Structured and Low-Rank Decompositions for Large-Scale Imaging Datasets

Moens, R.A.R.

DOI
10.4233/uuid:c1057109-773e-4613-916d-237c459452f7

Publication date
2026

Document Version
Final published version

Citation (APA)
Moens, R. A. R. (2026). *Structured and Low-Rank Decompositions for Large-Scale Imaging Datasets*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:c1057109-773e-4613-916d-237c459452f7

Structured and Low-Rank Decompositions for Large-Scale Imaging Datasets

Roger A.R. Moens

Structured and Low-Rank Decompositions for Large-Scale Imaging Datasets

STRUCTURED AND LOW-RANK DECOMPOSITIONS FOR LARGE-SCALE IMAGING DATASETS

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus prof. dr. ir. H. Bijl
chair of the Board for Doctorates,
to be defended publicly on Friday 6 February 2026 at 15:00

by

Roger Amaury Rutger MOENS

Master of Science in Systems and Control,
Master of Science in Aerospace Engineering,
Delft University of Technology, Delft, Netherlands,
born in Brussels, Belgium.

This dissertation has been approved by

promotor: dr. ing. R. Van de Plas
promotor: prof. dr. ir. B. De Schutter

Composition of the doctoral committee:

Rector Magnificus,	chairperson
dr. ing. R. Van de Plas,	Delft University of Technology
prof. dr. ir. B. De Schutter,	Delft University of Technology
Independent members:	
dr. ir. K. Batselier	Delft University of Technology
prof. dr. N. Gillis	University of Mons, Belgium
prof. dr. ir. G. Leus	Delft University of Technology
dr. Y. Mohammed	Leiden University Medical Center
prof. dr. I. Styles	Queen's University Belfast, United Kingdom
Reserve member:	
prof. dr. ir. T. Keviczky	Delft University of Technology

Keywords: low-rank, sparsity, non-negativity, matrix decomposition, imaging

Printed by: Koninklijke Rijnja BV

Front & Back: Images generated by DALL-E.

ISBN 978-94-6518-231-5

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

Contents

SUMMARY

A common strategy to address new scientific challenges consists of abstracting the underlying problem, recasting it to an existing problem formulation and applying an established methodology. In this dissertation, we offer a variation on this familiar academic theme. The setting we will focus on is primarily found within image-acquiring instruments, characterized by producing vast quantities of data, from several hundreds up to more than half a million images per experiment. The challenges that we address throughout this work will mainly consist of (a) reducing dimensionality and (b) denoising, which have a direct and significant impact on the analysis and thus interpretation of these extensive image sets.

We investigate computational methods for two specific imaging instruments: (1) a time-of-flight imaging mass spectrometer, employed in biochemical research to visualize molecular distributions across very small organic tissues, and (2) a mid-infrared imager, utilized in astronomical research to study very large protostars, temperate exoplanets, and objects within our solar system. Despite their considerable promise in acquiring detailed molecular maps and critical astronomical insights, respectively, the practical analysis and interpretation of their image sets face substantial obstacles, namely their dimensionality and the effect of noise. Addressing these challenges may involve drawing on existing computational and storage capacity and harnessing any available prior information or problem-specific structure. Fortunately, analytical solutions to the obstacles across imaging instruments often bear a resemblance to each other, as we distil them to abstract mathematical models and eventually formulate those problems as optimization problems. The computational methods we are interested in are so-called low-rank methods, they can simultaneously provide insight in data structure (analysis), as well as reduce the dimensionality of the data and denoise it.

In a first paper, we investigate the limitations of traditional low-rank methods such as Principal Component Analysis (PCA) for analyzing inherently noisy, high-dimensional imaging mass spectrometry (IMS) datasets when spatially or spectrally sparse signals are abundant. We know that these signals can heavily distort the low-rank approximation, and therefore propose two established matrix decomposition techniques, Principal Component Pursuit (PCP) and Stable Principal Component Pursuit (SPCP), that explicitly separate sparse biological signals and dense noise from the low-rank structure in the data. These methods are applied to MALDI IMS datasets of human cornea and retina tissue, demonstrating improved performance in dimensionality reduction and denoising over PCA. The findings highlight that SPCP achieves more interpretable and compressed representations by accommodating sparse residuals and minimizing signal overestimation, which is essential for preserving biologically meaningful patterns in IMS data.

In a second paper, we use the same low-rank and sparse modeling technique, but here to overcome limitations in mid-infrared astronomical imaging caused by dominant, variable thermal background noise. Unlike traditional chopping and nodding methods,

which may be impractical for future large telescopes, such as the Mid-infrared ELT Imager and Spectrograph (METIS) instrument on the Extremely Large Telescope (ELT), we decompose data into low-rank background, sparse point sources and dense background noise components using SPCP. We verified our method on ground-based VLT Imager and Spectrometer for mid-InfraRed (VISIR) data with synthetically injected sources and airborne Stratospheric Observatory for Infrared Astronomy (SOFIA) data. For VISIR, our results show that the proposed methodology improves photometric consistency and detection precision in low signal-to-noise conditions, albeit with a bias in the retrieved source flux. This bias undermines precise photometry and needs further study. However, it does not directly affect the detection of faint sources, another important application within astronomical imaging. For SOFIA data, we achieved up to 100-fold background reduction while preserving source signal comparably to traditional methods. As such, we provide an efficient, computational background subtraction method that enhances sensitivity for faint source detection, paving the way for more robust mid-IR astronomical observations.

In a third paper, we propose a sparse-format-aware framework using Singular Value Thresholding (SVT) and Fixed-Point Continuation (FPC) algorithms, which models IMS data as incomplete matrices and poses the underlying problem as a matrix completion problem. Traditional data reduction methods like peak picking can sometimes miss low-intensity or near-isobaric molecular species and reduce the data selectively, potentially biasing downstream analysis in IMS. These novel methods capture the full profile of the raw IMS data, retaining more low-intensity and near-isobar signals than conventional techniques. Through case studies on different IMS datasets, the approach shows lower reconstruction errors than peak picking, even in the presence of missing data. Additionally, higher compression rates are achievable, up to 2500-fold, while minimizing information loss with respect to peak picking. The approach also mitigates selection bias and retains biologically relevant signals, enhancing the sensitivity and specificity of downstream analysis.

In a fourth paper, we present a novel framework for unmixing IMS measurements by integrating constraints derived from microscopy data. This approach leverages complementary spatial information to enhance traditional matrix factorization techniques, improving signal separation and noise reduction. By incorporating microscopy-informed constraints, the method effectively distinguishes mixed spectral signatures and isolates subtle molecular distributions within complex biological tissues. We detail the development and application of an optimization algorithm that regularizes the unmixing process through a least-squares approach, ensuring more accurate attribution of molecular components to their respective spatial locations. Experimental validation on a synthetic and real IMS dataset demonstrates enhanced retrieval of biological relevant signals compared to conventional approaches.

In summary, focusing on image-acquiring instruments that generate vast, high-dimensional datasets, we tackle the dual challenges of dimensionality reduction and denoising. Our work spans both biochemical imaging mass spectrometry and astronomical mid-infrared imaging, developing low-rank and sparse modeling strategies that exploit instrument- and experiment-specific structures. These methods yield more interpretable, compressed representations, reduced noise, and improved signal fidelity. By aligning

advanced optimization techniques with domain-specific needs, our contributions enable more accurate downstream analysis and lay groundwork for more sensitive and computationally efficient imaging pipelines that can be applied across other disciplines.

SAMENVATTING

Een gebruikelijke strategie om nieuwe wetenschappelijke uitdagingen aan te pakken, bestaat eruit het onderliggende probleem te abstraheren, het te herschikken naar een bestaande probleemformulering en een gevestigde methodologie toe te passen. In deze dissertatie bieden we een variatie op dit bekende academische thema. De context waarop we ons zullen richten, vindt men hoofdzakelijk binnen beeldvormende instrumenten, gekenmerkt door het produceren van enorme hoeveelheden data, variërend van enkele honderden tot meer dan een half miljoen beelden per experiment. De uitdagingen die we in dit werk behandelen, bestaan voornamelijk uit (a) het reduceren van dimensies en (b) het verminderen van ruis, die beiden een direct en significant effect heeft op de analyse en dus de interpretatie van deze omvangrijke beeldverzamelingen.

We onderzoeken computationele methoden voor twee specifieke beeldvormingsinstrumenten: (1) een time-of-flight beeldvormende massaspectrometer, gebruikt in biochemisch onderzoek om moleculaire distributies in zeer kleine organische weefsels te visualiseren, en (2) een midden-infrarood-imager, toegepast in astronomisch onderzoek om zeer grote protosterren, gematigde exoplaneten en objecten binnen ons zonnestelsel te bestuderen. Ondanks hun belofte om respectievelijk gedetailleerde moleculaire kaarten en cruciale astronomische inzichten te verkrijgen, stuit de praktische analyse en interpretatie van hun beeldverzamelingen op substantiële obstakels, namelijk de dimensies en de invloed van ruis. Het aanpakken van deze uitdagingen kan inhouden dat men gebruikmaakt van bestaande computationele en opslagcapaciteit en eventuele beschikbare voorkennis of probleem-specifieke structuur benut. Gelukkig vertonen analytische oplossingen voor de obstakels bij verschillende beeldvormingsinstrumenten vaak overeenkomsten, omdat we ze abstraheren tot wiskundige modellen en uiteindelijk die problemen formuleren als optimalisatieproblemen. De computationele methoden waarin we geïnteresseerd zijn, worden zogenaamde lage-rank methoden genoemd, ze kunnen tegelijkertijd inzicht verschaffen in de datastructuur (analyse), evenals de dimensies van de data reduceren en ruis onderdrukken.

In een eerste artikel onderzoeken we de beperkingen van traditionele lage-rank methoden zoals Principal Component Analysis (PCA) voor het analyseren van van nature ruisende, hoog-dimensionale datasets uit imaging mass spectrometry (IMS) wanneer er ruimtelijk of spectraal schaarse signalen aanwezig zijn. We weten dat deze signalen de lage-rank-benadering sterk kunnen vervormen, en daarom stellen we twee gevestigde matrix decompositie technieken voor, Principal Component Pursuit (PCP) en Stable Principal Component Pursuit (SPCP), die expliciet schaarse biologische signalen en dichte ruis van de lage-rank-structuur in de data scheiden. Deze methoden worden toegepast op MALDI IMS-datasets van menselijke hoornvlies- en netvliesweefsel, waarbij een verbeterde prestatie in dimensiereductie en ruisonderdrukking ten opzichte van PCA wordt aangetoond. De bevindingen benadrukken dat SPCP meer interpreteerbare en compacte representaties bereikt door schaarse residuen toe te staan en signaaloverschatting te

minimaliseren, wat essentieel is voor het behouden van biologisch relevante patronen in IMS-data.

In een tweede artikel gebruiken we dezelfde lage-rank- en schaarse modelleringsbenadering, maar nu om beperkingen in midden-infrarood-astronomische beeldvorming te overwinnen die worden veroorzaakt door overheersende, variabele thermische achtergrondruis. In tegenstelling tot traditionele chopping- en nodding-methoden, die wellicht onpraktisch zijn voor toekomstige grote telescopen zoals voor de Mid-infrared ELT Imager and Spectrograph (METIS)-instrument op de Extremely Large Telescope (ELT), ontbinden we de data met SPCP in lage-rank-achtergrond, schaarse puntbronnen en dichte achtergrondruiscomponenten. We hebben onze methode geverifieerd op grond-gebaseerde VLT Imager en Spectrometer for mid-InfraRed (VISIR)-data met synthetisch geïnjecteerde bronnen en op data vergaard vanuit een observatievliegtuig, de Stratospheric Observatory for Infrared Astronomy (SOFIA). Voor VISIR tonen onze resultaten aan dat de voorgestelde methodologie de fotometrische consistentie en detectieprecisie verbetert onder omstandigheden met een lage signaal-ruisverhouding, zij het met een bias in de teruggewonnen bronflux. Deze bias ondermijnt precieze fotometrie en vereist verder onderzoek, maar beïnvloedt de detectie van zwakke bronnen, een andere belangrijke toepassing binnen astronomische beeldvorming, niet. Voor SOFIA-data hebben we een achtergrondreductie tot wel 100-voud bereikt, terwijl het bronsignaal op vergelijkbare wijze behouden blijft als bij traditionele methoden. Als zodanig bieden we een efficiënte computationele methode voor achtergrondssubtractie die de gevoeligheid voor de detectie van zwakke bronnen verbetert en daarmee de weg effent voor robuustere midden-IR-astronomische waarnemingen.

In een derde artikel stellen we een framework voor dat rekening houdt met het sparce-formaat door gebruik te maken van de Singular Value Thresholding (SVT)- en Fixed-Point Continuation (FPC)-algoritmen, waarbij IMS-data worden gemodelleerd als incomplete matrices en het onderliggende probleem wordt geformuleerd als een matrixcompletieprobleem. Traditionele gegevensreductiemethoden zoals piekselectie kunnen soms lage-intensiteits- of bijna-isobare moleculaire soorten missen en de data selectief reduceren, dat verdere analyse in IMS kan vertekenen. Deze nieuwe methoden vangen het volledige profiel van de ruwe IMS-data op en behouden meer lage-intensiteits- en bijna-isobele signalen dan conventionele technieken. Bij verscheidene studies op verschillende IMS-datasets vertoont de benadering lagere reconstructiefouten dan piekselectie, zelfs in de aanwezigheid van ontbrekende data. Bovendien zijn hogere compressiewaarden tot wel 2500-voud haalbaar, terwijl het informatieverlies ten opzichte van piekselectie wordt geminimaliseerd. De benadering vermindert ook selectiebias en behoudt biologisch relevante signalen, waardoor de gevoeligheid en specificiteit van downstreamanalyse worden verbeterd.

In een vierde artikel presenteren we een nieuw framework voor het ontrafelen van IMS-metingen door integratie van optimalisatiebeperkingen afgeleid van microscoopdata. Deze benadering maakt gebruik van aanvullende ruimtelijke informatie om traditionele matrixfactorisatietechnieken te verbeteren, resulterend in betere signaalscheiding en ruisonderdrukking. Door microscoop-geïnformeerde optimalisatiebeperkingen toe te voegen, onderscheidt de methode effectief gemengde spectrale signalen en isoleert subtiele moleculaire verdelingen binnen complexe biologische weefsels. We belichten de

ontwikkeling en toepassing van een optimalisatie-algoritme dat het gproces regulariseert via een kleinste-kwadraten-benadering, waardoor een nauwkeurigere toewijzing van moleculaire componenten aan hun respectievelijke ruimtelijke locaties wordt verzekerd. Experimentele validatie op een synthetische en een echte IMS-dataset toont een verbeterde terugwinning van biologisch relevante signalen in vergelijking met conventionele benaderingen.

Samenvattend, met de focus op beeldvormende instrumenten die enorme, hoogdimensionale datasets genereren, pakken we de dubbele uitdagingen van dimensiereductie en ruisonderdrukking aan. Ons werk beslaat zowel biochemische imaging mass spectrometry als astronomische midden-infraroodbeeldvorming en ontwikkelt lage-rank- en schaarse modelleringsstrategieën die instrument- en experiment-specifieke structuren benutten. Deze methoden leveren meer interpreteerbare, gecomprimeerde representaties, verminderde ruis en verbeterde signaalgetrouwheid op. Door geavanceerde optimalisatietechnieken af te stemmen op domeinspecifieke behoeften, maken onze bijdragen meer nauwkeurige downstreamanalyse mogelijk en leggen ze de basis voor gevoeliger en computationeel efficiëntere beeldvormingspijpleidingen die in andere disciplines kunnen worden toegepast.

PREFACE

This dissertation marks the end of a journey that began when I first set foot in Delft in 2013. More than 60 years after my belated grandfather Roger Timmerman, I took the train towards the same station and walked over the same city cobbles. Over the past 12 years, I have had the privilege of contributing to a vibrant academic, social and sports community, guided by exceptional mentors and supported by exceptional friends and family. It is with gratitude that I look back and would like to thank you for making this possible.

First, I would like to thank my promotor, Raf Van de Plas, for his guidance, feedback, and generous support throughout different stages of my career from BSc final year's project, to MSc thesis, and finally to this dissertation. I would like to thank you for giving me these opportunities. I am also indebted to my promotor, Bart De Schutter, whose discretion, patience, and expertise were invaluable in steering me through the final challenges of this journey. In extension, I would like to thank the members of the reading committee for their valuable comments and constructive criticism and the opportunity to defend this dissertation.

Secondly, I would like to thank Jeffrey Spraggins, and the collaborators at Vanderbilt University, Kameron, Ólöf, Ally, Lauren, Claire, Chris, Heath, Melissa, Kate, Martin, Emilio, and all others. I have learned a lot from *y'all* during the weekly DA meetings, ASMS conferences and the visit at the Mass Spectrometry Research Center! Another special thanks to Lukasz for the support in the early stages and remote support at end. For the astronomical part of my journey, I would like to thank Bernhard Brandl and Alex: thank you for your support and for patiently guiding me through principles that were, quite literally, light-years away from my daily life.

Closer to home, I want to thank my office colleagues, Max Mendel and Coen. The life lessons learned in the office will never be forgotten! Special thanks to Sander (Bregman) for his 24/7 walking-in office service, the many talks, the padel games, foosball games and support when organizing events at DCSC. At the same time, also a big shout-out to Rogier and Mees, two outstanding foosball partners, I have had the honour to play with. More broadly, thanks to the whole foosball and lunch crew from the first, second, third and fourth floor. I am grateful to the DCSC secretaries for their help and support, with a special thanks to Sandra. To Paul-Louis, it has been a joy to observe your passion for mathematics and drive to share it with everyone around! Besides, the many laughs, coffees, ice creams, padel games and DP problems, you have been a big driver for me to come early to office! To Léonore, I am glad to have shared an office with you. Your collaborative spirit and loyalty have left a lasting impression on me!

Next to my colleagues, I also want to thank Timo, Jeroen, Folkert and Daniel for their support, the coffees, and endless discussions! Also a special thanks towards my friends in Belgium, in special Brecht, Dries, and Jente. Those many nights in Brussels during my Ph.D. at Xavier's place were mind-expanding and necessary for continuity.

Finally, I am profoundly grateful to my family. To my parents, Ingrid and Roger, for their unconditional love and belief in me, to my partner Hannah Roos, for constant encouragement, understanding, and for keeping me grounded. But also listening endlessly to the same boring stories over and over again. Also a special thanks to the parents of my partner, Ursula and Jan-Willem for all the support, listening to my stories, and giving me the ability to let some steam off! I cannot forget my sister Céline of course, for the many calls, fights and laughs! You are the reason, Rooibos has become my favourite tea! Last but not least, I want to thank my grandfather Roger Moens Sr. for the honest advice and critique at the different stages of my life. In the end, success appears to be the logical consequence of hard work, willpower, and perseverance.

All this support gave me the strength to persevere through challenges and moments of doubt. After 12 years in Delft, years filled with discovery, growth, and friendship, it feels fitting to close this chapter and embark on the next adventure.

Roger Amaury Rutger Moens
Den Haag, August 2025

1

INTRODUCTION

Modern imaging systems face a variety of challenges that, perhaps surprisingly, can be addressed with similar methods. Using imaging mass spectrometers, used in biochemical research, and the METIS mid-infrared instrument, used in astronomical research, as case studies, we illustrate how sensors give rise to massive, noisy data cubes, i.e., large, third-order tensor data structures containing noisy measurements. We examine the specific obstacles these data cubes present, survey promising research directions, and propose structured decompositions as effective solutions for several challenges for both instruments.

1

1.1. IMAGE-ACQUIRING INSTRUMENTS: FROM MICROSCOPIC TO TELESCOPIC SCALES

Image-acquiring instruments span a diverse set of application domains and operate on different spatial scales, from microscopes in biology at the nanometer (*i.e.*, 10^{-9} m) scale [1] to telescopes in cosmology at *arcsec*[1] scales [2], and electromagnetic spectral scales, from radio waves at meter wavelength [3] to Gamma ray waves at picometer (*i.e.*, 10^{-12} m) wavelength [4]. Yet, they all share a fundamental goal: mapping spatially resolved signals across a field of view, *i.e.*, acquiring images. In this dissertation, we are interested in two particular domains as they cover both the study of very small objects as well as the very large: (1) analytical chemistry and (2) mid-infrared astronomy. In particular, this dissertation deals with imaging datasets from two instruments: in analytical chemistry, imaging mass spectrometry (IMS) platforms, and in mid-infrared astronomy, the Mid-infrared ELT[2] Imager and Spectrograph (METIS) [6]. They are characterized by respectively extending mass spectrometry's molecular specificity into the spatial domain [7, 8], and exploiting a large aperture telescope to produce high-resolution mid-infrared images. Both modalities typically generate sets of images that are structured in discrete data cubes ($\mathcal{M}$ in **Fig. 1.1**), three-dimensional arrays/tensors whose two transverse mode correspond to spatial coordinates (x and y in **Fig. 1.1**) and whose third mode captures a spectral or temporal channel (mass-to-charge in IMS, usually inferred from time or frequency information, and time or wavelength information for METIS, ϕ in **Fig. 1.1**)).

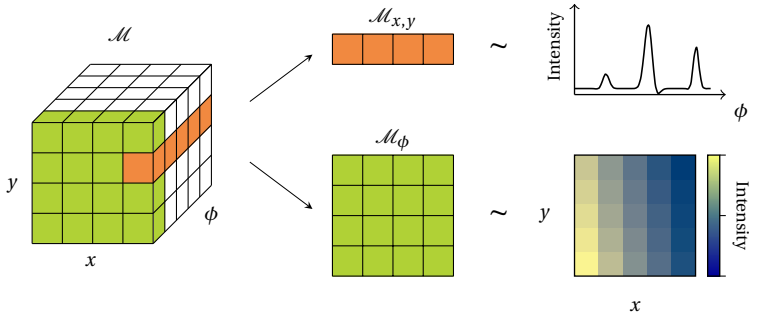

Figure 1.1: Schematic of an imaging data cube typically considered in this dissertation. A typical data cube can be represented by the tensor form (left) and contain a fiber (top right and orange color) and a slice (bottom right and green color). The variables x and y represent spatial axes, ϕ is a spectral axis.

1.1.1. IMAGING MASS SPECTROMETRY (IMS)

IMS merges mass spectrometry's label-free molecular specificity with spatial localization by raster-scanning a probe such as a laser or ion beam across a biochemical tissue sample [9, 10]. The variety of IMS that we work on, matrix-assisted laser desorption/ionization (MALDI), coats a thin tissue slice with a UV-absorbing organic matrix solution [9, 11], a laser pulse desorbs and ionizes co-crystallized analytes [9, 10] at each coordinate (x, y),

[1] The actual size will depend on the distance to the measured objects.
[2] Extremely Large Telescope [5]

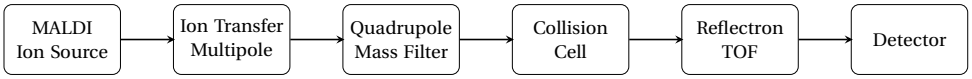

Figure 1.2: Schematic representation of the individual components of a typical MALDI-TOF instrument. The system begins with a MALDI (Matrix-Assisted Laser Desorption/Ionization) source for soft ionization of analytes. The ion beam is subsequently focused using an ion transfer multipole, followed by a quadrupole mass filter that enables selection of specific mass ranges. A collision cell removes interfering ions, such as isobaric species, before ions enter the reflectron TOF (Time-of-Flight) analyzer. The reflectron, with its extended ion path, enhances mass resolution and contributes to additional ion filtering. Finally, ions are detected based on their arrival time and intensity by a detector.

and the resulting ions are analyzed in a mass analyzer, *e.g.*, a time-of-flight (TOF) [12] or Fourier transform ion cyclotron resonance (FT-ICR) [13] mass analyzer. A schematic overview of a MALDI-TOF instrument is depicted in **Fig. 1.2**. A value at (x, y, z) thus reports an abundance of a certain molecule or set of molecules of weight z at a particular pixel location (x, y), while the output at every pixel location (x, y) is a full mass spectrum $\mathcal{M}_{x,y} \in \mathbb{R}^K$ (similar to **Fig. 1.1**). As such each experiment creates a data cube $\mathcal{M} \in \mathbb{R}^{I\times J\times K}$ (similar to **Fig. 1.1**) that encodes molecular distributions, typically of the lipidomic, glycomic, metabolomic, and proteomic content of the biological sample, at high spatial resolution, *i.e.*, the μm-scale [14]. Once each spatial location is scanned, one can reconstruct per mass-to-charge bin an ion image $\mathcal{M}_\phi \in \mathbb{R}^{I\times J}$ (similar to **Fig. 1.1**). In IMS, researchers continually strive to optimize a "golden triangle" of performance metrics: *specificity*, *sensitivity*, and *spatial resolution* [15]. Roughly speaking, they respectively correspond to the spectral axis resolution, the detection limit, and spatial axes' resolution. Enhancing one of these metrics often comes at the expense of another (whence the concept of a triangle). This expense is what will be of interest in this dissertation as it can be (partially) compensated for by methodological advances pushing these forward to achieve more precise molecular identification, lower detection limits, and finer spatial imaging details. The MALDI–IMS experiments considered in this dissertation acquire hundreds to thousands of pixels for each spatial dimension and each spectrum contains around 10^3–10^6 mass-to-charge bins. Summarizing, the IMS datasets we will consider can be represented as $\mathcal{M} \in \mathbb{R}^{I\times J\times K}$ with

- dimension I: spatial x (pixels) $\sim 10^2 - 10^3$ bins,
- dimension J: spatial y (pixels) $\sim 10^2 - 10^3$ bins,
- dimension K: mass-to-charge (m/z) $\sim 10^3 - 10^6$ bins.

Usually, $\sum_K \mathcal{M} \in \mathbb{R}^{I\times J}$ is called the total ion current image and $\sum_{I,J} \mathcal{M} \in \mathbb{R}^K$ the total ion current mass spectrum. Finally, the reader should be aware that IMS datasets are often preprocessed before we performed our analysis. Typical steps include baseline correction (*i.e.*, projection), normalization (*i.e.*, scaling), alignment (*i.e.*, translation/resampling), smoothing/denoising (*e.g.*, convolution), calibration (*i.e.*, axis transformation), and peak detection with or without accumulation (*i.e.*, subsampling with or without summation). The preprocessing specifics are provided in each particular chapter.

1

1.1.2. MID-INFRARED IMAGING: METIS

The Mid-infrared ELT Imager and Spectrograph (METIS) will be one of the first-generation instruments on ESO's Extremely Large Telescope (ELT, 39 m primary) [5], capable of direct imaging in the L, N, and M band, roughly between $3-12\ \mu m$, at milliarcsecond spatial sampling [16]. The thermal mid-infrared (mid-IR, $\lambda \approx 3-40\ \mu m$) spans a unique spectral range ideal for probing warm dust, interstellar molecules, obscured objects (*e.g.*, protostars), temperate exoplanets, solar system bodies, and distant redshifted sources. Mid-IR observations address key topics such as star formation, object composition and temperature, solar flare electron densities, and the role of molecular clouds in galactic dynamics. The imaging mode of METIS will create data cubes $\mathscr{M} \in \mathbb{R}^{I\times J\times K}$ (similar to **Fig. 1.1**) that, for our interest, encode the "intensity" of an object of interest at a particular wavelength band. The number of spatial pixels lies around 10^3 in each spatial dimension, while the temporal dimension can vary in size depending on the observational time and post-processing specifics [16]. To contend with the overwhelming thermal background originating from both the atmosphere and the warm telescope optics, METIS employs a fast, dedicated internal chopper to mitigate the highly time-variable component of the background. However, traditional nodding techniques, *i.e.*, including a primary mirror shift, which are effective against slower background variations, are not feasible on the ELT due to its size and the complexity of its dynamic, actively aligned five-mirror system [6]. In the absence of nodding, additional residuals, stemming from quasi-static instrumental artifacts, slowly varying atmospheric conditions, and evolving telescope configurations, can imprint persistent background patterns that complicate source detection, particularly in direct imaging applications [17, 18, 19]. These residuals must be addressed with alternative calibration strategies to reach the sensitivity limits necessary for detecting faint sources like exoplanets. As a result, METIS will produce large mid-infrared data cubes heavily affected by residual thermal and atmospheric noise, necessitating advanced background subtraction techniques to isolate weak astrophysical signals. Furthermore, the combination of long observational time, the large number of measured frames, and the joint presence of spatial and spectral information allows these data cubes to be analyzed along multiple dimensions, enabling customized co-addition schemes and tailored noise-mitigation pipelines that exploit the full structure of the data. As METIS has not seen first light yet, *i.e.*, has not yet been put into operation, the datasets we will consider are from VLT/VISIR [20] and SOFIA/FORCAST [21, 22] and can be represented as $\mathscr{M} \in \mathbb{R}^{I\times J\times K}$ with

- dimension I: spatial x (pixels) $\sim 10^2 - 10^3$ bins,
- dimension J: spatial y (pixels) $\sim 10^2 - 10^3$ bins,
- dimension K: temporal $\sim 10^2 - 10^4$ bins.

Usually, $\frac{1}{K}\sum_K \mathscr{M} \in \mathbb{R}^{I\times J}$ is called the time frame average. Finally, the reader should be aware that these datasets are often preprocessed before we performed our analysis. Typical steps can include co-addition (*i.e.*, summation), flat fielding (*i.e.*, projection), dark current subtraction (*i.e.*, projection), bad pixel correction (*e.g.*, projection), normalization (*i.e.*, scaling). The preprocessing specifics are provided in each particular chapter.

1.2. CURRENT CHALLENGES IN IMAGING

We extract three common challenges within our imaging applications that have a direct and significant impact on the analysis and thus interpretation of these extensive image sets: (1) dimensionality of data is high its volume large; (2) application of control strategies can be limited; (3) contamination of signal of interest by noise and (instrumental) artifacts. Note that the first challenge arises because, although the data are currently represented as a third-order tensor, the incorporation of additional acquisition modes and the subdivision of long temporal vectors into distinct dynamical modes will naturally elevate the data to higher-order tensor structures. Moreover, the overall data volume is substantial, often exceeding standard computational node memory (128–256 GB RAM).

1.2.1. HIGH DIMENSIONALITY AND THE CURSE OF DATA VOLUME

Let us consider a simple equation to evaluate this challenge: if each mode d in which we measure (with a total of n dimensions) is discretized with sampling distance ε_d across the measurement range l_d, the total number of measurements scales as:

$$\#\text{measurements} = \prod_{d=1}^{n} \left(\frac{l_d}{\varepsilon_d}\right) = \left(\frac{l_1}{\varepsilon_1}\right)\left(\frac{l_2}{\varepsilon_2}\right)\cdots\left(\frac{l_n}{\varepsilon_n}\right). \tag{1.1}$$

From Eq. 1.1, we make two observations: (1) the number of measurements grows *exponentially* in n; (2) an increase in a measurement range, l_d, or decrease in a single sampling distance, ε_d, leads to a *linear* increase in the number of measurements. The first observation is what is called a "curse of dimensionality" [23]. While for this thesis the implication fo the curse of

In IMS, the dimensionality may increase through the addition of a third spatial axis [24], the incorporation of an ion mobility separation [15] (*e.g.*, MALDI–timsTOF), or the integration of other complementary molecular-discriminatory modalities [25]. METIS, by contrast, is currently confined to two spatial axes and one spectral axis.

As dimensionality n increases, classical statistical and distance-based methods degrade in performance due to the "curse of dimensionality" [23], noise accumulation [26], and the phenomenon of concentration of measure effects rendering popular distance measures, *e.g.*, based on ℓ_p-norms, ineffective [27]:

$$\text{if} \quad \lim_{i\to\infty} \mathrm{Var}\left(\frac{\|X_i\|}{\mathbb{E}[\|X_i\|]}\right) = 0, \quad \text{then} \quad \frac{D_{\max} - D_{\min}}{D_{\min}} \to 0, \tag{1.2}$$

where i is the dimensionality of the data space, $\mathscr{F}$ is a 1-dimensional data distribution in $(0,1)$, X_d is a data point from $\mathscr{F}^d$, were each coordinate follows $\mathscr{F}$, $D_{\max}$ is the maximal distance of a data point to the origin, $D_{\min}$ is the minimal distance of a data point to the origin, and $\|\cdot\|$ is an ℓ_p-norm. This implies that the ℓ_p-norms between points become nearly indistinguishable, undermining, *e.g.*, nearest-neighbor search and clustering tasks in high-dimensional spaces.

Besides dimensionality, the total volume of data poses serious challenges. Under IMS' "golden triangle", and similar goals for METIS, the measurement range and resolution are altered, yielding a linear, rather than exponential, increase in the total number of measurements. Yet, in imaging experiments this already may increasingly produce tera-

to petascale datasets. For example, a single METIS time series will be able to exceed 10^{12} samples in total, while a MALDI–timsTOF IMS run can produce tens of terabytes per tissue section. This brings practical problems to storing, transferring, analyzing, visualizing and interpretating of the data in a current lab set-up. Storing, transferring, and sequentially accessing these data can overwhelm memory-bound algorithms and saturate I/O pipelines. Processing such volumes in a single batch is often infeasible, necessitating batch processing, streaming, or tiled methods [28, 29]. Moreover, the problem is intensifying. As instruments become more sensitive and efficient, data acquisition scales up, paralleling Jevons's paradox[3] [30].

1.2.2. LIMITED CONTROL POSSIBILITIES: OPEN-LOOP ACQUISITION

The application of classical control strategies to reduce imperfections in imaging-capable mass spectrometers and the METIS instrument can encounter both physical and practical limits. In many IMS setups for example, real-time control is effectively impossible, proprietary hardware and software lock down the system, and the ionization and detection process itself must run open-loop, so we have to rely entirely on "open-loop" output. However, for some of its individual components (see **Fig. 1.2**) control strategies have been designed and are being applied (see *e.g.*, [31, 32]). On the other hand, deformable mirrors and wavefront sensors in astronomical adaptive optics cannot correct distortions faster or smaller than their finite temporal response and spatial bandwidth, and even the brightest guide stars may be too faint or present aberration modes outside the sensor's view, leaving residual errors that cannot be driven to zero without sacrificing closed-loop stability. For both IMS and METIS, instrument-induced artifacts can also introduce patterns that were never anticipated, further complicating correction [33]. In those and some other cases, pre-processing these datasets becomes essential: by combining sets of acquired images, exploiting statistical priors about noise and system behaviour, and requiring desired structure in reconstructions, we can still recover information that we need for discovery.

1.2.3. NOISE AND ARTIFACTS

Both IMS and METIS datasets are subject to noise and a host of artifacts, ranging from instrumental drifts to sample-specific effects, that undermine classical processing pipelines unless explicitly modelled (see *e.g.*, [33, 34, 35]). IMS measurements introduce additional, sample-specific artifacts, *e.g.*, variations in matrix crystal deposition (*i.e.*, "matrix effects") creating inhomogeneous ionization patterns [36], high-abundance species inducing ion suppression [37], detector dead-time distortions skewing or lowering intensity values [38], and lipid delocalization, lateral diffusion of analytes during sample preparation, blurring true molecular distributions [39]. In METIS, incomplete chopping may leave a spatially correlated thermal background whose residual fluctuations can mimic faint point sources, while thermal drift over time shifts baselines and produces spurious features if not corrected [33]. In general, detector readout (electronic) noise and photon-counting (shot) noise add random fluctuations and occasional outliers, further degrading sensitivity to weak signals and causing feature selection or filtering algorithms to fail. As these noise sources propagate through normalization, feature extraction, and statistical testing, they

[3] Jevons's paradox states that increased efficiency in resource use, can lead to greater overall consumption of that resource.

amplify false positives, bias normalization factors, and obscure genuine features. Sporadic spikes or drops in intensity generate outliers that can warp multivariate decompositions and undermine reproducibility.

1.3. RESEARCH DIRECTIONS

With the main challenges in mind, *i.e., high dimensionality and large data volume, limited control possibilities,* and *interference of noise and artifact,* multiple research avenues emerge. In our effort to define a unifying solution, we structure our approach by first examining models and outlining considerations. As such, we will arrive to the main topic of this dissertation: *structured decompositions for large-scale imaging datasets.*

1.3.1. MODEL AND METHOD CONSIDERATIONS

Our methods must balance three forces: (i) the need for expressive models that capture the physics of signal formation under limited control, (ii) the computational tractability demanded by large volume data, and (iii) robustness to complex noise and artifacts. We address each in turn by constraining models, representations, and priors.

LINEAR MIXING MODELS

In general, we can model our measurements as

$$\mathcal{M} = \mathcal{A}(\mathcal{X}) + \mathcal{E}, \tag{1.3}$$

where $\mathcal{M} \in \mathbb{R}^{I\times J\times K}$ is the measured data cube (given), $\mathcal{X} \in \mathbb{R}^{I\times J\times K}$ is a latent data cube, often of interest (to be found), $\mathcal{A}$ is a forward operator acting on the $I \times J \times K$-tensor space (to be found or given/constructed), and $\mathcal{E} \in \mathbb{R}^{I\times J\times K}$ is a residual term. The forward operator $\mathcal{A}$ will be assumed to behave linearly, as this usually leads to favourable properties such as

- *convexity*: leading to global, sometimes unique optima and to usage of mature solvers,
- *interpretability*: modes of $\mathcal{A}$ can correspond to a "pure" response (*e.g.*, a (known) tissue-specific spectrum or point-source point spread function),
- *scalability*: matrix-based routines exploit highly-optimized linear-algebra libraries (*e.g.*, LAPACK [40]),
- *sufficiency for mild non-linearities*: small detector saturation or ion-suppression can be linearized or treated as residuals in $\mathcal{E}$.

Limiting ourselves to linear mixing balances model fidelity against the computational complexity that fully non-linear physics would introduce, which is critical when our data volume already overwhelms storage, I/O, and memory.

TENSOR VS. MATRIX REPRESENTATIONS AND MODE FLATTENING

Our imaging data are naturally expressed as order-3 tensors $\mathcal{M} \in \mathbb{R}^{I\times J\times K}$ (see **Fig. 1.1**). However, instead, we will flatten our tensor by unfolding along one mode, usually the

spectral dimension (in this case denoted by a subscript 3):

$$M := \mathcal{M}_{(3)} \in \mathbb{R}^{K\times(IJ)}. \tag{1.4}$$

Working with this matrix M brings several benefits. Firstly, it lets us tap directly into mature matrix-analysis tools [41], *e.g.*, the singular value decomposition and other well-studied factorizations [42, 43, 44], with their proven convergence guarantees and highly optimized implementations. Secondly, the algorithmic and theoretical machinery for matrices is often simpler than that for higher-order tensors [45, 46]. Finally, this unfolding acts as a natural bridge back to full tensor methods: usually it provides a good initializations for tensor decompositions [47, 48], so we retain a clear pathway to modelling genuine multi-way structures.

As such, we will focus in this dissertation on a reformulation of Eq. 1.3, namely

$$M = \mathcal{A}(X) + E, \tag{1.5}$$

where $M \in \mathbb{R}^{K\times(IJ)}$ and latent matrix $X \in \mathbb{R}^{K\times(IJ)}$, and $\mathcal{A}$ is a linear operator acting on the space of $K \times (IJ)$-matrices. Again, $E \in \mathbb{R}^{K\times(IJ)}$ is considered a residual term. Furthermore, this formulation will allow us to include prior information into a single framework.

PRIOR INFORMATION

The robustness of such unmixing models under limited control and complex instrumental or environmental artifacts crucially depends on the degree and form of prior or supervisory information injected into the model.

Deterministic base signals We treat each pure component, *i.e.*, each column of the latent matrix X, as a fixed, reproducible system response. Physically, this means that whenever a given source spectrum is measured by the instrument, the observed spectral shape or point spread function remains identical.

Structured priors Instrumental and sample backgrounds can sometimes exhibit a low-dimensional structure, whereas true signals are localized and auxiliary reference data can guide model selection. For example, in METIS mid-infrared imaging, thermal emission from optics and telescope surfaces varies smoothly across the field, and in IMS of biological tissue, the structure of functional tissue units is usually spatially coherent. At the same time, exoplanet point sources in METIS and molecular biomarkers in IMS could appear only in a few spatial pixels or spectral channels. We therefore identify a set of matrix properties and structures:

- *low rank:* the latent matrix $X \in \mathbb{R}^{O\times P}$ is constrained to $\text{rank}(X) \ll \min(O, P)$. In image processing this models the fact that smooth or slowly-varying signals lie in a low-dimensional subspace of the full pixel space. By enforcing low rank, we potentially reduce the effective high dimensionality of the problem, saving memory and computation when dealing with very large image sets.
- *sparsity:* the latent term X or residual term E can be constrained via $\|X\|_0 \ll OP$ or $\|E\|_0 \ll IJK$, where $\|A\|_0 := \#\{(i, j) \mid A_{ij} \neq 0\}$. In images this captures the fact that

only a small fraction of pixels (*e.g.*, edges, point sources, defects) differ from the background. Enforcing sparsity isolates those features, reduces sensitivity to noise, or could further limit the number of active variables, making analysis of large-scale or high-resolution data tractable.

- *non-negativity:* both $\mathscr{A}$ and X can be constrained to lie in a non-negative cone, reflecting non-negativity concentrations, intensities, or mixtures.
- *spectral smoothness:* penalizing "large" second-derivatives in, *e.g.*, each column of X to can enforce physically plausible, smooth spectral profiles.
- *labelling:* a small library of reference spectra or a handful of annotated spatial regions of interest could seed (initial) structures of $\mathscr{A}$ steering algorithms away from degenerate solutions.

DATA VOLUME VS. INFORMATION SCOPE

To respect data-volume constraints, we distinguish four data processing regimes that each have their advantages and disadvantages:

- **Global (batch) methods**: exploit all pairwise correlations for maximal accuracy, but incur often a large computational cost of $\mathscr{O}(n^3)$ for matrix dimension $(n \times n)$.
- **Streaming (online) methods**: process incoming spectra or frames one at a time with a constant memory footprint, though they may suffer from subspace drift, which is problematic for recovering the latent matrix, over time.
- **Tiled (local) methods**: partition data into overlapping blocks to limit per-tile computations, followed by reconciliation steps to mitigate edge effects.
- **Randomized methods**: randomized strategies, such as sketching, subsampling, or randomized projections, aim to reduce memory and computational footprint while approximatively preserving essential information.

1.3.2. STRUCTURED AND LOW-RANK DECOMPOSITIONS

Following our linear mixing model, structured as in Eq. 1.4, we can model our imaging data, enabling the use of prior information, *i.e.*, known structures. As such, structured decompositions tackle the three core challenges in a single framework:

- **Mitigating the curse of dimensionality and reducing large data volume.** By forcing the dominant variations in our data cube to lie in a low-dimensional subspace, low-rank constraints collapse billions of measurements into a compact representation. This directly reduces both storage and computational complexity, converting an intractable high-dimensional problem into the analysis of a small set of "dominant" modes. Besides, our linear model formulation admits a spectrum of solvers. This versatility lets us accommodate terabyte-scale image data sets.
- **Leveraging open-loop data via offline processing.** Since in our case in-loop feedback is infeasible, we instead collect all raw frames under open-loop conditions and defer correction to a post-acquisition stage.

- **Robustly suppressing noise and artifacts.** Random fluctuations and sporadic outliers degrade classical pipelines. By modelling them, as a sparse residual, structured decompositions automatically segregate unstructured noise from a coherent signal, improving downstream tasks such as peak detection, feature extraction, and statistical testing. Further constraints, *e.g.*, non-negativity, spectral smoothness, and library-guided priors, enable encoding physical knowledge and sharpen separation.

As we develop novel methods and algorithms for thermal-background suppression in METIS and full-profile compression in IMS, this unified framework of structured decompositions ensures that each advance addresses all three challenges, laying a solid foundation for large-scale imaging analysis.

1.4. DISSERTATION OVERVIEW AND CONTRIBUTIONS

This dissertation is organized in a paper-based format: Chapters 2-5 each take the form of a stand-alone manuscript that has been submitted or accepted for publication. All code implementations, hyper-parameters, variables, and other necessary prerequisites are fully defined within their respective chapters.

CHAPTER 2: SPARSE OUTLIERS AS OBSTRUCTION

We propose a low-rank modelling framework that incorporates sparse residuals to improve compression and denoising of MALDI IMS data. By explicitly separating sparse biological signals from dense noise, our approach yields lower-dimensional representations than traditional Principal Component Analysis (PCA) while preserving key features. Applied to human eye tissue datasets, it achieves stronger compression and more interpretable component images. Our main contribution is demonstrating that sparsity-separating low-rank methods outperform PCA for imaging mass spectrometry.

This chapter is based on the following publication:

> Moens, R. A. R., Migas, L. G., Anderson, D. M. G., Messinger, J. D., Ovchinnikova, O. S., Caprioli, R. M., Spraggins, J. M., & Van de Plas, R. (2025). Advanced Dimensionality Reduction for Imaging Mass Spectrometry of Human Eye Tissue Through Low-Rank Modeling with Sparse and Dense Residuals. *Analytical Chemistry*, 97.42, 23040-23049.

CHAPTER 3: SPARSE OUTLIERS AS SOLUTION

We introduce a method that models mid-infrared imaging data as a combination of a low-rank thermal background and sparse point sources, enabling robust background subtraction without requiring classical chopping or nodding. By decomposing each observation into a low-rank background term, a sparse source term, and noise, our approach more effectively isolates faint astrophysical signals. Applied to VISIR and SOFIA datasets, it achieves a substantial reduction in background flux compared to traditional methods (by factors of 3-10) while preserving source flux, leading to improved photometric accuracy and detection precision. The primary contribution is demonstrating that low-rank plus sparse modelling significantly enhances background reduction and source recovery

in mid-infrared astronomy.

This chapter is based on the following work:

> Moens, R. A. R., Pietrow, A. G. M., Brandl, B., & Van de Plas, R. (2025). Thermal Background Reduction for Mid-Infrared Imaging by Low-Rank Background and Sparse Point Source Modelling [Unpublished Manuscript, Submitted to Astronomy & Astrophysics].

CHAPTER 4: STRUCTURED DECOMPOSITIONS ON A TERABYTE SCALE

We propose a sparsity-aware low-rank matrix factorization framework that models missing-value IMS data via matrix completion, avoiding traditional peak picking. By combining sparse-format-aware singular value thresholding and fixed-point continuation, our method preserves full spectral profiles, including low-intensity and near-isobaric peaks, while achieving compression factors comparable to peak picking. Applied to large-scale FT-ICR and MALDI-TOF IMS datasets, it reduces full-spectrum information loss by up to 40% compared to peak picking, offering richer representations for downstream analysis.

This chapter is based on the following publication:

> Moens, R. A., Migas, L. G., Van Ardenne, J. M., Skaar, E. P., Spraggins, J. M., & Van de Plas, R. (2025). Preserving Full Spectrum Information in Imaging Mass Spectrometry Data Reduction. *Bioinformatics, 41*(5), btaf247.

CHAPTER 5: INCLUDING STRUCTURED PRIORS ON A TERABYTE SCALE

We propose a tissue-informed inverse-problem framework that "unmixes" blended IMS pixel spectra into component spectra of annotated regions (*e.g.*, single cells or functional tissue units) by leveraging high-resolution microscopy boundaries. Incorporating non-negativity, sparsity, and low-rank priors, we solve both overdetermined and underdetermined linear mixing models, outperforming ordinary least squares, non-negative least squares, and singular value thresholding in recovering biologically meaningful profiles. Applied to synthetic single-cell data and a large HuBMAP kidney FTU dataset, it yields more accurate region-specific spectra compared to weighted schemes and scales to millions of pixels without requiring singular value decompositions or explicit pseudo-inversion.

This chapter is based on the following work:

> Moens, R. A. R., Patterson, N.H., Migas, L.G., Esselman, A.B., Moser, F.A., Spraggins, J.M. & Van de Plas, R. (2025). Unmixing of Imaging Mass Spectrometry Measurements Using Microscopy-informed Constraints [Unpublished Manuscript].

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

Chapter 6 concludes with a summary of findings, practical guidelines for practitioners, and final remarks on the broader impact of structured decomposition in imaging science.

It synthesizes the lessons learned from Chapters 2–5 and provides recommendations for future work in scalable, full-spectrum signal processing across diverse high-dimensional imaging modalities.

REFERENCES

[1] S. W. Hell. Far-field Optical Nanoscopy. In: *Science* 316.5828 (2007), pp. 1153–1158.

[2] J. P. Gardner, J. C. Mather, M. Clampin, R. Doyon, M. A. Greenhouse, H. B. Hammel, J. B. Hutchings, P. Jakobsen, S. J. Lilly, K. S. Long, et al. The James Webb Space Telescope. In: *Space Science Reviews* 123 (2006), pp. 485–606.

[3] M. P. van Haarlem, M. W. Wise, A. Gunst, G. Heald, J. P. McKean, J. W. Hessels, A. G. de Bruyn, R. Nijboer, J. Swinbank, R. Fallows, et al. LOFAR: The Low-frequency Array. In: *Astronomy & Astrophysics* 556 (2013), A2.

[4] J. A. Sorenson, M. E. Phelps, et al. *Physics in Nuclear Medicine*. Grune & Stratton New York, 1987.

[5] R. Gilmozzi and J. Spyromilio. The European Extremely Large Telescope (E-ELT). In: *The Messenger* 127.11 (2007), p. 3.

[6] B. R. Brandl, R. Lenzen, E. Pantin, A. Glasse, J. Blommaert, M. Meyer, M. Guedel, L. Venema, F. Molster, R. Stuik, et al. METIS: the Thermal Infrared Instrument for the E-ELT. In: *Ground-based and Airborne Instrumentation for Astronomy IV*. Vol. 8446. SPIE. 2012, pp. 554–566.

[7] L. A. McDonnell and R. M. Heeren. Imaging Mass Spectrometry. In: *Mass Spectrometry Reviews* 26.4 (2007), pp. 606–643.

[8] R. M. Caprioli, T. B. Farmer, and J. Gile. Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. In: *Analytical Chemistry* 69.23 (1997), pp. 4751–4760.

[9] R. M. Heeren, D. F. Smith, J. Stauber, B. Kükrer-Kaletas, and L. MacAleese. Imaging Mass Spectrometry: Hype or Hope? In: *Journal of the American Society for Mass Spectrometry* 20.6 (2009), pp. 1006–1014.

[10] R. M. Caprioli. Imaging Mass Spectrometry: a Perspective. In: *Journal of Biomolecular Techniques: JBT* 30.1 (2019), p. 7.

[11] D. S. Cornett, M. L. Reyzer, P. Chaurand, and R. M. Caprioli. MALDI Imaging Mass Spectrometry: Molecular Snapshots of Biochemical Systems. In: *Nature Methods* 4.10 (2007), pp. 828–833.

[12] B. Mamyrin. Time-of-flight Mass Spectrometry (Concepts, Achievements, and Prospects). In: *International Journal of Mass Spectrometry* 206.3 (2001), pp. 251–266.

[13] A. G. Marshall, C. L. Hendrickson, and G. S. Jackson. Fourier Transform Ion Cyclotron Resonance Mass Spectrometry: a Primer. In: *Mass Spectrometry Reviews* 17.1 (1998), pp. 1–35.

[14] R. S. Young, A.-K. Piper, L. McAlary, J. C. McKinnon, J. S. Lum, J. Soltwisch, M. Niehaus, and S. R. Ellis. Subcellular Mass Spectrometry Imaging of Lipids and Nucleotides Using Transmission Geometry Ambient Laser Desorption and Plasma Ionisation. In: *bioRxiv preprint bioRxiv:2025.05.13.653655* (2025).

[15] J. M. Spraggins, K. V. Djambazova, E. S. Rivera, L. G. Migas, E. K. Neumann, A. Fuetterer, J. Suetering, N. Goedecke, A. Ly, R. Van de Plas, et al. High-performance Molecular Imaging with MALDI Trapped Ion-mobility Time-of-flight (TimsTOF) Mass Spectrometry. In: *Analytical Chemistry* 91.22 (2019), pp. 14552–14560.

[16] B. R. Brandl, F. Bettonvil, R. van Boekel, A. Glauser, S. Quanz, O. Absil, A. Amorim, M. Feldt, A. Glasse, M. Güdel, et al. METIS: The Mid-infrared ELT Imager and Spectrograph. In: *The Messenger* (2021), pp. 22–26.

[17] L. Burtscher, I. Politopoulos, S. Fernández-Acosta, T. Agocs, M. van den Ancker, R. van Boekel, B. Brandl, H.-U. Käufl, E. Pantin, A. G. Pietrow, et al. Towards a Physical Understanding of the Thermal Background in Large Ground-based Telescopes. In: *Ground-based and Airborne Instrumentation for Astronomy VIII*. Vol. 11447. SPIE, 2020, pp. 1678–1694.

[18] D. Petit Dit De La Roche, M. van den Ancker, M. Kissler-Patig, V. Ivanov, and D. Fedele. New Constraints on the HR 8799 Planetary System From Mid-infrared Direct Imaging. In: *Monthly Notices of the Royal Astronomical Society* 491.2 (2020), pp. 1795–1799.

[19] K. Wagner, A. Boehle, P. Pathak, M. Kasper, R. Arsenault, G. Jakob, U. Käufl, S. Leveratto, A.-L. Maire, E. Pantin, et al. Imaging Low-mass Planets Within the Habitable Zone of α Centauri. In: *Nature Communications* 12.1 (2021), p. 922.

[20] Y. Rio, P.-O. Lagage, D. Dubreuil, G. A. Durand, C. Lyraud, J.-W. Pel, J. C. de Haas, A. Schoenmaker, and H. Tolsma. VISIR: the Mid-infrared Imager and Spectrometer for the VLT. In: *Infrared Astronomical Instrumentation*. Vol. 3354. SPIE. 1998, pp. 615–626.

[21] A. Krabbe. SOFIA Telescope. In: *Airborne Telescope Systems*. Vol. 4014. SPIE. 2000, pp. 276–281.

[22] T. L. Herter, J. D. Adams, G. E. Gull, J. Schoenwald, L. D. Keller, B. E. Pirger, C. P. Henderson, G. J. Stacey, T. Nikola, J. M. De Buizer, W. D. Vacca, and K. Ennico. FORCAST: A Mid-Infrared Camera for SOFIA. In: *Journal of Astronomical Instrumentation* 7.4, 1840005-451 (2018), pp. 1840005–451.

[23] D. L. Donoho et al. High-dimensional Data Analysis: The Curses and Blessings of Dimensionality. In: *AMS Math Challenges Lecture* 1.2000 (2000), p. 32.

[24] A. D. Palmer and T. Alexandrov. Serial 3D Imaging Mass Spectrometry at Its Tipping Point. In: *Analytical Chemistry* 87.8 (2015), pp. 4055–4062.

[25] A. R. Buchberger, K. DeLaney, J. Johnson, and L. Li. Mass Spectrometry Imaging: a Review of Emerging Advancements and Future Insights. In: *Analytical Chemistry* 90.1 (2018), p. 240.

[26] J. Fan, F. Han, and H. Liu. Challenges of Big Data Analysis. In: *National Science Review* 1.2 (2014), pp. 293–314.

[27] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When Is "Nearest Neighbor" Meaningful? In: *International Conference on Database Theory*. Springer. 1999, pp. 217–235.

[28] X. Ruan, M. Mueller, G. Liu, F. Görlitz, T.-M. Fu, D. E. Milkie, J. L. Lillvis, A. Kuhn, J. Gan Chong, J. L. Hong, et al. Image Processing Tools for Petabyte-scale Light Sheet Microscopy Data. In: *Nature Methods* 21 (2024), pp. 1–11.

[29] T. Pietzsch, S. Saalfeld, S. Preibisch, and P. Tomancak. BigDataViewer: Visualization and Processing for Large Image Data Sets. In: *Nature Methods* 12.6 (2015), pp. 481–483.

[30] R. York and J. A. McGee. Understanding the Jevons Paradox. In: *Environmental Sociology* 2.1 (2016), pp. 77–87.

[31] A. Gapeev, A. Berton, and D. Fabris. Current-controlled Nanospray Ionization Mass Spectrometry. In: *Journal of the American Society for Mass Spectrometry* 20 (2009), pp. 1334–1341.

[32] S. Wright, R. R. Syms, S. O'Prey, G. Hong, and A. S. Holmes. Comparison of Ion Coupling Strategies for a Microengineered Quadrupole Mass Filter. In: *Journal of the American Society for Mass Spectrometry* 20.1 (2011), pp. 146–156.

[33] J. Sauter, W. Brandner, J. Heidt, and F. Cantalloube. Detection Limits of Thermal-infrared Observations with Adaptive Optics. I. Observational Data. In: *Publications of the Astronomical Society of the Pacific* 136.9 (2024), p. 095001.

[34] B. Balluff, C. Hopf, T. Porta Siegel, H. I. Grabsch, and R. M. Heeren. Batch Effects in MALDI Mass Spectrometry Imaging. In: *Journal of the American Society for Mass Spectrometry* 32.3 (2021), pp. 628–635.

[35] T. Boskamp, D. Lachmund, R. Casadonte, L. Hauberg-Lotte, J. H. Kobarg, J. Kriegsmann, and P. Maass. Using the Chemical Noise Background in MALDI Mass Spectrometry Imaging for Mass Alignment and Calibration. In: *Analytical Chemistry* 92.1 (2019), pp. 1301–1308.

[36] W. J. Perry, N. H. Patterson, B. M. Prentice, E. K. Neumann, R. M. Caprioli, and J. M. Spraggins. Uncovering Matrix Effects on Lipid Analyses in MALDI Imaging Mass Spectrometry Experiments. In: *Journal of Mass Spectrometry* 55.4 (2020), e4491.

[37] A. J. Taylor, A. Dexter, and J. Bunch. Exploring Ion Suppression in Mass Spectrometry Imaging of a Heterogeneous Tissue. In: *Analytical Chemistry* 90.9 (2018), pp. 5637–5645.

[38] B. J. Tyler and R. E. Peterson. Dead-time Correction for Time-of-flight Secondary-ion Mass Spectral Images: a Critical Issue in Multivariate Image Analysis. In: *Surface and Interface Analysis* 45.1 (2013), pp. 475–478.

[39] F. Fournelle, E. Yang, M. Dufresne, and P. Chaurand. Minimizing Visceral Fat Delocalization on Tissue Sections with Porous Aluminum Oxide Slides for Imaging Mass Spectrometry. In: *Analytical Chemistry* 92.7 (2020), pp. 5158–5167.

[40] E. Anderson, Z. Bai, C. Bischof, L. S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, et al. *LAPACK Users' Guide*. SIAM, 1999.

[41] V. Kuleshov, A. Chaganty, and P. Liang. Tensor Factorization via Matrix Factorization. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Vol. 38. Proceedings of Machine Learning Research. PMLR, 2015, pp. 507–516.

[42] G. H. Golub and C. F. Van Loan. *Matrix Computations*. JHU Press, 2013.

[43] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, 2022.

[44] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. John Wiley & Sons, 2009.

[45] V. De Silva and L.-H. Lim. Tensor Rank and the Ill-posedness of the Best Low-rank Approximation Problem. In: *SIAM Journal on Matrix Analysis and Applications* 30.3 (2008), pp. 1084–1127.

[46] C. J. Hillar and L.-H. Lim. Most Tensor Problems Are NP-hard. In: *Journal of the ACM (JACM)* 60.6 (2013), pp. 1–39.

[47] L. De Lathauwer, B. De Moor, and J. Vandewalle. On the Best Rank-1 and Rank-(r 1, R 2,..., Rn) Approximation of Higher-order Tensors. In: *SIAM Journal on Matrix Analysis and Applications* 21.4 (2000), pp. 1324–1342.

[48] T. G. Kolda and B. W. Bader. Tensor Decompositions and Applications. In: *SIAM Review* 51.3 (2009), pp. 455–500.

2

Sparse Outliers as Obstruction

Imaging mass spectrometry (IMS) yields high-dimensional and large datasets commonly exceeding 100 000 pixels, each reporting a mass spectrum of 200 000 intensity values or more. Reducing the dimensionality and size of IMS data is often necessary to enable downstream analysis, and matrix factorization-based approaches are often used for this purpose. However, the model underlying most of these techniques, decomposing measurements into the sum of a low-rank term (presumed signal) and a small entry-wise residuals term (presumed noise), is often not optimal for IMS. For example, while spatially or spectrally sparse signals are common in IMS data, they can heavily distort the low-rank approximation. Therefore, we propose capturing IMS data structure using low-rank models that, in addition to a dense residual, allow for sparse variation to be captured separately. We implement two such methods, principal component pursuit (PCP) and stable principal component pursuit (SPCP), apply them to IMS data, and compare them to a classical factorization method, principal component analysis (PCA). We investigate their dimensionality and noise reduction performance on MALDI Q-TOF IMS measurements of human cornea and retina tissue, since the human eye is a complex organ with lots of small, tightly packed tissue substructures that are spatially sparse. Our results suggest that, if parameters are set adequately, PCP and SPCP enable stronger dimensionality reduction and higher compression of IMS data compared to PCA while concurrently reducing signal overestimation.

The contents of this chapter are based on:

> Moens, R. A. R., Migas, L. G., Anderson, D. M. G., Messinger, J. D., Ovchinnikova, O. S., Caprioli, R. M., Spraggins, J. M., & Van de Plas, R. (2025). Advanced Dimensionality Reduction for Imaging Mass Spectrometry of Human Eye Tissue Through Low-Rank Modeling with Sparse and Dense Residuals. *Analytical Chemistry*, 97.42, 23040-23049.

2.1. INTRODUCTION

Imaging mass spectrometry (IMS) is an untargeted molecular imaging technique that measures the spatial distributions of a broad set of molecular species concurrently and throughout a sample [1, 2], with samples including biofilms [3], plant material [4], mammalian [5], and human tissue [6]. Different instrumental setups are used, with ionization sources ranging from desorption electrospray ionization (DESI) [7, 8] and matrix-assisted laser desorption/ionization (MALDI) [1] to laser ablation inductively coupled plasma (LAICP) [9] and secondary ion mass spectrometry (SIMS) [10], and mass analyzers based on time-of-flight (TOF) [11], Orbitrap [12], and Fourier transform ion cyclotron resonance (FTICR) [13] principles. While this chapter focuses on MALDI Q-TOF IMS for tissue samples, the described methods can be applied to other IMS experiment types as well.

MALDI Q-TOF IMS experiments commonly acquire more than 100 000 pixels. Each pixel reports a full mass spectrum, usually entailing more than 200 000 intensity values, resulting in more than 20 billion scalar ion intensity values per experiment. If stored exhaustively, these dataset sizes regularly exceed tens to hundreds of gigabytes per experiment. The large size and high-dimensional nature of IMS measurements often necessitate the use of dimensionality reduction techniques to condense the number of dimensions to keep track of, while incurring minimal loss of information. Dimensionality reduction is particularly important for aiding human interpretation of IMS data, denoising, and avoiding issues related to the curse of dimensionality [14] in downstream computational analysis.

Low-rank approximation is a common approach for IMS dimensionality reduction, utilized, *e.g.*, for feature selection, feature extraction, denoising, and visualization of trends that underlie high-dimensional measurements [15]. Principal component analysis (PCA) and non-negative matrix factorization (NMF) are typical low-rank modeling approaches that have delivered compelling results [16, 17, 18]. However, they often provide suboptimal representations when the IMS measurements contain sparse signals in addition to dense patterns. By sparse signals, we mean non-correlating signals that only appear in a few mass bins (*i.e.*, spectrally sparse) or in relatively small tissue areas (*i.e.*, spatially sparse). Sparse signals can occur, *e.g.*, due to sample preparation or instrument-induced noise perturbations or non-linear mixing effects in the underlying chemistry, but they can also be genuinely biological in nature due to the sample's content, structure, and orientation. For example, TOF-SIMS data are sometimes relatively sparse in the spatial domain due to limited ion yields per pixel, while Orbitrap and FTICR data often exhibit high sparsity in the spectral domain due to their high mass resolution and low noise baseline.

We propose capturing such structure inside IMS data more precisely using low-rank models that, in addition to a dense residual, allow for sparse variation to be captured separately. Such models offer advantages over traditional low-rank approaches, including better handling of outliers [19] and counteracting noise accumulation [14]. It allows them to deal more effectively with IMS-inherent sparsity, enables closer modeling of the true underlying data structure, and avoids loss of sparse biological signals in the process. These advantages are essential to approximating IMS measurements with sparse patterns well, and to factorization of full profile mass spectrometry data in general. These models have also been successfully applied in other domains such as video background

subtraction and collaborative filtering, underscoring their broad utility in capturing structured low-rank and sparse variation.

The methods are demonstrated on human eye tissue, specifically cornea and retina, for dimensionality reduction and noise reduction purposes. These tissues exhibit many small and fine-grained tissue substructures [20] that report inherently sparse intensity variations and thus are challenging to capture with traditional models (despite reporting genuine biological variation).

This chapter (a) explores how low-rank models with sparse residuals, in particular, principal component pursuit (PCP) [19] and stable principal component pursuit (SPCP) [21], can be applied to MALDI Q-TOF IMS data, (b) compares their outcomes to traditional approaches such as PCA, and (c) assesses the trade-offs required when modeling. We first introduce the two IMS datasets and the mathematical models behind PCP and SPCP, followed by a discussion of the obtained results and a concluding review of the use cases for this type of models.

2.2. DATASETS AND METHODS

Two MALDI Q-TOF IMS datasets from the human cornea and human retina tissue section are used to examine and compare the PCA, PCP, and SPCP dimensionality reduction models (Figs. 2.1 and 2.2). Figure 2.1 shows the total ion count spectrum and total ion count image of the cornea dataset, and Figure 2.2 shows the same for the retina dataset. A detailed description of these datasets and their preprocessing is provided in the Supplementary Information of this chapter..

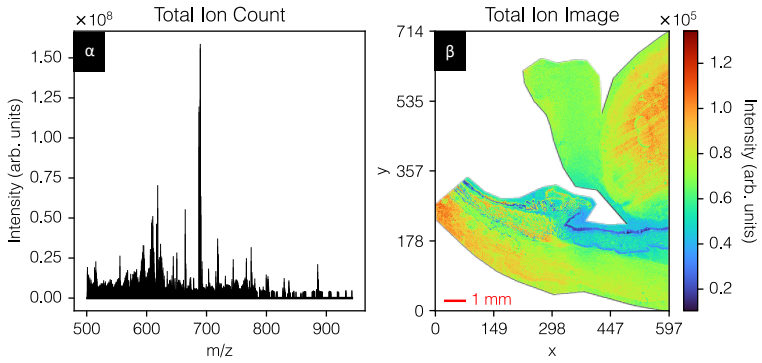

Figure 2.1: Human cornea dataset – total ion count spectrum (panel α) and total ion count image (panel β). The dataset entails 235 218 pixels by 2 381 peaks. The raw data has been exported, *m/z* aligned, calibrated, 5-95% TIC-normalized, and peak-picked. Fine sparse layers, *e.g.*, cornea, lens, ciliary processes, and iris, can be observed. An H&E stain is provided in Suppl. Fig. S2.1.

2.2.1. FACTORIZATION-BASED DIMENSIONALITY REDUCTION METHODS

We explore two dimensionality reduction models that extend the low-rank approximation model with an explicit and separate sparse residuals term, and we investigate how these methods fare in terms of capturing signal from IMS data and avoiding sparse signal loss.

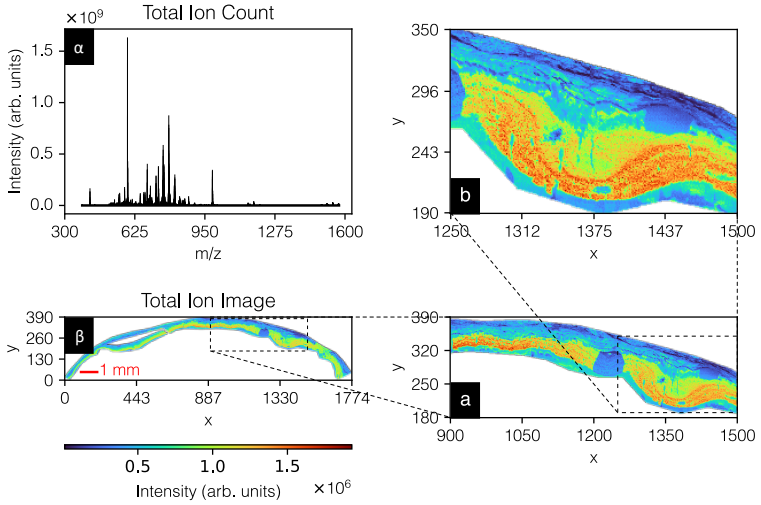

Figure 2.2: Human retina dataset – total ion count spectrum (panel α) and total ion count image (panel β). The dataset entails 137 923 pixels by 3 212 peaks. The raw data has been exported, m/z aligned, calibrated, 5-95% TIC-normalized, and peak-picked. Panels (a) and (b) provide zoomed-in views, revealing fine tissue structures including inner retina, choroid, and sclera. A brightfield image is provided in Suppl. Fig. S2.2.

The proposed methods are principal component pursuit (PCP) [19] and stable principal component pursuit (SPCP) [21], and we compare their results to a more traditional approach, namely principal component analysis (PCA) [22, 23]. The PCA [22, 23], PCP [19, 24], and SPCP [21, 25] implementations used in this chapter, are provided as an open-source Python 3 toolbox[1].

Model structure. The underlying models of all three methods can be considered members of the family $\mathscr{F}$ of extended linear mixture models [26]:

$$\mathscr{F}: \; A = B + C + D, \text{ with } B = YZ^T, \tag{2.1}$$

where $A \in \mathbb{R}^{m\times n}$ represents a measurement matrix to decompose, and $B \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{m\times n}$, and $D \in \mathbb{R}^{m\times n}$ are respectively defined as a low-rank term (B), a sparse residuals term (C), and a dense residuals term (D). Furthermore, B is a product of factors $Y \in \mathbb{R}^{m\times r}$ and $Z \in \mathbb{R}^{n\times r}$, where r is the matrix rank of B and it is expected that $r \ll \min(m,n)$. While there exist methods that optimize over Y and Z explicitly [27], we will obtain these factors implicitly by use of a nuclear norm surrogate.

Optimizations. The methods can be formulated as optimization problems:

- Principal Component Analysis (PCA),
  $f_{\text{PCA}} : \mathbb{R}^{m\times n} \times \mathbb{Z}_0^+ \to \mathbb{R}^{m\times n}; (A,r) \mapsto (B)$:

[1] https://github.com/vandeplaslab/hannibalspecter

$$\begin{aligned} \underset{B}{\text{minimize}} \quad & \|A-B\| \\ \text{subject to} \quad & \text{rank}(B) \leq r, \end{aligned} \tag{2.2}$$

- Principal Component Pursuit (PCP),
  $f_{\text{PCP}} : \mathbb{R}^{m\times n} \times \mathbb{R}^{+} \to \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n}; (A,\lambda) \mapsto (B,C)$:

$$\begin{aligned} \underset{B,C}{\text{minimize}} \quad & \|B\|_* + \lambda \|C\|_1 \\ \text{subject to} \quad & A = B + C, \end{aligned} \tag{2.3}$$

- Stable Principal Component Pursuit (SPCP),
  $f_{\text{SPCP}} : \mathbb{R}^{m\times n} \times \mathbb{R}^{+} \times \mathbb{R}^{+} \to \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n}; (A,\theta,\delta) \mapsto (B,C)$:

$$\begin{aligned} \underset{B,C}{\text{minimize}} \quad & \|B\|_* + \theta \|C\|_1 \\ \text{subject to} \quad & \|A-B-C\|_F \leq \delta. \end{aligned} \tag{2.4}$$

Our notation uses $\|\cdot\|_1$ as the entry-wise ℓ_1-norm, $\|\cdot\|$ as the operator norm, $\|\cdot\|_*$ as the nuclear norm, and $\|\cdot\|_F$ as the Frobenius norm. The PCA-method, f_{PCA}, has one parameter $r \in \mathbb{Z}_0^+$ and delivers a low-rank matrix $B = YZ^T$ explicitly and a dense residuals term D implicitly (as $D = A - B$). The PCP-method, f_{PCP}, also has one parameter $\lambda \in \mathbb{R}^+$ and delivers a low-rank matrix $B = YZ^T$ and a sparse residuals matrix C explicitly. With PCP, there is no (or only a trivially small) dense residuals term D since A is set to be equal to $B + C$. Also note that for PCP, the cardinality (number of non-zero entries, *i.e.*, sparsity) of C can be relaxed under certain conditions, such that it can contain dense noise [28]. The SPCP method has two parameters, namely $\theta \in \mathbb{R}^+$ and $\delta \in \mathbb{R}^+$. SPCP delivers a low-rank matrix $B = YZ^T$ and a sparse residuals matrix C explicitly and a dense residuals term D implicitly (as $D = A - B - C$). Further specifics on method conditions are provided in the Supplementary Information. While randomized PCA has a computational complexity of $\mathcal{O}(mnr)$, PCP and SPCP involve iterative procedures and k repeated SVD computations with estimated complexities of $\mathcal{O}(kmnr)$, making them more computationally demanding but still feasible with GPU acceleration and parallel processing.

Parameter Setting. All examined methods have parameters and a study under inexact recovery conditions [29] has suggested that, rather than a unique optimal parameter setting, the optima are different for different noise cases. We therefore explore method results across ranges of potential parameter values. These ranges were chosen on the basis of prior experiments [29], are intended to cover a wide array of scenarios, and deliver insight into the impact of parameters on a method's final result and performance. For PCA, we vary the rank r with unit steps from 1 to $\min(m,n)$. For PCP, a λ-multiplier parameter is varied across 2000 distinct values that are linearly spaced between 0.05 and 3. These values ensure that almost all rank values are represented in the retrieved PCP results, enabling a comparison to PCA and SPCP at a specific rank. The relation to PCP's λ parameter in Equation 2.3 is $\lambda = \frac{1}{\max(m,n)} \times \lambda$-multiplier. For SPCP, a θ-multiplier parameter is linearly varied across 100 values between 0.4 and 2.0. The relation to SPCP's θ parameter in Equation 2.4 is $\theta = \frac{1}{\max(m,n)} \times \theta$-multiplier. Additionally, for SPCP, we vary a σ-multiplier parameter across 100 values, logarithmically spaced between 10^{-4} and 10^{-1}. The relation to SPCP's

δ parameter in Equation 2.4 is $\delta = \sqrt{\min(m,n) + \sqrt{8\min(m,n)}}\,\|A\|_F \times \sigma$-multiplier [25]. SPCP's two parameter ranges result in a grid of $100 \times 100 = 10^4$ parameter pairs, explored exhaustively in both IMS case studies.

2.2.2. COMPARISON METRICS

When the three methods, f_{PCA}, f_{PCP}, and f_{SPCP}, are applied to the same IMS dataset, the differences between their models (Equations 2.2, 2.3, and 2.4) will ensure that each method yields a different decomposition of the same measurements A. This means that we will obtain distinct low-rank matrices (B_{PCA}, B_{PCP}, and B_{SPCP}), distinct sparse residuals terms (C_{PCP} and C_{SPCP}, there is no C_{PCA}), and distinct dense residuals terms (D_{PCA}, D_{PCP}, and D_{SPCP}). Since this chapter focuses on dimensionality reduction, we are particularly interested in the low-rank approximation matrix B and what is captured by its component vectors in the Y and Z matrices or, equivalently, by the U and V components of its singular value decomposition (SVD), *i.e.*, $B = USV^T$. We want to assess what the impact is on the low-rank approximation B of models that have an explicit sparse residuals term (PCP and SPCP) compared to methods that have no such term (PCA).

To compare different low-rank approximations of the same dataset, we formulate two sets of metrics to compare low-rank approximations. The first set of metrics is focused on the content of B, trying to capture how much overlap there is between the subspaces captured by PCA, PCP, and SPCP. The second set of metrics is focused on how well the low-rank approximation lines up with physical reality. More precisely, since ion counts are necessarily non-negative data, we can use the negativity of values in B as a heuristic for how far a low-rank approximation deviates from physical reality (which can have an impact on human interpretation for example). A detailed description of the content-based subspace overlap metrics and the non-negativity metrics can be found in the Supplementary Information.

2.3. RESULTS AND DISCUSSION

We conduct a comparison of PCA, PCP, and SPCP using two MALDI Q-TOF IMS case studies. The first case study is on human eye cornea tissue, and is focused on evaluating the dimensionality reduction performance of the different methods, exploring primarily the low-rank approximation term B of their decompositions. The second case study utilizes IMS measurements from human eye retina tissue, and is focused on evaluating the methods' noise reduction capabilities with illustrations on specific ion species' distributions. As such, Case Study 2 primarily examines the sparse residuals term C and the dense residuals term D of the methods' decompositions.

2.3.1. CASE STUDY 1: DIMENSIONALITY REDUCTION IN CORNEA IMS DATA

Reducing the dimensionality of IMS measurements using PCA, PCP, or SPCP, is primarily a matter of decomposing the measurements into a sum of B, C, and D-terms, and subsequently replacing the original matrix A by its low-rank approximation B for any downstream analysis. In this case study, we compare the low-rank terms provided by the different methods, *i.e.*, B_{PCA}, B_{PCP}, and B_{SPCP}. First, we investigate how much difference there is between the low-rank subspaces that the different methods deliver. We do this by

calculating the overlap between retrieved column and row subspaces of the different low-rank terms. Second, we assess the different methods' tendencies to utilize (non-physical) negative ion counts to obtain their low-rank approximations. To this end, we compare the B_{PCA}, B_{PCP} and B_{SPCP} terms of similar rank in terms of their percentage, sum, and mean of negative entries. The results of this latter analysis can be found in the Supplementary Information Results section.

COMMONALITY IN PCA, PCP, AND SPCP-DELIVERED SUBSPACES

We depict the results of the particular subspace overlap calculations in the same σ and θ-multiplier grid used to show the explored parameter space for SPCP (see Supplementary Information). Figure 2.3 shows the commonalities between the PCA, PCP, and SPCP-delivered dimensionality reduced subspaces. Panels (a) and (b) compare the PCP and SPCP approximations of the cornea dataset, respectively, reporting the overlap in their captured spatial and spectral patterns. Panels (c) and (d) compare the PCA and SPCP approximations of the cornea dataset, respectively, the overlap in their captured spatial and spectral patterns. The average subspace overlap can be defined as the mean of the singular values of the inner product of the compared subspaces (*i.e.*, $U^T_{\text{Method 1}} U_{\text{Method 2}}$ and $V^T_{\text{Method 1}} V_{\text{Method 2}}$). Values close to 0 on average signify that different information is captured by the particular orthonormal basis of the low-rank approximations, *i.e.*, both spaces tend to be orthogonal with respect to each other. Values close to 1 on average signify that the compared subspaces have overlapping orthonormal bases, *i.e.*, information is captured in a similar basis.

For SPCP and PCP (Figure 2.3a-b), gray regions around $\theta \approx 0.9$ and σ-multiplier $\approx 10^{-3}$ report almost full overlapping row and column subspaces between SPCP's and PCP's solutions, indicating that in that parameter range SPCP and PCP capture very similar low-rank approximations of the cornea data and describe the captured information in a similar fashion. However, we will note further on (see non-negativity metrics) that the SPCP components contain less negative entries than PCP components. Hence, it is not fully understood why this strong SPCP-PCP overlap appears in this part of the parameter space. Furthermore, in regions of relatively low rank, *i.e.*, below 20 % (see Suppl. Figure S2.3), there seems to be a valley of low overlap (dark green) between the SPCP and PCP approximations. This indicates that a rather large part of the obtained subspaces are orthogonal in this part of the parameter space, suggesting that SPCP and PCP are capturing different aspects of the data there, even though the methods are based on rather similar theory. We note that even though in some regions the overlap is small, the low-rank solutions themselves can still be "close", since we weigh singular values equally, while in the original low-rank approximations some bases are associated to very large singular values.

For SPCP and PCA (Figure 2.3c-d), four distinct regions are observed: (i) $\theta \approx 0.7$, (ii) below $\theta \approx 0.7$, (iii) above $\theta \approx 0.7$, and (iv) a region to the right of $\sigma \approx 10^{-2}$. Region (i) shows a very low match in column as well as row subspace overlap between SPCP and PCA (dark green), suggesting that, in this region of increasing rank, SPCP captures a different part of the spatial and spectral space of the IMS data than PCA does. Region (ii) exhibits a poor match between the PCA and SPCP-delivered column and row subspaces, indicating that there are quite large differences between the dimensionality reduced subspaces

delivered by these methods at low rank. Region (iii) shows a good match in row subspace, but a poor match in column subspace, indicating that while similar spectral trends are discerned by PCA and SPCP, their spatial component images are quite different. This could indicate that spatially sparse (and high intensive) features in the tissue, which we know to be present in this dataset, are captured by SPCP in its residuals terms, while PCA has no choice but to capture such variation in its low-rank term, causing the component images to deviate substantially. Finally, the gray coloring in region (iv) suggests an almost complete match between SPCP and PCA, which is probably due to the very low-rank in that area not allowing a lot of differentiation between the methods.

Overall, we are able to compare the content of the dimensionality reduced approximations delivered by PCA, PCP, and SPCP. We can distinguish different subareas within the parameter space where the approximating subspaces from the different methods are overlapping, and others where the subspaces are very different from each other. Between SPCP and PCP, we can observe a large overlap for most parameter settings, suggesting information is captured in a similar basis and thus similar underlying trends are found by these methods. Between SPCP and PCA, we observe that, while the row subspaces provided by these methods are quite similar, the overall column subspace overlap seems to be lower. This indicates that, while the underlying spectral patterns captured are similar, the corresponding spatial images indicating where these spectral patterns are active in the tissue are quite different between PCA and SPCP. One possible explanation could be that SPCP, in providing a sparse residuals term C_{SPCP}, offers a place for spatially sparse features in the data to go rather than into its low-rank approximation term B_{SPCP}. Since PCA has no such sparse residuals term, spatially sparse features in the tissue can only be captured in the low-rank approximation B_{PCA}, driving up the necessary rank to approximate a measurement set and modifying the spatial signatures of the principal components if the rank is set. This suggests that spatially sparse patterns in IMS data could substantially skew (at least the component images of) the low-rank approximation delivered by PCA, an issue SPCP does not suffer from. It should also be noted that if a spatially sparse tissue pattern ends up in the sparse residuals term in SPCP, that does not mean it is considered noise and is thrown away. It just means that SPCP's low-rank approximation has the option to disregard spatially sparse patterns (which could be considered spatial outliers from a global pattern perspective) and thus is able to model the global data patterns more tightly and robustly than PCA.

2.3.2. CASE STUDY 2: NOISE REDUCTION IN RETINA IMS DATA

The second case study moves focus from the methods' dimensionality reduction to their noise reduction capabilities. This means that in the context of the family $\mathscr{F}$ of extended linear mixture models (Equation 2.1) our attention shifts from the low-rank approximation term B (in Case Study 1) to the sparse residuals term C and dense residuals term D for Case Study 2. We explore the residuals terms delivered by PCA, PCP, and SPCP for five particular parameter settings (see Case Study 2 description in Supplementary Information).

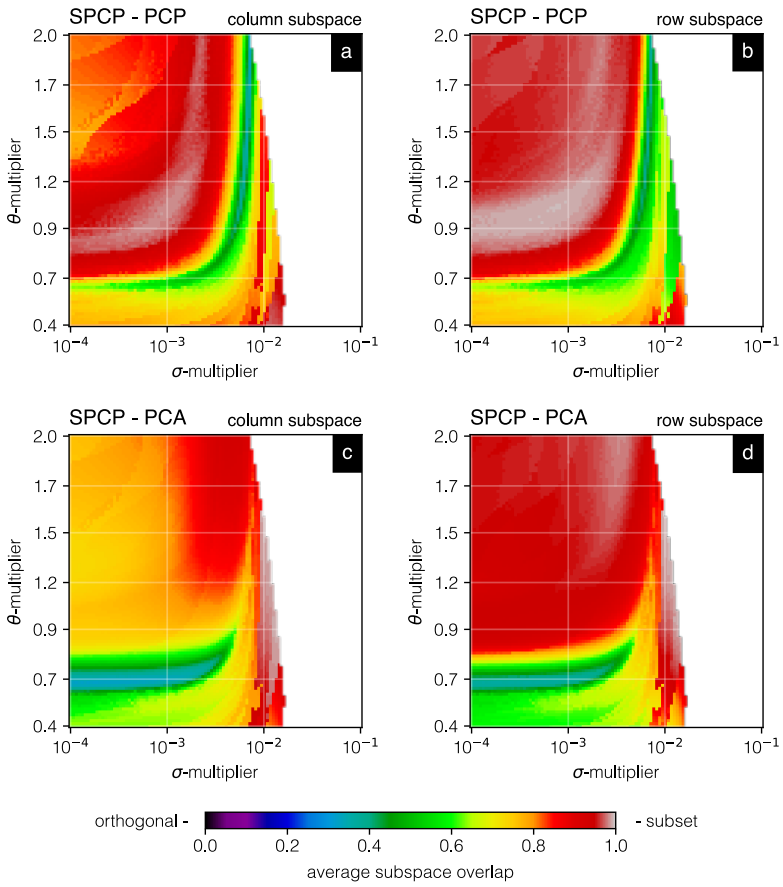

Figure 2.3: Commonality in PCA, PCP, and SPCP-delivered subspaces – Average overlap in column and row subspaces between low-rank approximation terms. (a) Overlap between the column subspaces of SPCP (B_{SPCP}) and PCP (B_{PCP}), *i.e.*, overlap in spatial patterns captured. (b) Overlap between the row subspaces of SPCP (B_{SPCP}) and PCP (B_{PCP}), *i.e.*, overlap in spectral patterns captured. (c) Overlap between the column subspaces of SPCP (B_{SPCP}) and PCA (B_{PCA}), *i.e.*, overlap in spatial patterns captured. (d) Overlap between the row subspaces of SPCP (B_{SPCP}) and PCA (B_{PCA}), *i.e.*, overlap in spectral patterns captured. The average subspace overlap can be defined as the mean of the singular values of the inner product of the compared subspaces (*i.e.*, $U^T_{\text{Method 1}} U_{\text{Method 2}}$ and $V^T_{\text{Method 1}} V_{\text{Method 2}}$). Values close to 0 on average signify that different information is captured by the particular orthonormal basis of the low-rank approximations, *i.e.*, both spaces tend to be orthogonal w.r.t. each other. Values close to 1 on average signify that the compared subspaces have overlapping orthonormal bases, *i.e.*, they are subsets of each other and similar information is captured on a similar basis.

ION INTENSITY DISTRIBUTION AMONG TERMS

The global entry-wise intensity histograms (representing values across all *m/z* bins) for the B, C, and D terms delivered by PCA, PCP, and SPCP are shown in Suppl. Figure S2.9. These allow us to observe the impact of the θ and σ-multipliers on the spread of the measurements' energy in A among the different decomposition terms. A detailed set of observations for each of the five cases is provided in the Supplementary Information.

Across all cases, the residuals term D of PCA (PCA has no C term) shows quite dense

intensity histograms (*i.e.*, lots of nonzero values spread across a broad range of intensity levels; in gray). The residuals term C of PCP (PCP has no D term) also shows relative wide and dense intensity histograms (*i.e.*, lots of nonzero values populating a broad range of intensity levels; in red). However, for SPCP, which has both a C and D term, we see the communication between the sparse and dense residuals terms in action. In all five cases, whenever the dense residuals term D_{SPCP} is allowed to siphon off dense variation that is not very structured (so not easily captured by B) and not very sparse (so not easily captured by C), it allows SPCP's sparse residuals term C_{SPCP} to capture a more sparse intensity distribution (see blue traces in C column of Suppl. Figure S2.9) than PCP's sparse residuals term C_{PCP} (see red traces in C column of Suppl. Figure S2.9). This is readily visible in Case 3 where the blue C_{SPCP} histogram contains less energy than the red C_{PCP} histogram and reports a smaller number of high intensity peaks, indicating sparse intensity distributions. Presumably, the "escape valve" offered by SPCP's dense residuals term allows also the low-rank approximation B to capture a tighter low-rank model of the data, but this is a hypothesis and is hard to assess from the intensity histograms in this figure.

This assessment of how ion intensity variation is distributed among decomposition terms shows that unlike PCA and PCP, at least for retina data, SPCP has the ability to (a) obtain a low-rank representation of an IMS dataset, which is useful for dimensionality reduction, compression, and interpretation, (b) capture spatially and spectrally sparse features in the data, which are not necessarily noise, but would require a lot of extra dimensions if they were forced to be represented by the low-rank approximation, and (c) separate out a layer of low-intensity dense variation, which can usually be considered noise and be thrown out. SPCP performs this decomposition of measurements in one optimization run and delivers the different signal components concurrently to the user, going beyond what PCA and PCP can deliver for the same dataset.

ION SPECIES-SPECIFIC EFFECTS

After examining what PCA, PCP, and SPCP do with the content of an IMS measurement set on a dataset-wide scale, we now investigate how SPCP's parameter setting influences individual ion species, mass bins, or ion images. Supplementary Figure S2.10 shows the distribution of energy among the C and D terms for six distinct m/z bins: m/z 601.53 (ion image with sparse spatial structures); m/z 1 007.01 (low intensity ion image); m/z 790.52 (low to average intensity ion image); m/z 666.43 (average intensity ion image); m/z 554.57 (high-intensity ion image without strong outliers); and m/z 591.01 (high-intensity ion image with strong outliers). Each panel shows the (histogram-ed) content of the residuals terms of PCA, PCP, and SPCP for a specific parameter setting/decomposition (Cases 1 through 5 specified in the Supplementary Information) and for a specific m/z bin. The measured variation that ends up in the residuals terms C and D is the difference between the raw measurement A and its low-rank low-dimensional representation B. If one considers the low-dimensional representation as the 'important' part of the dataset, the content of C and D could potentially be labeled as the 'non-important' part of the measurement and thus, could be removed to effectively 'denoise' the measurement. The dense residuals term D is not optimized towards a particular type of content, simply capturing whatever is not already represented by the other terms in the decomposition, so in almost all use cases the removal of D is a useful denoising technique. The sparse

residuals term C is optimized towards capturing sparse features in the measurements, which are not necessarily noise, so whether or not to remove C to denoise is an application-specific consideration conditioned on whether we think that sparse signal should be retained for downstream analysis or not. A more detailed treatment of the histogram traces in Suppl. Figure S2.10 is provided in the Supplementary Information.

We observe that under certain parameter settings the residuals of PCA, PCP, and SPCP for a specific m/z bin can be very similar, *e.g.*, for m/z 1 007.01 and m/z 790.52 for Cases 2 and 3, suggesting a similar denoising capability for those ion species. However, this behavior seems primarily correlated to low or low-to-average intensity ion species. When examining high-intensity ion species (m/z 666.43 and 554.57) or ion distributions that exhibit sparse signals (m/z 591.01), much larger differences between the methods can be discerned. We also see that in most m/z bins the residuals of PCP and SPCP contain inherently fewer negative entries, which correspond to our findings in Case Study 1. Finally, Suppl. Figure S2.10 shows that PCA's residual intensity distribution is mostly symmetric around the origin, where the PCP and SPCP residuals exhibit mostly asymmetric intensity distributions. Different parameter settings lead to different residuals distributions local to specific ion species, and some of these effects are tied to the nature of the ion species' intensity level (low versus high-intensity ions) and spatial distribution (*e.g.*, sparse features or not). This suggests that the parameters for PCA, PCP, and SPCP could and should be optimized differently in function of whether one wants to amplify or attenuate such ion species-specific effects. Overall, residuals of PCP and SPCP tend to contain fewer negative intensity values, suggesting less overestimation of the mass spectral signals in these methods' low-rank approximations.

EFFECTS ON IMAGES

While the histogram view on the distribution of measured ion counts among the different components gives a broad view into what is happening when a PCA, PCP, or SPCP-based dimensionality reduction is applied to IMS data, the final assessment comes down to what these methods mean in terms of the (denoised) images they provide. Specifically, we explore how each of these methods decomposes the same measured ion image into a low-rank approximated image, a sparse variation image, and a dense residuals image. Supplementary Figure S2.11 shows for the six ion species examined in the previous section what ends up in their sparse (*i.e.*, C) and dense (*i.e.*, D) residuals images, for five different parameter settings, and it allows us to make some general observations. The original ion images can be found in Suppl. Figure S2.8.

For PCA, it is expected that the D term captures small entry-wise noise (see Datasets and Methods). However, we observe large negative entries as well, *e.g.*, dark blue pixels in cases 2 and 3. The fact that PCA's D term is trying to negatively compensate against its B term to arrive at the measured ion intensity could potentially mean that the estimated signal in PCA's low-rank approximation B term is overestimated (*i.e.*, false signal creation in the low-rank term). This is generally undesirable since it could suggest tissue structure in the (denoised) B image that is not really present. We also observe genuinely biological sparse tissue features in PCA's sole residuals term, *e.g.*, small dark red colored areas for m/z 601.53 and 591.01. This suggests that if PCA's D term is simply labeled noise and thrown away without inspection, genuine biological information could be lost in the process.

For PCP, it is expected that its C term captures sparse residuals, which could be noise

or sparse biological signal. However, since the PCP model (Equation 2.3) has only two terms to represent measured ion intensity with, its C term also needs to capture whatever is not already represented by PCP's low-rank approximation term B, which means it has to occasionally capture dense noise patterns as well. In Suppl. Figure S2.11, we observe fewer negative entries (blue) in PCP's C than in PCA's D, suggesting less opportunity for overestimation in the low-rank approximation of PCP compared to that provided by PCA at the same rank. We observe more high-intensity positive (red) sparse features in PCP's C than in PCA's D, yet we do not see many truly sparse images (mostly or almost fully yellow, with the occasional sparse red feature) in PCP's C. This might be problematic when sparse features are under investigation, and PCP's model is simply assumed to deliver a direct notion of those features in its C, since this is shown to not necessarily pan out.

The problem of PCP's C term not necessarily capturing sparse variation is largely solved by SPCP's model (Equation 2.4). SPCP can dedicate its C term to capturing sparse noise and sparse biological signal by providing an extra D term to capture small entry-wise noise. This is reflected in Suppl. Figure S2.11's bottom two panels. We see SPCP's C images capture primarily sparse red features, *e.g.*, in cases 2, 3 and 4, among an image that consists largely of yellow pixels close to zero. We also observe that SPCP's D term captures dense noise, *e.g.*, in cases 1, 4 and 5, while for other cases it will contain negative entries, *e.g.*, for m/z 666.43, 554, 57 and 591.01. The discrepancy between PCP's C and SPCP's C seems to suggest that SPCP is better suited for capturing genuine sparse variation in the data. We therefore conclude that SPCP, in contrast to PCA and PCP, is able to factor out sparse signal more straightforwardly, but that parameter tuning is crucial for optimally exploiting that sparsity and for minimizing possible signal overestimation.

While we have discussed the content of the different decomposition terms in isolation, it is valuable to bring these different aspects of a measured ion image together for a specific ion species to see how the different dimensionality reduction methods, PCA, PCP, and SPCP, decompose the same ion image. Figure 2.4 shows, for ion species m/z 601.53, the measured ion image (A), the low-rank approximation image (B), the sparse residuals image (C), and the dense residuals image (D). This ion species is an example where the ion distribution contains a genuinely biological sparse tissue feature, and we see where this pattern ends up for filtering or retention. Figure 2.5 shows the same results for ion species m/z 1 007.01, whose ion distribution does not contain sparse tissue features. The strong similarity in low-rank approximations between PCA, PCP, and SPCP when no sparse variation is present in the data (*e.g.*, m/z 1 007.01), and PCP and SPCP's clear separation of sparse variation when sparse patterns are present in the data (*e.g.*, m/z 601.53) provide some guidance for why PCP and SPCP-based dimensionality reduction should be considered for IMS data, a data type where sparse variation is often prominently present.

In general, the low-rank approximation image B tends to capture the global tissue structure underlying the sometimes noisy ion image A. Using image B for downstream analysis rather than image A would effectively amount to denoising the ion image and potentially revealing tissue structure that might go unrecognized if it remained buried among noise. The dense residuals image D tends to capture unstructured variation across the whole dataset, is generally labeled as noise, and thus tends to be removed from further analysis. The sparse residuals image C tends to capture the sparse variation in the

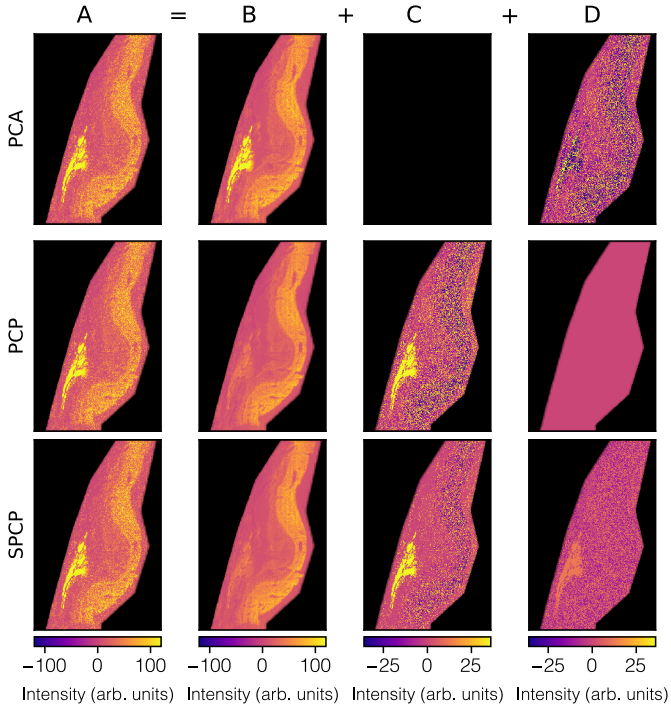

Figure 2.4: PCA, PCP, and SPCP decompositions of *m/z* 601.53's measured retina ion image (*A*) into the sum of a low-rank approximation image (*B*), a sparse residuals image (*C*), and a dense residuals image (*D*). We observe an example of a sparse tissue structure (in white) in ion image *A*. In this figure, *A*, *B*, *C*, and *D* show one column from, respectively, matrices A through D, refolded into an image. Note, however, that the methods have been applied on the whole set of ion images in one go (see Supplementary Information, Data Preprocessing). In PCA, this feature needs to be captured in the low-rank approximation image *B*, either costing additional dimensions, or skewing the axes of the lower-dimensional space to accommodate this high-intensity feature. In PCP and SPCP, the sparse tissue feature ends up primarily in the sparse residuals image *C*, where a call can be made whether one wants to remove or retain that sparse content. If the latter, SPCP's *C* image tends to be cleaner since it can funnel off dense noise to its dense residuals image *D*, allowing for dense noise removal even when sparse patterns are being retained. While at first it might seem that PCP and SPCP's low-rank approximation image does not leave much trace of the sparse tissue feature, it is actually still relatively well captured by those images if one takes the different color maps into account. Even if *C* is removed and only *B* is retained, the tissue feature would still be present in the lower-dimensional representation of the data, albeit at less high ion intensities.

measured ion image. Such sparse variation could be noise, but if spatially or spectrally sparse signals are present in the IMS data (*e.g.*, due to fine low-pixel-area tissue structures), that sparse variation could be genuinely biological in nature. The advantage of capturing such sparse variation as a separate image is that the user has the option to inspect that image and assess whether the sparse variation seems to delineate biological patterns. If deemed biological in nature, the user has the option to retain the sparse image and add it back to the low-rank image for downstream analysis. If deemed noise or not informative, the sparse image can be removed from further analysis. Regardless of whether the sparse image is retained, the ability to separate it in the PCP and SPCP cases and avoid sparse

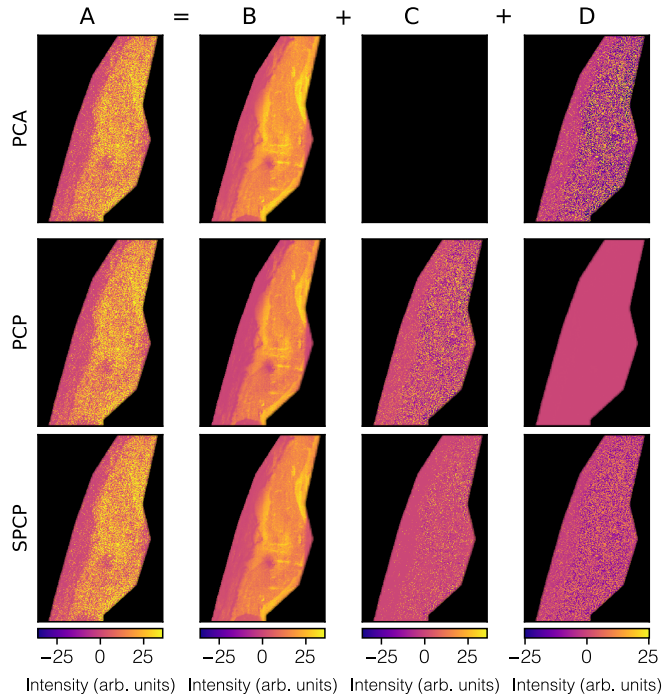

Figure 2.5: PCA, PCP, and SPCP decompositions of m/z 1 007.01's measured retina ion image (A) into the sum of a low-rank approximation image (B), a sparse residuals image (C), and a dense residuals image (D). In this figure, A, B, C, and D show one column from, respectively, matrices A through D, refolded into an image. Note, however, that the methods have been applied on the whole set of ion images in one go (see Supplementary Information, Data Preprocessing). This ion species is an example of an ion distribution without many sparse tissue features, and demonstrates the strong similarity between PCA, PCP, and SPCP results when sparse variation is absent.

variation from needing to be captured by the low-rank approximation allows for a tighter modeling of the underlying trends, a reduction in the number of dimensions that need to be retained (so better compression), and generally more physically interpretable axes for the lower-dimension subspace.

2.4. CONCLUSION

We introduced and compared PCP and SPCP as alternatives to the more traditional PCA dimensionality reduction (and noise reduction) method for IMS measurements. Using a cornea dataset, we investigated the commonality among PCA, PCP, and SPCP-delivered subspaces and their low-rank approximations. We also assessed in how far an added sparse residuals term C as in the PCP and SPCP decompositions impacts the non-negativity of these methods' low-rank approximations, and thus the physical feasibility and interpretability of the components and underlying trends retrieved from the data. In a retina case study, we explored the global as well as ion species-specific distribution of measured ion intensity among the different decomposition terms and the influence of

parameter settings on these distributions.

Our initial observations suggest that SPCP allows for a higher compression by separating out sparse residuals into a separate term, thus requiring fewer components to describe the remaining data. SPCP also seems to deliver more physically interpretable components for non-negative data than PCP and PCA. This was demonstrated on the cornea dataset, especially for data approximations of lower rank. The overlap in row and column subspaces of the low-rank approximations between all methods suggests that information is captured on a similar spectral basis. However, we could observe substantial differences in subspace overlap along the spatial domain, which might be an indicator of genuinely sparse biological features being more common on the spatial rather than spectral side for our cornea tissue data. From our second case study on retina data, we can conclude that parameter setting is highly influencing the recovered terms. For example, it shows that setting the σ-multiplier in combination with the θ-multiplier for SPCP can have a large effect on the cardinality of the sparse term. The latter is useful if sparse features are of interest. We can also conclude that SPCP and PCP in contrast to PCA have a tendency to approximate signal 'from below', *i.e.*, by a smaller value, minimizing possible signal overestimation.

This chapter demonstrates the beneficial capabilities of more advanced factorization methods in capturing approximations of (I)MS data. This can be a first step towards enabling full mass-profile factorization and analysis at scale. At the same time, we have shown that these advanced methods require more and better fine-tuning of parameters to find appropriate results. The latter is of increasing difficulty due to growing dataset sizes. In conclusion, our research has shown that exploiting sparsity is useful yet underappreciated in IMS data analysis and that it might be a key consideration for future dimensionality reduction solutions for IMS data and other molecular imaging modalities.

REFERENCES

[1] R. M. Caprioli, T. B. Farmer, and J. Gile. Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. In: *Analytical Chemistry* 69.23 (1997), pp. 4751–4760.

[2] L. A. McDonnell and R. M. Heeren. Imaging Mass Spectrometry. In: *Mass Spectrometry Reviews* 26.4 (2007), pp. 606–643.

[3] W. J. Perry, C. M. Grunenwald, R. Van de Plas, J. C. Witten, D. R. Martin, S. S. Apte, J. E. Cassat, G. B. Pettersson, R. M. Caprioli, E. P. Skaar, et al. Visualizing Staphylococcus Aureus Pathogenic Membrane Modification Within the Host Infection Environment By Multimodal Imaging Mass Spectrometry. In: *Cell Chemical Biology* (2022).

[4] S. Kaspar, M. Peukert, A. Svatos, A. Matros, and H.-P. Mock. MALDI-imaging Mass Spectrometry–an Emerging Technique in Plant Biology. In: *Proteomics* 11.9 (2011), pp. 1840–1850.

[5] M. Stoeckli, P. Chaurand, D. E. Hallahan, and R. M. Caprioli. Imaging Mass Spectrometry: a New Technology for the Analysis of Protein Expression in Mammalian Tissues. In: *Nature Medicine* 7.4 (2001), pp. 493–496.

[6] HuBMAP Consortium. The Human Body at Cellular Resolution: the NIH Human Biomolecular Atlas Program. In: *Nature* 574.7777 (2019), p. 187.

[7] R. G. Cooks, Z. Ouyang, Z. Takats, and J. M. Wiseman. Ambient Mass Spectrometry. In: *Science* 311.5767 (2006). Cited by: 1260, pp. 1566–1570.

[8] Z. Takáts, J. M. Wiseman, and R. G. Cooks. Ambient Mass Spectrometry Using Desorption Electrospray Ionization (DESI): Instrumentation, Mechanisms and Applications in Forensics, Chemistry, and Biology. In: *Journal of Mass Spectrometry* 40.10 (2005). Cited by: 742; All Open Access, Bronze Open Access, pp. 1261–1275.

[9] J. Koch and D. Günther. Review of the State-of-the-art of Laser Ablation Inductively Coupled Plasma Mass Spectrometry. In: *Applied Spectroscopy* 65.5 (2011). Cited by: 229; All Open Access, Bronze Open Access, 155A–162A.

[10] A. M. Belu, D. J. Graham, and D. G. Castner. Time-of-flight Secondary Ion Mass Spectrometry: Techniques and Applications for the Characterization of Biomaterial Surfaces. In: *Biomaterials* 24.21 (2003), pp. 3635–3653.

[11] B. Mamyrin. Time-of-flight Mass Spectrometry (Concepts, Achievements, and Prospects). In: *International Journal of Mass Spectrometry* 206.3 (2001), pp. 251–266.

[12] S. Eliuk and A. Makarov. Evolution of Orbitrap Mass Spectrometry Instrumentation. In: *Annu. Rev. Analytical Chemistry* 8 (2015), pp. 61–80.

[13] A. G. Marshall, C. L. Hendrickson, and G. S. Jackson. Fourier Transform Ion Cyclotron Resonance Mass Spectrometry: a Primer. In: *Mass Spectrometry Reviews* 17.1 (1998), pp. 1–35.

[14] J. Fan, F. Han, and H. Liu. Challenges of Big Data Analysis. In: *National Science Review* 1.2 (2014), pp. 293–314.

[15] N. Verbeeck, R. M. Caprioli, and R. Van de Plas. Unsupervised Machine Learning for Exploratory Data Analysis in Imaging Mass Spectrometry. In: *Mass Spectrometry Reviews* 39.3 (2020), pp. 245–291.

[16] R. Van de Plas, F. Ojeda, M. Dewil, L. Van Den Bosch, B. De Moor, and E. Waelkens. "Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis". In: *Biocomputing 2007*. World Scientific, 2007, pp. 458–469.

[17] P. W. Siy, R. A. Moffitt, R. M. Parry, Y. Chen, Y. Liu, M. C. Sullards, A. H. Merrill, and M. D. Wang. Matrix Factorization Techniques for Analysis of Imaging Mass Spectrometry Data. In: *2008 8th IEEE International Conference on BioInformatics and BioEngineering*. IEEE. 2008, pp. 1–6.

[18] M. Nijs, T. Smets, E. Waelkens, and B. De Moor. A Mathematical Comparison of Non-negative Matrix Factorization Related Methods with Practical Implications for the Analysis of Mass Spectrometry Imaging Data. In: *Rapid Communications in Mass Spectrometry* 35.21 (2021), e9181.

[19] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? In: *Journal of the ACM* 58.3 (2011), pp. 1–37.

[20] D. M. Anderson, J. D. Messinger, N. H. Patterson, E. S. Rivera, A. Kotnala, J. M. Spraggins, R. M. Caprioli, C. A. Curcio, and K. L. Schey. Lipid Landscape of the Human Retina and Supporting Tissues Revealed By High-resolution Imaging Mass Spectrometry. In: *Journal of the American Society for Mass Spectrometry* 31.12 (2020), pp. 2426–2436.

[21] Z. Zhou, X. Li, J. Wright, E. J. Candes, and Y. Ma. Stable Principal Component Pursuit. In: *IEEE international symposium on information theory*. IEEE. 2010, pp. 1518–1522.

[22] H. Hotelling. Analysis of a Complex of Statistical Variables Into Principal Components. In: *Journal of Educational Psychology* 24.6 (1933), p. 417.

[23] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[24] Z. Lin, M. Chen, and Y. Ma. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-rank Matrices. In: *arXiv preprint arXiv:1009.5055* (2010).

[25] N. S. Aybat and G. Iyengar. An Alternating Direction Method with Increasing Penalty for Stable Principal Component Pursuit. In: *Computational Optimization and Applications* 61.3 (2015), pp. 635–668.

[26] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. John Wiley & Sons, 2009.

[27] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex Optimization Meets Low-rank Matrix Factorization: An Overview. In: *IEEE Transactions on Signal Processing* 67.20 (2019), pp. 5239–5269.

[28] A. Ganesh, J. Wright, X. Li, E. J. Candes, and Y. Ma. Dense Error Correction for Low-rank Matrices Via Principal Component Pursuit. In: *2010 IEEE international symposium on information theory*. IEEE. 2010, pp. 1513–1517.

[29] R. Moens. *On the Atoms of Robustness: Robust Matrix Decomposition for Spectral Imaging*. https://repository.tudelft.nl/islandora/object/uuid:78817107-d34a-4d6f-9ea1-d7298436d63c. 2021.

[30] D. M. Anderson, A. Kotnala, L. G. Migas, N. H. Patterson, L. E. Tideman, D. Cao, B. Adhikari, J. D. Messinger, T. Ach, S. Tortorella, et al. Lysolipids Are Prominent in Subretinal Drusenoid Deposits, a High-risk Phenotype in Age-related Macular Degeneration. In: *Frontiers in Ophthalmology* 3 (2023), p. 1258734.

[31] K. Sládková, J. Houška, and J. Havel. Laser Desorption Ionization of Red Phosphorus Clusters and Their Use for Mass Calibration in Time-of-flight Mass Spectrometry. In: *Rapid Communications in Mass Spectrometry* 23.19 (2009), pp. 3114–3118.

[32] Y. Zhang, L. Huang, N. Pillar, Y. Li, Y. Li, L. G. Migas, R. Van de Plas, J. M. Spraggins, and A. Ozcan. Virtual Staining of Label-free Tissue in Imaging Mass Spectrometry. In: *Science Advances* 11.31 (2025), eadv0741.

[33] L. G. Migas. *msalign: Spectral alignment based on MATLAB's 'msalign' function*. https://github.com/lukasz-migas/msalign. Version 0.2.0. 2024.

[34] P. Monchamp, L. Andrade-Cetto, J. Y. Zhang, and R. Henson. Signal Processing Methods for Mass Spectrometry. In: *Systems Bioinformatics: An Engineering Case-Based Approach, Artech House Publishers* (2007).

[35] E. J. Candès and B. Recht. Exact Matrix Completion Via Convex Optimization. In: *Foundations of Computational Mathematics* 9.6 (2009), pp. 717–772.

[36] A. Björck and G. H. Golub. Numerical Methods for Computing Angles Between Linear Subspaces. In: *Mathematics of Computation* 27.123 (1973), pp. 579–594.

[37] D. D. Lee and H. S. Seung. Learning the Parts of Objects By Non-negative Matrix Factorization. In: *Nature* 401.6755 (1999), pp. 788–791.

SUPPLEMENTARY MATERIALS

SUPPLEMENTARY INFORMATION TO THE DATASETS

This section describes the sample preparation protocols for the retina and cornea tissues, acquisition details of the instrumentation used for MALDI Q-ToF IMS measurement, dataset characteristics, and preprocessing steps applied before analysis.

SAMPLE PREPARATION

The retina tissue sections were prepared as described by Anderson et al. [20]. Whole eyes were obtained from deceased human donors by Advancing Sight Network (Birmingham, AL) as part of ongoing studies on age-related macular degeneration (AMD) that are approved by institutional review at University of Alabama at Birmingham (protocol # N170213002), where tissues were collected. Whole globes were opened anteriorly and immersed in 4% phosphate-buffer paraformaldehyde (PFA) overnight. Globes were then placed in 1% PFA at 4°C for up to 48 h prior to sectioning. Dissected tissue containing central retina with the fovea was embedded in 2.25% carboxymethylcellulose (CMC) before sectioning a 93 year old retina donor tissue at 12-14 µm thickness using a Leica CM3050S (Leica, IL, USA) at -20 °C and mounting onto indium tin oxide (ITO) coated microscope slides (Delta Technologies ETC). Samples were vacuum sealed with oxygen absorbing packets and transported to Vanderbilt University on dry ice and stored in a -80 °C freezer. Before analysis, slides were brought to room temperature and dried in a vacuum desiccator for a minimum of 30 minutes. Cornea tissues were prepared from a whole eye globe from a 59 year old donor, the globe was fresh frozen and embedded in 15% fish Gelatin (Sigma Aldrich, St. Louis, MO, USA). The whole globe was then divided into two halves using a rotating cutting (Dremel 402 mandrel, Dremel, 402, Walnut Ridge, AR, USA) saw to reduce the size of the sample in order to improve section quality and reproducibility. Sections were taken around the nasal side of the mid point of the lens (were the nucleus of the lens and pupil were visible). Detailed method on this preparation can be found at (https://www.protocols.io/view/uab-vu-biomic-preparation-of-left-fresh-frozen-eye-3byl4jd38lo5/v1).

MALDI Q-TOF IMS DATASETS

The following is based on a description by the same co-authors in Anderson et al. [30]. The MALDI matrices, 1,5-diaminonaphthalene (DAN, 15 mg) for negative ion mode and 2,5-dihydroxyacetophenone (DHA, 20 mg) (Tokyo Chemical Industry CO, Tokyo, Japan.) for positive ion mode, were applied to tissue sections using a custom designed sublimation device. MALDI IMS data were acquired with a 10 μm pixel size for the retina with a 10 μm pitch, while data for the cornea were acquired at 20 μm pixel size with a 20 μm pitch in full scan mode using a timsTOF Pro for the retina and a timsTOF Flex for the cornea, MALDI imaging platform in QTOF mode (Bruker Daltonik, Bremen, Germany), laser parameters such as laser power and beam scan for the 20 μm pitch were optimized for each experiment varying pixel size. Data were acquired with 250 laser shots per pixel and within a mass range of m/z 300-2 000 for the retina and 500-2 000 for the cornea. The mass spectrometer was calibrated with red phosphorus prior to data acquisition [31].

Specifics	Cornea	Retina
Section thickness	12 μm	12-14 μm
Slide type	Poly-lysine coated ITO	Poly-lysine coated ITO
MALDI matrix	1,5-diaminonaphalene (DAN)	2,5-dihydroxyacetophenone (DHA)
Ionization mode	Negative	Positive
Pixel size	20 μm	10 μm
Instrument	Bruker MALDI timsTOF Flex	Bruker MALDI timsTOF Flex
Preprocessing	*m/z* aligned, calibrated and normalized by 5-95% TIC	*m/z* aligned, calibrated and normalized by 5-95% TIC
Mass-to-charge range	*m/z* 500-2 000	*m/z* 300-2 000
Data table	235 218 pixels×2 381 peaks	137 923 pixels×3 212 peaks
Variable type	32 bits int	32 bits int
Data size	2.24 GB	1.77 GB
Raw data size	6.28 GB	30.7 GB
Peak picked	Yes	Yes

Table 2.1: Table containing the specifics of our two datasets, including information on the wet-lab specifics and data footprint specifics.

DATA PREPROCESSING

The following is based on a description by the same co-authors in Zhang et al. [32]. Data were exported from the Bruker timsTOF file format (.d) to a custom binary format for easy access and improved performance. Each pixel/frame contains between 10^4 and 10^5 centroid peaks covering the entire acquisition range, which can be reconstructed into a pseudo-profile mass spectrum using Bruker's SDK (v2.21). The dataset was m/z-aligned using six internally identified peaks (appearing in at least 50% of the pixels) through the msalign library (v0.2.0) [33, 34]. This step corrects spectral misalignment (drift along the m/z axis), increasing overlap between spectral features (peaks) across the experiment. Subsequently, the mass axis of the data set was calibrated using the theoretical masses of the six peaks, achieving a precision of approximately ±1 ppm. Normalization correction factors were computed following the preprocessing steps and an outlier-insensitive total ion current variant that only includes data lying between the 5^{th} and 95^{th} percentiles of each mass spectrum (5/95% TIC) was used for mass spectral and ion image normalization. Subsequently, an average mass spectrum based on all pixels was calculated for each dataset. Each mass spectrum was peak-picked independently of each other, producing a feature list of 2 381 peaks for the cornea and 3 212 for the retina. It is important to note that isotopic peaks were not removed prior to proceeding with the analysis. The feature lists were used to extract data into a two-dimensional matrix of shape (N x M) where N is the number of pixels and M is the number of features. An image was created by summing the intensity of a specified peak with a ± 3-5ppm extraction window. Before further analysis, the image intensity matrix is normalized using the 5/95% TIC normalization factors.

MICROGRAPHS

SUPPLEMENTARY INFORMATION TO THE METHODS

This section outlines the theoretical conditions and assumptions underlying the PCP and SPCP methods, including sparsity, incoherence, and noise-bound constraints.

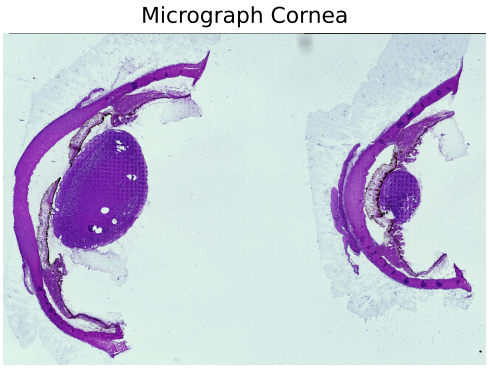

Figure S2.1: Human cornea H&E-stained image where fine sparse layers, *e.g.*, lens, ciliary processes, and iris, can be observed.

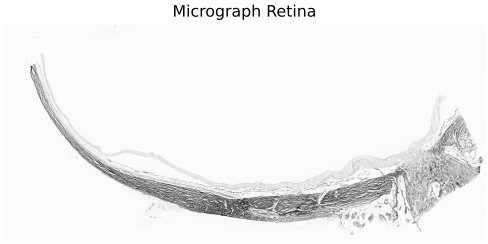

Figure S2.2: Human retina brightfield image revealing major and fine tissue structures including inner retina, choroid, and sclera.

METHOD CONDITIONS

For PCP and SPCP, several conditions for exact or bounded retrieval are imposed, including incoherence conditions on the low-rank term, *i.e.*, $\mu(A) \leq \eta$ for some η [35]; a sparsity constraint for the sparse residuals term, *i.e.*, $\|B\|_0 \leq \gamma$ for some γ, (although this constraint can be relaxed to allow for dense noise [28]); and a size constraint (entry-wise) of the dense residuals term, *i.e.*, $\|C\|_F \leq \delta$ for some δ. Note that conditions like incoherence are difficult to verify. In Chapter 2, we therefore assume that those conditions are satisfied. For a more elaborate sensitivity analysis with respect to these conditions, we refer the reader to prior work [29].

SUPPLEMENTARY INFORMATION TO THE CASE STUDIES

In this section, the rationale, parameter selection, and methodology for conducting both dimensionality reduction and noise reduction case studies are elaborated on in-depth.

CASE STUDY 1: FOCUS ON DIMENSIONALITY REDUCTION

The focus of Case Study 1 is on comparing the methods' dimensionality reduction performance. We will be using the human cornea dataset to illustrate this aspect.

To make a fair comparison of the quality of low-rank data approximations by PCA, PCP, and SPCP, would require that the low-rank data approximations are all given the same

number of components, *i.e.*, the same rank, to work with. Moreso, we would like to assess these methods across a large range of possible ranks, so the challenge becomes how to obtain PCA, PCP, and SPCP low-rank approximations that are matched in rank and this for different ranks. Since PCA has the rank r as an explicit parameter, it is not hard to obtain PCA results for a specific rank at which we want to compare. More precisely, the rank of a PCA result can be easily and uniquely set by truncating the underlying SVD. However, the rank of the B-matrix is not an explicit parameter of PCP and SPCP, and thus it is much less straightforward to retrieve PCP and SPCP results that employ a specific rank. In other words, it is complicated to construct bijective maps for PCP and SPCP, such that for PCP a unique λ-parameter value maps to a specific rank, or for SPCP a unique (θ,σ)-parameter value set yields a specific rank result. Instead, we approach the problem the other way around and perform an extensive parameter sweep for all methods to obtain the different rank results needed to drive a fair comparison.

Therefore, regarding SPCP, we execute a GPU-accelerated implementation of the SPCP algorithm on a randomly selected 10% subset (heuristically determined) of the total number of pixels in the cornea dataset, and repeat this for the 10 000 possible parameter settings of σ and θ-multipliers. The 10 000 recovered low-rank terms B_{SPCP} are saved and their rank and relative rank are depicted in Suppl. Figure S2.3 for each combination of the σ and θ-multipliers. The white entries in this figure correspond to trivial solutions, *e.g.*, $B_{\text{SPCP}} = 0$, which yield zero-rank approximations of the data that are not useful for dimensionality reduction purposes (not for compression or feature extraction, nor for human interpretation). In the non-white (non-trivial) parameter combinations in Suppl. Figure S2.3, we can observe that for an increasing θ-multiplier the rank tends to increase. This is to be expected since a larger θ-multiplier translates into a larger θ-value in SPCP's model in Equation 2.4, a growing need to keep the nuclear norm in the optimization's objective function low, and thus a higher barrier for variation to be captured by the sparse residuals term. In cases where the σ-multiplier and its corresponding dense residuals term D_{SPCP} are kept the same, the variation that can no longer be captured by the sparse residuals term C_{SPCP} will increasingly need to be captured by the low-rank term B_{SPCP}, resulting in an increase in the rank of that matrix. This increase in rank from close to zero up to ~2 400 with an increasing θ-multiplier is visible along the vertical axis of Suppl. Figure S2.3. We also observe that as the σ-multiplier increases, the rank goes down. This is also expected since an increased σ-multiplier value translates into a higher value for parameter δ in Equation 2.4, a higher allowed amplitude for the dense residuals in D_{SPCP}, and thus less need for the low-rank term B_{SPCP} and the sparse residuals term C_{SPCP} to capture the actual variation found in the measurements of A. Regardless of whether the barrier to enter the sparse residuals term C_{SPCP} is kept the same or not by the θ-multiplier, the lack of pressure to capture what is in the measurement matrix A that comes with an increased σ-multiplier tends to result in a decreased rank of the low-rank approximation, as is visible along the horizontal axis of Suppl. Figure S2.3. Furthermore, it can be observed that the relationship between the σ and θ-multipliers is not linear, and that the parameters they control tend to drive SPCP's model in Equation 2.4 to utilize the B_{SPCP}, C_{SPCP}, and D_{SPCP} terms as communicating barrels to capture A's content. Overall, an increasing θ-multiplier and decreasing σ-multiplier correspond to higher-rank approximations, while a decreasing θ-multiplier and increasing σ-multiplier

yield lower-rank approximations. Which parameter combination and corresponding decomposition is best for a practical application depends on the particular needs of that application: the lower the rank, the smaller the dimensionality of the approximation of the measurement set, the higher the compression ratio, but the fewer degrees-of-freedom the model has; the higher the rank, the more signal patterns are captured, but the higher the dimensionality of the approximation of the measurements and the worse the compression ratio. Finally, it is useful to point out the smoothness of the rank-space depicted in Suppl. Figure S2.3, which essentially depicts the content of the low-rank term B_{SPCP}. The smooth transitions suggest that the rank-space is differentiable and thus could be directly exploited for hyper-parameter optimization towards particular applications and constraints in future research.

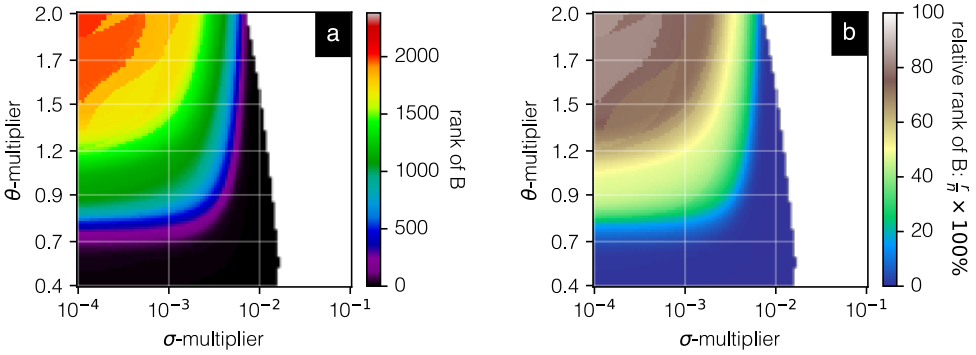

Figure S2.3: Recovered rank (a) and relative recovered rank (b) of the low-rank term (B) for the cornea dataset in function of the SPCP algorithm's θ- and σ-multiplier parameters. The relative recovered rank is taken with respect to the maximal possible matrix rank, *i.e.*, the $\min(m,n)$, where $B \in \mathbb{R}^{m \times n}$. White entries in the heat maps correspond to trivial solutions, *i.e.*, rank = 0. The rank and relative rank increase smoothly when moving from low θ and high σ values to high θ and low σ values. With the compression application in mind, we are interested in (θ,σ)-pairs that yield a relative rank below 20%.

Secondarily, for PCP results, a GPU-accelerated implementation of the PCP algorithm is run on the same 10% subset of pixels, and repeated for 2000 distinct values of the λ-multiplier. The 2000 recovered low-rank terms B_{PCP} are also saved for further processing. Third, PCA is run for all possible rank values and its corresponding recovered low-rank terms B_{PCA} are also stored to disk.

In order to match up PCA, PCP, and SPCP results in terms of rank, the rank of the SPCP results is chosen as a reference point. Next, for each SPCP parameter combination, the closest approximation of the same rank was sought within the PCP and PCA results. For selecting a corresponding PCP result, an additional constraint was set in that its rank must be equal or higher than the SPCP rank. Also, a second constraint was imposed, necessary to achieve a one-to-one map, stating that out of all PCP results with the rank we are looking for, the PCP solution with the highest λ value is selected. The latter constraint is necessary as different λ settings might yield the same rank in their solutions, and this rule ensures that the solution with largest nuclear norm is chosen. In conclusion, for each SPCP parameter value set (and thus B_{SPCP} approximation of A), we obtain a corresponding and unique PCP approximation B_{PCP} and PCA approximation B_{PCA} of that

same A, all with the same rank enabling comparison of their content.

CASE STUDY 2: FOCUS ON NOISE REDUCTION

Contrary to Case Study 1, here we do not consider all possible combinations of the θ and σ-multipliers, but instead restrict ourselves to five distinct parameter settings (Case 1 through 5) that are representative of different solution areas within the parameter space (see Suppl. Figure S2.4). Similar to the first case study, the low-rank approximations obtained at these five locations within the parameter space for SPCP (B_{SPCP}) are matched to low-rank approximations provided by PCP (B_{PCB}) and PCA (B_{PCA}), possessing the same rank. For each of these five parameter cases, we also investigate effects local to a specific

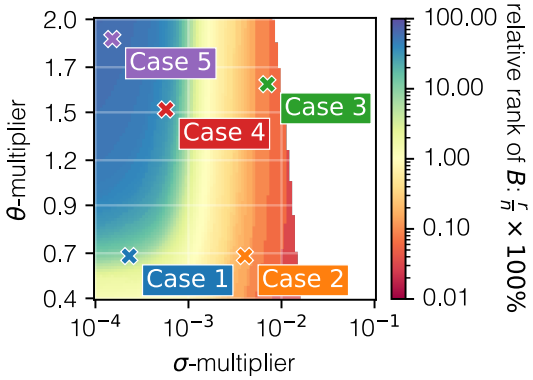

Figure S2.4: Varying parameter settings lead to different results for each method for the retina dataset. In Case Study 2, quantitative and qualitative results are explore for 5 cases, each representing a different part of the solution space. The different cases are: (Case 1) low rank (r=126) through low θ-multiplier and low σ-multiplier; (Case 2) very low rank (r=5) through low θ-multiplier and high σ-multiplier; (Case 3) very low rank (r=3) through high θ-multiplier and high σ-multiplier; (Case 4) middle rank (r=1 018) through middle θ-multiplier and middle σ-multiplier; (Case 5) high rank (r=1 825) through high θ-multiplier and low σ-multiplier.

m/z bin, examining ion distributions that exhibit interesting relevant aspects such as being dominated by low or high ion intensities or the presence of sparse spatial structures. This selection of six m/z-bins consists of (1) m/z 601.53, a representative of ion species whose distributions exhibit sparse spatial structures; (2) m/z 1 007.01, an example of ion distributions with only low intensity values; (3) m/z 790.52, a representative of ion images with low to average intensity values; (4) m/z 666.43, representing ion species with average intensity values; (5) m/z 554.57, an example of ion images with relatively high intensity values without strong outliers (here considered to be a spatially sparse features); and (6) m/z 591.01, representing high intensity ion distributions with strong outliers.

SUPPLEMENTARY INFORMATION TO THE COMPARISON METRICS

Below, we first formulate a content-based set of metrics, capturing the overlap between PCA, PCP, and SPCP captured subspaces. Then, we formulate a set of non-negativity based metrics as heuristics for how strongly the different low-rank approximations differ from the physical reality of ion counts.

CONTENT-BASED METRICS: METRICS REPORTING OVERLAP IN RECOVERED COLUMN AND ROW SUBSPACES

Let us consider two low-rank approximations of the same measurement set A, namely a low-rank matrix $B_{\text{Method 1}}$ obtained by applying a Method 1 and another matrix $B_{\text{Method 2}}$ provided by a Method 2. Even if Methods 1 and 2 were to capture largely the same subspace with their low-rank approximations of the data, there is little reason for the basis vectors captured in $B_{\text{Method 1}}$ and $B_{\text{Method 2}}$ to be identical, and so it is not really an option to compare components directly between methods. Therefore, we need a method to capture the overlap between two recovered subspaces, while remaining insensitive or invariant to the specific basis vector entries. Below, we formulate a principal angles/subspace similarity approach that uses the singular value decomposition to find correspondence between two sets of basis vectors [36].

Furthermore, a r-rank matrix B of size $m \times n$ can be written as its singular value decomposition, *i.e.*, the product of a matrix U of size $m \times r$, a diagonal matrix S of size $r \times r$ and the transpose of a matrix V of size $n \times r$, such that $B = USV^T$. As such, we can examine the subspace overlap along the spatial domain, *i.e.*, U, separately from the subspace overlap along the spectral domain, *i.e.*, V. The spatial subspaces are provided by $U_{\text{Method 1}}$ and $U_{\text{Method 2}}$, and report the "component images" encoded into the columns of $B_{\text{Method 1}}$ and $B_{\text{Method 2}}$ respectively. These images describe where in the tissue a certain component is active. The spectral subspaces are provided by $V_{\text{Method 1}}$ and $V_{\text{Method 2}}$, and report the "component spectra" encoded into the rows of $B_{\text{Method 1}}$ and $B_{\text{Method 2}}$ respectively. These pseudo-spectra describe which m/z bins or features in the IMS data are involved in a certain component.

The method we will use to assess overlap between two sets of basis vectors is to calculate a singular value decomposition of the product between the transpose of one set of orthogonal basis vectors and the other set of orthogonal basis vectors. The result reflects the alignment of the provided bases, and we calculate two metrics using this approach. Metric α reports the overlap between the (spatial) column subspaces of $B_{\text{Method 1}}$ and $B_{\text{Method 2}}$, while metric β reports the overlap between the (spectral) row subspaces of $B_{\text{Method 1}}$ and $B_{\text{Method 2}}$. These metrics are calculated as follows:

$$\begin{aligned} U_c \Sigma_c V_c^T &= \text{svd}(U_{\text{Method 1}}^T U_{\text{Method 2}}), \\ \alpha &= \frac{1}{r}\mathbf{tr}(\Sigma_c), \end{aligned} \tag{2.5}$$

and

$$\begin{aligned} U_r \Sigma_r V_r^T &= \text{svd}(V_{\text{method 1}}^T V_{\text{method 2}}), \\ \beta &= \frac{1}{r}\mathbf{tr}(\Sigma_r). \end{aligned} \tag{2.6}$$

Singular values, here captured in Σ_c and Σ_r, close to 1 indicate bases that are captured in both subspaces, while singular values close to 0 correspond to orthogonal basis vectors that differ between the subspaces. To report the closeness of the mutually recovered column and row subspaces as a single number, we define the mean of singular values as our final metric. The latter ensures that every singular value is equally weighted.

NON-NEGATIVITY METRICS

Ion intensity values measured by mass spectrometry are inherently non-negative. Therefore, any low-rank approximation of an IMS dataset that leans substantially on negative values in its underlying components might be mathematically a valid decomposition into vectors (potentially suited, *e.g.*, for compression), but it is unlikely to be a good representation of the underlying biological trends present in the data (*i.e.*, not suited for human interpretation). One could set non-negativity as a constraint for a low-rank decomposition of IMS data, as in, *e.g.*, non-negative matrix factorization, but in this work the goal is not to enforce non-negativity, but rather to focus on the sparsity related issues that come with using PCA in mass spectrometry data analysis, and PCA does not have such a non-negativity constraint built in. Moreso, with the focus on addressing sparse signal capture, none of the three models explored here optimize over a non-negative manifold or have a non-negativity constraint. Since we know molecular and ion species at a particular tissue location are either not present (zero ion count) or present (positive ion counts), and negative intensity counts have no physical meaning, we can use the presence of negative intensity values in these methods' low-rank components and approximations as a heuristic for an approximation's deviation from instrumental physics and biology. Deviations of a method's low-rank approximation from a non-negative model are captured using three different metrics: the percentage of negative entries in the low-rank term B of each method, the total sum of those entries, and the mean of those same entries. They reflect respectively the relative number of negative (and therefore non-biological) values in a low-rank approximation, the total magnitude of this mathematically correct, yet biologically improbable deviation, and the average magnitude of the deviation. Note that many more such metrics could be constructed. This particular set of metrics is selected for their simplicity and independence from the three methods explored here and their underlying models.

SUPPLEMENTARY INFORMATION TO THE RESULTS OF CASE STUDY 1

This section explores how different parameter settings influence the occurrence, magnitude, and distribution of negative entries in the low-rank approximations, evaluating their impact on interpretability.

NEGATIVE ENTRIES IN THE LOW-RANK APPROXIMATION

A low-rank approximation of a measurement set that is inherently non-negative can carry negative entries and still be useful for downstream analysis that does not require a physical interpretation of its components (*e.g.*, for data compression or preceding a supervised machine learning model). However, if the goal is to obtain a low-rank approximation of these non-negative measurements that can be biochemically interpreted, physical feasibility of the recovered components becomes more important. One could enforce non-negativity on a low-rank approximation of a dataset, as in non-negative matrix factorization [37]. However, here we are interested in gauging, when methods that do not enforce non-negativity are applied to inherently non-negative data, whether an added sparse residuals term C as in PCP and SPCP impacts the non-negativity of the low-rank approximation, and thus the physical feasibility and interpretability of the components and underlying trends retrieved from the data. Since negative ion counts are not physically possible, the percentage, sum, and mean of negative entries in a method's low-rank approximation that is not constrained to be non-negative could be used as heuristics to quantify that method's tendency to deviate from a physically feasible low-rank model. The results in terms of negative entries for PCA, PCP, and SPCP approximations are provided in the Supplementary Information, and discussed in-depth in sections dedicated to the percentages, sums, and means of those entries.

The results of Supplementary Figures S2.5, S2.6, and S2.7 suggest that, at least for this cornea dataset, SPCP outperforms both PCP and PCA for almost all parameter settings. Concurrently, the rank of the low-rank approximation term of SPCP and PCP seems positively correlated to the number of negative entries, *i.e.*, a higher rank will lead to more negative entries. However, the average negative entry will be small in magnitude. Our hypothesis is that this phenomenon is caused by noise being captured by the low-rank approximation term.

PERCENTAGE OF NEGATIVE ENTRIES

Supplementary Figure S2.5 shows the percentage of negative entries in the low-rank term B of SPCP (panel a), PCP (panel b), and PCA (panel c), and this for all explored parameter settings. The value at a particular location in panel (a) can be directly compared to the value found in panels (b) and (c) at that same location. For SPCP, we observe that for all parameter settings fewer than 10 % of the entries in B_{SPCP} contain negative values, while for PCP and PCA this percentage goes up to 18 %. While SPCP outperforms PCP and PCA for this data set for nearly all considered parameter values, PCP also outperforms PCA, but only for high-rank cases. Furthermore, in the region of approximations of relatively low rank (*i.e.*, below 20 %; see Suppl. Figs. S2.3b and S2.5c), we see that for high σ-multiplier values PCA and SPCP both produce a similar percentage of negative entries. This could indicate that in cases where the deviation from dataset A is allowed to be larger, that allowance for dense residuals could translate into less need to introduce negative entries into the low-rank approximation. Since PCA and SPCP both have a dense residuals term

D, this relieves their low-rank term B and sparse residuals term C from having to capture non-structured, non-sparse variation in the data, and presumably could allow the low-rank approximations to stay closer to the non-negative physical reality. While we have no proof of this hypothesis, it does seem to be confirmed by the relatively worse PCP results in Suppl. Figure S2.5b. Since the PCP model does not have a dense residuals term D, its low-rank term B and sparse residuals term C are forced to absorb non-structured, non-sparse variation in the dataset (by PCP's constraint that $A = B + C$), which could be the reason that PCP has many more negative entries in its low-rank approximations than SPCP has. Finally, we remark that a very similar smooth pattern can be observed in Suppl. Figure S2.5 as in the rank pattern of Suppl. Figure S2.3. This suggests that an increasing percentage of negative entries tends to correlate to increasing rank. This could mean that as more components are added to the low-rank approximation, there are more degrees of freedom to do a mathematically optimal approximation that nevertheless deviates from physical feasibility. It also suggests that if physical feasibility is important, a lower-rank approximation is probably more suitable than a higher-rank approximation, at least for cornea data. Overall, for all three methods, rank of their low-rank approximation seems to be positively correlated to the percentage of negative entries. Furthermore, in all tested parameter settings, SPCP has a smaller number of negative entries in the components of its low-rank approximations than PCP and PCA. Thus, SPCP tends to deliver more physically interpretable low-rank approximations than PCP and PCA, at least for this dataset.

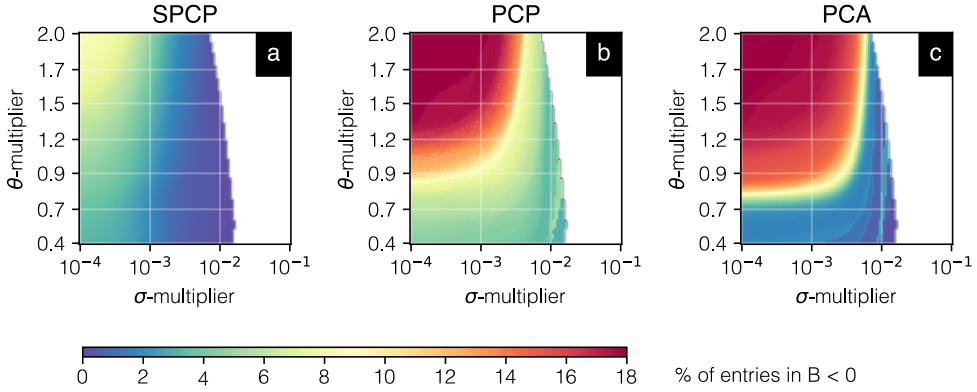

Figure S2.5: Percentage of negative entries in the low-rank term B of SPCP (a), PCP (b), and PCA (c). The percentage is calculated for each B_{PCA}, B_{PCP}, and B_{SPCP} that are rank matched for every θ and σ-multiplier parameter set. It reports the ratio (in %) of the number of negative entries in B (*i.e.*, $B_{ij} < 0$) to the total number of matrix entries (*i.e.*, mn for $B \in \mathbb{R}^{m \times n}$). A lower percentage of negative entries corresponds to a more physically feasible and interpretable approximation of the IMS data. For this dataset, SPCP seems to outperform both PCP and PCA for most of its parameter settings, requiring much fewer negative entries in its low-rank approximations to achieve the same rank and reduction of dimensionality.

SUM OF NEGATIVE ENTRIES

After investigating the number of negative entries in the different low-rank approximations, there is also value in assessing the magnitude of these deviations from physical

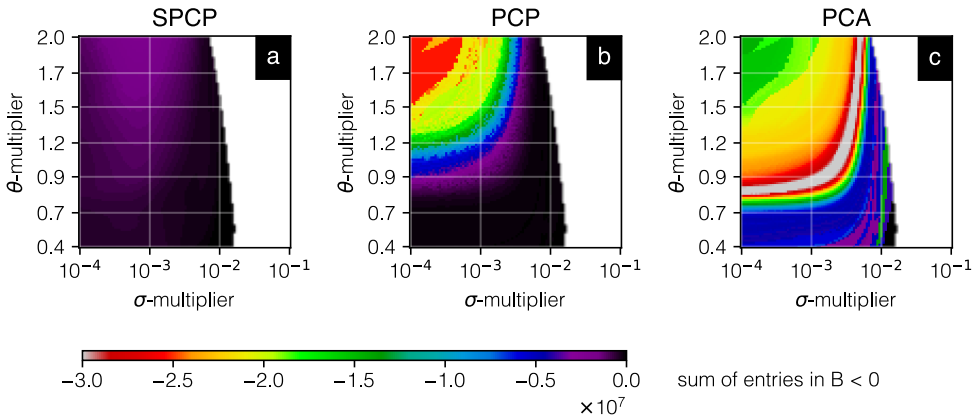

Figure S2.6: Sum of negative entries in the low-rank term B of SPCP (a), PCP (b), and PCA (c). The sum is calculated for each B_{PCA}, B_{PCP}, and B_{SPCP} that are rank matched for every θ and σ-multiplier parameter set. The sum reports the magnitude of the number of negative entries in B (*i.e.*, $B_{ij} < 0$). A value closer to zero corresponds to a more physically feasible and interpretable approximation of the IMS data. For this dataset, and similar to Suppl. Fig. S2.5, SPCP seems to outperform both PCP and PCA, for most of its parameter settings roughly by a factor 2 to 10.

feasibility. The sum of negative entries in B reflects the total dataset-wide magnitude of the deviation from easy interpretability of an approximation's components. We plot this sum for every examined parameter setting in Suppl. Figure S2.6.

For this data set, and similar to our observations regarding the percentage in Suppl. Figure S2.5, SPCP seems to outperform both PCP and PCA, for most of its parameter settings roughly by a factor 2 to 10. We also see that for PCP (Suppl. Figure S2.5b) with increasing rank the sum of negativity increases as well, while the inverse is observed for PCA (Suppl. Figure S2.5c). For rank-1 solutions (bottom-right), the negativity of all methods is comparable, while for low-rank settings (*i.e.*, below 20 % relative rank), SPCP's magnitude of negativity is the least of all three methods. Finally, we distinguish discrete steps in the very low-rank areas for PCA (bottom-right). This could be originating from high-intensity sparse features in the dataset being captured by the first few components of a low-rank approximation by PCA, and subsequent principal components trying to 'off-set' these components with a lot of negative entries to 'fulfill' PCA's orthogonality constraint (without the availability of a sparse residuals term). In conclusion, SPCP tends to accumulate the least total negative magnitude in its low-rank approximations when compared to PCP and PCA, and for this dataset we observe a positive correlation between the sum of negative entries and the rank of the approximations. This seems to further confirm that, given the same rank or dimensionality to work with, SPCP delivers more physically interpetable low-rank approximations than PCP and PCA, at least for this dataset.

MEAN OF NEGATIVE ENTRIES

The sum of negative entries gives an idea of the total deviation of a low-rank approximation from physical interpretability for inherently non-negative data. The percentage of negative entries gives an impression of how widespread this deviation is. Now, the

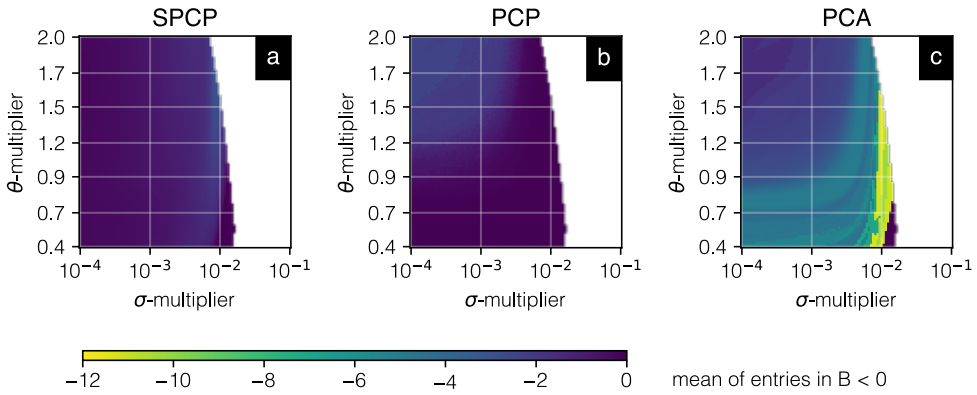

Figure S2.7: Mean of negative entries in the low-rank term B of SPCP (a), PCP (b), and PCA (c). The mean is calculated for each B_{PCA}, B_{PCP}, and B_{SPCP} that are rank matched for every θ and σ-multiplier parameter set. The mean reports the average magnitude of negative entries in B (*i.e.*, $B_{ij} < 0$). A value closer to zero corresponds to a more physically feasible and interpretable approximation of the IMS data. For this dataset, and similar to Suppl. Figs. S2.5 and S2.6, SPCP seems to outperform both PCP and PCA, for most of its parameter settings except for regions of 'large' σ-multiplier values. In these latter regions, SPCP outperforms PCP in terms of percentage and sum of negative entries in its low-rank approximation and thus has fewer negative values to begin with, but they do tend to be of larger average magnitude compared to PCP.

mean of negative entries can show us whether only few large magnitude negative values are present or whether there is a large number of smaller magnitude negative values. Supplementary Figure S2.7 shows the mean of negative entries in the low-rank term B of SPCP (panel a), PCP (panel b), and PCA (panel c), and this for all explored parameter settings. This figure can be interpreted in a manner similar to Suppl. Figs. S2.5 and S2.6.

For SPCP, a slightly higher mean is observed for σ-multiplier values close to 10^{-2}. This might be an artifact of the relatively lower number of negative entries in that area due to the very low-rank approximations achieved there. For PCP, one can discern again a correlation with the rank, albeit it a very slight one. In general, the mean negative entries for SPCP and PCP seem to be low compared to PCA. For PCA, we observe in the low rank region, *i.e.*, where the relative rank is below 20%, a relatively high mean of the negative entries, indicating that PCA's low-rank approximation is deviating quite a lot from a physically interpretable decomposition. Overall, the average negative entries seem the smallest for SPCP and PCP in comparison to PCA, at least for this dataset.

SUPPLEMENTARY INFORMATION TO THE RESULTS OF CASE STUDY 2

This section investigates how ion intensity is distributed among the decomposition terms (B, C, and D) for specific parameter settings and m/z-bins, and how these affect denoising and biological signal preservation.

We explore the residuals terms delivered by PCA, PCP, and SPCP for five select parameter settings:

- (Case 1) low rank B (r=126) through low θ-multiplier and low σ-multiplier;
- (Case 2) very low rank B (r=5) through low θ-multiplier and high σ-multiplier;

- (Case 3) very low rank B (r=3) through high θ-multiplier and high σ-multiplier;
- (Case 4) middle rank B (r=1 018) through middle θ-multiplier and middle σ-multiplier;
- (Case 5) high rank B (r=1 825) through high θ-multiplier and low σ-multiplier.

First, we compare the different matrix terms in general, using their element-wise histograms across all *m/z* bins to better understand how parameter settings influence the distribution of measured ion intensity among the different terms. Then, we compare the element-wise histograms within the scope of specific *m/z* bins to examine the impact of parameter settings on the ion species-specific intensity distributions. In the context of noise removal/reduction, we investigate how different parameter settings can impact the spatial distributions of the residuals terms' images. To this end, we provide a visualization of a measured ion image and its decomposition into low-rank, sparse, and dense images for the same set of *m/z* bins and parameter settings as before, albeit cropped to a sub-area within the retina for easier viewing (see Suppl. Figure S2.8). Finally, we demonstrate for two ion species, how the PCA, PCP, and SPCP methods decompose their measured ion images into a low-rank component, generally carrying tissue signal, a sparse component, sometimes carrying signal (*e.g.*, from small tissue structures), and a dense component that generally carries noise and can be removed. The comparison can be helpful in providing intuition to the reader to decide for their own datasets whether a sparse-signal-aware alternative to PCA, such as PCP or SPCP, can be useful for dimensionality and/or noise reduction purposes.

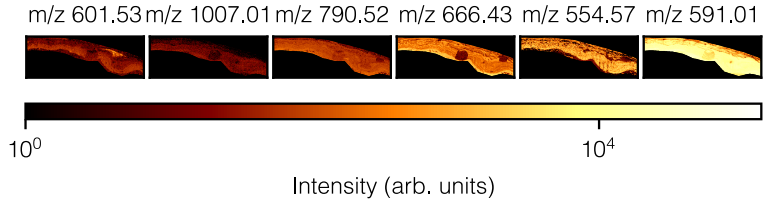

Figure S2.8: Six distinct *m/z* bins: *m/z* 601.53 (ion image with sparse spatial structures); *m/z* 1 007.01 (low intensity ion image); *m/z* 790.52 (low to average intensity ion image); *m/z* 666.43 (average intensity ion image); *m/z* 554.57 (high intensity ion image without strong outliers); and *m/z* 591.01 (high intensity ion image with strong outliers).

ION INTENSITY DISTRIBUTION AMONG TERMS

For Case 1 (low rank B), one can observe that the B and C terms of PCP and SPCP seem to match closely. This is expected, as when the σ-multiplier is low, very little to no energy can be captured by SPCP's D term. This makes SPCP's model (Equation 2.4) approximate PCP's model (Equation 2.3), leading to the match observed in the first row of Suppl. Figure S2.9. In noise reduction use cases, the measurement variation that is not structured and therefore cannot be captured well by a low-rank approximation B can be labeled as noise, suggesting that the dense residuals in D can be thrown out. Setting the σ-multiplier too high, as in Case 3 (very low rank B), releases SPCP from the need for $B+C$

to closely approximate A and results in a high amount of measurement energy to be captured by the dense residuals term D. As σ and its corresponding δ grow larger, there is no inherent mechanism preventing higher intensity biological signals from ending up in D. If biological signal variation is captured by D, removing D for noise reduction purposes is clearly undesirable. Overall, the presence of the D term gives SPCP decidedly an advantage over PCP, allowing tighter low-rank modeling (see Case Study 1 and Suppl. Figs. S2.5-S2.7). However, Case 3 illustrates that, in noise reduction use cases, if σ is set too high, there is a risk that genuine biological signal is removed. Case 1 and 3 together indicate that, when using SPCP, it is important to keep σ (and the corresponding δ) large enough to obtain the tighter modeling advantages that come with the presence of a dense residuals term, but small enough to avoid that higher intensity biological signals start entering D and are removed as noise. For SPCP in Case 2 (very low rank B), two peaks (in blue) can be observed at roughly -2 000 and +2 000 ion intensity in the histogram of the D term (more clearly visible in the D zoom-in). These are the result of the SPCP model (Equation 2.4) putting an upper intensity bound δ on the Frobenius norm of the difference between A and $B+C$, effectively clipping intensity variations that are not captured by the low-rank and sparse residuals terms to a maximum intensity. A similar set of clipping peaks can be observed in Case 4, albeit at ion intensities closer to zero since Case 4 uses a lower σ (and thus lower δ) than Case 2. Case 2 and 4 indicate that the σ-multiplier or direct specification of the δ-parameter in Equation 2.4 can be used to define an ion intensity threshold, below which dense measurement variation can be captured by D and be removed as noise, and above which dense measurement variation is forced to be captured by either the low-rank approximation term B or the sparse residuals term C, both of which tend to be kept as non-noise variation.

ION SPECIES-SPECIFIC EFFECTS

In the panels of Supplementary Figure S2.10, we show four histogram traces. We see in gray what remains after PCA's low-rank approximation, namely D_{PCA} (there is no C_{PCA}). In red, we see what remains after PCP has extracted its low-rank approximation, namely C_{PCP} (there is no D_{PCP}). In principle, C_{PCP} is optimized to capture sparse patterns. However, the red C_{PCP} traces consistently being wider than the blue C_{SPCP} profiles for the same data suggest that the lack of a dense residuals term D_{PCP} in PCP is a real impediment and forces C_{PCP} to capture more than just sparse variation. For SPCP, two traces are shown because what remains after SPCP's low-rank approximation is captured by $C_{\text{SPCP}} + D_{\text{SPCP}}$, *i.e.*, the sum of the blue and the yellow trace. The yellow trace represents non-low-rank non-sparse (dense) residuals and its content is usually suited for removal and denoising. Whether to remove the blue variation as noise depends on whether sparse features are important for the downstream analysis.

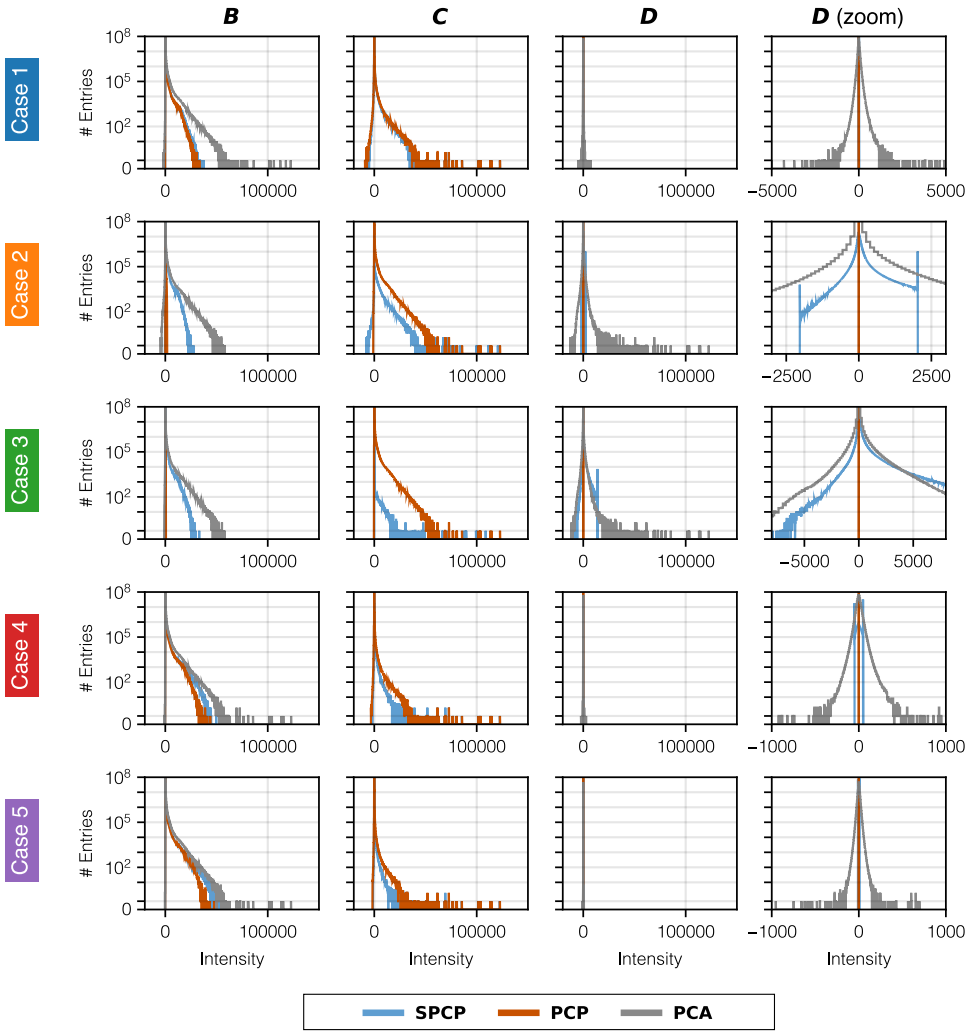

Figure S2.9: Global entry-wise intensity histograms for the B, C, and D terms delivered by SPCP, PCP, and PCA. For D an extra zoomed-in plot is shown. The impact of the different parameter settings is primarily visible in the C and D columns. We observe, *e.g.*, that the σ-multiplier for SPCP controls a dense residuals term as an "escape valve" for non-low-rank non-sparse variation in the measurements. It has a large influence on how the energy is distributed between C and D, and as such can be an extra handle to make C retrieve truly sparse patterns from the data.

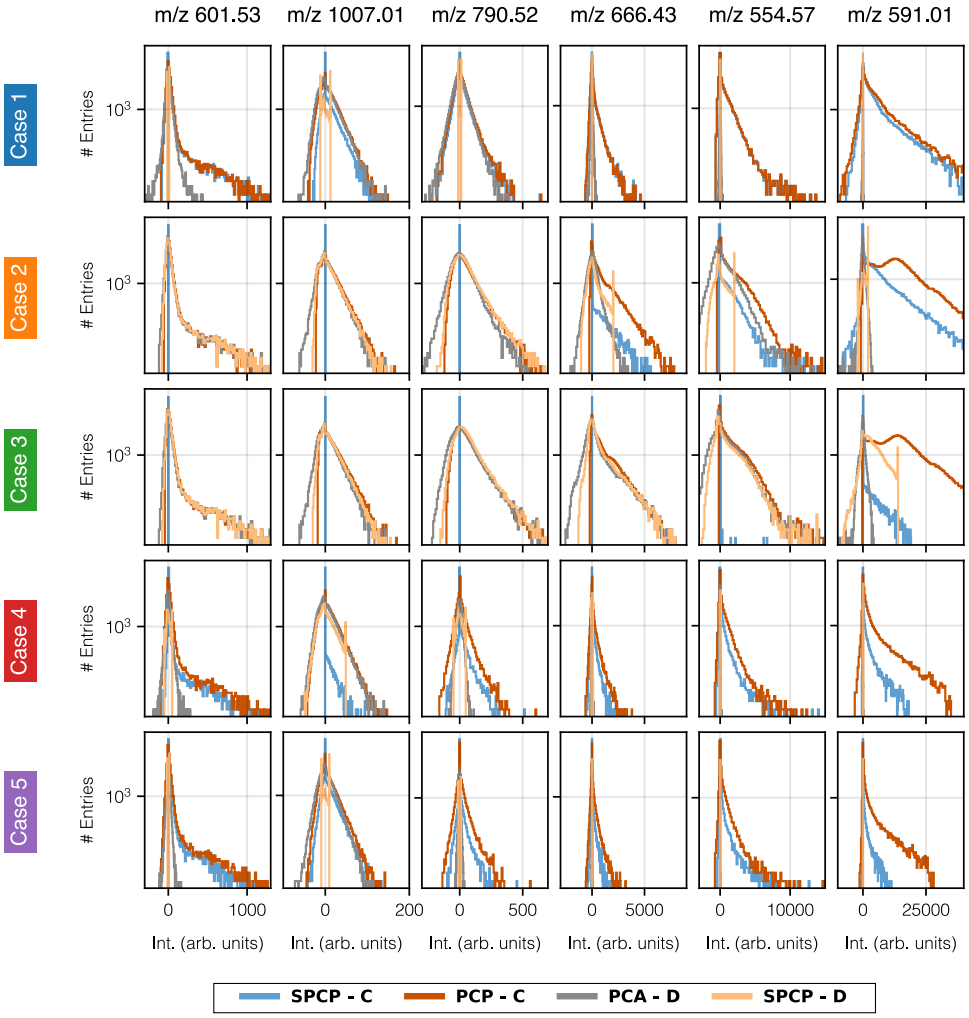

Figure S2.10: Mass-to-charge-specific entry-wise (for the whole data matrix) intensity histograms for the C and D terms delivered by SPCP, PCP, and PCA. The distribution of energy among C and D is examined for six distinct m/z bins: m/z 601.53 (ion image with sparse spatial structures); m/z 1 007.01 (low intensity ion image); m/z 790.52 (low to average intensity ion image); m/z 666.43 (average intensity ion image); m/z 554.57 (high intensity ion image without strong outliers); and m/z 591.01 (high intensity ion image with strong outliers). Different parameter settings lead to different residuals distributions local to specific ion species, and some of these effects are tied to the nature of the ion species' intensity level (low versus high intensity ions) and spatial distribution (*e.g.*, sparse features or not). Overall, residuals of PCP and SPCP tend to contain fewer negative intensity values, suggesting less overestimation of the mass spectral signals in these methods' low-rank approximations.

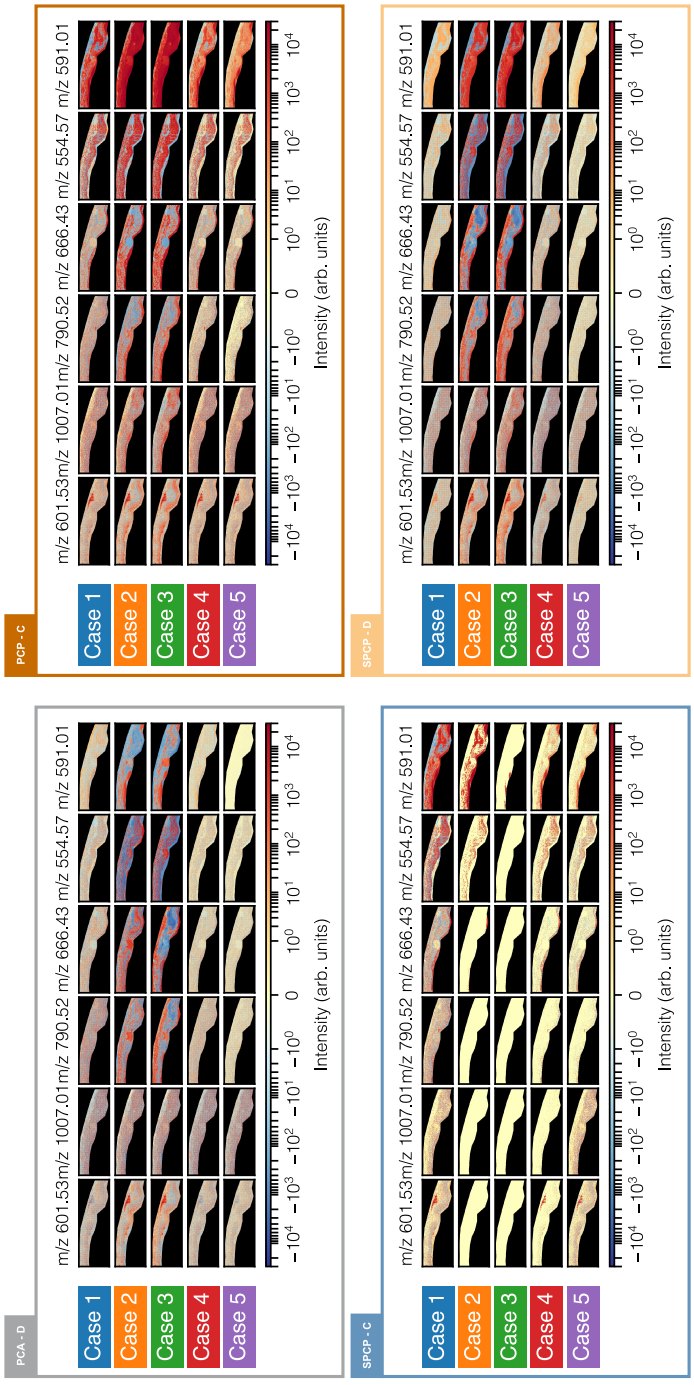

Figure S2.11: Images of the content of the sparse and dense residuals, *i.e.*, *C* and *D* terms, of the retina dataset for different *m/z* bins under different conditions for PCA, PCP and SPCP. Different parameter settings tend to lead to different results. One notable observation is that despite PCP's model trying to make the content of its *C* term contain sparse variation, the lack of a *D* term in PCP means that the *C* term nevertheless needs to absorb a lot of non-sparse variation as well (reflected by a lot of non-zero values in the 'PCP-*C*' panel, top-right). Since SPCP's model has an explicit *D* term for any variation that does not fit its *B* or *C* terms, SPCP's *C* term can be dedicated to capturing sparse variation (reflected by a lot of values close to zero in the 'SPCP-*C*' panel, bottom-left). The discrepancy between PCP's *C* and SPCP's *C* seems to suggest that SPCP is better at capturing genuine sparse variation.

REFERENCES

[1] R. M. Caprioli, T. B. Farmer, and J. Gile. Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. In: *Analytical Chemistry* 69.23 (1997), pp. 4751–4760.

[2] L. A. McDonnell and R. M. Heeren. Imaging Mass Spectrometry. In: *Mass Spectrometry Reviews* 26.4 (2007), pp. 606–643.

[3] W. J. Perry, C. M. Grunenwald, R. Van de Plas, J. C. Witten, D. R. Martin, S. S. Apte, J. E. Cassat, G. B. Pettersson, R. M. Caprioli, E. P. Skaar, et al. Visualizing Staphylococcus Aureus Pathogenic Membrane Modification Within the Host Infection Environment By Multimodal Imaging Mass Spectrometry. In: *Cell Chemical Biology* (2022).

[4] S. Kaspar, M. Peukert, A. Svatos, A. Matros, and H.-P. Mock. MALDI-imaging Mass Spectrometry–an Emerging Technique in Plant Biology. In: *Proteomics* 11.9 (2011), pp. 1840–1850.

[5] M. Stoeckli, P. Chaurand, D. E. Hallahan, and R. M. Caprioli. Imaging Mass Spectrometry: a New Technology for the Analysis of Protein Expression in Mammalian Tissues. In: *Nature Medicine* 7.4 (2001), pp. 493–496.

[6] HuBMAP Consortium. The Human Body at Cellular Resolution: the NIH Human Biomolecular Atlas Program. In: *Nature* 574.7777 (2019), p. 187.

[7] R. G. Cooks, Z. Ouyang, Z. Takats, and J. M. Wiseman. Ambient Mass Spectrometry. In: *Science* 311.5767 (2006). Cited by: 1260, pp. 1566–1570.

[8] Z. Takáts, J. M. Wiseman, and R. G. Cooks. Ambient Mass Spectrometry Using Desorption Electrospray Ionization (DESI): Instrumentation, Mechanisms and Applications in Forensics, Chemistry, and Biology. In: *Journal of Mass Spectrometry* 40.10 (2005). Cited by: 742; All Open Access, Bronze Open Access, pp. 1261–1275.

[9] J. Koch and D. Günther. Review of the State-of-the-art of Laser Ablation Inductively Coupled Plasma Mass Spectrometry. In: *Applied Spectroscopy* 65.5 (2011). Cited by: 229; All Open Access, Bronze Open Access, 155A–162A.

[10] A. M. Belu, D. J. Graham, and D. G. Castner. Time-of-flight Secondary Ion Mass Spectrometry: Techniques and Applications for the Characterization of Biomaterial Surfaces. In: *Biomaterials* 24.21 (2003), pp. 3635–3653.

[11] B. Mamyrin. Time-of-flight Mass Spectrometry (Concepts, Achievements, and Prospects). In: *International Journal of Mass Spectrometry* 206.3 (2001), pp. 251–266.

[12] S. Eliuk and A. Makarov. Evolution of Orbitrap Mass Spectrometry Instrumentation. In: *Annu. Rev. Analytical Chemistry* 8 (2015), pp. 61–80.

[13] A. G. Marshall, C. L. Hendrickson, and G. S. Jackson. Fourier Transform Ion Cyclotron Resonance Mass Spectrometry: a Primer. In: *Mass Spectrometry Reviews* 17.1 (1998), pp. 1–35.

[14] J. Fan, F. Han, and H. Liu. Challenges of Big Data Analysis. In: *National Science Review* 1.2 (2014), pp. 293–314.

[15] N. Verbeeck, R. M. Caprioli, and R. Van de Plas. Unsupervised Machine Learning for Exploratory Data Analysis in Imaging Mass Spectrometry. In: *Mass Spectrometry Reviews* 39.3 (2020), pp. 245–291.

[16] R. Van de Plas, F. Ojeda, M. Dewil, L. Van Den Bosch, B. De Moor, and E. Waelkens. "Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis". In: *Biocomputing 2007*. World Scientific, 2007, pp. 458–469.

[17] P. W. Siy, R. A. Moffitt, R. M. Parry, Y. Chen, Y. Liu, M. C. Sullards, A. H. Merrill, and M. D. Wang. Matrix Factorization Techniques for Analysis of Imaging Mass Spectrometry Data. In: *2008 8th IEEE International Conference on BioInformatics and BioEngineering*. IEEE. 2008, pp. 1–6.

[18] M. Nijs, T. Smets, E. Waelkens, and B. De Moor. A Mathematical Comparison of Non-negative Matrix Factorization Related Methods with Practical Implications for the Analysis of Mass Spectrometry Imaging Data. In: *Rapid Communications in Mass Spectrometry* 35.21 (2021), e9181.

[19] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? In: *Journal of the ACM* 58.3 (2011), pp. 1–37.

[20] D. M. Anderson, J. D. Messinger, N. H. Patterson, E. S. Rivera, A. Kotnala, J. M. Spraggins, R. M. Caprioli, C. A. Curcio, and K. L. Schey. Lipid Landscape of the Human Retina and Supporting Tissues Revealed By High-resolution Imaging Mass Spectrometry. In: *Journal of the American Society for Mass Spectrometry* 31.12 (2020), pp. 2426–2436.

[21] Z. Zhou, X. Li, J. Wright, E. J. Candes, and Y. Ma. Stable Principal Component Pursuit. In: *IEEE international symposium on information theory*. IEEE. 2010, pp. 1518–1522.

[22] H. Hotelling. Analysis of a Complex of Statistical Variables Into Principal Components. In: *Journal of Educational Psychology* 24.6 (1933), p. 417.

[23] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[24] Z. Lin, M. Chen, and Y. Ma. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-rank Matrices. In: *arXiv preprint arXiv:1009.5055* (2010).

[25] N. S. Aybat and G. Iyengar. An Alternating Direction Method with Increasing Penalty for Stable Principal Component Pursuit. In: *Computational Optimization and Applications* 61.3 (2015), pp. 635–668.

[26] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. John Wiley & Sons, 2009.

[27] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex Optimization Meets Low-rank Matrix Factorization: An Overview. In: *IEEE Transactions on Signal Processing* 67.20 (2019), pp. 5239–5269.

[28] A. Ganesh, J. Wright, X. Li, E. J. Candes, and Y. Ma. Dense Error Correction for Low-rank Matrices Via Principal Component Pursuit. In: *2010 IEEE international symposium on information theory*. IEEE. 2010, pp. 1513–1517.

[29] R. Moens. *On the Atoms of Robustness: Robust Matrix Decomposition for Spectral Imaging*. https://repository.tudelft.nl/islandora/object/uuid:78817107-d34a-4d6f-9ea1-d7298436d63c. 2021.

[30] D. M. Anderson, A. Kotnala, L. G. Migas, N. H. Patterson, L. E. Tideman, D. Cao, B. Adhikari, J. D. Messinger, T. Ach, S. Tortorella, et al. Lysolipids Are Prominent in Subretinal Drusenoid Deposits, a High-risk Phenotype in Age-related Macular Degeneration. In: *Frontiers in Ophthalmology* 3 (2023), p. 1258734.

[31] K. Sládková, J. Houška, and J. Havel. Laser Desorption Ionization of Red Phosphorus Clusters and Their Use for Mass Calibration in Time-of-flight Mass Spectrometry. In: *Rapid Communications in Mass Spectrometry* 23.19 (2009), pp. 3114–3118.

[32] Y. Zhang, L. Huang, N. Pillar, Y. Li, Y. Li, L. G. Migas, R. Van de Plas, J. M. Spraggins, and A. Ozcan. Virtual Staining of Label-free Tissue in Imaging Mass Spectrometry. In: *Science Advances* 11.31 (2025), eadv0741.

[33] L. G. Migas. *msalign: Spectral alignment based on MATLAB's 'msalign' function*. https://github.com/lukasz-migas/msalign. Version 0.2.0. 2024.

[34] P. Monchamp, L. Andrade-Cetto, J. Y. Zhang, and R. Henson. Signal Processing Methods for Mass Spectrometry. In: *Systems Bioinformatics: An Engineering Case-Based Approach, Artech House Publishers* (2007).

[35] E. J. Candès and B. Recht. Exact Matrix Completion Via Convex Optimization. In: *Foundations of Computational Mathematics* 9.6 (2009), pp. 717–772.

[36] A. Björck and G. H. Golub. Numerical Methods for Computing Angles Between Linear Subspaces. In: *Mathematics of Computation* 27.123 (1973), pp. 579–594.

[37] D. D. Lee and H. S. Seung. Learning the Parts of Objects By Non-negative Matrix Factorization. In: *Nature* 401.6755 (1999), pp. 788–791.

3

SPARSE OUTLIERS AS SOLUTION

Context. *Mid-infrared astronomy from the ground faces critical challenges in accurately detecting and quantifying sources due to the dominant, time- and spatially-variable background noise. Moreover, chopping and nodding - the traditional methods for dealing with these background issues - will not be technically feasible on the next generation of extremely large telescopes. This limitation requires the development of novel computational methods for robust background reduction.*

Aims. *We present and evaluate a novel method named LOw-RAnk Background ELimination (LORABEL) to improve the sensitivity of mid-infrared astronomical observations, without the need for classical telescope nodding, source masking, or other overheads in observing time.*

Methods. *We applied a low-rank background reduction strategy to (1) data taken on the ground with VISIR (VLT Imager and Spectrometer for mid-InfraRed) with synthetically injected sources, and (2) airborne data from SOFIA (Stratospheric Observatory for Infrared Astronomy). We compared the performance of our new method to classical chopping and nodding techniques, and analyzed the impact on source photometry and detection precision for different observational scenarios.*

Results. *In low signal-to-noise regimes (SNR* < 5*) in the ground-based VISIR data, LORABEL reduces variation in the photometric error with respect to chopping-differences only and even the classical chop-nod sequence, at the cost of introducing a bias. Secondly, we demonstrate that LORABEL increases detection precision in comparison to traditional background reduction methods. For the SOFIA dataset, we achieve a* $20-100$ *fold decrease in mean background flux with respect to the traditional chop-nod method, while preserving most of the source flux. Our findings suggest that LORABEL is applicable to a wider range of instrumental observation, i.e., both ground-based and airborne, and a suitable tool in the context of faint source detection.*

The contents of this chapter are based on:

> Moens, R. A. R., Pietrow, A. G. M., Brandl, B., & Van de Plas, R. (2025). Thermal Background Reduction for Mid-Infrared Imaging by Low-Rank Background and Sparse Point Source Modelling [Unpublished Manuscript, Submitted to Astronomy & Astrophysics].

3.1. INTRODUCTION

The thermal mid-infrared (mid-IR, $\lambda \approx 3-40\ \mu$m) covers an astronomically unique part of the electromagnetic spectrum, ideal for studies of warm dust, molecules in the interstellar medium, heavily obscured objects (protostars, active galactic nuclei), temperate exoplanets, solar system objects, and redshifted objects at large distances. Key questions addressed by mid-IR instruments include the evolution of star-forming regions, the compositions and surface temperatures of various celestial objects, solar flare electron densities, and the influence of molecular clouds on star formation and galactic dynamics. Historically, mid-IR imaging systems have evolved from earth-based, airborne, to space-based platforms, with notable instruments including VISIR [1] and TEXES [2], CanaryCam [3], SOFIA [4], IRAS [5] and Spitzer [6]. The most impressive images at mid-IR wavelengths have recently been delivered by the James Webb Space Telescope, [JWST, 7] with its 6.5m-aperture. While the boundary conditions for observing at mid-IR wavelength with a cooled telescope in space are unsurpassed in terms of thermal background and stability, the mission costs are usually extremely high and the telescope aperture size (key to achieving high angular resolution) has to be much smaller than a comparable facility on the ground. Very high angular resolution, which, for instance, is essential to directly image Earth-like exoplanets, requires a telescope aperture, much larger than 10 meters. Furthermore, ground-based and airborne instruments are easily accessible, allowing for modifications, and extended operational lifespans at relatively low costs. However, warm, ground-based telescopes face additional challenges, most notably the large thermal background, which can dominate the source signal by a large factor, and complicates the source detection [8]. In the best case, observational strategies and subsequent calibration measures permit the subtraction of the average thermal background, leaving "only" the Poissonian-distributed photon shot noise. However, the complexity and large number of variable thermal emitters in (or close to) the telescope beam have often led to residual background patterns, on top of the photon shot noise floor, which significantly reduced the detection limit. This leads to an unfortunate situation: while the new facilities on the ground will provide the angular resolution to directly image exoplanets near their host stars, insufficient background subtraction could lead to insufficient sensitivity to detect these exoplanets. Therefore, the elimination of residual background patterns is crucial for the detection of exoplanets and many other faint sources of interest [9, 10, 11]. The origin and exact timescales of these thermal background patterns have many different causes, are often even unknown, and differ from telescope to telescope. This makes their elimination difficult, and sometimes impossible. Generally, the thermal background can be categorized into (1) highly time-variable (sub-seconds scale) photon shot noise, (2) slowly variable (seconds scale) background emission patterns of low spatial frequencies, and (3) underlying quasi-static (minutes scale) patterns. To first order, the origin of (1) is the thermal (gray-body) emission from the atmosphere and the warm telescope, amplified by variations in detector responsivity, (2) is due to changes in telescope configuration (*e.g.*, due to tracking) and environmental conditions, and (3) is due to the optical configuration (*e.g.*, baffles). Traditionally, the background is reduced through the so-called “chopping and nodding” technique [12]. However, this method is not always feasible, especially not on the next generation of telescopes, like ESO’s Extremely Large Telescope (ELT). There, chopping is impossible due to the sheer size of the secondary mirror, and classical

nodding is not feasible due to the fact that ELT is not a stiff structure when tracking the object on the sky, but rather an actively controlled, dynamic five-mirror-telescope, that is constantly being realigned. Hence, new instruments, like the Mid-infrared ELT Imager and Spectrograph [METIS, 13] need alternative background reduction techniques. METIS will use a compact and fast internal chopper to take care of the fast variations (1), but requires additional methods to account for the slower variations, which have traditionally required nodding. Recent work [14, 15] on slightly more sophisticated chopping sequences has shown that novel reduction schemes can provide good results with chopping only and no additional telescope nodding. Other techniques are based on computationally modeling the chop residuals by PCA [16] or by drift scanning [17, 18, 19]. However, these methods only partially take the available information into account, require knowledge of source positions, timescales of background variation, and drift parameters, or decrease the observational efficiency altogether. To address these issues, we propose an algorithm that splits the data into a low-rank background image, a sparse signal and a random noise component, similar to a method used in high-contrast imaging [20]. The method presented here is utilizing chop-only measurements of point sources. The advantage of our approach is that it leverages the mathematical properties of the available data without requiring specific knowledge of, *e.g.*, point source positions, the background patterns, or their variability timescales. In addition, our developed method provides a path forward to model and reduce residual background patterns due to yet unknown boundary conditions, *e.g.*, for METIS at the ELT, where background structures, which cannot be reduced by classical methods, may occur [21].

We evaluate the photometric performance of our proposed method for point sources by (1) comparing it to artificially injected point sources in "empty" ground-based VISIR data, and (2) comparing it to chop-only and chop-nod methods in airborne SOFIA data to show its versatility. The structure of the chapter is as follows: we first introduce the definition of terms used throughout this chapter, followed by the introduction of the observational data and the methodology, including the specifics of aperture photometry. Subsequently, we present the results for both case studies, and finally, summarize and discuss our findings in the conclusions.

3.2. DEFINITIONS

3.2.1. NOMENCLATURE

Throughout our analysis we use a number of specific terms. The following glossary provides brief definitions of these terms in the context of this chapter.

- **Background Signal**: any undesired signal. This can be of astronomical (*e.g.*, clouds in front of sources, but also satellite flybys and perturbations caused by the Earth's atmosphere), instrumental (*e.g.*, interaction of signal with instrument components or telescope) and computational (*e.g.*, introduced by the data reduction pipeline) nature.
- **Chop-Nod Scheme**: scheme used to perform subtractions of measurements or time frames in different chop positions and nod positions to obtain an optimal background reduction.

- **Chop-Nod Subtraction**: subtraction obtained by subtracting two chop subtractions obtained from nodding positions.
- **Chop-Only Regime**: regime under which only chopping is possible and no nodding.
- **Chop Position**: off-axis tilt position of the secondary mirror. We use chop position in this chapter to denote the secondary mirror position of a measurement or time frame.
- **Chop Subtraction**: subtraction of two time frames in different chop positions.
- **Chopping**: performing a secondary mirror tilt to record off-axis measurements. Due to the off-axis tilt, sources move relative to the time frame. Chopping is oftentimes performed at a frequency of 1 Hertz or more.
- **Flux**: Average number of photons passing through a unit area per unit time.
- **Flux Density**: the density of flux (or photons) over a given area, measured in units of photons per square meter per second, or Analog-to-Digital Units (ADU) per square arcsecond.
- **Co-addition**: the process of combining multiple measurements over time to improve signal-to-noise ratio and reduce random variations in the data.
- **Measurement**: a non-integrated/non-added recording of a particular field of view of interest, in our case, a 2D image, measuring flux.
- **Nod Position**: position of the primary mirror. We use nod position in this chapter to denote the primary mirror position of a measurement or time frame.
- **Nodding**: performing a primary mirror shift. Due to this shift, sources move relatively to the time frame. Nodding is oftentimes performed on a timescale of 1 to 2 minutes.
- **Point Source**: a source without spatial structure, whose morphology is simple, given by the optical performance of the imaging system. In this chapter, we solely deal with point sources. Ideally, they appear as airy patterns in the measurements or time frames. We consider them to be small with respect to the field of view. Perfect background subtracted measurements, or time frames, can thus be considered to be sparse.
- **Source Signal**: signal of interest. In this chapter, we only deal with point sources as signal of interest.
- **Pixel**: we consider a pixel a single element/entry of an individual time frame.
- **Time Frame**: an individual time frame is the result of the co-addition of measurements, *e.g.*, to counteract noise variations.
- **Time Frame Average**: average of a time series, *i.e.*, representing flux densities across

a 2D image.

- **Time Series**: a set of time frames recorded over a particular time span. Time series consist of different time frames, one for each point in time.

3.2.2. MATHEMATICAL NOTATIONS

We denote a scalar with a lowercase letter $a \in \mathbb{R}$, a vector of size m with a bold lowercase letter $\mathbf{a} \in \mathbb{R}^m$, and a matrix of size m-by-n with a capital letter $A \in \mathbb{R}^{m \times n}$. The vectorizing operation vec(A) acts on a matrix and transforms it into a vector by stacking all columns or rows into a single column or row: $\mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{mn}$. Furthermore, we will use $||A||_*$ to denote the nuclear norm of a matrix A, $||A||_F$ for the Frobenius norm and $||\mathbf{a}||_2$ as vector 2-norm. The absolute value is denoted as $|\cdot|$. The rank of a matrix A, rank(A), denotes the dimension of the subspace containing the matrix A. Generally, we talk about low-rank, when $\text{rank}(A) \ll \min(m, n)$. The low-rank property of matrices has several favorable mathematical properties and can be used to describe data from the smallest possible set of bases. The sparsity of a matrix is considered to be the number of non-zero values in its entries. High sparsity implies many zero values. When we talk about a sparse regime, this implies that a matrix or vector contain many zero values in their entries. When we consider a small dense noise matrix A, we imply that $||A||_F \leq \delta$, *i.e.*, small in Frobenius norm, and that it has non-zero values in (almost) all entries (low sparsity), *i.e.*, dense. Finally, we use superscript letters $\mathbf{a}^j$ to denote chop positions, subscripts letters $\mathbf{a}_t$ to denote time frames, and bars ($\bar{\mathbf{a}}$) and hats ($\hat{\mathbf{a}}$) for different nod positions. Furthermore, when we consider a dense noise matrix, the noise is considered to be independent and identically distributed (i.i.d.).

3.2.3. NOISE AND ALL OF ITS FRIENDS

Mid-infrared astronomical observations are impacted by two primary noise categories: systematic and statistical noise, both considered background signal in this chapter. Systematic noise arises from instrumental artifacts, detector imperfections, and optical path limitations, creating pseudo-consistent, structured variations across measurements that depend primarily on local conditions and telescope orientation. Statistical noise, on the other hand, follows a Poissonian detection process and is present in sources as well as in background. We will model this noise for the background using an independent and identically distributed (i.i.d.) Gaussian approximation. In this framework, i.i.d. means each noise sample is statistically independent and drawn from the same Gaussian probability distribution. While Poissonian noise fundamentally represents discrete, integer-based photon counting with variance equal to the mean, the Gaussian approximation provides a continuous, computationally tractable representation for signal reconstruction, particularly valid when photon counts are sufficiently high.

3.3. OBSERVATIONAL DATA, SOURCE DETECTION, PHOTOMETRY AND METRICS

3.3.1. VISIR DATA CASE STUDY

The VLT Imager and Spectrometer for mid-InfraRed [VISIR, 1] is a ground-based mid-infrared observational instrument mounted to the Cassegrain focus of the second unit telescope of the Very Large Telescope (VLT). This telescope is one of four 8.2-meter-diameter telescopes that make up the core of the Paranal Observatory in Chile. VISIR provides high-resolution imaging and spectroscopy in the mid-IR wavelength range [1]. Our VISIR dataset was acquired as technical time observations [14]. The dataset used in our first case study, referrred to here as "empty" VISIR data, contains only background and no source signal, as it was taken pointing close to zenith without any target. The objective is to recreate an idealized dataset based on VISIR data to study our methodology under idealized circumstances. This VISIR data is therefore firstly low-rank approximated by performing a rank-7 singular value decomposition (SVD), defining a low-rank ground truth background. Secondly, we add Gaussian i.i.d. noise to our measurements with $\mu = 0$ and $\sigma = 1$ to mimic Poisson noise-related effects. Both the low-rank and the injection of Gaussian i.i.d. noise allows us to optimally set the method parameters (see Methodology). Finally, we inject sources (see specifics below and in **Table 3.1**)

INJECTED SOURCES

The injected sources are modeled by two 13-pixel full width at half maximum (FWHM) Gaussian profiles, mimicking positive and negative sources of chopped VISIR images. The scale of these Gaussian profiles is derived by calculating the FWHM as

$$\text{FWHM} \approx 1.22 \frac{\lambda}{D}, \tag{3.1}$$

where D is 8.2 m for VISIR and $\lambda = 20$ μm is defined to be in the Q-band. This leads to an FWHM of approximately 2.98×10^{-6} radians or around 0.6 arcsec. With VISIR's pixel scale being 0.045 arcsec/pixel in the small-field mode [SFM, 22], this leads to a 13-pixel FWHM. The two injections are then scaled, one to +1 and the other source to −1, such that

$$\begin{aligned} |F_{\text{source}}| &= \sum_{t \in n_t} \sum_{(x,y) \in A} \left| I(x,y,t) \right|, \\ &= 1. \end{aligned} \tag{3.2}$$

F_{source} is the total source flux, $I(x, y, t)$ is the flux density at position (x, y) at time frame t (with a total of n_t time frames) and A denotes the spatial area considered source. In a last stage, this unit source flux is then multiplied by the square root of the number of pixels used in the aperture photometry (n_{pix}), the standard deviation σ_{bkg} of a square aperture at the position of injection in the chop residual and the injection signal-to-noise ratio (SNR) to obtain the final source flux

$$F_{\text{source}} = \text{SNR}_{\text{injection}} \times \sqrt{n_{\text{pix}} \cdot \sigma_{\text{bkg}}^2} \tag{3.3}$$

The location of the source injection is randomly generated for each experiment and the positive and negative sources are generated minimally $3 \times \text{FWHM} + 30$ pixels apart. The

Table 3.1: Specifications for the VLT/VISIR observations of the "empty" field with injected sources and SOFIA/FORCAST observations of γ Cygni. The table provides a comprehensive overview of the astronomical, instrumental, preprocessing, and mathematical specifics for both datasets. Astronomical specifics include target coordinates and observation times. Instrumental details describe the chop/nod configurations, frequencies, and spectral filters utilized. The preprocessing section covers integration times, co-addition, flat-fielding, and bad pixel correction, ensuring optimized data quality. Mathematical specifics detail the data matrix structures used for analysis, including spatial and temporal specifics of the recorded data.

Specifications	**VLT/VISIR**	**SOFIA/FORCAST**
Astronomical Specifics		
Object	Injected point source of varying intensity	γ Cygni
Pointing	RA 276.743637 deg, DEC -25.43501 deg	RA 330.472792 deg, DEC 48.731228 deg
Time of Recording	2016-03-22T23:37:35.7488	2022-05-11T09:22:10.584
Instrumental Specifics		
Chop Directions	0, 90, 180, and 270 deg	0, 90, 180, and 270 deg
Chop Frequencies	4 Hz	1, 4, and 5 Hz
Nod Directions	Up-down and left-right configuration	Up-down and left-right configuration
Spectral Filter(s)	J8.9	F088, F111, F197
Preprocessing Specifics		
Co-addition	0.0125s integration time, 7 sub-integrations	48.4406 sampled measurement rate, 22 co-additions in original data production
Other Preprocessing	Flat-fielding, dark current subtraction, bad pixel correction	Flat-fielding, dark current subtraction, bad pixel correction, data is normalized for co-adding
Mathematical Specifics		
Data Sizes	900×1024 pixels for 171 time frames, *i.e.*, a data matrix $A \in \mathbb{R}^{921\,600\times 171}$	256×256 pixels for 24 time frames for 2 nod positions and 2 chop directions, *i.e.*, 4 data matrices $A_i \in \mathbb{R}^{65\,536\times 24}$ $\forall\, i \in [1,4]$

size of the background square at each injection position (for which the standard deviation is calculated) is defined as $3 \times \text{FWHM} + 15$ pixels. To yield a variety of source strengths for evaluating our methods, the $\text{SNR}_{\text{injection}}$ takes the following values: 1, 2, 3, 5, 10, 20 and 50 for evaluating the photometric error and values 1 to 10 for detection evaluation. Importantly, note that we use the chop residual as reference for background standard deviation calculation. An example of an injection in the time frame average is given in **Fig. 3.1**.

3.3.2. SOFIA DATA CASE STUDY

The Stratospheric Observatory for Infrared Astronomy [SOFIA, 4] was an airborne ∅2.5 m. telescope inside a modified Boeing 747SP, which was active from 2010 to 2022. One of its instruments was the Faint Object infraRed CAmera for the SOFIA Telescope [FORCAST, 23], a dual-channel mid-infrared camera and spectrograph sensitive from $5-40\ \mu$m. SOFIA could operate at altitudes of up to 14 km, above most of the water vapor in Earth's atmosphere, thereby providing clear views of the infrared sky. The observations taken under the mission ID `2022-05-11_FO_F867` and plan ID `71_0025` were obtained during

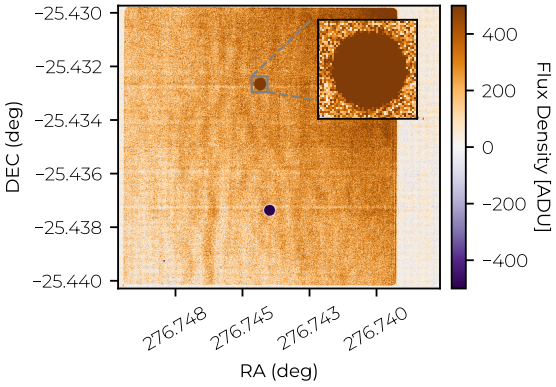

Figure 3.1: Time frame average for VLT/VISIR case study with injected source. See **Table 3.1** for the data specifics.

technical time, and observed γ Cygni. The specifics of the SOFIA dataset are also given in **Table 3.1**. The objective is to show performance of our methodology in non-ideal circumstances and show its versatility (not only restricted to ground-based observations). The time frame average is given in **Fig. 3.2**

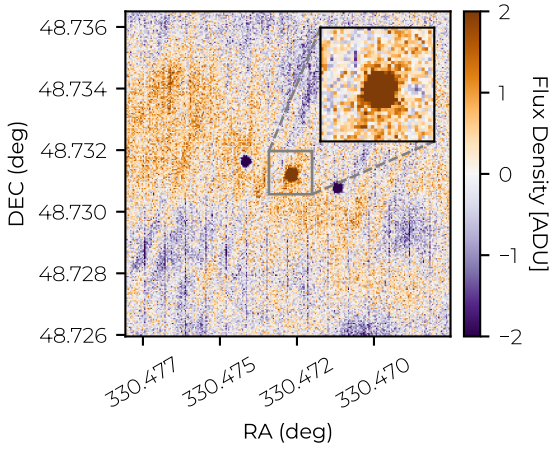

Figure 3.2: Time frame average for SOFIA/FORCAST case study. See **Table 3.1** for the data specifics.

3.3.3. SOURCE DETECTION

In this chapter, source detection was carried out using the DAOStarFinder algorithm, provided by the Python package Photutils [24]. The DAOStarFinder parameters considered were: (1) the FWHM, and (2) the detection threshold. The FWHM parameter was fixed at 13 pixels for both Chop-Only and Chop-Nod methods, whereas it was set to 11 pixels for

LORABEL. The threshold, denoted by ϵ, was computed using the following relation:

$$\epsilon = f \times \sigma_c \times \mathrm{SNR}_{\mathrm{injection}}, \tag{3.4}$$

where f is a factor used for tuning the detection sensitivity, σ_c is the standard deviation derived from a 3-sigma clipping procedure of the whole time frame average, and $\mathrm{SNR}_{\mathrm{injection}}$ is the resulting signal-to-noise of the injected source. The values of the threshold factor f used in this study are summarized in Table 3.2. This threshold factor and threshold have an influence on the outcomes of the DAOStarFinder algorithm. As we inject sources using a local standard deviation measure this might sometimes conflict with the global standard deviation, i.e., defined over the whole field of view, used in Eq. 3.4, as such lowering of the threshold is often necessary to obtain better results.

Table 3.2: Threshold factor (f) values used for different detection methods as a function of the injected source SNR. The very low values, i.e., 10^{-10}, at low injection SNRs reflect the idea of at least detecting the true positive source, where at higher injection SNRs the false negative rate needs to be controlled to improve the precision. These factors were partially heuristically set on a separate training set of 10 injections.

$\mathrm{SNR}_{\mathrm{injection}}$	Chop-Only	Chop-Nod	LORABEL
1	10^{-10}	10^{-10}	10^{-10}
2	10^{-10}	10^{-10}	10^{-10}
3	10^{-10}	10^{-10}	10^{-10}
4	0.6	0.8	0.7
5	0.6	0.8	0.7
6	0.5	0.7	0.6
7	0.4	0.7	0.5
8	0.4	0.7	0.5
9	0.4	0.7	0.5
10	0.4	0.7	0.5

Finally, we filter the results obtained from DAOStarFinder to reduce false positives by requiring the *roundness1* and *roundness2* to be < 0.3 and the *peak SNR* of the detected object > 2.

3.3.4. CLASSICAL APERTURE PHOTOMETRY (CAP)

Classical aperture photometry (CAP) was performed using the Python toolbox Photutils [24]. The aperture and annulus settings are given in **Table 3.3** and an example for VLT/VISIR, is given in **Fig. 3.3**.

This geometric configuration is empirically set for VISIR's and SOFIA's diffraction-limited PSF and mid-infrared sky background characteristics. The source positions were known a priori for VLT/VISIR and manually set for SOFIA/FORCAST.

The classical aperture photometry (CAP) is defined as

$$\hat{F}_{\mathrm{source}} = \sum_{t \in n_t} \sum_{(x,y) \in A} \left[I(x,y,t) - \mu_{\mathrm{annulus}} \right], \tag{3.5}$$

Table 3.3: Aperture photometry parameters used for the VISIR and SOFIA datasets. Measurements include the central aperture size and the inner and outer sizes of the background annulus, specified in pixels and arcsecs. Both aperture and annulus are square. These values are based on measured source sizes in the data. For VLT/VISIR we use a value of 3 × FWHM to ensure that almost all flux is captured for a reliable photometric error measurement.

Parameter	VLT/VISIR	SOFIA/FORCAST
Aperture size	39 pixels (1.76″)	15 pixels (11.52″)
Annulus inner size	39 pixels (1.76″)	17 pixels (13.06″)
Annulus outer size	69 pixels (3.10″)	30 pixels (23.04″)

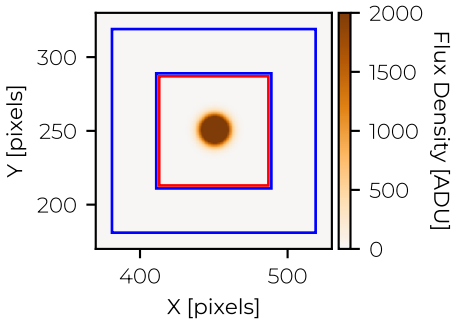

Figure 3.3: Square aperture photometry methodology applied to VISIR data. The source flux is measured within the red square aperture centered on the target position, while the surrounding blue square annulus defines the region used for background estimation and subtraction. The aperture and annulus are offset for visualization purposes.

where $I(x, y, t)$ is the flux density at position (x, y) at time t, A denotes the spatial area considered source, defined by the aperture, and μ_{annulus} is the median flux density in the annulus. $\hat{F}$ is the estimated source flux density. For chop-only, we apply CAP on the chop time frame average, for the chop-nod data we apply CAP on the chop-nod time frame average, and for LORABEL we apply CAP on the C time frame average. For all methods we keep integration time equal.

3.3.5. METRICS

We have two main metrics, being the photometric error and the SNR. The photometric error (ϕ) is defined as

$$\phi = \frac{\hat{F}_{\text{source}} - F_{\text{source}}}{F_{\text{source}}} \times 100\%, \tag{3.6}$$

where $\hat{F}_{\text{source}}$ is the estimated source flux density from CAP after background removal and F_{source} is the injected source flux. We used a simple SNR relation defined as

$$\text{SNR} = \frac{\hat{F}_{\text{source}}}{\sqrt{n_{\text{pix}} \cdot \sigma^2_{\text{bkg}}}}, \tag{3.7}$$

where we assume no read-out noise, no dark current noise and no photon shot noise associated to the point source. Finally, for the detection precision (π) we use the following definition

$$\pi = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

where TP (True Positives) counts the correct detections and FP (False Positives) counts the incorrect detections.

3.4. METHODOLOGY

We assume that a measurement at chop position j, time step t and one particular nod position α or β, $X_t^{j,\alpha} \in \mathbb{R}^{m \times n}$, consists of source signal $S_t^j \in \mathbb{R}^{m \times n}$, and additive background signal $E_t^{j,\alpha} \in \mathbb{R}^{m \times n}$, such that we can model it as

$$X_t^{j,\alpha} = S_t^{j,\alpha} + E_t^{j,\alpha}. \tag{3.8}$$

For each chop position, the source moves in the focal plane due to a tilt of the secondary mirror. With nodding, the entire telescope structure is moved, such that point sources also shift and chop positions will move sources in different directions relative to the time frame (see **Fig. 3.4**). An example image is given in **Fig. 3.1** (VLT/VISIR), where we observe a positive and negative point source that make up $S_t^{j,\alpha}$ and the background signal $E_t^{j,\alpha}$ that appears as a gradient from left bottom to right top.

3.4.1. TRADITIONAL CHOP-NOD SUBTRACTION

The traditional way to perform background subtraction is by performing a simple chop-nod scheme. These techniques are used to mitigate the effects of strong and variable background radiation in mid-infrared astronomy. The chop sequence is performed with the smaller secondary mirror, which allows for fast offsets to track variability of the background on sub-second time scales. However, the tilt of the secondary mirror breaks the symmetry of the optical system (with respect to the optical axis) and introduces residuals, which require a "mirrored" observing sequence. The offset for this mirrored sequence is performed with the primary mirror. This "nodding" with the bigger and more massive, primary mirror takes more time, which is acceptable, as the above-mentioned residuals vary on a much slower timescale of minutes. Here, we consider four different measurements, that is, two measurements at different chop positions (X_t^1 and X_t^2) and two measurements at different nod positions (*e.g.*, $X_t^{1,\alpha}$ and $X_t^{1,\beta}$) for each chop position. First, we obtain a chop subtraction for each nod position by simple subtraction of the two

chop images for each nod position

$$\begin{aligned} X_t^{\alpha} &= X_t^{1,\alpha} - X_t^{2,\alpha} \\ &= \underbrace{(S_t^{1,\alpha} - S_t^{2,\alpha})}_{\text{chop sources: } S_t^{\alpha}} + \underbrace{(E_t^{1,\alpha} - E_t^{2,\alpha})}_{\text{chop residual: } \Delta E_t^{\alpha}} . \end{aligned} \tag{3.9}$$

Note that we obtain two copies of the original source, one positive and one negative pattern. To further reduce imperfections left by the chop subtraction, so-called chop residuals and denoted by ΔE_t^{α} and ΔE_t^{β}, nod subtraction is applied. The resulting chop-nod subtraction at time step t, i.e., Y_t, can then be formulated as

$$\begin{aligned} Y_t &= \underbrace{\overbrace{(X_t^{1,\alpha} - X_t^{2,\alpha})}^{\text{chop subtraction}} - (X_t^{1,\beta} - X_t^{2,\beta})}_{\text{chop-nod subtraction}} \\ &= \underbrace{(S_t^{\alpha} - S_t^{\beta})}_{\text{chop-nod sources: } S_t} + \underbrace{(\Delta E_t^{\alpha} - \Delta E_t^{\beta})}_{\text{chop-nod residual: } \Delta E_t} . \end{aligned} \tag{3.10}$$

We now obtain three source patterns in our chop-nod subtraction Y_t of the original source, two negatives and one positive sources, where the positive source has double intensity with respect to the negative sources. This scheme is usually individually applied on all available time frames and then subsequently time averaged, *e.g.*, to reduce for photon shot noise, i.e, $Y = \frac{1}{n_t}\sum_t Y_t$, where n_t is the total number of available time frames. Note that this method exploits source signal shifts relative to the background signal to prevent source subtraction or removal. The assumption is that (1) a large part of the background signal in each chop position is similar (i.e., remains constant due to a fast chop rate), and (2) that chop residuals (here considered residual background signal) in two different nod positions are similar in nature. As such, no mathematical (structural) properties of the background are exploited. This scheme starts from sub-noise sources and is able to result in sup-noise sources, with large enough SNR for safe use further down the analysis.

3.4.2. PROPOSED METHOD: LOW-RANK BACKGROUND ELIMINATION (LORABEL)

For LORABEL, we require a time series of n_t chopped/chop-only frames (so no nodding). Ideally, the frames are sufficiently integrated[1]. This is largely dependent on the instrumental set-up, experimental design and preprocessing pipeline. The number of chopped frames n_t required for the working of LORABEL depends on the complexity (rank) of the background structure[2]. The latter can be estimated by performing an SVD on an "empty" dataset and looking at leading singular values. In general, more frames are better for a more accurate noise reduction.

The method we propose is based on a low-rank background signal and sparse source signal model. As such, it acts on the chop-only time frames of point sources, i.e., $X_t =$

[1] as a rule of thumb, the structural residual patterns should be visible in individual frames, not purely randomly distributed noise/photon shot noise

[2] as a rule of thumb, we take at least 10-fold number of frames with respect to the rank

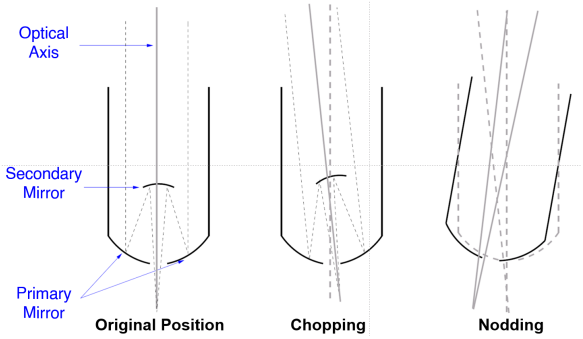

Figure 3.4: Abstract visualization of chopping and nodding, adapted from Lenzen [25]. The left panel shows the original position of the telescope and secondary mirror. The middle panel illustrates the chopping technique, where the secondary mirror is rapidly tilted back and forth to observe the target and a nearby reference position. The right panel depicts the nodding technique, where the entire telescope assembly is periodically tilted to observe the target and a nearby reference position.

$X_t^1 - X_t^2$ (for simplicity we omit the α and β nod position notation as we have only a single position α or β), where we will assume that the chop residuals (i.e., residual background signal, ΔE_t) remains (nearly) constant over different time frames. To apply LORABEL, we first group the chop subtractions of the available time series as follows

$$X = \begin{bmatrix} \text{vec}(X_1) & \text{vec}(X_2) & \cdots & \text{vec}(X_{n_t}) \end{bmatrix} \in \mathbb{R}^{mn \times n_t}. \tag{3.11}$$

Each column in the matrix X consists of a vectorized chop subtraction at time step j, while each row represents a time series of a single imaging detector pixel (that is also chop subtracted). Using Eq. 3.9, we can decompose our matrix as

$$X = S + \Delta E, \tag{3.12}$$

where $S \in \mathbb{R}^{mn \times n_t}$ contains the point sources signal and $\Delta E \in \mathbb{R}^{mn \times n_t}$ consists of quasi-static instrumental patterns (i.e., low-rank), slowly variable background emission patterns of low spatial frequency (i.e., low-rank) and highly time-variable photon shot noise (considered to be Poisson/Gaussian noise). As such, our proposed model can be considered a member of the family $\mathscr{F}$ of extended linear models:

$$\mathscr{F}: \; A = B + C + D, \tag{3.13}$$

where $A \in \mathbb{R}^{mn \times n_t}$ represents a measurement matrix that is decomposed into $B \in \mathbb{R}^{mn \times n_t}$, $C \in \mathbb{R}^{mn \times n_t}$, and $D \in \mathbb{R}^{mn \times n_t}$ which are matrix terms differing in mathematical properties. Here, we define them as a low-rank term (B), a sparse term (C), and a residual term ($D = A - B - C$). For our purposes, B and D are expected to model the content of ΔE, as we will respectively assume the quasi-static instrumental patterns and slowly variable background emission patterns of low spatial frequency to be low-rank and highly time-variable photon shot noise to be modelled as small dense noise (see Mathematical Notations). At the same time, the point sources are considered to be spatially sparse even though a second copy is created during chopping, as such we can model them by C. **Fig. 3.5** shows

this decomposition of input data (A) into three components: (B) a low-rank thermal background, (C) a source signal containing both positive and negative sources, and (D) Gaussian i.i.d. noise. This decomposition illustrates how the various signal components interact (with different flux densities), with (B) the low-rank background representing a large-scale thermal structure (considered noise), (C) a source signal capturing the astrophysical objects of interest, and (D) a randomly distributed noise component simulating observational uncertainties. A schematic overview of the framework is given in **Fig. 3.6**.

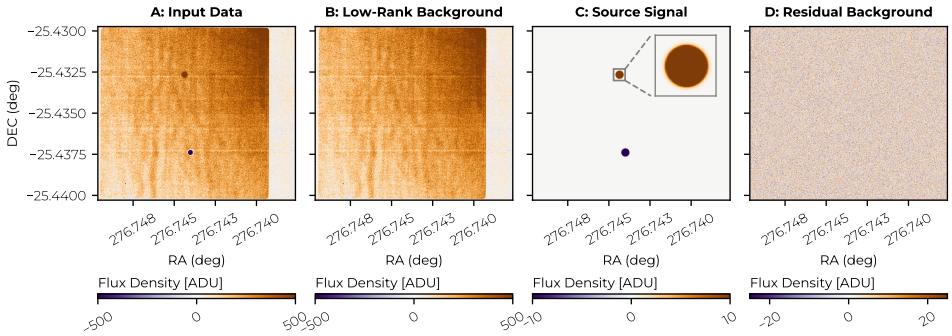

Figure 3.5: Ideal decomposition of the input image with very large injection SNR (A) into three components: (B) low-rank thermal background, (C) injected source signal with positive and negative sources, and (D) Gaussian noise. The combined model follows A = B + C + D, illustrating how each component contributes to the final observed data. Note that these images are on a different intensity scale.

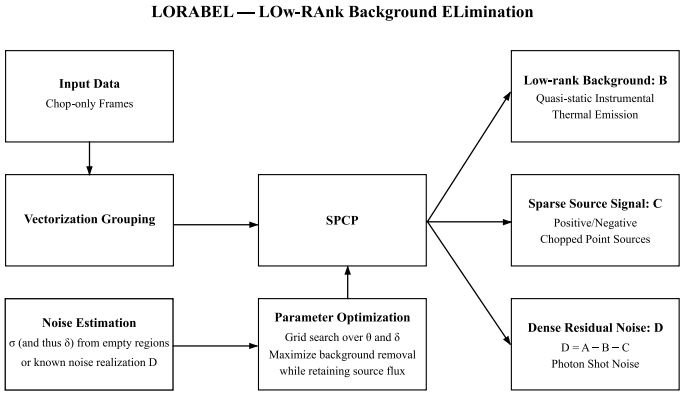

Figure 3.6: Schematic overview of the proposed LOw–RAnk Background ELimination (LORABEL) framework. A time series of chop-only frames is vectorized and stacked into a data matrix $A \in \mathbb{R}^{mn \times n_t}$. At the core of LORABEL, Stable Principal Component Pursuit (SPCP) decomposes the data into a low-rank component B capturing quasi-static instrumental and thermal background structures, a sparse component C containing the astrophysical point-source signal (including positive and negative chop copies), and a dense residual term $D = A - B - C$ modelling photon shot noise. The SPCP parameters θ (controlling the sparsity–rank trade-off) and δ (noise tolerance) are optimized via a grid search, where δ is linked to the estimated noise level σ obtained from signal-free regions, ensuring efficient background suppression while preserving source flux.

POINT SOURCES: LOW-RANK AND SPARSE

In an idealized setting, aside from their inherent spatial sparsity, point source signals would appear as a rank-1 feature in the time series data as presented in Eq. 3.11. However, since the flux of these point sources is typically much smaller than the overall background flux, especially when they have low SNR, their corresponding component is relatively weak in an ℓ_2-sense and does not dominate the leading components that constitute the B term. A key assumption here is that the temporal components of the point sources remain uncorrelated with the quasi-static instrumental and slowly varying, low-spatial-frequency background emissions represented by the B term. Our first case study reveals that this assumption can be violated, causing some point source signal to leak into the B term. A similar effect was observed in exoplanet detection for a similar class of methods [26]. In real-world observations, such as those from SOFIA/FORCAST, factors like atmospheric jitter, variations in instrument response, and imperfect preprocessing can alter this rank-1 property of point sources as its spatial shape, location and temporal pattern can alter in a non-linear fashion. Rather than remaining confined to a rank-1 representation, point sources tend then to "bleed" across multiple components and are thought to become less correlated to the stable low-rank background. Interestingly, this effect could actually lead to improved retrieval of the source flux density compared to what the idealized model predicts. This property is actually exploited by some drift scanning techniques [18].

OPTIMIZATION

The above problem of decomposing low-rank, sparse and small dense terms can be solved by stable principal component pursuit [SPCP, 27]. In their paper, they provide conditions related to (1) the incoherence of bases [28], (2) the rank of the low-rank term B, (3) the sparsity of the sparse term, C, and (4) the magnitude of the dense term, $||A-B-C||_F$, that leads to an approximate solution (close to the exact solution). We will assume these conditions are valid for our problem. The optimization program defined in **Equation 3.14** is proposed to solve this problem. We will denote Stable Principal Component Pursuit (SPCP) as, $f_{\text{SPCP}} : \mathbb{R}^{m\times n} \times \mathbb{R}^{+} \times \mathbb{R}^{+} \to \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n}; (A,\theta,\delta) \mapsto (B,C)$:

$$\begin{aligned} \underset{B,C}{\text{minimize}} \quad & \|B\|_* + \theta \|C\|_1 \\ \text{subject to} \quad & \|A-B-C\|_F \leq \delta. \end{aligned} \tag{3.14}$$

One observes that an input matrix A should be provided along with two parameters, θ and δ. θ trades-off the sparsity of the matrix C with the rank of B (i.e, a trade-off between source signal retention and background signal reduction), while δ acts as a cut-off related to noise. For i.i.d. element-wise, Gaussian noise with mean zero and standard deviation σ, δ is defined as $\delta = \sqrt{mn}\sigma$.

Also, observe that for background modelling we do not have to explicitly deal with bad pixels, *e.g.*, caused by pixels saturation, cosmic rays, satellite flyby's, as they would naturally be filtered and end up in the C term as long as they are occurring as sparse and not coherent phenomena. Additionally, this model does not require masking, removing point sources or having information about dithering settings.

PARAMETER TUNING

As previously outlined in **Equation 3.14**, the model involves two key parameters that require tuning: $\delta \in \mathbb{R}$ and $\theta \in \mathbb{R}$. The parameter δ can be estimated by the standard deviation σ of the i.i.d. Gaussian noise assumed to be present in the data. In our first case study $sigma$ is set manually and therefore known exactly, whereas in the other case studies it can be estimated from signal-free, *i.e.*, "empty", regions of the data. Meanwhile, an initial estimate for θ is given by $\theta = \sqrt{1/\max(m,n)}$, where m, n are the matrix sizes of the input matrix $A \in \mathbb{R}^{m \times n}$ [27]. These estimates assume that ideal conditions are met. At the same time, it has been suggested that fine-tuning these parameters can yield improved results [29].

For the first case study, we therefore performed a grid search between $0.1 \times \bar{\theta} \leq \theta \leq 1.1 \times \bar{\theta}$ where $\bar{\theta} = \sqrt{1/\max(m,n)}$ and fixed $\delta = \|E\|_F$, where E is the matrix assumed exclusively containing i.i.d. Gaussian noise (known a priori), as by definition $\|E\|_F = \sigma^2 mn$, where σ is defined as the standard deviation of the noise matrix D. For detection, these parameters are slightly adjusted to a $\theta = 1.1 \times \sqrt{1/\max(m,n)}$ and a $\delta = .7 \times \|A\|_F$. This ensured that most of the low-rank background was removed, and that source signal is retained as much as possible.

For the second case study, we conducted a parameter tuning exercise by performing a grid search over a 100×100 parameter setting range for both parameters around $\theta = \sqrt{1/\max(m,n)}$ and $\delta = \|A\|_F$ (note that A is the input matrix, as E is not known a priori). This grid search was executed using two A6000 NVIDIA GPUs to accelerate the process. The optimal parameter set was selected based on the criteria of maximal background reduction and maximal retention of the source signal by performing aperture photometry for each parameter setting. Heuristically, we found that setting the search space to $0.1 \times \sqrt{1/\max(m,n)} \leq \theta \leq \sqrt{1/\max(m,n)}$ and $10^{-10} \times mn \leq \delta \leq 10 \times mn$ is most efficient. As a rule of thumb, when increasing θ more background flux is removed and less source signal is kept, while lowering θ leads to less background flux removal, but more source signal retained. The parameter δ has an influence on both. When increasing δ, it allows for more background subtraction but probably leads to less source flux, while lowering δ results most of the time in more source flux but also more background flux.

3.4.3. AVAILABILITY OF ALGORITHMS AND DATA

All methods are implemented in object-oriented Python and bundled in our *LORABEL* package, available from `https://github.com/vandeplaslab/lorabel`. The data used in this work is public and can be found on the ESO and SOFIA archives, or requested by contacting the authors.

3.5. RESULTS AND DISCUSSION

In the first case study, our LORABEL method is compared to the chop-only and chop-nod method on ground-based data with injected sources. This enables an absolute evaluation for accurate source flux measurement as well as source detection. In the second case study, we compare the proposed method to a chop-only and chop-nod method on airborne SOFIA data. This enables an evaluation in a practical setting at different wavelengths and with a different telescope structure. The overall goal of both case studies is to research

the method's ability in ideal circumstances, i.e, with injected sources, and in practical settings.

3.5.1. QUANTITATIVE STUDY OF GROUND-BASED VISIR DATA WITH INJECTED SOURCES OF VARYING INTENSITY

The photometric error for the three methods with different injected source flux densities (expressed as a function of the injection SNR, $SNR_{injection}$) is given in **Fig. 3.7**. It consists of the percentage difference between the injected source flux and recovered source flux of the positive point source with respect to the injected source flux. We used classical aperture photometry (see Section 3.3.4) to obtain the recovered source flux. The experiment is performed on 300 different injection locations, where the local noise statistics change and thus the injection, to assess robustness. Finally, we kept integration time for all methods equal. The following observations were made:

1. **Injection SNR** ≤ 5: at very low injection SNRs, LORABEL shows a considerably smaller distribution of photometric errors compared to both chop-only and chop-nod. Its boxplot indicates a smaller interquartile range and less pronounced outliers, suggesting that LORABEL is less susceptible to noise fluctuations in this regime. In contrast, the chop-only and chop-nod methods tend to exhibit a wider distribution, implying less consistent performance when the source signal is very weak.

2. **Injection SNR** > 5: for higher injection SNR values, the LORABEL performance improves slowly. Its error distribution becomes much tighter, but is caught up by chop-nodding.

3. **LORABEL's bias**: the median photometric error for LORABEL does not center around zero, but rather is systematically shifted to approximately $\approx -30\%$. This negative median indicates a systematic underestimation of the recovered source flux density relative to the injected flux.

4. **Decrease in error with increasing injection SNR**: all methods exhibit a general decrease in photometric error as the injection SNR increases. This trend is expected because, at high source flux densities, the background contribution becomes relatively negligible.

The systematic underestimation, observed in **Fig. 3.7**, is thought to be caused by the injection of a perfect low-rank source, which in reality is probably partially reduced (see Section 3.4.2). This bias further depends on the data, i.e, instrument type and observation specifics, as well as parameter settings of LORABEL. As the bias appears to be systematic and only weakly related to the injection SNR, one could estimate this bias by synthetically injecting one or multiple sources with known source fluxes, estimating this bias and compensating for it. Of course, this would only be a solution in the case of perfect low-rank sources, such as our injections. In the case of "imperfect" point sources, one could bound spatial point sources shapes and temporal patterns and search for lower and upper bounds on the resulting variation in bias. This is, however, considered to be out-of-scope for this chapter.

The notably reduced variability in photometric error for LORABEL at low SNR values potentially greatly impacts the detection rate of very faint sources that other methods may not find. To evaluate the detection performance, we conducted a comparative detection experiment in which data processed by all three methods were subsequently analyzed using our detection algorithm (see Section 3.3.3). The superior performance of the LORABEL method is demonstrated clearly in **Fig. 3.8**, where LORABEL achieves consistently higher precision across all tested injection SNR levels. The displayed precisions represent averages calculated over 100 randomly selected injection locations, utilizing a uniform set of algorithm parameters across locations. It is important to note that using a fixed parameter set, rather than optimizing parameters individually for each injection site, occasionally leads to scenarios where no detection is achieved (resulting in true positives, TP = 0) or an increase in false positives (FP), thereby reducing overall precision.

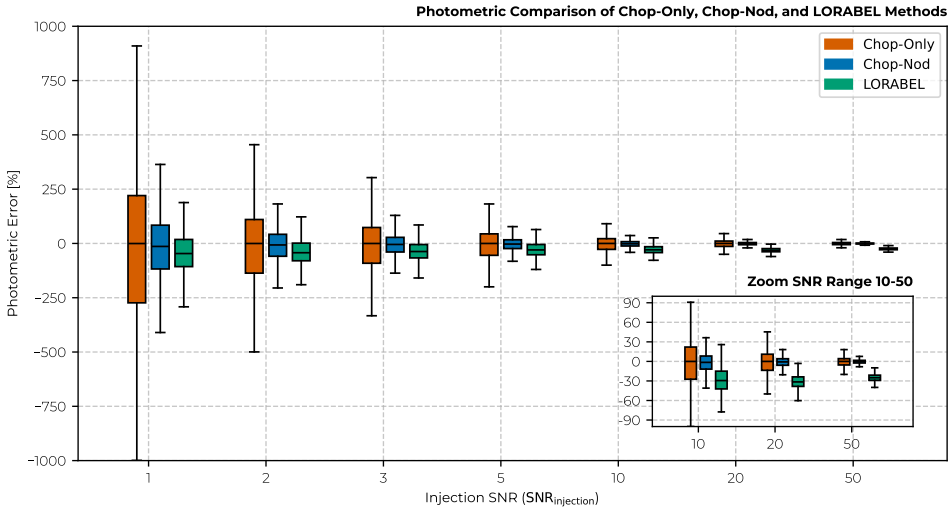

Figure 3.7: Comparison of photometric error using our proposed LORABEL method, the Chop-Only method, and the Chop-Nod method across different injection Signal-to-Noise Ratio (SNR) levels. The y-axis displays the photometric error in percentage. This comparison reveals the relative performance of each method under various observing conditions, particularly highlighting their accuracy in challenging low-SNR scenarios. The error variation for chop-only is seen to be the greatest across all injection SNR levels, whereas LORABEL performs better than Chop-Nod in low injection SNR regions (up to around 5). Finally, it is observed that LORABEL introduces a bias in the photometric error, a systematic underestimation of the source flux density.

3.5.2. COMPARATIVE STUDY WITH TRADITIONAL CHOP-NOD ON AIRBORNE SOFIA DATA

We evaluated the versatility of LORABEL by comparing it with traditional chop-nod techniques using airborne SOFIA observations. For this dataset, optimizing parameters (see 3.4.2) proved challenging due to the lack of sufficient time frames. In addition, the parameter setting was strongly influenced by dataset characteristics. Therefore, we combined the approach described earlier with heuristic tuning to achieve reasonable parameter

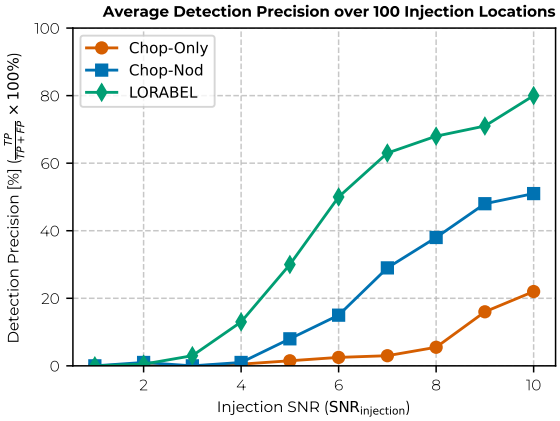

Figure 3.8: Detection precision comparison among the traditional Chop-Only and Chop-Nod methods and our proposed LORABEL approach. The figure presents the average detection precision computed over 100 random injection locations, where the local noise statistics vary with each injection. To avoid bias from selective parameter tuning, we employ a fixed set of detection parameters (see Section 3.3.3) derived from a training set. The constraints *roundness1*, *roundness2* and *peak SNR* are kept identical across all methods to reduce false positives, which results in "sub-optimal" outcomes as they could be tuned for each individual method. Furthermore, the use of local noise statistics for the injection and global noise statistics for detection threshold (see Section 3.3.3) further leads to a reduced detection precision. Finally, we note that this detection precision goes to 100% for high SNRs, not considered relevant for this experiment. Nonetheless, this setup ensures a fair comparison, and LORABEL consistently demonstrates superior precision at all SNR levels, especially excelling in lower SNR regimes (SNR < 5)

settings. In **Table 3.4**, the photometric analysis across multiple spectral filters reveals several key performance characteristics for LORABEL. The aperture photometry yields comparable mean flux values between LORABEL and the traditional chop-nod technique for most spectral filters, except in F111, where the chop-nod approach exhibits significant variability ($\sigma = 3211.2$ counts), likely due to temporal source misalignment. In contrast, background estimation with LORABEL proves to achieve superior stability across all bands ($\mu = -0.002$, $\sigma = 0.139$), substantially outperforming both chop-only and chop-nod approaches. Signal-to-noise analysis further demonstrates enhanced performance relative to traditional chop-nodding across all filters, with particular improvements in F088 and F111. Although the SNR in F197 is slightly reduced due to decreased aperture flux, it remains competitive. These results validate the practical applicability of LORABEL to airborne infrared observations, successfully extending its utility from theoretical framework to real observational data.

Visual inspection of the results, as shown in the decomposition of the time-integrated SOFIA F197 spectral filter in **Fig. 3.9**, demonstrates that LORABEL effectively separates the background, source signal, and noise components. Although the source signal component (C) does not exhibit a perfectly flat background across all pixels, it nonetheless facilitates a clearer analysis of the source under challenging observational conditions. This is further confirmed by the examination of the background distribution around the source, as illustrated in **Fig. 3.10**.

Table 3.4: Comparison of different spectral filters for the chop-only, LORABEL, and traditional chop-nod methods on aperture, peak, annulus, and SNR statistics. The reported averages (μ) and standard deviations (σ) are obtained with classical aperture photometry by considering the time frame average of each of the 4 available repeated measurements as a separate realization (see Table 3.1). "Aperture" refers to the source flux density, "Peak" denotes the maximum pixel brightness detected by the DAOStarFinder routine in the aperture, and "Annulus" reflects the mean flux density in the annulus. The SNR in this table is obtained by averaging source flux density and background flux density standard deviation over all 4 repeated measurements.

Method	Filter	Aperture $\mu \pm \sigma$	Peak $\mu \pm \sigma$	Annulus $\mu \pm \sigma$	SNR
Chop-Only	F088	4736.8 ± 39.2	394.7 ± 28.8	0.133 ± 0.838	376.8
	F197	1338.4 ± 59.5	90.5 ± 6.4	-0.095 ± 0.462	193.1
	F111	3353.8 ± 663.8	254.1 ± 30.0	-0.516 ± 1.066	209.7
Chop-Nod	F088	4227.6 ± 61.4	386.4 ± 16.7	0.100 ± 1.051	268.2
	F197	877.1 ± 62.3	93.0 ± 4.3	0.500 ± 0.626	93.4
	F111	2425.5 ± 3211.2	308.3 ± 53.1	-1.000 ± 1.975	81.9
LORABEL	F088	4247.0 ± 119.8	391.3 ± 28.5	-0.002 ± 0.139	445.9
	F197	1040.0 ± 84.6	87.1 ± 6.7	-0.001 ± 0.026	176.0
	F111	2931.0 ± 545.6	248.6 ± 27.8	0.032 ± 0.259	880.2

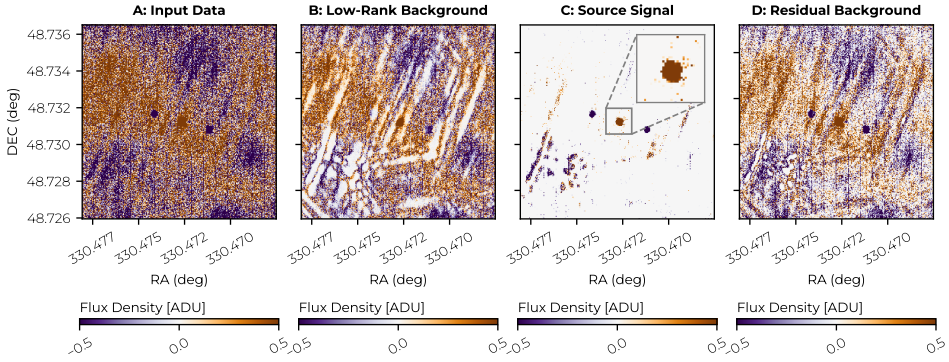

Figure 3.9: Time-integrated data from a single F197 spectral filter, illustrating the separation of low-rank background, sparse source signal, and photon shot noise achieved by LORABEL. This decomposition allows for clearer identification of the source signal even in the presence of significant background fluctuations. Some residuals still appear (mainly) in the bottom-left of the *C* image, which is undesirable. In this dataset, it is caused by a lack of time frames and the variability of this part of the background structure (i.e., only appearing in a few frames). Changing the parameters of our method, in this case, could not prevent this.

3.6. CONCLUSIONS

In this work, we introduced a novel computational technique, LORABEL, for background subtraction in mid-infrared astronomy, and benchmarked its performance against both chop-only and traditional chop-nod methods. Using simulated VISIR data with injected point sources as well as real SOFIA observations, our results reveal several key findings. For the VISIR case study, LORABEL delivered more consistent flux measurements in low signal-to-noise regimes compared to the conventional methods. Although the method

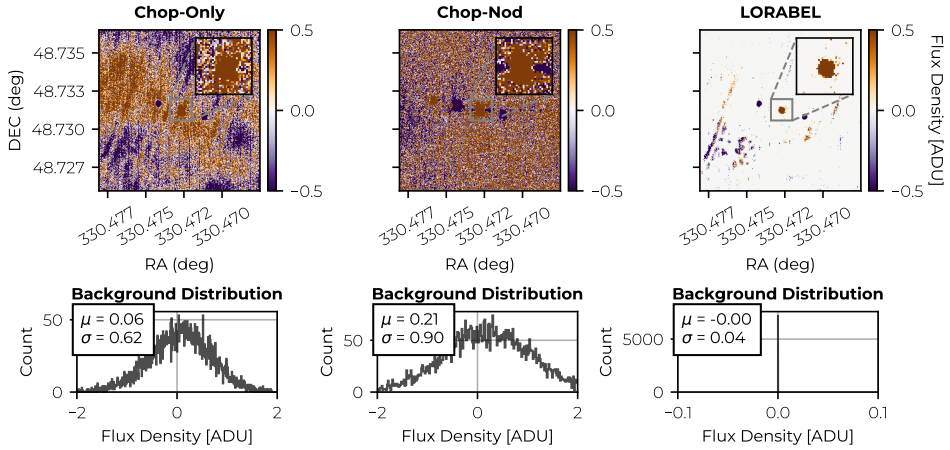

Figure 3.10: Comparison of background subtraction methods for SOFIA mid-infrared observations. Top panels: Time-integrated images showing the sky background residuals for the Chop-Only (left), Chop-Nod (center), and LORABEL (right). Insets provide zoomed views around the source. Bottom panels: Flux density distributions measured in an annular region of 117 pixels surrounding the source (with the source masked). The mean (μ) and standard deviation (σ) of the background residuals are indicated. LORABEL exhibits superior background subtraction, with a notably reduced spread in background flux density compared to traditional techniques. Note that the chop-nod method shows increased photon shot noise due to the combination of two measurements (increased by a factor of $\sqrt{2}$, i.e, $\sqrt{2} \times 0.62 \approx 0.90$). Additionally, small misalignments in the SOFIA data may contribute to incomplete background reduction. Note that the Chop-Only here corresponds to the same "A: Input Data" as in **Fig. 3.9** (due to scale differences the colours might not appear equal).

systematically underestimates the source flux, suggesting that a portion of the true signal is inadvertently removed, it substantially reduces variability in the photometric error. While the percentage of flux loss is dependent on the parameter settings and dataset specific, a way to compensate for this loss in a real-world setting could be by introducing artificial sources with known source flux in a dataset. Furthermore, we show that LORABEL is particularly promising for detecting faint sources under challenging conditions. We showed in our second experiment that LORABEL outperforms both chop-only and chop-nod methods in detection precision for all SNR cases. In the airborne SOFIA dataset, LORABEL achieved a reduction in the mean background flux (by factors of 3 to 10) while largely preserving the source signal, leading to improved signal-to-noise ratios across all spectral bands with respect to traditional chop-nodding. Despite a slight increase in the standard deviation of the source flux measurements, the results confirm that LORABEL can perform even in non-ideal scenarios with limited time frames. While the optimization of LORABEL can be challenging due to its complex parameter space and sensitivity to various observational factors, our results clearly show that effective background subtraction is achievable without relying on source masking or additional nodding. In summary, LORABEL shows significant potential for enhancing mid-infrared observations, especially for source detection in low SNR regimes when nodding is not available. Nonetheless, further refinements are required to mitigate its systematic flux underestimation and to fully optimize its performance under varying observational con-

ditions.

REFERENCES

[1] Y. Rio, P.-O. Lagage, D. Dubreuil, G. A. Durand, C. Lyraud, J.-W. Pel, J. C. de Haas, A. Schoenmaker, and H. Tolsma. VISIR: the Mid-infrared Imager and Spectrometer for the VLT. In: *Infrared Astronomical Instrumentation*. Vol. 3354. SPIE. 1998, pp. 615–626.

[2] J. Lacy, M. Richter, T. Greathouse, D. Jaffe, and Q. Zhu. TEXES: A Sensitive High-resolution Grating Spectrograph for the Mid-infrared. In: *Publications of the Astronomical Society of the Pacific* 114.792 (2002), p. 153.

[3] C. M. Telesco, D. Ciardi, J. French, C. Ftaclas, K. T. Hanna, D. B. Hon, J. H. Hough, J. Julian, R. Julian, M. Kidger, C. C. Packham, R. K. Pina, F. Varosi, and R. G. Sellar. CanariCam: a multimode mid-infrared camera for the Gran Telescopio CANARIAS. In: *Instrument Design and Performance for Optical/Infrared Ground-based Telescopes*. Ed. by M. Iye and A. F. M. Moorwood. Vol. 4841. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. 2003, pp. 913–922.

[4] A. Krabbe. SOFIA Telescope. In: *Airborne Telescope Systems*. Vol. 4014. SPIE. 2000, pp. 276–281.

[5] W. Irace and D. Rosing. The IRAS Telescope. In: *Journal of the British Interplanetary Society* 36 (1983), pp. 27–33.

[6] M. W. Werner, T. L. Roellig, F. Low, G. H. Rieke, M. Rieke, W. Hoffmann, E. Young, J. Houck, B. R. Brandl Brandl, G. Fazio, et al. The Spitzer Space Telescope Mission. In: *The Astrophysical Journal Supplement Series* 154.1 (2004), p. 1.

[7] G. H. Rieke, G. Wright, T. Böker, J. Bouwman, L. Colina, A. Glasse, K. Gordon, T. Greene, M. Güdel, T. Henning, et al. The Mid-infrared Instrument for the James Webb Space Telescope, I: Introduction. In: *Publications of the Astronomical Society of the Pacific* 127.953 (2015), p. 584.

[8] L. Burtscher, I. Politopoulos, S. Fernández-Acosta, T. Agocs, M. van den Ancker, R. van Boekel, B. Brandl, H.-U. Käufl, E. Pantin, A. G. Pietrow, et al. Towards a Physical Understanding of the Thermal Background in Large Ground-based Telescopes. In: *Ground-based and Airborne Instrumentation for Astronomy VIII*. Vol. 11447. SPIE, 2020, pp. 1678–1694.

[9] D. Petit Dit De La Roche, M. van den Ancker, M. Kissler-Patig, V. Ivanov, and D. Fedele. New Constraints on the HR 8799 Planetary System From Mid-infrared Direct Imaging. In: *Monthly Notices of the Royal Astronomical Society* 491.2 (2020), pp. 1795–1799.

[10] K. Wagner, A. Boehle, P. Pathak, M. Kasper, R. Arsenault, G. Jakob, U. Käufl, S. Leveratto, A.-L. Maire, E. Pantin, et al. Imaging Low-mass Planets Within the Habitable Zone of α Centauri. In: *Nature Communications* 12.1 (2021), p. 922.

[11] E. Matthews, A. Carter, P. Pathak, C. Morley, M. Phillips, S. K. PM, F. Feng, M. Bonse, L. Boogaard, J. Burt, et al. A Temperate Super-Jupiter Imaged with JWST in the Mid-infrared. In: *Nature* (2024), pp. 1–4.

[12] R. Papoular. The processing of infrared sky noise by chopping, nodding and filtering. In: *Astronomy and Astrophysics* 117.1 (1983), pp. 46–52.

[13] B. R. Brandl, R. Lenzen, E. Pantin, A. Glasse, J. Blommaert, M. Meyer, M. Guedel, L. Venema, F. Molster, R. Stuik, et al. METIS: the Thermal Infrared Instrument for the E-ELT. In: *Ground-based and Airborne Instrumentation for Astronomy IV*. Vol. 8446. SPIE. 2012, pp. 554–566.

[14] A. G. M. Pietrow. "Mid-IR background calibrations for the E-ELT's METIS instrument". MA thesis. Leiden University, 2016.

[15] A. G. M. Pietrow, L. Burtscher, and B. Brandl. Inverse Chop Addition: Thermal IR Background Subtraction without Nodding. In: *Research Notes of the American Astronomical Society* 3.2, 42 (2019), p. 42.

[16] H. Rousseau, S. Ertel, D. Defrère, V. Faramaz, and K. Wagner. Improving Mid-infrared Thermal Background Subtraction with Principal Component Analysis. In: *Astronomy & Astrophysics* 687 (2024), A147.

[17] S. Heikamp, B. R. Brandl, C. U. Keller, L. Venema, E. Pantin, R. Siebenmorgen, D. Ives, and F. Kerber. Drift scanning technique for mid-infrared background subtraction. In: *Ground-based and Airborne Instrumentation for Astronomy V*. Ed. by S. K. Ramsay, I. S. McLean, and H. Takami. Vol. 9147. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. 2014, 91479T, 91479T.

[18] R. Ohsawa, S. Sako, T. Miyata, T. Kamizuka, K. Okada, K. Mori, M. S. Uchiyama, J. Yamaguchi, T. Fujiyoshi, M. Morii, et al. "Slow-scanning" in Ground-based Mid-infrared Observations. In: *The Astrophysical Journal* 857.1 (2018), p. 37.

[19] A. R. Torres-Quijano, C. Packham, and S. F. Acosta. CanariCam Mid-infrared Drift Scanning: Improved Sensitivity and Spatial Resolution. In: *Publications of the Astronomical Society of the Pacific* 133.1029 (2021), p. 114501.

[20] C. G. Gonzalez, O. Absil, P.-A. Absil, M. Van Droogenbroeck, D. Mawet, and J. Surdej. Low-rank Plus Sparse Decomposition for Exoplanet Detection in Direct-imaging ADI Sequences-The LLSG Algorithm. In: *Astronomy & Astrophysics* 589 (2016), A54.

[21] J. Sauter, W. Brandner, J. Heidt, and F. Cantalloube. Detection Limits of Thermal-infrared Observations with Adaptive Optics. I. Observational Data. In: *Publications of the Astronomical Society of the Pacific* 136.9 (2024), p. 095001.

[22] ESO. *ESO - Overview — eso.org*. https://www.eso.org/sci/facilities/paranal/instruments/visir/overview.html. [Accessed 27-02-2025]. 2024.

[23] T. L. Herter, J. D. Adams, G. E. Gull, J. Schoenwald, L. D. Keller, B. E. Pirger, C. P. Henderson, G. J. Stacey, T. Nikola, J. M. De Buizer, W. D. Vacca, and K. Ennico. FORCAST: A Mid-Infrared Camera for SOFIA. In: *Journal of Astronomical Instrumentation* 7.4, 1840005-451 (2018), pp. 1840005–451.

[24] L. Bradley, B. Sipocz, T. Robitaille, E. Tollerud, C. Deil, Z. Vinícius, K. Barbary, H. M. Günther, A. Bostroem, M. Droettboom, et al. Photutils: Photometry Tools. In: *Astrophysics Source Code Library* (2016), ascl–1609.

[25] F. Lenzen. *Statistical regularization and denoising*. Leopold Franzens University, Innsbruck, Austria, 2006.

[26] L. Pueyo. Detection and Characterization of Exoplanets Using Projections on Karhunen–loeve Eigenimages: Forward Modeling. In: *The Astrophysical Journal* 824.2 (2016), p. 117.

[27] Z. Zhou, X. Li, J. Wright, E. J. Candes, and Y. Ma. Stable Principal Component Pursuit. In: *IEEE international symposium on information theory*. IEEE. 2010, pp. 1518–1522.

[28] E. J. Candes and J. Romberg. Sparsity and Incoherence in Compressive Sampling. In: *Inverse problems* 23.3 (2007), p. 969.

[29] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? In: *Journal of the ACM (JACM)* 58.3 (2011), pp. 1–37.

4

STRUCTURED DECOMPOSITIONS ON A TERABYTE SCALE

Motivation. *Imaging mass spectrometry (IMS) has become an important tool for molecular characterization of biological tissue. However, IMS experiments tend to yield large datasets, routinely recording over 200 000 ion intensity values per mass spectrum and more than 100 000 pixels, i.e., spectra, per dataset. Traditionally, IMS data size challenges have been addressed by feature selection or extraction, such as by peak picking and peak integration. Selective data reduction techniques such as peak picking only retain certain parts of a mass spectrum, and often these describe only medium-to-high-abundance species. Since lower-intensity peaks and, for example, near-isobar species are sometimes missed, selective methods can potentially bias downstream analysis towards a subset of species in the data rather than considering all species measured.*

Results. *We present an alternative to selective data reduction of IMS data that achieves similar data size reduction while better conserving the ion intensity profiles across all recorded m/z-bins, thereby preserving full spectrum information. Our method utilizes a low-rank matrix completion model combined with a randomized sparse-format-aware algorithm to approximate IMS datasets. This representation offers reduced dimensionality and a data footprint comparable to peak picking, but also captures complete spectral profiles, enabling comprehensive analysis and compression. We demonstrate improved preservation of lower signal-to-noise-ratio signals and near-isobars, mitigation of selection bias, and reduced information loss compared to current state-of-the art data reduction methods in IMS.*

The contents of this chapter are based on:

Moens, R. A., Migas, L. G., Van Ardenne, J. M., Skaar, E. P., Spraggins, J. M., & Van de Plas, R. (2025). Preserving Full Spectrum Information in Imaging Mass Spectrometry Data Reduction. *Bioinformatics, 41*(5), btaf247.

4.1. INTRODUCTION

Imaging mass spectrometry (IMS) is an analytical imaging technology that enables molecular mapping of complex biological samples, such as tissues, biofilms, or dispersed cells [1, 2, 3, 4, 5]. IMS combines the sensitivity and specificity of mass spectrometry with spatial information. It enables researchers to concurrently measure the distribution of hundreds to thousands of molecular species throughout tissue sections or other heterogeneous samples without the need for labeling target molecules [1, 2, 6, 7]. This capability holds strong potential for probing the lipidomic, glycomic, metabolomic, and proteomic content of biological samples across a wide range of applications, spanning from fundamental research in biology and medicine to the development of novel diagnostics and therapeutics [8, 9, 10].

However, the outstanding multiplexing capability of IMS-capable instruments generates vast amounts of data, often containing spatially resolved information for thousands of molecular species in a single experiment [1, 11, 12]. The volume and high-dimensionality of IMS data present significant challenges in data processing, analysis, and interpretation [12]. One of the primary challenges is data reduction, as raw IMS datasets typically consist of hundreds of thousands to millions of spatial locations, *i.e.,* pixels, each associated with a mass spectrum containing hundreds of thousands of ion intensities. These datasets contain a mixture of high- and low-intensity peaks as well as features with varying signal-to-noise ratios. Managing such large datasets requires effective data reduction techniques that extract meaningful information while minimizing computational burden and storage demands [13].

Current data reduction in IMS can be broadly categorized into acquisition-time and post-acquisition approaches (see Supplementary Materials of this chapter for a more elaborate overview). Acquisition-time methods reduce data during collection, typically resulting in a sparse representation rather than describing spectra in the full mass domain. Post-acquisition methods, often user-controlled, include peak picking, spectral integration, and spatial cropping [12, 14, 15]. These methods aim to further convert IMS data into more manageable representations [12, 13]. However, methods like peak picking can miss low-intensity, low signal-to-noise-ratio (*SNR*) peaks, and near-isobars, potentially introducing bias. Recent efforts have improved peak-picking accuracy [16], but challenges remain in handling near-isobaric species and low-intensity peaks. Here, we introduce a novel data model for IMS measurements that addresses missing values through sparse-format-aware low-rank matrix approximation. This approach offers an alternative to traditional data reduction methods for IMS, mitigating selection bias and minimizing information loss early in the analysis pipeline by preserving full spectrum information rather than only selective sub-windows of the measured mass range. Additionally, to handle the computational demands and large memory requirements of modern IMS datasets, we explore the use of randomization strategies to optimize low-rank factorization methods and implement obtaining such a representation of an IMS dataset.

4.2. METHODS

Consider a MALDI-TOF IMS dataset M ($\in \mathbb{R}^{m \times n}$), where m is the number of pixels/spectra, n is the number of m/z-bins recorded by the instrument, and M_{ij} is the ion intensity value associated with the i-th pixel and j-th m/z-bin. This dataset consists of real-valued ion intensities, where a row of M is a spectrum associated with a specific spatial location in the tissue and where a column of M is a particular m/z-bin considered across all pixels. The latter can be reconstructed into a so-called ion image, reporting the spatial distribution and abundance of a specific m/z-bin's intensities. For some IMS experiments, the intensity values M_{ij} are clipped by the instrument (*e.g.,* explicitly by acquisition-time data reduction or implicitly by the instrument's limit-of-detection), effectively not reporting intensity values below a certain relative ion count k. Ion intensity clipping is sometimes expressly performed to induce a sparse regime[1] on the recorded signals, often to save disk space. Regardless of the reason for clipping to occur, we want to explicitly deal with the missing values introduced by it. Therefore, we propose to model clipping as a function $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n}$,

$$f(M) = [f_{ij}(M)]_{m \times n} \tag{4.1a}$$

where $f_{ij}(M)$ is defined for each entry (i, j) as:

$$f_{ij}(M) = \begin{cases} M_{ij} & \text{if } M_{ij} \geq k \\ 0 & \text{if } M_{ij} < k \end{cases} \tag{4.1b}$$

for $i \in [1, m]$ and $j \in [1, n]$. The resulting (sparsified) dataset $f(M) \in \mathbb{R}^{m \times n}$ can be stored in a sparse matrix format, a data structure that only explicitly stores non-zero values and their locations in the matrix, leaving zero values to be implicitly represented without consuming memory. Most post-acquisition data reduction methods ignore the non-linear operator, $f(\cdot)$ (see Supplementary Materials). By applying a sampling operator $\mathscr{P}_\Omega(\cdot)$ to M, essentially a relaxation for $f(M)$, we acknowledge that there are missing values in M and avoid the assumption that those missing values are necessarily zeroes when modeling. The sampling operator also avoids that those missing values (potentially) negatively impact the model. Specifically [17], $\mathscr{P}_\Omega : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n}$ is defined as

$$[\mathscr{P}_\Omega(M)]_{ij} = \begin{cases} M_{ij} & \text{if } (i, j) \in \Omega \\ 0 & \text{if } (i, j) \notin \Omega \end{cases} \tag{4.2}$$

for $i \in [1, m]$, $j \in [1, n]$, and where Ω is the set of indices corresponding to the (known, reported) sampled values, as obtained from the instrument and after any acquisition-time data reduction. We denote $(i, j) \notin \Omega$ as the set Ω_c, making Ω and Ω_c complementary subsets of all entries in M. We can formulate the modeling of an IMS dataset M implicitly as a missing value problem, wherein a low-rank matrix approximation X of M is sought in the presence of missing data. The M-approximating matrix X can be considered an underlying model for the observed measurements in M, and the rank of X denotes the

[1] In this context, sparsity refers to the number of non-zero values in measurements, *i.e.,* high sparsity implies many zero values. A sparse regime implies that measurements contain many zero values.

dimension of the subspace containing the approximating matrix (see Supplementary Materials).

Since the proposed rank-optimization problem is non-convex, NP-hard, and thus difficult to calculate, we instead solve a convex relaxation of the problem (see Eq. 1 in the Supplementary Materials) using the singular value thresholding (SVT) algorithm [17, 18]:

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & ||X||_*, \\ \text{subject to} \quad & \mathscr{P}_\Omega(M) = \mathscr{P}_\Omega(X). \end{aligned} \tag{4.3}$$

This program is shown to exactly recover the solution of the original problem (see Eq. 1 in the Supplementary Materials) under specific conditions, *e.g.*, incoherence of bases and sampling distribution [19]. As proving a condition's validity is considered to be as hard as solving the original problem (see Eq. 1 in the Supplementary Materials), the conditions cannot be verified. Thus, we will assume that conditions are met. Nevertheless, we will discuss the sampling distribution assumption in the Case Study section of this chapter, as it is closely intertwined with the clipping mechanism during acquisition-time data reduction.

The SVT's advantage lies in utilizing matrices in sparse and low-rank format without requiring dense memory storage, crucial for large IMS datasets. However, the SVD's time complexity [20] remains a bottleneck for MALDI-TOF IMS datasets, leading to the adoption of a divide-factor-conquer approach to address this issue (see Supplementary Materials). Besides the sparse-format-aware SVT approach and its SVD-related modification to calculate our low-rank approximation of IMS data, we also explore a second method that has a similar solving program as in Eq. 4.3, but it relaxes the equality into an inequality constraint:

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & ||X||_*, \\ \text{subject to} \quad & ||\mathscr{P}_\Omega(M) - \mathscr{P}_\Omega(X)||_F^2 \leq \sigma. \end{aligned} \tag{4.4}$$

The latter enables the use of a fixed point continuation (FPC) algorithm for solving instead [17, 21]. This second program accounts, in addition to missing values, for low-intensity dense noise (*e.g.*, Gaussian noise) in the measurements, which is also inherently present in IMS data. The disadvantage of this algorithm is that it requires a dense-format matrix of similar size as M to be stored in memory (for the MALDI-TOF IMS dataset in this chapter, this amounts to 1 649.769 GB). As such, it is clear that whether SVT or FPC are the better choice for solving these optimization problems depends on the resources available and the needs of the subsequent analysis. Finally, note that both SVT and FPC's practical implementations have a δ and τ parameter that arise as part of their solving algorithms. These hyperparameters require *a priori* setting (or optimization). We specify their setting for each experiment in the Supplementary Materials.

Furthermore, to deal with both SVD complexity and memory load, we make use of the divide-factor-conquer approach (DFC) [22]. It consists of three steps and provides a framework that we can apply to both the SVT and FPC algorithms for obtaining an approximation $\bar{X}$ of matrix M with completion for the complete dataset M. Its specifics are provided in the Supplementary Materials.

4.3. CASE STUDIES

We demonstrate the method's applicability across different but relatively common instrumental platforms for IMS. Although these case studies focus on specific datasets, the algorithms have not been customized to any particular instrumental setup or IMS dataset type, suggesting that the basic approach could be useful in other types of IMS experiments as well. In a first case study, we establish that an IMS dataset representation using a low-rank matrix factorization approach can outperform an equally small IMS dataset representation using traditional peak picking in a no-missing value case. We demonstrate this on Fourier-transform ion cyclotron resonance (FT-ICR) imaging mass spectrometry data (**Fig. 4.1**). In the second case study, we investigate the reconstruction error (*i.e.*,

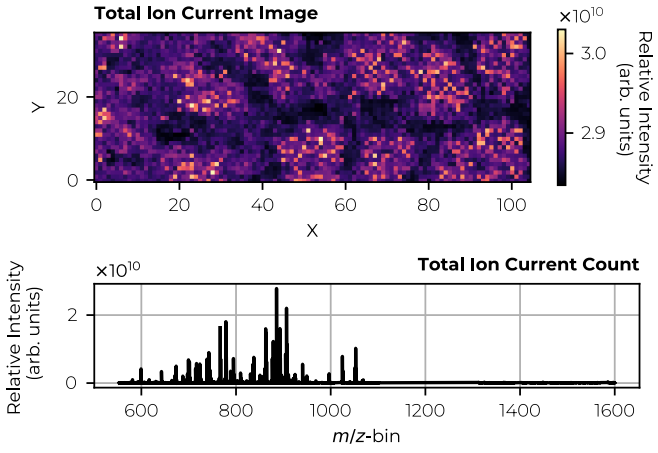

Figure 4.1: MALDI FT-ICR IMS measurement of human kidney tissue. The experiment was conducted using a 15T Bruker MALDI FT-ICR mass spectrometer (Bruker Daltonics, Billerica, MA, USA) with 50-μm pixel size, covering the m/z range from 552 to 1 600 in negative ionization mode. For further sample preparation specifics, see the Supplementary Materials. The raw data were exported to a custom file format and normalized using 5-95%-TIC. The dataset contains 3 780 spectra, each consisting of 1 372 421 m/z-bins. For further data preprocessing specifics, see the Supplementary Materials. The top panel shows the spatial distribution, represented as a total ion current image (*i.e.*, the summation of the normalized spectral axis). The bottom panel displays the average mass spectrum.

on sampled/known values, $\in \Omega$), the imputation error (*i.e.*, on missing values, $\in \Omega_c$) and the global error (*i.e.*, on all entries, $\in (\Omega \cup \Omega_c)$) on the same FT-ICR IMS dataset as in the first case study (**Fig. 4.1**). To mimic missing entries in the FT-ICR data, we implement two sampling schemes. The goal of the third case study is to evaluate the methodology, specifically the SVT and FPC algorithms with the divide-factor-conquer approach, directly on TOF IMS data that inherently includes missing values (**Fig. 4.2**). This dataset consists of 312 249 m/z-bins for 1 320 876 spectra (1.65 TB in dense matrix format). The evaluation is both quantitative, using an error score and compression factor, and qualitative, with a focus on visualizing advantages and limitations that are relevant to analytical chemists. The data preprocessing can be found in the Supplementary Materials.

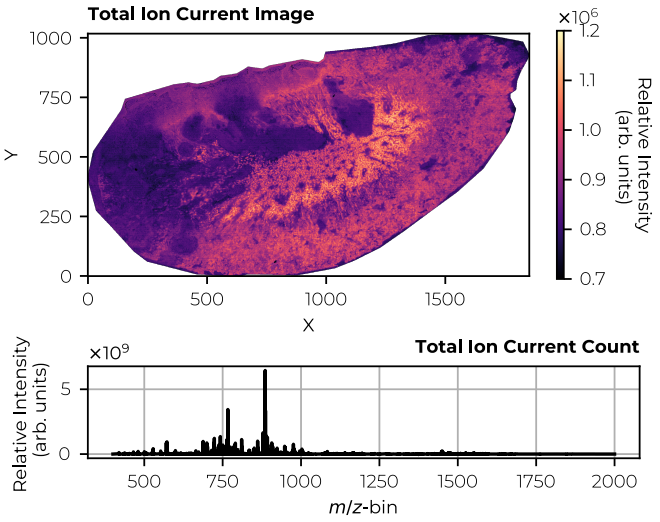

Figure 4.2: MALDI qTOF IMS measurement of *Staphylococcus aureus*-infected mouse kidney tissue. The infection-induced abscesses are visible as dark areas in the total ion current image. The experiment was performed using a Bruker timsToF Flex mass spectrometer (Bruker Daltonics, Billerica, MA, USA) with 5-μm pixel size, covering *m/z* range from 400 to 2 000 in negative ionization mode. For further sample preparation specifics, see the Supplementary Materials. The raw data were exported to a custom file format and normalized using 5-95%-TIC. The dataset contains 1 320 876 spectra, each consisting of 312 249 *m/z*-bins. For further data preprocessing specifics, see the Supplementary Materials. The top panel shows the spatial distribution of the total ion current image. The bottom panel displays the average spectrum.

4.3.1. CASE STUDY 1: LOW-RANK MATRIX FACTORIZATION OUTPERFORMS TRADITIONAL PEAK PICKING

We first demonstrate that, in addition to retaining full spectrum information, a low-rank matrix approximation can achieve a lower reconstruction/global error compared to peak picking in a no-missing value case. Having established this baseline, we can then expand our problem setting with missing values in the second case study.

The best rank-k approximation with respect to the Frobenius norm (a measure we will use throughout this chapter) is given by the truncated singular value decomposition (SVD) [23, 24]. Since peak picking can be viewed as a form of (low-rank) matrix approximation by selecting specific columns from a dataset (with a column representing a selected peak), we can assert that peak picking is, at best, as effective as the truncated SVD. In **Table 4.1**, we observe a 39.1% difference in reconstruction error (which is equal to the global error in the absence of missing values) between the truncated SVD (factorization) and peak picking (100 peaks, see Supplementary Materials for extended numbers of picked peaks). Even if we compare the reconstruction error for a similar data footprint, this still amounts to a difference of 36.8%. While a reconstruction score can be a rather abstract form of capturing full spectrum information content, we highlight in **Fig. 4.3** a concrete difference between the raw IMS data, a low-rank representation (rank-100), and a conventional peak-picked representation. This example demonstrates that not only is the overall ion

intensity profile preserved in the factorization representation, but also a low-abundant peak at *m/z* 778.524 is effectively retained, where the peak-picked representation misses this peak entirely. Both the metric used in **Table 4.1** and the example of missing low-abundance peaks in **Fig. 4.3** illustrate that low-rank matrix factorization can outperform traditional peak picking when it comes to IMS dimensionality reduction. However, while SVD is a strong factorization method, it may not always be optimal, *e.g.*, when dealing with missing values in the data.

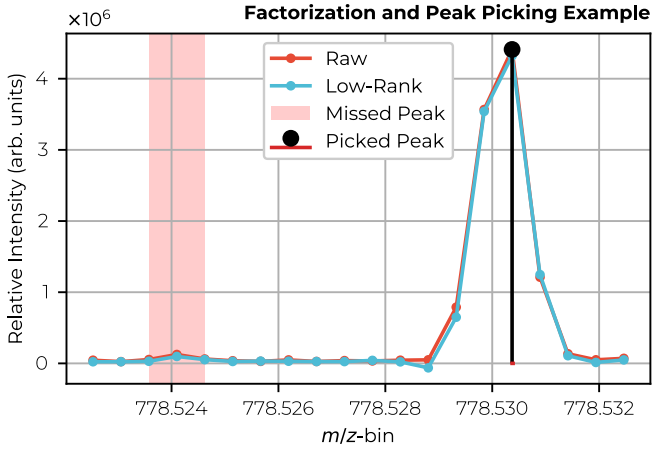

Figure 4.3: MALDI FT-ICR spectrum of a particular pixel showing the raw, low-rank and peak-picked data in a particular *m/z*-window of the mass spectrum. The area, highlighted in red, shows a low-abundant peak missed by the peak-picking representation, yet accurately captured and reconstructed by the low-rank factorization representation (see Supplementary Materials for the corresponding raw and imputed ion images). This is achieved at identical compression ratios for both representations. If low-abundant peaks are of interest, factorization-based representations are probably better suited to reduce the dimensionality of IMS datasets.

4.3.2. CASE STUDY 2: RECONSTRUCTION AND IMPUTATION QUALITY WHEN DEALING WITH MISSING VALUES

Having established that a factorization approach is favourable over peak picking in a no-missing value situation, our factorization approach is now evaluated in a missing value scenario. For this case study, we therefore implement two sampling schemes to mimic missing values in IMS data:

(α) Selects the top 8.9% of intensity values (to establish the in-sampling set Ω, *i.e.*, the known values), with all other (lower) intensity values removed (making up the out-of-sampling set Ω_c, *i.e.*, missing values),

(β) Selects 8.9% of entries, not based on intensity but, uniformly at random (in-sampling set Ω, *i.e.*, the known values), with all other values removed (out-of-sampling set Ω_c, *i.e.*, missing values).

Scheme α mimics commonly employed IMS acquisition-time data reduction, while scheme β examines discrepancies related to incoherence conditions imposed by most

Table 4.1: **Comparison of truncated SVD and peak picking results**. The reconstruction error is used throughout this chapter as a metric to measure how well (full) spectrum information is captured. A larger error implies that more information is lost. Hence, a low error is desired. However, note that an error of 0% is (probably) not desired as the data contains noise and it would be desirable to filter off this noise, leading to a (small) error. From this table, we observe that a factorization approach, the truncated SVD, leads to a substantial decrease in reconstruction error (up to 39.1%) compared to peak picking. The truncated SVD is carried out by truncating an SVD performed by the GESDD-routine. Peak picking is performed by matching (1) the rank and (2) the data footprint, *i.e.*, MBs on disk. The raw data in dense matrix format has a storage footprint of 20.751 GB.

Method	**Rank**	**Reconstruction Error** $\frac{\|\|M-X\|\|_F}{\|\|M\|\|_F} \times 100\%$	**Compression Factor w.r.t. Dense Format**	**Dense Data Footprint**
Raw	–	0	–	20.7510 GB
Peak Picking (100 peaks)	100	60.5	13 724	0.0015 GB
Peak Picking (123 peaks)	123	58.2	11 158	0.0019 GB
Truncated SVD	100	21.4	11 158	0.0019 GB

Table 4.2: **Comparison of SVT and FPC results, with threshold sampling scheme** α. A low reconstruction error is observed for all methods for both raw and low-rank inputs and references, comparable to the no-missing values case. The imputation error is more substantial. However, since these consist mostly of low-intensity values (caused by the clipping operator), their impact is small on the global error. For SVT, we set parameters $\delta = 1$ and $\tau = 10^{-3}$ and for FPC, we set $\delta = 1.4$ and $\tau = 10^{-3}$ (see the Supplementary Materials). Peak picking is performed by picking the 100 highest peaks of the total ion current count of the raw data. For SVT with raw input data, we obtain a 171 rank solution, and for FPC, a 271 rank solution. For SVT with low-rank input data, we obtain a 131 rank solution, and for FPC, a 111 rank solution. We truncate all solutions to a rank of 100 for fair comparison. The SVT took on average 64.74 minutes to converge, while the FPC algorithm took on average 41.89 minutes.

Input M	**Reference** $\tilde{M}$	**Method**	**Rank**	**Reconstruction Error** $\frac{\|\|P_\Omega(\tilde{M}-X)\|\|_F}{\|\|P_\Omega(\tilde{M})\|\|_F} \times 100\%$	**Imputation Error** $\frac{\|\|P_{\Omega_c}(\tilde{M}-X)\|\|_F}{\|\|P_{\Omega_c}(\tilde{M})\|\|_F} \times 100\%$	**Global Error** $\frac{\|\|\tilde{M}-X\|\|_F}{\|\|\tilde{M}\|\|_F} \times 100\%$
Raw	*Raw*	*Peak Picking*	100	–	–	60.5
Raw	Raw	SVT	100	26.6	66.7	34.7
Raw	Raw	FPC	100	13.7	59.2	25.0
Raw	Low-Rank	SVT	100	26.5	51.5	31.4
Raw	Low-Rank	FPC	100	6.2	33.2	13.8
Low-Rank	Low-Rank	SVT	100	4.9	93.1	34.8
Low-Rank	Low-Rank	FPC	100	2.6	83.4	31.0

matrix completion algorithms [17, 19]. The 8.9% sampling rate was chosen for IMS fidelity, matching the real-world TOF IMS dataset sampling rate in the third case study. **Table 4.2** presents error scores for both the SVT and FPC algorithms (without the divide-factor-conquer approach) using sampling scheme α on the FT-ICR IMS dataset. It reports:

- Reconstruction error: modeling error for known entries.
- Imputation error: modeling error for missing values.
- Global error: modeling error for both known and missing entries.

Error scores were calculated with respect to both raw data and its low-rank approximation. Therefore, the input matrix is defined as the matrix used as input to our algorithms (Eq. 4.3, 4.4). The reference matrix is defined as the matrix used as reference in the error scores. We considered two types of input (M) and reference ($\tilde{M}$) matrices:

- Raw: A dataset with missing values sampled directly from the raw data (thus including high-rank noise variation),
- Low-rank: A dataset with missing values sampled from a low-rank version of raw data, obtained through truncated singular value decomposition with rank 100.

Using the low-rank approximation as input (M) and reference ($\tilde{M}$) ensures that the low-rank conditions imposed by SVT and FPC are met, reducing unwanted (often noisy) variation from impacting the evaluation process.

Finally, note that the presented results stem from single experiments influenced by various factors (*e.g.*, tissue type, sample preparation, detector type, raw data structure, noise levels, algorithmic parameters). Consequently, they are only evaluated relative to each other.

RECONSTRUCTION, IMPUTATION AND GLOBAL ERROR

As shown in **Table 4.2**, both methods exhibit relatively low *reconstruction error* for non-missing values (between 2.6 and 26.6%) across all input and reference matrices. This is comparable/a slight improvement with respect to the results found in the first case study. Generally, FPC outperforms SVT on the raw input matrix, which is expected due to FPC's ability to filter out small dense noise. On the other hand, both methods show only moderate performance on *imputation error* for missing values (between 33.2 and 93.1%) across all input and reference combinations, with FPC showing a slight advantage. This trend, along with similar results from the uniform sampling scheme β (see the Supplementary Materials), suggests that while a low-rank factorization representation captures the non-missing value entries effectively, its performance as a predictor for missing values is limited and further investigation is needed to better understand the underlying causes.

Interestingly, contrary to expectations, the uniform sampling scheme β does not outperform the threshold-based sampling α, despite its closer alignment with incoherence conditions. This highlights the need to carefully consider the implications of different sampling strategies. Threshold sampling scheme α removes low-intensity entries, primarily associated with noise. Hence, it requires the imputation of noisy features by a low-rank model. However, this scheme generally fails to satisfy incoherence conditions, leading to poor imputation error in general. The poor performance could, for example, be caused by the spatial correlation of low-abundance values, which is particularly evident in specific m/z-bins and distinct (positive) spatial areas across the tissue. As illustrated in **Fig. 4.4**, under an intensity-magnitude driven sampling scheme α, a high-intensity m/z-bin at 885.571 is missing only a few values (white entries), while a low-intensity m/z-bin at 756.254 can be missing many values.

In contrast, a uniform sampling scheme β is expected to perform better because it more closely aligns with incoherence conditions and treats all ion species the same. However, we observe worse reconstruction and imputation errors compared to scheme α. This is probably related to (1) the low 8.9% sampling rate (*i.e.*, 91.1% of all intensity entries are missing in this dataset and there is relatively little signal to model with), and (2) the predominant number of low signal-to-noise m/z-bins in the raw data. Consequently, uniform sampling leads to a significant loss in high-valued, "informative" entries. We

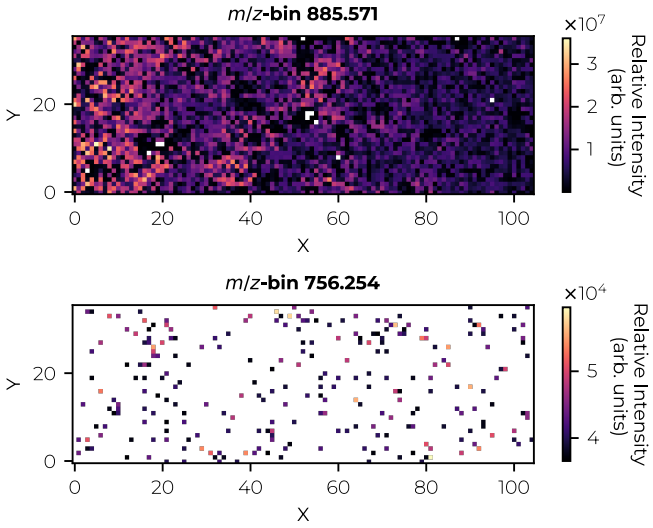

Figure 4.4: MALDI FT-ICR ion images from a single *m/z*-bin of the low-rank input matrix with threshold sampling scheme α. The top image depicts an *m/z*-bin with high intensity and thus, with the scheme α, a low number of missing values. The bottom image depicts an *m/z*-bin with low intensity and thus, with scheme α, a high number of missing values.

expect that the imputation error for sampling procedure α and both reconstruction and imputation error for sampling procedure β could benefit from advanced feature scaling. Additionally, introducing chemical noise (*e.g.*, speckle noise) into our data before applying the sampling scheme could cause high-intensity values to become missing, albeit at a lower rate, which may further reduce the imputation error.

Nevertheless, both methods achieve a *global error* that is 25.7 to 46.7% lower than that of peak picking, which has a global error of 60.5%, representing a substantial improvement in terms of full spectrum information, even in the presence of missing values. Moreover, note that the imputation error does not significantly influence the global error score since low-intensity features (at least 100-fold smaller than the base peak) contribute less to the global error—due to the properties of the Frobenius norm and because the sampling scheme α retains high-intensity values (see Supplementary Materials for ion image examples). Overall, this result implies that full spectrum information is better captured by the proposed low-rank factorization methodology than by peak picking, both when missing values are present as well as when they are absent (see Supplementary Materials for k-means clustering comparison).

4.3.3. CASE STUDY 3: ADVANTAGES AND DISADVANTAGES OF LOW-RANK MATRIX COMPLETION FOR MISSING VALUE TOF IMS DATA

In this case study, we forgo synthetically generated missing values, for which the ground truth is known, and apply our approach on IMS data with intrinsic missing values.

Table 4.3: **Comparison of SVT and FPC results, with randomized projection divide-factor-conquer approach.** A substantial improvement for the reconstruction error is observed for both SVT and FPC in comparison to peak picking, while compression factors and data footprint are equal. For SVT, we set parameters $\delta = 1.7$ and $\tau = .5$ and for FPC, we set $\delta = 1$ and $\tau = 1.5 \times 10^{-2}$ (see the Supplementary Materials). Peak picking matches the data footprint, *i.e.*, MBs on disk. The raw data in dense matrix format has a storage footprint of 1 649.769 GB, and storing it in a sparse matrix format (*e.g.*, compressed sparse column) amounts to 279.648 GB. The full process of dividing, factoring and combining took around 12 hours for the SVT and around 8 hours for the FPC.

Method	Rank	Reconstruction Error $\frac{\|P_\Omega(\tilde{M}-X)\|_F}{\|P_\Omega(\tilde{M})\|_F} \times 100\%$	Compression Factor w.r.t. Dense Format	Compression Factor w.r.t. Sparse Format	Dense Data Footprint
Raw	–	0	–	–	1 649.769 GB
Peak Picking (123 peaks)	–	52.0	2 525	642	0.653 GB
SVT	100	21.6	2 525	642	0.653 GB
Raw	–	0	–	–	1 649.769 GB
Peak Picking (129 peaks)	–	51.4	2 405	611	0.686 GB
FPC	105	23.6	2 405	611	0.686 GB

RECONSTRUCTION ERROR AND COMPRESSION FACTOR

The performance metrics for this third case study, including the reconstruction error and compression factors, are summarized in **Table 4.3**. They highlight that both SVT and FPC exhibit comparable performance. While a reconstruction error of approximately 20% might seem substantial at first, it actually represents a 30% reduction in information loss compared to traditional peak picking, all while maintaining the same data footprint and enabling full profile analysis in downstream workflows. Furthermore, all methods demonstrate high compression factors, with the IMS representation's footprint being approximately 2 500 times smaller compared to a dense matrix format and 600 times smaller than a sparse matrix format, comparable to those achieved by peak picking.

SPECTRAL ERROR DISTRIBUTION AND BIOLOGICAL INTERPRETATION

We further investigate the distribution of reconstruction errors for individual spectra, referred to as the spectral error score (see **Fig. S4.14.a** and **Fig. S4.14.c** in Supplementary Materials). This score is calculated both for

(*a*) the 100 *m*/*z*-bins with the largest total ion current count across the dataset; and

(*b*) the largest 100 *m*/*z*-bins per spectrum, *i.e.*, the top peaks in each individual spectrum.

The distributions of spectral error scores reveal patterns that correlate with biology for both SVT and FPC methods under both scoring criteria (*a* and *b*). Interestingly, the spectral error is slightly lower for the largest individual spectrum peaks (*b*), as the top dataset-wide peaks (*a*) might not be present in every spectrum. The error distributions appear to be a mixture of two Gaussian-like distributions with different means and standard deviations. Spatial reconstruction of these distributions (**Fig. S4.14.b** and **Fig. S4.14.d** in the Supplementary Materials) reveal distinct tissue regions that may correlate with the total ion current count (**Fig. 4.2**). Moreover, no clear relationship is observed between these distributions and the number of non-zero values per spectrum (see **Fig. S4.18** in the Supplementary Materials). This suggests significant heterogeneity in molecular distributions within the tissue, rather than issues related to incoherence, might

be influencing the reconstruction quality, especially in *Staphylococcus aureus*-infected regions.

METHODOLOGICAL EFFECTS ON RECONSTRUCTED ION IMAGES AND SPECTRA

Although high-intensity ion images and peaks are recovered well (see reconstruction error and spectral error score, and the Supplementary Materials for ion image examples), distortions may occur in the reconstructed spectra and individual ion images of very low intensity (**Fig. S4.15** in Supplementary Materials). Commonly observed distortions included (1) small peak shifts, *i.e.,* shifting of peak distribution along the *m/z* axis, (2) peak widening, *i.e.,* smearing peaks over larger *m/z* ranges, and (3) peak prediction, *i.e.,* imputation of peaks not present in the raw spectrum, but predicted on the basis of dataset-wide observed patterns. It should be noted that these effects are not necessarily incorrect, that they may arise from genuine corrections for small non-linear misalignments due to instrumentation or noise, or from other instrumental artifacts.

Notably, these distortions are more prominent in low-intensity peaks (mostly around and below 10^3 relative intensity in peak height), which is consistent with the optimization process focused on minimizing the Frobenius norm. This introduces (4) a recovery bias that favours better reconstruction of high-intensity peaks, as also observed in this TOF IMS dataset. From a spatial perspective, caution is warranted when interpreting very sparse ion images as biologically meaningful. For example, the predicted ion image of *m/z* 1 284.10, is based on only very few measurements (see raw ion image of *m/z* 1 284.10, **Fig. S4.15** in the Supplementary Materials). These low-abundant species-centric effects, whether desired or undesired, can potentially be mitigated in the future through advanced feature scaling and an improved model. They should also always be considered within the context of peak picking approaches, which often leave no record of low-abundant species to begin with.

PRESERVATION OF NEAR-ISOBARIC SPECIES

Near-isobaric species, molecular species with nearly identical mass-to-charge ratios but different chemical compositions, pose significant challenges for accurate peak detection. These species are often overlooked in peak picking due to their low intensity relative to dominant peaks, or may be incorrectly integrated as a single species. In **Fig. S4.16** in the Supplementary Materials, we present an example of such a near-isobaric species, at approximately *m/z* 725.53 (blue area), located close to a dominant species at *m/z* 725.51 (orange area). Due to their proximity, near-isobaric species are frequently neglected. Integrating the orange and blue areas separately, reveals different spatial molecular distributions, indicating that these *m/z*-ranges correspond to distinct molecular species. When applying peak picking, the best-case scenario consists of integrating the orange area and neglecting the blue area, potentially missing the near-isobaric species. In the worst-case scenario, both areas are integrated as one, and the dominant peak's intensity overshadows that of the near-isobaric species, leading to the loss of unique spatial information. In both cases, the unique near-isobaric information is lost. However, our methods successfully preserve this information by approximating the full spectrum without requiring prior specification of their positions on the *m/z*-axis. This ensures that near-isobaric species are more accurately captured and represented, maintaining the unique spatial and molecular information they provide. Preserving near-isobaric species is an important factor in

improving analysis specificity. Specificity on an instrumental and/or (bio)chemical level is an important driver in IMS [11], being able to maintain it in the analysis is thus of utter importance.

RETENTION OF LOWER-INTENSITY ION SPECIES AND BIAS MITIGATION

Our approach effectively mitigates bias by retaining lower-intensity ion species that are commonly disregarded by peak picking, especially when only the largest peaks are retained. We identified several *m/z*-bins representing peaks corresponding to biologically relevant lipids and adducts, which were preserved in our analysis despite their low intensity (**Fig. S4.17** in Supplementary Materials). The specific *m/z* values include:

- The lipid LPE 18:1 at *m/z* 478.29 (confirmed by liquid chromatography mass spectrometry),
- A 4-(dimethylamino)cinnamic acid (DMACA, see Suppl. Materials) adduct of PE O-(36:3) at *m/z* 917.54,
- [CL(77:2)+Na-2H]- at *m/z* 1552.12.

These *m/z*-bins are not isotopic peaks and thus provide unique information about species abundant in different spatial regions in the tissue. These examples are only a few from a large group of peaks (1000+) that are preserved for this TOF IMS dataset. Retaining low signal-to-noise ratio signals enhances the detection of species in downstream analyses, reducing confirmation bias by reporting on nearly all instrument-detected peaks rather than narrowing the analysis pre-maturely to a set of high-abundant species. Retention of lower-intensity ion species in the computational representation and analysis is an important factor in maintaining sensitivity throughout the chain from sample preparation to instrument to computational analysis to biological insight.

4.4. CONCLUSIONS

This chapter explored the application of matrix factorization algorithms on IMS data, focusing on the goal of dimensionality and data footprint reduction, addressing the issue of missing values, and evaluating both quantitative and qualitative outcomes. For a no-missing value case, a low-rank factorization-based representation of IMS data improved the reconstruction error by 39.1% over peak picking while concurrently maintaining a full spectrum profile for all spectra in the dataset. In the missing value case, we achieved low reconstruction errors for both SVT and FPC based approaches, comparable to the no-missing case. We also highlighted the persistent challenge of reducing imputation errors, which could potentially be mitigated through advanced feature scaling that accounts for the specific characteristics of IMS data and an improved data model. For the missing value case, we demonstrated a substantial reduction in full spectrum information loss (global error) up to 40% compared to traditional peak picking methods, while achieving compression factors similar to peak picking. Our experiments revealed that matrix completion algorithms offer significant advantages in maintaining sensitivity by preserving lower-signal-to-noise ratio signals and mitigating selection bias. At the same time, we demonstrated the preservation of specificity by retention of near-isobaric species in the

analysis through our full profile approach. These improvements are expected to enhance downstream analysis by providing a richer, more complete reduced representation of IMS data while also providing dimensionality reduction capabilities comparable to traditional peak picking. The importance of this research lies in the introduction of a framework for IMS data reduction by factorization in an early stage and with awareness of missing values. This framework enables high compression rates, up to 2 500-fold compared to dense matrix storage formats and up to 600-fold compared to sparse matrix storage formats, while preserving substantially more full profile information than peak picking. We emphasize the importance of utilizing full spectra in downstream analysis to avoid premature or biased information loss, as often occurs with peak integration or peak picking. However, our methods also have limitations, such as peak shifting and widening, low-intensity peak prediction, and the prediction of very sparse ion images.

Looking forward, future work could focus on exploring on-the-fly low-rank approximation schemes that can be employed during data acquisition to enhance accuracy and reduce computational burden. Additionally, it will be important to incorporate considerations for non-negativity, measurement sparsity, and uncertainty. Addressing issues related to peak shifting, widening, and normalization also emerges as a critical area for further research.

In conclusion, our study shows that low-rank factorization-based representations of IMS data can substantially advance the field by reducing full spectrum information loss by 30 to 40% compared to traditional peak picking methods. This work highlights the potential of matrix factorization and, in particular, completion algorithms for avoiding premature feature selection and for lifting IMS data analysis to the full profile level.

REFERENCES

[1] R. M. Caprioli, T. B. Farmer, and J. Gile. Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. In: *Analytical Chemistry* 69.23 (1997), pp. 4751–4760.

[2] L. A. McDonnell and R. M. Heeren. Imaging Mass Spectrometry. In: *Mass Spectrometry Reviews* 26.4 (2007), pp. 606–643.

[3] A. B. Esselman, N. H. Patterson, L. G. Migas, M. Dufresne, K. V. Djambazova, M. E. Colley, R. Van de Plas, and J. M. Spraggins. Microscopy-directed Imaging Mass Spectrometry for Rapid High Spatial Resolution Molecular Imaging of Glomeruli. In: *Journal of the American Society for Mass Spectrometry* 34.7 (2023), pp. 1305–1314.

[4] W. J. Perry, C. M. Grunenwald, R. Van de Plas, J. C. Witten, D. R. Martin, S. S. Apte, J. E. Cassat, G. B. Pettersson, R. M. Caprioli, E. P. Skaar, et al. Visualizing Staphylococcus Aureus Pathogenic Membrane Modification Within the Host Infection Environment By Multimodal Imaging Mass Spectrometry. In: *Cell Chemical Biology* (2022).

[5] T. Bien, K. Koerfer, J. Schwenzfeier, K. Dreisewerd, and J. Soltwisch. Mass Spectrometry Imaging to Explore Molecular Heterogeneity in Cell Culture. In: *Proceedings of the National Academy of Sciences* 119.29 (2022), e2114365119.

[6] A. R. Buchberger, K. DeLaney, J. Johnson, and L. Li. Mass Spectrometry Imaging: a Review of Emerging Advancements and Future Insights. In: *Analytical Chemistry* 90.1 (2018), p. 240.

[7] M. Aichler and A. Walch. MALDI Imaging Mass Spectrometry: Current Frontiers and Perspectives in Pathology Research and Practice. In: *Laboratory Investigation* 95.4 (2015), pp. 422–431.

[8] S. S. Rubakhin, J. C. Jurchen, E. B. Monroe, and J. V. Sweedler. Imaging Mass Spectrometry: Fundamentals and Applications to Drug Discovery. In: *Drug Discovery Today* 10.12 (2005), pp. 823–837.

[9] P.-M. Vaysse, R. M. Heeren, T. Porta, and B. Balluff. Mass Spectrometry Imaging for Clinical Research–latest Developments, Applications, and Current Limitations. In: *Analyst* 142.15 (2017), pp. 2690–2712.

[10] S. Kaspar, M. Peukert, A. Svatos, A. Matros, and H.-P. Mock. MALDI-imaging Mass Spectrometry–an Emerging Technique in Plant Biology. In: *Proteomics* 11.9 (2011), pp. 1840–1850.

[11] J. M. Spraggins, K. V. Djambazova, E. S. Rivera, L. G. Migas, E. K. Neumann, A. Fuetterer, J. Suetering, N. Goedecke, A. Ly, R. Van de Plas, et al. High-performance Molecular Imaging with MALDI Trapped Ion-mobility Time-of-flight (TimsTOF) Mass Spectrometry. In: *Analytical Chemistry* 91.22 (2019), pp. 14552–14560.

[12] T. Alexandrov. MALDI Imaging Mass Spectrometry: Statistical Data Analysis and Current Computational Challenges. In: *BMC Bioinformatics* 13.Suppl 16 (2012), S11.

[13] N. Verbeeck, R. M. Caprioli, and R. Van de Plas. Unsupervised Machine Learning for Exploratory Data Analysis in Imaging Mass Spectrometry. In: *Mass Spectrometry Reviews* 39.3 (2020), pp. 245–291.

[14] D. M. Anderson, R. Van de Plas, K. L. Rose, S. Hill, K. L. Schey, A. C. Solga, D. H. Gutmann, and R. M. Caprioli. 3-D Imaging Mass Spectrometry of Protein Distributions in Mouse Neurofibromatosis 1 (NF1)-associated Optic Glioma. In: *Journal of Proteomics* 149 (2016), pp. 77–84.

[15] P. Monchamp, L. Andrade-Cetto, J. Y. Zhang, and R. Henson. Signal Processing Methods for Mass Spectrometry. In: *Systems Bioinformatics: An Engineering Case-Based Approach, Artech House Publishers* (2007).

[16] A. González-Fernández, A. Dexter, C. J. Nikula, and J. Bunch. NECTAR: A New Algorithm for Characterizing and Correcting Noise in QToF-Mass Spectrometry Imaging Data. In: *Journal of the American Society for Mass Spectrometry* 34.11 (2023), pp. 2443–2453.

[17] E. J. Candes and Y. Plan. Matrix Completion with Noise. In: *Proceedings of the IEEE* 98.6 (2010), pp. 925–936.

[18] J.-F. Cai, E. J. Candès, and Z. Shen. A Singular Value Thresholding Algorithm for Matrix Completion. In: *SIAM Journal on Optimization* 20.4 (2010), pp. 1956–1982.

[19] E. J. Candes and B. Recht. Exact Matrix Completion Via Convex Optimization. In: *Communications of the ACM* 55.6 (2012), pp. 111–119.

[20] J. Dongarra, M. Gates, A. Haidar, J. Kurzak, P. Luszczek, S. Tomov, and I. Yamazaki. The Singular Value Decomposition: Anatomy of Optimizing an Algorithm for Extreme Scale. In: *SIAM Review* 60.4 (2018), pp. 808–865.

[21] S. Ma, D. Goldfarb, and L. Chen. Fixed Point and Bregman Iterative Methods for Matrix Rank Minimization. In: *Mathematical Programming* 128.1 (2011), pp. 321–353.

[22] L. W. Mackey, A. Talwalkar, and M. I. Jordan. Distributed Matrix Completion and Robust Factorization. In: *Journal of Machine Learning Research* 16.1 (2015), pp. 913–960.

[23] C. Eckart and G. Young. The Approximation of One Matrix By Another of Lower Rank. In: *Psychometrika* 1.3 (1936), pp. 211–218.

[24] L. Mirsky. Symmetric Gauge Functions and Unitarily Invariant Norms. In: *The Quarterly Journal of Mathematics* 11.1 (1960), pp. 50–59.

SUPPLEMENTARY MATERIALS

INTRODUCTION: CURRENT DATA REDUCTION

Current data reduction approaches can be broadly classified into acquisition-time and post-acquisition data reduction methods. Acquisition-time data reduction involves reducing data during acquisition, typically resulting in a selective representation rather than describing full spectra. For instance, time-of-flight (TOF) mass spectrometers may retain only ion intensity values above a certain threshold, discarding other measured *m/z*-bins. These representations, while reduced, are often not well-suited for downstream analysis due to their still relatively large size, unstructured feature selection, and sparse-matrix-format storage, commonly requiring reconstruction to the full mass domain (*i.e.*, intensity values for all *m/z*-bins) and subsequent data re-reduction. Moreover, this process is usually "hard-coded" into the instrument and not under user control.

Post-acquisition data reduction is user-controlled and includes techniques such as peak picking (*i.e.,* selecting a subset of features or masses), spectral integration (*i.e.,* combining features over small mass ranges), and spatial cropping (*i.e.,* selecting a spatial subset of interesting pixels and spectra) [1, 2, 3]. Note that ion intensity integration sometimes already happens at the detector and/or instrument level, for example, to counteract space-charge effects. These post-acquisition methods usually aim to convert full spectrum IMS data, that is IMS data with ion intensity values for each *m/z*-bin in the measured mass range, into a more manageable representation reporting only the signal features deemed important, often certain peaks, and taking up a reduced memory footprint such that it is practical for subsequent analysis [1, 4]. However, methods like peak picking can miss or inadequately capture certain peaks in the full spectra. These missed signals often include low-intensity or low *SNR* peaks and near-isobars (*i.e.*, peaks with nearly identical *m/z* but representing different molecular species, and sometimes presenting as 'shoulders' in a peak profile). This is often due to their focus on the abundance of signals rather than considering the structured signal presence or absence across measurements. Furthermore, their selective nature can introduce bias by limiting downstream analysis to a subset of (often more abundant) molecular species rather than considering all measured species. Although recent efforts have been made to improve peak-picking accuracy and robustness [5], challenges persist, particularly in handling near-isobaric species and low-intensity peaks.

METHODS

NON-LINEAR OPERATOR

Most post-acquisition data reduction methods ignore the non-linear operator, $f(\cdot)$, and treat the dataset $f(M)$ as if it was M by assuming zero ion intensity for entries where $M_{ij} < k$. This is, for example, common in most peak picking algorithms, as they often require a full spectrum profile with an intensity value for each *m/z*-bin to determine

where peaks are located. This approach may lead to information loss by assuming zero intensity where the abundance was low but not zero, and it can ultimately lead to biased feature subset selection, overemphasizing the importance of medium-to-high-abundant molecular species and underrepresenting or ignoring low-abundant species. In contrast, the methods presented here do not ignore the non-linear clipping operator, but rather seek to take it explicitly into account and avoid some of the assumptions listed above. Concretely, we propose to model the non-linear clipping function and describe it as a sampling operator.

LOW-RANK APPROACH

Low-rank matrices have several favourable mathematical properties in the context of underdetermined systems of equations and are oftentimes used to describe data from the smallest possible set of basis vectors, *i.e.*, "the simplest representation for the given measurements". We formulate the modeling of IMS data, *i.e.*, the missing value problem, as an optimization problem that seeks to capture the IMS data using as little rank as possible, while concurrently being aware of the sampling operator and thus missing values:

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \text{rank}(X), \\ \text{subject to} \quad & \mathscr{P}_\Omega(M) = \mathscr{P}_\Omega(X), \end{aligned} \tag{4.5}$$

where $X \in \mathbb{R}^{m \times n}$ and $\mathscr{P}_\Omega(\cdot) : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n}$ can be seen as an orthogonal projection projecting a matrix onto the space of $\mathbb{R}^{m \times n}$ matrices with support Ω.

SINGULAR VALUE THRESHOLDING

A strong advantage of the SVT is that during the optimization it utilizes the matrices in either a sparse or low-rank format, and it does not require a dense-format copy to be stored in memory. This is an extremely important aspect of this modeling effort, since many IMS datasets simply do not fit in memory when considered in a dense data format. A downside, however, is the singular value decomposition (SVD) at the center of each iteration of the optimization. Namely, the time complexity of the SVD, $\mathscr{O}(mn^2)$, becomes a bottleneck at the scale of MALDI-TOF IMS datasets, mainly due to a number of BLAS level-2 operations at the heart of the SVD [6]. However, different solutions exist to reduce, *e.g.*, [7]), or completely remove the SVD, *e.g.*, [8]. We opt for a divide-factor-conquer approach [9], as it is theoretically well-studied and acts as a framework that we can adapt for a second method.

DIVIDE-FACTOR-CONQUER APPROACH

Furthermore, to deal with both SVD complexity and memory load, we make use of the divide-factor-conquer approach (DFC) [9]. It consists of three steps and provides a framework that we can apply to both the SVT and FPC algorithms for obtaining an approximation $\bar{X}$ of matrix M with completion for the complete dataset M. Its first step consists of dividing M into t different matrices C_e, consisting of a subset of M's columns, sampled uniformly at random:

$$C_e = [\mathscr{P}_\Omega(M)]_{\Theta_e}, \ \forall \ e \in [1, t], \tag{4.6}$$

where Θ_e consists of a set of column indices of size $l = \frac{n}{t}$ and Θ contains t such sets. We assume for ease that l is an integer. As such, we obtain t matrices $C_e \in \mathbb{R}^{m \times l}$. Note that each C_e also has a particular $\mathscr{P}_\Omega^{\Theta_i}$ associated with it, namely the sampling associated to those columns in Θ_i. In the second step, these subsampled matrices are factorized separately using the matrix completion methods, either SVT or FPC. Hence, t low-rank matrices $\hat{X}_e$ are obtained. The rank of the overall approximation $\bar{X}$ is calculated by taking the median of all matrix ranks of $\hat{X}_e$. This rank is utilized in the final step, which consists of reconstructing the factored solutions $\hat{X}_e$ into a final approximate factorization $\bar{X}$. A standard Gaussian matrix $G \in \mathbb{R}^{m \times (k+p)}$ is constructed, with p as oversampling parameter, generally used to improve the reconstruction [7]. Next, a power iteration scheme is implemented as $Y = (\hat{X}\hat{X}^T)^q \hat{X}G$, with q as the number of iterations and $\hat{X}$ consists of re-ordering and stacking all low-rank approximations. Finally, the top k singular values of Y are obtained, *e.g.,* by QR decomposition, to form $Q \in \mathbb{R}^{m \times k}$. The final solution $\bar{X}$ is then obtained by

$$\bar{X} = QQ^\dagger \hat{X}, \tag{4.7}$$

where † is representing the pseudo-inverse. The advantage of the divide-factor-conquer approach is that conditions that arise in matrix completion methods, as presented earlier, also guarantee strong estimation properties for divide-factor-conquer [9]. We implemented the SVT and FPC algorithms as well as the divide-factor-conquer approach for both in an efficient Python object-oriented toolbox, with an eye towards saving memory where most needed and accelerating calculations where possible.

Case Study 1: Low-Rank Matrix Factorization Outperforms Traditional Peak Picking

Reconstruction Error for Different Numbers of Picked Peaks

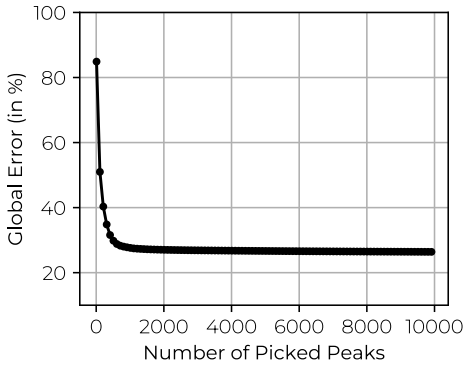

Figure S4.1: Reconstruction error plot for different numbers of picked peaks.

Raw and Imputed Ion Images for m/z 778.524

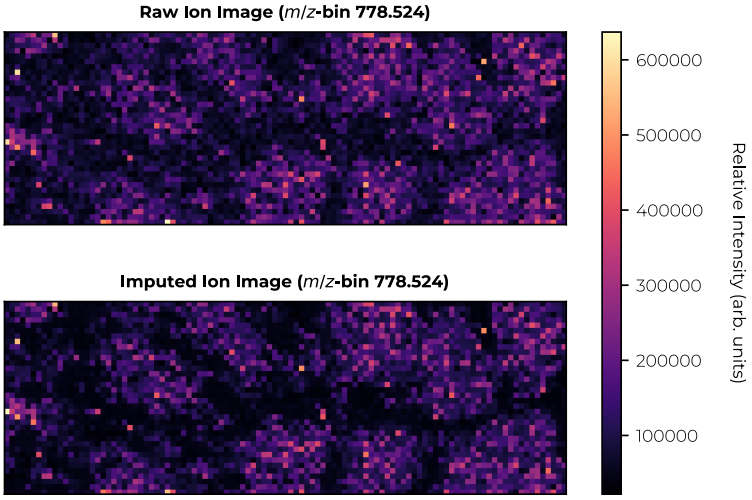

Figure S4.2: Raw and imputed ion images of *m/z*-bin 778.524.

CASE STUDY 2: RECONSTRUCTION AND IMPUTATION QUALITY WHEN DEALING WITH MISSING VALUES

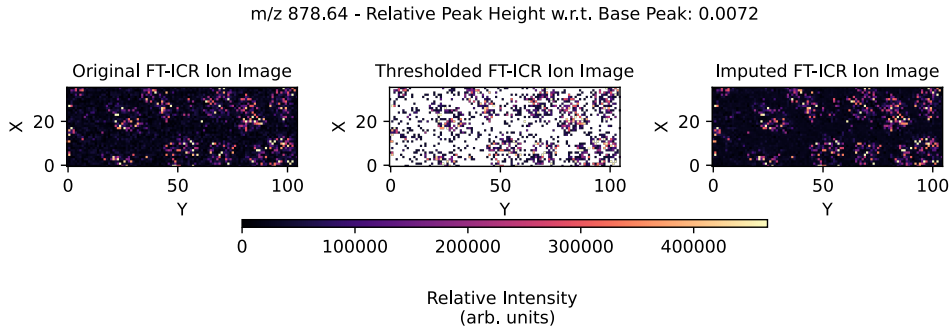

Figure S4.3: Raw, thresholded and imputed ion image of *m/z*-bin 878.640. This is a very low-intensity peak w.r.t. the base peak (0.0072 of the base peak height). The right image shows good imputation when visually compared to the original FT-ICR ion image.

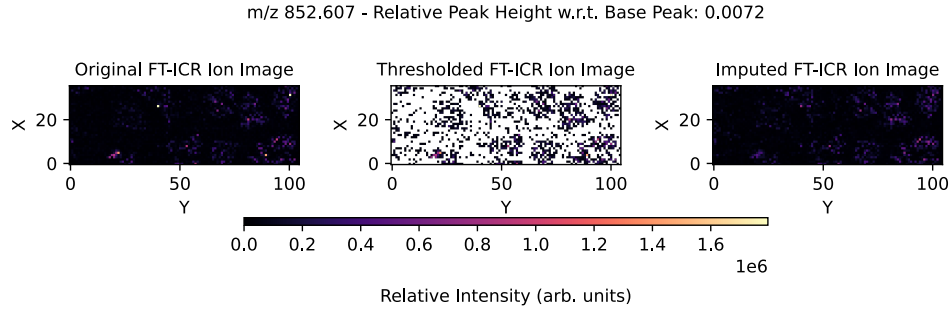

Figure S4.4: Raw, thresholded and imputed ion image of *m/z*-bin 852.607. This is a very low-intensity peak w.r.t. the base peak (0.0072 of the base peak height). The right image shows good imputation when visually compared to the original FT-ICR ion image, but it does noticably underestimate high-intensity features.

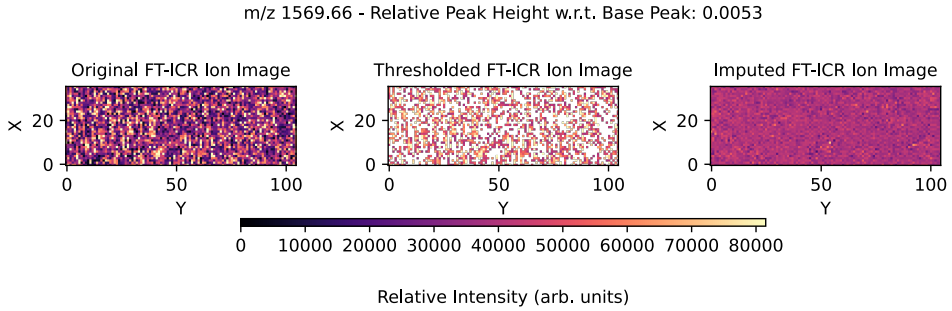

Figure S4.5: Raw, thresholded and imputed ion image of *m/z*-bin 1569.66. This is an extremely low-intensity peak w.r.t. the base peak (0.0053 of the base peak height). The right image shows imputation of a noise ion image. This result shows the denoising effect of the approach in action.

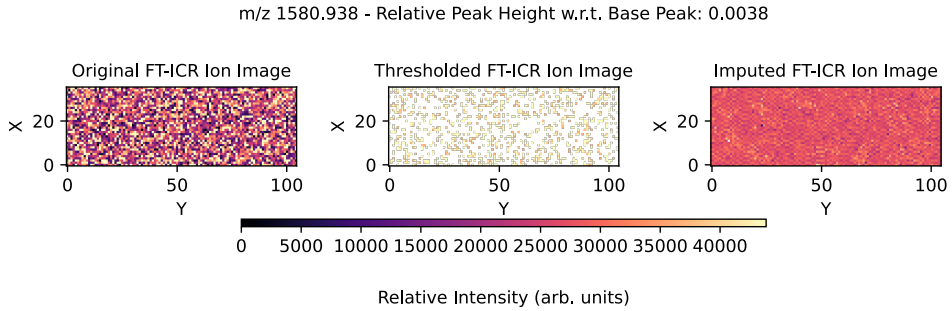

Figure S4.6: Raw, thresholded and imputed ion image of *m/z*-bin 1580.938. This is an extremely low-intensity peak w.r.t. the base peak (0.0038 of the base peak height). The right image shows imputation of a noise ion image. This result shows the denoising effect of the approach in action.

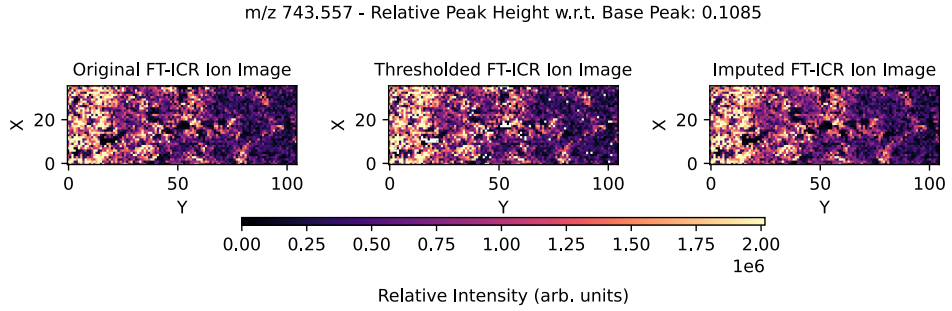

Figure S4.7: Raw, thresholded and imputed ion image of *m/z*-bin 743.557. This is a relatively high-intensity peak w.r.t. the base peak (0.1085 of the base peak height). The right image shows visually good imputation when compared to the original ion image.

POST-PROCESSING EXAMPLE

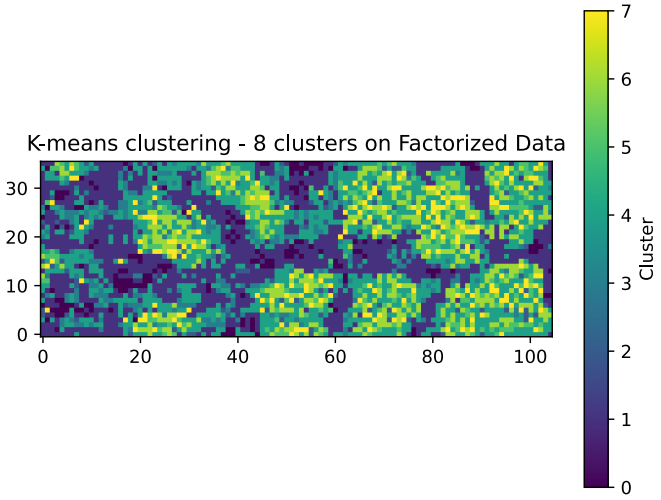

Figure S4.8: *k*-means clustering with 8 clusters applied to factorized FT-ICR data. The visualization of clusters, segmenting the tissue in the process, shows good correspondence to (presumed to be biological) patterns also visible in the total ion image.

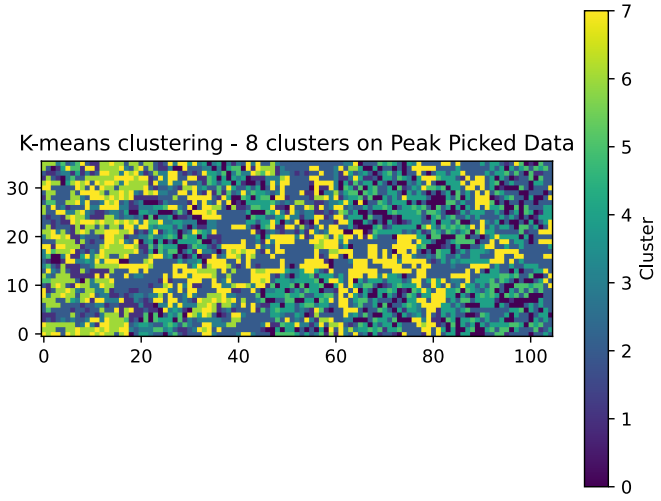

Figure S4.9: *k*-means clustering with 8 clusters applied to raw peak picked (100 peaks) FT-ICR data. The visualization of clusters, segmenting the tissue, shows some correspondence with respect to the total ion image, but less compared to the results on factorized FT-ICR data (Fig. S4.8).

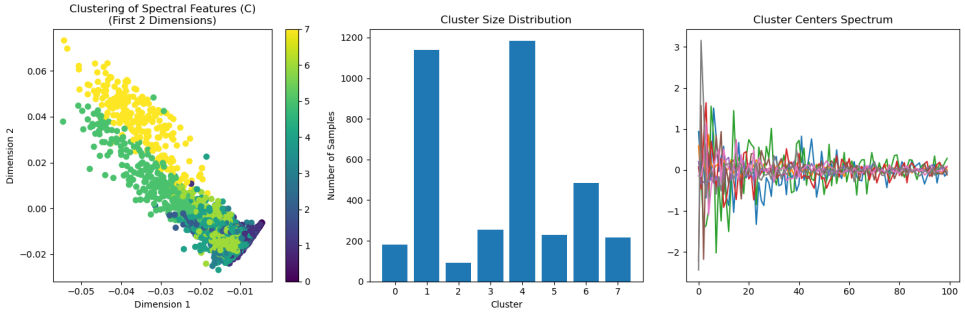

Figure S4.10: *k*-means clustering with 8 clusters applied to the factorized FT-ICR data. We provide some further details on the results shown in Fig. S4.8, namely a visualization of the cluster-labeled data points along the first two latent dimensions, the distribution of cluster sizes, and the centers of cluster spectra.

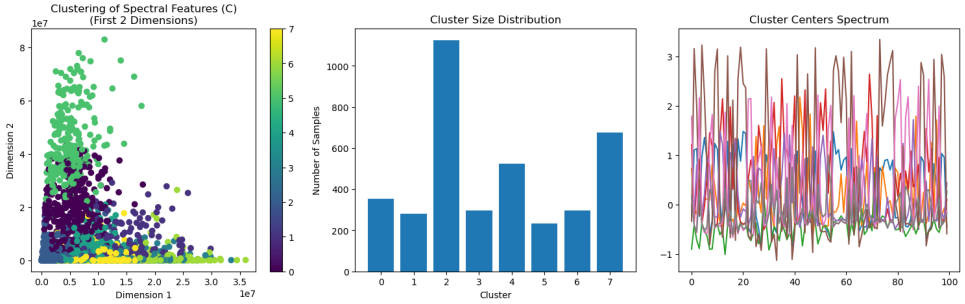

Figure S4.11: *k*-means clustering with 8 clusters applied to the raw peak picked (100 peaks) FT-ICR data. We provide some further details on the results shown in Fig. S4.9, namely a visualization of the cluster-labeled data points along the first two latent dimensions, the distribution of cluster sizes, and the centers of cluster spectra.

CASE STUDY 3: ADVANTAGES AND DISADVANTAGES OF LOW-RANK MATRIX COMPLETION FOR MISSING VALUE TOF IMS DATA

ION IMAGE EXAMPLES

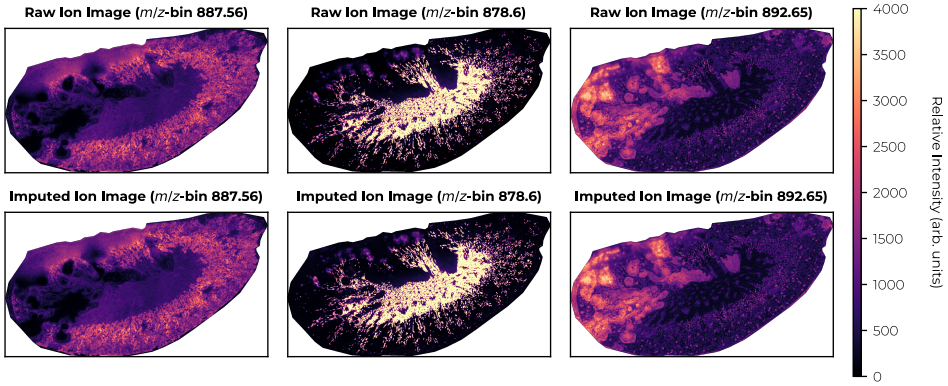

Figure S4.12: In the first row, the three columns depict different ion images retrieved from the raw data, namely *m/z* 887.56, *m/z* 878.60 and *m/z* 892.65. These ion species are approximated by the SVT, and shown in the second row. These ion images are shown as representatives of high-intensity peaks.

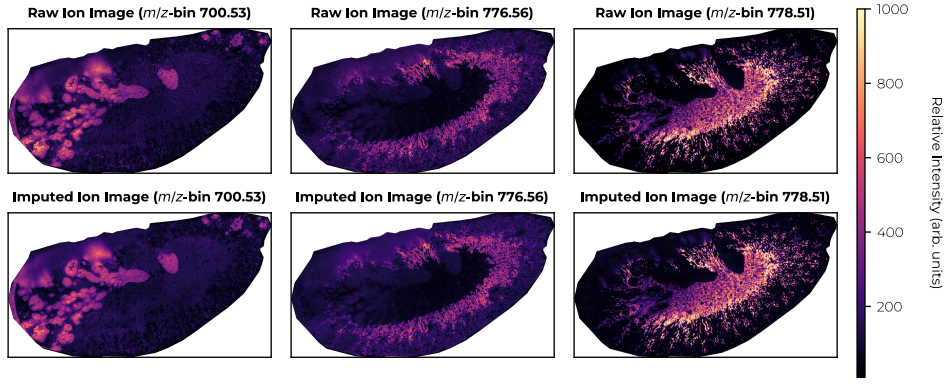

Figure S4.13: In the first row, the three columns depict different ion images retrieved from the raw data, namely *m/z* 700.53, *m/z* 776.56 and *m/z* 778.51. These ion species are approximated by the SVT, and shown in the second row. These ion images are shown as representatives of lower-intensity peaks.

SPECTRAL ERROR DISTRIBUTION AND BIOLOGICAL INTERPRETATION

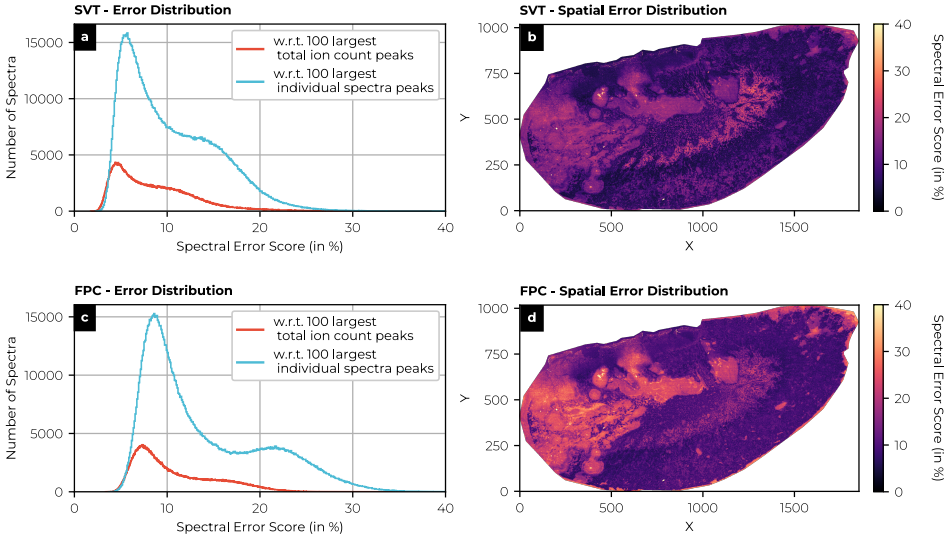

Figure S4.14: Spectral error score, *i.e.*, $\frac{||\tilde{M}_{i\bullet} - X_{i\bullet}||_2}{||\tilde{M}_{i\bullet}||_2}$, reports on the error of individual spectra. The distribution of the spectral error score is given (left) for both SVT and FPC methods with respect to the 100 largest total ion current count peaks and with respect to the 100 largest individual spectral peaks. The spatial distribution for the spectral error score with respect to the 100 largest individual spectra peaks is also depicted for the SVT and FPC (right).

METHODOLOGICAL EFFECTS ON RECONSTRUCTED ION IMAGES AND SPECTRA

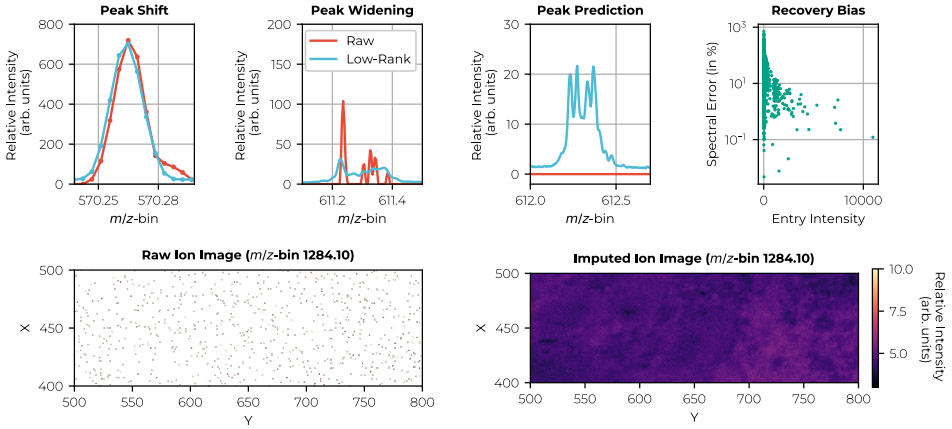

Figure S4.15: The first row depicts four potential effects of the low-rank approximation of a complete dataset on individual spectra, namely peak shifting (*i.e.*, small shifts of the peak center), peak widening, peak prediction (potential introduction of low-intensity peaks in line with patterns observed in the rest of the dataset), and bias in the recovery error for large peaks (better recovery/representation of large peaks). In the second row, an extreme example is given of an individual ion image at *m/z* 1284.10, both raw (left) and recovered (right). The predicted image depicts a biological scene, even though the raw image contains barely any data points (very sparse). The missing values in the image are imputed, based on the available information from the rest of the dataset. This could lead to a substantial imputation error for those ion species. These effects are a price we pay for retaining full spectrum information in our dimensionality reduction methods, and they affect low-abundant and missing pixel-dominated species first. As such, we robustly keep the full spectrum profile intact for higher intensity peaks, in contrast to selective peak picking. At the same time, these recognized shortcomings can be taken into account for future model improvements.

PRESERVATION OF NEAR-ISOBARIC SPECIES

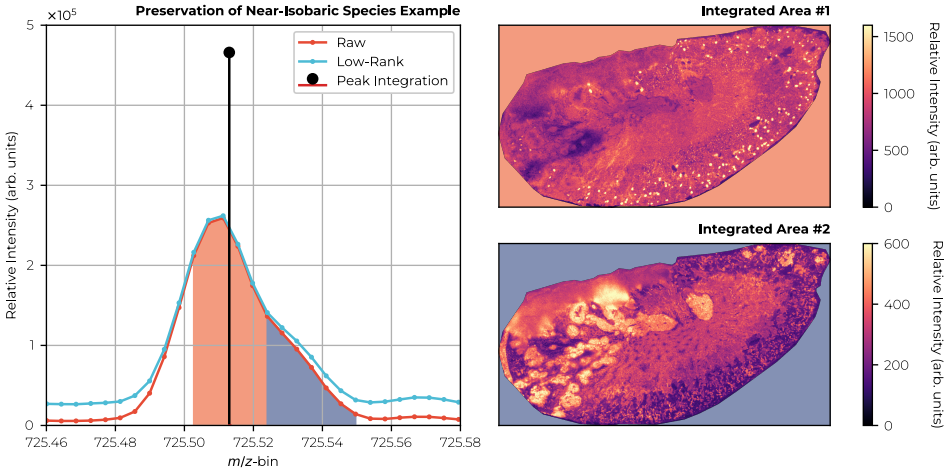

Figure S4.16: A particular spectrum is examined (left) in the range between m/z 725.46 and 725.58 from the perspective of both the low-rank approximation and raw data, along with a peak integrated version. The highlighted (orange and blue) areas under the curve are integrated and spatially depicted (right). We observe that the shoulder (blue) differs spatially from the peak (orange). With peak integration, relative information of the shoulder (blue) is lost, due to its weak signal.

RETENTION OF LOWER-INTENSITY ION SPECIES AND BIAS MITIGATION

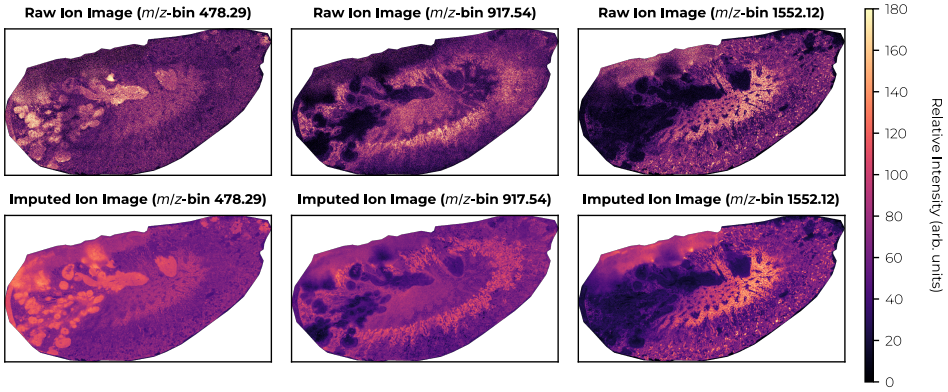

Figure S4.17: In the first row, the three columns depict different ion images as found in the raw data, namely m/z 479.29, m/z 917.54 and m/z 1 552.12. These ion species are detected in more than 80% of the IMS pixels, they are not isotopes, and their distributions suggest a biology-driven distribution. Nevertheless, they would be disregarded in a simple peak picking procedure, if *e.g.*, only the 1 000 largest peaks would be retained. Our approach (see second row) retains these ion species, albeit approximately.

SAMPLE PREPARATION: HUMAN KIDNEY TISSUE (FT-ICR IMS)

Human kidney tissue was surgically removed during a full nephrectomy, and remnant tissue was processed for research purposes by the Cooperative Human Tissue Network (CHTN) at Vanderbilt University Medical Center. Human biospecimens were collected in compliance with the CHTN protocols, institutional IRB policies, and the National Cancer Institute's best practices for procurement of remnant surgical research material. The tissue was flash-frozen over an isopentane-dry ice slurry and embedded in carboxymethylcellulose. Tissue sections were cryosectioned with a thickness of 10 μm and thaw-mounted onto indium tin-oxide-coated glass slides (Delta Technologies, Loveland, CO, USA). 1,5-Diaminonaphthalene (DAN) was applied to the tissue surface using a TM Sprayer M3 (HTX Technologies, Chapel Hill, NC, USA). The sample was imaged (50 μm pitch) directly after matrix application with a 15T MALDI Fourier-transform ion cyclotron resonance (FT-ICR) mass spectrometer (Solarix, Bruker Daltonics, Billerica, MD, USA). Briefly, data were generated from m/z 300-2 000 with a 4M file size in negative ion mode.

SAMPLE PREPARATION: MOUSE KIDNEY TISSUE (QTOF IMS)

C57BL6/J mice were retro-orbitally infected with *S. aureus* Newman and sacrificed humanely five days post-infection. Kidneys were flash frozen on an isopentane/dry ice slurry and embedded in 2.6% carboxymethylcellulose. 5 μm thick sections were collected using a Leica Biosystems CM3050S cryostat and thaw-mounted onto indium tin oxide coated glass slides. Sections were washed with cold 150 mM ammonium formate for 45 seconds for three total washes. 5 mg of 4-(dimethylamino)cinnamic acid matrix was applied using an in-house sublimation device. MALDI IMS data were collected in negative ion mode from m/z 400-2 000 using a Bruker MALDI timsTOF fleX platform with a 5 μm step size, 25% laser power, and 25 shots per pixel.

DATA PREPROCESSING

The TOF dataset was *m/z*-aligned with a custom program [3, 10, 11]. For all case studies, we used a 5-95% TIC pixel normalization method, *i.e.,* scaling each row by the sum of its entries between 5%-percentile and 95%-percentile [3, 10, 11]. We did not statistically normalize (*i.e.*, mean subtraction and scaling) the features, *i.e., m/z*-bins (columns), as we observed that it hindered low-rank recovery, presumably due to the majority of measured *m/z*-bins consisting of low signal-to-noise measurements, sub-noise features or noise features. Statistically normalizing the *m/z*-bins amplifies the singular values associated with the noise subspace while diminishing those linked to the signal subspace, particularly in high-intensity *m/z*-bins. Further exploration of advanced normalization schemes was deemed beyond the scope of Chapter 4. All calculations were performed on a Dell Precision 7920 workstation with 56 cores at 2.7 GHz and 1.5 TB of memory, and two NVIDIA A6000 GPUs connected via NVLink Bridge.

PARAMETER SETTING

For SVT, convergence for the completion problem is guaranteed if $0 < \delta < 2$ [12]. However, it was also noted that this choice can be too conservative, and the convergence slow [12]. For our datasets, we found that setting the parameter to $\delta > 2$ breaks the convergence and

Table 4.4: SVT and FPC, with random sampling scheme β. A moderate reconstruction error is observed for all methods for both raw and low-rank input and references. The imputation error is substantial. For sampling scheme β, the impact of the imputation error is larger on the global error. For SVT, we set parameters $\delta = 1$ and $\tau = 10^{-3}$ and for FPC, we set $\delta = 1.4$ and $\tau = 10^{-3}$ (see Supplementary Materials 4.4). For SVT with raw input data we obtain a 111 rank solution and for FPC a 111 rank solution. We truncate all solutions to a rank of 100 for fair comparison. The SVT took on average 106 minutes to converge, the FPC algorithm on average 19 minutes.

Input M	**Reference** $\tilde{M}$	**Method**	**Rank**	**Reconstruction Error** $\frac{\|P_\Omega(\tilde{M}-X)\|_F}{\|P_\Omega(\tilde{M})\|_F} \times 100\%$	**Imputation Error** $\frac{\|P_{\Omega_c}(\tilde{M}-X)\|_F}{\|P_{\Omega_c}(\tilde{M})\|_F} \times 100\%$	**Global Error** $\frac{\|\tilde{M}-X\|_F}{\|\tilde{M}\|_F} \times 100\%$
Raw	Raw	SVT	100	50.0	92.6	89.6
Raw	Raw	FPC	100	29.3	84.9	81.6
Raw	Low-Rank	SVT	100	48.3	92.2	89.2
Raw	Low-Rank	FPC	100	23.5	84.2	80.7
Low-Rank	Low-Rank	SVT	100	67.9	92.4	90.5
Low-Rank	Low-Rank	FPC	100	22.8	84.4	80.9

that values of $\delta < 1$ lead to slow convergence. Hence, we manually tuned the values per experiment between $1 < \delta < 1.7$. For FPC, different suggestions are made for setting δ [13, 14]. We found by manually tuning that convergence was ensured for our datasets settings values, similar to SVT, between $1 < \delta < 2$.

Although different suggestions for SVT's and FPC's τ exist (note that τ is described by μ for FPC [13]), we defined τ relative to $\max(m,n)$. By trial-and-error, we found values between 10^{-1} and 10^{-3} perform adequate to obtain low-rank solutions. We did not observe large deviations in recovery when setting the parameters in those ranges. A parameter sensitivity analysis is, however, advisable to further tune the outcomes.

For the divide, factor and conquer approach, the raw data matrix $M \in \mathbb{R}^{312\,249 \times 1\,372\,421}$ was divided into 3 303 subsampled matrices, $C_i \in \mathbb{R}^{312\,249 \times 400}$, where $i \in [1, 3\,303]$. Subsampling was performed along the spatial axis, which outperformed spectral axis subsampling due to the presence of many noise-dominant *m/z*-bins. Spectral subsampling risked aggregating noise features into the same subsampled matrices, leading to faulty results. Sampling sizes of $300 - 500$ spectra were heuristically found to ensure good recovery, *i.e.,* fulfilling the low-rank constraint while allowing for efficient computation, completing within a 12-hour runtime.

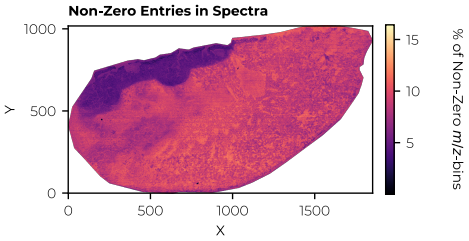

Figure S4.18: Spatial distribution of the number of non-zero values per spectrum. Dark regions correspond to spectra with few values (after clipping), while bright regions reflect spectra with more measured values. Note that most spectra only contain $2 - 30\%$ of measured values. Some spatial regions contain more zeroes, *i.e.,* fewer features are captured.

REFERENCES

[1] T. Alexandrov. MALDI Imaging Mass Spectrometry: Statistical Data Analysis and Current Computational Challenges. In: *BMC Bioinformatics* 13.Suppl 16 (2012), S11.

[2] D. M. Anderson, R. Van de Plas, K. L. Rose, S. Hill, K. L. Schey, A. C. Solga, D. H. Gutmann, and R. M. Caprioli. 3-D Imaging Mass Spectrometry of Protein Distributions in Mouse Neurofibromatosis 1 (NF1)-associated Optic Glioma. In: *Journal of Proteomics* 149 (2016), pp. 77–84.

[3] P. Monchamp, L. Andrade-Cetto, J. Y. Zhang, and R. Henson. Signal Processing Methods for Mass Spectrometry. In: *Systems Bioinformatics: An Engineering Case-Based Approach, Artech House Publishers* (2007).

[4] N. Verbeeck, R. M. Caprioli, and R. Van de Plas. Unsupervised Machine Learning for Exploratory Data Analysis in Imaging Mass Spectrometry. In: *Mass Spectrometry Reviews* 39.3 (2020), pp. 245–291.

[5] A. González-Fernández, A. Dexter, C. J. Nikula, and J. Bunch. NECTAR: A New Algorithm for Characterizing and Correcting Noise in QToF-Mass Spectrometry Imaging Data. In: *Journal of the American Society for Mass Spectrometry* 34.11 (2023), pp. 2443–2453.

[6] J. Dongarra, M. Gates, A. Haidar, J. Kurzak, P. Luszczek, S. Tomov, and I. Yamazaki. The Singular Value Decomposition: Anatomy of Optimizing an Algorithm for Extreme Scale. In: *SIAM Review* 60.4 (2018), pp. 808–865.

[7] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. In: *SIAM Review* 53.2 (2011), pp. 217–288.

[8] T. Zhou and D. Tao. Godec: Randomized Low-rank & Sparse Matrix Decomposition in Noisy Case. In: *Proceedings of the 28th International Conference on Machine Learning*. 2011.

[9] L. W. Mackey, A. Talwalkar, and M. I. Jordan. Distributed Matrix Completion and Robust Factorization. In: *Journal of Machine Learning Research* 16.1 (2015), pp. 913–960.

[10] L. G. Migas. *msalign: Spectral alignment based on MATLAB's 'msalign' function*. https://github.com/lukasz-migas/msalign. Version 0.2.0. 2024.

[11] M. A. Farrow, L. E. M. Tideman, E. K. Neumann, L. G. Migas, N. H. Patterson, M. E. Colley, J. L. Allen, E. L. Pingry, M. Dufresne, H. Yang, M. Brewer, E. S. Rivera, C. E. Romer, K. Djambazova, K. Sharman, A. R. S. Kruse, D. B. Gutierrez, R. C. Harris, A. B. Fogo, M. P. de Caestecker, R. M. Caprioli, R. V. de Plas, and J. M. Spraggins. A Lipid Atlas of the Human Kidney. In: *Science Advances* 11.24 (2025), eadu3730.

[12] J.-F. Cai, E. J. Candès, and Z. Shen. A Singular Value Thresholding Algorithm for Matrix Completion. In: *SIAM Journal on Optimization* 20.4 (2010), pp. 1956–1982.

[13] E. J. Candes and Y. Plan. Matrix Completion with Noise. In: *Proceedings of the IEEE* 98.6 (2010), pp. 925–936.

[14] S. Ma, D. Goldfarb, and L. Chen. Fixed Point and Bregman Iterative Methods for Matrix Rank Minimization. In: *Mathematical Programming* 128.1 (2011), pp. 321–353.

5

INCLUDING STRUCTURED PRIORS ON A TERABYTE SCALE

Imaging mass spectrometry (IMS) provides spatially resolved molecular information of organic tissue but can be limited by pixel signals mixing contributions from adjacent biological structures, e.g. of single cells and multicellular functional tissue units (FTUs). This chapter proposes computational methods to predict mass spectral profiles of biological structures on the basis of IMS data by "unmixing" pixel-level signals, leveraging microscopy-based boundary information of these structures. By modeling each biological structure as having a unique mass spectrum, we formulate a linear mixing model and solve the corresponding inverse problem that unmixes blended signals. In particular, we cover both overdetermined and underdetermined linear system scenarios and compare ordinary least squares, non-negative least squares, and singular value thresholding to a custom algorithm, coined Tissue-informed Unmixing of Labeled regions by Inverse Problem (TULIP), specifically tailored to IMS data. Validation on a synthetic in-situ single cell dataset and demonstration on a large-scale kidney FTU dataset illustrate the potential of these methods for enhanced in-situ tissue structure analysis, e.g. in cellular and tissue studies.

The contents of this chapter are based on:

Moens, R. A. R., Patterson, N.H., Migas, L.G., Esselman, A.B., Moser, F.A., Spraggins, J.M. & Van de Plas, R. (2025). Unmixing of Imaging Mass Spectrometry Measurements Using Microscopy-informed Constraints [Unpublished Manuscript].

5.1. INTRODUCTION

Imaging mass spectrometry (IMS) has emerged as a powerful tool for spatially resolved, label-free molecular analysis across a range of biological samples, from large tissue sections to individual cells [1, 2, 3, 4, 5, 6]. It enables *in-situ* molecular mapping of proteins, lipids, and metabolites, without the need for chemical stains or a priori information [1, 2], making it valuable for investigating biological processes such as disease pathology, tissue heterogeneity, and metabolic changes [7, 8, 9]. Although there are various surface sampling approaches, MALDI is one of the most common due to its combination of high spatial fidelity and molecular coverage.

A commonly encountered issue within MALDI, with potentially significant impact, is the imperfect spatial alignment between the sampling pattern, here simplified as a measured IMS pixel, and the underlying biological structures. Many structures or regions of interest, such as organelles (depending on the type, below ≈ 1 μm), small-scale cells (depending on the type, at around or below 5 μm), groups of cells (depending on the type and number, at or above 5 μm), or larger-scale functional tissue units (FTUs, depending on the type ≈ 100 μm), may be smaller than the pixel size (at or around 5 μm) and/or span across neighboring pixels, leading to blended or mixed spectra [10, 11, 12]. This blending limits the specificity of molecular assignments to these structures or regions of interests, particularly in spatially heterogeneous contexts. While instrumental advances like oversampling [13] and transmission geometry MALDI [14, 15] have pushed pixel sizes downwards as a partial solution, they often compromise throughput, sensitivity, or molecular coverage.

Several computational approaches have sought to resolve these limitations by unmixing coarse IMS pixels into regions of interest using prior information, such as provided by microscopy modalities [10, 16, 17, 18, 19]. They offer higher spatial resolution but most of the time present either insufficiently powerful (local) weighing schemes, are tightly coupled to a specific instrumental resolution, rely on physical cell isolation, or are not easily generalized across tissues or imaging modalities. To address these, we propose TULIP (Tissue-informed Unmixing of Labeled regions by Inverse Problem), a framework that aims to predict the molecular profiles of structures of interest from IMS data by using microscopy-informed constraints. TULIP formulates spectral unmixing as an inverse problem that incorporates spatial priors, *e.g.*, from microscopy, without requiring physical isolation of the regions of interest. It combines biologically motivated constraints such as sparsity, non-negativity, and low-rank assumptions [20, 21], and is designed to operate across different spatial scales and biological hierarchies, from (sub)cellular domains to tissue structures to anatomical regions. We take a heuristic approach leveraging nuclear norm minimization as a relaxation of rank minimization [22], and bridge ideas from maximum-margin matrix factorization [23], rewriting the nuclear norm as a non-convex Frobenius norm loss [24] to reduce the memory complexity. Finally, we considered advances in non-negative matrix completion [25], as well as regularized non-negative matrix factorization (NMF) [26] to construct our optimization problem. Although solving our proposed constrained low-rank matrix approximation remains generally NP-hard [27], (non-)convex relaxations and heuristic optimizers have provided a practical path forward for our setting.

5.2. METHODS

5.2.1. INVERSE PROBLEM FORMULATION

Our methodological framework defines the combination of IMS with high spatial resolution microscopy data as a forward model. Each biological structure is modelled as a region of interest (ROI), *e.g.*, a single cell, a cell type group, or an FTU. Each ROI is assumed to be deterministic and homogeneous in molecular/spectral content, thus represented by a unique mass spectrum. Our model considers IMS pixels to represent spectra as linear mixtures of the underlying ROI profiles, *e.g.*, weighted proportionally by their spatial overlap with IMS pixels. Formally, we express the forward model as:

$$Y = AX, \tag{5.1}$$

where $Y \in \mathbb{R}^{p \times q}$ represents IMS measurements (known) across p pixels and q mass-to-charge (m/z) bins, $A \in \mathbb{R}^{p \times r}$ is the microscopy-derived spatial mixing matrix (known) indicating overlaps between IMS pixels and annotated ROIs (*e.g.*, individual cells, cell type groups, or FTUs), and $X \in \mathbb{R}^{r \times q}$ contains the predicted spectra corresponding to each annotated ROI (unknown) with r the total number of such ROIs (known). Note that the mixing matrix A could include a more advanced weighting scheme, incorporating prior knowledge on *e.g.*, the laser ablation spot morphology or accounting for a non-uniform grid of pixels. Finally, the heterogeneity of the samples, the quality of annotations in A, and available computational resources could also be considered when constructing the model and selecting the right solution method.

SINGLE CELL EXAMPLE

To account for non-annotated regions (NARs, see **Fig. 5.1**), arising *e.g.*, due to the morphology of the tissue, we extend the formulation in Eq. 5.1 by:

$$Y = A'X' + \Lambda R, \tag{5.2}$$

where $Y \in \mathbb{R}^{p \times q}$ represents IMS measurements, $A' \in \mathbb{R}^{p \times r}$ is the microscopy-derived spatial mixing matrix, $X' \in \mathbb{R}^{r \times q}$ contains the predicted ROIs (here single cells spectra), $\Lambda \in \mathbb{R}^{p \times k}$ captures residual spatial weights for k non-annotated regions ensuring that rows of A (see Eq. 5.3) sum to unity[1], and $R \in \mathbb{R}^{k \times q}$ represents the estimated spectra for these non-annotated regions. Note that we model the NAR in each IMS pixel separately ($k = p$), but we could also model it as a single spectrum by setting $k = 1$.

[1] In our case, this constraint follows the nature of laser-based ablation, where the same amount of energy is deposited for each IMS pixel surface area.

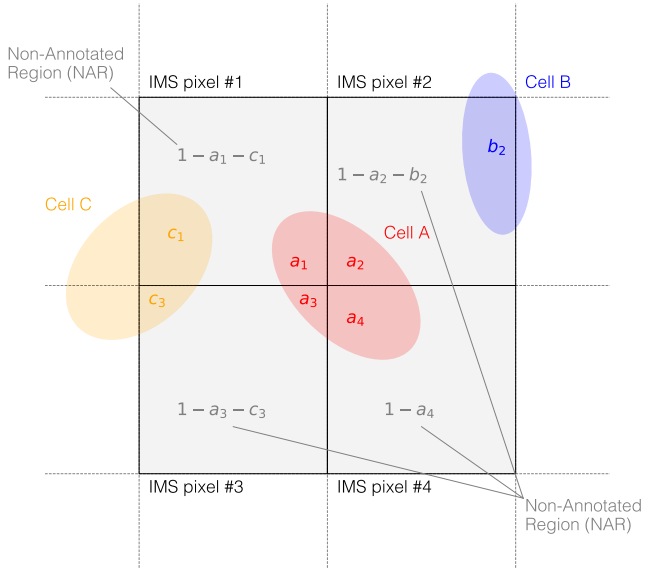

Figure 5.1: Simplified representation of IMS pixels and annotated ROIs (in this case single cells). Lowercase letters indicate relative overlap areas within an IMS pixel, normalized to unity.

As such,

$$\begin{aligned} Y &= A'X' + \Lambda R \\ &= \begin{bmatrix} A' & \Lambda \end{bmatrix} \begin{bmatrix} X' \\ R \end{bmatrix} \\ &= AX, \end{aligned} \tag{5.3}$$

where for the particular example shown in **Fig. 5.1**

$$A' = \begin{bmatrix} a_1 & 0 & c_1 \\ a_2 & b_2 & 0 \\ a_3 & 0 & c_3 \\ a_4 & 0 & 0 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \alpha & 0 & 0 & 0 \\ 0 & \beta & 0 & 0 \\ 0 & 0 & \gamma & 0 \\ 0 & 0 & 0 & \delta \end{bmatrix}, \tag{5.4}$$

with $\alpha = 1 - a_1 - c_1$, $\beta = 1 - a_2 - b_2$, $\gamma = 1 - a_3 - c_3$, and $\delta = 1 - a_4$. Each a_i, b_i and c_i correspond to the relative overlap in area of the ROIs, *i.e.* cell, with the particular IMS pixel i such that the following holds:

$$\sum_{j=1}^{r+k} A_{ij} = \sum_{j=1}^{r+k} (A'_{ij} + \Lambda_{ij}) = \mathbf{1}. \tag{5.5}$$

Note that depending on the setting and the modelling, the condition number of the matrix A can be affected by modelling NARs. In turn, this can lead to instabilities if not properly addressed by the method.

5.2.2. LINEAR SYSTEM

Given the data sizes typical of these IMS problems, where matrices with more than 100 000 rows and columns are not uncommon [20], we opt for a straightforward and efficient linear system approach. Commonly, two scenarios are encountered in such an approach:

OVERDETERMINED LINEAR SYSTEM

This scenario encompasses situations where the mixing matrix A is tall and narrow, *i.e.*, more rows than columns. For example, it can, but not exclusively, consists of setups in which non-annotated regions (NARs) are either absent or only partially present, and/or where ROIs are relatively large with respect to the IMS pixel size. Following the notation from Eq. 5.1, this case is described by:

$$Y = AX, \quad \text{with} \quad p \geq r \quad (\text{with } A \in \mathbb{R}^{p \times r}). \tag{5.6}$$

Usually, no exact solution exists[2]. However, a unique least squares solutions can exist[3] and be efficiently obtained. While computationally efficient, the least squares approach may *e.g.*, not fully capture detailed residual contributions in highly spatially heterogeneous samples, such as the NAR in the single cell example of **Fig. 5.1**.

UNDERDETERMINED LINEAR SYSTEM

This scenario applies where the mixing matrix A is short and wide, *i.e.*, more columns than rows. For example, it can, but not exclusively consists of setups where the comprehensive contributions from NARs or multiple annotated regions per single IMS pixels are modelled. In this case, again following the notation from Eq. 5.1, the number of unknowns exceeds the number of equations, this case is described by:

$$Y = AX, \quad \text{with} \quad p < r \quad (\text{with } A \in \mathbb{R}^{p \times r}). \tag{5.7}$$

Because the system has more unknowns than equations, additional constraints or prior knowledge, *e.g.*, regularization or sparsity constraints, are necessary to ensure uniqueness of a solution. Although often more computationally demanding to solve, this model allows for a detailed representation, enabling more rigorous downstream analyses and possibly increasing model fit of the NAR and ROI spectra. The overdetermined model is preferable when residual contributions are thought to be minimal or computational efficiency is prioritized, whereas the underdetermined model can be useful, *e.g.*, for capturing substantial non-annotated variability.

5.2.3. PRIOR KNOWLEDGE ON IMS AND MICROSCOPY DATA

In many problems involving IMS data, leveraging relevant prior knowledge can significantly improve the accuracy, robustness, and interpretability of the results by constraining the solution space [20, 21], but it can also improve the computational and memory complexity. Examples of such prior information include:

1. **Measurement matrix Y:**

[2] *I.e.*, not consistent, *e.g.*, when Y is not fully in the column space of A.
[3] Usually when rank(A) is full.

- Measured IMS data in Y are sometimes subjected to a non-linear transformation $f(\cdot)$, *e.g.*, due to detector clipping, thresholding measured ion intensity values M_{ij} that fall under a particular threshold $\epsilon \in \mathbb{R}_+$ [20]:

$$Y_{ij} = f(M_{ij}) = \begin{cases} M_{ij}, & \text{if } M_{ij} \geq \epsilon, \\ 0, & \text{if } M_{ij} < \epsilon, \end{cases} \tag{5.8}$$

  resulting in sparsity levels of the full profile IMS data below 10%.
- Contains a measurement error, *e.g.*, originating from sample preparation procedures and instrumental artifacts (*e.g.*, delocalization, matrix-solution binding, laser pattern deformation, detector jitter) assumed to be distributed in a Gaussian fashion.
- High-dimensional: commonly more than 200 000 *m/z*-bins per pixel, and usually more than 100 000 pixels.

2. **Mixing matrix A:**
   - Usually very high sparsity,
   - Full-rank, possibly high condition number,
   - Bounded, *i.e.*, $\|A\|_\infty \leq 1$ with entries $0 \leq A_{ij} \leq 1$,
   - Possibly disturbed by an error term ΔA, *i.e.* $A + \Delta A$.
3. **ROI spectra X:**
   - Consists of sums of dependent and independent Poisson-distributed variables,
   - Spectra measured independently, but *m/z*-bins exhibit interdependencies,
   - Exhibits low-rank and sparsity characteristics.

Using this information and knowing that measured ion intensities are non-negative values, we propose a novel method for solving inverse problems occurring in IMS by using microscopy-informed constraints.

5.2.4. TISSUE-INFORMED UNMIXING OF LABELED REGIONS BY INVERSE PROBLEM (TULIP)

We propose the TULIP method, incorporating prior knowledge and related to regularized least squares with rank, sparsity and non-negativity constraints.

PROPOSED PROGRAM

$$\begin{aligned} \underset{L\in\mathscr{C}_L, R\in\mathscr{C}_R}{\text{minimize}} \quad & \frac{1}{2}(\|L\|_F^2 + \|R\|_F^2) \\ \text{subject to} \quad & \|\mathscr{P}_\Omega(ALR^T - Y)\|_F^2 \leq \sigma^2 pq, \end{aligned} \tag{5.9}$$

where

- $\mathscr{C}_L = \{L \in \mathbb{R}_+^{r\times t} : \|L\|_1 \leq s_L\}$ with t denoting the factorization rank, and non-negativity promoting biological interpretability;
- $\mathscr{C}_R = \{R \in \mathbb{R}_+^{q\times t} : \|R\|_1 \leq s_R\}$ with t denoting the factorization rank, and non-negativity promoting biological interpretability;
- s_L, s_R are sparsity parameters that limit the ℓ_1-norm of L and R respectively, promoting solution uniqueness and alignment to physical constraints and thus interpretability [21, 28];
- $\mathscr{P}_\Omega$ is the projection operator onto the set of observed indices Ω, *i.e.*, for any matrix $Z \in \mathbb{R}^{m\times n}$,

$$(\mathscr{P}_\Omega(Z))_{ij} = \begin{cases} Z_{ij}, & \text{if } (i,j) \in \Omega, \\ 0, & \text{otherwise;} \end{cases}$$

- σ represents the standard deviation of the measurement noise (under the assumption that it is modelled by a Gaussian distribution), so that the constraint $\|\mathscr{P}_\Omega(ALR^T - Y)\|_F^2 \leq \sigma^2\, pq$ limits the reconstruction error to a level consistent with the noise floor;
- the term pq corresponds to the total number of elements in $Y \in \mathbb{R}^{p\times q}$, ensuring that the error tolerance scales with the size of the problem;
- $\|\cdot\|_F^2$ is the squared Frobenius norm.

TULIP builds on a low-rank approach based on a nuclear norm minimization [22], a proxy for the non-convex rank minimization, and makes use of concepts from maximum-margin matrix factorization [23] to obtain the Burer-Monteiro approximation [24]. For the latter, it has been proven, under specific conditions, that such non-convex formulations can avoid spurious local minima and recover global optima in polynomial time, see *e.g.*, [29]. However, we will not assume any of these conditions to hold for our proposed problem. Merely, its explicit factorization of $X = LR^T$ simplifies our large-scale optimization and reduces the memory complexity. In addition, parallels between our proposed program and a regularized non-negative matrix factorization [26, 30, 31, 32] and non-negative matrix completion can be drawn [21, 25]. Our approach is purely heuristic, combining elements of (non-negative) matrix completion and low-rank matrix approximation together with prior information to provide a practical pathway for addressing our problem in its high-dimensional biological setting. Note that we consider the completion projector $\mathscr{P}_\Omega$ as a practical construct, as the sampling pattern in IMS is often intensity-based. This means that low intensity values are removed, leading to non-uniform sampling schemes and thus typically no adherence to incoherence or restricted isometry properties [20]. The formulation including the projector $\mathscr{P}_\Omega$ allows, however, for the sparse format of Y and ALR^T to be efficiently exploited and reduce the memory footprint, as no dense matrix of size $p \times q$ is required in memory during optimization.

PROJECTED GRADIENT DESCENT

Given the inherent sparsity and scale of the data, we employ a projected gradient descent with Adam optimizer (for a parameter discussion see Appendix: Parameter Tuning) to set

adaptive learning rates [33]. The primary advantages of this approach are its flexibility, simplicity, and efficient memory usage without the need for matrix inverses. An outline of the optimization is given in **Algorithm 1**, where the projection onto $\mathscr{C}_e = \{Q : Q \geq 0, \|Q\|_1 \leq s_e\}$ is performed by a soft-thresholding with threshold λ_e. We choose to check the fidelity constraint $\|\mathscr{P}_\Omega(ALR^T - Y)\|_F^2 \leq \sigma^2\, pq$ at the final iteration, but this can also be done during the optimization.

Algorithm 1 Projected Gradient Descent for TULIP

Require: Initial points L_0, R_0, step sizes α_L, α_R, tolerance ϵ
Require: Measurement matrix Y, mask Ω, parameters $\gamma, \sigma, \lambda_i$
$k \leftarrow 0$
while not converged **do**
    // Compute gradients
    $G_L \leftarrow L_k + \gamma A^T(\mathscr{P}_\Omega(ALR^T - Y))R$
    $G_R \leftarrow R_k + \gamma(AL)^T(\mathscr{P}_\Omega(ALR^T - Y))$

    // Gradient descent step
    $\tilde{L}_{k+1} \leftarrow L_k - \alpha_L G_L$
    $\tilde{R}_{k+1} \leftarrow R_k - \alpha_R G_R$

    // Project onto constraints
    $L_{k+1} \leftarrow \text{proj}_{\mathscr{C}_L}(\tilde{L}_{k+1})$ ▷ Project onto $\mathscr{C}_L$
    $R_{k+1} \leftarrow \text{proj}_{\mathscr{C}_R}(\tilde{R}_{k+1})$ ▷ Project onto $\mathscr{C}_R$

    // Check convergence
    if $\max(\|L_{k+1} - L_k\|_F, \|R_{k+1} - R_k\|_F) < \epsilon_s$ **then**
        break
    end if
    $k \leftarrow k + 1$
end while
Ensure: L_{k+1}, R_{k+1} satisfying all constraints

To prevent dense reconstruction of the matrix $\mathscr{P}_\Omega(ALR^T - Y)$ when computing the gradients, we implemented a sparse error calculation method leveraging Numba's just-in-time (JIT) compilation and CPU parallelization. Specifically, we compute $ALR^T - Y$ only at the known (observed) indices, Ω, provided by the sparse representation of the matrix Y, to significantly reduce memory consumption and computational complexity. Future strategies to expand to even larger datasets could make use of a distributed framework (see *e.g.*, [34]). However, for our problem we opted for an in-memory approach.

Computational Complexity Analysis The primary computational cost per iteration in Algorithm 1 stems from matrix multiplications and sparse matrix operations involved in the gradient computations. Specifically, the approximation of the product $\mathscr{P}_\Omega(ALR^T - Y)$ occurs in two stages. Initially, the intermediate product AL is calculated with a complexity of $\text{nnz}(A) \cdot t$, where $A \in \mathbb{R}^{p \times r}$ is a sparse matrix and $\text{nnz}(A)$ denotes the number of com-

puted non-zero elements stored in A. Subsequently, each column of R^T is individually multiplied by selected rows from the intermediate product, according to a given sparsity pattern defined by the sparse data structure. This step has a complexity proportional to the number of non-zero elements, specifically $\text{nnz}(Y) \cdot t$. This step is parallelized, but remains a matrix-vector operation (BLAS level 2) in our implementation. Finally, the computed intermediate results are subtracted from the known Y, which is negligible in computational cost. The gradient calculation for L involves computing:

$$G_L \leftarrow L_t + A^T\Big(\mathscr{P}_\Omega(ALR^T - Y)\Big)R,$$

which has two major computational costs. First, computing the residual $\mathscr{P}_\Omega(ALR^T - Y)$ costs $(\text{nnz}(A) + \text{nnz}(Y)) \cdot t$ operations. Then, multiplying this sparse residual by R incurs an additional cost of $\text{nnz}(Y) \cdot t$, and subsequently multiplying by A^T is in the worst case bounded by rpt. An analogous complexity can be derived to compute the gradient G_R. An overview with comparison to an all dense complexity is given in **Table 5.1**. Generally

Table 5.1: Per iteration computational complexity comparison between dense and sparse matrix format of Y and A for computing the gradient of L, $G_L \leftarrow L_t + A^T\Big(\mathscr{P}_\Omega(ALR^T - Y)\Big)R$.

Operation	**Dense**	**Sparse**
Compute residual $\mathscr{P}_\Omega(ALR^T - Y)$	$prt + pqt$	$(\text{nnz}(A) + \text{nnz}(Y)) \cdot t$
Multiply residual by R	pqt	$\text{nnz}(Y) \cdot t$
Multiply result by A^T	rpt	rpt
Summed Complexity	$2pt(r+q)$	$(\text{nnz}(A) + 2\text{nnz}(Y) + rp)t$

speaking, TULIP is of computation complexity $\mathscr{O}(prt)$. This comparison underscores the efficiency of the algorithm when handling large, sparse matrices and highlights its scalability given that the latent dimensions r and t are mostly of modest size relative to the overall problem dimensions p and q. Of course, the algorithm is practically limited by the number of BLAS level 2 operations in the calculation of the error term $\mathscr{P}_\Omega(ALR^T - Y)$, which will be our bottleneck. Finally, note that this is the per iteration complexity.

Memory Complexity Analysis The memory cost for Y is on the order of $\mathscr{O}(\text{nnz}(Y))$. The same $\text{nnz}(Y)$ is true for $\mathscr{P}_\Omega(ALR^T - Y)$. The matrix $A \in \mathbb{R}^{p \times r}$ is a sparse matrix with associated cost $\mathscr{O}(\text{nnz}(A))$. The low-rank factors L and R are defined by the factorization rank t. The memory needed to store these matrices is $\mathscr{O}(rt)$ for L and $\mathscr{O}(qt)$ for R. The same costs are required for the gradients and an additional copy needs to be stored from the previous time step k. Collecting all the storage costs, we have a memory complexity:

$$\mathscr{O}(\text{nnz}(Y) + \text{nnz}(A) + rt + qt).$$

Since we assume that the size t is much smaller than p, q, and r (*i.e.*, $t \ll \min(p,q,r)$) and that the number of nonzeros in Y satisfies $\text{nnz}(Y) \ll pq$, and the $\text{nnz}(A) \ll \text{nnz}(Y)$, the following simplifications hold:

$$\mathscr{O}\Big(\text{nnz}(Y)\Big).$$

This expression indicates that, in our sparse and low-rank setting, the memory required is much lower than that of a dense formulation, which would be of the order $\mathcal{O}(pq)$. This also emphasizes the importance of the efficient exploitation of the projection operator $\mathcal{P}_\Omega$. Practically, we observe a peak memory footprint of ~ 3 times the data footprint of Y, assumed to be represented as a sparse matrix.

Parameter Optimization We have the following parameters in our optimization:

- $\gamma \in \mathbb{R}_+$: defines the trade-off between smoothness of L and R/low-rank property[4] and the data fidelity constraint;
- $t \in \mathbb{N}$: factorization rank, with the intuition to set it higher than the true underlying rank [29], also plays an important role in the convergence rate and computational cost per step (see **Table 5.1**);
- $\sigma \in \mathbb{R}_+$: standard deviation of the noise in the measurements, can be estimated *a priori*;
- $\epsilon_s \in \mathbb{R}_+$: stopping criterion;
- $\lambda_L \in \mathbb{R}_+$: element-wise soft-threshold for L, influences the sparsity pattern of L;
- $\lambda_R \in \mathbb{R}_+$: element-wise soft-threshold for R, influences the sparsity pattern of R.

In practice, the most important parameters to adjust are γ, t, λ_L, and λ_R. For λ_L and λ_R the intuition is straightforward: increasing these parameters leads to increased sparsity in, respectively, L and R. On the other hand, γ and t will play a role in the reconstruction error (see Appendix: Metrics). The Ordinary Least Squares (see Section 5.2.4) solution can usually provide a lower bound for this reconstruction error. We apply data subsampling and a Bayesian parameter optimization over γ, t, λ_L, and λ_R to speed up the parameter tuning. The parameter tuning per case study is presented in Appendix: Parameter Tuning.

BENCHMARKING METHODS

For benchmarking our novel method, we used three "off-the-shelf" methods: Ordinary Least Squares (LS), Non-negative Least Squares (NNLS), and Singular Value Thresholding (SVT) [35]. The ordinary least squares is implemented through an LU decomposition exploiting the sparse matrices [36], while the NNLS is implemented through an alternating direction method of multipliers (ADMM) [37]. Finally, the singular value thresholding is making use of the GESDD singular value decomposition implementation.

- **Ordinary Least Squares (LS):**

$$\min_X \|Y - AX\|_F^2 \tag{5.10}$$

- **Non-Negative Least Squares (NNLS):**

$$\min_X \|Y - AX\|_F^2, \quad \text{subject to } X \geq 0 \tag{5.11}$$

[4] as $\|X\|_* = \min_{X=LR^T} \frac{1}{2}(\|L\|_F^2 + \|R\|_F^2)$

- **Singular Value Thresholding (SVT):**

$$\min_X \|\mathscr{P}_\Omega(Y - AX)\|_F^2 + \lambda\|X\|_* \tag{5.12}$$

A nice property of LS is that it will theoretically provide us a lower bound for both the NNLS, SVT, as well as TULIP, when considering the metric $\|Y - AX\|_F^2$. Note that LS and NNLS do not include completion of missing values, but can still be useful when assuming $Y_{ij} = M_{ij}$, *i.e.*, missing values were assumed to be zero. Reciprocally, for SVT and TULIP, we can assume all values to be measured. As such, any of these methods can be run on either sparse raw (non-feature-selected) data or on dense peak-picked (feature-selected) data. In general, the overall computational complexity of LS will be considered here as $\mathcal{O}(pr^2)$ [38], similar to NNLS as $\mathcal{O}(pr^2)$, and the complexity of the SVT as $\mathcal{O}(pqt)$. All methods, including TULIP, can be used both in overdetermined as well as underdetermined scenarios. However, note that the properties of the solution will differ.

AVAILABILITY

The implementations of the methods used in this chapter, are provided as an open-source Python 3 toolbox at `https://github.com/vandeplaslab/tulip`.

5.3. NUMERICAL RESULTS

The first numerical study provides a ground truth comparison for both under- and overdetermined scenarios, the second study allows us to investigate the algorithm's performance in a large-scale setting. For the first study, synthetic single cell datasets were generated by spatial segmentation of multiplexed immunofluorescence (MxIF) microscopy images. IMS spectra for the single cells were simulated using k-means clustering of the corresponding IMS spectra (obtained after the MxIF imaging) to obtain ground truth spectra. The downsampling of microscopy data to IMS resolutions (0.65 μm to 5 μm)was performed linearly with a Gaussian-weighting. Methods were validated through 100-fold bootstrapping, assessing reconstruction accuracy (*i.e.*, optimization fit), cell fit scores (*i.e.*, ground truth fit), sparsity (*i.e.*, indicator for sparsity in the solution space), clustering quality (*i.e.*, closeness to ground truth classification), and computational efficiency. For further details on the synthetic data, statistical resampling and metrics, we refer the reader to the Appendix. A large-scale FTU dataset (details in Appendix), previously employed in a study [39] and part of the Human BioMolecular Atlas Program (HuBMAP) study of Human Kidney [40], was integrated into our analysis for the second study. The parameter setting for both studies can likewise be found in the Appendix. All calculations were performed on CPU on a Dell Precision 7920 workstation equipped with 56 cores at 2.7 GHz and 1.5 TB of memory.

5.3.1. SYNTHETIC SINGLE CELL

By generating ground-truth spectra through clustering real IMS spectra and mixing those, we do not ensure that they replicate the sparsity patterns inherent to the raw IMS data matrix Y. This is no problem as we can define Ω accordingly and prove the applicability of our methodology on both clipped and non-clipped data. At the same time, note that we do not optimize for X to be sparse, but rather L and/or R for optimization reasons.

OVERDETERMINED LINEAR SYSTEM

In the overdetermined case we modelled the NAR using a single spectrum. We subsampled the raw IMS spectra to create our ground truth spectra and varied the underlying spatial cell distribution by altering the section used of the microscopy image to different spatial locations to construct the mixing matrix A. The results are presented in Table 5.2. We observed that all methods achieved comparable reconstruction errors (*i.e.*, optimization error). TULIP exhibited a slightly higher reconstruction error, which is expected due to the nature of its gradient descent optimization strategy. The cell fit scores (*i.e.*, ground truth fit) revealed that the non-negative constrained methods (NNLS and TULIP) performed approximately 5% better than LS and 10% better than SVT. This suggests that enforcing non-negativity enhances biological interpretability and alignment with known ground truth distributions. Notably, TULIP achieved the highest average cell fit score, implying better recovery of biologically relevant spatial structures despite its slightly elevated reconstruction error. We also observed a substantially larger confidence interval for SVT in both the cell fit and clustering scores. This variability is caused by the problem becoming ill-conditioned for some spatial settings. And although that we regularize both the LS and SVT by a small factor, it seems to affect the SVT more than the LS. Sparsity was observed similar across all methods, as expected, since it was not considered a central concern in this case study. This confirms that none of the methods excessively pruned the solution space under this setup. Clustering scores, while generally decent, showed greater variability, especially for SVT. This metric is merely used here, however, to demonstrate a post-processing capability, rather than a dedicated clustering. Nonetheless, consistent clustering scores across methods support the robustness of spectral decomposition approaches in preserving cell-type distinctions. As such, it messages that a good clustering can be obtained from the estimated spectra provided by the different methods. Finally, run times varied based on both the number of iterations required by each method and their implementation specifics. While this metric only provides a rough estimate, it is evident that LS is the most computationally efficient, followed by NNLS, TULIP, and SVT, which incurs the highest computational cost.

Table 5.2: Comparative Performance Metrics of LS, NNLS, SVT and TULIP methods in an overdetermined linear system setting for a raw, full profile (415 007 features) dataset. Metrics include reconstruction error, cell fit score, non-negativity, sparsity, and clustering score, with computational performance assessed through run times (see Appendix: Metrics). Methodology validated via 100-fold bootstrapping using 10 000 IMS pixels and 500 m/z-bins with Gaussian-weighted mixing (see Appendix: Synthetic Data).

	Performance Metrics - Overdetermined Scenario					
Method	**Recon. Error (%)**	**Cell Fit Score (%)**	**Non-Negativity (%)**	**Sparsity (%)**	**Clustering Score (%)**	**Run Time (s)**
LS	$24.8540^{+4.8861}_{-8.2822}$	$75.0359^{+11.5675}_{-22.6169}$	$90.7511^{+0.9268}_{-0.9540}$	$9.3480^{+0.8589}_{-0.8602}$	$85.2260^{+10.3246}_{-28.2804}$	$0.0434^{+0.3193}_{-0.0226}$
NNLS	$24.8691^{+4.8834}_{-8.2837}$	$80.5795^{+7.6214}_{-24.1341}$	$100.0000^{+0.0000}_{-0.0000}$	$10.0708^{+0.8425}_{-1.0307}$	$91.8206^{+4.0296}_{-24.4886}$	$3.2044^{+0.4922}_{-0.4969}$
SVT	$24.8562^{+4.8910}_{-8.2819}$	$70.5367^{+15.4892}_{-53.4948}$	$90.5926^{+0.8669}_{-0.9769}$	$9.4617^{+0.9449}_{-0.8477}$	$79.1175^{+16.3560}_{-27.0554}$	$7.7211^{+2.0974}_{-1.8037}$
TULIP	$25.3831^{+4.8147}_{-7.7523}$	$80.9436^{+6.3865}_{-28.2552}$	$100.0000^{+0.0000}_{-0.0000}$	$9.1826^{+2.6137}_{-2.4993}$	$87.9596^{+7.3921}_{-29.6487}$	$6.9157^{+0.3808}_{-0.3025}$

By application of a similar bootstrap method (see Appendix) on the LS data, uncertainty intervals were obtained for the reconstruction of individual cell spectra (see **Fig.**

5.2). We observe there that the confidence interval is small with respect to the relative intensities for this particular class. In practice, this might change from cell to cell and depends heavily on the setting, *e.g.* conditioning of the A matrix. Note, that this uncertainty is related to the mixing process, other sources of uncertainties are not considered, but could impact the overall uncertainty.

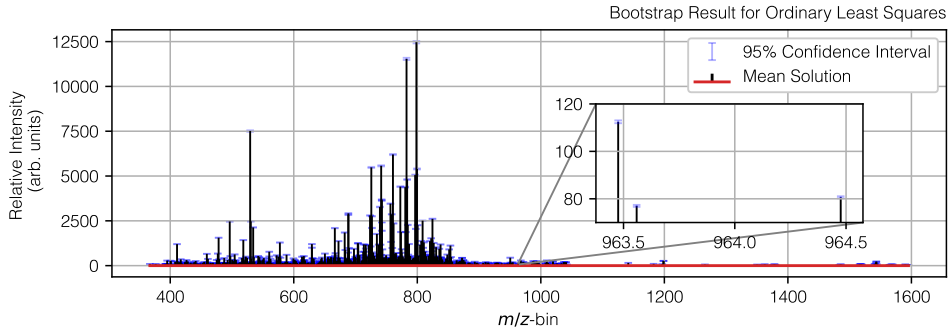

Figure 5.2: Bootstrap results for ordinary least squares estimates in the overdetermined scenario, showing the result for an individual retrieved ROI, *i.e.*, cell, spectrum and its corresponding confidence interval. One observes that the confidence interval is relatively small with respect to the relative intensities for this particular single cell, even in the inset example on the right. This will mainly depend on the conditioning of the A matrix. Similar plots are obtainable for the other methods.

5

UNDERDETERMINED LINEAR SYSTEM

In the underdetermined case (see **Table** 5.3), where each NAR is modelled as an individual spectrum, all methods achieved a low reconstruction error overall as well. As there are more degrees-of-freedom to the model, the reconstruction error is remarkably smaller than the overdetermined linear system. TULIP has slighly higher reconstruction error, as previously observed in the overdetermined case. This is thought to originate from slow convergence in the neighborhood of its local minimum. In terms of cell fit score NNLS performs best, suggesting improved biological relevance. As in the overdetermined case, non-negativity appears crucial: TULIP and NNLS strictly enforces this constraint, unlike LS and SVT, and this likely contributes to a superior interpretability. The sparsity is, however, notably higher in LS and SVT, and consistently lower for TULIP. However, since the latter is mainly driven by parameter tuning, it has the potential to be improved at the cost of a higher reconstruction error. Clustering scores were comparable, with NNLS slightly ahead, yet TULIP remained robust across trials. While TULIP incurred substantially longer run times—2 to 6 times slower than SVT, this trade-off is balanced by its gradient-based nature and better scalability in high-dimensional settings without relying on the SVD. By application of a similar bootstrap method for TULIP, uncertainty intervals are obtained for the individual cell spectra (see **Fig.** 5.3).

5.3.2. HUBMAP STUDY ON HUMAN KIDNEY (FTU)

To demonstrate the scalability of our algorithm and validate it using non-synthetic data, we conducted a numerical experiment on an IMS dataset [39], $Y \in \mathbb{R}^{869\,851 \times 507\,429}$, stored

Table 5.3: Comparative Performance Metrics of LS, NNLS, SVT and TULIP methods in an underdetermined linear system setting for a raw, full profile (415 007 features) dataset. Metrics include reconstruction error, cell fit score, non-negativity, sparsity, and clustering score, with computational performance assessed through run times (see Appendix: Metrics). Methodology validated via 100-fold bootstrapping using 10 000 IMS pixels and 500 *m/z*-bins with Gaussian-weighted mixing (see Appendix: Synthetic Data.

	Performance Metrics - Underdetermined Scenario					
Method	**Recon. Error (%)**	**Cell Fit Score (%)**	**Non-Negativity (%)**	**Sparsity (%)**	**Clustering Score (%)**	**Run Time (s)**
LS	$0.2771^{+0.0916}_{-0.0545}$	$86.6961^{+4.6787}_{-9.6768}$	$83.8352^{+2.2559}_{-2.4131}$	$18.5638^{+2.5923}_{-2.5577}$	$92.8343^{+3.0519}_{-9.6459}$	$0.0705^{+0.2082}_{-0.0233}$
NNLS	$0.2406^{+0.1021}_{-0.0898}$	$88.4062^{+4.3031}_{-10.7151}$	$100.0000^{+0.0000}_{-0.0000}$	$12.1890^{+1.5724}_{-1.5943}$	$94.0690^{+2.5314}_{-4.4621}$	$12.9890^{+4.1889}_{-2.3521}$
SVT	$0.0057^{+0.0036}_{-0.0024}$	$86.8179^{+4.6750}_{-9.7444}$	$82.8575^{+2.2767}_{-2.4731}$	$18.4844^{+2.9039}_{-2.7507}$	$93.1346^{+2.9369}_{-9.8467}$	$3.6418^{+0.9488}_{-0.4707}$
TULIP	$2.0967^{+0.2924}_{-0.2380}$	$85.2511^{+4.6268}_{-9.8960}$	$100.0000^{+0.0000}_{-0.0000}$	$7.4594^{+2.5942}_{-2.1238}$	$89.5560^{+5.5831}_{-4.5504}$	$86.7397^{+27.8654}_{-6.2593}$

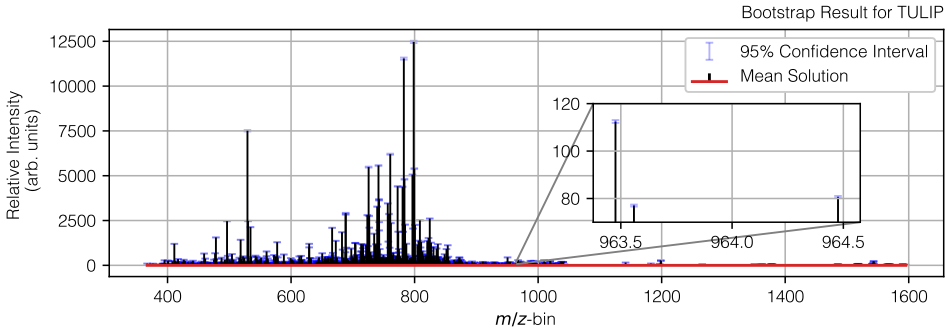

Figure 5.3: Bootstrap results for TULIP in the underdetermined scenario, showing the result for an individual retrieved cell spectrum and its corresponding confidence interval. One observes that the confidence interval is relatively small with respect to the relative intensities for this particular single cell, even in the inset example on the right, and comparable to LS, see **Fig. 5.2**. This will mainly depend on the conditioning of the A matrix. Similar plots are obtainable for the other methods.

in Compressed Sparse Column (CSC) format with a memory footprint of 124.24 GB (equivalent to 1 765.55 GB in dense format). The associated mixing matrix $A \in \mathbb{R}^{869\,851 \times (6+503\,926)}$ is also stored in CSC format, occupying 0.014 GB (or 1 753.39 GB in dense format). This setup includes 6 ROIs, corresponding to different clusters (*i.e.*, different cell types inside the kidney's glomerulus). A visualization of a single glomerulus is provided in **Fig. 5.4**, and individual segment/ROI names are listed in **Table 5.4**, alongside 503 926 NARs.

Our implementations of LS and NNLS require an explicit computation of $A^T A$, making them infeasible for this dataset. Additionally, the computational complexity of SVT becomes prohibitive and would necessitate either an out-of-memory or a randomized strategy, as previously suggested [20]. SVT also requires computing the pseudo-inverse of the matrix A, which is similarly impractical under these conditions and therefore not considered here.

However, TULIP remains applicable: with an average runtime of 37 minutes per iteration over 100 iterations, the total computation time amounted to approximately 60

hours to obtain the following results.

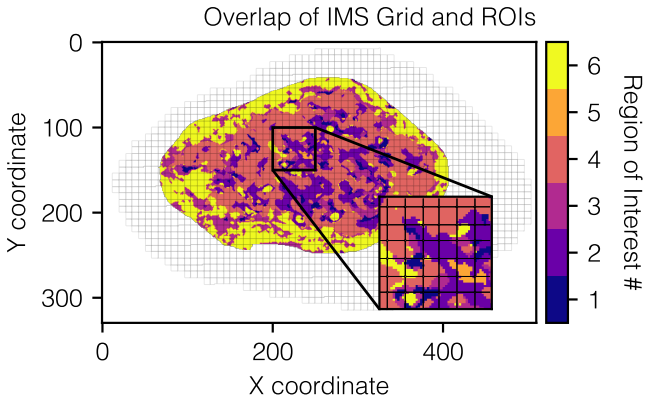

Figure 5.4: Example of a single glomerulus with highlighted glomerular segments. Observe that the IMS pixel grid overlaps also outside the glomerulus, this padding was added to ensure that the whole glomerulus was covered. The total dataset consists of dozens of these glomeruli. See **Table 5.4** for the glomerular segment names.

Table 5.4: Glomerular segment naming as seen in **Fig. 5.4**. The segments are similar to a previous study [39].

ROI #	Glomerular Segment
1	Mesangial cells
2	Endothelial cells
3	Non-cell specific cluster
4	Podocytes
5	Mesangial matrix
6	Zero/low immunofluorescence staining

QUANTITATIVE ANALYSIS

To validate our approach, we compared results with a prior study [39], in which IMS pixels associated with a single cluster (*i.e.*, no mixing) were averaged per cluster to generate ROI spectra. While one could argue that obtaining spectra in this manner is computationally cheaper, our method offers a scalable alternative suitable for more challenging environments. As such, this analysis validates both the accuracy and scalability of our approach.

Table 5.5 reports the reconstruction error (see Appendix: Metrics for more information) across all pixels, as well as for pure (single cluster) and mixed (multiple clusters or

cluster/NAR) pixel categories. TULIP significantly reduces the reconstruction error in all categories, with the most substantial improvement observed in the "All" group (from 80.09% to 22.39%), demonstrating advanced overall reconstruction performance. The improvement for "Pure" pixels is smaller, as expected, since these cases are simpler to model (see **Fig. S5.4** for the reconstruction score per cluster). Additionally, the spectra estimated by TULIP closely match the averaged spectra obtained from pure pixels, lending empirical support to the validity of our method. These average spectra may also serve as effective initializations for TULIP, potentially improving convergence and stability. Finally, substantial gains are observed in the "Mixed" category, further highlighting TULIP's robustness in handling spectral mixtures. In general, the overall reconstruction error is believed to be constrained by the expressiveness of the mixing matrix A, *i.e.* how well A can represent the possible values of Y through linear combinations of X. Future work may explore models that explicitly incorporate ΔA to further improve reconstruction accuracy.

Table 5.5: Reconstruction error comparison between the results used in a previous study [39] and TULIP. For "All", all pixels (*i.e.*, 869 851) are considered, for "Pure", only pixels that have a single underlying cluster (*i.e.*, 66 367) are used and for "Mixed", only pixels having a mix of underlying clusters and clusters and NARs (*i.e.* 317 565) are used.

Method	**Reconstruction Error (%)**		
	All	**Pure**	**Mixed**
Previous Study [39]	80.09	34.89	33.18
TULIP	22.71	34.43	29.84

QUALITATIVE ANALYSIS

The spectrum for each individual cluster is shown in **Fig. S5.2**. While these spectra appear visually similar, a more detailed comparison using the mass spectral difference plots, presented in **Fig. 5.5**, reveals distinct differences between the spectra of different clusters. This figure is analogous to **Fig. S5.3**, which was obtained using a procedure similar to that of the previous study [39] (and also mentioned in that study), and thus further supports the validity of our approach.

5.4. CONCLUSION

This work introduces a novel computational framework that leverages microscopy-informed constraints to overcome a key challenge in IMS: sampling patterns mixing underlying biological structures. We provide a strategy to overcome this limitation in the specificity of molecular assignments at the regions of interest level, particularly interesting in spatially heterogeneous contexts. By formulating spectral unmixing as an inverse problem, our approach effectively disentangles mixed measurements from IMS pixels into the underlying molecular spectra of distinct biological regions, ranging from single cells to larger functional tissue units. Through the integration of biologically meaningful

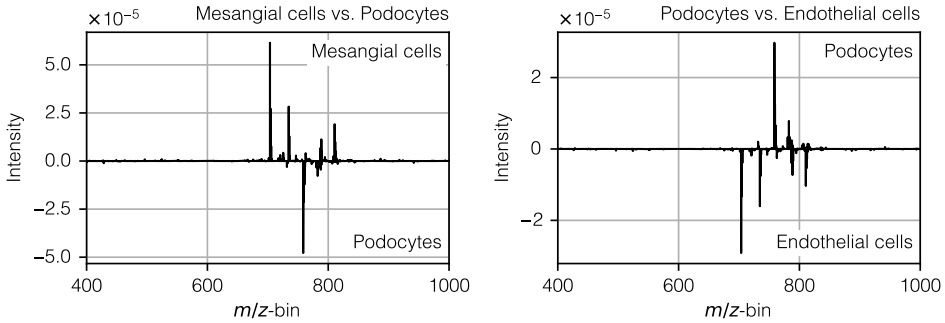

Figure 5.5: Mass spectral difference (of total sum normalized spectra) plots illustrating ion intensity differences between ROI #1 (primarily mesangial cells) and ROI #4 (primarily podocytes) (left), and between ROI #4 and ROI #2 (primarily endothelial cells) (right) as obtained by TULIP. These results highlight that each glomerular segment exhibits a distinct profile of IMS-detected molecular species. Comparing these results to **Fig. S5.3** generated with a procedure identical to the previous study [39], we observe that most peaks are further validating our approach.

constraints such as non-negativity, sparsity, and low-rank assumptions, a specifically tailored algorithm, TULIP, offers a robust strategy to address both overdetermined and underdetermined scenarios. Comparative studies against established methods, including Ordinary Least Squares (LS), Non-negative Least Squares (NNLS), and Singular Value Thresholding (SVT), demonstrate that, while each method has its merits, TULIP consistently achieves superior cell fit scores and enforces non-negativity in the recovered spectra for the overdetermined case. At the same time, it reports competitive scores and errors to these methods in the underdetermined scenario. These advantages are particularly pronounced in complex settings involving full profile data and significant non-annotated contributions. Furthermore, the evaluation on a challenging large-scale dataset highlights the method's scalability and its potential to yield biologically interpretable results. Overall, our results underscore the benefit of integrating high-resolution microscopy information to guide the unmixing process in IMS, thereby enhancing molecular specificity and interpretability in tissue imaging. Future work will focus on further refining the parameter tuning process, extending the framework to additional imaging modalities, and addressing challenges posed by expressiveness of the mixing matrix. Ultimately, TULIP paves the way for more detailed *in-situ* analyses of heterogeneous biological samples, opening new avenues for research in cellular and tissue research.

REFERENCES

[1] R. M. Heeren, D. F. Smith, J. Stauber, B. Kükrer-Kaletas, and L. MacAleese. Imaging Mass Spectrometry: Hype or Hope? In: *Journal of the American Society for Mass Spectrometry* 20.6 (2009), pp. 1006–1014.

[2] R. M. Caprioli. Imaging Mass Spectrometry: a Perspective. In: *Journal of Biomolecular Techniques: JBT* 30.1 (2019), p. 7.

[3] L. Li, R. W. Garden, and J. V. Sweedler. Single-Cell MALDI: A New Tool for Direct Peptide Profiling. In: *Trends in Biotechnology* 18.4 (2000), pp. 151–160.

[4] A. R. Buchberger, K. DeLaney, J. Johnson, and L. Li. Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. In: *Analytical Chemistry* 90.1 (2017), p. 240.

[5] A. Körber, I. G. Anthony, and R. M. Heeren. Mass Spectrometry Imaging. In: *Analytical Chemistry* (2025).

[6] K. V. Djambazova, J. M. Van Ardenne, and J. M. Spraggins. Advances in Imaging Mass Spectrometry for Biomedical and Clinical Research. In: *TrAC Trends in Analytical Chemistry* 169 (2023), p. 117344.

[7] M. Aichler and A. Walch. MALDI Imaging Mass Spectrometry: Current Frontiers and Perspectives in Pathology Research and Practice. In: *Laboratory Investigation* 95.4 (2015), pp. 422–431.

[8] B. Balluff, M. Hanselmann, and R. Heeren. Mass Spectrometry Imaging for the Investigation of Intratumor Heterogeneity. In: *Advances in Cancer Research* 134 (2017), pp. 201–230.

[9] D. Miura, Y. Fujimura, and H. Wariishi. In Situ Metabolomic Mass Spectrometry Imaging: Recent Advances and Difficulties. In: *Journal of Proteomics* 75.16 (2012), pp. 5052–5060.

[10] K. Ščupáková, F. Dewez, A. K. Walch, R. M. Heeren, and B. Balluff. Morphometric Cell Classification for Single-Cell MALDI-Mass Spectrometry Imaging. In: *Angewandte Chemie* 132.40 (2020), pp. 17600–17603.

[11] M. Nijs, T. Smets, E. Waelkens, and B. De Moor. A Mathematical Comparison of Non-negative Matrix Factorization Related Methods with Practical Implications for the Analysis of Mass Spectrometry Imaging Data. In: *Rapid Communications in Mass Spectrometry* 35.21 (2021), e9181.

[12] P.-L. Delacour, S. Wahls, J. M. Spraggins, L. Migas, and R. Van de Plas. Signal Recovery Using a Spiked Mixture Model. In: *IEEE Transactions on Signal Processing* (2025), pp. 1–14.

[13] A. Maimó-Barceló, J. Garate, J. Bestard-Escalas, R. Fernández, L. Berthold, D. H. Lopez, J. A. Fernández, and G. Barceló-Coblijn. Confirmation of Sub-Cellular Resolution Using Oversampling Imaging Mass Spectrometry. In: *Analytical and Bioanalytical Chemistry* 411 (2019), pp. 7935–7941.

[14] A. Zavalin, E. M. Todd, P. D. Rawhouser, J. Yang, J. L. Norris, and R. M. Caprioli. Direct Imaging of Single Cells and Tissue at Sub-Cellular Spatial Resolution Using Transmission Geometry MALDI MS. In: *Journal of Mass Spectrometry* 47.11 (2012), pp. 1473–1481.

[15] R. S. Young, A.-K. Piper, L. McAlary, J. C. McKinnon, J. S. Lum, J. Soltwisch, M. Niehaus, and S. R. Ellis. Subcellular Mass Spectrometry Imaging of Lipids and Nucleotides Using Transmission Geometry Ambient Laser Desorption and Plasma Ionisation. In: *bioRxiv preprint bioRxiv:2025.05.13.653655* (2025).

[16] T.-H. Ong, D. J. Kissick, E. T. Jansson, T. J. Comi, E. V. Romanova, S. S. Rubakhin, and J. V. Sweedler. Classification of Large Cellular Populations and Discovery of Rare Cells Using Single Cell Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry. In: *Analytical Chemistry* 87.14 (2015), pp. 7036–7042.

[17] E. T. Jansson, T. J. Comi, S. S. Rubakhin, and J. V. Sweedler. Single Cell Peptide Heterogeneity of Rat Islets of Langerhans. In: *ACS Chemical Biology* 11.9 (2016), pp. 2588–2595.

[18] Y. R. Xie, V. K. Chari, D. C. Castro, R. Grant, S. S. Rubakhin, and J. V. Sweedler. Data-Driven and Machine Learning-Based Framework for Image-Guided Single-Cell Mass Spectrometry. In: *Journal of Proteome Research* 22.2 (2023), pp. 491–500.

[19] L. Rappez, M. Stadler, S. Triana, R. M. Gathungu, K. Ovchinnikova, P. Phapale, M. Heikenwalder, and T. Alexandrov. SpaceM Reveals Metabolic States of Single Cells. In: *Nature Methods* 18.7 (2021), pp. 799–805.

[20] R. A. Moens, L. G. Migas, J. M. Van Ardenne, E. P. Skaar, J. M. Spraggins, and R. Van de Plas. Preserving Full Spectrum Information in Imaging Mass Spectrometry Data Reduction. In: *Bioinformatics* 41.5 (2025), btaf247.

[21] F. A. Battjes, K. Olling, R. Van de Plas, and R. A. R. Moens. Enforcing Physical Constraints in Full Spectrum Imaging Mass Spectrometry Data Reduction. In: *[Unpublished manuscript]* (2025).

[22] M. Fazel. "Matrix Rank Minimization with Applications". PhD thesis. PhD thesis, Stanford University, 2002.

[23] N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin Matrix Factorization. In: *Advances in neural information processing systems* 17 (2004).

[24] S. Burer and R. D. Monteiro. A Nonlinear Programming Algorithm for Solving Semidefinite Programs Via Low-Rank Factorization. In: *Mathematical Programming* 95.2 (2003), pp. 329–357.

[25] O. V. Thanh and N. Gillis. Minimum-Volume Nonnegative Matrix Completion. In: *2024 32nd European Signal Processing Conference (EUSIPCO)*. IEEE. 2024, pp. 2452–2456.

[26] L. Taslaman and B. Nilsson. A Framework for Regularized Non-Negative Matrix Factorization, With Application to the Analysis of Gene Expression Data. In: *PLOS ONE* 7.11 (2012), e46331.

[27] N. Gillis. Introduction to Nonnegative Matrix Factorization. In: *arXiv Preprint arXiv:1703.00663* (2017).

[28] H. Kim and H. Park. Sparse Non-negative Matrix Factorizations Via Alternating Non-negativity-Constrained Least Squares for Microarray Data Analysis. In: *Bioinformatics* 23.12 (2007), pp. 1495–1502.

[29] R. Ge, J. D. Lee, and T. Ma. Matrix Completion Has No Spurious Local Minimum. In: *Advances in Neural Information Processing Systems* 29 (2016).

[30] A. Cichocki, R. Zdunek, and S.-i. Amari. Nonnegative Matrix and Tensor Factorization. In: *IEEE Signal Processing Magazine* 25.1 (2007), pp. 142–145.

[31] P. O. Hoyer. Non-negative Matrix Factorization With Sparseness Constraints. In: *Journal of machine learning research* 5.Nov (2004), pp. 1457–1469.

[32] V. P. Pauca, J. Piper, and R. J. Plemmons. Nonnegative Matrix Factorization For Spectral Data AAnalysis. In: *Linear algebra and its applications* 416.1 (2006), pp. 29–47.

[33] S. J. Reddi, S. Kale, and S. Kumar. On the Convergence of Adam and Beyond. In: *arXiv Preprint arXiv:1904.09237* (2019).

[34] S. Eswar, K. Hayashi, G. Ballard, R. Kannan, M. A. Matheson, and H. Park. PLANC: Parallel Low-Rank Approximation with Nonnegativity Constraints. In: *ACM Transactions on Mathematical Software (TOMS)* 47.3 (2021), pp. 1–37.

[35] J.-F. Cai, E. J. Candès, and Z. Shen. A Singular Value Thresholding Algorithm for Matrix Completion. In: *SIAM Journal on Optimization* 20.4 (2010), pp. 1956–1982.

[36] J. W. Demmel. *SuperLU Users' Guide, Version 2.0*. Tech. rep. UCB/ERL M99/09. University of California at Berkeley, 1999.

[37] C. Zheng, M. Yu, J. Shan, A. Wang, and H. Chen. Fast Sparse Non-negative Least Squares Via ADMM for High-Resolution DOA Estimation. In: *IEEE Sensors Journal* 23.4 (2023), pp. 3901–3910.

[38] G. H. Golub and C. F. Van Loan. *Matrix Computations*. JHU Press, 2013.

[39] A. B. Esselman, F. A. Moser, L. E. Tideman, L. G. Migas, K. V. Djambazova, M. E. Colley, E. L. Pingry, N. H. Patterson, M. A. Farrow, H. Yang, et al. In Situ Molecular Profiles of Glomerular Cells By Integrated Imaging Mass Spectrometry and Multiplexed Immunofluorescence Microscopy. In: *Kidney International* 107.2 (2025), pp. 332–337.

[40] S. Jain, L. Pei, J. M. Spraggins, M. Angelo, J. P. Carson, N. Gehlenborg, F. Ginty, J. P. Gonçalves, J. S. Hagood, J. W. Hickey, et al. Advances and Prospects for the Human BioMolecular Atlas Program (HuBMAP). In: *Nature Cell Biology* 25.8 (2023), pp. 1089–1100.

SUPPLEMENTARY MATERIALS

SYNTHETIC DATA GENERATION

The procedure to generate synthetic data is not specific to single cell IMS data. As long as regions of interest are annotated in microscopy or other modality data of high-spatial resolution, this procedure can be used. It comprises three steps:

1. **Spatial View Generation:** A TIFF image containing the high-resolution image with segmented/clustered/classified regions of interest, is processed using the `SPATIAL` module of our `TULIP` toolbox. The TIFF image can be cropped spatially to a certain view and cells that are annotated within this view receive a unique integer IDs. Small cells, *i.e.* only contained in a small number of high-resolution pixels, below a cutoff threshold (*e.g.*, `cut_off`=50 pixels) are removed to reduce the condition number, and the remaining non-cellular regions are designated as non-annotated regions (NARs). Most of the time this cutoff is set to get rid of *e.g.* regions that fall partly inside the view, but only by a few pixels. Finally, the NAR can be subdivided into smaller regions or maintained as a single region.

2. **Spectral View Generation:** Using IMS data, the `SPECTRAL` module extracts spectral signatures by clustering all the spectra. For the cellular component, the spectra are clustered with traditional k-means clustering into a fixed number of classes (*e.g.*, `n_clus`=10), while for the background (NAR) a single average spectrum is computed. Therefore, we use only IMS spectra that are fully spanning the background, leaving "mixed" pixels aside. While this might generate a bias towards large cell spectra, we do not consider this aspect as a problem for validation of our method. Another option, is to provide a fixed number of cluster spectra and their probability of appearance, *e.g.* relative cluster sizes.

3. **Synthetic Data Mixing:** The `MIX` module links the spatial and spectral information. Spectral classes are assigned to each cell and NAR via random sampling and a sparse linking matrix. Therefore, we keep the same clustering distribution, *i.e.* keep corresponding cluster sizes. The spatial view is then downsampled (by a factor, *e.g.*, `scale`=16) to mimic the lower resolution of IMS with respect to the microscopy modality used. Optionally, Gaussian-weighted downsampling or other position-specific weights can be applied to emphasize central pixels (at microscopy level) more than pixels (at microscopy level) near the border of the IMS pixel.

Table 5.6 summarizes the key parameters and their roles in the synthetic data generation pipeline. A visual overview of the generation is given in **Fig. S5.1**.

For our synthetic dataset, we set `cut_off` at 50 pixels, `n_clus (cell)` at 10, `n_clus (NAR)` at 4, `scale` at 16, `dtype` at Gaussian with `sigma` at 4, and no additional `weights`.

Table 5.6: Key parameters for synthetic data generation of a synthetic single cell dataset.

Parameter	Description
`cut_off`	Minimum cell size threshold (pixels)
`n_clus (cell)`	Number of clusters for cell spectra
`n_clus (NAR)`	Number of clusters for NAR spectra
`scale`	Downsampling factor (microscopy-to-IMS)
`dtype`	Downsampling type (linear/gaussian)
`sigma`	Standard deviation for Gaussian kernel
`weights`	Positional weights for weighted downsampling

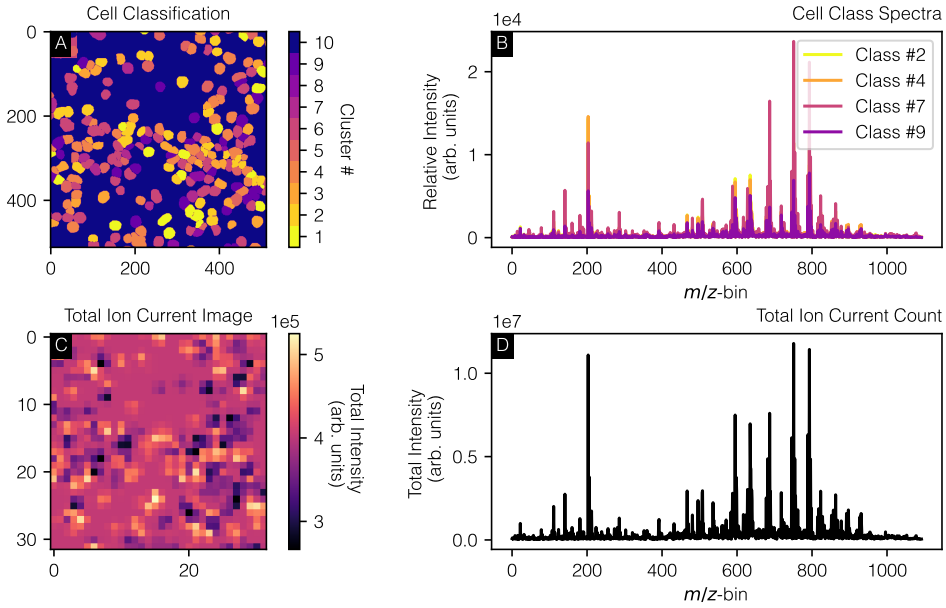

Figure S5.1: Synthetic imaging mass spectrometry (IMS) data generation pipeline. (A) A microscopy image is processed to create a spatial map, annotating individual cells and non-annotated regions (NARs), and a cell class is assigned to each cell based on the clustering distribution. (B) Spectral signatures extracted from real IMS data are clustered to define cell classes and a background (NAR) class. These spatial and spectral components are combined through a mixing step, assigning spectral profiles to each spatial region and applying spatial downsampling (linear) to simulate realistic IMS pixel sizes. The resulting synthetic data (C, D) facilitates validation of our unmixing methods for single-cell IMS.

STATISTICAL RESAMPLING METHOD

To obtain robust estimates of performance metrics and corresponding confidence intervals, we employed a separate bootstrapping procedure for the metrics as well as the spectra.

METRICS

Specifically, we performed random subsampling (sampling without replacement) of rows and columns from the original data array. In each iteration, a subset of rows was randomly selected, ensuring variability in the input data across repetitions. We used a fixed subset of columns for all iterations. The standard subsample sizes were set to 10 000 rows and 500 columns per iteration, balancing computational efficiency with statistical robustness. For each subsample, TULIP and the benchmarking methods were applied, and several evaluation metrics were calculated. By repeating this process systematically over multiple iterations, empirical distributions for each metric were obtained. Confidence intervals were subsequently derived using the 0.025 and 0.975 quantiles (CI-0.025 and CI-0.975) to define the lower and upper bounds, respectively, and we report this variation with respect to the median. This method closely aligns with the standard non-parametric bootstrap technique [1, 2], but differs slightly in the use of subsampling rather than full resampling with replacement.

SPECTRA

For the generation of confidence intervals on the spectra, bootstrap replicates were generated by resampling the original observations with replacement, thereby preserving the underlying distributional properties of the dataset. Specifically, each bootstrap iteration involved the following two steps:

1. Random selection (with replacement) of indices corresponding to the rows of matrices A and Y.

2. Computation of the solution for each bootstrap sample.

This process was repeated for 10 iterations, resulting in a distribution of spectral estimates. The mean and standard deviation across bootstrap replicates provided measures of central tendency and variability, respectively. Confidence intervals were constructed using the normal approximation method, calculated as:

$$\mathrm{CI}_{95\%} = \bar{\theta} \pm 1.96 \times \hat{\sigma}_\theta \tag{5.13}$$

where $\bar{\theta}$ denotes the mean of bootstrap solutions, and $\hat{\sigma}_\theta$ is their standard deviation.

HUBMAP STUDY ON HUMAN KIDNEY (FTU)

The methods (mainly wet-lab specifics) can be found in the original paper [3] and an extended methods section are provided in the supplemental material of the same paper.

PARAMETER TUNING

Here, we describe the parameter tuning details per case study. The parameter for NNLS is set heuristically at $\rho = 1$ [4] and number of iterations to 500, and for SVT we set $\delta = 1.2\frac{qp}{\mathrm{nnz}(Y)}$ and $\tau = q$, consistent to advised parameter setting, *e.g.* in [5]. The stopping criteria for SVT are set to be a reconstruction error of 10^{-4} or 1 000 iterations. To tackle ill-conditioning issues with LS, we use a small regularization term of $\epsilon = 0.001$, that is used in the normal equation as $A^T A + \epsilon I$. For TULIP, we further optimize the parameters

of the Adam optimizer and apply an exponential decay on the learning rate $\alpha = \alpha_0 e^{-r\,k}$, where α_0 is specified below, k is the iteration counter and r is a heuristically set decay rate by calculating the halving time to be around $600-700$ steps, leading to $k = 0.001$. Finally, in the first case study we initialize L as a random matrix, with entries sampled from a uniform distribution between 0 and 1 and R through a random *Xcol* initialization [6]. In the second case study, we initialize both L and R as random matrices, with entries sampled from a uniform distribution between 0 and 1 and R.

CASE STUDY 1: OVERDETERMINED SCENARIO

We set the total number of iterations to 300 and the number of restarts at 20.

Adam Optimizer $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\alpha_0 = 2$.

Synthetic Single Cell For the synthetic single cell data, the true underlying rank of the problem is known *a priori*, namely 14, hence we fix the rank to that value. For the other parameters, *i.e.* λ_L, λ_R and, γ we used a Bayesian optimization with Gaussian process based minimization through `scikit-optimize`. We use a Gaussian kernel and 100 function evaluations. Therefore, we optimize for

$$\theta = \phi_1 + \zeta\left(100 - \bar{\phi}_4(L,R)\right),$$

where ϕ_1 is the reconstruction score (see Appendix: Metrics), $\zeta \in \mathbb{R}$ is a parameter that trades-off sparsity for reconstruction error, and $\bar{\phi}_4(L,R)$ is defined by

$$\bar{\phi}_4(L,R) = \frac{\phi_4(L) + \phi_4(R)}{2},$$

where ϕ_4 is the sparsity metric (see Appendix: Metrics). We set $\lambda_L = [0,1]$, $\lambda_R = [0,1]$, $\gamma = [0.1, 1\,000]$, $\zeta = 0.1$ and optimize θ. We found the following optimal parameters for peak picked data: $\lambda_L^* = 0.1127$, $\lambda_R^* = 0.1380$ and $\lambda^* = 119.0036$. These values were obtained on a separate training set consisting of a different spatial location on the microscopy lay out and a different set of spectra.

CASE STUDY 1: UNDERDETERMINED SCENARIO

We set the total number of iterations to 300 and the number of restarts to 20.

Adam Optimizer $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\alpha_0 = 2$.

Synthetic Single Cell For the synthetic single cell data, the true underlying rank of the problem is known *a priori*, namely 14 (10 different cell classes and 4 different background classes), hence we fix the rank to that value. For the other parameters, *i.e.* λ_L, λ_R and, γ we used the same Bayesian optimization with Gaussian process as in the previous scenario. We found the following optimal parameters for peak picked data: $\lambda_L^* = 0$, $\lambda_R^* = 0.4211$ and $\lambda^* = 1\,000.000$. For λ^*, we noted that increasing the upper bound of its range did not change results, hence, we kept this boundary value. These values were obtained on a separate training set consisting of a different spatial location on the microscopy lay-out and different set of spectra.

HUBMAP STUDY ON HUMAN KIDNEY (FTU)

For the HuBMAP data, we use a similar method as for the synthetic single cell. However, the rank is here not known *a priori*, and thus becomes an extra parameter to tune. Secondly, given the size of the problem, we train on a single glomerulus (this limits the number of pixels to be below 10 000) and also subsample the spectral axis by randomly selecting 500 *m/z*-bins. These *m/z*-bins were manually checked to represent high and low intensity peaks. We employ the same Bayesian optimization strategy and obtain the following optimal parameters: $t^* = 100$, $\lambda_L^* = 0$, $\lambda_R^* = 0.3147$ and $\lambda^* = 0.001$. In this case study, we again hit the limits of the search space values for t. Heuristically, we found that increasing the rank further only slows the process of finding solutions and does not decrease the reconstruction error substantially. We limited the number of iterations to 100 as on average a step takes around 37 minutes to complete.

Adam Optimizer $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\alpha_0 = 2$ and a decay rate $k = 0.01$.

METRICS

Here, we describe the metrics used for evaluation.

5

RECONSTRUCTION ERROR

The reconstruction score (ϕ_1) measures the accuracy of the unmixing reconstruction, calculated as the relative error between the reconstructed signal and the original mixed signal, expressed as a percentage:

$$\phi_1 = 100\% \times \frac{\|Y - AX\|_F}{\|Y\|_F},$$

where Y contains the IMS spectra, A is the mixing matrix, and X contains the individual estimated spectra of the ROI. In case we are dealing with missing values, we consider the reconstruction error to be defined as:

$$\phi_1 = 100\% \times \frac{\|\mathscr{P}_\Omega(Y - AX)\|_F}{\|\mathscr{P}_\Omega(Y)\|_F}.$$

CELL FIT SCORE

Fit quality measures how accurately the cell spectra are reconstructed. Specifically, the cell fit percentage is calculated using the Frobenius norm:

$$\phi_2 = 100\% - 100\% \times \frac{\|X - \hat{X}\|_F}{\|X\|_F},$$

where X contains the ground truth cell spectra and $\hat{X}$ represents the estimated spectra from unmixing.

NON-NEGATIVITY

This metric reports the fraction of non-negative entries in the unmixed spectra matrix X, expressed as a percentage:

$$\phi_3 = 100\% \times \frac{\text{number of entries } \geq 0}{\text{total number of entries}}.$$

SPARSITY

Sparsity measures the proportion of nearly-zero entries (less than 2ε, where ε is the machine precision for float32 format) in the estimated unmixed spectra, representing how sparse the recovered solution is:

$$\phi_4 = 100\% \times \frac{\text{number of entries } < 2\varepsilon}{\text{total number of entries}}.$$

CLUSTERING SCORE

This metric evaluates the accuracy of clustering after unmixing, calculated as the percentage of correctly matched clusters:

$$\phi_5 = 100\% \times \frac{\text{number of correctly matched clusters}}{\text{total number of clusters}}.$$

A simple k-means clustering is applied, from the `sklearn.cluster` module, with the known number of clusters, *i.e.* cell types. We then look at the number of matches of closest clusters making use of the "true" cluster centres and memberships.

RUN TIME

The run time is the duration for running of the algorithm. It is given as:

$$\phi_6 = t_{\text{end}} - t_{\text{start}}.$$

HUBMAP STUDY ON HUMAN KIDNEY (FTU) - ADDITIONAL FIGURES

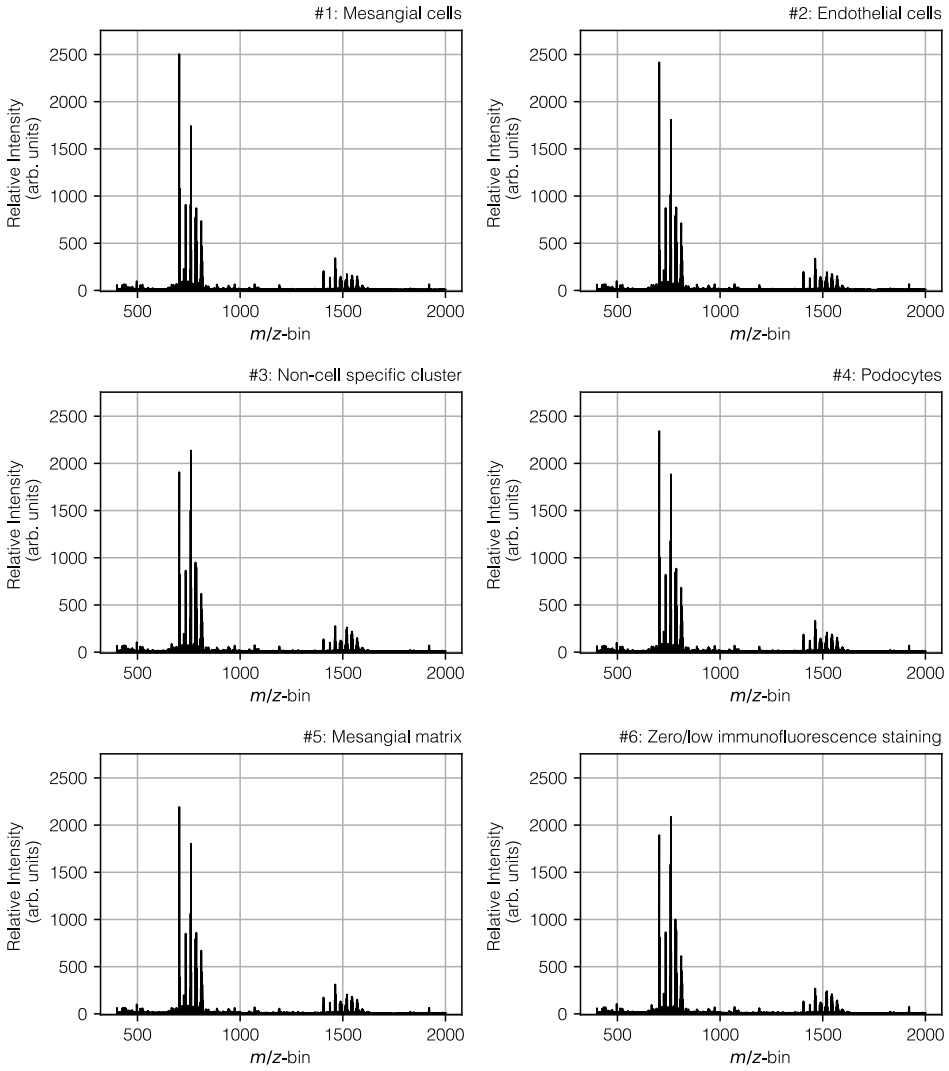

Figure S5.2: Estimated spectra for the different glomerular segments. While visually not immediately apparent, all spectra show distinct patterns, *e.g.*, in the region between 500 and 1 000, we can observe different peaks and intensities.

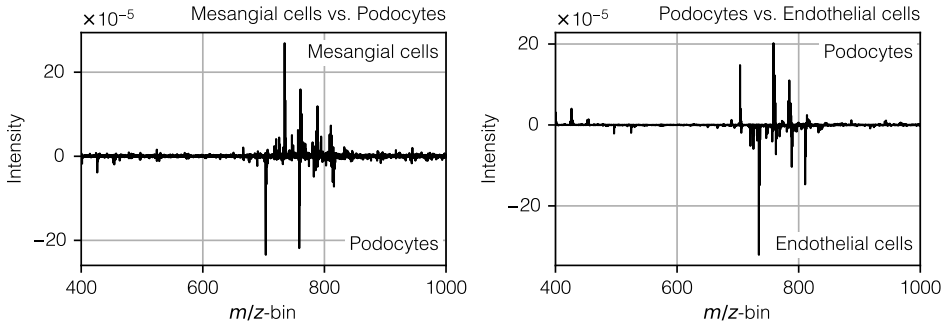

Figure S5.3: Mass spectral difference (of total sum normalized spectra) plots illustrating ion intensity differences between ROI #1 (primarily mesangial cells) and ROI #4 (primarily podocytes) (left), and between ROI #4 and ROI #2 (primarily endothelial cells) (right) as obtained in a previous study [3]. These results highlight that each glomerular segment exhibits a distinct profile of IMS-detected molecular species.

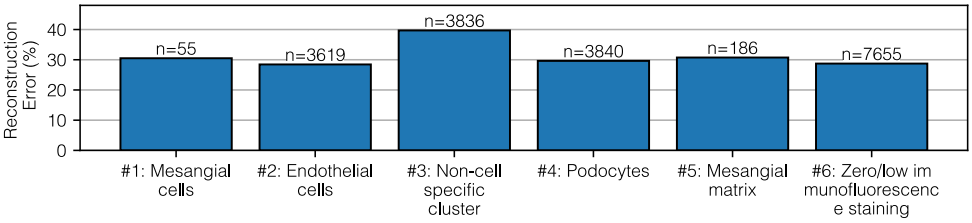

Figure S5.4: Bar plot of reconstruction error (%) across six clusters. Each bar represents the percentage error in Frobenius norm, with the sample size n displayed above each bar. The plot highlights how reconstruction accuracy varies by cluster. The reconstruction error is only considered on the pure pixels, only containing a single cluster. Cluster 3 (non-cell specific) comprises primarily non-targeted (with the markers) regions within the glomerulus, often capturing a heterogeneous mix of substructures, resulting in a relatively elevated reconstruction error.

REFERENCES

[1] B. Efron. “Bootstrap Methods: Another Look at the Jackknife”. In: *Breakthroughs in Statistics: Methodology and Distribution*. Springer, 1992, pp. 569–593.

[2] D. N. Politis and J. P. Romano. Large Sample Confidence Regions Based on Subsamples Under Minimal Assumptions. In: *The Annals of Statistics* (1994), pp. 2031–2050.

[3] A. B. Esselman, F. A. Moser, L. E. Tideman, L. G. Migas, K. V. Djambazova, M. E. Colley, E. L. Pingry, N. H. Patterson, M. A. Farrow, H. Yang, et al. In Situ Molecular Profiles of Glomerular Cells By Integrated Imaging Mass Spectrometry and Multiplexed Immunofluorescence Microscopy. In: *Kidney International* 107.2 (2025), pp. 332–337.

[4] C. Zheng, M. Yu, J. Shan, A. Wang, and H. Chen. Fast Sparse Non-negative Least Squares Via ADMM for High-Resolution DOA Estimation. In: *IEEE Sensors Journal* 23.4 (2023), pp. 3901–3910.

[5] J.-F. Cai, E. J. Candès, and Z. Shen. A Singular Value Thresholding Algorithm for Matrix Completion. In: *SIAM Journal on Optimization* 20.4 (2010), pp. 1956–1982.

[6] G. Casalino, N. Del Buono, and C. Mencar. Subtractive Clustering For Seeding Non-negative Matrix Factorizations. In: *Information Sciences* 257 (2014), pp. 369–387.

6

Conclusions and Recommendations

Throughout this dissertation, we have addressed challenges related to data dimensionality, data volume, and signal contamination with noise and artifacts in the processing of IMS and METIS data, proposing structured and low-rank decomposition techniques that balance data fidelity, interpretability, and computational efficiency. Beginning in Chapter 1 with a discussion of the curse of dimensionality and the imperative for robust models, we have developed and tested a suite of methods, low-rank plus sparse decompositions, matrix completion approaches, and cross-modality integration, tailored to the unique demands of imaging mass spectrometry and mid-infrared astronomy.

This dissertation focused on developing structured and low-rank decompositions to overcome the three identified challenges in image-acquiring instruments, *i.e.*, high dimensionality and large data volume, limited control possibilities, and interference of noise and artifacts. We have investigated in this dissertation different matrix decomposition methods to achieve this objective. The conclusions for each contribution are summarized as follows:

In Chapter 2, we set out to address the need for *robust noise and artifact suppression* in an imaging mass spectrometry (IMS) setting. To this end, we investigated two well-studied, structured, low-rank methods, Principal Component Pursuit (PCP) and Stable PCP (SPCP), as extensions of traditional Principal Component Analysis (PCA) that explicitly model sparse residuals. In a first case study using a human cornea IMS dataset, we assessed communality among PCA, PCP, and SPCP regarding the recovered subspaces of their respective low-rank approximations. We observed that they all share a very similar spectral subspace, but that there are substantial differences with respect to the subspace recovered in the spatial domain. The latter might be an indication that sparse features are more common in the spatial rather than the spectral domain in this type of data. In a second case study on human retina data, we further explored how tuning the noise-threshold parameter (σ-multiplier) and the sparse-penalty parameter (θ-multiplier) in SPCP governs the cardinality and distribution of the extracted sparse features. We found that the features captured by the sparse term can contain genuine biological signals and that careful parameter selection is therefore essential: too small a threshold can suppress genuine signals in the sparse term, while too large a threshold reverts the model toward PCA-like behavior. Nevertheless, PCP and SPCP both tended to approximate signal intensities from below in their low-rank terms, *i.e.*, minimizing the risk of overestimating molecular intensities, whereas PCA showed a tendency to overestimate these intensities in low-rank reconstructions. Conclusively, low-rank and sparse decomposition methods offer powerful means to approximate IMS data and pave the way for high-dimensional, full mass-profile analysis. However, they demand careful parameter tuning, a challenge that grows with dataset size. Overall, our findings highlight the under-appreciated value of leveraging sparsity in IMS, for noise and artifact suppression, and suggest it should be a central consideration in future dimensionality-reduction approaches for molecular imaging.

In Chapter 3, we set out to address the challenge of extracting faint, yet sparse, astrophysical signals from low-rank mid-infrared backgrounds and doing so without sacrificing observing efficiency. This aligns mainly with the need for precise *noise and artifact removal*. In this chapter, we presented LORABEL, an SPCP-based background-subtraction algorithm tailored for mid-IR astronomy, and evaluated it against both chop-only and chop-nod techniques on simulated Earth-based VLT/VISIR and real airborne SOFIA/FORCAST datasets. Our VISIR simulations demonstrated that LORABEL delivers markedly more stable photometry at low signal-to-noise ratios, reducing the scatter of flux measurements even though it incurs a systematic underestimation of the absolute flux (up to $10-15\%$, depending on parameters). This trade-off can partially be calibrated out in practice by injecting artificial sources of known brightness, thereby preserving accuracy while retaining precision, addressing the need for robust noise suppression and unbiased source recovery. In a second case study involving detection tests across varying signal-

to-noise ratios (SNRs), LORABEL outperformed both chop-only and chop-nod methods in detection precision, which is particularly interesting for the faintest sources. This confirms that SPCP's explicit modelling of sparse "source" terms alongside low-rank backgrounds can enhance sensitivity under challenging observing conditions. When applied to real airborne SOFIA data, which is methodologically speaking non-ideal (*e.g.*, limited observing frames), LORABEL reduced mean background levels by factors of three to ten while largely preserving target signals. Although we observed a modest increase in the per-source measurement variance, the net improvement in SNR across all spectral bands validates LORABEL's applicability in non-ideal, time-constrained scenarios. Conclusively, these results show that by exploiting sparsity as a source signal property instead of a noise property, LORABEL can achieve effective background removal without requiring prior knowledge on source position, additional observational time or nodding frames, advancing the SPCP framework beyond IMS data into the realm of astronomical imaging. At the same time, the complexity of its hyperparameter space remains challenging and its tendency to not completely subtract background signal remains problematic for accurate photometry applications.

In Chapter 4, we introduced a unified low-rank decomposition framework that explicitly exploits the intrinsic sparse structure (*i.e.*, a matrix with mostly zero elements) of IMS measurements to reduce its memory footprint. Through this reduction, full profile datasets can be processed and the *large data volume* challenge can be addressed. We achieved compression factors of up to 2 500-fold relative to dense storage and 600-fold relative to selective peak list storage, yet still preserving full profile information for every pixel. In the no-missing-value scenario, our decomposition strategy reduced reconstruction error by 39.1% compared to traditional peak picking while retaining all spectral features. In the missing-value scenario, it cut global information loss by up to 40% at comparable compression ratios. Crucially, our approach maintained sensitivity to low-SNR and near-isobaric signals, avoiding the selection bias of peak-selection methods by preserving specificity through the retention of closely spaced *m/z* features. These enhancements should improve downstream analyses by furnishing a more comprehensive reduced representation of IMS data, while still delivering dimensionality reduction on par with conventional peak picking. Conclusively, these developments have paved the way for a large volume, full-profile IMS analysis pipeline that compresses data without sacrificing the rich signal content, which is essential for high-fidelity lipidomic, metabolomic and proteomic discovery.

In Chapter 5, we developed TULIP, a microscopy-informed spectral unmixing framework for imaging mass spectrometry (IMS), to resolve *high-dimensional* mixed pixel signals into distinct molecular spectra tied to biological regions, from single cells to functional tissue units. By casting unmixing as an inverse problem enriched with non-negativity, sparsity, and low-rank constraints, TULIP robustly handles both overdetermined and underdetermined linear system cases. Benchmarking against Ordinary Least Squares (LS), Non-negative Least Squares (NNLS), and Singular Value Thresholding (SVT) showed that TULIP delivers advanced cell fit scores and enforced non-negative spectra in overdetermined settings, while matching their performance in underdetermined scenarios. Its scalability and biological interpretability were confirmed on a large functional tissue unit dataset. Conclusively, these findings established that integrating high-resolution

microscopy-constraints into IMS unmixing via TULIP enhances molecular specificity and interpretability, achieving robust, non-negative spectral recovery, superior cell fit performance, and scalability across both over- and underdetermined contexts. By guiding unmixing and ultimately decompositions with biologically meaningful priors, TULIP paved the way for scalable, in situ analyses of heterogeneous samples and suggests a novel means for matrix-based dimensionality reduction in molecular imaging pipelines.

RECOMMENDATIONS FOR FUTURE WORK

While this dissertation advances structured and low-rank decompositions for large imaging datasets, there are still several outstanding challenges that are situated in the modelling of the data, the gathering of the data itself, the methods, and future case studies and applications.

TENSOR AND MULTIWAY DECOMPOSITIONS

Flattening high-dimensional data into matrices discards its multiway structure. We therefore advocate extending our matrix-based framework to tensor-based models, such as presented in Chapter 1, Eq. 1.3, which natively separate spatial, spectral, and/or temporal dimensions. Besides the acknowledged physics-informed constraints, these models could also incorporate complementary measurements, such as high-resolution microscopy in IMS. For METIS, spatial properties of point and extended sources could be taken into account, as well as more complex reduction schemes.

NON-LINEAR AND DEEP LEARNING MODELS

Linear subspace methods have inherent limitations when the signal manifold is, *e.g.*, curved or hierarchical. Kernels could offer a first step toward capturing non-linear relationships in the image data. On the other hand, auto-encoder architectures with explicit sparsity and low-rank constraints at the bottleneck layer can be trained end-to-end on large IMS or METIS datasets, learning complex non-linear embeddings. Note, however, that there often exists a trade-off between interpretability, scalability, and model complexity, making full profile analysis in IMS or before-reduction data analysis for METIS challenging. Therefore, method scalability techniques (see below) can be considered.

PROBABILISTIC AND VARIATIONAL APPROACHES

For applications where quantifying uncertainty is as important as recovering signal, probabilistic formulations can be adopted to yield posterior distributions over the recovered latent factors. Variational inference techniques can approximate these posteriors efficiently, providing, *e.g.*, intervals on reconstructed spectra and images that inform downstream decision-making and experimental design. Model selection, choosing optimal ranks or sparsity levels, can then be carried out automatically by maximizing the evidence lower bound or marginal likelihood, further reducing researcher bias in hyperparameter choices.

Physics-Informed and Application-Driven Constraints

Embedding more extensively physics-based priors directly into the decomposition objective has the potential to further boost interpretability. To this aim, known point spread function shapes for METIS, molecular profiles or combinations for IMS, or known detector or other upstream instrument component artifacts can be encoded in the optimization process. Furthermore for IMS, one possible prior to investigate is the sparsity of features that are more common in the spatial rather than the spectral domain. However, note again that there is a trade-off here in model complexity, and scalability, and that any bias introduction should be avoided.

Data Preprocessing

Ensuring reliable inputs to any decomposition algorithm begins with rigorous preprocessing. In the context of IMS, we recommend researching per-feature scaling and alignment to counteract systematic *m/z* drift and peak broadening before any decomposition is performed, especially for full profile methods. Secondly, to preserve low-intensity or very sparse features, extra effort needs to be put in the review of the ℓ_2-norm penalty, that is currently often used. Reviewing wavelet-based decompositions might also be a way to preserve low-intensity or very sparse features.

Method Scalability

Until now, we have assumed that more data leads to better decision making. However, we may not need to process all the data at once. If, however, we want to scale our algorithms beyond the current limits, we can turn to block methods, randomized approaches, or online/streaming techniques. These reduce memory complexity, by exchanging it for extra computation or by sacrificing aspects such as exactness or convergence speed, but they spare IMS vendors and the METIS project from storing large datasets. In particular, online algorithms like incremental SVD or randomized subspace tracking can be adapted for streaming data. Deploying these solvers on GPUs enables on-edge decompositions within acquisition hardware, slashing data-transfer and storage costs. Moreover, integrating low-rank compression on the fly into the acquisition pipeline further alleviates I/O bottlenecks by writing compressed representations directly to disk as the data arrive.

Case Studies and Future Applications

We have considered a variety of case studies in this dissertation, ranging from human eye tissue to point sources in the mid-infrared. However, across all these cases and for both application areas, we identified a recurring challenge: the lack of clear, universally accepted scoring metrics due to the absence of field-wide standards. A crucial first step toward standardizing both domains is to establish and define such metrics for decomposition methods. To this end, we propose using a representative selection of tissue sections, *i.e.*, human, mammalian, and plant, for IMS, and a diverse range of objects for mid-infrared observations, including point sources and extended objects of various forms and characteristics. These could serve as a foundational set for defining robust metrics. Additionally, defining clear standards for synthetic dataset creation would further support fair and consistent evaluation of different algorithms. As such, we would

propose future case studies to include a variety of datasets and a comparison with respect to these metrics and synthetic ground truth datasets.

CURRICULUM VITÆ

Roger Amaury Rutger MOENS

22-05-1995 Born in Jette, Belgium.

EDUCATION

2007–2013 Grammar School
Sint-Jozefscollege, Brussels, Belgium

2013–2017 B.Sc. in Mechanical Engineering
Delft University of Technology, Delft, Netherlands

2015-2016 International Diploma in Mechanical Engineering
Imperial College London, London, United Kingdom

2017-2021 M.Sc. in Aerospace Engineering
Delft University of Technology, Delft, Netherlands

2017-2021 M.Sc. in Systems and Control
Delft University of Technology, Delft, Netherlands

2021-2025 Ph.D. in Machine Learning for Multimodal Imaging
Delft University of Technology, Delft, Netherlands
Thesis: Structured and Low-Rank Decompositions for Large-Scale Imaging Datasets
Promotors: dr. ing. R. Van de Plas
prof. dr. ir. B. De Schutter

LIST OF PUBLICATIONS

5. Battjes, F.A., Olling, K., Delacour, P.L., Migas L.G., Spraggins, Lorenz, M.W., Ievlev, A.V., J.M., Ovchinnikova, O.S., Raf Van de Plas, R., & **Moens, R.A.R** (2025). Enforcing Physical Constraints in Full Spectrum Imaging Mass Spectrometry Data Reduction [Unpublished Manuscript].

4. **Moens, R. A. R.**, Patterson, N.H., Migas, L.G., Esselman, A.B., Moser, F.A., Spraggins, J.M. & Van de Plas, R. (2025). Unmixing of Imaging Mass Spectrometry Measurements Using Microscopy-informed Constraints [Unpublished Manuscript].

3. **Moens, R. A.**, Migas, L. G., Van Ardenne, J. M., Skaar, E. P., Spraggins, J. M., & Van de Plas, R. (2025). Preserving Full Spectrum Information in Imaging Mass Spectrometry Data Reduction. *Bioinformatics, 41*(5), btaf247.

2. **Moens, R. A. R.**, Pietrow, A. G. M., Brandl, B., & Van de Plas, R. (2025). Thermal Background Reduction for Mid-Infrared Imaging by Low-Rank Background and Sparse Point Source Modelling [Unpublished Manuscript, Submitted to Astronomy & Astrophysics].

1. **Moens, R. A. R.**, Migas, L. G., Anderson, D. M. G., Messinger, J. D., Ovchinnikova, O. S., Caprioli, R. M., Spraggins, J. M., & Van de Plas, R. (2025). Advanced Dimensionality Reduction for Imaging Mass Spectrometry of Human Eye Tissue Through Low-rank Modeling with Sparse and Dense Residuals. *Analytical Chemistry*, 97.42, 23040-23049.