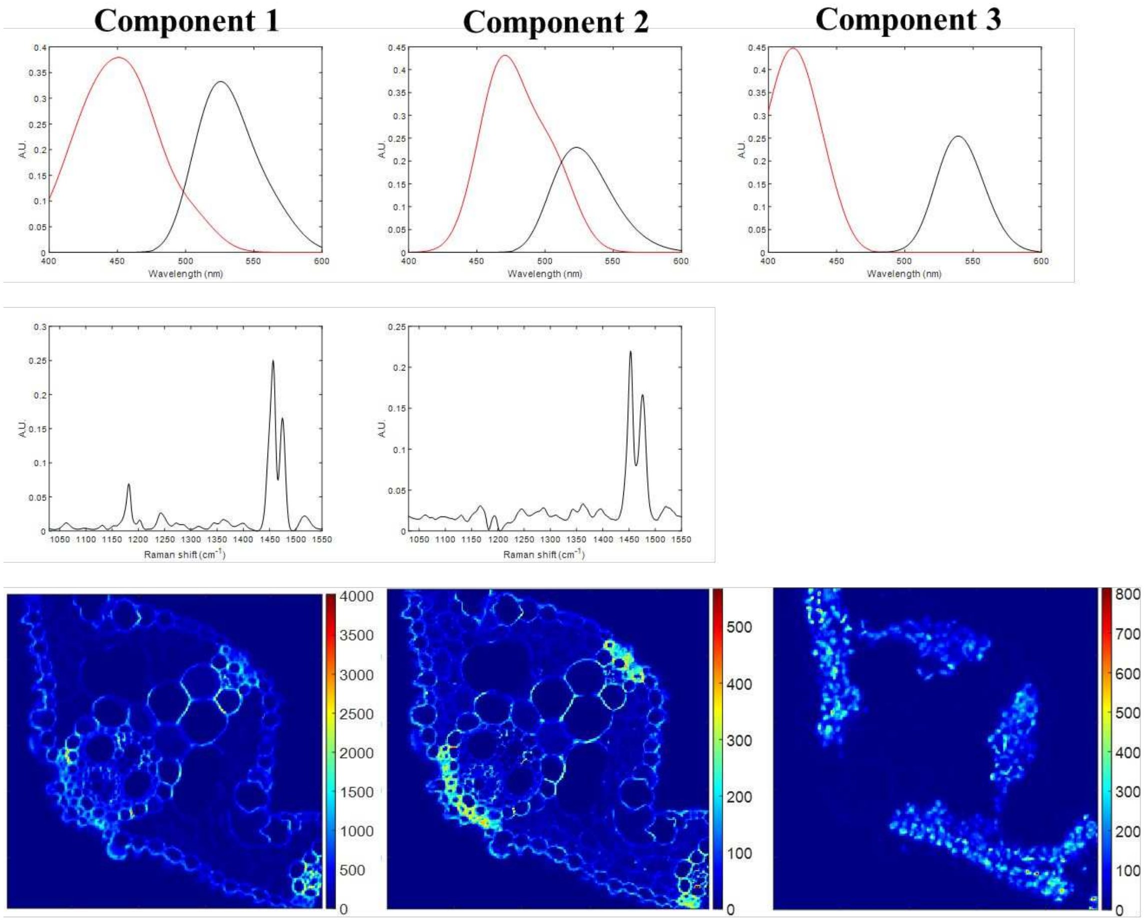


Figure S1. Pure profiles used to generate the EEM and Raman images. Top plots: pure excitation (red) and emission (black) spectra, middle plots: Raman spectra; bottom plots: related distribution maps.

The data matrix related to the 3D Raman image ($\mathbf{D}_{\mathbf{RS}}$) was generated unfolding the maps from components 1 and 2 into linear concentration profiles to form the matrix $\mathbf{C}_{\mathbf{RS}}$ and then multiplying them by the matrix $\mathbf{S}_{\mathbf{RS}}{}^{\mathbf{T}}$, formed by the two related pure Raman spectra, according to the bilinear model of equation $\mathbf{D}_{\boldsymbol{RS}} = \mathbf{C}_{\boldsymbol{RS}}\mathbf{S}_{\mathbf{RS}}^{\mathbf{T}}$ (see Figure 1a for a better understanding). i.i.d noise was added to the Raman data set $\mathbf{D}_{\mathbf{RS}}$. If the 3D Raman image needs to be recovered, the data matrix $\mathbf{D}_{\mathbf{RS}}$, sized (x

× y, nr. Raman spectral channels) can be easily refolded into a data cube $\underline{\mathbf{D}_{\mathbf{RS}}}$, sized (x, y, nr. Raman spectral channels) (see Figure 1a in main manuscript).

To generate the data matrix related to the 4D fluorescence image, $\mathbf{D}_{\mathbf{FS}}$, first the pure 2D EEM spectrum per each of the three components was generated multiplying $\mathbf{s}_{\mathbf{ex,i}}\mathbf{s}_{\mathbf{em,i}}^{\mathbf{T}}$ where $\mathbf{s}_{\mathbf{ex,i}}$ is the excitation spectrum for the ith component and $\mathbf{s}_{\mathbf{em,i}}^{\mathbf{T}}$ the related emission spectrum. For each component, the 2D EEM landscape, sized (λ_{ex}, λ_{em}), is vectorized concatenating in a single row vector the emission spectra related to the different excitation wavelengths. The vectorized EEM spectra of the three components were put together to form the matrix, $\mathbf{S}_{\mathbf{FS}}^{\mathbf{T}}$. The distribution maps of the three components were unfolded into linear concenrtration profiles to form the matrix $\mathbf{C}_{\mathbf{FS.}}$ The data matrix $\mathbf{D}_{\mathbf{FS}}$ was obtained through the product $\mathbf{D}_{\mathbf{FS}} = \mathbf{C}_{\mathbf{FS}}\mathbf{S}_{\mathbf{FS}}^{\mathbf{T}}$. If the 4D image hypercube wants to be recovered, $\mathbf{D}_{\mathbf{FS}}$, sized (x × y, λ_{ex}, λ_{em}), can be refolded into the 4D hypercube $\underline{\mathbf{D}_{\mathbf{FS}}}$, sized (x, y, λ_{ex}, λ_{em}). Poisson noise was added to $\mathbf{D}_{\mathbf{FS}}$ using the Matlab function *poissrnd* to simulate the natural noise in the fluorescence hyperspectral images.

The 3D/4D fused multiset [$\mathbf{D}_{\mathbf{RS}}$ $\mathbf{D}_{\mathbf{FS}}$] was formed as indicated in Figure 2a of the main manuscript. The amount of noise added to the multiset has been calculated according to Eq 1.

$$E(\%) = \sqrt{\frac{\Sigma_i\Sigma_j(d_{ij,n} - d_{sij,nf})^2}{\Sigma_i\Sigma_j(d_{sij,n})^2}} \cdot 100 \qquad (1)$$

Where $d_{sij,n}$ and $d_{ij,nf}$ denote the ijth element of the raw multiset [$\mathbf{D}_{\mathbf{RS}}$ $\mathbf{D}_{\mathbf{FS}}$] with noise added and the ijth element of the noise-free multiset, respectively. The E(%) added was 14.49%, similar to the noise level found in the real analyses for this kind of multisets. This means that when this multiset is analysed by MCR-ALS, a lack of fit very similar to E(%) should be obtained.

MCR-ALS analysis is performed in the multiset using a pure bilinear model and a hybrid bilinear/trilinear model. In both analyses, non-negativity and equality constraints in **C** and normalisation of the profiles in the augmented spectral direction were used. The main results from MCR-ALS analyses, i.e., the distribution maps and recovered spectra, are in Figures 3 and 4 of the manuscript, together with the main text.

For a more detailed comparison of the simulated profiles and the profiles recovered by MCR-ALS analyses, table S1 is included.

Table S1. Correlation coefficients between simulated and resolved MCR profiles.

Model	Components	Concentration profile	Raman spectrum	Excitation spectrum	Emission spectrum
Bilinear	1	0.994	0.997	0.996	0.997
	2	0.938	0.979	0.830	0.891
	3	0.984	-	1.000	1.000
Hybrid bilinear/trilinear	1	1.000	1.000	1.000	1.000
	2	0.998	1.000	1.000	1.000
	3	0.999	-	1.000	1.000

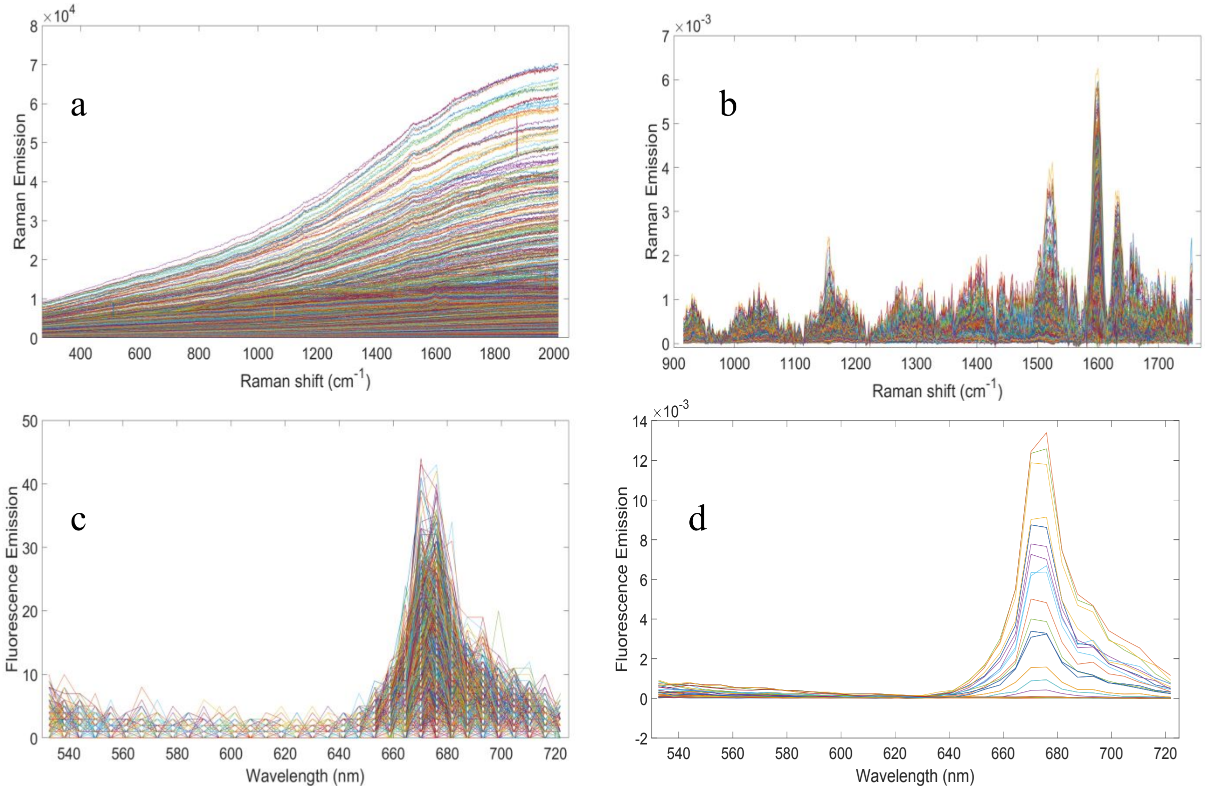


Figure S2. a) Raw Raman spectra. Note the big fluorescence baseline contribution from vegetal tissues. b) Corrected Raman spectra by Asymmetric Least Squares. c) Raw fluorescence spectra collected at 494 nm excitation-wavelength. Note the low and noisy signal detected. d) Fluorescence spectra after binning.

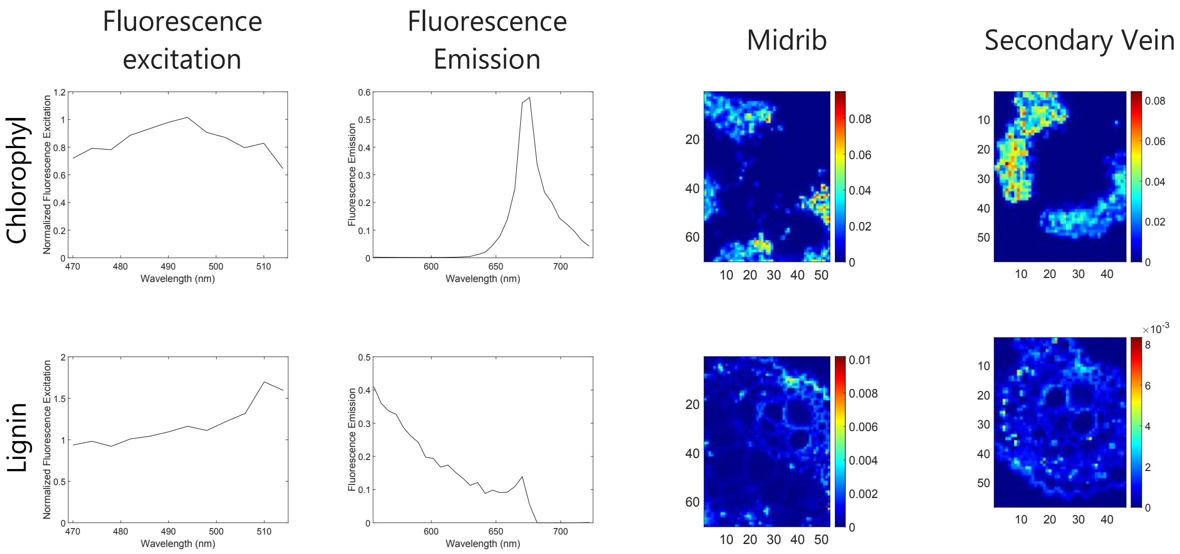


Figure S3. Pure fluorescence profiles and pure distribution maps from MCR-ALS analysis. Plots in columns indicate from left to right: resolved excitation spectra, emission spectra, maps from midrib and from secondary vein of rice leaves.

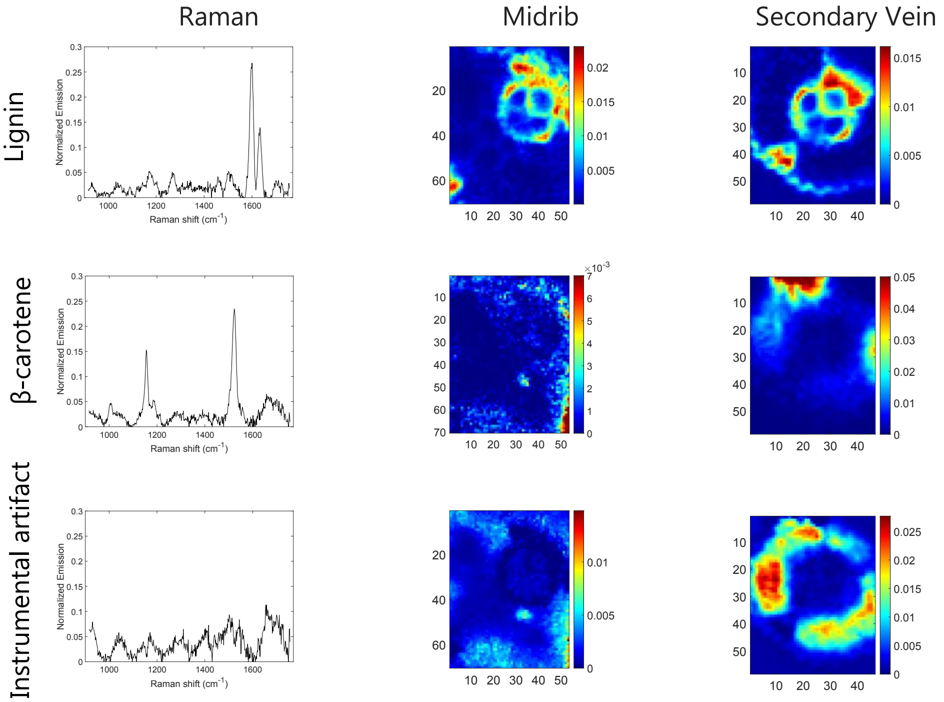


Figure S4. Pure Raman profiles and pure distribution maps from MCR-ALS analysis.

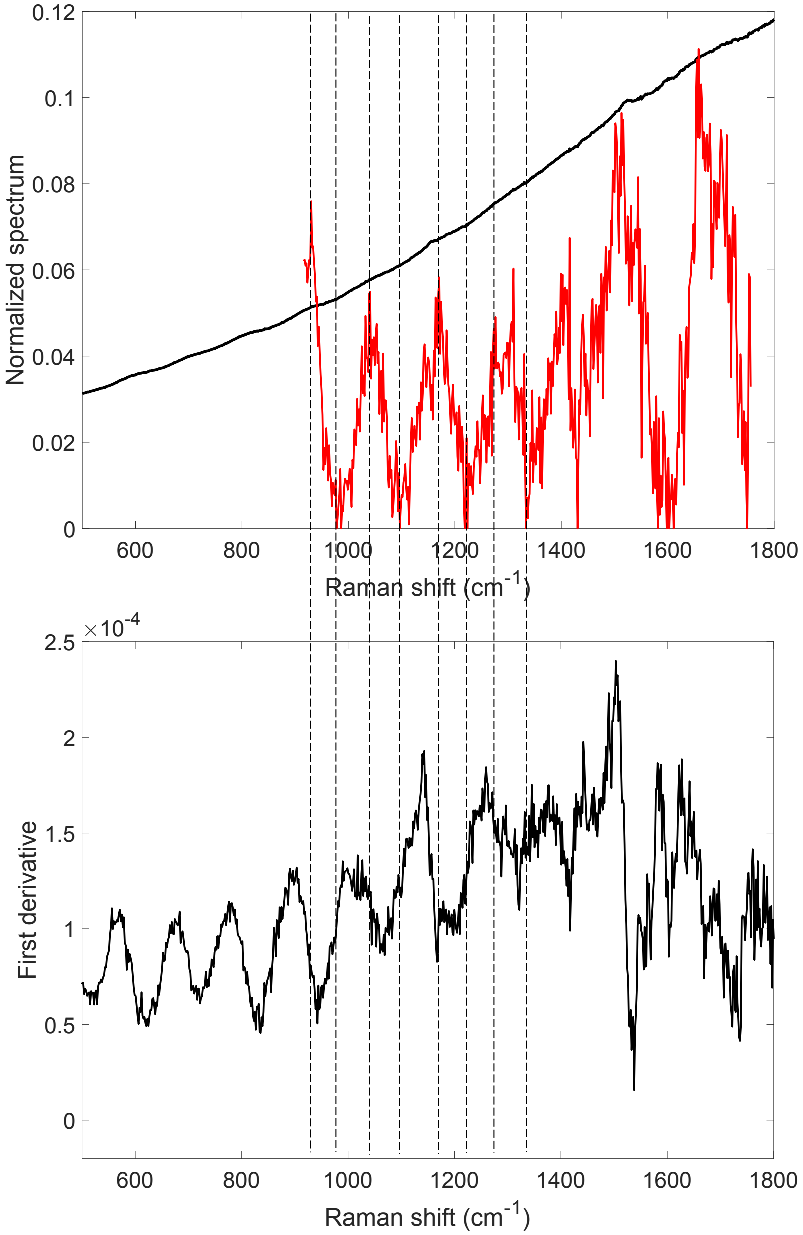


Figure S5. Top plot: in black, mean spectrum of the secondary vein Raman image; in red, pure spectral profile of the instrumental artifact extracted by MCR-ALS. It is possible to observe the match between the sinusoidal shape of the mean spectrum and the peaks of the pure spectrum profile. Bottom plot: first derivative of the mean spectrum of the secondary vein. It can be observed the matching between the maxima and minima in the red spectrum from the artifact with the inflexion points of the first derivative of the mean spectrum of the secondary vein image. This observation discards the hypothetic unintended side effect of the AsLS baseline correction in the emergence of the artifact extracted by MCR.

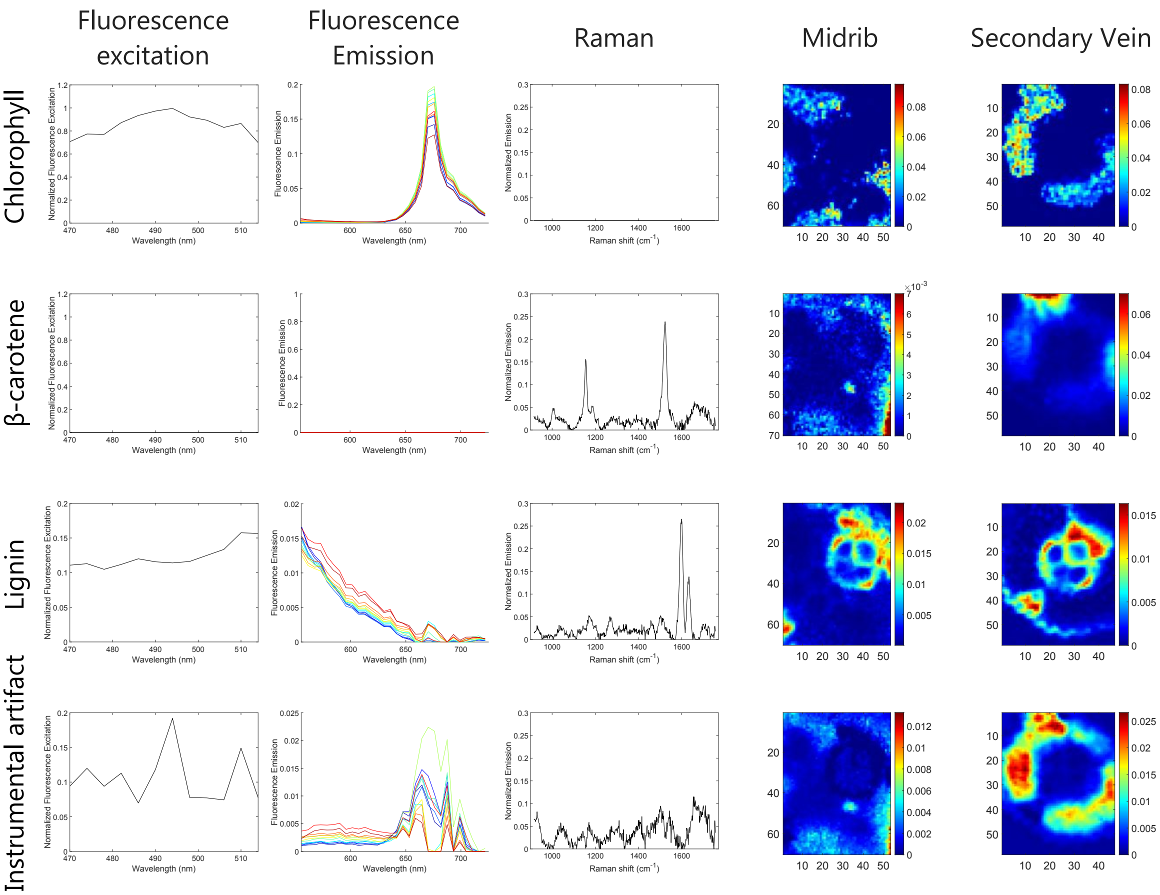


Figure S6. Pure Raman profiles, pure fluorescence spectra and pure distribution maps from MCR-ALS analysis. Note that even if equality constraint is applied, an intracomponent variation of the shape in the emission spectra can be observed.

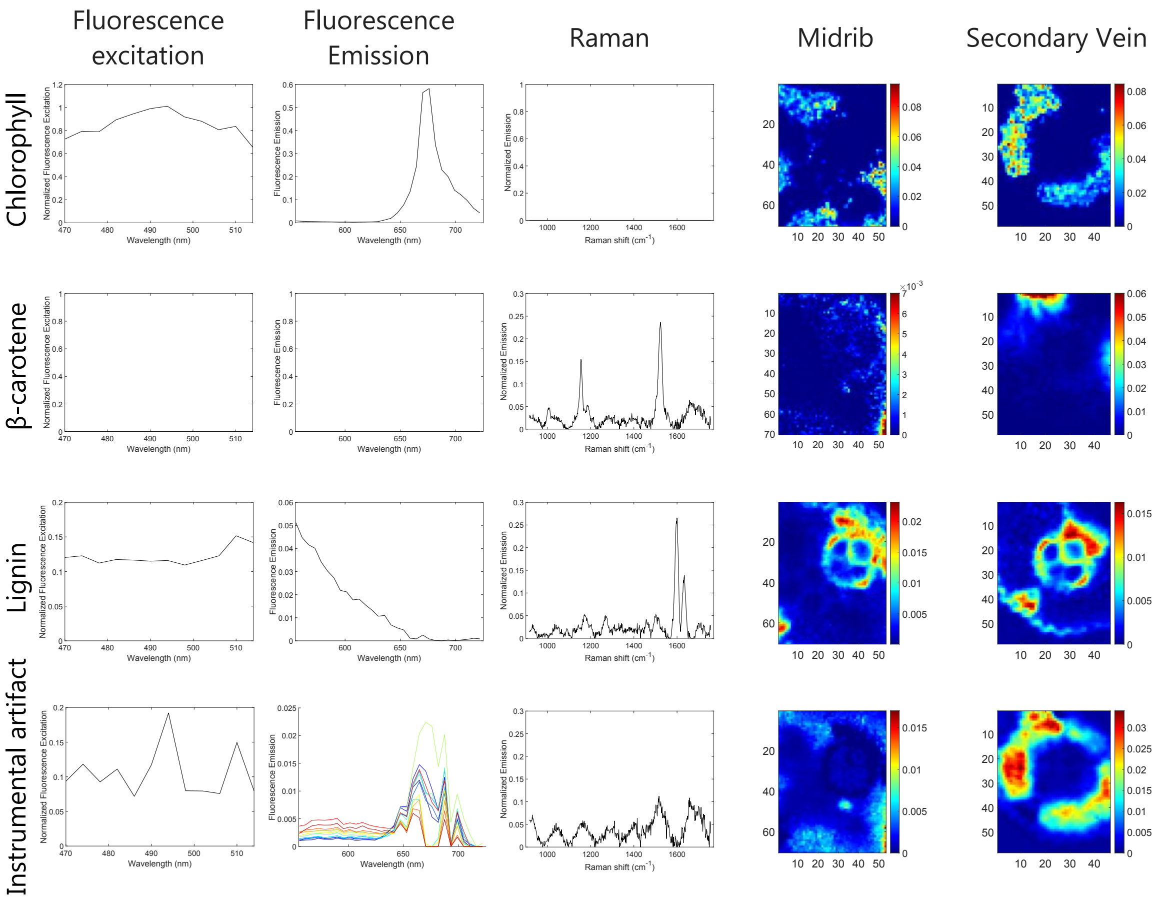


Figure S7. Pure Raman profiles and pure maps distributions output of MCR-ALS analysis. Note that even if equality constraint is applied, no substantial improvement in the results is perceived compared with Figure 6 in the main manuscript since trilinearity constraint is sufficient to recover the correct maps and spectral profiles.

Erratum

3D and 4D Image Fusion: Coping with Differences in Spectroscopic Modes among Hyperspectral Images

Adrián Gómez-Sánchez, Mónica Marro, Maria Marsal, Pablo Loza-Alvarez, and Anna de Juan.

Please, note that Figure 5 and Figure 6 of the article have been duplicated. Figure 6 should be replaced with the following one.

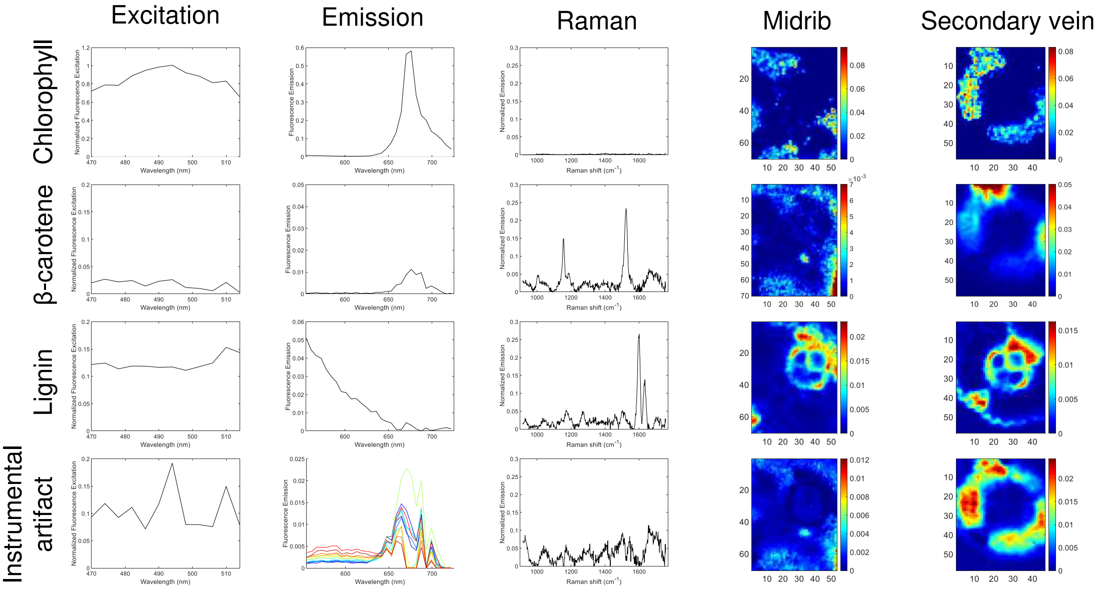
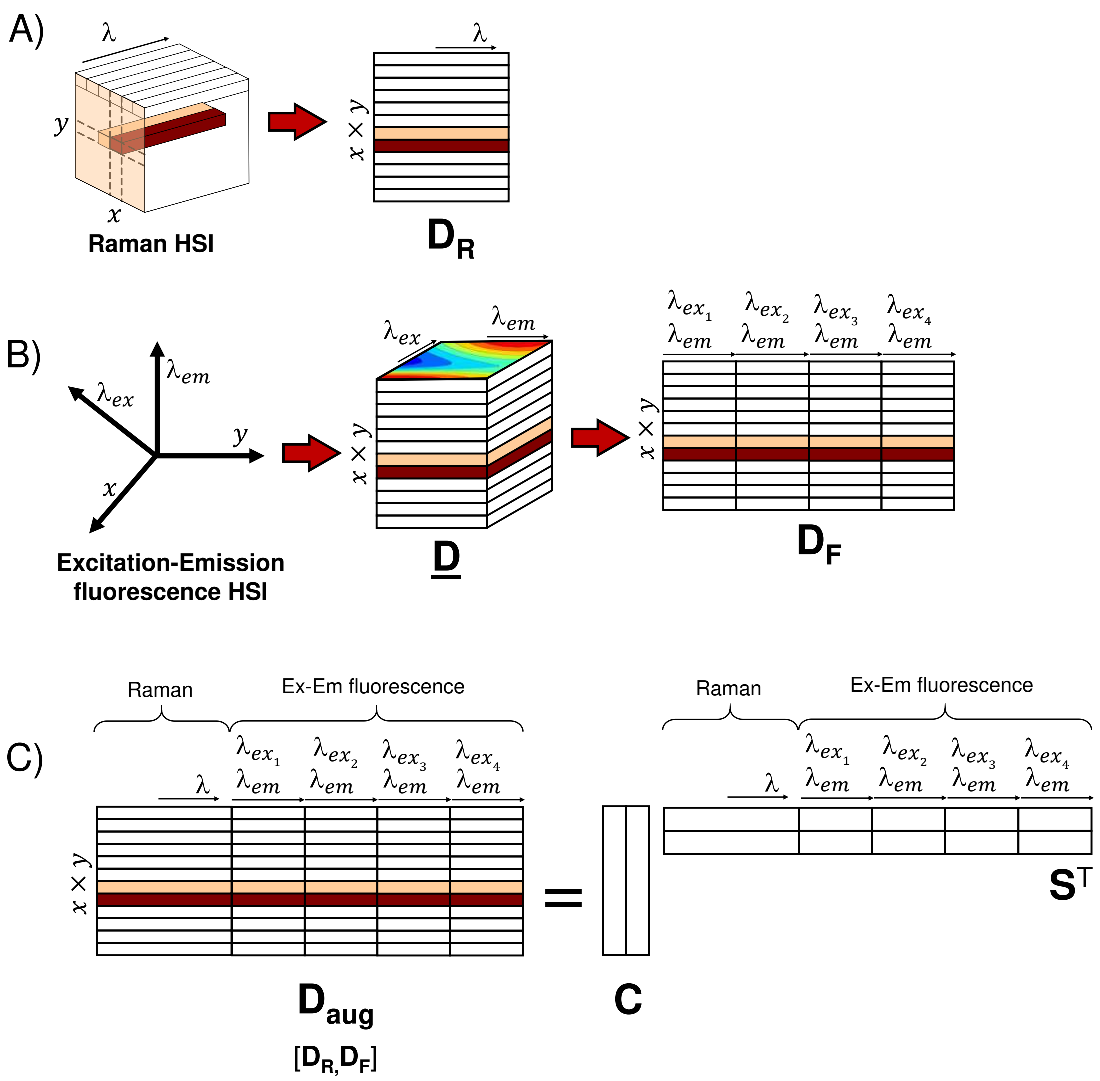


Figure 6. MCR results obtained on the Raman/fluorescence image multiset using a hybrid bilinear/trilinear model. Plots in columns indicated from left to right: resolved excitation spectra, emission spectra, Raman spectra, map from midrib and from secondary vein of rice leaves.

Publication VII addresses the problem of fusing HSIs with different spectral dimensionality. For example, Raman images can be represented as a data cube with three dimensions 3D (x, y, λ), while EEM images are described with a data array of four dimensions 4D (x, y, λ_{ex}, λ_{em}). The fusion of both kinds of images is not trivial, since they have different dimensions and the Raman image follows a bilinear model and the EEM image a trilinear model.

Back to the section 2.4 about challenges in data fusion, Figure 18 showed a challenging data structure where a matrix (representing an unfolded 3D image) and a tensor (representing an unfolded 4D image) were connected. The proposal in this thesis goes in the direction of solving the 3D/4D image fusion problem by a dedicated variant of MCR-ALS multiset analysis that requires a single factorization model, does not require any prior information about the components in the images to be fused (merely the global number of components) and is able to accommodate the underlying bilinear and trilinear models of the images to be combined.

To build a multiset structure merging the information of 3D and 4D images, both images need to be unfolded to form a data matrix. The 3D image is unfolded in the pixel direction, placing the Raman pixel spectra one on top of each other. The 4D image is first unfolded in the pixel direction to form a cube and, afterwards, further unfolded by vectorizing the 2D spectrum of every pixel as well until a data matrix is formed (Fig. 43A and B). The vectorization of a 2D Excitation-Emission Fluorescence landscape entails concatenating all emission spectra related to the different excitation wavelengths in the same row. This transforms the original 4D array into a data matrix, where each row represents the vectorized EEM spectrum of a pixel. Once the necessary image unfoldings have been carried out, the data matrices derived from the 3D and 4D images can be connected to create a row-wise augmented multiset (Fig. 43C).



A)
λ
y
x
Raman HSI
λ
$x \times y$
$\mathbf{D_R}$
B)
λ_{em}
λ_{ex}
y
x
Excitation-Emission fluorescence HSI
λ_{ex}
λ_{em}
$x \times y$
$\underline{\mathbf{D}}$
λ_{ex_1} λ_{em}
λ_{ex_2} λ_{em}
λ_{ex_3} λ_{em}
λ_{ex_4} λ_{em}
$x \times y$
$\mathbf{D_F}$
C)
Raman
Ex-Em fluorescence
λ
λ_{ex_1} λ_{em}
λ_{ex_2} λ_{em}
λ_{ex_3} λ_{em}
λ_{ex_4} λ_{em}
$x \times y$
=
$\mathbf{S}^T$
$\mathbf{D_{aug}}$
$[\mathbf{D_R}, \mathbf{D_F}]$
$\mathbf{C}$

Figure 43. A) Unfolding of a 3D Raman image into data matrix $\mathbf{D_R}$ B) Unfolding of a 4D EEM fluorescence image into $\mathbf{D_F}$. C) Row-wise augmented multiset $\mathbf{D_{aug}}$ formed by matrices $\mathbf{D_R}$ and $\mathbf{D_F}$.

Now, the multiset $\mathbf{D_{aug}}$ could be analyzed by MCR-ALS assuming bilinearity for all the multiset. However, it is known that the Raman block of the multiset follows a bilinear model, while the emission fluorescence blocks follows a trilinear model. Although analyzing trilinear data using a bilinear model is a correct option [Chaumel et al., 2021], the uniqueness advantage of trilinear models is lost. Therefore, the ideal scenario would be preserving the natural model of each imaging measurement to increase the accuracy of the final solutions.

To do so, a modification of the MCR-ALS algorithm is proposed to introduce a hybrid bilinear/trilinear model that can preserve the natural underlying spectroscopic model of the fused 3D and 4D images. The proposed approach involves a modification of the current implementation of the trilinearity constraint to enable the optional application *per block*, achieving a partial trilinearity in the multiset model if required.

In the Raman-EEM fusion, the trilinearity constraint applies only to the emission spectra blocks of matrix **S**^{T}**,** while the spectral block associated with the Raman data are exempt from this requirement (Fig. 44). The implementation of the trilinearity constraint in the blocks selected follows the procedure described in Fig. 20B. The hybrid bilinear/trilinear model obtained ensures that the trilinear model is respected for the excitation-emission fluorescence data and, hence, the uniqueness for the retrieved profiles, while preserving the bilinear model for the Raman measurement.

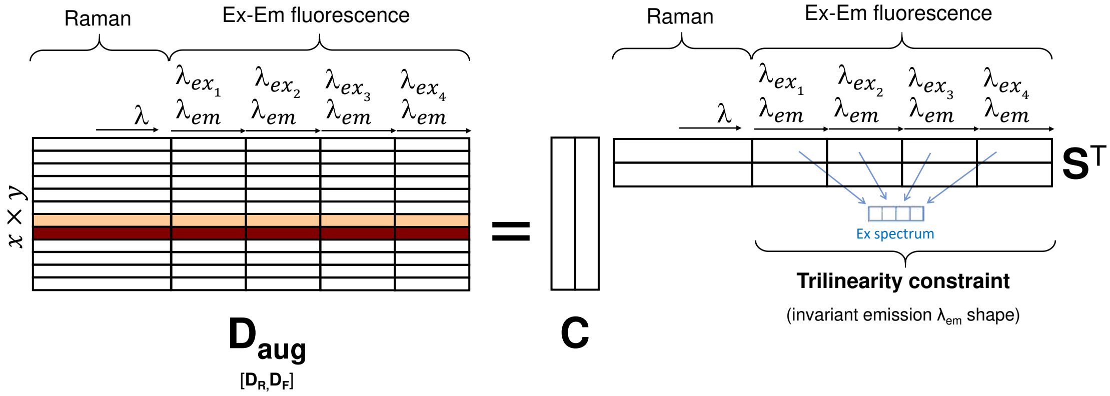

Raman
Ex-Em fluorescence
λ
λ_{ex_1} λ_{em}
λ_{ex_2} λ_{em}
λ_{ex_3} λ_{em}
λ_{ex_4} λ_{em}
$x \times y$
=
$\mathbf{D_{aug}}$
$[\mathbf{D_R},\mathbf{D_F}]$
C
$\mathbf{S^T}$
Ex spectrum
Trilinearity constraint
(invariant emission λ_{em} shape)

Figure 44. Multiset formed by a 3D Raman image and a 4D fluorescence image and related MCR model, with only emission blocks constrained to follow a trilinear behavior.

The new partial trilinearity constraint has been tested for the study of cross-sections of rice leaves measured by Raman imaging (3D) and EEM imaging (4D). Additionally, a simulated example is included to rigorously validate the proposed methodology in controlled conditions.

Results of the simulated dataset

The partial trilinearity constraint was tested in a simulated dataset consisting of the fusion of a 3D Raman and a 4D EEM image, where the Raman block follows a bilinear model and the emission fluorescence blocks follow a trilinear model. The simulated multiset contained three components, where one of them has no detectable signal in Raman. Both 3D and 4D simulated images were concatenated as in Fig. 43C.

The multiset was analyzed by MCR-ALS in two different ways: a) using a bilinear model and b) using a hybrid bilinear/trilinear model through the application of the trilinearity constraint in the emission fluorescence blocks. Both analyses showed the same lack of fit, approximately 14.5%, consistent with the added noise in the simulation. Fig. 45 and 46 show the retrieved pure distribution maps and pure spectra for the pure bilinear model and the hybrid bilinear-trilinear model, respectively. Table xxx shows the correlation coefficients between the simulated profiles and the retrieved profiles by MCR for both MCR analyses.

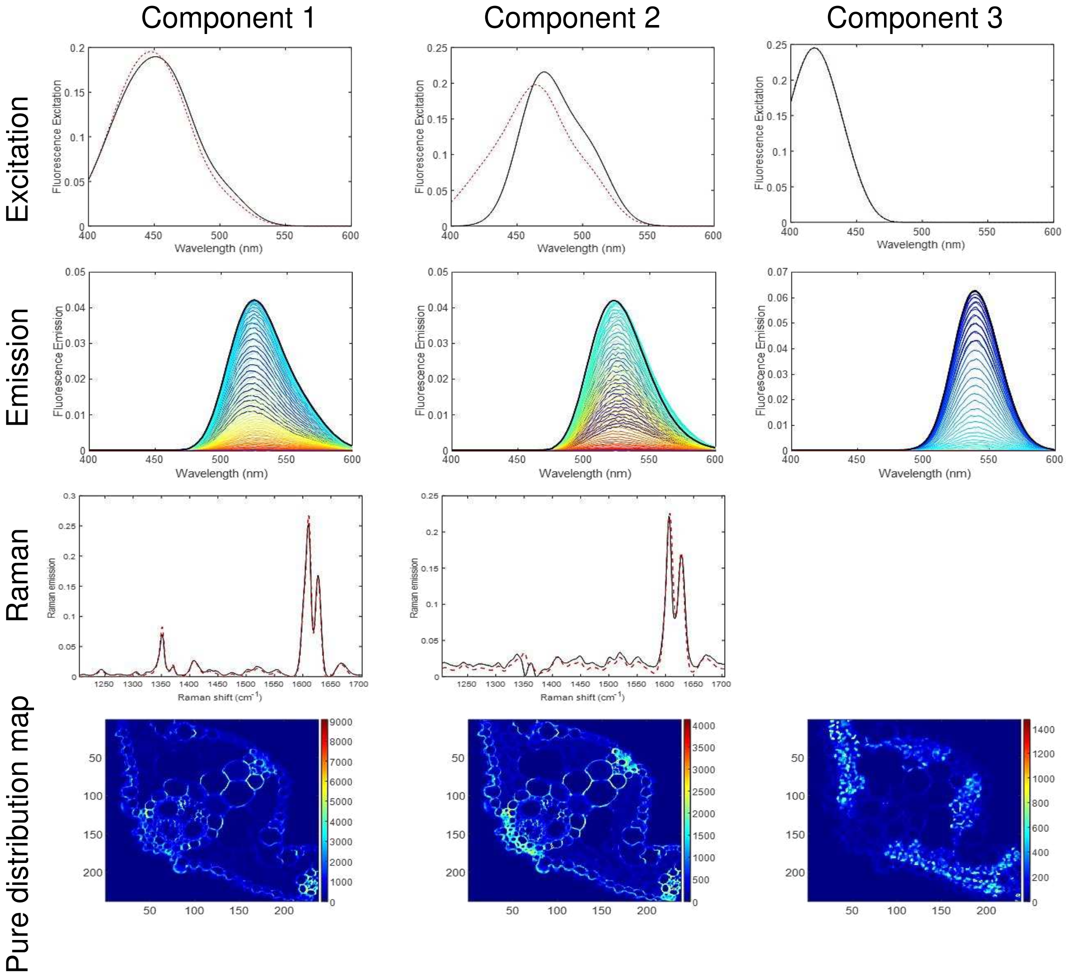


Figure 45. MCR results on the simulated dataset fusing a 3D Raman image and a 4D EEM fluorescence image using a bilinear model. Top plots: pure excitation spectra (black line: simulated profiles; red dashed lines: MCR spectra). Second row plots: pure emission spectra (black line: simulated profile shape; colored spectra from blue to red, MCR emission spectra from lowest to highest excitation wavelength). Third row plots: Raman pure spectra (black line: simulated profiles; red dashed lines: MCR spectra). Bottom plots: MCR pure distribution maps.

Figure 45 reveals that the pure emission spectra of components 1 and 2 exhibit slight differences in shape for different excitation wavelengths, which is unexpected for fluorescence measurements. These variations impact the recovery of excitation spectra, clearly erroneous for component 2 and with some deviations for the other components. These differences between simulated and recovered profiles can also be confirmed with the correlation coefficients presented in Table 1, which also point to discrepancies between the simulated and recovered distribution maps.

Instead, Figure 46 and Table 2 show that the use of the partial trilinearity constraint makes that all pure profiles recovered by MCR-ALS perfectly match the simulated ones. Table 2 clearly confirms that the hybrid model consistently provides high correlation coefficients, almost close to 1, while the coefficients for the bilinear model are notably lower.

Table 2. Correlation coefficients between simulated and resolved MCR profiles.

Model	Components	Concentration profile	Raman spectrum	Excitation spectrum	Emission spectrum
Bilinear	1	0.994	0.997	0.996	0.997
	2	0.938	0.979	0.830	0.891
	3	0.984	-	1.000	1.000
Hybrid bilinear/trilinear	1	1.000	1.000	1.000	1.000
	2	0.998	1.000	1.000	1.000
	3	0.999	-	1.000	1.000

The improvement of the results can be attributed to the use of the partial trilinear model that causes a drastic reduction, if not complete elimination, of rotational ambiguity. N. Omidikia (2022) posteriorly investigated and quantified the rotational ambiguity of a system under the influence of the partial trilinearity constraint. The study demonstrated that applying the partial trilinear constraint always helps significantly to reduce the rotational ambiguity of the pure resolved profiles by MCR-ALS. In fact, if all components are present in the bilinear and trilinear blocks, uniqueness is achieved, as it happens in the simulated data set of Fig. 46. This research also established rules to understand how shared components among blocks impact the reduction of rotational ambiguity. This study contributes to validate the proposed methodology, emphasizing the substantial improvement of results with the use of a hybrid bilinear/trilinear model in 3D/4D image fusion compared to a pure bilinear model.

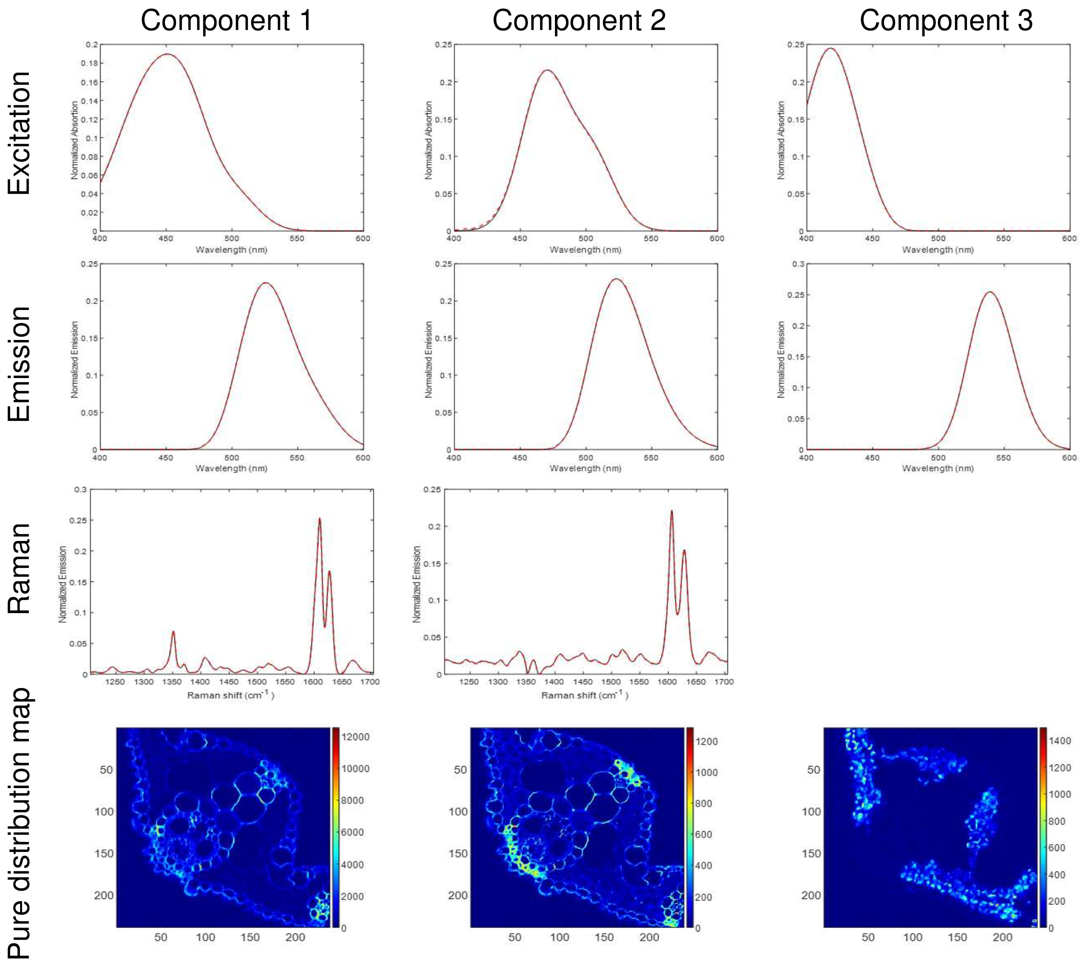


Figure 46. MCR results on the simulated dataset fusing a 3D Raman image and a 4D EEM fluorescence image from the same sample using the hybrid bilinear-trilinear model. Top plots: excitation spectra (black line: simulated profiles; red dashed lines: MCR spectra). Second row plots: emission spectra (black line: simulated profiles; red dashed lines: MCR spectra) Third row plots: Raman spectra (black line: simulated profiles; red dashed lines: MCR spectra). Bottom plots: MCR distribution maps.

Results of the analysis of real images of leaf cross sections

Two regions (named midrib and secondary vein) of a rice leave cross-section were measured by Raman and excitation-emission fluorescence imaging. After appropriate preprocessing, the images were aligned following the protocol of **Publication V** and concatenated in a single multiset as in Fig. 43C.

The analysis by MCR-ALS of the multiset was performed also using a bilinear model and a hybrid bilinear/trilinear model, with the trilinearity constraint employed in the blocks associated with the fluorescence signal. As in the simulated example, the results obtained by applying only a bilinear model (shown in **Publication VI**) showed inconsistencies such as modifications in the shape of emission spectra for the same component, which indicated that the solutions were not the expected ones. The inclusion of additional constraints within the MCR bilinear framework did not provide a clear improvement of the results obtained.

Figure 47 presents the MCR-ALS results obtained using a hybrid bilinear/trilinear model. Four components were needed to describe the system under study, three biological components and an additional contribution associated with a Raman artifact. The trilinear constraint was implemented on all fluorescence emission spectra blocks, excluding the Raman block. Within the fluorescence blocks, trilinearity was applied to all components, except to the fluorescence part related to a Raman artifact, where noise contribution is not expected to follow a trilinear behavior. This real system is a good example of the flexibility of the trilinear constraint, which is optionally applied *per block* and *per component.*

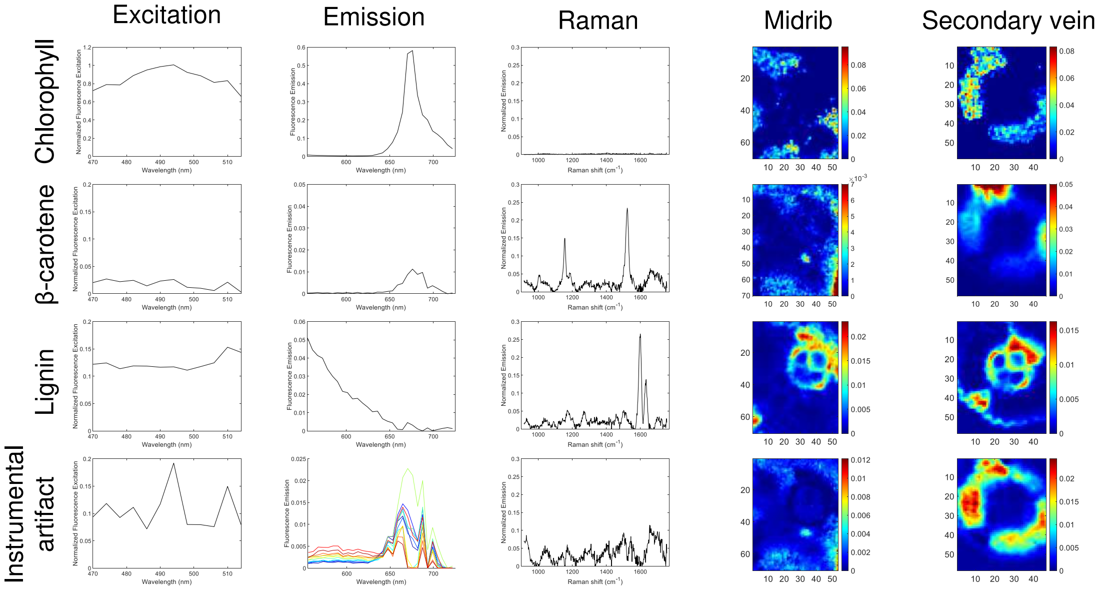


Figure 47. MCR results obtained on the Raman/fluorescence image multiset using a hybrid bilinear-trilinear model. Plots in columns indicated from left to right: resolved excitation spectra, emission spectra, Raman spectra, map from midrib and from secondary vein of rice leaves.

The MCR-ALS analysis successfully identified four components related to chlorophyll, carotene, lignin, and an instrumental artefact, already detected when Raman image analysis was carried out independently. The distribution maps locate the components as expected from a biological point of view, i.e., chlorophyll in the mesophyll cells, as well as carotenes and lignin consistently in vascular tissues. The spectroscopic features of the resolved spectra also agree with the identity of the biological compounds modelled [Zhang et al., 2017; Tschirner et al., 2009; Donaldson, 2020] (for more detail in the biological description of the components, see the original publication).

To conclude, the hybrid bilinear/trilinear models obtained through the new implementation of the partial trilinearity constraint offer a very good solution to address differences in spectral dimensionality and underlying measurement models for the 3D/4D image fusion problem. This approach preserves the appropriate underlying model for each image platform, reduces the rotational ambiguity in the resolved profiles and eliminates the need to identify common components between 3D and 4D images. The application of this approach can be extended to various applications fields beyond image analysis, such as multitechnique process monitoring where 2D EEM landscapes and other 1D spectroscopic techniques (circular dichroism, UV-Vis, mass spectroscopy…) need to be combined.

Connecting the present publication with earlier chapters of this thesis, the improvements made to the trilinearity constraint to handle missing data presented in **Publications I** and **II** could be equally applicable in this framework. Note that, in the present publication, the excitation-emission measurements were complete excitation-emission landscapes, but the incorporation of the partial trilinearity constraint presented in this work is perfectly compatible for EEM data with missing values. If that was the case, the only change would be that the trilinearity constraint would be applied to the selected blocks with missing values applying the implementation presented in **Publication II**.

3.5 Fusion of images showing differences in scanned area and spatial resolution

The spatial characteristics of images may differ very often in scanned area and spatial resolution. As described in section 3.3, classical image fusion in these instances requires identical spatial properties for all the images to be merged, i.e., working at the lowest spatial resolution and analyzing only the common scanned area. The challenge in this context is improving image fusion by using all the available information provided by the different platforms. As discussed in section 3.4, variations among scanned areas or differences in spatial resolution among platforms can result in incomplete multisets, characterized by the presence of missing blocks of information. This kind of structure challenges both unmixing methodologies, such as MCR, and exploratory methods, such as PCA, since the classical algorithms for these methods cannot handle missing data and there is room for improvement in the methodologies that have been designed for this purpose. Therefore, the two last publications of this thesis are focused on providing modifications of MCR-ALS and PCA for a more efficient analysis of incomplete multisets.

Publication VIII proposes a modification of MCR-ALS to analyze incomplete multisets by doing a single factorization and without the need of data imputation. The approach is versatile and can adapt to any pattern of missing entries. It can be used for any data fusion application involving blocks of information obtained in non-equivalent experimental conditions.

Publication IX proposes a novel algorithm, called Orthogonalized Alternating Least Squares (O-ALS), to perform PCA in data sets with missing blocks of information. O-ALS is an iterative ALS method that estimates the scores and loadings subject to the Gram-Schmidt orthogonalization constraint. It works using a single factorization and does not require any data imputation step. O-ALS is compared with PCA approaches designed to tackle the missing value problem.

Publication VIII. Dealing with missing data blocks in Multivariate Curve Resolution. Towards a general framework based on a single factorization model.
Authors: A. Gómez-Sánchez, C. Ruckebusch, R. Tauler, A. de Juan.
Trends in Analytical Chemistry (2024) (submitted).

Publication IX. Solving the missing value problem in PCA by Orthogonalized-Alternating Least Squares (O-ALS)
Authors: A. Gómez-Sánchez, R. Vitale, C. Ruckebusch, A. de Juan.
Chemometrics and Intelligent Laboratory Systems (2024) (submitted).

Dealing with missing data blocks in Multivariate Curve Resolution. Towards a general framework based on a single factorization model.

Adrián Gómez-Sánchez[1,2(*)], Cyril Ruckebusch[2], Romà Tauler[3], Anna de Juan[1(*)]

[1]Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028, Barcelona, Spain

[2]LASIRE - Laboratory of Advanced Spectroscopy, Interactions, Reactivity and Environment Université Lille, CNRS, UMR 8516, Cité Scientifique, Bâtiment C5, 59000, Lille, France

[3]Department of Environmental Chemistry, Institute of Environmental Assessment and Water Research (IDAEA-CSIC), 08034 Barcelona, Spain

(*) Corresponding authors: aderegomez@gmail.com, anna.dejuan@ub.edu.

Abstract

Multivariate Curve Resolution (MCR) deals with the mixture analysis problem by decomposing a data set with mixed information into a bilinear model of pure component contributions. Multiset analysis deals with fused data blocks linked to related experiments and/or techniques. Nevertheless, experiments and techniques often show differences that lead, when concatenated, to incomplete multisets with missing blocks of information. Incomplete multisets aim at incorporating all available information in the initial blocks of measurements but require adapted algorithms to be properly handled. This work presents the evolution of the different perspectives adopted to analyze incomplete multisets with advantages and drawbacks. Finally, a new methodology is proposed that adapts to any data configuration with missing entries without the need to perform data imputation or multiple factorizations. The new method adapts very well to analytical applications where the blocks of information to be fused are not acquired in equivalent experimental conditions.

Keywords: Incomplete multiset, Multivariate Curve Resolution Alternating Least Squares (MCR-ALS), missing values.

The MCR method. Concept and data configurations.

Multivariate Curve Resolution encloses a family of methods devoted to solving the mixture analysis problem through a bilinear decomposition that incorporates dyads of profiles describing the pure contributions of the components involved in the mixed information [1-3].

The bilinear model is expressed through the equation 1:

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

where **D** represents the matrix of mixed information and $\mathbf{S}^T$ and **C** matrices contain pure component profiles describing responses (qualitative information) and related concentration (or contribution) profiles, respectively. **E** accounts for the non-explained variance by the bilinear $\mathbf{CS}^T$ model. Although the paradigm of MCR models is associated with the description of spectroscopic data, where **D** contains spectroscopic mixtures, $\mathbf{S}^T$ pure component spectra and **C** the related concentration profiles, it is well known that MCR finds application in a wide variety of contexts ranging from a large diversity of instrumental responses to environmental or -omics data [1].

MCR was first envisioned as a two-way data analysis methodology, focused on the analysis of a single data table of information [2,3]. However, soon became clear that concatenating several data tables with related information provided more accurate information about the systems and the related pure components to be described, i.e., the inherent ambiguity of the bilinear MCR decomposition drastically decreased. This evolution gave rise to what we know today as MCR multiset analysis [1,4,5]. Multisets are very flexible data structures, defined as augmented matrices formed by several blocks of information that may have different dimensions and chemical meanings. To mention few examples, a multiset can be formed by the blocks of different instrumental sensor outputs monitoring the same process [6,7], or by a set of experiments monitored with the same technique [7-10], or by the combination of environmental data tables collected along time or on different environmental compartments [11-13]. Building a multiset always requires that the row and/or the column mode of the matrices to be connected are common, but other dimensions can be completely free. Figure 1 shows the architecture of the most complex multiset, defined as a row-and column-wise augmented matrix. The example in the figure could represent the combination of two experiments monitored by two different instrumental techniques, although augmentation can happen in only one of the directions of the bilinear model and the number and configuration of blocks forming a multiset does not have any kind of restriction.

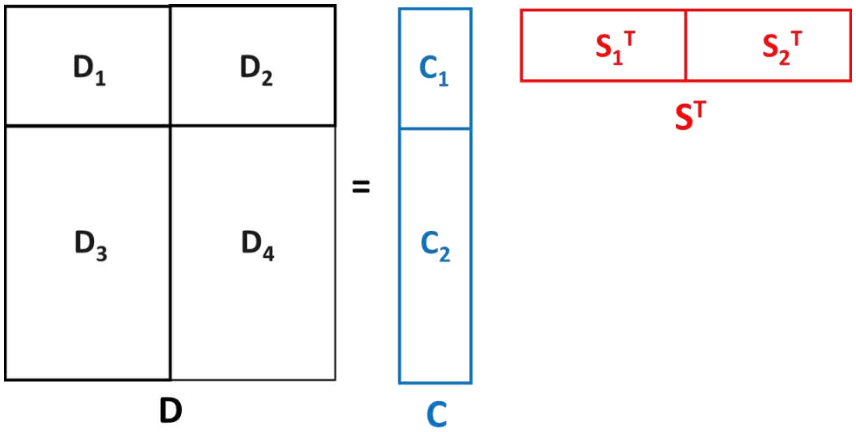


Figure 1. Bilinear MCR model related to a row-and column-wise augmented multiset.

Formally speaking, a major asset of MCR multiset analysis is that the bilinear model in equation 1 still holds and no mathematical complexity is added to handle this kind of data structure. The gain in multiset analysis is the incorporation of diversity in the initial information and the completely different meaning and shapes allowed for the different $\mathbf{C}_i$ and $\mathbf{S}_i^T$ profiles.

Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) is the MCR methodology that introduced the idea of multiset analysis and has extensively exploited the concept in many different applications [1,5]. MCR-ALS works optimizing the profiles in **C** and $\mathbf{S}^T$ under constraints. In a multiset context, there is a total flexibility to tune the constraints needed to define the profiles in each of the $\mathbf{C}_i$ and $\mathbf{S}_i^T$ blocks. In MCR-ALS, the two least-squares steps used in every iterative cycle to retrieve **C** and $\mathbf{S}^T$ are expressed in equations 2 and 3, respectively:

$$\mathbf{C} = \mathbf{D}\,\mathbf{S}\left(\mathbf{S}^T\mathbf{S}\right)^{-1} \qquad (2)$$

$$\mathbf{S}^T = \left(\mathbf{C}^T\mathbf{C}\right)^{-1}\mathbf{C}^T\mathbf{D} \qquad (3)$$

The goal of the method is minimizing the residuals issued from the comparison of the initial **D** information and the one retrieved by the $\mathbf{CS}^T$ bilinear model, $\min\left\|\mathbf{D} - \mathbf{CS}^T\right\|$. In the multiset context of Figure 1, which represents a row and column-wise augmented matrix **D**, $\mathbf{D} = [\mathbf{D}_1\ \mathbf{D}_2;\ \mathbf{D}_3\ \mathbf{D}_4]$, $\mathbf{C} = [\mathbf{C}_1;\ \mathbf{C}_2]$ and $\mathbf{S}^T = [\mathbf{S}_1^T\ \mathbf{S}_2^T]$. In this notation, the semi-colon (;) indicates that the augmentation is in the column direction (one matrix block on top of each other) and the blank space that augmentation is in the row direction (one matrix block besides the other). Steps 2 and 3 work as defined as long as no missing values exist in the initial structure **D**. In the presence of missing information, different perspectives and kinds of solutions have been adopted until the general framework proposed in this work has been proposed.

MCR for incomplete multisets. The evolution around the missing block problem.

Multisets are built connecting blocks of information that share the row and/or column direction. However, the available information does not always allow building a full regular data structure because some of the necessary blocks of information are missing. When this happens, a new kind of data configuration appears, defined as incomplete multisets [14]. Incomplete multisets arise from the fact that it is not always easy to get blocks of information in equivalent experimental/time conditions. Many scenarios providing incomplete multisets have been reported in the literature. For instance, in environmental monitoring campaigns [14], blocks of information related to a specific month may be missing because measurements could not be carried out; in a multitechnique multiset, the measurements provided by a specific technique may be unfeasible in certain experimental conditions [15], or when coupling hyperspectral images from different platforms, not all of them can provide the same spatial resolution [16, 17]. Nevertheless, it is important to find ways to use actively all available information, even if this implies working with challenging data structures. An additional complexity in incomplete multisets is the pattern of the missing information, i.e., missing values are not randomly scattered within the multiset, but they are distributed as complete blocks of missing information.

There are two different perspectives to address the missing information problem in MCR, namely: a) filling the initial missing information with estimated values and proceed with the data analysis on the completed structure (see Figure 2a and b) recovering the bilinear model only with the real available information and, eventually, reproduce the full complete structure if required (see Figure 2b). It is important to look carefully at the advantages and drawbacks of the existing approximations around these two perspectives before describing the new methodology proposed.

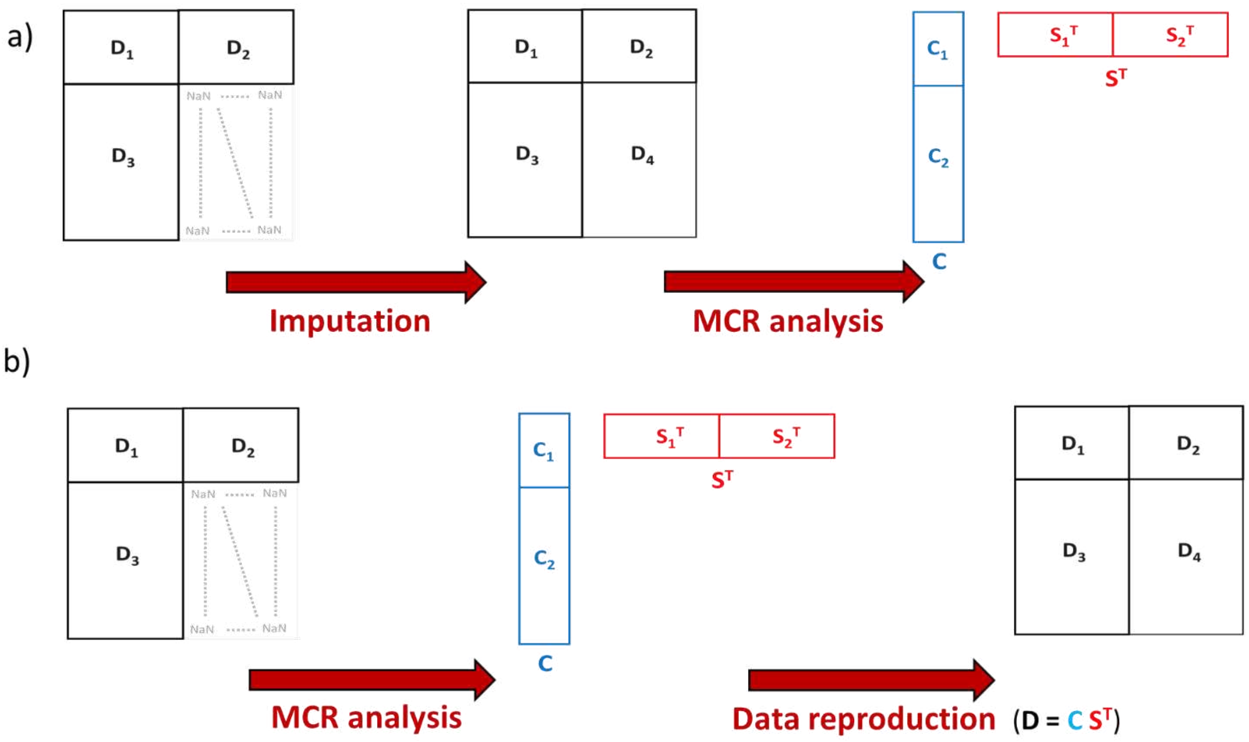


Figure 2. Different approaches to perform MCR analysis of incomplete multisets. a) Imputation and MCR analysis on the complete multiset structure. b) Retrieval of the bilinear model based on the available incomplete multiset (real information) and subsequent reconstruction of the complete multiset. Missing values are represented with NaN (meaning Not a Number) in the gray block of missing information.

The approach presented in Figure 2a has the advantage of getting a complete multiset structure onto which the MCR analysis can proceed with a single factorization model, as described in equations 2 and 3. The critical step is the first one, oriented to estimate the missing information in the data set, which is based on the use of imputation methods. Imputation can result from the simple estimation of missing values by using averages (or median values) of neighboring available information or from the use of more advanced PCA-based approaches [18-21]. Imputation methods work very efficiently when the missing information is randomly scattered in the data set and the percentage of missing information is low to moderate. However, the performance of this methodology is more compromised when the amount of missing information increases and, particularly, when the pattern of missing information is very systematic, as it happens in the missing block pattern of incomplete multisets. In these instances, the convergence of the imputation method becomes extremely slow and some biases in the estimated values can occur that may affect significantly the results of the subsequent MCR analysis [20,21].

The general approach proposed in Figure 2b offers the great advantage of skipping the imputation step, since the bilinear model is recovered using only the original available information. However, equations 2 and 3 cannot be straightforwardly applied because the multiset, i.e., the matrix **D**, is not complete. The first methodology proposed to perform the direct analysis of incomplete multisets exploited the idea that a multiset with missing blocks of information can be seen as a set of several smaller complete multiset structures connected among them (see Figure 3) [1,14-16]. Bearing this concept in mind, the final bilinear model is obtained through the smart combination of multiple factorizations issued from the analysis of the complete submultisets forming the original incomplete structure. The multiple factorization concept has also been used in other contexts, such as in Combined Tensor and Matrix Factorization (CMTF) methods, another example where the pieces of information to be connected, i.e., matrices and tensors, do not have an equivalent data configuration [22,23].

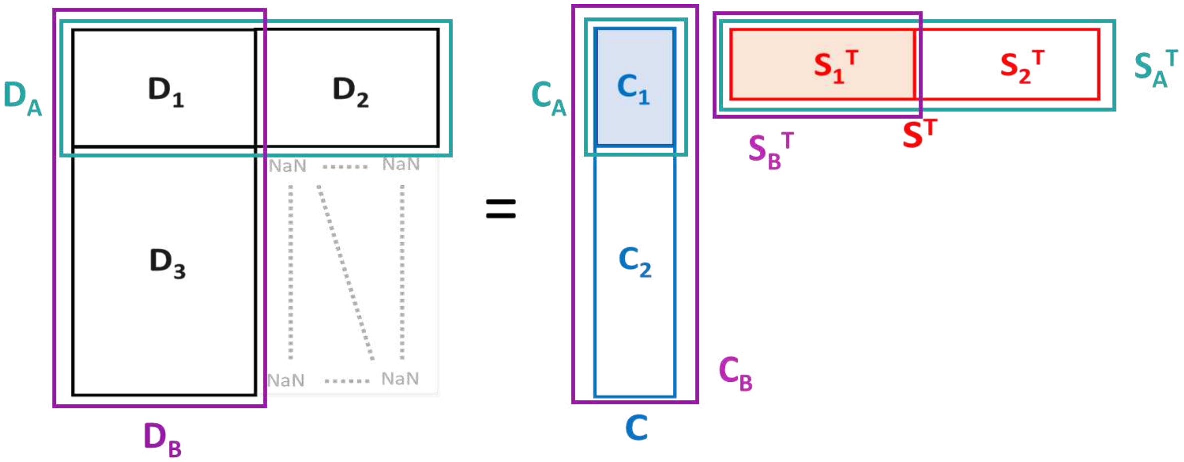


Figure 3. MCR analysis of an incomplete multiset using a multiple factorization approach.

Following the multiple factorization concept, the MCR analysis of the incomplete multiset in Figure 3 would take into consideration that the complete multisets $\mathbf{D_A}$ and $\mathbf{D_B}$ define all the available information in the initial incomplete structure. Thus, in every iterative cycle, the factorizations of $\mathbf{D_A}$ and $\mathbf{D_B}$ would be performed and the error function to be minimized would incorporate the terms associated with each of the factorizations carried out. Equations 4 and 5 for the $\mathbf{D_A}$ factorization and equations 6 and 7 for the $\mathbf{D_B}$ factorization mimic formally equations 2 and 3. Equation 8 expresses the error function used in the analysis. As can be seen in Figure 3, all the information of the full bilinear model $\mathbf{CS}^T$ can be obtained by the combination of the bilinear models issued from the $\mathbf{D_A}$ and $\mathbf{D_B}$ factorizations.

$$\mathbf{C_A} = \mathbf{D_A}\,\mathbf{S_A}\left(\mathbf{S_A^T}\,\mathbf{S_A}\right)^{-1} \qquad (4)$$

$$\mathbf{S_A^T} = \left(\mathbf{C_A^T}\,\mathbf{C_A}\right)^{-1}\mathbf{C_A^T}\,\mathbf{D_A} \qquad (5)$$

$$\mathbf{C_B} = \mathbf{D_B}\,\mathbf{S_B}\left(\mathbf{S_B^T}\,\mathbf{S_B}\right)^{-1} \qquad (6)$$

$$\mathbf{S_B^T} = \left(\mathbf{C_B^T}\,\mathbf{C_B}\right)^{-1}\mathbf{C_B^T}\,\mathbf{D_B} \qquad (7)$$

$$\min\left(\left\|\mathbf{D_A} - \mathbf{C_A}\,\mathbf{S_A^T}\right\| + \left\|\mathbf{D_B} - \mathbf{C_B}\,\mathbf{S_B^T}\right\|\right) \qquad (8)$$

The multiple factorization approach for MCR analysis of incomplete multisets has several advantages, such as: a) the bilinear model obtained is based solely on real information and b) since MCR analysis is carried out directly on the incomplete structure, biases stemming from the imputation step are avoided. Although this methodology has been successfully applied in different examples, some limitations associated with the difficulty of reconciling the bilinear models of several factorizations need to be pointed out. A first problem appears when some blocks of the $\mathbf{C}$ and $\mathbf{S}^T$ matrices are obtained with more than one factorization model, e.g., shaded blocks $\mathbf{C_1}$ and $\mathbf{S_1^T}$ in Figure 3 are calculated with $\mathbf{D_A}$ and $\mathbf{D_B}$ factorizations. In this case, there are different ways to obtain the profiles in the final bilinear model, either as a mean of the different factorization results [14,15] or favoring one factorization solution over the rest because it came from more complete information [16]. The option adopted is data-dependent and the decision is not trivial. The combination of factorizations becomes also more complex when the number of required factorizations to cover the incomplete multiset increases. The last difficulty concerns the need to have previous information about the number and identity of components in the different submultisets factorized for a proper combination of results.

Looking at the two approaches described and displayed in Figure 2, the ideal option would be having a methodology able to work with a single factorization model based only on the real available information. The new methodology proposed meets these two requirements, as will be described in the next section.

The missing block problem solved with a single factorization model.

Understanding the single factorization approach proposed to deal with the missing block problem in MCR requires going back to the least-squares fundamentals. Equations 2 and 3, used to analyze complete multisets, work with the full **D**, **C** and **S**^{T} matrices, but equivalent expressions could be formulated if the **C** and **S**^{T} matrices were recovered row-by-row and column-by column, respectively. Thus, every *i* row of **C**, **c**(*i*,:), could be recovered using the related row of **D**, **d**(*i*,:), and **S**^{T}, as shown in equation 9. Analogously, every *j* column of **S**^{T}, **s**(:,*j*), could be recovered using the related column of **D**, **d**(:,*j*), and **C**, as in equation 10. Obviously, the error function is min$\left\|\mathbf{D} - \mathbf{CS}^T\right\|$ since a single bilinear model is calculated (to calculate the residuals, the elements with missing information in **D** are not taken into consideration).

$$\mathbf{c}(i,:) = \mathbf{d}(i,:)\ \mathbf{S}\left(\mathbf{S}^T\mathbf{S}\right)^{-1} \qquad (9)$$

$$\mathbf{s}(:,\mathrm{j})^T = \left(\mathbf{C}^T\mathbf{C}\right)^{-1}\mathbf{C}^T\mathbf{d}(:,\mathrm{j}) \qquad (10)$$

This stepwise definition of the least-squares steps required to calculate the MCR bilinear model has clear advantages for data sets that have rows and columns with different structures, i.e., some with full complete information and some with missing values. Setting the problem in this manner, equations 9 and 10 simply adapt to work with the available elements in every row or column of **D** and the matching counterparts in **S**^{T} and **C** to recover the full row or column of the **C** and **S**^{T} matrix, respectively.

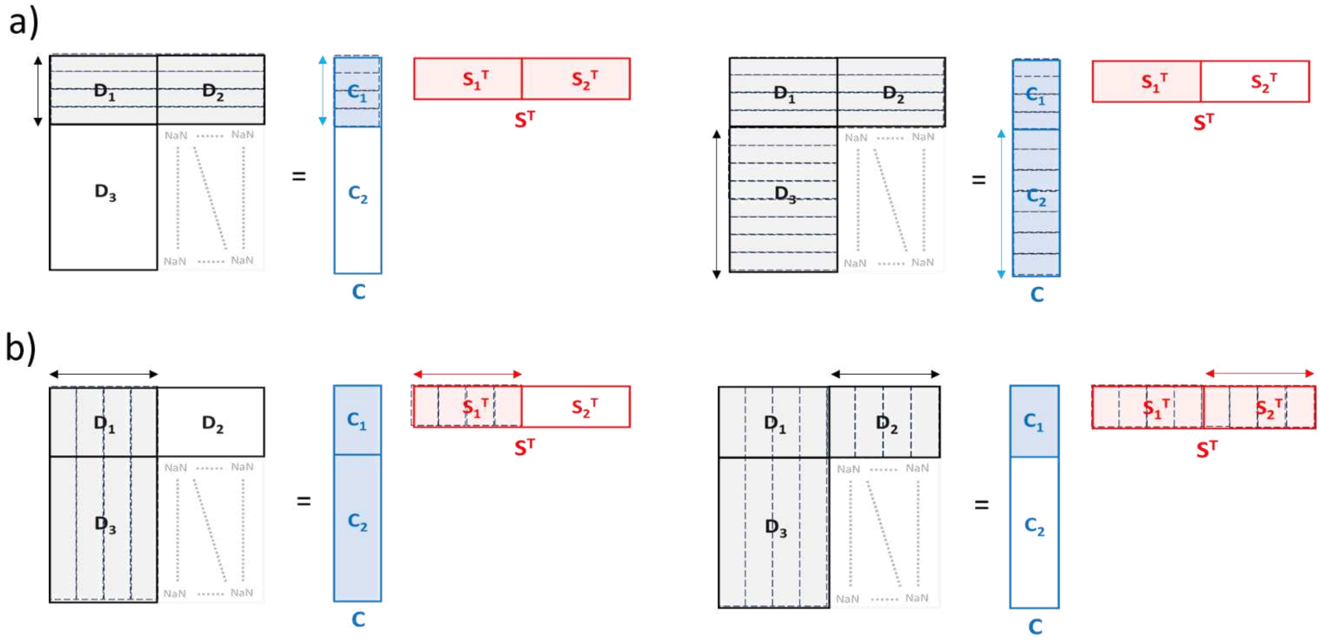


Figure 4. MCR analysis of an incomplete multiset with a single factorization model. a) **C** and b) $\mathbf{S}^T$ calculation.

Figure 4 clearly displays how this single factorization approach would work to analyze the same incomplete multiset shown in Figure 3. Figure 4a represents the row-by-row least-squares operations required to calculate the full **C** matrix, formed by the $\mathbf{C_1}$ and $\mathbf{C_2}$ blocks. All rows in the $\mathbf{C_1}$ block associate with rows in **D** that do not have any missing value and, hence, they could be calculated using the complete $\mathbf{S}^T$ matrix, as in equation 9 (left plot in figure 4a). Conversely, rows in $\mathbf{C_2}$ relate to rows of **D** where only the information of the $\mathbf{D_3}$ block is present, which is associated with the spectral information of the $\mathbf{S_1}^\mathbf{T}$ block. Hence, the equation required to retrieve the full rows of $\mathbf{C_2}$ would only use the matching information of rows of **D** and $\mathbf{S_1}^\mathbf{T}$, as in equation 11 (right plot in figure 4a).

$$\mathbf{c}(i,:) = \mathbf{d}(i,:)\ \mathbf{S_1}\left(\mathbf{S}_1^T\mathbf{S}_1\right)^{-1} \qquad (11)$$

Analogously, Figure 4b represents the column-by-column least-squares operations required to calculate the full $\mathbf{S}^\mathbf{T}$ matrix, formed by $\mathbf{S_1}^\mathbf{T}$ and $\mathbf{S_2}^\mathbf{T}$ blocks. In this case, the columns of the block $\mathbf{S_1}^\mathbf{T}$ would be calculated as in equation 10 because the related columns of **D** have no missing information (left plot in figure 4b). Instead, the calculation of the full columns of block $\mathbf{S_2}^T$ would only use the available column information in block $\mathbf{D_2}$ and the matching information of block $\mathbf{C_1}$, as expressed in equation 12 (right plot in Figure 4b).

$$\mathbf{s}(:,j)^T = \left(\mathbf{C}_\mathbf{1}^T\mathbf{C_1}\right)^{-1}\mathbf{C}_\mathbf{1}^T\mathbf{d}(:,j) \qquad (12)$$

The important fact about this new methodology to analyze incomplete multisets is that no matter the amount and structure of the missing information in **D**, the adapted row-by-row and column-by-column least-squares calculation of **C** and $\mathbf{S}^\mathbf{T}$ will always provide a single and complete bilinear model. This stepwise use of the least-squares calculation was first proposed in 2013 by Maeder et al. to deal with the presence of scattered missing values linked to saturated signals in regression problems [24], but was never envisioned to be used in the context of MCR models with full blocks of missing information. In the MCR context, this methodology to deal with incomplete multisets not only avoids imputation and the complexity related to multiple factorizations, but provides a generalized framework to analyze incomplete multisets with any kind of missing block configuration.

To complement the theoretical description of the missing block problem in MCR analysis of incomplete multisets, the new approach is illustrated in a paradigmatic context of missing block data sets, the image fusion scenario. However, the formulation proposed would work for any other application with incomplete multiset structures.

Testing the MCR single factorization approach. The image fusion paradigm.

Hyperspectral images (HSI) are analytical measurements that combine spatial and spectral information. They are usually displayed as data cubes with two spatial dimensions (pixel coordinates) and a spectral dimension. The analysis of an HSI by MCR implies unfolding the image cube into a spectroscopic data matrix [25]. The bilinear model provides pure spectra of the image constituents and the related concentration profiles, which can be refolded into concentration maps (see Figure 5a). Combining measurements that come from different imaging platforms has a lot of value because of the complementary perspective that offer the different spectroscopic techniques. This may be achieved with a multiset, where the pixel mode is common among platforms and the spectroscopic responses per pixel are concatenated (see Figure 5b) [7,16,25-27].

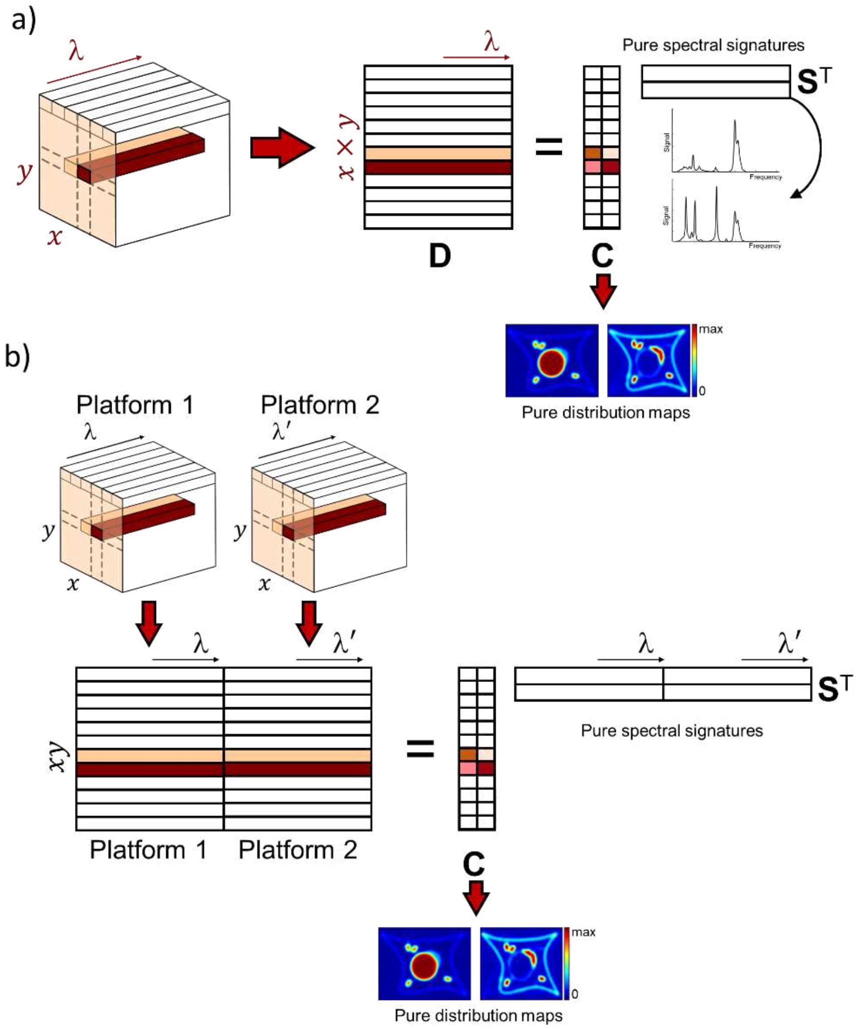


Figure 5. a) MCR model for a hyperspectral image. b) Multiset formed by image fusion of two platforms.

However, the requirement of having a common pixel mode for image fusion is not trivial. Other than the necessary preprocessing oriented to image coregistration, i.e., spatial transformations to shift/rotate images so that they spatially match [28,29], there may be many other differences among the information provided by imaging platforms. Notably, it is extremely common to obtain images that do not cover the same sample area. Additionally, differences of spatial resolution among imaging platforms often occur, i.e., a single pixel in a platform with low spatial resolution covers a sample area represented by several pixels in a platform with higher spatial resolution.

To illustrate the situations mentioned above, let us consider the easy simulated example in Figure 6, where the images obtained by two different platforms do not cover the same area. In this case, the fluorescence image covers all the sample area (see the yellow frame in Figure 6). On the other hand, the Raman image covers a smaller region of the sample (see the pink frame in figure 6). Although the total scanned area of both images is different, a common region, delimited by the pink frame, overlaps and has equivalent pixels measured with fluorescence and Raman. Thus, a multiset augmented in the row direction, as in Figure 5b, can be built only for those pixels present in both fluorescence and Raman images, leading to the multiset [$\mathbf{D_1}$,$\mathbf{D_2}$]. However, the fluorescence pixels of the non-common scanned area can still be used for the analysis by the concatenation of $\mathbf{D_1}$ to $\mathbf{D_3}$, building the incomplete multiset [$\mathbf{D_1}$,$\mathbf{D_2}$;$\mathbf{D_3}$,**NaN**] (Fig. 6, right).

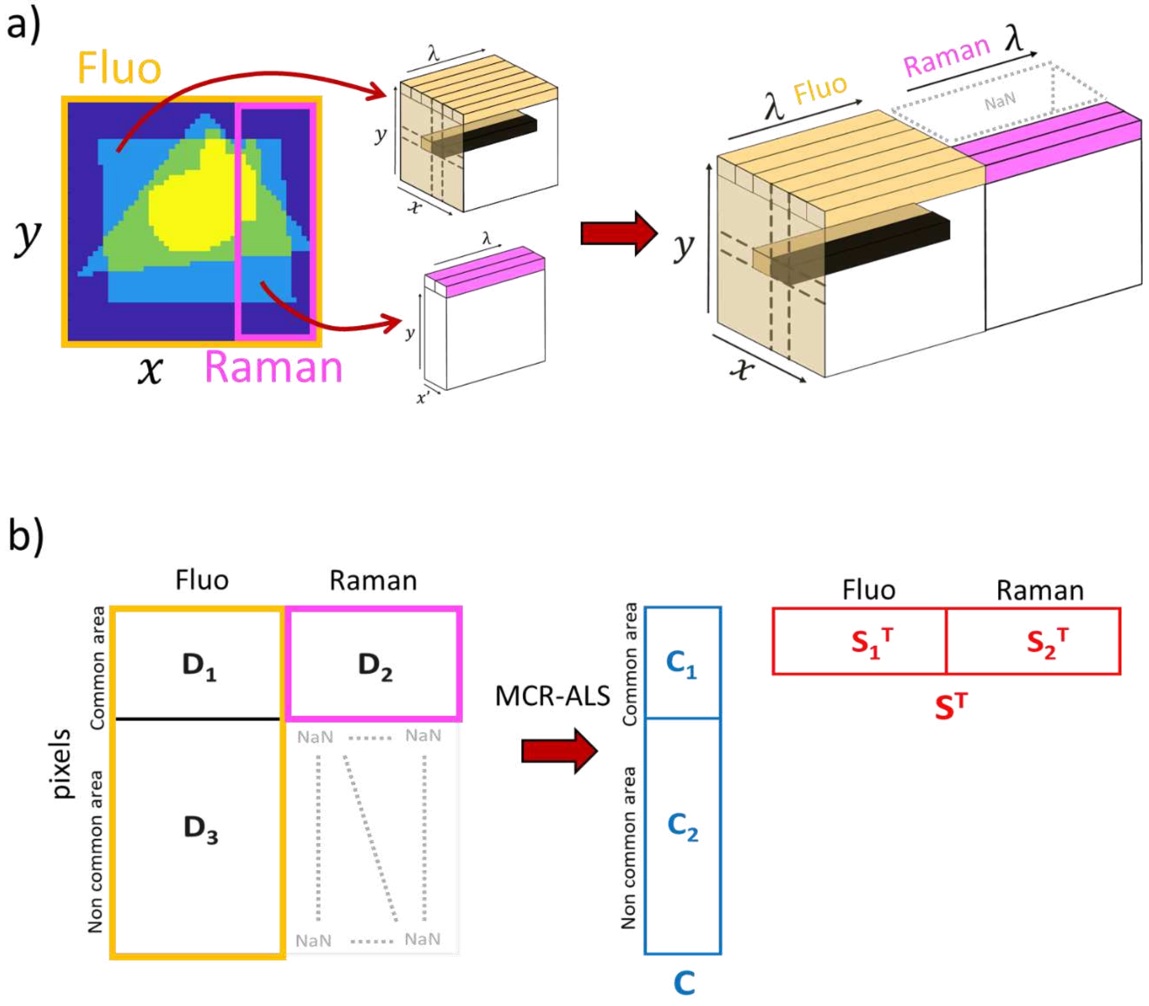


Fig. 6. a) Simulated sample with three components and scanned areas for fluorescence (yellow) and Raman (pink) images. The fluorescence and Raman hyperspectral images are concatenated. Since not all fluorescence spectra

have the corresponding Raman spectra, a part of the data is missing. b) The incomplete multiset and the related bilinear model connected with the images shown in a).

The performance of the new MCR-ALS algorithm proposed to deal with incomplete multisets by a single factorization model is tested in simulated cases that correspond to Figure 6. As seen in the figure, the images contain three different components. For the sake of simplicity, there is no difference in pixel size between the Raman and fluorescence images and no coregistration actions are required. The fluorescence image has a size of 50 x 50 pixels, whereas the Raman image is sized 20 x 50 pixels. The pure Raman profiles span the spectral range from 1000 to 1600 cm^{-1}, represented by 60 spectral channels. The pure fluorescence profiles cover the wavelength range from 400 to 600 nm, with a total of 60 channels. To mimic real conditions, Poisson noise was added to the spectra in matrix **D** (around 6-7% of the total signal). The incomplete multiset formed by fusing the fluorescence and Raman images as in Figure 6 contains a 29% of missing entries. Figure 7 shows the distribution maps and the related Raman and fluorescence spectra for the three components used to produce the different simulated case studies.

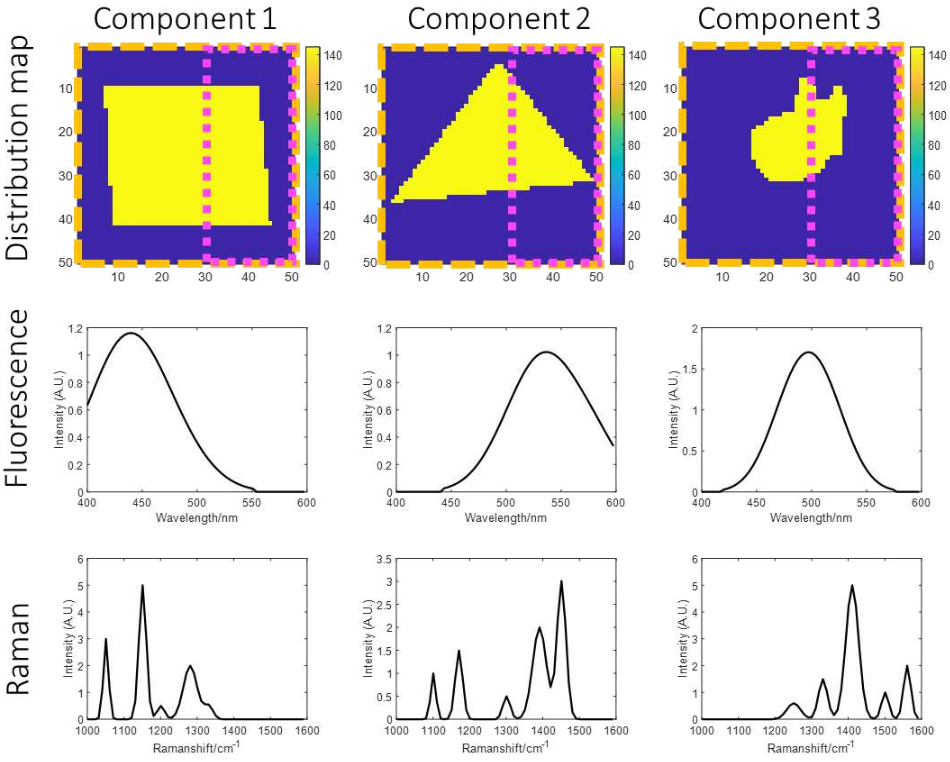


Fig. 7. Data set simulated with three components to study the MCR-ALS analysis of incomplete multisets. Top) Pure distribution maps. In yellow dashed line, scanned area of fluorescence image; in pink dashed line, scanned area of Raman. Middle) Pure fluorescence signatures. Bottom) Pure Raman signatures.

Based on this simulated data set, three realistic scenarios have been considered: case 1 assumes that all components have signal on all techniques and case 2 and 3 represent different scenarios where one component has a null spectroscopic signal (as detailed in Table 1).

Table 1. Information about the contributions of individual components (X marks presence) in the simulated data sets. Case 1 assumes all components are present in both images. Cases 2 and 3 assumes that component 3 does not show signal in Raman and fluorescence, respectively.

	Fluorescence hyperspectral image			**Raman hyperspectral image**		
Case	**Component 1**	**Component 2**	**Component 3**	**Component 1**	**Component 2**	**Component 3**
1	X	X	X	X	X	X
2	X	X	X	X		X
3	X	X		X	X	X

For all MCR-ALS analyses, initial estimates were obtained with a SIMPLISMA-based approach [30] using [$\mathbf{D_1}$, $\mathbf{D_2}$], selecting three components, and the convergence criterion was set at 10^{-10}%. Non-negativity was applied to both **C** and $\mathbf{S}^T$ matrices. In all MCR-ALS analyses, the lack of fit obtained matches the noise level added in the simulations.

Figure 8 shows the resolved pure maps and pure spectral profiles obtained for the three components for case 1. As it can be seen, a perfect match with the true simulated solution is obtained, with correlation coefficients > 0.99 between the MCR resolved profiles and the true solution. These results show the ability of MCR-ALS to deal with missing data without imputation and by a single factorization model using the row-by-row and column-by-column least squares calculation, despite the significant amount of missing entries.

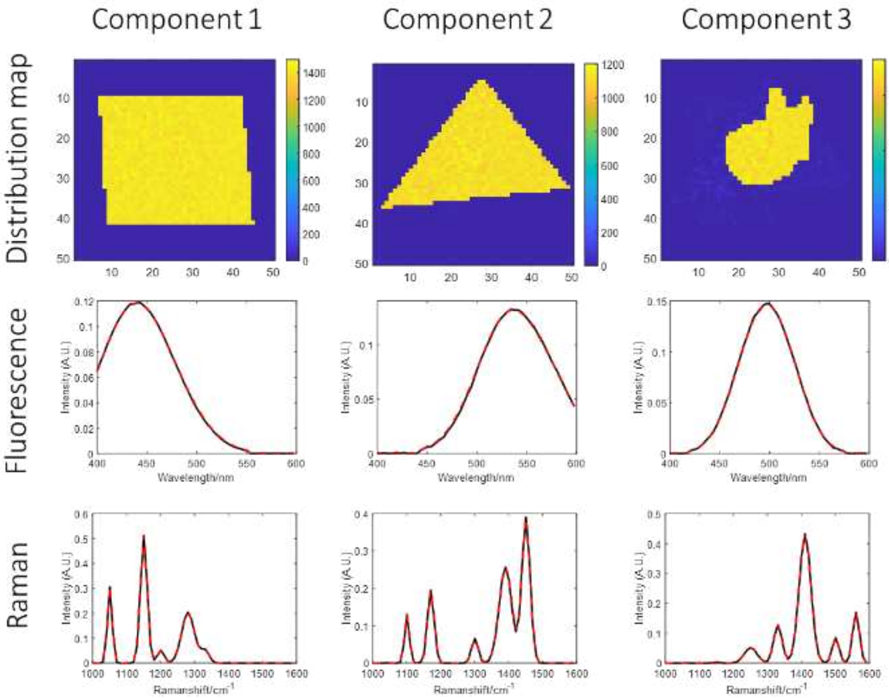


Fig. 8. MCR results on the simulated case 1. Top plots: pure distribution maps. Second row plots: fluorescence spectra (red dashed lines: simulated profile shape; black lines: MCR spectra. Third row plots: Raman spectra (red dashed lines: simulated profiles; black lines: MCR spectra).

Similar results can be observed for case 2 (Figure 9), with all resolved profiles perfectly matching with the true simulated solution (correlation coefficients > 0.99). It is important to note that the fact that component three has a null signal in Raman does not imply any problem, since all Raman pixel spectra are connected with a fluorescence signal in the multiset and, hence, the map for the component three is, nevertheless, well described. It is also relevant to mention that the algorithm is able to model the null Raman signal for the component three without setting any prior information about this fact.

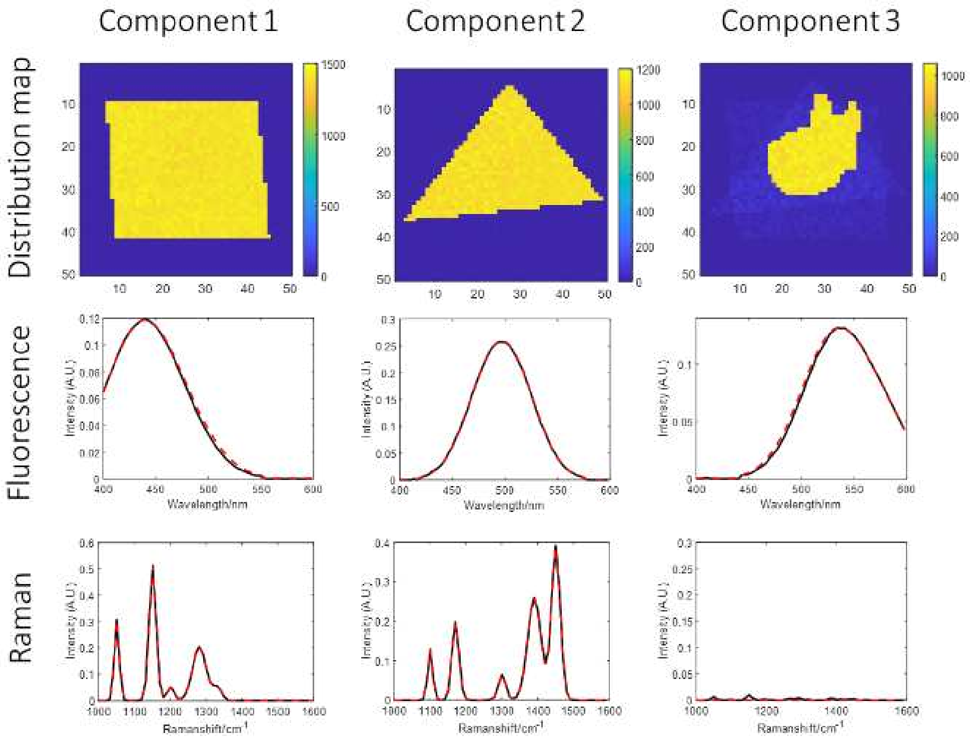


Fig. 9. MCR results on the simulated case 2. Top plots: pure distribution maps. Second row plots: fluorescence spectra (red dashed lines: simulated profile shape; black lines: MCR spectra. Third row plots: Raman spectra (red dashed lines: simulated profiles; black lines: MCR spectra).

Figure 10 shows the pure maps and pure spectral profiles resolved by MCR-ALS for case 3. The pure spectra of all compounds are almost perfectly resolved, with correlation coefficients > 0.99 and so are the maps for components 1 and 2, with signal for both fluorescence and Raman images. It is important to pay attention to the map resolved for component 3, to which is associated a null fluorescence signal. In such a case, the map region related to the common area scanned by both images is well recovered (this region would correspond to the block $\mathbf{C_1}$ in Figure 4a), since information about the component is provided by the Raman signal. We refer to the top left plot in Figure 4a to understand that the full $\mathbf{S}^T$ matrix information is used to calculate $\mathbf{C_1}$. However, the map region related to the area only scanned by the fluorescence image (related to $\mathbf{C_2}$ in Figure 4a) cannot be recovered since there is no signal available for this component (see top right plot in Figure 4a to understand that only the $\mathbf{S_1}^T$ submatrix, related to fluorescence, is used to calculate $\mathbf{C_2}$). This limitation is intrinsically related to the amount of information in the initial data set. No algorithm would successfully retrieve the information for a component with null signal unless this signal is connected to another technique that compensates this lack of information.

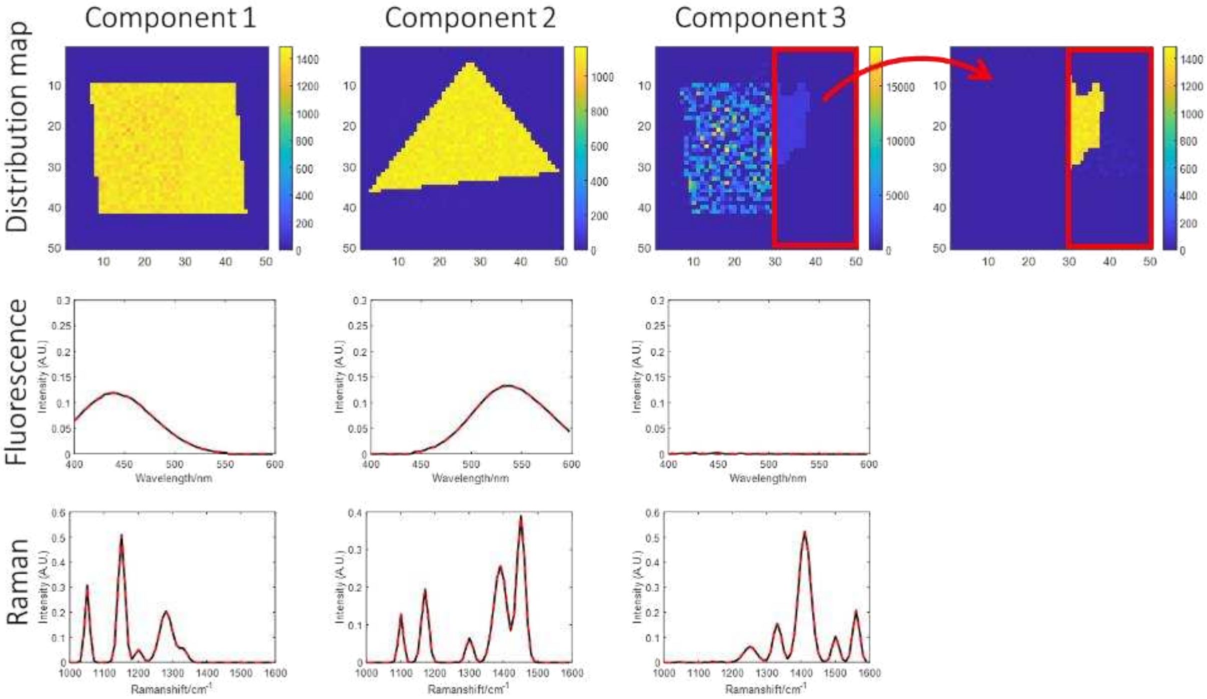


Fig. 10. MCR-ALS results on the simulated data set of case 3. Top plots: pure distribution maps. For component 3, the part without the undefined part is also show for a clear illustration of the resolved part of this component. Second row plots: fluorescence spectra (red dashed lines: simulated profile shape; black lines: MCR spectra. Third row plots: Raman spectra (red dashed lines: simulated profiles; black lines: MCR spectra).

Real image fusion scenarios tend to be more complex than the simulated example in Figure 6 because, additionally to differences in scanned area, there is often a difference in the spatial resolution (pixel size) of the images provided by the different platforms to merge in a multiset. Obtaining a complete multiset in this context implies cropping the images to finally use only the overlapped sample area by the different platforms and downsampling the highest spatial resolution images by binning the pixels to achieve the pixel size of the least detailed image. In doing so, a large part of available information is lost and the capacity to distinguish components decreases, since the binning operation mixes the information associated with neighbouring pixels and the selectivity of the initial high spatial resolution image decreases. Instead, all available information can be incorporated and actively used in an incomplete multiset, where differences among image platforms (specific scanned areas by a single technique and pixel spectra with the initial highest resolution) can be included as additional blocks of information with a missing counterpart block (see Figure 11).

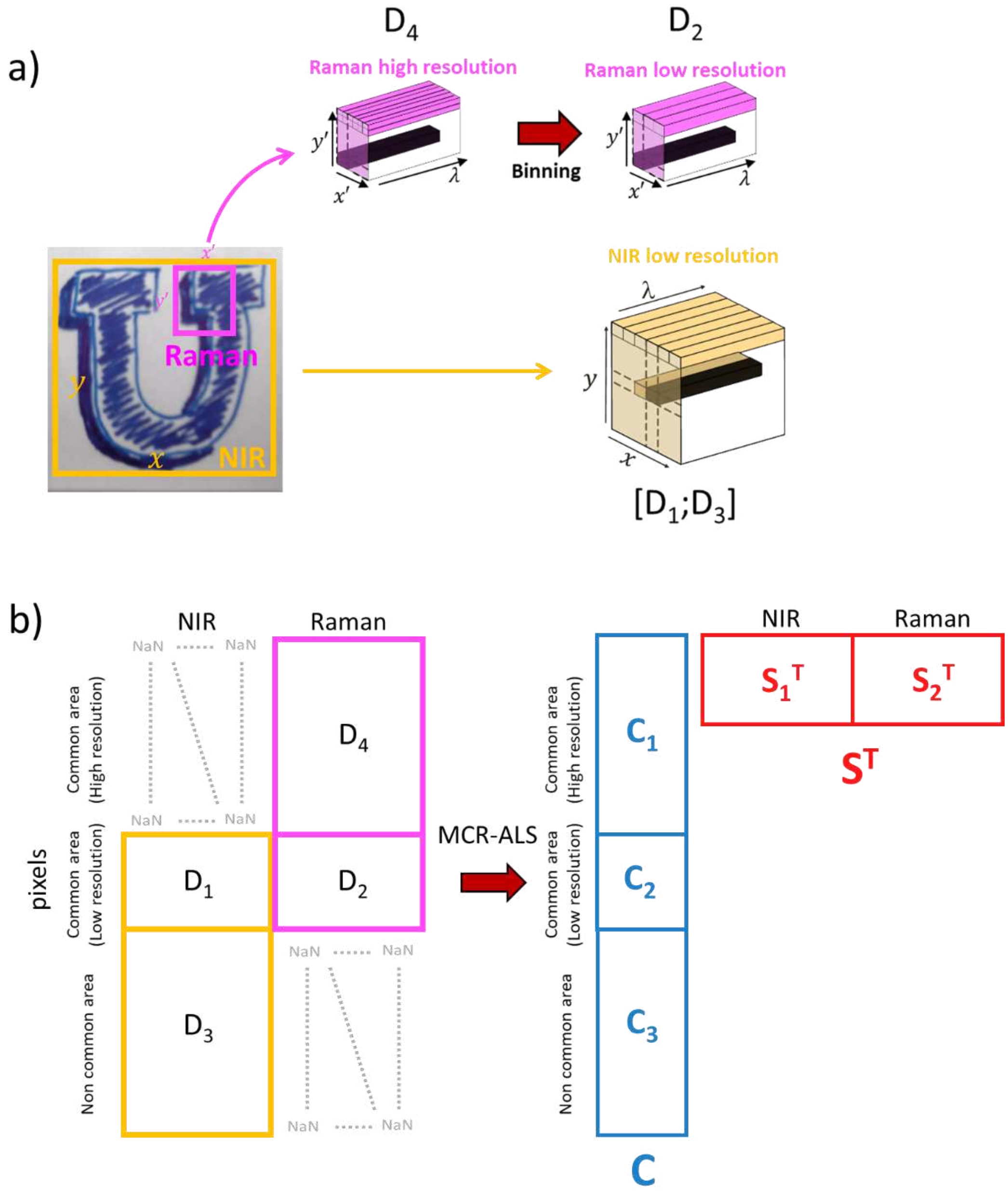


Fig. **11.** Real example of the construction of an incomplete multiset where different scanned areas and spatial resolution exist among imaging platforms. a) On the left, a NIR image of a sample formed by different inks is acquired, covering the full sample area (yellow), while the Raman measurement only considers a small region (pink). b) Incomplete multiset structure and related bilinear model for the NIR/Raman image fusion. The Raman image ($\mathbf{D}_4$) is binned ($\mathbf{D}_2$) to match the pixel size of the common area scanned by the NIR image ($\mathbf{D}_1$). The NIR pixels of the non-common scanned area (**D_3**) are concatenated below **D_1**, while the Raman image with the original spatial resolution is concatenated above **D_2**.

Figure 11 shows the real sample where three blue pens (Pilot V-Ball Grip, *PVG,* Uniball Signo, *US*, and Bic Velocity, *BV*) are used to write a letter “U” on a paper. The different inks are located inside the letter, depicting the contour and forming the shade, respectively. The sample is recorded by NIR and Raman images. In this case, the whole “U” has been scanned with NIR, providing an image with 300 microns of pixel size, while a smaller part of the letter has been scanned with Raman due to the long acquisition time associated with this imaging technique, providing an image with a pixel size of 100 microns.

For all MCR analyses, NIR and Raman spectra were preprocessed using the first derivative, calculated by the Savitzky-Golay method. Initial estimates were obtained with a SIMPLISMA-based approach. The convergence criterion was set at 10^{-10}% and non-negativity was applied to the **C** matrix.

Before discussing the benefit of image fusion, we report the MCR-ALS results provided by analyzing each image platform separately.

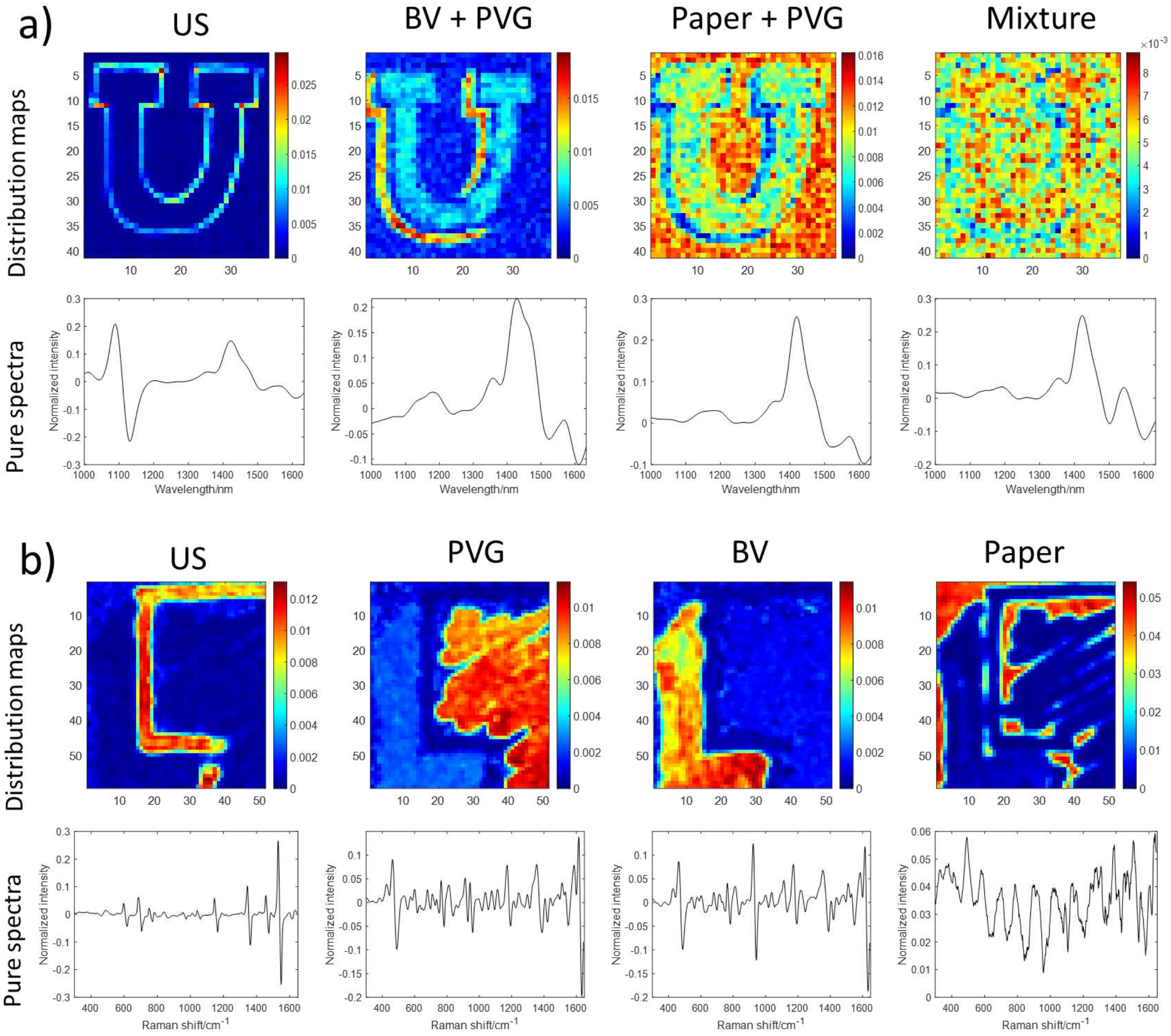


Fig. **12.** MCR results for the a) NIR image and b) Raman image of inks. Top plots: pure distribution maps. Bottom plots: resolved pure spectra.

Figure 12a shows the pure distribution maps and pure spectral profiles of the MCR analysis for the NIR dataset. Four components were detected and the lack of fit of the model was 5.4%, satisfactory for NIR hyperspectral images. As it can be observed in the pure resolved distribution map, the pure component related to the US ink is well recovered. However, the three other components are completely mixed. Since NIR pure spectral signatures are not very selective, the rotational ambiguity in this dataset is high and, thus, the pure profiles are mixed. In addition, the strong signal of the paper and its similar spectrum to the BV and PVG signal makes even more difficult the analysis. On the other hand, the MCR results shown in Figure 12b for the Raman image dataset are excellent. The lack of fit of the model was 10.5%

and four components were detected. The three different inks US, PVG and BV can be clearly located on the distribution maps and match the original ink distribution, as well as the area where the paper signal has a significant contribution. In addition, the spectral signatures extracted are in good agreement with the bands of the pure inks. The drawback is that only a small area of the sample is observed.

The incomplete multiset structured as in Figure 11 contains a 40% of missing entries, but allows the analysis of all available data in the original images without losing any information. Figure 13 shows the distribution maps of the full scanned area (top plots) and the resolved spectral signatures (bottom plots) of the MCR-ALS analysis of the incomplete multiset. The lack of fit of the model was 9.5%. As can be observed, the complementary information of Raman and NIR affects positively the results obtained for the full scanned sample area.

The pure spectral signatures represent the connected NIR/Raman pure spectra for the different components. The complementary information of both techniques and the capability of the Raman spectroscopy to differentiate the components translates into the extraction of the correct full NIR/Raman spectral signatures for the pure inks and of the distribution maps associated in the full scanned area. Thus, the top plots in Figure 13a, which represent the information of the full scanned sample area defined by blocks $\mathbf{C_2}$ and $\mathbf{C_3}$ in Figure 11, display correctly the regions of presence of the three inks and the paper. Note that the good definition of the maps is not only associated with the small area scanned by the Raman image, but to all the scanned area by NIR, due to the use of the fused NIR/Raman information.

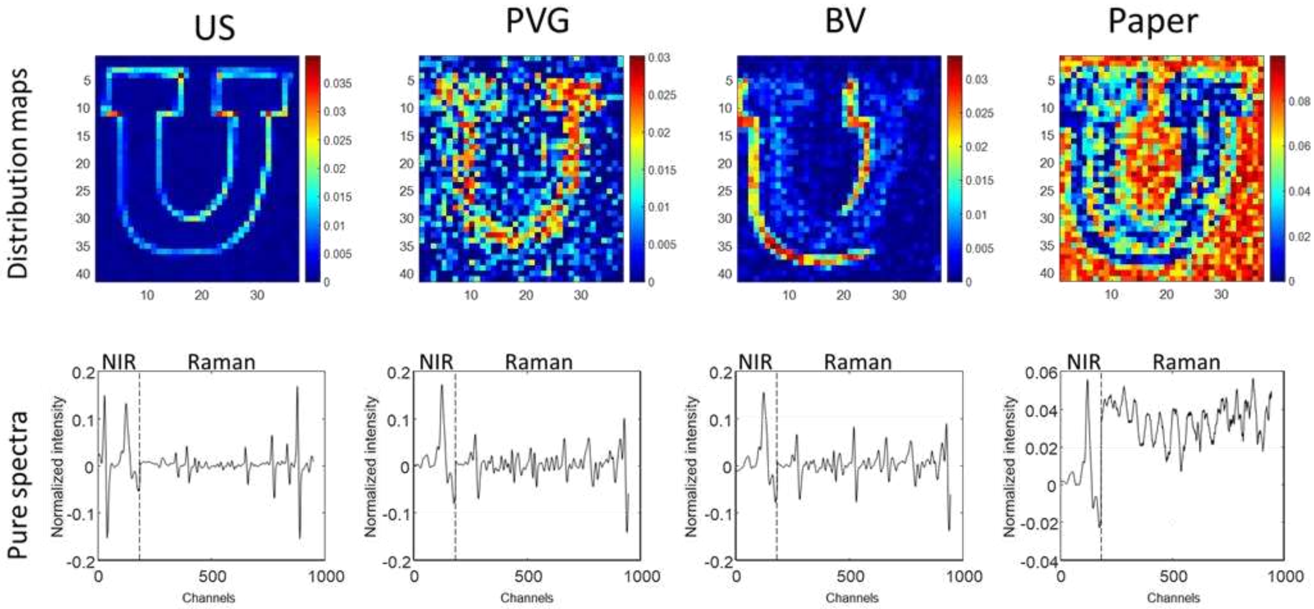


Fig. **13.** MCR results for the incomplete multiset of inks. Top plots: pure distribution maps. Bottom plots: resolved pure spectra signatures (fused NIR and Raman).

Note that the proposed MCR methodology based on least squares row-by-row and column-by-column calculation of **C** and $\mathbf{S}^T$ matrices provides excellent results and works in a very simple manner for the complex incomplete structure in Figure 11. Using the former multiple factorization approach, three different factorizations should have been combined, a difficult task to reconcile the resolved profiles obtained by all bilinear submodels.

Conclusions

Addressing the presence of missing blocks of information in incomplete multisets is a hard task for Multivariate Curve Resolution analysis due to the high percentage of missing entries and the non-random pattern of the absent data. The evolution of approaches based on the combination of multiple factorizations towards the proposed single factorization approach provides a simple and generalized framework adapted to any data configuration with missing information.

The new methodology does not require imputation approaches and can handle large amounts of missing data as long as the present information describes adequately the components in some of the blocks of the incomplete data structure. The gradual row-by-row and column-by-column least- squares calculation of **C** and $\mathbf{S}^T$ matrices based on the use of the available information in the initial data set and the adapted information of the counterpart matrix of the bilinear model is a simple concept that can be extended to many other algorithms involving bilinear or multilinear decompositions on data sets with missing information.

Acknowledgements

A. G.-S. and A.J. acknowledge financial support from the Spanish government Project PID 2019-1071586 B-IOO and the Catalan government (2021 SGR 00449). A. G.-S. acknowledges scholarships from the MOBLILEX U Lille program, from the Santander bank and from Fundació Montcelimar. R. T. acknowledges the financial support from Spanish Ministry of Science and Innovation for project PID 2019-105732 GB-C21.

Author contributions

Adrián Gómez-Sánchez: Conceptualization, Methodology, Data acquisition, Software, Formal analysis, Data curation, Investigation, Discussion, Writing – original draft, Visualization.

Cyril Ruckebusch: Discussion, Writing – review & editing, Supervision, Funding acquisition.

Romà Tauler: Conceptualization, Software, Investigation, Discussion, Writing – review & editing, Discussion.

Anna de Juan: Conceptualization, Investigation, Discussion, Writing – original draft, Visualization. Writing – original draft, Visualization, Discussion, Supervision, Project administration, Funding acquisition.

Declaration of interests

The authors declare no competing interests.

References

[1] A. de Juan, R. Tauler. Multivariate Curve Resolution: 50 years addressing the mixture analysis problem–A review. Anal. Chim. Acta 1145 (2021) 59-78. https://doi.org/10.1016/j.aca.2020.10.051

[2] A. de Juan, S.C. Rutan, R. Tauler. Two-Way Data Analysis: Multivariate Curve Resolution–Iterative Resolution Methods. In: S. Brown, R. Tauler, and R. Walczak (Eds.), Comprehensive Chemometrics, Elsevier: Oxford, 2020, Vol. 2, 153-171. https://doi.org/10.1016/B978-0-12-409547-2.14752-3

[3] Z. Zhang, P. Ma, H. Lu. Two-Way Data Analysis: Multivariate Curve Resolution: noniterative Resolution Methods. In: S. Brown, R. Tauler, and R. Walczak (Eds.), Comprehensive Chemometrics, Elsevier: Oxford, 2020, Vol. 2, 137-152. https://doi.org/10.1016/B978-0-12-409547-2.14875-9

[4] R. Tauler. Multivariate curve resolution applied to second order data. Chemom. Intell. Lab. Sys. 30 (1995) 133-146. https://doi.org/10.1016/0169-7439(95)00047-X

[5] R. Tauler, M. Maeder, A. de Juan. Multiset data analysis: Extended Multivariate Curve Resolution. In: S. Brown, R. Tauler, and R. Walczak (Eds.), Comprehensive Chemometrics, Elsevier: Oxford, 2020, Vol. 2, 305-336. https://doi.org/10.1016/B978-0-12-409547-2.14702-X

[6] B. Debus, M. Orio, J. Rehault, G. Burdzinski, C. Ruckebusch, M. Sliwa. Fusion of Ultraviolet–Visible and Infrared Transient Absorption Spectroscopy Data to Model Ultrafast Photoisomerization. J. Phys. Chem. Lett. 8(15) (2017) 3530-3535. https://doi.org/10.1021/acs.jpclett.7b0125

[7] A. de Juan, A. Gowen, L. Duponchel, C. Ruckebusch. Image Fusion. In: M. Cocchi (Ed.), Data Handling in Science and Technology, Elsevier: Volume 31, 2019, Pages 311-344. https://doi.org/10.1016/B978-0-444-63984-4.00011-9

[8] A. de Juan, R. Tauler. Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution. Anal. Chim. Acta 500(1-2) (2003) 195-210. https://doi.org/10.1016/S0003-2670(03)00724-4

[9] C. Ruckebusch, A. De Juan, L. Duponchel, J. P. Huvenne. Matrix augmentation for breaking rank-deficiency: A case study. Chemom. Intell. Lab. Sys. 80(2) (2006) 209-214. https://doi.org/10.1016/j.chemolab.2005.06.009

[10] C. Ruckebusch, M. Sliwa, P. D. Pernot, A. De Juan, R. Tauler. Comprehensive data analysis of femtosecond transient absorption spectra: A review. J. Photochem. Photobiol. C: Photochem. Rev. 13(1) (2012) 1-27. https://doi.org/10.1016/j.jphotochemrev.2011.10.002

[11] R. Tauler. Interpretation of Environmental Data Using Chemometrics. In: D. Barceló (Ed.), Techniques and Instrumentation in Analytical Chemistry, Elsevier: Volume 21, 2000, Pages 689-736. https://doi.org/10.1016/S0167-9244(00)80022-0.

[12] S. Mas, A. de Juan, R. Tauler, A. C. Olivieri, G. M. Escandar. Application of chemometric methods to environmental analysis of organic pollutants: A review. Talanta 80(3) (2010) 1052-1067. https://doi.org/10.1016/j.talanta.2009.09.044

[13] R. Tauler, M. Viana, X. Querol, A. Alastuey, R. M. Flight, P. D. Wentzell, P. K. Hopke. Comparison of the results obtained by four receptor modelling methods in aerosol source apportionment studies. Atmos. Environ. 43(26) (2009) 3989-3997. https://doi.org/10.1016/j.atmosenv.2009.05.018

[14] M. Alier, R. Tauler. Multivariate curve resolution of incomplete data multisets. Chemom. Intell. Lab. Sys. 127 (2013) 17-28. https://doi.org/10.1016/j.chemolab.2013.05.006

[15] M. De Luca, G. Ragno, G. Ioele, R. Tauler. Multivariate curve resolution of incomplete fused multiset data from chromatographic and spectrophotometric analyses for drug photostability studies. Anal. Chim. Acta 837 (2014) 31-37. https://doi.org/10.1016/j.aca.2014.05.056

[16] S. Piqueras, C. Bedia, C. Beleites, C. Krafft, J. Popp, M. Maeder, R. Tauler, A. de Juan. Handling different spatial resolutions in image fusion by multivariate curve resolution-alternating least squares for incomplete image multisets. Anal. Chem. 90 (2018) 6757-6765. https://doi.org/10.1021/acs.analchem.8b00630

[17] C. Bedia, A. Sierra, R. Tauler. Multimodal multisample spectroscopic imaging analysis of tumor tissues using multivariate curve resolution. Chemom. Intell. Lab. Sys. 215 (2021) 104366.

[18] B. Walczak, D.L. Massart. Dealing with missing data: Part I. Chemom. Intell. Lab. Sys. 58.1 (2001) 15-27. https://doi.org/10.1016/S0169-7439(01)00131-9

[19] B. Grung, R. Manne. Missing values in principal component analysis. Chemom. Intell. Lab. Sys. 42 (1998) 125–139. https://doi.org/10.1016/S0169-7439(98)00031-8

[20] G. Tomasi, R. Bro. PARAFAC and missing values. Chemom. Intell. Lab. Sys. 75(2) (2005) 163-180. https://doi.org/10.1016/j.chemolab.2004.07.003

[21] A. Ilin, T. Raiko. Practical approaches to principal component analysis in the presence of missing values. J. Mach. Learn. Res. 11 (2010) 1957-2000. https://dl.acm.org/doi/10.5555/1756006.1859917

[22] E. Acar, M.A. Rasmussen, F. Savorani, T. Næs, R. Bro. Understanding data fusion within the framework of coupled matrix and tensor factorizations. Chemom. Intell. Lab. Sys. 129 (2013) 53-63. https://doi.org/10.1016/j.chemolab.2013.06.006

[23]. K. Huang, N. D. Sidiropoulos, A. P. Liavas. A flexible and efficient algorithmic framework for constrained matrix and tensor factorization. IEEE Trans. Signal Process. 64(19) (2016) 5052-5065. https://doi.org/10.1109/TSP.2016.2576427

[24] Y. Beyad, M. Maeder. Multivariate linear regression with missing values. Anal. Chim. Acta 796 (2013) 38-41. https://doi.org/10.1016/j.aca.2013.08.027

[25] A. de Juan. Multivariate curve resolution for hyperspectral image analysis. In: Data Handl. Sci. Tech. Vol. 32, Elsevier, 2020, pp. 115-150. http://doi.org/10.1016/B978-0-444-63977-6.00007-9

[26] A. Gómez-Sánchez, M. Marro, M. Marsal, P. Loza-Alvarez, A. de Juan. Linear unmixing protocol for hyperspectral image fusion analysis applied to a case study of vegetal tissues. Sci. Rep. 11(1) (2021) 18665. https://doi.org/10.1038/s41598-021-98000-0

[27] A. Gómez-Sánchez, M. Marro, M. Marsal, P. Loza-Alvarez, A. de Juan. 3D and 4D image fusion: coping with differences in spectroscopic modes among hyperspectral images. Anal. Chem. 92(14) (2020) 9591-9602. http://doi.org/10.1021/acs.analchem.0c00780.

[28] T.G. Schaaff, J.M. McMahon, P.J. Todd. Semiautomated analytical image correlation. Anal. Chem. 74.17 (2002) 4361-4369. https://doi.org/10.1021/ac025693b

[29] S. P. Solsona, M. Maeder, R. Tauler, A. de Juan. A new matching image preprocessing for image data fusion. Chemom. Intell. Lab. Sys. 164 (2017) 32-42. https://doi.org/10.1016/j.chemolab.2017.02.013

[30] W. Windig, J. Guilment. Interactive self-modeling mixture analysis. Anal. Chem. 63(14) (1991) 1425-1432. https://doi.org/10.1021/ac00014a016

Solving the missing value problem in PCA by Orthogonalized-Alternating Least Squares (O-ALS)

Adrián Gómez-Sánchez[1,2*], Raffaele Vitale[2], Cyril Ruckebusch[2], Anna de Juan[1*]

[1]Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028, Barcelona, Spain

[2]LASIRE - Laboratory of Advanced Spectroscopy, Interactions, Reactivity and Environment Université Lille, CNRS, UMR 8516, Cité Scientifique, Bâtiment C5, 59000, Lille, France

Corresponding authors: agomezsa29@alumnes.ub.edu, anna.dejuan@ub.edu

Abstract

Dealing with missing data poses a challenge in Principal Component Analysis (PCA) since the most common algorithms are not designed to handle them. Several approaches have been proposed to solve the missing value problem in PCA, such as Imputation based on SVD (I-SVD), where missing entries are filled by imputation and updated in every iteration until convergence of the PCA model, and the adaptation of the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm, able to work skipping the missing entries during the least-squares estimation of scores and loadings. However, some limitations have been reported for both approaches. On the one hand, convergence of the I-SVD algorithm can be very slow for datasets with a high percentage of missing data. On the other hand, the orthogonality properties among scores and loadings might be lost when using NIPALS.

To solve these issues and perform PCA of datasets with missing values without the need of imputation steps, a novel algorithm called Orthogonalized-Alternating Least Squares (O-ALS) is proposed. The O-ALS algorithm is an alternating least-squares algorithm that estimates the scores and loadings subject to the Gram-Schmidt orthogonalization constraint. The way to estimate scores and loadings is adapted to work only with the available information.

In this study, the performance of O-ALS is tested and compared with NIPALS and I-SVD in simulated data sets and in a real case study. The results show that O-ALS is a very accurate and fast algorithm to analyze data with any percentage and distribution pattern of missing entries, being able to provide correct scores and loadings in cases where I-SVD and NIPALS do not perform satisfactorily.

Keywords: Principal Component Analysis (PCA), missing values, NIPALS, Imputation, SVD, Orthogonalized Alternating Least Squares (O-ALS).

1. Introduction

Principal Component Analysis (PCA) [1-3] is one the most fundamental tools in chemometrics for data compression and visualization of complex data sets. It is widely employed as an exploratory tool for all kinds of data, such as hyperspectral imaging, [4,5], quality control [6], complex mixtures [7], as a basic tool for classification methods [8] or in PAT technologies [9,10], and for an endless number of different scenarios.

PCA is a bilinear factorization method that decomposes the data into the so-called principal components, obtained using orthogonality and normalization constraints (Eq. 1). The non-random variance of the initial data set can be described using a limited number of principal components, *N*, which allows for an easy visualization of the information. Each principal component is represented by a score and a loading vector, linked to the representation of samples (or row information) and variables (or column information) in the data matrix.

The PCA model is defined as in Eq. 1:

$$\begin{aligned} &\mathbf{D} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \\ \textit{subject to:}\quad &\begin{cases} \mathbf{T}^{\mathrm{T}}\mathbf{T} = \mathbf{W}, \mathbf{W}_{ij} = \mathbf{0}\, for\ i \neq j \\ \mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I} \end{cases} \end{aligned} \qquad \text{Eq. 1}$$

where **D** is the data matrix sized $R \times C$, **T** is the scores matrix sized $R \times N$, **P** is the loadings matrix sized $C \times N$, and **E** is the residual matrix sized $R \times C$. R, C and N stand for the number of rows and columns of the initial data set and the number of relevant principal components, respectively. **W** is a diagonal matrix, sized $N \times N$, containing the eigenvalues in its diagonal, while **I** is the identity matrix sized $N \times N$. The PCA solution is unique up to sign and permutation ambiguity when principal components have non-equal eigenvalues [11].

Data sets often have missing values due to measurement errors, incomplete data collection or because of the nature of the measurement [12-14]. The presence of missing values represents a significant challenge when applying PCA, since standard approaches, such as Singular Value Decomposition (SVD), cannot be straightforwardly applied.

The different patterns that missing values can adopt within a data sets can affect differently and significantly the performance of the algorithms used (Fig. 1) [15]. Thus, randomly distributed missing values (Fig. 1A) do not tend to affect much the analysis because the estimation of missing entries is, in this case, quite easy. Instead, systematic patterns of missing values (Fig. 1B), e.g., those found in excitation-emission matrices where no signal is recorded below the excitation wavelength range and scattering contributions have to be suppressed, provide a much more challenging scenario. Within the non-random patterns of missing entries, the missing block pattern is specifically associated with data fusion scenarios [16]. For instance, this pattern happens when blocks of several experiments monitored with different techniques are concatenated into a single structure (multiset) and the data coming from a particular technique may be missing for a specific experiment (Fig. 1C).

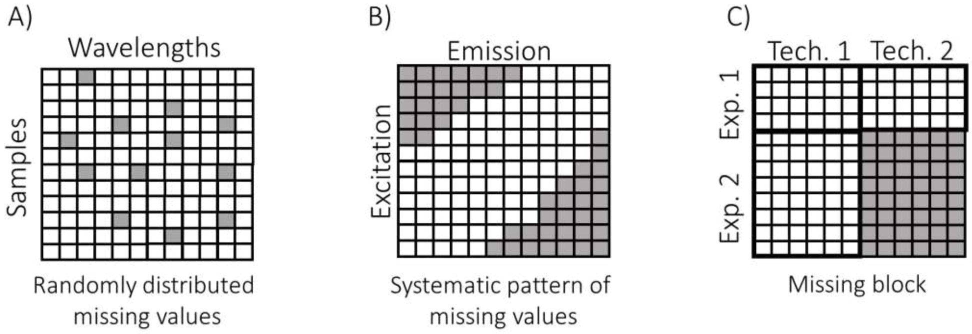


Figure 1. Usual patterns of missing data found in data sets. In white, available entries. In gray, missing entries. A) Random distribution of missing values. B) Systematic pattern of missing values. C) Missing block pattern, where two experiments were conducted using Technique 1 and Technique 2, but the block of data corresponding to Experiment 2 measured with Technique 2 is missing.

Depending on the percentage and pattern of missing values, different strategies may be adopted. For instance, if the pattern of missing data is randomly distributed (Fig. 1A) and the data presents a soft continuity across the variables, such as in spectroscopic data, an interpolation based on the nearest neighbour entries can be a simple and reliable solution to replace the missing observations [17]. However, when missing values show a systematic pattern, interpolation is not an option due to the absence of neighboring available entries, while the extrapolation can be very risky and not recommended. This problem is even more dramatic in the case of missing data blocks in data fusion (Fig. 1C). In this scenario, there are some regions of the data set with full concatenated blocks of information, e.g., when measurements with different techniques were acquired in identical conditions, and some others where a block of information does not have an equivalence and is connected with a block of missing entries, providing the so-called incomplete multisets.

To perform PCA on an incomplete multiset, two main strategies can be adopted, either estimating the missing entries and work with traditional PCA algorithms, such as the Imputation based on SVD (I-SVD) does [15,18], or adapting existing algorithms to work only with the available information, such as Nonlinear Estimation by Iterative Partial Least Squares (NIPALS) could do [19].

I-SVD relies on applying the SVD algorithm after imputing the missing values with estimates. With I-SVD, estimates are subsequently updated using the prediction of the SVD model, and a new SVD is conducted, until convergence. While I-SVD yields accurate orthogonal scores and loadings, the computational cost associated with the method is exceptionally high and convergence is not achieved in a reasonable time for challenging patterns of missing data.

NIPALS is a widely used method for the sequential extraction of principal components in multivariate data analysis. The technique is based on a one-by-one extraction of principal components through an alternating least squares (ALS) approach. Although the mathematical operations of NIPALS are easily adapted to handle missing values, it has been reported that the algorithm fails to provide orthogonal decompositions [18], and it only works properly when the (pseudo)rank of the data is 1, as might have pointed out by Anderson Christoffersson [20]. For higher ranks, the extraction one-by-one of the subsequent components is not adequate as it results in non-orthogonal scores and loadings among the extracted components.

To overcome the limitations of the previous approaches, this work proposes a novel algorithm called Orthogonalized-Alternating Least Squares (O-ALS). O-ALS is adapted to work with missing entries, preserves orthogonality among components and achieves accurate results with a low computational cost. This algorithm operates iteratively performing adapted least-squares row-by-row and column-by-column calculations of the scores and loadings of all components, respectively. In every iteration, the full matrices of scores and loadings are subject to the Gram-Schmidt orthogonalization [21, 22] to ensure orthogonality among the estimated profiles. The proposed O-ALS algorithm offers a promising solution for handling missing values in PCA and provides accurate results in few seconds.

In the remainder of this paper, the *modus operandi* of I-SVD, NIPALS and O-ALS for the analysis of data sets with missing information is first described and their benefits and limitations assessed from a theoretical perspective. Afterwards, the performance of these algorithms is evaluated on simulated and real hyperspectral imaging data fusion examples, where missing block patterns are often encountered.

2. Algorithm description

In this section, the calculation of PCA models in the presence of missing values with I-SVD, NIPALS and O-ALS algorithms is described in detail. All of them were in-house encoded in MATLAB and are available on request.

Imputation based on SVD (I-SVD)

Let us consider $\mathbf{D}$ a full data set and $\mathbf{D}_m$ a derived data set with missing entries. To obtain the PCA model of the matrix $\mathbf{D}_m$, I-SVD works by estimating the missing entries of the data set [15,19]. The approach tries to impute values so that the original data space of the full data set $\mathbf{D}$ is preserved. The steps to apply SVD are as follows:

Step 1) *Initial imputation of missing values in the data matrix* $\boldsymbol{D}_m$. This can be done by replacing missing entries with random values or by the mean of the observed values.

Step 2) *SVD of the imputed data matrix.*

Step 3) *Update of the missing entries with the new predicted values from the SVD model calculated in step 2*. This is done reconstructing the full data $\widehat{\mathbf{D}}$ using a specified number of components equal to the rank of the data set. The missing entries on $\mathbf{D}_m$ are replaced by the predicted values of $\widehat{\mathbf{D}}$. Convergence is typically determined by a small change in the SVD components or by a predefined number of iterations.

This algorithm typically converges to the correct data space of $\mathbf{D}$ since the imputed entries must maintain the structure of the data space of the available data in $\mathbf{D}_m$. In other words, the final imputed values are estimated so that they represent the original space of $\mathbf{D}$. The only assumption for I-SVD to work is that the rank chosen for the SVD reproduction of the data is correct.

Non-Iterative Partial Least Squares (NIPALS)

NIPALS is a well-known iterative algorithm that can be used to estimate sequentially the principal components of a given dataset **D** [19]. The steps of the algorithm are as follows:

Step 1) *Find an initial guess for the loading vector* $\boldsymbol{p}\,(C \times 1)$ *of the first principal component*. This estimate can be a random vector, the column mean of the available data or the most intense column of the data set **D**. This initial estimate is used to start Step 2. Note that the score vector $\mathbf{t}$ $(R \times 1)$ can be also taken as initial estimate. If this is the case, the steps 2A and 2B will be swapped

Step 2) *Iterative ALS optimization of the score and the loading vector.*

2A) Given **D** and $\mathbf{p}^{\mathrm{T}}$, calculate the score vector by least squares.

$$\mathbf{t} = \mathbf{D}(\mathbf{p}^{\mathrm{T}})^{+} \qquad \text{Eq. 2}$$

2B) Given **D** and $\mathbf{t}$, calculate the loading vector.

$$\mathbf{p}^{\mathrm{T}} = \mathbf{t}^{+}\mathbf{D} \qquad \text{Eq. 3}$$

Where "+" indicates the pseudoinverse matrix. Steps 2A and 2B are repeated until convergence is reached, usually defined by a small change in the residual matrix **E** or by a maximum number of iterations. When the convergence is achieved, $\mathbf{t}$ and $\mathbf{p}^{\mathrm{T}}$ are retrieved.

Step 3) *Deflation of the initial data matrix **D***. The data set is deflated by removing the variance explained by the first component (Eq. 4).

$$\mathbf{D}_{\boldsymbol{new}} = \mathbf{D} - \mathbf{t}\mathbf{p}^{\boldsymbol{T}} \qquad \text{Eq. 4}$$

The deflation process ensures that the variance of the next principal component will be orthogonal to the previous one, since the variation described by the first component has been removed from the data. If more components need to be calculated, once the data is deflated, steps 2 and 3 are repeated replacing **D** by $\mathbf{D}_{\boldsymbol{new}}$ in Eq. 2 and 3.

One of the claimed benefits of NIPALS is that it can handle missing values by skipping them during the ALS procedure, as represented in Fig. 2. To deal with the missing entries, the least squares calculations of steps 2A and 2B are done now row-by-row (Eq. 5) and column-by-column (Eq. 6), adapting the least-squares calculations to the available information in $\mathbf{D}_m$.

Step 2A)

$$\mathbf{t}(i, \mathbf{1}) = \mathbf{d}_m(i, :)(\mathbf{p}^{\mathrm{T}})^{+} \qquad \text{Eq. 5}$$

Step 2B)

$$\mathbf{p}^{\mathrm{T}}(\mathbf{1}, j) = \mathbf{t}^{+}\mathbf{d}_m(:, j) \qquad \text{Eq. 6}$$

Where i and j go from 1 to R and 1 to C, respectively. Then, when missing values are encountered in the vector $\mathbf{d}_m(i,:)$ of Eq. 5, the calculation of $\mathbf{t}(i,\mathbf{1})$ is done using only the available entries in $\mathbf{d}_m(i,:)$ (Fig. 2A) and the analogous values of $\mathbf{p}^{\mathrm{T}}$**,** which share position with the available entries in $\mathbf{d}_m(i,:)$. Such an operation is done to obtain all the elements of **t,** using every time the available entries in row $\mathbf{d}_m(i,:)$ and the analogous information in $\mathbf{p}^{\mathrm{T}}$**.** As can be inferred from the calculation, every row can have a completely different number and position of available entries; therefore, the algorithm adapts to any pattern of missing information. Analogously, the same approach is done in Eq. 6 (Fig. 2B) to calculate every element of $\mathbf{p}^{\mathrm{T}}$, using only the available entries in the column $\mathbf{d}_m(:,j)$ and the related information in **t**. Although this simple and elegant approach allows NIPALS to adapt the least-squares calculations for missing data, it will be shown that there are consequences in the orthogonality of the components estimated.

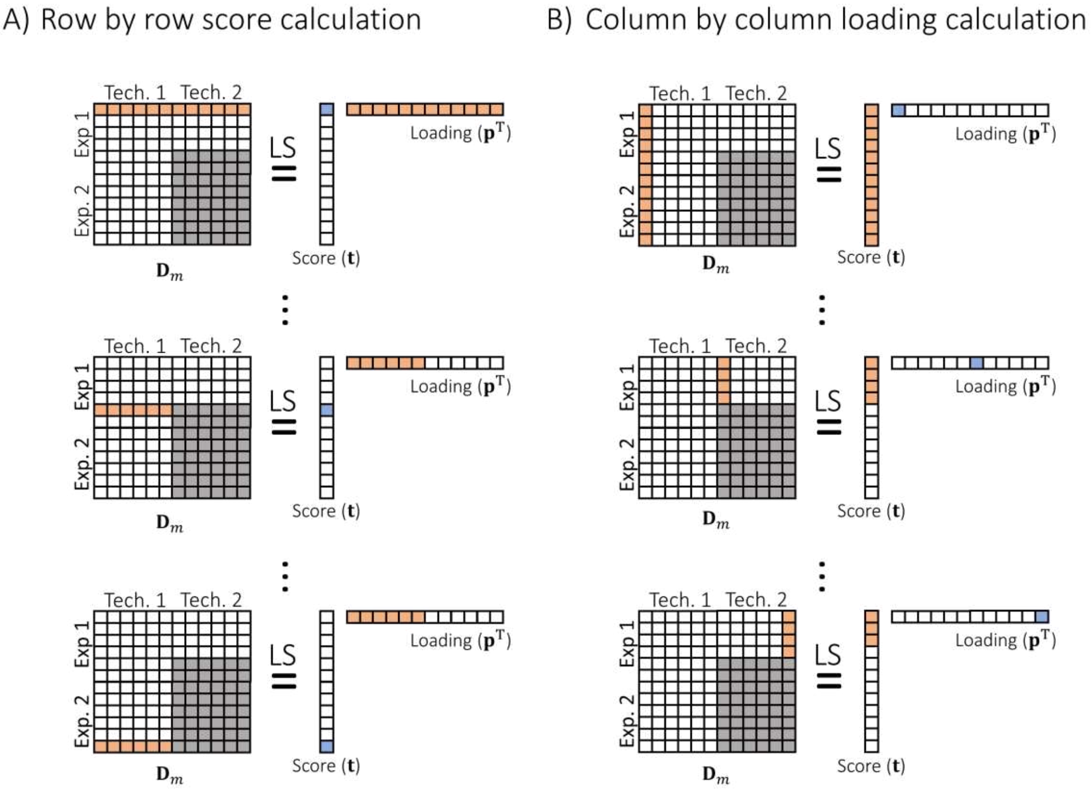

A) Row by row score calculation
B) Column by column loading calculation
Tech. 1
Tech. 2
Exp 1
Exp. 2
LS
=
Loading ($\mathbf{p}^{\mathrm{T}}$)
Score ($\mathbf{t}$)
$\mathbf{D}_m$

Figure 2. A) Row-by-row calculation of the score profile by the NIPALS algorithm. Given a loading $\mathbf{p}^{\mathrm{T}}$ and given a row $\mathbf{d}_m(i,:)$ (both in orange), the corresponding score value $\mathbf{t}(i,1)$ is calculated (blue). If missing values are encountered, the loading is adapted to match the corresponding available entries in $\mathbf{d}_m(i,:)$. B) Column-by-column calculation of the loading profile by the NIPALS algorithm. Given a score **t** and given a column $\mathbf{d}_m(:,j)$ (both in orange), the corresponding score value $\mathbf{p}^{\mathrm{T}}(1,j)$ is calculated (blue). Similarly, if missing values are encountered, the score is adapted to match the corresponding available entries in $\mathbf{d}_m(:,j)$. The calculation is performed for all i and j which go from 1 to R and 1 to C, respectively.

Orthogonalized-Alternating Least Squares (O-ALS)

The new algorithm Orthogonalized-Alternating Least Squares (O-ALS) is designed to work only with the available information to estimate the PCA model. In contrast to the NIPALS approach, this algorithm stands out for its ability to keep the orthogonality of the scores and loading profiles during the alternating least squares calculation.

Summarizing, O-ALS is an iterative alternating least squares bilinear factorization that operates applying a Grand-Schmidt orthogonalization constraint. A key difference with the NIPALS algorithm is that all N components required to describe the variance of the data according to the desired rank are estimated simultaneously. The O-ALS steps are described below:

Step 1) *Generation of an initial estimate for the loadings matrix* $\boldsymbol{P}^T$ *(*$N \times C$*)*. This can be done with random numbers. Note that the score matrix $\mathbf{T}$ ($R \times N$) can also be taken as initial estimates. Then, the steps 2A and 2B will be swapped.

Step 2) *Iterative ALS optimization of the scores and loadings under the Gram-Schmidt orthogonalization constraint*. This includes steps 2A and 2B, visualized in Fig. 3 and described below.

> Step 2A) Row-by-row LS estimation of the scores matrix $\mathbf{T}$ (Fig. 3A). For each data row, the corresponding row of the scores vector is calculated using a least squares optimization process. The calculation of $\mathbf{t}(i,:)$ is performed by using only the available entries in $\mathbf{d}_m(i,:)$ and the analogous values of $\mathbf{P}^{\mathrm{T}}$ that share position with the available entries in $\mathbf{d}_m(i,:)$ (Eq. 7).
>
> $$\mathbf{t}(i,:) = \mathbf{d}_m(i,:)(\mathbf{P}^{\mathrm{T}})^{+} \qquad \text{Eq. 7}$$
>
> where i goes from 1 to R. Once the calculation in Eq. 7 is finished, the Gram-Schmidt orthogonalization constraint is applied to $\mathbf{T}$ to preserve the orthogonality among the score profiles.
>
> Step 2B) Column-by-column LS estimation of the loadings matrix, $\mathbf{P}^{\mathrm{T}}$ (Fig. 3B). Every column of the loading matrix is estimated using only the available entries in $\mathbf{d}_m(:,j)$ and the analogous values of $\mathbf{T}$ that share position with the available entries in $\mathbf{d}_m(:,j)$ (Eq. 8).
>
> $$\mathbf{p}^{\mathrm{T}}(:,j) = \mathbf{T}^{+}\mathbf{d}_m(:,j) \qquad \text{Eq. 8}$$
>
> where j goes from 1 to C. Similarly, the Gram-Schmidt orthogonalization constraint is applied to the matrix $\mathbf{P}^{\mathrm{T}}$ to maintain orthogonality among the loadings, which are individually normalized using the Euclidean norm.

Steps 2A and 2B are iteratively repeated until convergence is achieved. Convergence is typically defined by a small change in the principal component estimates or by a predetermined maximum number of iterations. Upon achieving convergence, the final scores matrix $\mathbf{T}$ and loading matrix $\mathbf{P}^{\mathrm{T}}$**,** formed each of them by orthogonal profiles, are obtained.

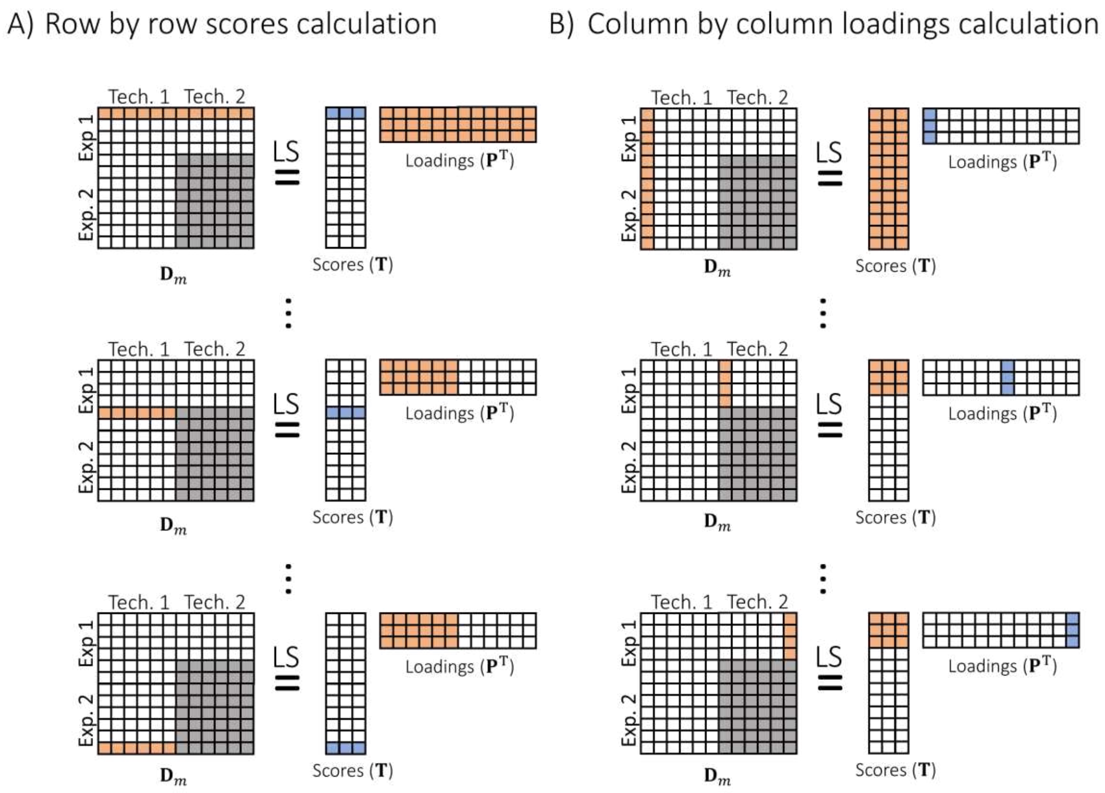

A) Row by row scores calculation
B) Column by column loadings calculation
Tech. 1
Tech. 2
Exp 1
Exp. 2
LS
=
Loadings ($\mathbf{P}^{\mathrm{T}}$)
Scores ($\mathbf{T}$)
$\mathbf{D}_m$

Figure 3. A) Row-by-row calculation of scores by O-ALS for a three-component system. Given the loadings $\mathbf{P}^{\mathrm{T}}$ and a row $\mathbf{d}_m(i,:)$ (both in orange), the corresponding score values $\mathbf{T}(i,:)$ are calculated (blue). If missing values are encountered, the loadings are adapted to match the corresponding available entries in $\mathbf{d}_m(i,:)$. B) Column-by-column calculation of loadings by O-ALS for a three-component system. Given the scores $\mathbf{T}$ and a column $\mathbf{d}_m(:,j)$ (both in orange), the corresponding score values $\mathbf{P}^{\mathrm{T}}(:,j)$ are calculated (blue). Similarly, if missing values are encountered, the score is adapted to match the corresponding available entries in $\mathbf{d}_m(:,j)$. The calculation is performed for all i and j which go from 1 to R and 1 to C, respectively.

3. Datasets

This section includes the details of the simulated and real examples of incomplete multisets with missing blocks as shown in Fig. 1C. Since hyperspectral image fusion is a field where incomplete multisets are easily encountered, the simulations have been performed mimicking the fusion of a NIR and a Raman hyperspectral image, considering various noise levels and missing data patterns. Additionally, a real example of NIR and Raman image fusion is studied.

To set the scene, a HSI consists of large number of spectra associated with a grid of points (pixels) spanning a scanned sample surface (Fig. 4A). A HSI can be represented as a data cube, with two spatial dimensions, sized x and y, that represent the pixel coordinates, and a third spectral dimension, sized λ. To analyze an HSI, the image cube is generally unfolded in the pixel direction by stacking each spectrum one under the other one to form a data matrix (Fig. 4B). When two or more HSI need to be analyzed simultaneously (data fusion), a multiset is built by concatenating the spectra related to the same pixel for each HSI (blocks $\mathbf{D}_2$ and $\mathbf{D}_3$ in Fig. 4C) [16]. The multiset formed integrates spectral information from all the individual HSIs, facilitating joint analysis and exploration of the combined data. Classical image fusion leading to a complete multiset, as in Fig. 4C, requires that the image datasets to be combined cover the same scanned area and have the same pixel size, thus discarding the non-common measured areas and lowering the spatial resolution of some the employed techniques to equal the pixel size among platforms. When pixels of one image without equivalent information in another platform are to be kept, an incomplete multiset must be built by concatenating these pixels (block $\mathbf{D}_1$) with a missing block of information (Fig. 4D).

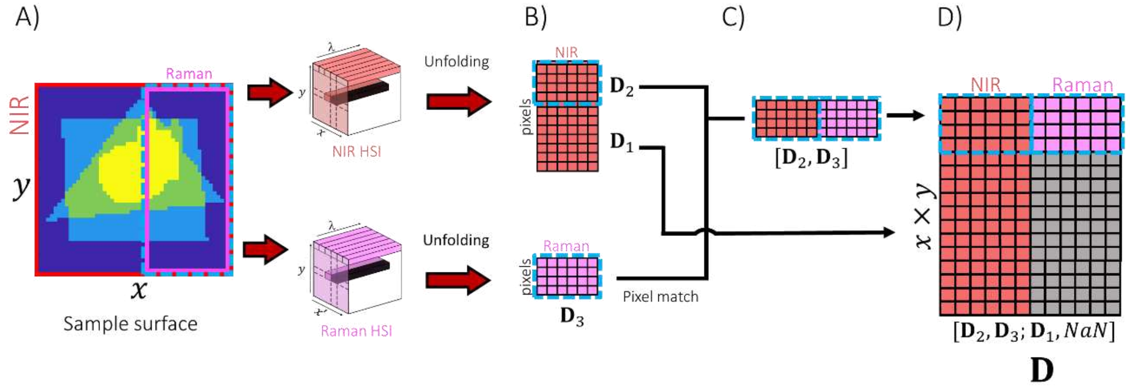


Figure 4. Simulated case of image fusion A) Picture of the simulated scanned sample. The red square corresponds to the area scanned by NIR. The pink square corresponds to the area scanned by Raman. In dashed blue line, the common scanned area by both Raman and NIR techniques. Both HSI have the same pixel size. B) The NIR and Raman spectra corresponding to the same scanned area, $\mathbf{D}_2$ and $\mathbf{D}_3$ respectively, are fused in a single complete multiset (C). Finally, the incomplete multiset is built by concatenating the multiset $[\mathbf{D}_2, \mathbf{D}_3]$ and $\mathbf{D}_1$ (representing the sample area scanned only by NIR). The missing block (in gray) is filled with Not-a-Number (NaN) and corresponds to the non-scanned area by Raman.

Simulated datasets

The simulated incomplete multisets represent a fusion of an NIR and a Raman image following the scheme of Fig. 4. The images contain three different components and cover a total scanned area of 50×50 pixels. The pure Raman spectra span the spectral range from 700 to 1600 cm^{-1} and include 600 spectral channels. On the other hand, the NIR spectra cover the wavelength range from 935 to 1720 nm, with a total of 60 spectral channels. The pure components employed to generate the simulated data are shown in Fig. 5A. Two different levels of Poisson noise were proposed (noise-free and 7% of the total signal) to study the algorithms in noise-free conditions and mimicking the uncertainty present in real photon-counting techniques. For each noise level, two distinct missing data patterns were considered: a) randomly distributed missing values with varying proportions of missing entries (0%, 5%, 30%, and 80%) and b) missing block pattern, where an entire block of missing data is introduced mimicking a scenario where the Raman measurement was covering only part of the sample area. Similar to the randomly distributed case, missing blocks were introduced at proportions of 0%, 5%, 30%, and 80% of the total entries of the data set.

Real dataset

A real sample was measured by NIR and Raman imaging. This example is a controlled sample formed by a drawing done with three commercial blue pens, Uniball Signo (US), Bic Velocity Gel (BV) and Pilot V Ball Grip (PVG) on a conventional paper surface (Fig. 5B) [23]. This case is similar to that in Fig. 4, where all the surface was scanned by NIR imaging, but only a small part of it by Raman imaging. This generates an incomplete multiset with 70% of missing values. For this system, a 4-component PCA model is expected to properly describe the signals of the three inks and the paper.

The NIR hyperspectral image was acquired by a pushbroom NIR camera (Specim FX17 by Spectral Imaging Ltd., Oulu, Finland) and it was formed by 224 spectral channels covering the 935–1720 nm spectral range and 246 × 225 pixels with a pixel size approximately of 106× 106 μm^2. A Savitzky-Golay first derivative was applied (second order, 15 window points) to remove the spectral baseline.

The Raman hyperspectral image was collected using an INVIA RAMAN Microscope spectrometer (RENISHAW, Gloucestershire, UK). The studied spectral range goes from 270 to 2015 cm−1, with a spectral resolution of 1.55–1.95 cm^{-1} depending on the Raman shift scanned. Pixel size was approximately 26.5 × 26.5 μm^2. A Savitzky-Golay first derivative was applied in the spectra direction (second order, 15 window points) to remove spectra baselines. Raman pixels were binned to achieve the same pixel size as in the NIR image.

Following the scheme of Fig. 4, both images were fused in a single incomplete multiset.

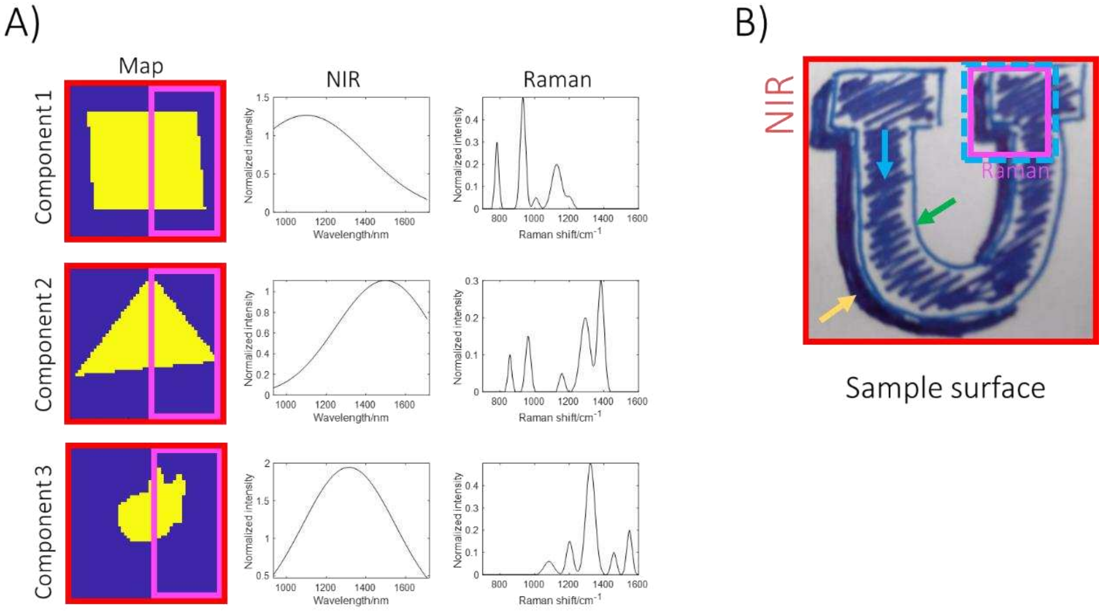


Figure 5. A) Pure distribution maps and pure spectral profiles of the components used to simulate the image fusion case. The area enclosed by the pink rectangle corresponds to the common scanned area by NIR and Raman images, while the red area corresponds to the area scanned only by NIR. B) Picture of the real scanned sample. The yellow, blue and green arrows correspond to BV, PVG and US inks, respectively. The red square corresponds to the area scanned by NIR. The pink square corresponds to the area scanned by Raman. In dashed blue line is the common scanned area among techniques. The image fusion was performed similarly to Fig. 4.

4. Results

The simulated data sets were used to compare and understand the differences between the NIPALS, I-SVD and O-ALS algorithms. The performance of the algorithms was tested by comparing the scores and loadings profiles retrieved on the data set with missing values with those obtained in the full dataset. Besides, the variance explained by the PCA model with a number of components equal to the rank of the simulated data is provided.

Results of simulated datasets

The simulated datasets cover the relevant scenarios where the performance of the NIPALS, I-SVD and O-ALS algorithms can be properly tested, i.e., different noise levels (noise- free or 7% of noise respect to the total signal), different patterns of missing data (random entries or missing block entries) and several percentages of missing data (5, 30 and 80%). Since all algorithms require an iterative optimization, the convergence criterion based on differences in error among consecutive iterations was set to 10^{-11}% (close to the machine limit precision) to avoid prematurely stopping the calculations. In most analyses, initial estimates were chosen as random values. Results are summarized in Table 1 and 2.

Table 1. Correlation coefficients among recovered and true scores and loadings (from the complete matrix) for the random missing pattern case. Explained variance with a model of 3 components (rank of the simulated data).

				Noise-free			**Noise 7%**		
Missing pattern	**Algorithm**	**Component**		**Percentage of missing data**			**Percentage of missing data**		
				5%	**30%**	**80%**	**5%**	**30%**	**80%**
Random	**NIPALS**	**1**	**Score**	1.0000	0.9999	0.9998	1.0000	0.9998	0.9996
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		2	**Score**	1.0000	0.9998	0.9993	0.9999	0.9980	0.9966
			Loading	1.0000	1.0000	0.9999	1.0000	0.9999	0.9997
		3	**Score**	1.0000	1.0000	0.9996	0.9999	0.9937	0.9909
			Loading	1.0000	1.0000	0.9999	1.0000	0.9984	0.9982
		Explained variance (%)		100.00	99.99	99.97	99.53	99.53	99.51
	I-SVD	**1**	**Score**	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		2	**Score**	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		3	**Score**	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		Explained variance (%)		100.00	100.00	100.00	99.53	99.53	99.54
	O-ALS	**1**	**Score**	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		2	**Score**	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		3	**Score**	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		Explained variance (%)		100.00	100.00	100.00	99.53	99.53	99.54

When the missing entries are distributed randomly (Table 1), NIPALS, I-SVD and O-ALS perform very well (all correlation coefficients are >0.99). Here, some remarks must be considered. First, I-SVD and O-ALS perform in an excellent way, i.e., there is a perfect match between the retrieved and the true simulated scores and loadings for all scenarios, even with a high percentage of missing data when all components are represented by the available information. We found that the percentage of missing entries when a random pattern is used does not affect the accuracy of the solutions provided by the I-SVD and the O-ALS algorithms. The relevance of the percentage of missing data will be discussed in detail in the following subsections.

Regarding the NIPALS solution, the retrieved scores and loadings profiles show a slight degradation as the index of the principal component increases, as well as it can be observed in the explained variance. When the percentage of missing data and the noise level increase, this effect becomes even more noticeable.

Table 2. Correlation coefficients among recovered scores and loadings and true scores and loadings (from the complete matrix) for the missing block pattern case. Explained variance with a model of 3 components (rank of the simulated data).

				Noise-free			Noise 7%		
Missing pattern	**Algorithm**	**Component**		**Percentage of missing data**			**Percentage of missing data**		
				5%	**30%**	**80%**	**5%**	**30%**	**80%**
Missing block	**NIPALS**	**1**	**Score**	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	0.9997	0.9996	0.9993	0.9997	0.9996	0.9993
		2	**Score**	0.9967	0.9915	0.9829	0.9966	0.9914	0.9829
			Loading	0.9992	0.9976	0.9921	0.9992	0.9976	0.9921
		3	**Score**	0.9836	0.9835	0.9802	0.9826	0.9799	0.9738
			Loading	0.9987	0.9866	0.9600	0.9985	0.9861	0.9587
		Explained variance (%)		99.99	99.98	99.98	99.54	99.59	99.62
	I-SVD	**1**	**Score**	1.0000	1.0000	-*	1.0000	1.0000	1.0000*
			Loading	1.0000	1.0000	-*	1.0000	1.0000	1.0000*
		2	**Score**	1.0000	1.0000	-*	0.9999	0.9996	0.9994*
			Loading	1.0000	1.0000	-*	1.0000	1.0000	0.9998*
		3	**Score**	1.0000	1.0000	-*	1.0000	0.9956	0.9936*
			Loading	1.0000	1.0000	-*	0.9983	0.9998	0.9986*
		Explained variance (%)		100.00	100.00	-*	99.55	99.61	99.64
	O-ALS	**1**	**Score**	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		2	**Score**	1.0000	1.0000	1.0000	0.9999	0.9996	0.9994
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998
		3	**Score**	1.0000	1.0000	1.0000	0.9983	0.9956	0.9936
			Loading	1.0000	1.0000	1.0000	1.0000	0.9998	0.9986
		Explained variance (%)		100.00	100.00	100	99.55	99.61	99.64

*NIPALS solutions were used as initial estimates instead of random values due to the impossibility to converge of the I-SVD algorithm when initialized with random values.

When a missing block pattern is used (Table 2), NIPALS shows now a strong degradation on the retrieved scores and loadings for all cases. The results of I-SVD are excellent in absence of noise for the cases of 5% and 30% of missing data. However, I-SVD was unable to converge in a reasonable time (<6 h) for the case of 80% of missing data. Instead, the O-ALS algorithm has excellent results for all cases, providing the correct scores and loadings in less than a minute. Instead,

Looking more carefully at the correlation coefficients of Table 1 and 2, very interesting aspects can be investigated, e.g., why NIPALS fails, or why I-SVD and O-ALS achieve excellent results even for 80% missing entries for random patterns. The behavior of these results is addressed in the following specific subsections.

Presence of bias in the recovered scores and loadings

The bias of the NIPALS components in the presence of missing values was already reported by Bjørn Grung and Rolf Manne [18]. Indeed, the one-by-one calculation of components employed by NIPALS explains the incorrect retrieval of the scores and loadings with missing data, as will be explained in detail and can be perceived in Fig. 6.

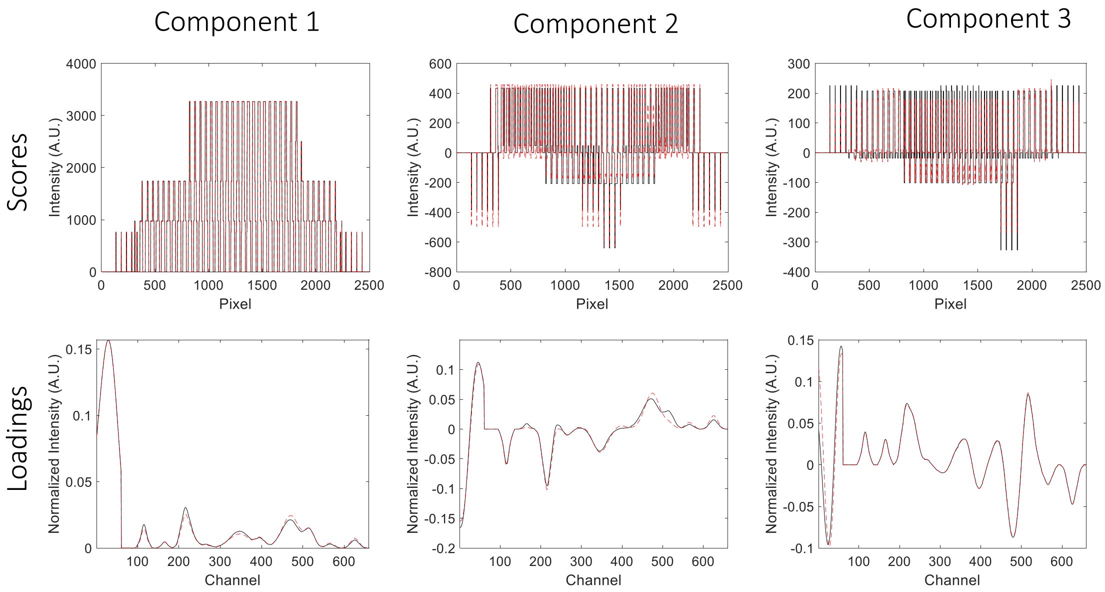


Figure 6. Scores and loadings extracted by NIPALS (dashed red line) and scores and loadings of the complete matrix (black line) for the free-noise missing block case with 80% of missing entries. Small deviations among NIPALS scores and loadings and the true simulated profiles can be seen already in the first principal component. The differences significantly increase in the second and third components.

To understand the reason for the bias in the scores and loadings retrieved by NIPALS, it is important to remind that when $\mathbf{D}_m$ is analyzed, I-SVD and O-ALS estimate a PCA model with a number of components equal to the rank of the data, *N,* i.e., the models obtained are always describing the real space of the data and no bias is detected in any instance. Instead, NIPALS computes the components sequentially and in every step tries to define $\mathbf{D}_m$, or the resulting deflated matrix, with a rank-1 approximation model. Fitting the data $\mathbf{D}_m$ skipping the missing entries generates a data space equivalent to impute the model $\mathbf{tp}^{\mathrm{T}}$ on the missing entries of $\mathbf{D}_m$. Thus, when calculating the first component, the missing part of $\mathbf{D}_m$ comes from a rank-1 approximation $\mathbf{tp}^{\mathrm{T}}$, while the rest of $\mathbf{D}_m$ is still rank N, being incoherent with the original data structure and generating incorrect components. The same effect occurs when the following components are calculated, but it is even aggravated because the deflation step performed using incorrect components adds to the bias due to the discrepancy between the rank of the available entries in the deflated $\mathbf{D}_m$ and the rank-1 approximation obtained with $\mathbf{tp}^{\mathrm{T}}$for the missing entries. It is expected then to observe an increase in the degradation of the scores and loadings as more components are extracted, as it can be observed in the results in Table 1 and Table 2. Thus, the only instance in which the use of the adapted NIPALS algorithm works in the presence of missing data is when the rank of $\mathbf{D}_m$ is equal to 1.

Additional facts that support the inaccurate extraction of scores and loadings by NIPALS are the non-orthogonality among the scores and loadings profiles obtained and the variance explained by the PCA model with the correct rank *N,* which is always lower than the variance expected and properly described by analogous I-SVD or O-ALS PCA models with the same number of components.

Although the bias mentioned above always occurs, the error induced by NIPALS can be more manageable when the number of missing entries and components are low, as it is shown in the results for the case of 5% of missing entries. However, even in these conditions, the resulting bias is strongly data-dependent.

Effect of the pattern and the percentage of missing data into the obtained scores and loadings.

The fundamental concept behind PCA is based on the assumption that the data has an underlying structure or pattern that can be captured by a low rank representation of the relevant information. When missing values are present, I-SVD works to estimate the missing values so that the low rank-*N* model required to describe the variation in the data set is preserved. Instead, O-ALS use only the available information to estimate the scores and loadings that preserve the underlying low rank-*N* structure. It is important to point out that both methods span the same rank-*N* data space (Fig. 7) and this means that if the available information describes well this space, a correct imputation will be achieved and, as a consequence, correct scores and loadings will be obtained, and vice versa.

Since the available information in the incomplete simulated data sets is enough to define properly the rank-*N* space, perfect scores and loadings are achieved as it can be observed in Table 1 and 2 and Fig. 7 for the noise-free case. Thus, there is no dependence either of the missing pattern or of the percentage of missing values, but of the information contained in the available data.

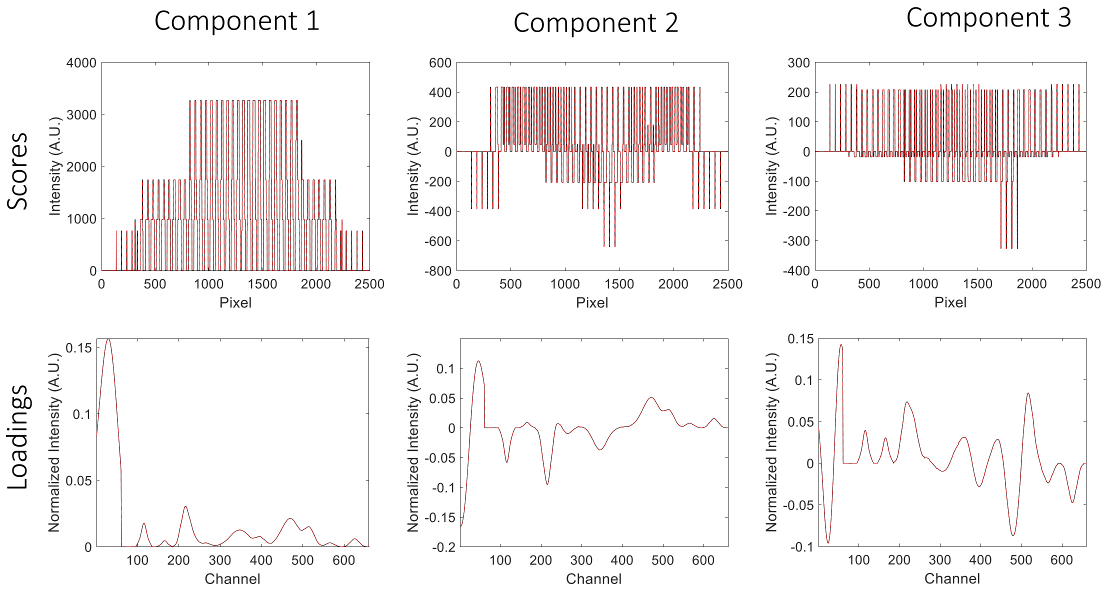


Figure 7. Scores and loadings extracted by O-ALS and I-SVD (dashed red line) and scores and loadings of the complete matrix (black line) for the noise-free missing block case with 30% of missing entries. Perfect match is achieved for both algorithms between the scores and loadings retrieved and those obtained when the full data matrix is available.

Despite the correct retrieval of scores and loadings by I-SVD, the imputation step required by this algorithm may lead to an extremely slow convergence in practice. Indeed, due to the nature of the I-SVD algorithm, the imputed values play a relevant role in the PCA model estimated in every iteration. If the percentage of missing values is high and the missing block pattern does not allow for an easy estimate of the missing entries, the high leverage of these estimated values in the PCA model compared to the weight of the originally available entries will result in an extremely long convergence time (hours or even days), making the use of this algorithm impractical. This is the case shown in Table 2 for the situation of 80% of missing data with a missing block pattern where a solution could not be found for the noise-free case even if the reconstructed data set by the NIPALS solution was used to initialize the algorithm. For the case in the presence of noise, the reconstructed model by the NIPALS solution had to be used as initial estimate for I-SVD, otherwise the algorithm did not converge in a reasonable time (hours at least).

The O-ALS method is not affected by the slow convergence problem since the *modus operandi* of the algorithm ignores the missing entries instead of pushing the predicted missing entries to the correct dataspace, thus saving time. In fact, O-ALS shows a very similar computation time for all percentages of missing entries, contrarily to algorithms based on imputation. To see clearly this effect, the datasets with a 7% noise added following a random and a missing block pattern were analyzed 500 times by O-ALS and I-SVD using different random initial estimates in each run (Table 3).

Table 3. Mean and standard deviation of the time spent by each algorithm in 500 runs for data sets with 7% noise added. Random values were used as initial estimates.

Algorithm	**Random pattern**			**Block pattern**		
	5%	**30%**	**80%**	**5%**	**30%**	**80%**
I-SVD	10±1s	27±3s	72±2s	114s±12s	17±1 min	-
O-ALS	4±0.8s	2.5±0.6s	2.2±0.5s	17±6s	20±21s	19±11s

The slow convergence of I-SVD for data sets with high percentage of missing entries is not observed when the missing pattern is random (Table 1), where the mean computation times were 10±1, 27±3 and 72±2s for the 5, 30 and 80% of random missing data respectively, while for O-ALS were 4±0.8, 2.5±0.6 and 2.2±0.5s. Despite O-ALS is always faster, no significant differences were found in practical terms. This is explained by the information pattern in the available data $\mathbf{D}_m$. The chance to find a good imputation per each iteration depends on the ability to properly capture the original space of **D** from the available data $\mathbf{D}_m$ when SVD is applied after filling the missing entries. A good imputation per iteration can be easily done when the pattern of missing entries is random since the data structure is very well preserved ($\mathbf{D}_m$ closely represents the complete data **D**).

On the other hand, when complete missing blocks of information are present, the available information in $\mathbf{D}_m$ does not allow for a fast estimate of the missing entries. In this context, the results show that O-ALS spent 17±6, 20±21 and 19±11s for the 5, 30 and 80% missing value-case respectively, whereas I-SVD spent 114±12s and 17±1 min for the 5 and 30% missing value case, respectively. However, I-SVD was unable to converge in a reasonable time (>6 h) for the 80% missing block pattern, becoming impractical for the analysis.

Whereas I-SVD can suffer from an excessively slow convergence, a possible limitation of the O-ALS algorithm is the possibility to fall into local minima. Possible local minima have been detected for the O-ALS algorithm when the missing block pattern is analyzed. From 500 runs, 257, 159 and 148 for the 5, 30 and 80%-case, respectively, were stuck in a possible local minimum, detected because the explained variance of the model is slightly lower than the one expected. No local minima were detected for data with a random missing pattern. The presence of local minima for bilinear factorizations when missing data are present was demonstrated by Ilin and Raiko in 2010 [24]. If this is the case, the high speed of the O-ALS algorithm easily solves the problem. Thus, the algorithm can be initialized several times during few iterations using different initial estimates each time. Then, the initial estimates that provided the best fit are selected and used to perform the analysis until convergence is achieved, allowing the retrieval of the correct PCA model without increasing significantly the analysis time. For the extreme case (80% of missing data, missing block pattern) and using the best 10 different initializations during 35 iterations, the local minima encountered frequency is reduced from 32% to <1% while the analysis time is increased from 19s to 120s.

Results of a real case study

The O-ALS algorithm was applied to the real dataset described in section 3 which contains 70% of missing values. The convergence criterion based on differences in error among consecutive iterations was set as 10^{-11}% (close to the machine limit precision) to avoid stopping prematurely the algorithm. Initial estimates were chosen as random values using four components. Ten different initializations were performed to avoid possible local minima. Results are shown in Fig. 8.

Four components were detected. The variance explained for each one was 62.48, 25.87, 4.55 and 0.47% for PC 1, 2, 3 and 4, respectively. The total explained variance of the model was 93.37%, which is reasonable considering the noise of real NIR and Raman images.

Figure 8A displays the clusters associated with the pixels of the pen inks and with the pixels of the paper represented in the PCA space. The first PC was not shown because of the lack of informative relevance (related to describe the mean of the data since the data set was not mean-centered). The PC 3 vs PC 2 score plot allows differentiating the pixels of paper (black circle) and of the US ink (green circle) from pixels of the rest of inks (see Figure 8B for the corresponding representation of the position of the pixels of a cluster on the area scanned). The signal from the US pen is very clear and selective in both NIR and Raman techniques; hence the differentiation in lower PCs. On the other hand, PC 4 vs PC 3 score plot differentiates the pixels corresponding to US ink (green circle), BV ink (blue circle) and it can be slightly observed the cluster of the PVG ink (yellow circle). Finally, the PC 4 vs PC 2 score plot allows a clear separation of the pixels of the PVG ink (yellow circle) from others. The signals from PVG, BV and the paper are not very selective in Raman, and not selective at all in NIR. However, the differences are enough to cluster correctly the pixels of all components.

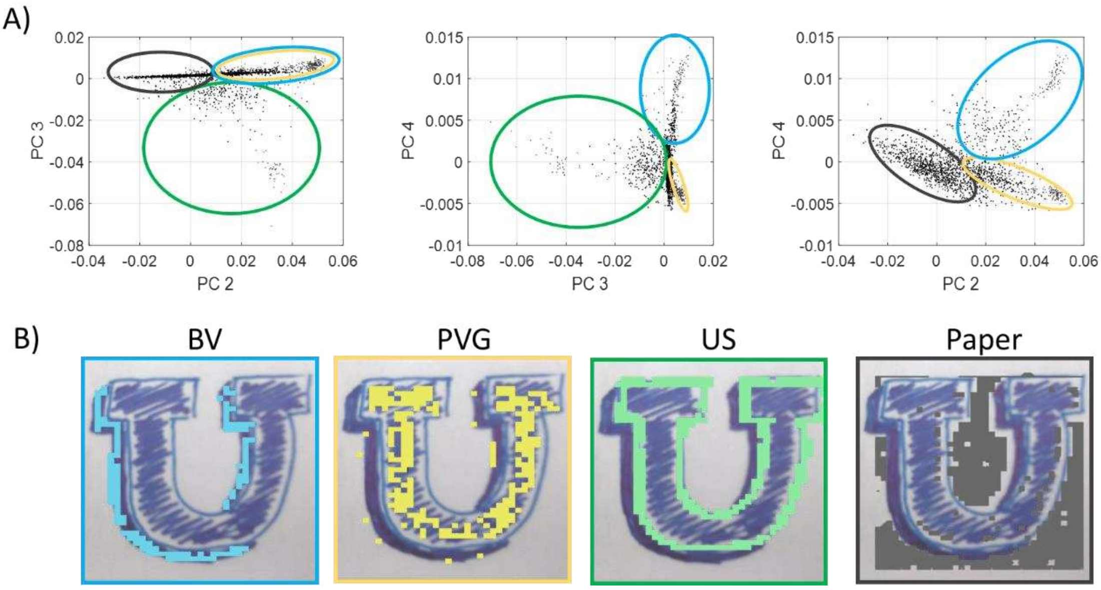


Figure 8. A) Scores extracted by O-ALS. Clusters related to pixels from different inks and paper can be observed. In black, the cluster related to the paper. In green, pixels related to the US ink. In blue, pixels related to the BV ink. In yellow, pixels related to the PVG ink. B) Pixels of the clusters selected (blue, green, yellow and cyan) displayed on the original image.

Summarizing, the analysis suggests that O-ALS effectively detected the four expected clusters related to the signals from the three pen inks and the paper and captured the PCA space of the dataset despite the high percentage of missing entries and the challenging missing block pattern.

5. Conclusions

In this study, the new Orthogonalized- Alternating Least Squares algorithm (O-ALS) has been presented as a fast and accurate algorithm to provide PCA models for data sets with any kind of pattern and percentage of missing entries. O-ALS works only with the available entries in the data set and estimates simultaneously all necessary components in the PCA model using an alternating least-squares optimization of the scores and loadings under the Gram-Schmidt orthogonalization constraint.

Comparing the performance of O-ALS with the NIPALS and I-SVD algorithms in the same scenarios, some characteristics are key to understand the differences in the results provided by the different methodologies. A relevant fact is that O-ALS and I-SVD work always extracting simultaneously the *N* components required for the PCA model, whereas NIPALS proceeds with the one-at-a-time sequential extraction of components. Working in the correct rank-*N* space results in the consistent retrieval of accurate scores and loadings of O-ALS and I-SVD across various scenarios differing in noise level, pattern and percentage of missing entries. NIPALS instead suffers from a bias in the retrieved scores and loadings due to the rank-1 approximation used in the extraction of the components and a loss of orthogonality among the extracted profiles.

A fundamental difference between I-SVD and O-ALS is that the former works imputing the missing entries and the latter using only the available information. Whereas the *modus* *operandi* of the algorithms does not influence the accuracy of the results obtained, it has a significant effect in the computation time and convergence among them. O-ALS remains a fast algorithm whatever the pattern and percentage of missing entries in the data. Instead, the convergence of the I-SVD algorithm gets compromised for data sets with high percentage of missing entries and complex missing block patterns.. Finally, the only limitation identified for the O-ALS algorithm is the possibility to fall in local minima in extreme cases of missing block patterns with high percentage of missing entries. In this instance, starting with a small number of initializations for a few iterations and selecting the model with the best fit to obtain the definitive PCA model clearly solves the problem.

Although the O-ALS algorithm is suitable for any kind of data set with missing entries, a very promising application field is image fusion, where missing block patterns always exist, the proportion of missing values can reach up to 80% and the size of the data sets can be massive. In this scenario, the fast and accurate O-ALS algorithm can be an excellent tool for exploratory purposes and to, eventually reconstruct the missing blocks of information if required.

Acknowledgements

A. G.-S. warmly thanks Martina Beese, PhD student in Leibniz-Institut für Katalyse and in Institut für Mathematik (Universität Rostock) for fruitful discussions. A. G.-S. and A.J. acknowledge financial support from the Catalan government (2021 SGR 00449). A. G.-S. acknowledges scholarships from the MOBLILEX U Lille program, from the Santander bank and from Fundació Montcelimar.

References

1. Jolliffe, I. T.; Cadima, J. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* (**2016**) 374.2065:20150202.
2. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 2.1-3 (**1987**): 37-52.
3. Camacho, J.; Picó, J.; Ferrer, A. Data understanding with PCA: structural and variance information plots. *Chemometrics and Intelligent Laboratory Systems* 100.1 (**2010**): 48-56.
4. Grahn, H.; Geladi, P. (Eds.). (**2007**). *Techniques and applications of hyperspectral image analysis*. John Wiley & Sons.
5. Amigo, J. M.; Martí, I.; Gowen, A. Hyperspectral imaging and chemometrics: A perfect combination for the analysis of food structure, composition and quality. *Data handling in science and technology*. Vol. 28. Elsevier, **2013**. 343-370.
6. Torres-Cobos, B.; et al. Varietal authentication of virgin olive oil: Proving the efficiency of sesquiterpene fingerprinting for Mediterranean Arbequina oils. *Food Control* 128 (**2021**): 108200.
7. Tauler, R.; Casassas, E. Principal component analysis applied to the study of successive complex formation data in Cu (II)–ethanolamine systems. *Journal of Chemometrics* 3.S1 (**1989**): 151-161.

8. Wold, S.; Sjöström, M. SIMCA: a method for analyzing chemical data in terms of similarity and analogy. Vol. 52. ACS Symposium Series. **1977**. 243-282
9. Avila, C. R.; et al. Process monitoring of moisture content and mass transfer rate in a fluidised bed with a low-cost inline MEMS NIR sensor. *Pharmaceutical Research* 37 (**2020**): 1-19.
10. Kourti, T. Quality by design in the pharmaceutical industry: process modelling, monitoring and control using latent variable methods. IFAC Proceedings 42(11) (2009): 36-41.
11. Trefethen, L. N.; Bau, D. Numerical Linear Algebra. *SIAM*, Philadelphia, PA, **1997**.
12. Alier, M.; Tauler, R. Multivariate curve resolution of incomplete data multisets. *Chemometrics and Intelligent Laboratory Systems* 127 (**2013**): 17-28.
13. De Luca, M.; Ragno, G.; Ioele, G.; Tauler, R. Multivariate curve resolution of incomplete fused multiset data from chromatographic and spectrophotometric analyses for drug photostability studies. *Analytica Chimica Acta* 837 (**2014**): 31-37.
14. Piqueras, S.; et al. Handling different spatial resolutions in image fusion by multivariate curve resolution-alternating least squares for incomplete image multisets. *Analytical Chemistry* 90 (**2018**): 6757-6765.
15. Walczak, B.; Massart, D. L. Dealing with missing data: Part I. *Chemometrics and Intelligent Laboratory Systems* 58.1 (**2001**): 15-27.
16. de Juan, A.; de Oliveira, R. R.; Gómez-Sánchez, A. Multiset analysis by multivariate curve resolution: The unmixing methodology to handle hyperspectral image fusion scenarios. Data Handling in Science and Technology. Vol. 33. Elsevier, **2024**. 111-132.
17. Bahram, M.; et al. Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation. *Journal of Chemometrics: A Journal of the Chemometrics Society* 20.3-4 (**2006**): 99-105.
18. Grung, B.; Manne, R. Missing values in principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 42.1-2 (**1998**): 125-139.
19. Wold, H. Soft modeling by latent variables: the nonlinear iterative partial least squares (NIPALS) approach. *Journal of Applied Probability* 12.S1 (**1975**): 117-142.
20. Christoffersson, A. The One Component Model with Incomplete Data. Doctoral Thesis, University of Uppsala, Sweden. (**1970**).
21. Schmidt, E. Zur Theorie der linearen und nichtlinearen Integralgleichungen I. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener. *Mathematische Annalen* 63 (**1907**): 433–476.
22. Gram, J. P. Ueber die Entwickelung reeller Functionen in Reihen mittels der Methode der kleinsten Quadrate. *Journal for die Reine und Angewandte Mathematik* 94 (**1883**): 41–73.
23. Borba, F.D.S.L.; Honorato, R.S.; de Juan, A. Use of Raman spectroscopy and chemometrics to distinguish blue ballpoint pen inks. *Forensic science international*, 249 (**2015**): 73-82.
24. Ilin, A.; Raiko, T. Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research* 11 (**2010**): 1957-2000.

The missing block problem in image fusion

Multisets are formed by connecting information blocks along the row and/or the column direction. The absence of certain blocks of data can result in an incomplete multiset, which shows a strong systematic pattern of missing values (Fig. 48). Several scenarios can originate incomplete multisets, such as missing monthly data in environmental monitoring campaigns for some measurements [Alier and Tauler, 2013] or absent or unfeasible experiments for certain techniques in multitechnique spectroscopic monitoring [De Luca et al., 2014].

In image fusion, the incomplete structures often respond to the natural characteristics of the imaging platforms. For instance, as stated in subsection 2.2.4, HSIs acquired by different imaging platforms may cover different scanned areas. This usually happens when the measurement of one platform is very time consuming and the image should be reduced to specific areas of interest. The resulting image fusion leads to an incomplete multiset, where non-common scanned areas do not have an equivalent block of information to be connected to (Fig. 16). The fusion of HSIs acquired through spectroscopic techniques with different spatial resolution can lead to incomplete multiset structures as well. In the classical image fusion, the spatial resolution differences are overcome by a binning process. However, the binning has drawbacks, such as the loss of the spatial resolution and a possible increased rotational ambiguity in the MCR-ALS solutions, as commented **Publication VI**. A solution to this problem is including the pixel spectra of the high resolution-HSI appended to the fused images at lowest spatial resolution to form an incomplete multiset structure, as the one shown in (Fig. 17) [Piqueras et al., 2018; de Juan et al., 2024]. Although some algorithms have been proposed to deal with this kind of structures, the present thesis presents two alternatives for unmixing and exploratory purposes.

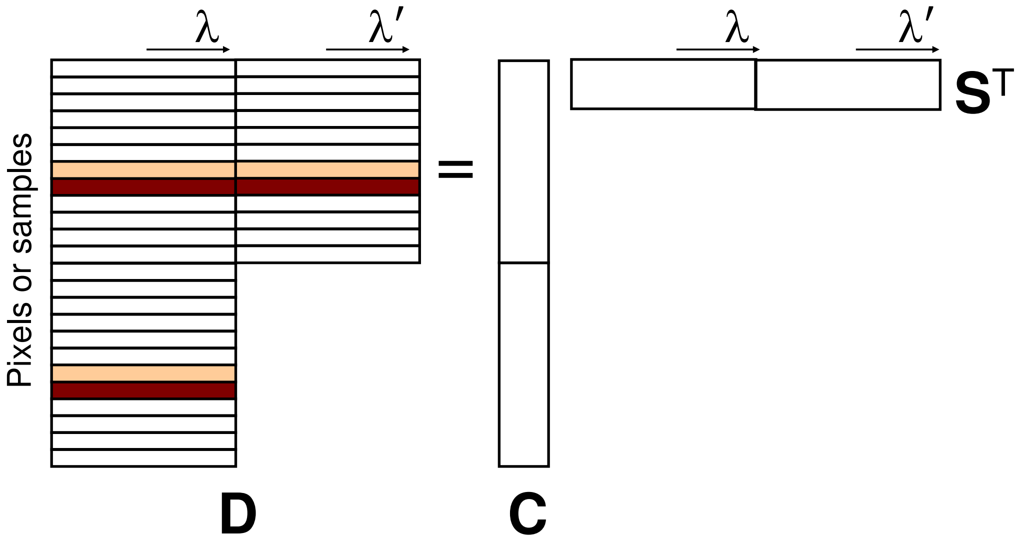


Figure 48. Schematic representation of an incomplete multiset structure and the associated bilinear model.

The evolution of the MCR-ALS algorithm to deal with missing data

Publication VIII provides a comprehensive review of the evolution of MCR algorithms to handle missing data. If the goal is to retain as much information as possible while addressing the missing value problem in MCR, there are typically two different perspectives to consider: a) filling the missing information with estimated values and analyzing the completed structure (the so-called *imputation*), and b) recovering the bilinear model based solely on the real available information, i.e., avoiding the use of predicted or imputed values when the model is calculated.

The imputation approach offers the advantage of obtaining a complete multiset structure for MCR analysis, which can be used to obtain a single factorization model. The imputed values can be calculated as averages or median values of neighbouring available information. More advanced PCA-based imputation approaches, such as updating the imputed values during iterations [Grung and Manne, 1998; Walczak and Massart, 2001], can also be used to provide the initial complete multiset submitted to MCR. However, the performance of these imputation approaches becomes more compromised when the amount of missing information increases, particularly in cases of systematic patterns such as the missing block pattern seen in incomplete multisets. In such instances, the convergence of imputation methods becomes slow, and biases in estimated values may significantly impact the results if the imputation approach is stopped prematurely.

On the other hand, there are approaches that avoid the imputation step to recover the bilinear model and rely only on the available information in the initial incomplete multiset. This is the case of the methodology proposed by Alier and Tauler in 2013. Here, to address the presence of missing blocks during the MCR iterations, the incomplete multiset is split in smaller complete submultiset structures. In every iteration, MCR-ALS is individually applied to the complete submultisets and the pure profiles provided by the different bilinear submodels are suitably combined to obtain the total bilinear model describing the full incomplete multiset. Although this algorithm has been successfully applied in different examples, it presents several drawbacks. First, this approach requires multiple factorizations and the complexity to combine the results obtained increases dramatically as the number of submultisets grows. Additionally, a proper combination of the pure profiles obtained requires the prior identification of the number and identity of components of each submultiset, which is not always an easy task.

Hence, the novel methodology proposed combines the benefits of both imputation and multiple factorization approaches, allowing for the analysis of incomplete multisets with a single factorization model based solely on the available information.

The missing block problem solved with a single factorization model

The proposed single factorization approach to address the missing block problem in MCR requires going back to fundamental least-squares principles and recovering the equations 7 and 8 of subsection 2.2.2, used for the least-squares calculation of **C** and $\mathbf{S}^{\mathrm{T}}$ given **D**.

Those least-squares equations, which are designed to analyze complete multisets, can be reformulated in equivalent expressions that obtain the **C** and $\mathbf{S}^{\mathrm{T}}$ matrices row-by-row and column-by-column, respectively (Eq. 24 and Eq. 25). This reformulation does not change the scheme of MCR-ALS algorithm, which remains exactly the same as in Subsection 3.2.

$$\mathbf{c}(i,:) = \mathbf{d}(i,:)\left(\mathbf{S}^{\mathrm{T}}\right)^{+} \qquad \text{Eq. 24}$$

$$\mathbf{s}^{\mathrm{T}}(:,j) = \mathbf{C}^{+}\mathbf{d}(:,j) \qquad \text{Eq. 25}$$

Here, each row of **C** ($\mathbf{c}(i,:)$) is recovered using the corresponding row of **D** ($\mathbf{d}(i,:)$) and $\mathbf{S}^{\mathrm{T}}$, and each column of $\mathbf{S}^{\mathrm{T}}$ ($\mathbf{s}^{\mathrm{T}}(:,j)$) is recovered using the related column of **D** ($\mathbf{d}(:,j)$) and **C**. This formulation provides a clearer view into the mechanics of alternating least squares, which is now viewed from a vectorized perspective.

With this in mind, the essence of the ALS method to handle missing data lies in the fact that, when $\mathbf{c}(i,:)$ is calculated, if the row $\mathbf{d}(i,:)$ contains missing data, the pure matrix $\mathbf{S}^{\mathrm{T}}$ can be accommodated to this situation by using only the matching columns available in $\mathbf{d}(i,:)$ for the calculation in Eq. 24. Likewise, when $\mathbf{s}^{\mathrm{T}}(:,j)$ is calculated, if the column $\mathbf{d}(:,j)$ contains missing data, the pure matrix **C** can be accommodated by using only the matching rows available in $\mathbf{d}(:,j)$ for the calculation in Eq. 25. This procedure allows calculating **C** and $\mathbf{S}^{\mathrm{T}}$ even if **D** contains missing values using a single factorization.

Figure 49 illustrates the application of this single factorization approach to analyze an incomplete multiset as the one in Figure 48. Here, the least squares step starts by calculating the **C** matrix row by row. It can be seen that all the calculations regarding the submatrix $\mathbf{C}_1$ correspond to the use of rows [$\mathbf{D}_1$,$\mathbf{D}_2$], where no missing values are encountered and the full matrix $\mathbf{S}^{\mathrm{T}}$ (left plot in Figure 49A). However, when the first row of $\mathbf{C}_2$ is calculated, the related row of $\mathbf{D}_3$ contains missing values. Therefore, the least squares step to recover the rows of $\mathbf{C}_2$ uses only the corresponding information from the rows of $\mathbf{D}_3$ and $\mathbf{S}_1{}^{\mathrm{T}}$, as shown in Eq. 26.

$$\mathbf{c}(i,:) = \mathbf{d}(i,:)\left(\mathbf{S}_1{}^{\mathrm{T}}\right)^{+} \qquad \text{Eq. 26}$$

$$\mathbf{s}^{\mathrm{T}}(:,j) = \mathbf{C_1}^{+}\mathbf{d}(:,j) \qquad \text{Eq. 27}$$

Similarly, Figure 49B illustrates the least-squares calculations performed column-by-column to compute the complete **S**$^{\mathrm{T}}$ matrix. In this case, the columns of **S**$_1^{\mathrm{T}}$ are computed using the columns of [**D**$_1$;**D**$_3$], with no missing values and the full matrix **C** (left plot in Figure 49B). However, the calculation of the full columns of the **S**$_2^{\mathrm{T}}$ block relies only on the available column information in the **D**$_2$ block and the corresponding information in the **C**$_1$ block, as shown in Eq. 27.

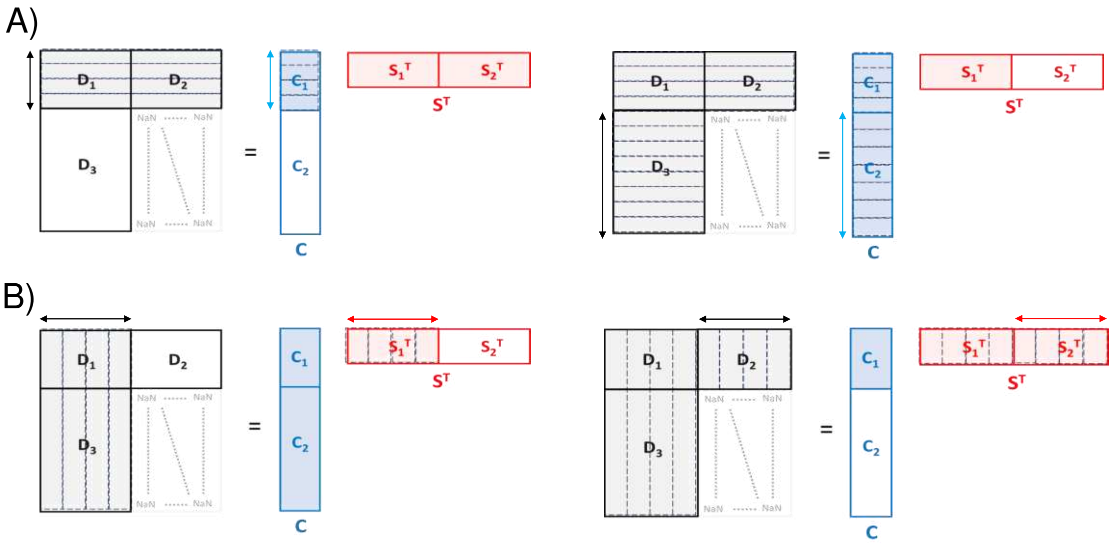


Figure 49. MCR analysis of an incomplete multiset with a single factorization model. a) the rows of **C** are calculated using the full **S**$^{\mathrm{T}}$ matrix. When a missing value in the rows of **D** is encountered, **S**$^{\mathrm{T}}$ is accommodated to match the same available entries than the selected row from **D**. b) The columns of **S**$^{\mathrm{T}}$ are calculated using the full **C** matrix and, as before, when a missing value is encountered in the columns of **D**, **C** is accommodated.

This new methodology to analyze incomplete multisets is able to provide a complete bilinear model independently of the quantity and structure of missing information in **D**, without data imputation and in a single factorization step. This advantageous row-by-row and column-by-column least squares calculation was first proposed by Beyad and Maeder in 2013 to deal with the presence of scattered missing values linked to saturated signals in regression problems, but was never envisioned for the MCR analysis of incomplete multisets, with full blocks of missing information.

The MCR single factorization approach applied to image fusion problems

Analysis of simulated incomplete multisets

The simulated dataset shows a case of image fusion of emission fluorescence and Raman HSIs, where the sample scanned areas are not totally coincident among platforms. Therefore, the non-common areas are added to the corresponding part of the multiset, generating an incomplete structure (Fig. 50).

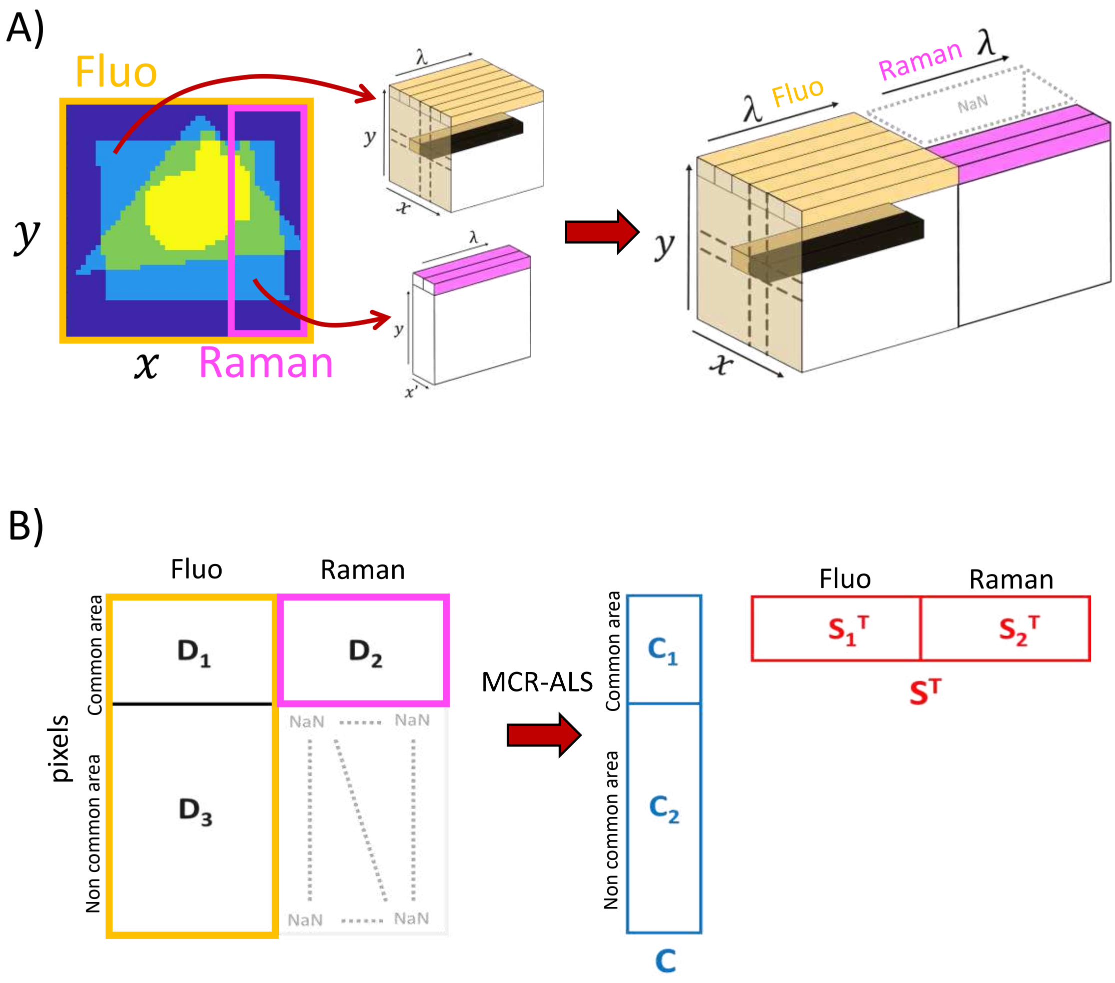


Figure 50. A) Left, simulated sample with three components and scanned areas for fluorescence (yellow) and Raman (pink) images. Right, the fluorescence and Raman hyperspectral images are concatenated in a row-wise direction. Since not all fluorescence spectra have the corresponding Raman spectra, a part of the data is empty (*NaN*, in dashed gray lines). B) Incomplete multiset with the MCR bilinear model.

In this simulation, the fluorescence image covers the entire sample area (yellow), while the Raman image covers a smaller region (pink), having a common area measured by both techniques (Fig. 50A). The incomplete multiset is then built by concatenating the fluorescence and Raman spectra of the common area [$\mathbf{D}_1$,$\mathbf{D}_2$], and appending the fluorescence pixels of the non-common area D3 to D1 to provide the incomplete multiset [**D**1,**D**2;**D**3,*NaN*] (Fig. 50B, right). This incomplete multiset contains around 30% of missing entries. The pure components used in this simulation are shown in Fig. 51. Poisson noise was added to mimic the real conditions of the measurements representing around the 7% of the total signal.

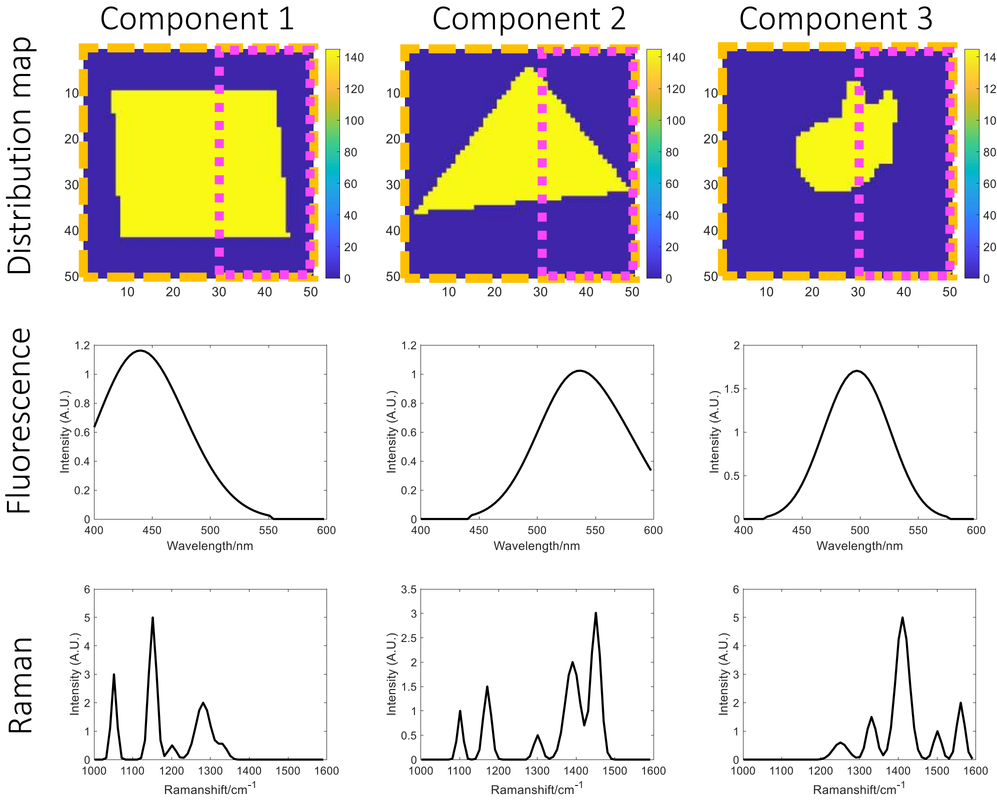


Figure 51. Dataset simulated with three components to study the MCR-ALS analysis of incomplete multisets. Top plots: pure distribution maps. In yellow dashed line, scanned area of fluorescence image; in pink dashed line, scanned area of Raman. Middle plots: pure fluorescence signatures. Bottom plots: pure Raman signatures.

This simulated dataset serves to show three realistic scenarios that may happen in image fusion: case 1 assumes that all components have signal on all techniques, case 2 assumes that component 3 has null Raman signal and case 3 assumes that component 3 has null fluorescence signal.

In all MCR analyses, initial estimates were obtained using a SIMPLISMA-based approach on [**D**1,**D**2], selecting three components and a convergence criterion of 10^{-10}%. Non-negativity constraints were applied to both C and S^T matrices.

The lack of fit in all MCR analyses corresponded to the noise level added in the simulations. In Fig. 52-54, the resolved pure maps and spectral profiles by MCR-ALS for cases 1, 2 and 3, respectively, are shown. The analysis by MCR of case 1 showed that both spectra and maps perfectly match with the true simulated solution, with correlation coefficients higher than 0.99 between the

MCR results and the true solution. This shows the capacity of MCR-ALS to deal with missing data when the row-by-row and column-by-column least squares calculation is used, despite the significant amount of missing entries.

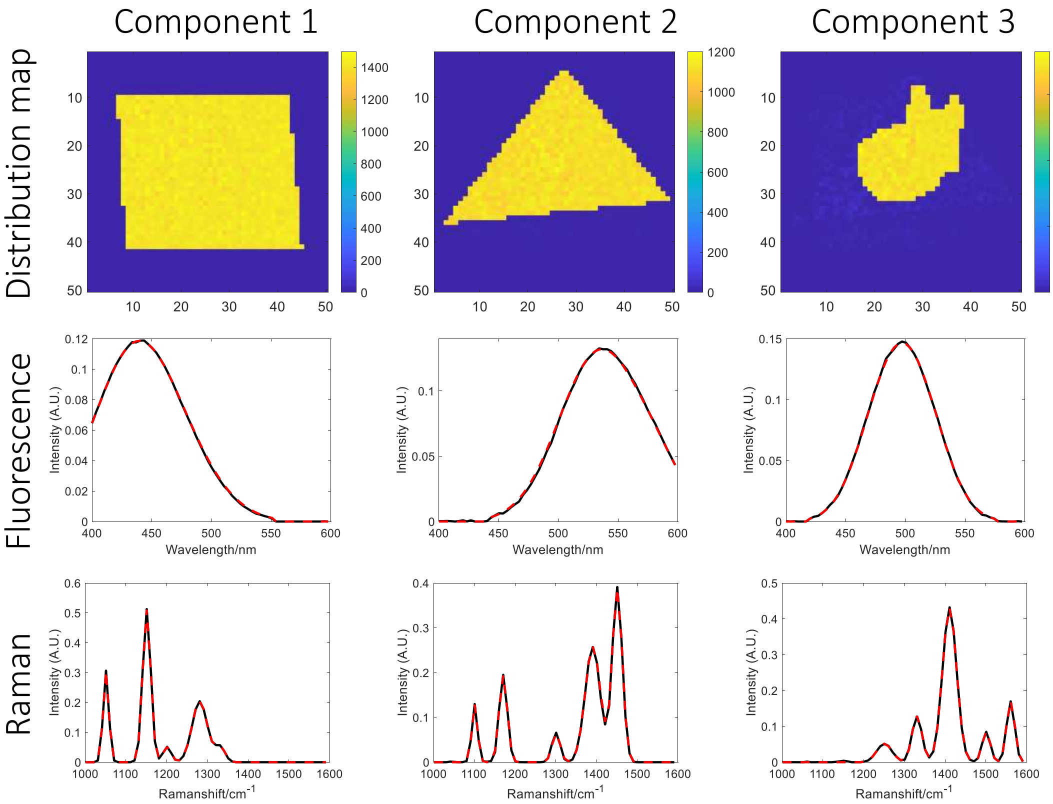


Figure 52. MCR results on the simulated case 1. Top plots: pure distribution maps. Second row plots: fluorescence spectra (red dashed lines: simulated profile shape; black lines: MCR spectra. Third row plots: Raman spectra (red dashed lines: simulated profiles; black lines: MCR spectra).

Similar results are provided for case 2 (Fig. 53), with all resolved profiles matching perfectly with the true simulated solution (correlation coefficients > 0.99). The presence of a null signal in Raman for one component does not pose any problem in this case. This is due to the connection of all Raman pixel spectra with a fluorescence signal in the multiset. Hence, the map for the component with null Raman signal is, nevertheless, well described.

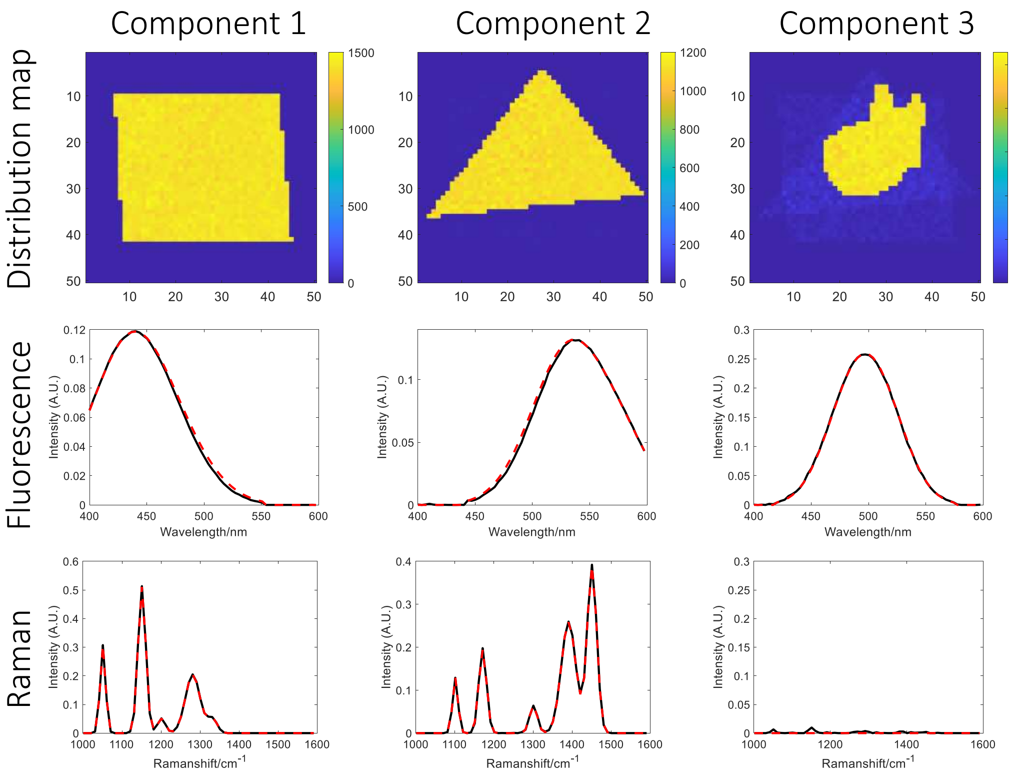


Figure 53. MCR results on the simulated case 2. Top plots: pure distribution maps. Second row plots: fluorescence spectra (red dashed lines: simulated profile shape; black lines: MCR spectra. Third row plots: Raman spectra (red dashed lines: simulated profiles; black lines: MCR spectra).

Figure 54 shows the resolved pure maps and pure spectral profiles by MCR-ALS for case 3. The pure spectra for all compounds are well recovered with correlation coefficients above 0.99, and the maps for components 1 and 2 show the correct distribution shapes. However, something interesting happens with the distribution map of component 3. It is possible to observe that there is a region not well-defined, which corresponds to the region where the Raman signal has not been measured. If we consider the true pure profiles, it is possible to see that component 3 has not been detected outside the common scanned area because it does not present fluorescence signal. Therefore, only on the region with non-null signal measured, i.e., the region where Raman provides signal, the distribution map is correctly recovered.

This limitation is inherent to the information available in the initial dataset, and no algorithm can recover information for a component with a null signal unless it is connected to another technique compensating for this lack of information.

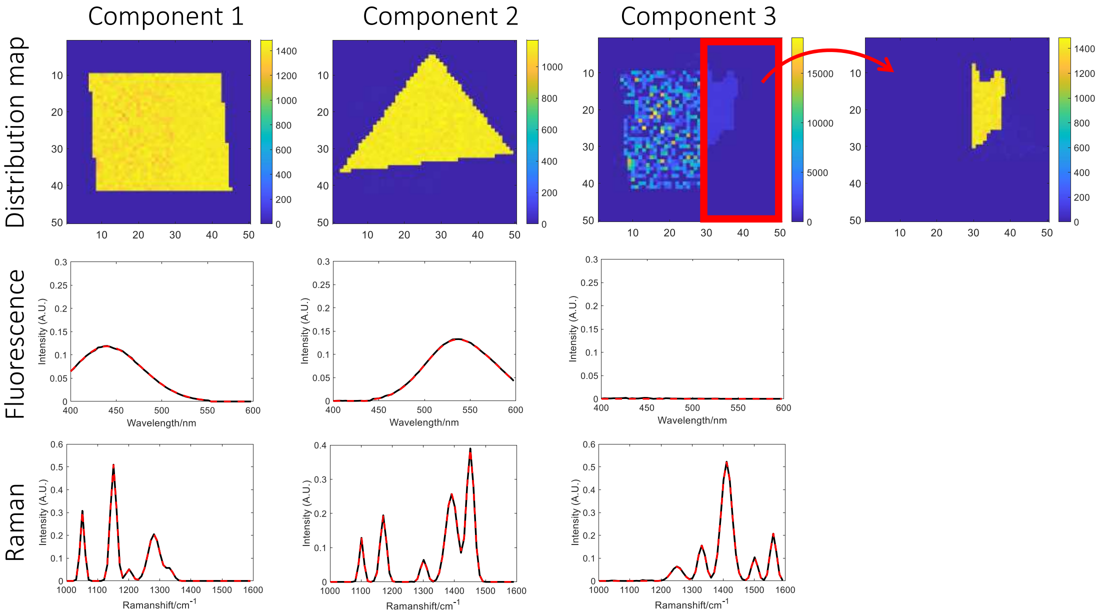


Figure 54. MCR results on the simulated dataset of case 3. Top plots: pure distribution maps. For component 3, the part without the undefined part is also show for a clear illustration of the resolved part of this component. Second row plots: fluorescence spectra (red dashed lines: simulated profile shape; black lines: MCR spectra. Third row plots: Raman spectra (red dashed lines: simulated profiles; black lines: MCR spectra).

Analysis of a real incomplete multiset

In real image fusion scenarios, the complexity increases even more than in the previous simulated case. This happens when the imaging techniques have different spatial resolution and different scanned area. As stated in subsection 3.4, building a complete multiset in such cases involves restricting the analysis to the common sample area covered by the different platforms and downsampling higher-resolution images through pixel binning to match the pixel size of the imaging technique with the worst resolution. That was the case of the example shown in **Publication VI** with fluorescence and Raman HSIs, matching the scanned area and infrared pixel size. In this same scenario, an incomplete multiset approach retains all available information, allowing the inclusion of pixel spectra of non-common scanned areas and the pixel spectra with highest resolution as additional blocks. Figure 55 illustrates this strategy for the real case of image fusion using Raman and NIR imaging.

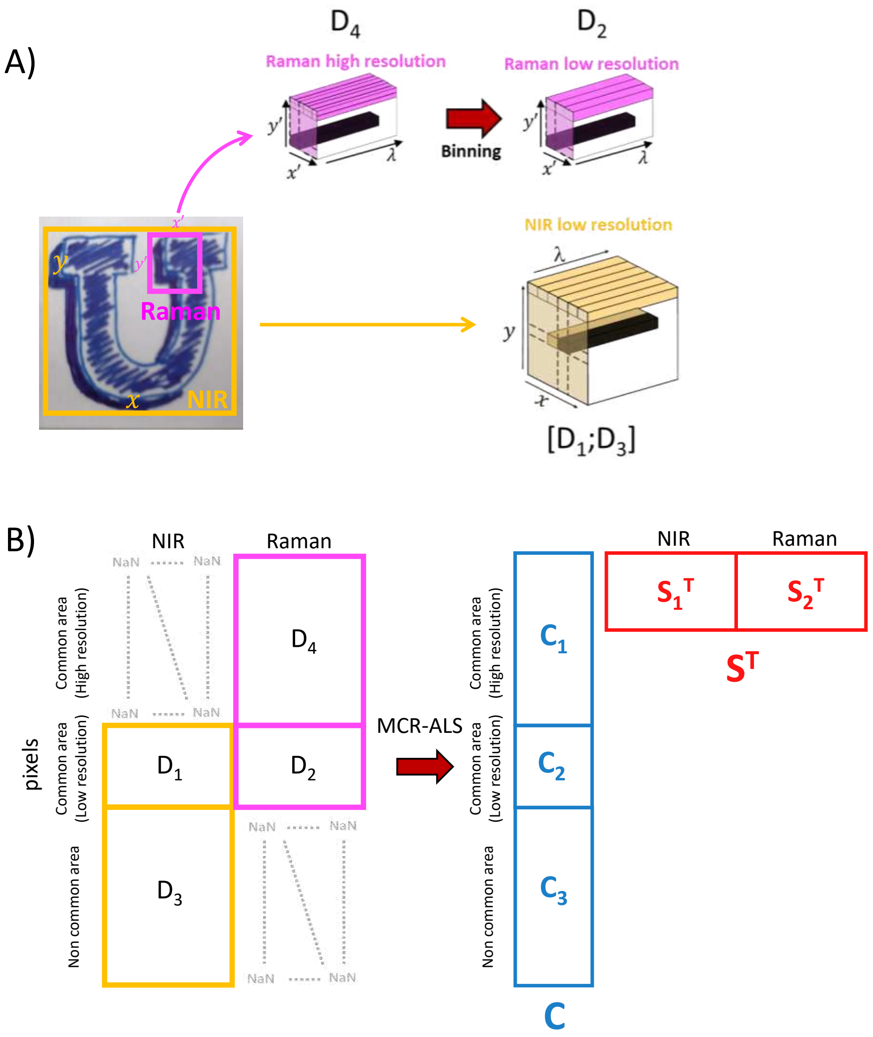


Figure 55. Real example of the construction of an incomplete multiset where different scanned areas and spatial resolution exist among imaging platforms. A) On the left, a NIR image of a sample formed by different inks is acquired, covering the full sample area (yellow), while the Raman measurement only considers a small region and having a smaller pixel size (pink). B) Incomplete multiset structure and related bilinear model for the NIR/Raman image fusion. The Raman image ($\mathbf{D}_4$) is binned ($\mathbf{D}_2$) to match the pixel size of the common area scanned by the NIR image ($\mathbf{D}_1$). The NIR pixels of the non-common scanned area ($\mathbf{D}_3$) are concatenated below $\mathbf{D}_1$, while the Raman image with the original spatial $\mathbf{D}_4$ resolution is concatenated above $\mathbf{D}_2$.

Figure 55A shows a real sample, where three different blue pens (Pilot V-Ball Grip, PVG; Uniball Signo, US; Bic Velocity, BV) write the letter "U" on paper. The PVG pen is used to fill the inner part of the letter U. The US pen is used to delimitate the inner part, while the shadow was drawn by the BV pen. This simple sample allows controlling the number, identity and the spatial distribution of the components, facilitating a comparison between the true solution and the MCR-recovered pure profiles. The incomplete multiset is built as in Fig. 55B.

Before going into the image fusion results, a brief description of the results obtained by the individual MCR analyses using NIR and Raman images is provided. The distribution maps and pure spectra for the NIR image were not very well defined because the pure spectra of two of the inks and the paper are very similar. In contrast, the analysis of the Raman image allowed a perfect separation of components, but the area scanned by this technique was very small (for more detail on the results, see related figures in **Publication VIII**). To summarize, in none of the cases the results provided by the individual techniques were fully satisfactory.

The incomplete multiset, shown on Fig. 55 was analysed by MCR-ALS. Initial estimates were derived through a SIMPLISMA-based approach on [$\mathbf{D}_1$,$\mathbf{D}_2$], with a convergence criterion set at 10^{-10}%. and non-negativity constraints applied to the concentration profiles **C**. Upon analyzing this incomplete multiset, the positive impact of the complementary Raman and NIR information on the results for the entire scanned sample area becomes evident. Figure 56 displays the distribution maps (top plots) and the resolved spectral signatures (bottom plots) from the MCR analysis of this multiset, which provided a model with 9.5% of lack of fit. The pure spectral signatures, representing connected NIR/Raman pure spectra for the different components, match very well the known pure ink spectra. Moreover, the distribution maps from the full scanned sample area (defined by blocks [$\mathbf{C}_2$;$\mathbf{C}_3$] in Figure 55) show correctly the regions of each ink and the presence of paper. It is important to highlight that the well-defined maps result not only from the small area scanned by the Raman image but from the entire scanned area by NIR, thanks to the incorporation of fused NIR/Raman information. The distribution maps related to the high-resolution block $\mathbf{C}_1$ are identical to those obtained in the individual Raman image analysis (not shown).

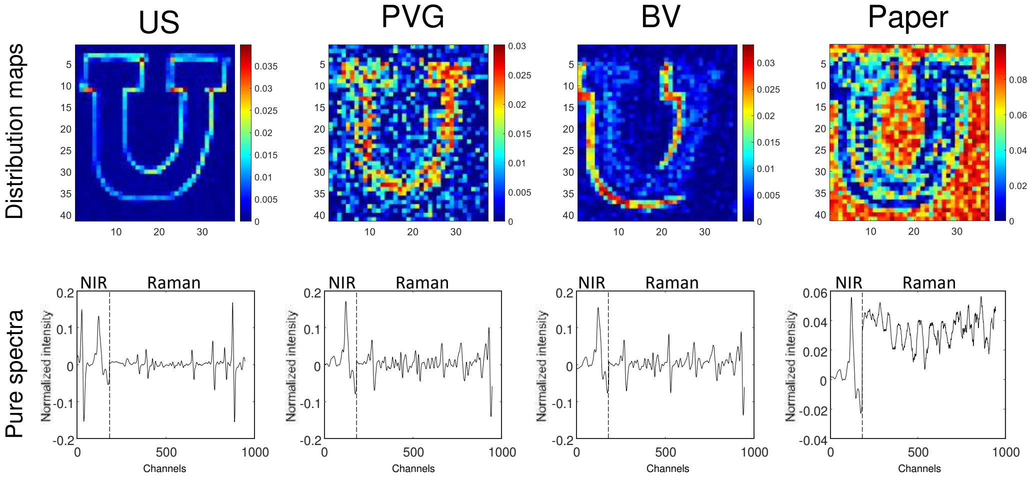


Figure 56. MCR results for the incomplete multiset of inks. Top plots: pure distribution maps. Bottom plots: resolved pure spectra signatures (fused NIR and Raman).

These results show the power of the new variant of MCR-ALS to provide a single and accurate bilinear model even in the presence of a large amount of block missing information.

The modification of the least squares steps to operate in a row-by-row and column-by-column modes is fully compatible with the use of all available constraints in MCR and also for the new trilinearity constraint implementations explained in this thesis in **Publication II**, aimed to analyze trilinear data with missing values, and **Publication VII**, aimed to provide hybrid bilinear-trilinear models. The *modus operandi* of all proposed approaches allows MCR to become a modular algorithm easily scalable to analyze complex data sets in a single multiset structure, where different constraints and models can simultaneously be used. A recent example of scalability is the work published by Queral-Beltran et al. (2024), where an incomplete multiset structure was analyzed by implementing the partial trilinearity constraint of **Publication VII**.

PCA adaptation to the missing block problem

The analysis of incomplete multisets is also a challenge for exploratory methods like PCA. PCA is a widely employed bilinear decomposition method in multivariate data analysis, where the dataset is described by a set of orthogonal scores **T** and loadings **P**. Its main applications are exploratory purposes [Mas et al., 2010; Amigo et al., 2013] and data classification [Wold and Sjöström, 1997]. The original algorithms to perform PCA were not designed to handle missing data and some strategies were proposed to solve this problem. Two adaptations of PCA to work with missing data are the Nonlinear Iterative Partial Least Squares (NIPALS) [Wold, 1975] and the Imputation based on SVD (I-SVD) [Grung and Manne, 1998; Walczak and Massart, 2001]. These two algorithms are briefly explained to understand better the differences with the new Orthogonalized Alternating Least Squares approach.

Non-Iterative Partial Least Squares (NIPALS)

NIPALS is a method that performs a sequential calculation of the principal components using an alternating least squares optimization of the scores and loadings. NIPAS calculates the first principal component through ordinary ALS, followed by a deflation step to remove its variance. Additional components are then extracted through iterative ALS on the deflated matrix and the component extraction/deflation cycle is repeated until reaching the desired rank. To deal with the missing data, NIPALS uses the row-by-row and column-by-column least squares calculation skipping the missing values during the ALS calculations [Christoffersson, 1970; Grung and Manne, 1998]. This algorithm has been thoroughly described in **Publication II**, where it is used to implement the trilinearity constraint in the presence of missing data. However, NIPALS is unsuitable for data sets with missing values and rank higher than one and provides biased non-orthogonal scores and loadings.

Imputation based on SVD

PCA models are often calculated with the SVD algorithm. Nevertheless, as highlighted in **Publications I** and **II**, SVD cannot handle missing data unless the algorithm is appropriately modified. To solve this problem, a proposed solution is the use of an iterative imputation algorithm based on SVD, or I-SVD. This algorithm relies on applying the SVD algorithm after imputing the missing values with estimates. These estimates are subsequently updated using the prediction of the model itself, until convergence. While I-SVD yields orthogonal scores and loadings, the computational cost associated with the method is exceptionally high for datasets with a significant number of missing entries following a non-random pattern and convergence is sometimes not achieved.

Orthogonalized alternating least squares (O-ALS)

In this work, the Orthogonalized Alternating Least Squares (O-ALS) algorithm is introduced to address the problem of missing values in PCA. In comparison with NIPALS and I-SVD, O-ALS has the ability to obtain orthogonal scores and loadings in a short computation time, irrespective of the percentage and pattern of the missing entries on the dataset.

Essentially, O-ALS works performing a row-by-row and column-by-column alternating least squares estimation of the scores and loadings subject to a Gram-Schmidt orthogonalization constraint. The ALS calculation of scores and loadings mimics exactly the row-by-row and column-by-column least squares procedure of **Publication VIII**. The key difference relies on the subsequent application of Gram-Schmidt orthogonalization as a constraint instead of non-negativity or other MCR-related constraints. The Gram-Schmidt orthogonalization, named after mathematicians Jørgen Pedersen Gram and Erhard Schmidt, is an algorithm used to transform a set of linear independent vectors into a set of orthogonal vectors spanning the same space. The idea behind O-ALS is to use Gram-Schmidt orthogonalization on **T** and **P** profiles after their calculation. The steps of O-ALS are defined below:

Step 1) *Generation of initial estimates*. Start with initial estimates for the loadings (P) or scores (T) matrices.

Step 2) *Iterative ALS estimation of scores and loadings under the Gram-Schmidt orthogonalization constraint*.

Step 2A: *Row-by-row Least Squares calculation of scores (LS).* For each data row of D, $\mathbf{d}(i,:)$, the corresponding scores row $\mathbf{t}(i,:)$ is calculated by least squares (see Eq. 28). This calculation skips missing values by adapting the loadings matrix to use only the same columns as in the available entries in the row of **D**. After all rows are calculated, the Gram-Schmidt orthogonalization constraint is applied to **T** to preserve orthogonality.

$$\mathbf{t}(i,:) = \mathbf{d}(i,:)\left(\mathbf{P}^{\mathrm{T}}\right)^{+} \qquad \text{Eq. 28}$$

Step 2B: *Column-by-column Least Squares calculation of loadings (LS).* For each data column $\mathbf{d}(:,j)$, calculate the related loading column $\mathbf{p}^{\mathrm{T}}(:,j)$ using a least-squares process (see Eq. 29). This calculation skips missing values by adapting the scores matrix to use only the same rows as in the available entries in the column of **D**. After all columns are calculated, the Gram-Schmidt orthogonalization constraint is applied to **P** to preserve orthogonality.

$$\mathbf{p}^{\mathrm{T}}(:,j) = \mathbf{T}^{+}\mathbf{d}(:,j) \qquad \text{Eq. 29}$$

Repeat steps 2A and 2B iteratively until convergence is achieved and the final scores matrix (**T**) and loading matrix (**P**) are obtained.

Comparison of NIPALS, I-SVD and O-ALS on simulated datasets

The basic dataset used is very similar to the simulated Case 1 (Fig. 51) of **Publication VIII**. The conditions tested are different noise levels (noise-free or 7% noise), patterns of missing data (random or missing block entries) and percentages of missing data (5%, 30%, and 80%). The analyses were performed under a convergence criterion of $10^{-13}\%$ to prevent premature termination of the algorithms. Random values were chosen as initial estimates for all analyses. The results are summarized in Tables 3 and 4.

Table 3. Correlation coefficients among recovered scores and loadings and the analogous simulated profiles (from the complete matrix) for the random missing pattern case.

				Noiseless			Noise 7%		
Missing pattern	Algorithm	Component		Percentage of missing data			Percentage of missing data		
				5%	30%	80%	5%	30%	80%
Random	NIPALS	1	Score	1.0000	0.9999	0.9998	1.0000	0.9998	0.9996
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		2	Score	1.0000	0.9998	0.9993	0.9999	0.9980	0.9966
			Loading	1.0000	1.0000	0.9999	1.0000	0.9999	0.9997
		3	Score	1.0000	1.0000	0.9996	0.9999	0.9937	0.9909
			Loading	1.0000	1.0000	0.9999	1.0000	0.9984	0.9982
	I-SVD	1	Score	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		2	Score	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		3	Score	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	O-ALS	1	Score	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		2	Score	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		3	Score	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

In the case of randomly distributed missing entries (Table 3), NIPALS, I-SVD, and O-ALS showed a very good performance, with all correlation coefficients >0.99. Both I-SVD and O-ALS showed an exceptional performance, achieving perfect matches with the true scores and loadings across all scenarios, even in cases with high percentages of missing data. Interestingly, it is possible to observe that the percentage of missing data in a random pattern does not significantly affect the quality of the solutions provided by the I-SVD algorithm, a conclusion that also applies to O-ALS. However, NIPALS solutions showed degradation with increasing percentages of missing data, particularly in noisy scenarios, although they still remained close to the true solutions in the absence of noise.

Table 4. Correlation coefficients among recovered scores and loadings and the analogous simulated profiles (from the complete matrix) for the missing block pattern case.

				Noiseless			Noise 7%		
Missing pattern	Algorithm	Component		Percentage of missing data			Percentage of missing data		
				5%	30%	80%	5%	30%	80%
Missing block	NIPALS	1	Score	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	0.9997	0.9996	0.9993	0.9997	0.9996	0.9993
		2	Score	0.9967	0.9915	0.9829	0.9966	0.9914	0.9829
			Loading	0.9992	0.9976	0.9921	0.9992	0.9976	0.9921
		3	Score	0.9836	0.9835	0.9802	0.9826	0.9799	0.9738
			Loading	0.9987	0.9866	0.9600	0.9985	0.9861	0.9587
	I-SVD	1	Score	1.0000	1.0000	-*	1.0000	1.0000	1.0000*
			Loading	1.0000	1.0000	-*	1.0000	1.0000	1.0000*
		2	Score	1.0000	1.0000	-*	0.9999	0.9996	0.9994*
			Loading	1.0000	1.0000	-*	1.0000	1.0000	0.9998*
		3	Score	1.0000	1.0000	-*	1.0000	0.9956	0.9936*
			Loading	1.0000	1.0000	-*	0.9983	0.9998	0.9986*
	O-ALS	1	Score	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		2	Score	1.0000	1.0000	1.0000	0.9999	0.9996	0.9994
			Loading	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998
		3	Score	1.0000	1.0000	1.0000	0.9983	0.9956	0.9936
			Loading	1.0000	1.0000	1.0000	1.0000	0.9998	0.9986

*NIPALS solutions were used as initial estimates instead random values due to the impossibility to converge of the I-SVD algorithm.

Regarding the results of the algorithms for data with missing block patterns (Table 4), I-SVD showed excellent performance in the absence of noise for cases of 5% and 30% missing data. However, I-SVD failed to converge within a reasonable time (<6 hours) for the 80% missing data case. To improve the convergence of I-SVD, the scores and loadings provided by NIPALS were used as initial estimates. Even then, I-SVD did not converge in a reasonable time for the noiseless case. In contrast, the O-ALS algorithm consistently provided excellent results across all cases. The solutions provided by NIPALS were always worse than for the two previous algorithms and the distortion of scores and loadings became more evident as the index of the component calculated increased.

Presence of bias in the recovered scores and loadings

The NIPALS algorithm showed a systematic bias in the recovered scores and loadings. The bias is practically inappreciable for the random pattern of missing values and becomes evident for the data sets with a systematic pattern of missing entries (Fig. 57). However, the presence of small deviations in the analysis even in an ideal scenario (low rank space and no noise) is a hint that NIPALS is not addressing the missing data problem correctly, although from a practical point of view, the errors may be negligible in some cases.

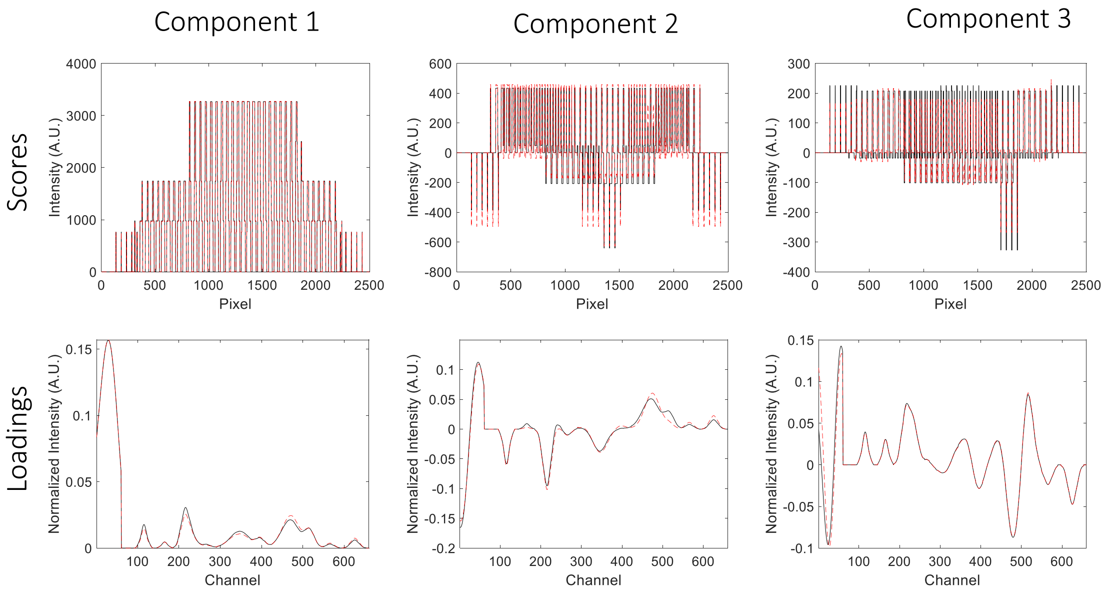


Figure 57. Scores and loadings extracted by NIPALS (dashed red line) and scores and loadings of the complete matrix (black line) for the missing block case with 80% of missing entries and no noise. Small deviations can be seen in the first principal component, and significant differences in the second and third.

The bias of NIPALS when dealing with missing data was highlighted by Bjørn Grung and Rolf Manne in 1998. This error comes from the *per-component* calculation inherent to NIPALS, where every component is calculated at a time, unlike I-SVD and O-ALS, which estimates the *N* components that define the rank of the data matrix simultaneously.

To understand this fact, consider the computation of the first principal component using NIPALS both in the absence and presence of missing data. In a complete matrix, the first principal component is computed using an ALS approach. However, in an incomplete matrix, deriving the first principal component by skipping the missing entries is equivalent to replace the missing data with values predicted from the PCA model. In other words, the NIPALS solution is equivalent to impute missing data with predictions that belong to a rank-1 space, while the missing values of the original matrix are present in a space with different rank. Consequently, NIPALS essentially imputes missing data using a rank-1 model, leading to a bias. This error becomes more evident with subsequent component extractions due to the cumulative errors in each deflation step and explains why the error increases when the number of missing entries increases (more missing values are incorrectly imputed) and why the subsequent components are more biased (error accumulation). Please note that the use of NIPALS in **Publication II** is justified, since the case study involves a matrix of rank 1 and, therefore, the calculation is correct.

Effect of the pattern and the percentage of missing data into the obtained PCA model.

In the context of PCA, the rank of the data matrix defines its underlying structure, with the assumption that this structure can be captured by a lower-dimensional representation without information loss. When dealing with missing data, I-SVD and O-ALS perform the suitable calculations using the real rank of the data, *N*. Essentially, both methods span the same dataspace. Therefore, an algorithm that accurately estimates missing values will yield correct scores and loadings, and vice versa.

In scenarios where the available information adequately defines the rank-N basis of the incomplete data set, both I-SVD and O-ALS achieve perfect scores and loadings, irrespective of the missing pattern or percentage of missing values in the data set. This kind of results identifies the information encoded on the available data as the key factor over the missing patterns or the percentage of missing entries to ensure a correct solution. Thus, when the available information in incomplete data is enough to define the dataspace, perfect scores and loadings are obtained, as shown in Tables 3 and 4 and Fig. 58 for noiseless cases.

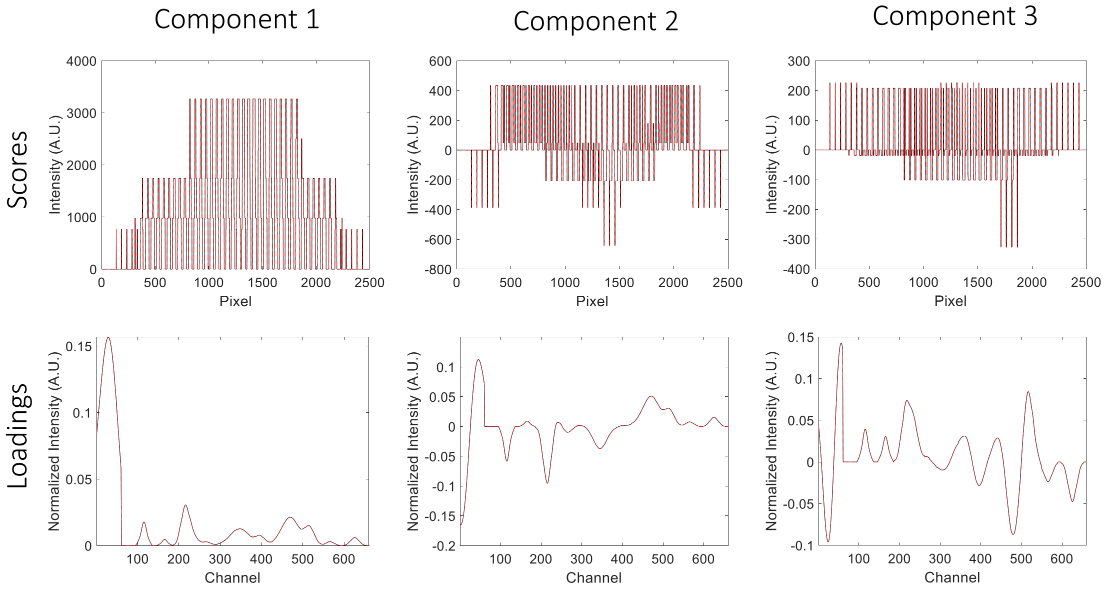


Figure 58. Scores and loadings extracted by O-ALS and I-SVD (dashed red line) and scores and loadings of the complete matrix (black line) for the missing block case with 30% of missing entries. Perfect match between calculated and simulated profiles is achieved for both algorithms.

Effect of missing values on the algorithm convergence

I-SVD has limitations due to its algorithmic nature. During the iterations, the provisional imputed values tend to move the scores and loadings from the correct data space. However, across iterations, the imputed values improve to get closer to the correct dataspace in each iteration. Therefore, when there is a high percentage of missing data and they follow a missing block pattern, the imputed values cannot be easily estimated and have a strong influence in the model to be optimized, increasing the convergence time up to hours or more. This can be observed in Table 4, where for 80% missing data in the noiseless case, I-SVD failed to converge even when initialized with the solution provided by NIPALS, which is a reasonably good initial estimate. In the presence of noise, the solutions provided by NIPALS had to be used as initial estimates for I-SVD to achieve convergence within a reasonable time (hours). On the other hand, O-ALS achieved correct solutions in a computation time significantly lower. To illustrate this effect, the datasets with a 7% noise added following a random and a missing block pattern were analyzed 500 times by O-ALS and I-SVD using different random initial estimates in each run (Table 5).

Table 5. Mean and standard deviation of the time spent per each algorithm per case in 500 runs for the 7% noise-added case. Random values were used as initial estimates.

Algorithm	Random pattern			Block pattern		
	5%	30%	80%	5%	30%	80%
I-SVD	10±1s	27±3s	72±2s	114s±12s	17±1 min	-
O-ALS	4±0.8s	2.5±0.6s	2.2±0.5s	17±6s	20±21s	19±11s

On the other hand, a limitation of the O-ALS algorithm is the possibility to fall into local minima, particularly when analyzing multisets with missing block patterns. From 500 runs, a considerable number of them found possible local minima, especially pronounced in scenarios with higher percentages of missing data with non-random pattern. No local minima were found for the analysis of data with random missing patterns. Ilin and Raiko (2010) previously warned about the presence of local minima for bilinear factorization of datasets in the presence of missing entries. In this scenario, the O-ALS algorithm offers a straightforward solution. The algorithm can undergo multiple initializations within a few iterations, employing different initial estimates each time. Subsequently, the initial estimates yielding the most accurate fit are chosen and used for the analysis until convergence is achieved. This approach enables the extraction of the correct PCA model without significantly extending the computation time. For instance, in an extreme case with 80% missing data and a missing block pattern, employing the best 10 initializations across 35 iterations reduces the frequency of encountering local minima from 32% to less than 1%, and the increase in analysis time goes from 19 seconds to 120 seconds.

Results of a real case study

The O-ALS algorithm was applied to a real incomplete multiset with 70% missing values. This dataset was built by the NIR and Raman hyperspectral images from **Publication VIII**, where NIR covers all the sample area and Raman covers only a small amount of the sample area, as in Fig. 55. In this case, to simplify the analysis, the Raman hyperspectral image with the original high spatial resolution was not included.

A convergence criterion of 10^{-13}% was used to ensure that O-ALS properly converges. Ten runs were conducted to mitigate potential local minima, with initial estimates chosen randomly for four components. The run showing best fit was chosen as final solution.

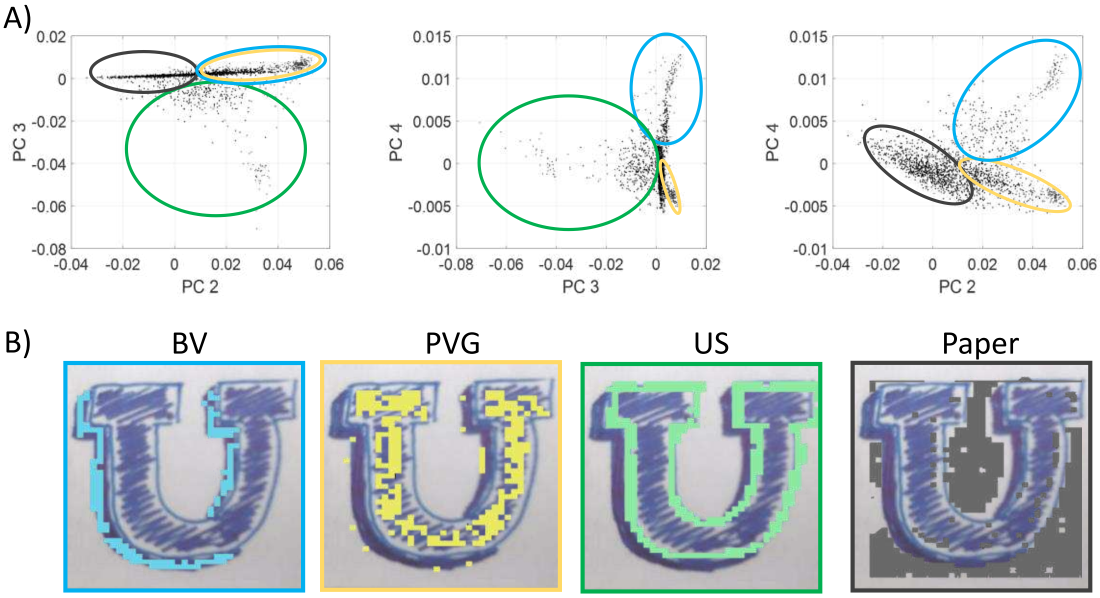


Figure 59. A) Scores extracted by O-ALS. Different clusters related to the inks can be observed. In black, the cluster related to the paper. In green, pixels related to US signal. In blue, pixels related to BV signal. In yellow, pixels related to PVG signal. B) Clusters selected (blue, green, yellow and cyan) showing the pixel position on the original image.

The results are shown in Fig. 59. In the PCA space, four components were clearly detected. The percentage of variance explained by each principal component was 62.48%, 25.87%, 4.55%, and 0.47% for PC 1, 2, 3, and 4, respectively. In total, the model accounted for 93.37% of the variance. Since the data were not mean-centered, the first principal component was not shown, since it reflects basically the mean of the data without highlighting informative clusters. In the PC 3 vs PC 2 plot, the paper (black circle) and the US pen (green circle) can be differentiated from the rest of the signal (see Fig. 59B for the corresponding pixel positions within the scanned area). On the other hand, the PC 4 vs PC 3 plot differentiates the signals corresponding to the US pen (green circle) and BV (blue circle). Additionally, it can be slightly observed the cluster of PVG signal (yellow circle). Lastly, the PC 4 vs PC 2 plot separates the PVG signal from the rest of the data (yellow circle). Therefore, the analysis suggests that O-ALS effectively captured the principal component dataspace, showing the four expected clusters related to the signal from the three pens and the paper despite the huge amount of missing entries.

Summarizing, the study has examined the performance of NIPALS, I-SVD, and a novel algorithm, the O-ALS, for PCA in datasets with missing values. The results indicate that I-SVD and O-ALS consistently provide accurate results across various scenarios, even with a high percentage of missing data, maintaining their ability to recover the true scores and loadings. On the other hand, special care must be taken into account when applying NIPALS to datasets containing missing values, since the PCA models provided are biased, especially when the percentage of missing entries is high and for systematic patterns. According to the results, the pattern and percentage of missing values do not significantly affect the performance of I-SVD and O-ALS. Additionally, the study explores the convergence behavior of the two algorithms, highlighting that I-SVD struggles with high percentages of missing entries, while O-ALS provides PCA models within reasonable times. However, further research is needed to investigate the occurrence of local minima in I-SVD and O-ALS solutions, since this question is still unclear.

CHAPTER 4. CONCLUSIONS

The conclusions of this thesis are divided into two main blocks, related to the conclusions drawn from the results presented in Section I and Section II of Chapter 3.

Addressing challenges of fluorescence images

1. Several algorithms have been proposed to enable the use of trilinear models for the analysis of Excitation-Emission Fluorescence (EEM) images in the presence of missing information. Special attention has also been paid to improve the interpretation of the information of fluorescence decay curves, the basic measurement of Fluorescence Lifetime Images (FLIM). To do so, a simple approach to extract the Instrument Response Function from the decay signals measured has been provided. A dedicated method to unmix multiexponential curves into their monoexponential contributions for lifetime estimation based on the trilinear decomposition of the transformed/kernelized original signals has been designed. The latter method is particularly suitable for decay curves with few sampling points.

2. Trilinear decomposition of Excitation-Emission Fluorescence (EEM) images by MCR-ALS in the presence of missing data calls for the modification of the current implementation of the trilinearity constraint. In absence of missing data, the constant pure emission spectral shape required in trilinear models for a specific component is obtained by doing SVD on a complete matrix $\mathbf{S}_{\mathbf{fn}}$ of equally sized pure emission spectra obtained at different excitation wavelengths. In the presence of missing data, the matrix of pure emission spectra has missing entries and SVD cannot be straightforwardly used. The first implementation of trilinearity to handle missing data involves combining the results of sequential SVD analyses on complete submatrices of the ragged matrix $\mathbf{S}_{\mathbf{fn}}$ of pure emisision spectra. This implementation can be applied to any systematic pattern of missing values. However, the selection of the submatrices is dataset-dependent and the algorithm may be complex and difficult to implement for higher-order models than trilinear models. The second implementation of the trilinearity constraint proposed works employing NIPALS instead of SVD to obtain the common emission spectral shape. NIPALS allows handling efficiently missing entries for rank-1 matrices. This algorithm can estimate row-by-row and column-by-column the elements of the score vector and the loading vector, respectively adapting the least-squares calculations required to the available information. This implementation is much simpler than the first one proposed and can be applied to any random or systematic pattern of missing entries encountered in EEM data. This NIPALS-based implementation of trilinearity can be easily incorporated in existing MCR-

ALS interfaces, codes and is easily scalable to be used in higher-order models.

3. A fluorophore-specific protocol to study the photobleaching phenomenon in fluorescence images has been presented. The measurement of 4D excitation-emission fluorescence images often requires prolonged acquisition times, where photobleaching may occur. The study of photobleaching has been addressed by the analysis of time-series of consecutive 3D and 4D images by MCR-ALS. This study has benefited from the use of the trilinearity implementation to handle missing data to enable an accurate description of photobleaching in 3D and 4D images, providing distribution maps, pure fluorescence profiles, and decay curves for each resolved fluorophore while ensuring unique solutions. Once photobleaching was characterized, strategies such as selecting few excitation and emission channels while increasing bandwidth detection were applied to acquire a 4D image with limited exposure time, ensuring the comprehensive characterization of all fluorescent compounds. Up to seven compounds related to biological regions of plant tissues were detected in the real example investigated. This approach is particularly effective for complex biological samples and it can be extended to various sample types containing natural or stained fluorescent compounds.

4. The kernelizing approach has been proposed as an efficient method to generate trilinear data arrays from bilinear data matrices of multiexponential decays. The results showed that the trilinear decomposition methods applied on data sets obtained by kernelizing effectively provided the correct decay constants of the underlying components in different scenarios, being particularly useful to handle multiexponential measurements with few sampling points. Despite the positive results, further exploration to optimize the kernel width selection is required. Although tested in time-resolved spectroscopic data and in FLIM mages, kernelizing can be generalized to deal with any analytical measurement expressed by multiexponential decay curves.

5. A novel method for estimating the IRF in TRFS measurements, named Blind Instrumental Response Function Identification (BIRFI) has been proposed. BIRFI requires only the experimental measurement of an exponential decay to estimate the IRF by combining the full signal and its exponential part in a deconvolution process. The method demonstrated to provide accurate predictions of the IRF, improving the accuracy of lifetime determination across different scenarios. BIRFI is offered as a much simpler and accurate way to estimate IRF skipping the direct

measurement of fluorophores with ultrashort lifetimes or of the elastic scattering of the laser pulse for IRF estimation.

6. The proposed algorithms and methodologies presented in Section I of Chapter 4 are oriented to improve the analysis of fluorescence measurements. These improvements have a direct synergy with image fusion, since they contribute not only to a more accurate characterization and extraction of components, but to reduce the rotational ambiguity of the global dataset when fused with other imaging modalities.

Addressing challenges of image fusion

7. The image fusion strategies developed in this thesis were oriented to enable merging measurements from different image platforms taking advantage of all available information, the common among platforms and the specific of every measurement, which can be a different area scanned or the definition of the sample area analyzed with high resolution. The image fusion strategies proposed allow combining image platforms with spectroscopic and spatial differences among them via chemometric variants of MCR-ALS that are prepared to accommodate different underlying measurement models or to handle incomplete multisets with missing blocks of information, originated from the presence of platform-specific information that does not have an equivalence in other image measurements.

8. The description of a case study based on the fusion of fluorescence, Raman, and infrared hyperspectral images for the characterization of constituents of rice leaves cross-sections has been the starting point to present the classical protocol of image fusion based on multiset analysis by MCR-ALS and to identify the advantages and drawbacks of the approach. Comparing the results of individual analyses with the global multiset analysis of the fused images, a higher number of well-characterized components could be obtained due to the joint analysis of the highly diverse spectroscopic information of the multiset. However, the classical image fusion protocol requires a complete multiset where all the information connected among platforms should have the same spatial conditions, i.e., same pixel size and area scanned. These limitations generate some drawbacks that must be considered. First, there is a significant loss of spatial resolution for some of the techniques, which causes that closely located components become indistinguishable and that selective information be lost by the required binning procedures. Second, non-common scanned areas are discarded during analysis, which may lead to the loss of valuable information. The disadvantages

identified triggered the proposal of new methodologies that can overcome the limitations associated with classical image fusion approaches.

9. A challenge to be solved was the design of an image fusion protocol to merge images with different spectroscopic dimensionality, e.g., 3D Raman images and 4D EEM fluorescence images. The methodology had to address the different dimensionality of the data arrays, i.e., an unfolded matrix for a 3D image and a 3D tensor for the 4D image and, most important, the different underlying model of the image measurements, bilinear for Raman images and trilinear for EEM images. An initial multiset could be easily built by unfolding as much as required the 4D image until a data matrix formed by rows with vectorized 2D EEM landscapes was produced and subsequently connected with the Raman image matrix. To model the fused multiset with the suitable underlying image measurements, a partial trilinearity constraint for MCR-ALS was proposed. Thus, the flexible nature of MCR-ALS allowed for the application of the trilinear constraint to the vectorized emission spectra blocks of matrix $\mathbf{S}^T$, while the Raman block was modelled in a bilinear fashion. The hybrid bilinear/trilinear model obtained allowed preserving the natural models of the image measurements and provided more accurate solutions thanks to the reduction of rotational ambiguity induced by the application of trilinearity This methodology was successfully applied to simulated and real datasets based on the measurement of a leaf cross-section sample using Raman and 4D excitation-emission fluorescence images.

10. The last contribution to improve image fusion was focused on the connection of image measurements with spatial differences in terms of spatial resolution and/or area scanned. As mentioned above, the spatial differences among measurements originate incomplete multisets, with missing blocks present. A new MCR-ALS algorithm for incomplete multiset analysis, based on the row-by-row and column-by-column least-squares estimation of matrices **C** and $\mathbf{S}^T$, has been proposed. This simple *modus operandi* allows adapting the least-squares calculations to use only the available entries in each row or column calculation. In comparison with existing algorithms meant for the same purpose, this adapted MCR algorithm does not require any data imputation step and is based on a single factorization model. Besides, it adapts to any percentage and pattern of missing entries. The algorithm was tested in a controlled example were three inks were measured by NIR and Raman images that were showing different spatial resolution and area scanned, providing an incomplete multiset. When the NIR HSI was analyzed individually by MCR-ALS, the results revealed four components, but the

rotational ambiguity associated with the non-selective nature of NIR spectral signatures prevented the correct extraction of pure maps and spectra. Conversely, the MCR results for the individual Raman image showed an excellent recovery of the pure components, but only a small sample area was scanned. The results of the image fusion analysis of the incomplete multiset showed a perfect recovery of all components due to the complementary Raman and NIR information and the correct maps could be defined for the full sample area scanned.

11. The concept of the row-by-row and column-by-column least-squares calculation of the bilinear model in MCR-ALS has also been proposed in a new PCA algorithm for analysis of incomplete multisets, called Orthogonalized-Alternating Least Squares (O-ALS). O-ALS operates performing the alternating row-by-row and column-by-column least-squares calculation of scores and loadings under the Gram-Schmidt orthogonalization constraint. A comparison with PCA algorithms adapted to handle missing data, such as NIPALS and Imputation-SVD revealed better performance of O-ALS in complex scenarios. A first conclusion was that NIPALS suffers from biases in the estimation of scores and loadings due to the sequential component extraction, resulting in a loss of orthogonality among profiles. Instead, O-ALS and I-SVD extract all required components simultaneously and provide good and orthogonal estimates for scores and loadings across diverse scenarios. A clear difference between I-SVD and O-ALS is that I-SVD imputes missing entries, while O-ALS utilizes only available information, being thus unaffected by the amount or pattern of missing entries. This difference causes that I-SVD presents very slow convergence when the amount of missing entries increases and the pattern of missing value sis not random. Instead, O-ALS requires a short computation time in all scenarios tested, but may suffer from local minima in extreme cases. However, this problem can be easily mitigated by using multiple initializations followed by the selection of the best model. The algorithm is clearly promising for image fusion, where missing block patterns are common, and offers a fast and accurate solution for exploratory analysis and potential reconstruction of missing information in these scenarios.

REFERENCES

Adão et al., 2017.
Adão, T., Hruška, J., Pádua, L., Bessa, J., Peres, E., Morais, R., & Sousa, J. J. (**2017**). Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry. *Remote sensing*, 9(11), 1110.

Alcaraz et al., 2019.
Alcaraz, M. R., Monago-Maraña, O., Goicoechea, H. C., & de la Peña, A. M. (**2019**). Four-and five-way excitation-emission luminescence-based data acquisition and modeling for analytical applications. A review. *Analytica Chimica Acta*, 1083, 41-57.

Alier et al., 2011.
Alier, M., Felipe, M., Hernández, I., & Tauler, R. (**2011**). Trilinearity and component interaction constraints in the multivariate curve resolution investigation of NO and O 3 pollution in Barcelona*. Analytical and bioanalytical chemistry*, 399, 2015-2029.

Alier and Tauler, 2013.
Alier, M., & Tauler, R. (**2013**). Multivariate curve resolution of incomplete data multisets. *Chemometrics and Intelligent Laboratory Systems*, 127, 17-28.

Amigo et al., 2008.
Amigo, J. M., Cruz, J., Bautista, M., Maspoch, S., Coello, J., & Blanco, M. (**2008**). Study of pharmaceutical samples by NIR chemical-image and multivariate analysis. TrAC Trends in Analytical Chemistry, 27(8), 696-713.

Amigo, 2010.
Amigo, J. M. (**2010**). Practical issues of hyperspectral imaging analysis of solid dosage forms. *Analytical and bioanalytical chemistry*, 398, 93-109.

Amigo et al., 2013
Amigo, J. M., Martí, I., & Gowen, A. (**2013**). Hyperspectral imaging and chemometrics: A perfect combination for the analysis of food structure, composition and quality. In *Data handling in science and technology* (Vol. 28, pp. 343-370). Elsevier.

Amigo et al., 2015.
Amigo, J. M., Babamoradi, H., & Elcoroaristizabal, S. (**2015**). Hyperspectral image analysis. A tutorial. *Analytica Chimica Acta*, 896, 34-51.

Amigo and Grassi, 2019.
Amigo, J. M., & Grassi, S. (**2019**). Configuration of hyperspectral and multispectral imaging systems. In *Data handling in science and technology* (Vol. 32, pp. 17-34). Elsevier.

Andersen and Bro, 2003.
Andersen, C. M., & Bro, R. (**2003**). Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *17*(4), 200-215.

Appalaneni et al., 2014.

Appalaneni, K., Heider, E. C., Moore, A. F., & Campiglia, A. D. (**2014**). Single fiber identification with nondestructive excitation–emission spectral cluster analysis. *Analytical Chemistry*, *86*(14), 6774-6780.

Bahram et al., 2006.

Bahram, M., Bro, R., Stedmon, C., & Afkhami, A. (**2006**). Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *20*(3-4), 99-105.

Blanchet et al., 2007.

Blanchet, L., Ruckebusch, C., Huvenne, J. P., & de Juan, A. (**2007**). Hybrid hard-and soft-modeling applied to difference spectra. *Chemometrics and Intelligent Laboratory Systems*, 89(1), 26-35.

Bernard et al., 2008.

Bernard, S., Beyssac, O., & Benzerara, K. (**2008**). Raman mapping using advanced line-scanning systems: geological applications. *Applied spectroscopy*, *62*(11), 1180-1188.

Bec et al., 2020.

Bec, J., Shaik, T. A., Krafft, C., Bocklitz, T. W., Alfonso-Garcia, A., Margulies, K. B., ... & Marcu, L. (**2020**). Investigating origins of FLIm contrast in atherosclerotic lesions using combined FLIm-Raman spectroscopy. *Frontiers in Cardiovascular Medicine*, *7*, 122.

Becker, 2012.

Becker, W. (**2012**). Fluorescence lifetime imaging–techniques and applications. *Journal of microscopy*, *247*(2), 119-136.

Beyad and Maeder, 2013.

Beyad, Y., & Maeder, M. (**2013**). Multivariate linear regression with missing values. *Analytica Chimica Acta*, *796*, 38-41.

Borgen and Kowalski, 1985.

Borgen, O. S., & Kowalski, B. R. (**1985**). An extension of the multivariate component-resolution method to three components. *Analytica Chimica Acta*, *174*, 1-26.

Borràs et al., 2015.

Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., & Busto, O. (**2015**). Data fusion methodologies for food and beverage authentication and quality assessment–A review. *Analytica Chimica Acta*, 891, 1-14.

Brereton et al., 2017.
Brereton, R. G., Jansen, J., Lopes, J., Marini, F., Pomerantsev, A., Rodionova, O., ... & Tauler, R. (**2017**). Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools. *Analytical and Bioanalytical Chemistry*, *409*, 5891-5899.

Bro, 1997.
Bro, R. (**1997**). PARAFAC. Tutorial and applications. *Chemometrics and intelligent laboratory systems*, *38*(2), 149-171.

Bro and De Jong, 1997.
Bro, R., & De Jong, S. (**1997**). A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *11*(5), 393-401.

Bro and Sidiropoulos, 1998.
Bro, R., & Sidiropoulos, N. D. (**1998**). Least squares algorithms under unimodality and non-negativity constraints. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 12(4), 223-247.

Buchberger et al., 2018.
Buchberger, A. R., DeLaney, K., Johnson, J., & Li, L. (**2018**). Mass spectrometry imaging: a review of emerging advancements and future insights. *Analytical Chemistry*, *90*(1), 240.

Burger, 2006.
Burger JE (**2006**). Hyperspectral NIR Image Analysis. Data Exploration, Correction, and Regression. Doctoral thesis. Swedish University of Agriculutral Sciences, Umeå.

Burger and Geladi, 2006.
Burger, J., & Geladi, P. (**2006**). Hyperspectral NIR imaging for calibration and prediction: a comparison between image and spectrometer data for studying organic and biological samples. Analyst, 131(10), 1152-1160.

Caprioli, 2015.
Caprioli, R. M. (**2015**). Imaging mass spectrometry: enabling a new age of discovery in biology and medicine through molecular microscopy. Journal of the American Society for Mass Spectrometry, 26, 850-852.

Chan and Kazarian, 2016.
Chan, K. A., & Kazarian, S. G. (**2016**). Attenuated total reflection Fourier-transform infrared (ATR-FTIR) imaging of tissues and live cells. *Chemical Society Reviews*, *45*(7), 1850-1864.

Chaumel et al., 2021.
Chaumel, J., Marsal, M., Gómez-Sánchez, A., Blumer, M., Gualda, E. J., de Juan, A., ... & Dean, M. N. (**2021**). Autofluorescence of stingray skeletal cartilage: hyperspectral imaging as a tool for histological characterization. *Discover Materials*, *1*(1), 16.

Christoffersson, 1970.
Christoffersson, A. (**1970**). The One Component Model with Incomplete Data. Doctoral Thesis, University of Uppsala, Sweden.

Clark and Dines, 1986.
Clark, R. J., & Dines, T. J. (**1986**). Resonance Raman spectroscopy, and its application to inorganic chemistry. New analytical methods (27). *Angewandte Chemie International Edition in English*, *25*(2), 131-158.

Clarke, 1819.
Clarke, E. D. (1819). Account of a newly discovered variety of green fluor spar, of very uncommon beauty, and with remarkable properties of colour and phosphorescence. *The Annals of Philosophy*, *14*(34-36), 2.

Cocchi et al., 2018.
Cocchi, M., Biancolillo, A., & Marini, F. (**2018**). Chemometric methods for classification and feature selection. In *Comprehensive analytical chemistry* (Vol. 82, pp. 265-299). Elsevier.

Coic et al., 2022.
Coic, L., Sacre, P. Y., Dispas, A., De Bleye, C., Fillet, M., Ruckebusch, C., ... & Ziemons, E. (**2022**). Selection of essential spectra to improve the multivariate curve resolution of minor compounds in complex pharmaceutical formulations. *Analytica Chimica Acta*, *1198*, 339532.

Cozzolino, 2012.
Cozzolino, D. (**2012**). Recent trends on the use of infrared spectroscopy to trace and authenticate natural and agricultural food products. *Applied Spectroscopy Reviews*, *47*(7), 518-530.

Danezis et al., 2016.
Danezis, G. P., Tsagkaris, A. S., Camin, F., Brusic, V., & Georgiou, C. A. (**2016**). Food authentication: Techniques, trends & emerging approaches. *TrAC Trends in Analytical Chemistry*, *85*, 123-132.

de Juan et al., 2005.
de Juan, A., Maeder, M., Hancewicz, T., & Tauler, R. (**2005**). Local rank analysis for exploratory spectroscopic image analysis. Fixed size image window-evolving factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 77(1-2), 64-74.

de Juan et al., 2008.
de Juan, A., Maeder, M., Hancewicz, T., & Tauler, R. (**2008**). Use of local rank-based spatial information for resolution of spectroscopic images. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 22(5), 291-298.

de Juan and Tauler, 2016.
de Juan, A., & Tauler, R. (**2016**). Multivariate curve resolution-alternating least squares for spectroscopic data. In *Data Handling in Science and Technology* (Vol. 30, pp. 5-51). Elsevier.

de Juan, 2018.
de Juan, A. (**2018**). Hyperspectral image analysis. When space meets Chemistry. *Journal of Chemometrics*, *32*(1), e2985.

de Juan, 2019.
de Juan, A. (**2019**). Multivariate curve resolution for hyperspectral image analysis. In *Data Handling in Science and Technology* (Vol. 32, pp. 115-150). Elsevier.

de Juan et al., 2019.
de Juan, A., Gowen, A., Duponchel, L., & Ruckebusch, C. (**2019**). Image fusion. In *Data handling in science and technology* (Vol. 31, pp. 311-344). Elsevier.

de Juan and Tauler, 2021.
de Juan, A., & Tauler, R. (**2021**). Multivariate Curve Resolution: 50 years addressing the mixture analysis problem–A review. *Analytica Chimica Acta*, *1145*, 59-78.

de Juan et al., 2024.
de Juan, A., de Oliveira, R. R., & Gómez-Sánchez, A. (**2024**). Multiset analysis by multivariate curve resolution: The unmixing methodology to handle hyperspectral image fusion scenarios. In *Data Handling in Science and Technology* (Vol. 33, pp. 111-132). Elsevier.

De Luca et al., 2014.
De Luca, M., Ragno, G., Ioele, G., & Tauler, R. (**2014**). Multivariate curve resolution of incomplete fused multiset data from chromatographic and spectrophotometric analyses for drug photostability studies. *Analytica Chimica Acta*, *837*, 31-37.

Devos et al., 2021.
Devos, O., Ghaffari, M., Vitale, R., de Juan, A., Sliwa, M., & Ruckebusch, C. (**2021**). Multivariate curve resolution slicing of multiexponential time-resolved spectroscopy fluorescence data. *Analytical Chemistry*, *93*(37), 12504-12513.

Dochow et al., 2015.
Dochow, S., Ma, D., Latka, I., Bocklitz, T., Hartl, B., Bec, J., ... & Popp, J. (**2015**). Combined fiber probe for fluorescence lifetime and Raman spectroscopy. *Analytical and bioanalytical chemistry*, *407*, 8291-8301.

Donaldson and Williams, 2018.
Donaldson, L., & Williams, N. (**2018**). Imaging and spectroscopy of natural fluorophores in pine needles. *Plants*, *7*(1), 10.

Donaldson, 2020.
Donaldson, L. (**2020**). Autofluorescence in plants. *Molecules*, *25*(10), 2393.

Duponchel et al., 2003.
Duponchel, L., Elmi-Rayaleh, W., Ruckebusch, C., & Huvenne, J. P. (**2003**). Multivariate curve resolution methods in imaging spectroscopy: influence of extraction methods and instrumental perturbations. *Journal of Chemical information and computer sciences*, 43(6), 2057-2067.

Efremov et al., 2008.
Efremov, E. V., Ariese, F., & Gooijer, C. (**2008**). Achievements in resonance Raman spectroscopy: Review of a technique with a distinct analytical chemistry potential. *Analytica chimica acta*, *606*(2), 119-134.

Eilers and Kroonenberg, 2014.
Eilers, P. H., & Kroonenberg, P. M. (**2014**). Modeling and correction of Raman and Rayleigh scatter in fluorescence landscapes. *Chemometrics and Intelligent Laboratory Systems*, *130*, 1-5.

El-Hagrasy et al., 2001.
El-Hagrasy, A. S., Morris, H. R., D'amico, F., Lodder, R. A., & Drennen III, J. K. (**2001**). Near-infrared spectroscopy and imaging for the monitoring of powder blend homogeneity. *Journal of pharmaceutical sciences*, *90*(9), 1298-1307.

Elcoroaristizabal et al., 2015.
Elcoroaristizabal, S., Bro, R., García, J. A., & Alonso, L. (**2015**). PARAFAC models of fluorescence data with scattering: A comparative study. *Chemometrics and Intelligent Laboratory Systems*, *142*, 124-130.

Engelsen et al., 2003.
Engelsen, S. B., & Bro, R. (**2003**). PowerSlicing. *Journal of Magnetic Resonance*, *163*(1), 192-197.

Gemperline, 1989.
Gemperline, P. J. (**1989**). Mixture analysis using factor analysis I: Calibration and quantitation. *Journal of chemometrics*, 3(4), 549-568.

Gierlinger and Schwanninger, 2007.
Gierlinger, N., & Schwanninger, M. (**2007**). The potential of Raman microscopy and Raman imaging in plant research. *Journal of Spectroscopy*, *21*, 69-89.

Golshan et al., 2016.
Golshan, A., Abdollahi, H., Beyramysoltan, S., Maeder, M., Neymeyr, K., Rajkó, R., ... & Tauler, R. (**2016**). A review of recent methods for the determination of ranges of feasible solutions resulting from soft modelling analyses of multivariate data. *Analytica Chimica Acta*, *911*, 1-13.

Gowen et al., 2008.
Gowen, A. A., O'donnell, C. P., Cullen, P. J., & Bell, S. E. J. (**2008**). Recent applications of chemical imaging to pharmaceutical process monitoring and quality control. *European journal of pharmaceutics and biopharmaceutics*, *69*(1), 10-22.

Griffiths, 2002.
Griffiths, P. R. (**2002**). *Handbook of vibrational spectroscopy* (Vol. 4). J. M. Chalmers (Ed.). Wiley.

Grung and Manne, 1998.
Grung, B., & Manne, R. (**1998**). Missing values in principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *42*(1-2), 125-139.

Gualda et al., 2015.
Gualda, E. J., Pereira, H., Vale, T., Estrada, M. F., Brito, C., & Moreno, N. (**2015**). SPIM-fluid: open source light-sheet based platform for high-throughput imaging. *Biomedical optics express*, *6*(11), 4447-4456.

Haaland et al., 2007.
Haaland, D. M., Jones, H. D., Sinclair, M. B., Carson, B., Branda, C., Poschet, J. F., ... & Brasier, A. R. (**2007**). Hyperspectral confocal fluorescence imaging of cells. In *Next-Generation Spectroscopic Technologies* (Vol. 6765, pp. 50-58). SPIE.

Hagen et al., 2012.
Hagen, N., Kester, R. T., Gao, L., & Tkaczyk, T. S. (**2012**). Snapshot advantage: a review of the light collection improvement for parallel high-dimensional measurement systems. *Optical Engineering*, *51*(11), 111702-111702.

Hamilton and Gemperline, 1990.
Hamilton, J. C., & Gemperline, P. J. (**1990**). Mixture analysis using factor analysis. II: self-modeling curve resolution. *Journal of chemometrics*, *4*(1), 1-13.

Herschel, 2013.
Herschel, W. (**2013**). *The Scientific Papers of Sir William Herschel*. Cambridge University Press.

Hirschfeld, 1980.
Hirschfeld, T. (**1980**). The hy-phen-ated methods. Analytical Chemistry, 52(2), 297A-312A.

Hoebe et al., 2007.
Hoebe, R. A., Van Oven, C. H., Gadella Jr, T. W. J., Dhonukshe, P. B., Van Noorden, C. J. F., & Manders, E. M. M. (**2007**). Controlled light-exposure microscopy reduces photobleaching and phototoxicity in fluorescence live-cell imaging. Nature biotechnology, 25(2), 249-253.

Hruska et al., 2014.
Hruska, Z., Yao, H., Kincaid, R., Brown, R., Cleveland, T., & Bhatnagar, D. (**2014**). Fluorescence excitation–emission features of aflatoxin and related secondary metabolites and their application for rapid detection of mycotoxins. Food and Bioprocess Technology, 7, 1195-1201.

Hugelier et al., 2015.
Hugelier, S., Devos, O., & Ruckebusch, C. (**2015**). On the implementation of spatial constraints in multivariate curve resolution alternating least squares for hyperspectral image analysis. *Journal of Chemometrics*, *29*(10), 557-561.

Hugelier et al., 2015b.
Hugelier, S., Devos, O., & Ruckebusch, C. (**2015**). Constraining shape smoothness in multivariate curve resolution–alternating least squares. *Journal of Chemometrics*, *29*(8), 448-456.

Huang et al., 2009.
Huang, B., Bates, M., & Zhuang, X. (**2009**). Super-resolution fluorescence microscopy. *Annual review of biochemistry*, *78*, 993-1016.

Ilin and Raiko, 2010.
Ilin, A., & Raiko, T. (**2010**). Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research*, *11*, 1957-2000.

Jamin et al., 1998.
Jamin, N., Dumas, P., Moncuit, J., Fridman, W. H., Teillaud, J. L., Carr, G. L., & Williams, G. P. (**1998**). Highly resolved chemical imaging of living cells by using synchrotron infrared microspectrometry. *Proceedings of the National Academy of Sciences*, *95*(9), 4837-4840.

Jones et al., 2019.
Jones, R. R., Hooper, D. C., Zhang, L., Wolverson, D., & Valev, V. K. (**2019**). Raman techniques: fundamentals and frontiers. *Nanoscale research letters*, *14*, 1-34.

Kruskal, 1977.
Kruskal, J. B. (**1977**). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, *18*(2), 95-138.

Joliffe and Morgan, 1992.
Joliffe, I. T., & Morgan, B. J. T. (**1992**). Principal component analysis and exploratory factor analysis. *Statistical methods in medical research*, *1*(1), 69-95.

Kawata et al., 2017.
Kawata, S., Ichimura, T., Taguchi, A., & Kumamoto, Y. (**2017**). Nano-Raman scattering microscopy: resolution and enhancement. *Chemical reviews*, *117*(7), 4983-5001.

Koenig, 1975.
Koenig, J. L. (**1975**). Application of Fourier transform infrared spectroscopy to chemical systems. *Applied Spectroscopy*, *29*(4), 293-308.

Lakowicz, 2006.
Lakowicz, J. R. (**2006**). In Principles of Fluorescence Spectroscopy.

Landgrebe, 1999.
Landgrebe, D. (**1999**). Information extraction principles and methods for multispectral and hyperspectral image data. In *Information processing for remote sensing* (pp. 3-37).

Lasch and Naumann, 2006.
Lasch, P., & Naumann, D. (**2006**). Spatial resolution in infrared microspectroscopic imaging of tissues. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1758*(7), 814-829.

Lawson and Hanson, 1995.
Lawson, C. L., & Hanson, R. J. (**1995**). *Solving least squares problems*. Society for Industrial and Applied Mathematics.

Lawton and Sylvestre, 1971.
Lawton, W. H., & Sylvestre, E. A. (**1971**). Self modeling curve resolution. *Technometrics*, *13*(3), 617-633.

Lemmetyinen et al., 2014.
Lemmetyinen, H., Tkachenko, N. V., Valeur, B., Hotta, J. I., Ameloot, M., Ernsting, N. P., ... & Boens, N. (**2014**). Time-resolved fluorescence methods (IUPAC Technical Report). *Pure and Applied Chemistry*, *86*(12), 1969-1998.

Lewis and Edwards, 2001.
Lewis, I. R., & Edwards, H. (**2001**). *Handbook of Raman spectroscopy: from the research laboratory to the process line*. CRC press.

Liput et al., 2020.

Liput, D. J., Nguyen, T. A., Augustin, S. M., Lee, J. O., & Vogel, S. S. (**2020**). A guide to fluorescence lifetime microscopy and Förster's Resonance Energy Transfer in Neuroscience. *Current protocols in neuroscience*, *94*(1), e108.

Luchowski et al., 2009.
Luchowski, R., Gryczynski, Z., Sarkar, P., Borejdo, J., Szabelski, M., Kapusta, P., & Gryczynski, I. (**2009**). Instrument response standard in time-resolved fluorescence. *Review of Scientific Instruments*, *80*(3).

Maeder, 1987.
Maeder, M. (**1987**). Evolving factor analysis for the resolution of overlapping chromatographic peaks. *Analytical chemistry*, *59*(3), 527-530.

Malik and Tauler, 2013.
Malik, A., & Tauler, R. (**2013**). Extension and application of multivariate curve resolution-alternating least squares to four-way quadrilinear data-obtained in the investigation of pollution patterns on Yamuna River, India—a case study. *Analytica chimica acta*, *794*, 20-28.

Malik and Tauler, 2014.
Malik, A., & Tauler, R. (**2014**). Performance and validation of MCR-ALS with quadrilinear constraint in the analysis of noisy datasets. *Chemometrics and Intelligent Laboratory Systems*, *135*, 223-234.

Malinowski, 1992.
Malinowski, E. R. (**1992**). Window factor analysis: Theoretical derivation and application to flow injection analysis data. *Journal of chemometrics*, *6*(1), 29-40.

Mangold et al., 2008.
Mangold, N., Gendrin, A., Gondet, B., LeMouelic, S., Quantin, C., Ansan, V., ... & Neukum, G. (**2008**). Spectral and geological study of the sulfate-rich region of West Candor Chasma, Mars. *Icarus*, 194(2), 519-543.

Manne, 1995.
Manne, R. (**1995**). On the resolution problem in hyphenated chromatography. *Chemometrics and Intelligent Laboratory Systems*, *27*(1), 89-94.

Marín-García and Tauler, 2020.
Marín-García, M., & Tauler, R. (**2020**). Chemometrics characterization of the Llobregat river dissolved organic matter. *Chemometrics and Intelligent Laboratory Systems*, *201*, 104018.

Mas et al., 2010.
Mas, S., de Juan, A., Tauler, R., Olivieri, A. C., & Escandar, G. M. (**2010**). Application of chemometric methods to environmental analysis of organic pollutants: A review. *Talanta*, *80*(3), 1052-1067.

McCreery, 2005.
McCreery, R. L. (**2005**). *Raman spectroscopy for chemical analysis*. John Wiley & Sons.

McDonnell and Heerem, 2007.
McDonnell, L. A., & Heeren, R. M. (**2007**). Imaging mass spectrometry. *Mass spectrometry reviews*, *26*(4), 606-643.

Mertz, 2019.
Mertz, J. (**2019**). Strategies for volumetric imaging with a fluorescence microscope. *Optica*, *6*(10), 1261-1268.

Mintenig et al., 2017
Mintenig, S. M., Int-Veen, I., Löder, M. G., Primpke, S., & Gerdts, G. (**2017**). Identification of microplastic in effluents of waste water treatment plants using focal plane array-based micro-Fourier-transform infrared imaging. *Water research*, *108*, 365-372.

Nagy et al., 2009.
Nagy, J. A., Chang, S. H., Dvorak, A. M., & Dvorak, H. F. (**2009**). Why are tumour blood vessels abnormal and why is it important to know?. *British journal of cancer*, *100*(6), 865-869.

Naumann et al., 1991.
Naumann, D., Helm, D., & Labischinski, H. (**1991**). Microbiological characterizations by FT-IR spectroscopy. *Nature*, *351*(6321), 81-82.

Neal et al., 2023.
Neal, S. N., Stacchiola, D., & Tenney, S. A. (**2023**). Spatially resolved multimodal vibrational spectroscopy under high pressures. *Physical Chemistry Chemical Physics*, *25*(46), 31578-31582.

Occhipinti et al., 2023.
Occhipinti, M., Alberti, R., Parsani, T., Dicorato, C., Tirelli, P., Gironda, M., ... & Frizzi, T. (**2023**). IRIS: A novel integrated instrument for co-registered MA-XRF mapping and VNIR-SWIR hyperspectral imaging. *X-Ray Spectrometry*.

Offroy et al., 2010.
Offroy, M., Roggo, Y., Milanfar, P., & Duponchel, L. (**2010**). Infrared chemical imaging: Spatial resolution evaluation and super-resolution concept. *Analytica chimica acta*, 674(2), 220-226.

Olivieri et al., 2004.
Olivieri, A. C., Arancibia, J. A., Muñoz de la Peña, A., Duran-Meras, I., & Espinosa Mansilla, A. (**2004**). Second-order advantage achieved with four-way fluorescence excitation− emission− kinetic data processed by parallel factor analysis and trilinear least-squares. Determination of methotrexate and leucovorin in human urine. *Analytical Chemistry*, *76*(19), 5657-5666.

Olmos et al., 2018.
Olmos, V., Marro, M., Loza-Alvarez, P., Raldúa, D., Prats, E., Padrós, F., ... & de Juan, A. (**2018**). Combining hyperspectral imaging and chemometrics to assess and interpret the effects of environmental stressors on zebrafish eye images at tissue level. *Journal of biophotonics*, *11*(3), e201700089.

Omidikia, 2022.
Omidikia, N. (**2022**). The effect of multilinear data fusion on the accuracy of multivariate curve resolution outputs. *Analytica Chimica Acta*, *1227*, 340325.

Omrani et al., 2014.
Omrani, H., Dudelzak, A. E., Hollebone, B. P., & Loock, H. P. (**2014**). Assessment of the oxidative stability of lubricant oil using fiber-coupled fluorescence excitation–emission matrix spectroscopy. *Analytica chimica acta*, *811*, 1-12.

Patel and Mehta, 2010.
Patel, BD., & Mehta, PJ. (**2010**). An overview: application of Raman spectroscopy in pharmaceutical field. *Current Pharmaceutical Analysis*, *6*(2), 131-141.

Petry et al., 2003.
Petry, R., Schmitt, M., & Popp, J. (**2003**). Raman spectroscopy—a prospective tool in the life sciences. *chemphyschem*, *4*(1), 14-30.

Piqueras et al., 2017.
Piqueras, S., Maeder, M., Tauler, R., & De Juan, A. (**2017**). A new matching image preprocessing for image data fusion. *Chemometrics and Intelligent Laboratory Systems*, *164*, 32-42.

Piqueras et al., 2018.
Piqueras, S., Bedia, C., Beleites, C., Krafft, C., Popp, J., Maeder, M., ... & de Juan, A. (**2018**). Handling different spatial resolutions in image fusion by multivariate curve resolution-alternating least squares for incomplete image multisets. *Analytical chemistry*, *90*(11), 6757-6765.

Piqueras et al., 2020.
Piqueras, S., Füchtner, S., Rocha de Oliveira, R., Gomez-Sanchez, A., Jelavić, S., Keplinger, T., ... & Thygesen, L. G. (**2020**). Understanding the formation of heartwood in larch using synchrotron infrared imaging combined with multivariate analysis and atomic force microscope infrared spectroscopy. *Frontiers in plant science*, *10*, 1701.

Primpke et al., 2017.
Primpke, S., Lorenz, C., Rascher-Friesenhausen, R., & Gerdts, G. (**2017**). An automated approach for microplastics analysis using focal plane array (FPA) FTIR microscopy and imge analysis. *Analytical Methods*, 9(9), 1499-1511.

Qin et al., 2013.
Qin, J., Chao, K., Kim, M. S., Lu, R., & Burks, T. F. (**2013**). Hyperspectral and multispectral imaging for evaluating food safety and quality. *Journal of Food Engineering*, *118*(2), 157-171.

Queral-Beltran et al. (2024).
Queral-Beltran, A., Marin-Garcia, M., Lacorte, S., & Tauler, R. (**2024**). Multivariate curve resolution of incomplete and partly trilinear multiblock datasets. *Chemometrics and Intelligent Laboratory Systems*, 105081.

Raman, 1928.
Raman, C. V., & Krishnan, K. S. (**1928**). A new type of secondary radiation. *Nature*, *121*(3048), 501-502.

Rodríguez-Vidal et al., 2020.
Rodríguez-Vidal, F. J., García-Valverde, M., Ortega-Azabache, B., González-Martínez, Á., & Bellido-Fernández, A. (**2020**). Characterization of urban and industrial wastewaters using excitation-emission matrix (EEM) fluorescence: Searching for specific fingerprints. *Journal of environmental management*, *263*, 110396.

Ring, 2000.
Ring, E. F. J. (**2000**). The discovery of infrared radiation in 1800. *The Imaging Science Journal*, *48*(1), 1-8.

Ruckebusch and Blanchet, 2013.
Ruckebusch, C., & Blanchet, L. (**2013**). Multivariate curve resolution: a review of advanced and tailored applications and challenges. *Analytica chimica acta*, 765, 28-36.

Savitzky and Golay, 1964.
Savitzky, A., & Golay, M. J. (**1964**). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, *36*(8), 1627-1639.

Sawall et al., 2019.
Sawall, M., Schröder, H., Meinhardt, D., & Neymeyr, K. (**2019**). On the ambiguity underlying multivariate curve resolution methods. *Comprehensive Chemometrics, 2nd edition*, pages 199-231.

Sawall et al., 2022.
Sawall, M., Ruckebusch, C., Beese, M., Francke, R., Prudlik, A., & Neymeyr, K. (**2022**). An active constraint approach to identify essential spectral information in noisy data. *Analytica Chimica Acta*, *1233*, 340448.

Selci, 2019.
Selci, S. (**2019**). The future of hyperspectral imaging. *Journal of Imaging*, *5*(11), 84.

Schie et al., 2021.
Schie, I. W., Stiebing, C., & Popp, J. (**2021**). Looking for a perfect match: multimodal combinations of Raman spectroscopy for biomedical applications. *Journal of Biomedical Optics*, *26*(8), 080601-080601.

Schlücker et al., 2003.
Schlücker, S., Schaeberle, M. D., Huffman, S. W., & Levin, I. W. (**2003**). Raman microspectroscopy: a comparison of point, line, and wide-field imaging methodologies. *Analytical Chemistry*, *75*(16), 4312-4318.

Shashkova and Leake, 2017.
Shashkova, S., & Leake, M. C. (**2017**). Single-molecule fluorescence microscopy review: shedding new light on old problems. *Bioscience reports*, *37*(4), BSR20170031.

Shipp et al., 2017.
Shipp, D. W., Sinjab, F., & Notingher, I. (**2017**). Raman spectroscopy: techniques and applications in the life sciences. *Advances in Optics and Photonics*, *9*(2), 315-428.

Simon et al., 2018.
Simon, M., van Alst, N., & Vollertsen, J. (**2018**). Quantification of microplastic mass and removal rates at wastewater treatment plants applying Focal Plane Array (FPA)-based Fourier Transform Infrared (FT-IR) imaging. *Water research*, *142*, 1-9.

Smith et al., 2016.
Smith, R., Wright, K. L., & Ashton, L. (**2016**). Raman spectroscopy: an evolving technique for live cell studies. *Analyst*, *141*(12), 3590-3600.

Smolinska et al., 2019.
Smolinska, A., Engel, J., Szymanska, E., Buydens, L., & Blanchet, L. (**2019**). General framing of low-, mid-, and high-level data fusion with examples in the life sciences. In *Data Handling in Science and Technology* (Vol. 31, pp. 51-79). Elsevier.

Soriano-Disla et al., 2014.
Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., Macdonald, L. M., & McLaughlin, M. J. (**2014**). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied spectroscopy reviews*, *49*(2), 139-186.

Stedmon et al., 2008.
Stedmon, C. A., & Bro, R. (**2008**). Characterizing dissolved organic matter fluorescence with parallel factor analysis: a tutorial. *Limnology and Oceanography: Methods*, *6*(11), 572-579.

Suhling et al., 2015.
Suhling, K., Hirvonen, L. M., Levitt, J. A., Chung, P. H., Tregidgo, C., Le Marois, A., ... & Krstajic, N. (**2015**). Fluorescence lifetime imaging (FLIM): Basic concepts and some recent developments. *Medical Photonics*, 27, 3-40.

Swinehart, 1962.
Swinehart, D. F. (**1962**). The beer-lambert law. *Journal of chemical education*, *39*(7), 333.

Szabelski et al., 2009.
Szabelski, M., Ilijev, D., Sarkar, P., Luchowski, R., Gryczynski, Z., Kapusta, P., ... & Gryczynski, I. (**2009**). Collisional quenching of erythrosine B as a potential reference dye for impulse response function evaluation. *Applied spectroscopy*, *63*(3), 363-368.

Szabelski et al., 2009b.
Szabelski, M., Luchowski, R., Gryczynski, Z., Kapusta, P., Ortmann, U., & Gryczynski, I. (**2009**). Evaluation of instrument response functions for lifetime imaging detectors using quenched Rose Bengal solutions. *Chemical Physics Letters*, *471*(1-3), 153-159.

Talari et al., 2017.
Talari, A. C. S., Martinez, M. A. G., Movasaghi, Z., Rehman, S., & Rehman, I. U. (**2017**). Advances in Fourier transform infrared (FTIR) spectroscopy of biological tissues. *Applied Spectroscopy Reviews*, *52*(5), 456-506.

Tauler and Barceló, 1993.
Tauler, R., & Barceló, D. (**1993**). Multivariate curve resolution applied to liquid chromatography—diode array detection. *TrAC Trends in Analytical Chemistry*, *12*(8), 319-327.

Tauler, 1995.
Tauler, R. (**1995**). Multivariate curve resolution applied to second order data. *Chemometrics and intelligent laboratory systems*, *30*(1), 133-146.

Tauler et al., 1995.
Tauler, R., Smilde, A., & Kowalski, B. (**1995**). Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *Journal of Chemometrics*, *9*(1), 31-58.

Tauler et al., 1998.
Tauler, R., Marqués, I., & Casassas, E. (**1998**). Multivariate curve resolution applied to three-way trilinear data: Study of a spectrofluorimetric acid–base titration of salicylic acid at three excitation wavelengths. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *12*(1), 55-75.

Tauler and Maeder, 2009.
Tauler, R., & Maeder, M. (**2009**). Two-way data analysis: multivariate curve resolution–error in curve resolution.

Tauler et al., 2020.
Tauler, R., Maeder, M., & De Juan, A. (**2020**). Multiset data analysis: extended multivariate curve resolution.

Tauler, 2021.
Tauler, R. (**2021**). Multivariate curve resolution of multiway data using the multilinearity constraint. *Journal of Chemometrics*, *35*(2), e3279.

Tschirner et al., 2009.
Tschirner, N., Schenderlein, M., Brose, K., Schlodder, E., Mroginski, M. A., Thomsen, C., & Hildebrandt, P. (**2009**). Resonance Raman spectra of β-carotene in solution and in photosystems revisited: an experimental and theoretical study. *Physical chemistry chemical physics*, *11*(48), 11471-11478.

Tuck et al., 2020.
Tuck, M., Blanc, L., Touti, R., Patterson, N. H., Van Nuffel, S., Villette, S., ... & Desbenoit, N. (**2020**). Multimodal imaging based on vibrational spectroscopies and mass spectrometry imaging applied to biological tissue: a multiscale and multiomics review. *Analytical chemistry*, *93*(1), 445-477.

Turrel and Corset, 1996.
Turrell, G., & Corset, J. (Eds.). (**1996**). *Raman microscopy: developments and applications*. Academic Press.

Valeur and Berberan-Santos, 2012.
Valeur, B., & Berberan-Santos, M. N. (**2012**). *Molecular fluorescence: principles and applications*. John Wiley & Sons.

Vishal et al., 2019.
Vishal, B., Krishnamurthy, P., Ramamoorthy, R., & Kumar, P. P. (**2019**). Os TPS 8 controls yield-related traits and confers salt stress tolerance in rice by enhancing suberin deposition. *New Phytologist*, *221*(3), 1369-1386.

Walczak and Massart, 2001.
Walczak, B., & Massart, D. L. (2001). Dealing with missing data: Part I. Chemometrics and Intelligent Laboratory Systems, 58(1), 15-27.

Wang et al., 2014.
Wang, W., & Paliwal, J. (**2014**). A multimodal spectrometer for Raman scattering and near-infrared absorption measurement. *Vibrational Spectroscopy*, *74*, 13-19.

Wen et al., 2020.
Wen, Z., Wang, L., Zhang, X., Ma, Y., Liu, X., Kaminski, C. F., & Yang, Q. (**2020**). Fast volumetric fluorescence imaging with multimode fibers. *Optics Letters*, *45*(17), 4931-4934.

Wightman et al., 2019.
Wightman, R., Busse-Wicher, M., & Dupree, P. (**2019**). Correlative FLIM-confocal-Raman mapping applied to plant lignin composition and autofluorescence. *Micron*, *126*, 102733.

Williams and Norris, 2001.
Williams, P. and Norris, K. (**2001**) Near-Infrared Technology in the Agricultural and Food Industries. American Association of Cereal Chemists, USA.

Williams et al., 2021.
Williams, G. O., Williams, E., Finlayson, N., Erdogan, A. T., Wang, Q., Fernandes, S., ... & Bradley, M. (**2021**). Full spectrum fluorescence lifetime imaging with 0.5 nm spectral and 50 ps temporal resolution. *Nature communications*, *12*(1), 6616.

Windig and Guilment, 1991.
Windig, W., & Guilment, J. (**1991**). Interactive self-modeling mixture analysis. *Analytical chemistry*, *63*(14), 1425-1432.

Wold, 1975.
Wold, H. (**1975**). Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. *Journal of Applied Probability*, *12*(S1), 117-142.

Wold, 1977.
Wold, S., & Sjöström, M. (**1977**). SIMCA: a method for analyzing chemical data in terms of similarity and analogy.

Wolfe and Zissis, 1978.
Wolfe, W. L., & Zissis, G. J. (Eds.). (**1978**). *The infrared handbook*. The Office.

Yang and Ying, 2011.
Yang, D., & Ying, Y. (**2011**). Applications of Raman spectroscopy in agricultural products and food analysis: A review. *Applied Spectroscopy Reviews*, *46*(7), 539-560.

Zhang et al., 2016.
Zhang, X., de Juan, A., & Tauler, R. (**2016**). Local rank-based spatial information for improvement of remote sensing hyperspectral imaging resolution. *Talanta*, *146*, 1-9.

Zhang et al., 2017.
Zhang, X., Chen, S., & Xu, F. (**2017**). Combining Raman imaging and multivariate analysis to visualize lignin, cellulose, and hemicellulose in the plant cell wall. *JoVE (Journal of Visualized Experiments)*, (124), e55910.

Zavattini et al., 2003.
Zavattini, G., Vecchi, S., Leahy, R. M., Smith, D. J., & Cherry, S. R. (**2003**). A hyperspectral fluorescence imaging system for biological applications. In *2003 IEEE Nuclear Science Symposium. Conference Record (IEEE Cat. No. 03CH37515)* (Vol. 2, pp. 942-946). IEEE.

Zexer et al., 2020.
Zexer, N., & Elbaum, R. (**2020**). Unique lignin modifications pattern the nucleation of silica in sorghum endodermis. *Journal of Experimental Botany*, *71*(21), 6818-6829.