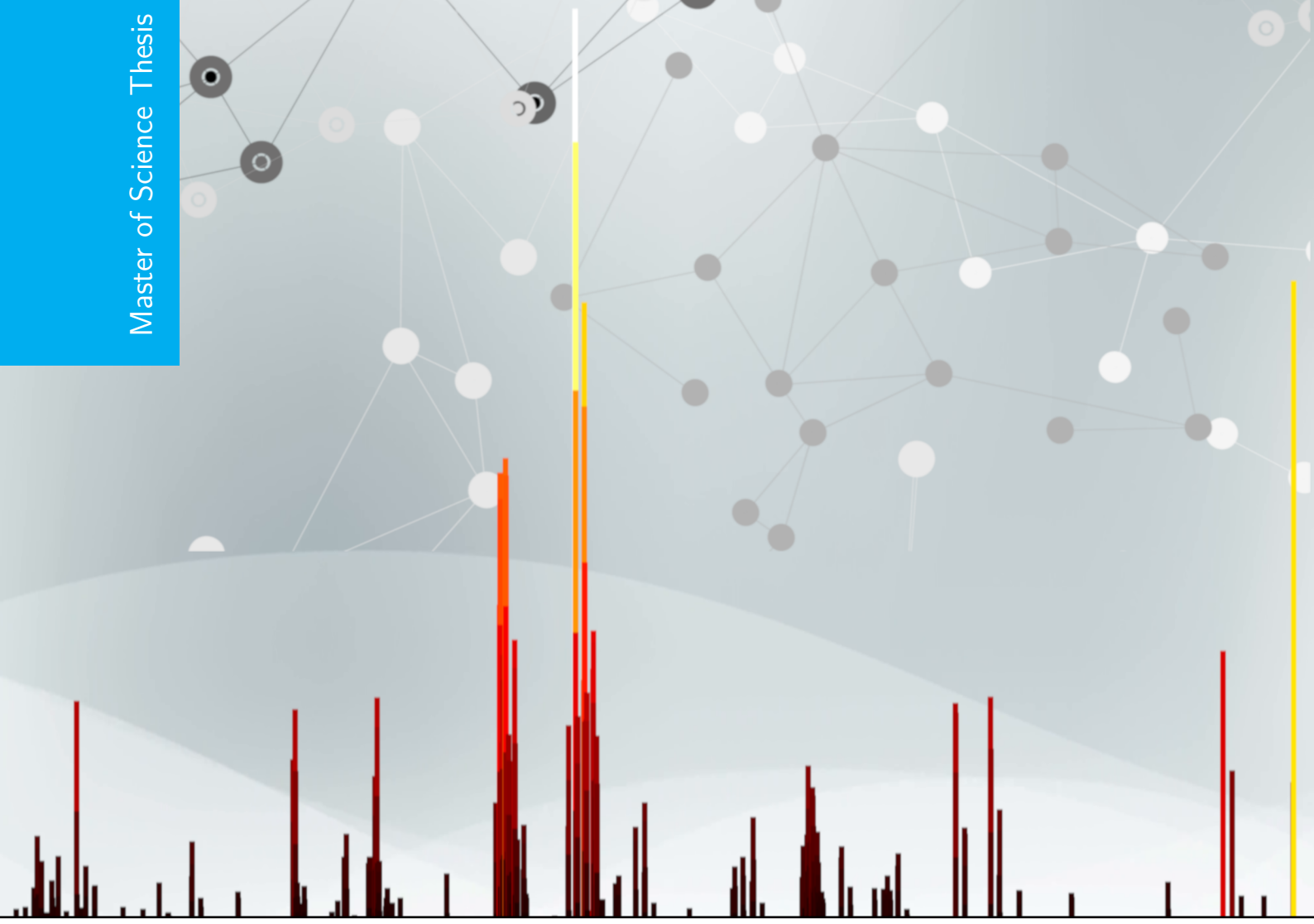
Intensity-Aware Rank Estimation for Dimensionality Reduction in Imaging Mass Spectrometry

M.H. van Winden

Master of Science Thesis



TU Delft Delft University of Technology

Delft Center for Systems and Control

Intensity-Aware Rank Estimation for Dimensionality Reduction in Imaging Mass Spectrometry

Master of Science Thesis

For the degree of Master of Science in Systems and Control at Delft University of Technology

M.H. van Winden

April 8, 2019

Faculty of Mechanical, Maritime and Materials Engineering (3mE) · Delft University of Technology

Abstract

Imaging Mass Spectrometry (IMS) is a spectral imaging technique, which enables detection of the spatial distribution of molecules by collecting a mass spectrum for every pixel across a tissue sample. As such, IMS enables the detection of disease-introduced anomalies in tissue samples as well as the gaining of deeper insight on a molecular level into biological processes.

The dimensionality of IMS data is high, considering that every bin (or ion) along a mass spectrum represents a separate image and the number of pixels per image is relatively high. Manual analysis of the data suffers from this high dimensionality as visualization becomes increasingly difficult. Furthermore, analysis of such large datasets becomes problematic or infeasible for computational techniques both in time and computational resources. Moreover, the dimensionality of current IMS measurements hampers new applications capturing even more data.

Linear dimensionality reduction methods, such as Principal Component Analysis (PCA) and Nonnegative Matrix Factorization (NMF), seek to reduce these datasets to a set of (principal) components. These components span an underlying feature subspace within the original measurement space. Rank estimation determines the quantity of such components, estimating the number needed to represent the original dataset in a lower-dimensional space while incurring minimal information loss. In the context of IMS, this task is typically performed without the use of domain-specific knowledge.

Intensity-aware rank estimation seeks to utilize domain knowledge - in the form of an ion intensity threshold - to help estimate the rank. This threshold emerges naturally from IMS, due to prior knowledge on instrument and ionization process inaccuracies in the low ion intensity region. The ion intensity threshold defines a lower bound for which variations in measurements are reliable.

Establishing an intensity-aware version of rank estimation requires the threshold, defined in the original measurement space, to be linked to the abstract feature subspace, defined by NMF or PCA, where the rank estimation takes place. This connection is nontrivial to make and is, therefore, a central topic of this thesis. Furthermore, intensity-aware rank estimation requires the abstract subspace to represent the majority of the information above the threshold in the first set of components, which is not guaranteed in pure NMF and PCA formulations.

In this thesis, we demonstrate threshold-aware rank estimation and residual-fraction rank estimation which make rank estimation for PCA intensity-aware. Threshold-aware rank estimation applies a histogram transformation to the intensities in the original measurement space to emphasize threshold-exceeding intensities. Consecutively, we estimate the rank based on the percentage of explained variance. Residual-fraction rank estimation uses untransformed measurements but instead estimates rank based on the ratio of the above- and below-threshold residuals. We demonstrate that both rank estimations are able to find the correct rank in a synthetic dataset. With threshold-aware rank estimation applied to an IMS dataset, we show that the transformation before application of PCA leads to a lower overall estimate of rank based on a percentage of the explained variance. With residual-fraction rank estimation applied to an IMS dataset, we show that we can obtain rank estimates based on the structure of dataset close to cross-validation rank estimates for the same dataset.

Table of Contents

List of Figures

Acknowledgements

This Master of Science graduation thesis came to be after a one-and-a-half year internship at CERN, Geneva. It originated from the desire to get in touch again with the mathematics after applying mostly rule-based control techniques in SCADA and PLC systems at CERN. My supervisor, dr.ing. R. Van de Plas, provided me with the topic to bring rank estimation closer to the physical process behind data. It supplied me with me plenty of mathematics, which has challenged me. It has frustrated and enlightened me greatly, which is mostly the result of there not being universal answer to the problem of rank estimation.

I would like to thank first and foremost my supervisor dr.ing. R. Van de Plas for his time, guidance and inspiration for this thesis. In the weekly meetings, he guided me through the process of finding an approach towards the problem of rank estimation. He gave me the freedom for giving direction to the research, experimenting and investigating other opportunities, while at the same providing explicit bounds and requirements.

Furthermore, I would like to thank my thesis colleagues in the "Squad" for encouraging me when I was demotivated or stuck. The strict thesis hours and cookie punishments got me through the last hard push of writing down my observations and discoveries. I want to thank especially my roommates Stijn Bosma and Peterke van der Zwaag with whom I shared my frustrations and achievements.

Lastly, I am grateful to my friends and family who have supported me through this process. I would like especially to express my appreciation to my parents for their active moral, emotional and financial support throughout my whole education. I would not have been able to achieve this without them. Also, I would like to thank my dear friends and peers Yvo Putter and Marc van Vliet for listening to my considerations about this topic as an intermezzo to good stories and laughter during the countless dinner evenings. Lastly, I would like to express my gratitude to my girlfriend, Maris Tali, who paired up with me when we both had to write our theses. Maris supported me when I was frustrated, allowed me to take a break, relax and enjoy other activities in life, and made me look at the problem from a different perspective.

Delft, University of Technology — M.H. van Winden
April 8, 2019

"Poetry is the art of giving different names to the same thing. Mathematics is the art of giving the same name to different things."

— *Henri Poincaré*

Chapter 1

Introduction

Spectral imaging covers a set of different imaging techniques that collect a spectrum of intensity values at every location or pixel of an image. A recently developed spectral imaging technique, Imaging Mass Spectrometry (IMS), collects a mass-over-charge spectrum per pixel covering the entire sample and thus allows detection of molecules across a biological tissue sample section [1]. Figure 1-1 shows a schematic overview of an IMS process.

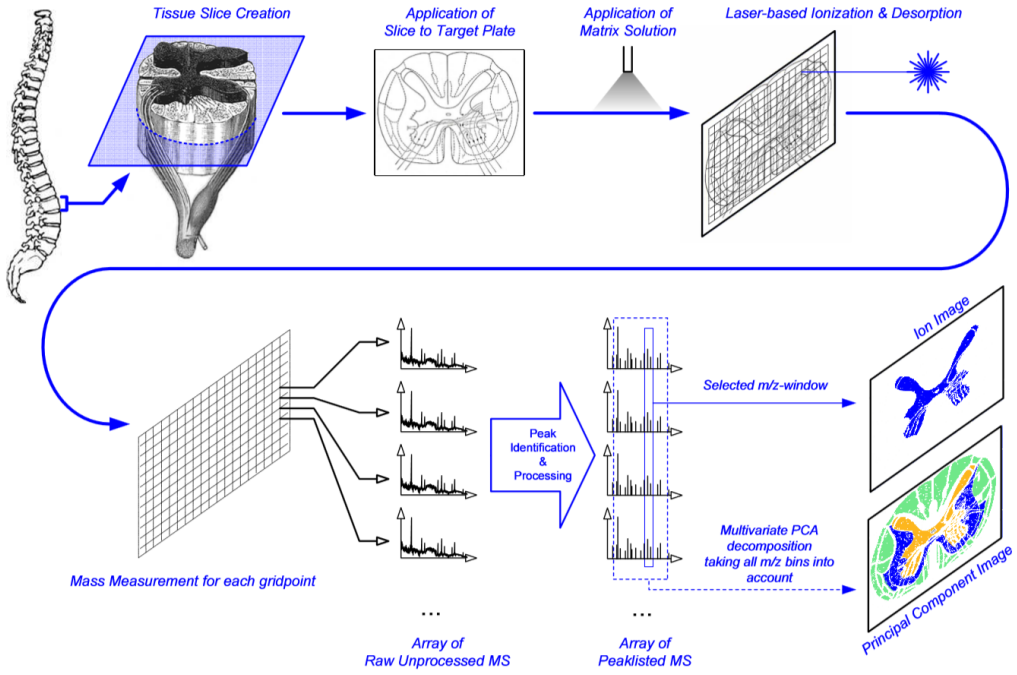


Figure 1-1: A schematic overview of an IMS experiment. The tissue section is obtained using a microtome, mounted on a target plate, and an appropriate chemical matrix solution is applied to enable ionization. The mass spectral measurements, data collection, and most low-level processing of the mass spectra take place inside the mass spectrometer or its instrument computer [2].

For a pixel, the mass-over-charge (m/z) spectrum, or mass spectrum (since the charge is often +1), reports the presence and abundance of ions (charged molecules) present at this pixel location in a sample. This mass spectral signal is described by a finite set of mass-over-charge bins, covering a mass range of interest. The intensity values along the spectral dimension represent the (relative) quantities of the respective ions present in the sample at a particular location. These spectra are acquired using a mass spectrometer, which pixel by pixel desorps and ionizes the molecules (puts a charge on them). Electromagnetic fields then accelerate the ions through a mass analyzer, which filters and sorts the ions by mass, and moves them towards a detector which counts their occurrences. Ordering the ion count values by their corresponding ion masses then creates the mass spectrum. This process continues until all pixels in the sample are measured. A view on the distribution of the molecules over the sample can be obtained by stacking the pixel mass spectra such that they form ion images for every m/z bin in the mass spectrum. Examples of these ion images are shown in figure 1-2.

The measurement of the spatial distribution of biomolecules in a sample enables detection and comparison of proteomic, peptidomic, lipidomic, and metabolomic content throughout organic tissue sections, based on the molecular masses and without requiring a priori labeling for target molecules [1]. As a result, IMS can be used as a tool to unravel and understand processes in cells at a molecular level, and it is a potent tool for disease-related biomolecular discovery [3]. For this reason, IMS is used in exploratory biomedical studies of diseases such as Parkinson's disease [4], Alzheimer's disease [4], and cancer [5].

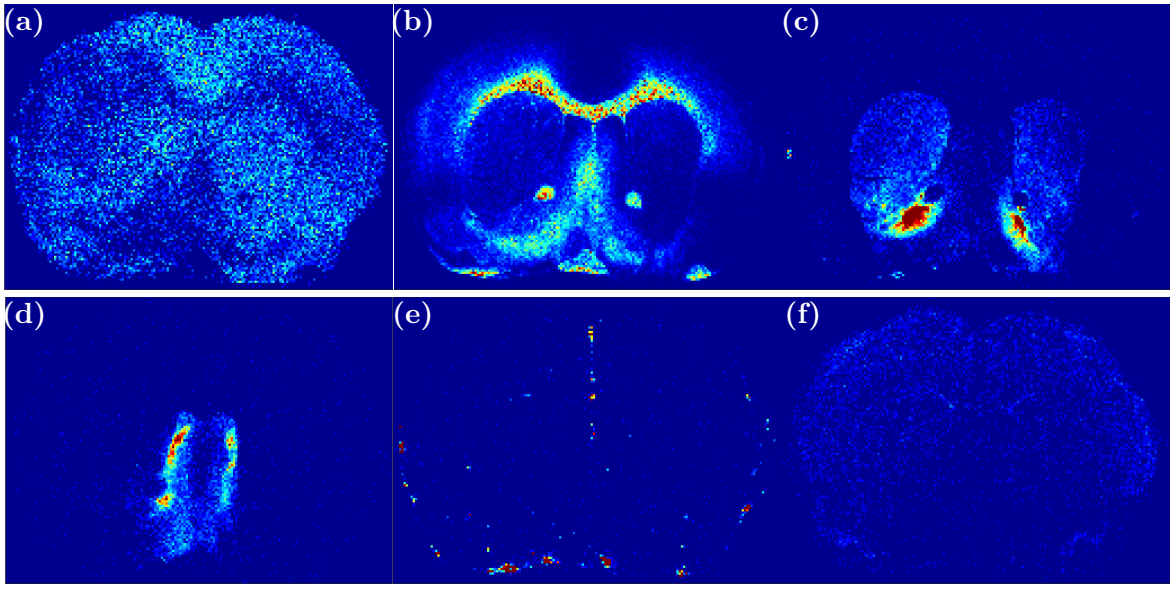


Figure 1-2: A set of typical ion images in a Coronal Rat Brain dataset. **a**: $m/z = 1608.8$, **b**: $m/z = 2029.1$, **c**: $m/z = 2063.0$, **d**: $m/z = 2396.2$, **e**: $m/z = 2490.1$, **f**: $m/z = 2759.4$.

A growing issue in the IMS field is the high dimensionality of the datasets. The high dimensionality in IMS originates from the number of bins in each spectrum and the number of pixels for which spectra are collected. Examples of problems arising with high dimensionality are:

- IMS data can range from gigabytes to a few terabytes, depending on the image area

and the spatial and spectral resolution [6]. As a result, analysis of such large datasets becomes problematic and rapidly infeasible for computational techniques both in time and computational resources [7].

- Manual analysis is impractical for humans for studies without an a priori hypothesis of a target molecule associated with an ion image, or a target pixel spectrum, due to the number of mass-bins in each spectrum and the number of pixels for which spectra are collected.
- IMS suffers from the curse of dimensionality [8, 9]. The high dimensionality results in a large volume of the measurement space and causes the available data only sparsely describe this space. As a result of this sparsity, it becomes problematic to achieve statistical significance and define similarity between measurements. Consequently, classification, segmentation, and other machine learning approaches require vast amounts of training data [10].
- The inability to effectively handle the current dimensionality of IMS measurements hampers the development of new applications capturing even more data, such as 3D IMS [11] or connecting ion mobility detection to IMS [12]. Furthermore, these problems are strengthened by the consistently increasing spatial and spectral resolutions of the instrumentation [13].

Dimensionality reduction could ease these problems. In general, the motivation of dimensionality reduction is to obtain a reduced set of components spanning an underlying lower dimensional feature subspace within the original measurement space. Figure 1-3 shows a schematic representation of a decomposition of the spectral measurements into a set of three principal components. The idea is that these principal components capture the majority of the information in the data, while requiring fewer dimensions to describe this information [6]. This class of dimensionality reduction is comparable to lossy compression. Consequently, this class of dimensionality reduction is always a trade-off between discarded information and dataset dimensionality.

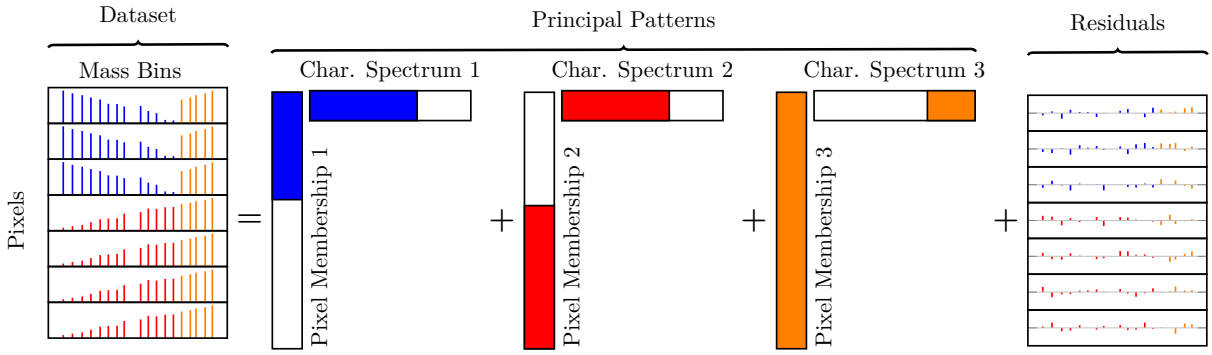


Figure 1-3: Schematic overview of linear dimensionality reduction in which the dataset consisting of stacked pixel spectra is decomposed into three components. The residuals on the right represent the lost information as a result of dimensionality reduction.

Rank estimation, the focus of this thesis, attempts to estimate in a linear setting the optimal number of these principal components to select for minimizing the dimensionality of a dataset,

while at the same capturing all relevant information. We expect that incorporating prior knowledge about on the relevant intensity range of the data into this process could further improve the low dimensional approximation and the associated rank estimate of a given dataset. We refer to this approach as intensity-aware rank estimation. This approach intends to incorporate this prior knowledge about the data into the process by taking into account known instrument and experimental design properties when constructing a lower dimensional representation and making an estimate of the rank.

In this thesis, we focus on intensity-aware rank estimation based upon Principal Component Analysis (PCA) and Nonnegative Matrix Factorization (NMF) for which a minimum signal intensity threshold for the obtained mass spectra is used as an example of such prior knowledge. In an IMS mass spectrum, the region below this minimum intensity threshold can typically not be used to support biological conclusions. This low-intensity region is largely defined by instrument and experimental design properties, such as instrument sensitivity, which cause this region to consist primarily of noise or low-reliability signals. The abstract subspace found by the dimensionality reduction method should, therefore, emphasize values above the intensity threshold to make the dimensionality reduction result consistent with the measurement instrument properties.

1-1 Research Goal: Intensity-Aware Rank Estimation

The objective of this thesis is to select a minimal number of principal components spanning the lower dimensional feature subspace, while capturing the majority of the intensity-threshold-exceeding intensity signal variation. With intensity-aware rank estimation, we take into account the biological significance and signal reliability of the intensities as defined by the ion intensity threshold.

Utilizing the ion intensity threshold for rank estimation is nontrivial due to the duality between the original measurement space, in which the threshold is defined, and the lower dimensional feature subspace produced by the dimensionality reduction. Figure 1-3 shows that the abstract subspace of rank K describes the original data as weighted combinations of K vectors. This description is not suited for thresholding, because one scalar value in the abstract subspace describes a linear combination of multiple values in the original measurement-space. Also, we have no guarantee on the existence of a rank for which all threshold-exceeding intensities are only described in the weighted combination of the first K vectors. On the other hand, utilizing a threshold directly in the original measurement space, before dimensionality reduction, changes the distribution of the data and could cause overly harsh cropping of the lower dimensional subspace. Particularly, an increase in the variance or area under the curve, due to the manipulation for a specific entry in the mass spectrum in the measurement space, could boost that elements significance or weight in the abstract space.

Furthermore, general rank estimation is nontrivial since it is not only affected by the prior knowledge about the data, but also by the motivation behind the reduction of dimensionality. In other words, the optimal number of principal components can be dependent on the application. For example, the optimal rank estimate to find distinct patterns in the preprocessing phase leading up to clustering with k-means differs from the optimal rank for compression of the original dataset focused on the reconstruction of that dataset. The former requires a

number of, potentially uncorrelated, principal components highlighting the distinctions between the measurements of interest. The latter requires a number of components resulting in a minimal residual of the intensities of interest.

In this thesis, we target the nature of the data, but we bear in mind the application-dependence of rank estimation. We have investigated rank estimation by two different approaches in the form of intensity-aware dimensionality reduction and intensity-aware rank estimation based on PCA. Within these approaches, we introduce respectively a threshold-shifted and residual-fraction rank estimation and argue that their applicability is dependent on the aforementioned context of the application.

In threshold-shifted rank estimation, we apply a transformation in the original measurement space before dimensionality reduction and thus modify the obtained abstract subspace. The applied transformation is a downwards-shift on the intensity values by a threshold τ (the ion intensity relevance threshold) and subsequently setting any negative values to zero. This transformation reduces the area under the curve for ions with low-intensity values more than for their high-intensity counterparts. As a result of this transformation, we can obtain a set of principal components emphasizing the intensities in the spectrum above the intensity threshold. Subsequently, we estimate the rank by specifying a required percentage of the captured variance in the transformed space.

In residual-fraction rank estimation, we focus on rank estimation without manipulation of the data in the original measurement space. Instead, we intend to estimate the rank by minimizing the fraction of residuals of threshold-exceeding intensities relative to the residuals of the below-threshold intensities as captured in the traditional threshold-unaware abstract subspace.

1-2 Related work and Contributions

The large body of work focused on the dimensionality reduction of IMS datasets, in order to ease the problems with analysis arising with dimensionality, demonstrates the importance of this topic. Furthermore, several publications have explicitly stressed the importance for more and improved dimensionality reduction techniques [7, 10, 5]. The application of different matrix factorization methods, such as PCA [6, 14, 15] and NMF [16, 17], other linear methods, such as Probabilistic Latent Semantic Analysis (PLSA) [14] and Linear Discriminant Analysis (LDA) [15] and random projections [8], and nonlinear methods, such as autoencoders [7] and t-Distributed Stochastic Neighbors Embedding (t-SNE) [5] to IMS datasets have been evaluated. In the context of IMS, a more memory efficient version of PCA has been proposed [6]. PCA has been used as an unsupervised dimensionality reduction before clustering [18] and trend detection [19].

For example, Klerk et al. have applied PCA, NN-PARAFAC, and PCA+VARIMAX for evaluation of large SIMS and LDI datasets [20] and estimated the rank based on the cumulative percentage of variance. McCombie et al. have used the cumulative variance for component selection in PCA [18]. Hanselmann et al. have estimated the rank via the Akaike Information Criterion (AIC) [21] for PLSA with nonnegativity constraints closely related to NMF for IMS data [22]. Harn et al. proposed a dictionary learning method in which the rank is imposed by the dictionary to unravel molecular structures [23]. In a similar spectral imaging technique,

Hyperspectral Imaging (HSI), commonly used in remote sensing, dimensionality reduction based on NMF was applied using an empirical rank estimate [24, 25].

In this thesis, we focus on linear dimensionality reduction methods PCA and NMF, The additional challenges posed by nonlinear methods, especially due to the dimensionality of the problem, are not the focus of this thesis. We deem linear methods a good starting point to derive an initial intensity-aware rank estimation method and leave a possible nonlinear generalization for future research.

We believe most applications of dimensionality reduction to IMS lack a formal rank estimation procedure for IMS datasets. For this reason, we propose two methods to establish a more formal rank estimation procedure for PCA-based dimensionality reduction for IMS with the help of the ion intensity threshold. This approach differs from the aforementioned techniques, where the rank is often determined empirically. In other cases, traditional methods available from the literature, that do not take into account contextual information about the datasets, were used.

Outside of the context of IMS, a wide range of rank estimation methods for dimensionality reduction based on decomposition has been developed. Contrary to Cross-Validation (CV) [26, 27, 28] and Bootstrapping [29, 30, 31], our approach towards intensity-aware rank estimation is not based on computationally intensive iterative decomposition for evaluation of a prediction error, nor does it assume a statistical model about the data, such as Maximum Likelihood Estimation (MLE) [32], Minimum Description Length (MDL) [33, 34], Stein's Unbiased Estimator (SURE) [35], Bayesian Model Selection [36] and Automatic Relevance Determination (ARD) [37, 38, 39]. We deem threshold-shifted rank estimation an extension of variance-based thresholding methods [40, 41] built around the intensity threshold. Residual-fraction rank estimation differs from the aforementioned methods, as it is solely based on the distribution of the residuals over the intensities.

With threshold-shifted rank estimation, we show that reducing dimensionality and estimating rank in a manner that reflects the biological significance of the intensity values can lead to a lower overall estimate of rank for a given dataset, when compared to threshold-unaware dimensionality reduction. The lower rank is a consequence of the abstract subspace that captures less of the intensity values below the intensity threshold. With the residual-fraction rank estimation, we show that the effect of low-reliability intensities in the data can be used to obtain a subset of principal components maximally reflecting above-threshold intensities in traditional PCA. Furthermore, we demonstrate that residual-fraction rank estimates are similar to rank estimates as obtained with cross-validation.

1-3 Outline

In this chapter, we have introduced dimensionality reduction, rank estimation and motivated why dimensionality reduction is essential. Furthermore, we have motivated why dimensionality reduction aware of the instrument properties could be beneficial in the case of IMS. For the remainder of the thesis, the outline is as follows:

- In chapter 2, we introduce the mathematical concept of rank, discuss the origin of the threshold, and outline the datasets used to evaluate our method.

- In chapter 3, we introduce PCA and NMF, show how PCA and NMF capture intensities based on the residuals and demonstrate why the intensity capturing is problematic for pure intensity-aware rank estimation. As an alternative, we propose residual-fraction rank estimation.
- In chapter 4, we outline Threshold-Aware Principal Component Analysis (TAPCA) and introduce the particular class of transformations to make the lower-dimensional abstract subspace emphasize threshold-exceeding intensities. Consecutively, we show three different examples of such transformations, that enable emphasis on the threshold-exceeding intensities. Furthermore, we discuss the problems in other dimensionality reduction methods that we have tested to create a lower-dimensional abstract subspace aware of the intensity threshold.
- In chapter 5, we compare the threshold-shifted method for constructing the abstract subspace with standerd dimensionality reduction based on PCA in terms of residuals. Furthermore, we use a synthetic dataset with known rank to evaluate both threshold-shifted and residual-fraction rank estimation. Finally, we apply both the threshold-shifted and residual-fraction rank estimation methods to a real IMS dataset and discuss the results.
- In chapter 6, we summarize our observations on intensity-aware rank estimation and list our conclusions and recommendations for future work.

Chapter 2

Fundamentals

2-1 Dimensionality Reduction

Feature extraction and feature selection are typical strategies for dimensionality reduction [31]. Feature selection intends to find a subset of the original measured variables by detecting and picking relevant features for a task at hand. Feature extraction, on the other hand, transforms the data typically from a high- to a lower-dimensional space, potentially generating new variables in the process. Feature extraction methods are the focus of this thesis since, one, many feature selection methods are supervised, and two, the goal is to retain as much of the original data, while reducing dimensionality in an unsupervised manner. In this section, we define the concept of rank in the context of linear dimensionality reduction based on the feature extraction methods Principal Component Analysis (PCA) [40] and Nonnegative Matrix Factorization (NMF) [42]. Furthermore, we introduce the concept of intensity-aware rank estimation based on this mathematical concept of rank.

Rank

The rank R of a matrix $\mathbf{M}$ is the number of linearly independent rows or columns of $\mathbf{M}$. In the context of decomposition or factorization, the rank R of $\mathbf{M}$ is equal to the minimal number of rank-one matrices $\mathbf{F}$ required to reconstruct $\mathbf{M}$ exactly. A matrix is rank-one if it can be expressed as the nonzero outer product $w \times h$ of a column vector w and a row vector h. For example, if $\mathbf{M}$ has rank three ($R = 3$), we require the sum of three rank-one matrices or equivalently the sum of three outer products of two vector pairs to reconstruct the matrix $\mathbf{M}$ exactly. Equation (2-1) shows a matrix notation of the R pairs of column vectors w_i and rows vectors h_i represented as matrices $\mathbf{W} \in \mathbb{R}^{N \times R}$ and $\mathbf{H} \in \mathbb{R}^{R \times M}$ if $\mathbf{M} \in \mathbb{R}^{N \times M}$. An alternative formulation of the rank of matrix $\mathbf{M}$ is the number of nonzero eigenvalues of $\mathbf{M}$ or $\mathbf{M}^T\mathbf{M}$.

$$\mathbf{M} = \sum_{i=1}^{R} \mathbf{F}_i = \sum_{i=1}^{R} w_i h_i = \mathbf{WH} \tag{2-1}$$

Throughout this thesis, and in the context of PCA, we also refer to the rank-one matrices F_i, or the combinations of w_i and h_i, as principal components, principal patterns or principal factors.

Linear Dimensionality Reduction and Rank Estimation

In Linear Dimensionality Reduction (LDR) based on feature extraction with decomposition or factorization, we strive towards a lower dimensional approximation $\hat{\mathbf{M}}$ of the matrix $\mathbf{M}$ such that rank K of $\hat{\mathbf{M}}$ is smaller than rank R of $\mathbf{M}$, while incurring minimal information-loss. For this reason, dimensionality reduction intends to utilize a lower number of components w_i and h_i to approximate the matrix $\mathbf{M}$.

$$\mathbf{M} \approx \hat{\mathbf{M}} = \sum_{i=1}^{K} \mathbf{F}_i = \sum_{i=1}^{K} w_i \cdot h_i = \mathbf{WH} \tag{2-2}$$

with $\mathbf{M} \in \mathbb{R}^{N \times M}$, $\hat{\mathbf{M}} \in \mathbb{R}^{N \times M}$, $\mathbf{W} \in \mathbb{R}^{N \times K}$, $\mathbf{H} \in \mathbb{R}^{K \times M}$, $\text{rank}(\hat{\mathbf{M}}) = K$ and $\text{rank}(\mathbf{M}) = R$ with $R \leq \min(N, M)$. This process is visualized in figure 2-1.

In equation (2-2), the components w_i and h_i span an underlying abstract feature subspace within the original measurement space of $\mathbf{M}$. In this representation the components w_i and h_i describe the matrix $\hat{\mathbf{M}}$ as a weighted combination of K vectors. The way the components w_i and h_i, and consequently these rank-one matrices $\mathbf{F}_i$, are constructed, is determined by the underlying LDR technique and discussed in the section about PCA and NMF.

In the context of LDR, rank estimation determines the number K of these rank-one matrices $\mathbf{F}_i$ required to obtain a sufficiently close approximation $\hat{\mathbf{M}}$ of the original matrix $\mathbf{M}$ for minimal information loss given that $K < R$. As introduced before, we consider the definition of minimal information loss depends on the context, such as the motivation behind the application of dimensionality reduction and the nature of the data.

Intensity-Aware Rank estimation

Intensity-aware rank estimation seeks to add some of the aforementioned context about the data in the form of a reliability-threshold on the intensity values. We can interpret this threshold as lowered interest in all sub-threshold intensities in matrix $\mathbf{M}$ in equation equation (2-2). Consequently, we intend the reconstruction matrix $\hat{\mathbf{M}}$ to particularly capture threshold-exceeding intensities, while neglecting the sub-threshold fluctuations. In this thesis, we approach this in two ways which can be explained based on equation (2-2):

Intensity-aware Rank Estimation does standard LDR, but attempt to use the intensity threshold in helping to pick a set of components that preferentially represent above-threshold structure in the signal. The components w_i and h_i are constructed using the traditional linear feature extraction methods PCA and NMF and intend to find a rank K for which matrix $\hat{\mathbf{M}}$ has sufficiently captured the threshold-exceeding intensities in $\hat{\mathbf{M}}$. In chapter 3, we outline why selecting a rank based on sufficient capturing of threshold-exceeding intensities is troublesome. As an alternative, we propose a residual-fraction-based method to estimate the rank.

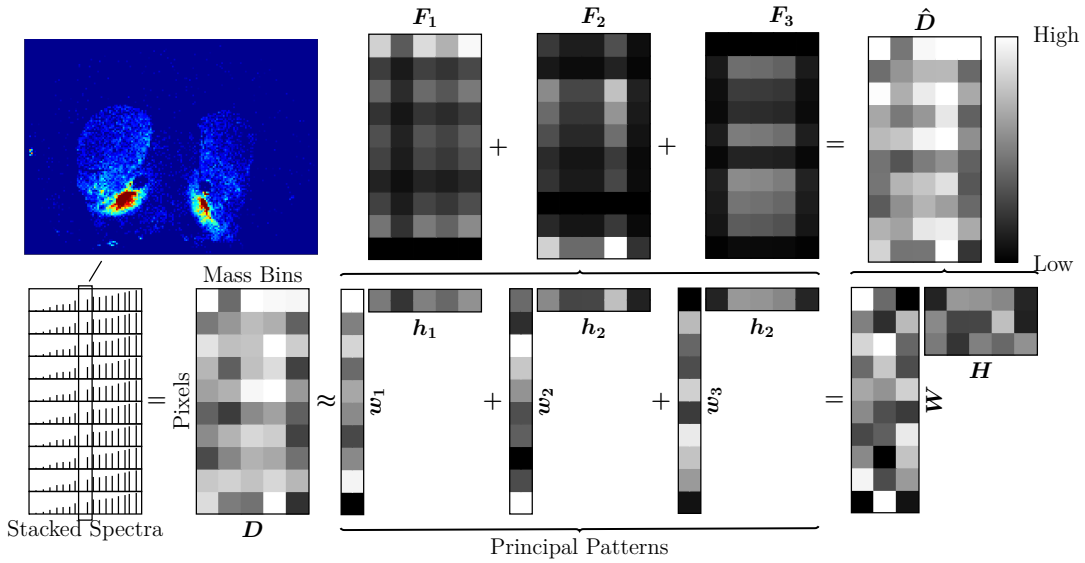


Figure 2-1: Schematic overview of a low-rank approximation of stacked pixel spectra in $\mathbf{D}$ by a decomposition into three rank-one matrices constructed by the components w_i and h_i resulting in a reduction of dimensionality. The choice of the factors w_i and h_i is dependent on the chosen dimensionality reduction algorithm.

Intensity-aware Dimensionality Reduction integrates the intensity threshold directly in the LDR process, steering towards customized components that inherently represent primarily above-threshold structure in the signal. The components w_i and h_i are constructed such that they specifically emphasize the threshold-exceeding intensities by customization of PCA. Consecutively, we intend to find a rank K for which matrix $\hat{\mathbf{M}}$ has sufficiently captured the majority threshold-exceeding intensities in $\mathbf{M}$. In chapter 4, we demonstrate a PCA-based threshold-shifted dimensionality reduction method to modify the component w_i and h_i supported by the ion intensity threshold. Subsequently, we estimate the rank based on the threshold-exceeding contribution to the explained variance.

Intensity-aware Rank Estimation uses the intensity-related information post-LDR, while Intensity-aware Dimensionality Reduction uses this information for the LDR process itself.

2-2 Datasets

2-2-1 MALDI Dataset

Over the past half-century different Imaging Mass Spectrometry (IMS) technologies with their own applications, advantages, and disadvantages have been developed. Major ones are Desorption Electrospray Ionization (DESI), Matrix Assisted Laser Desorption Ionization (MALDI), and Secondary Ion Mass Spectrometry Imaging (SIMS). In this thesis, we specifically demonstrate the application of the proposed intensity-aware dimensionality reduction techniques on a dataset obtained by MALDI IMS applied to mammalian tissue. The technique proposed in this thesis, is also readily applicable to other types of IMS.

MALDI IMS [1] is a specific IMS technique, which obtains a direct spatial mapping of ions from a tissue section by using the molecular specificity and sensitivity provided by a mass spectrometer. This mapping can obtained by the following consecutive steps. First a chemical, the matrix solution, is deposited on the sample. This matrix solution, when being irradiated with laser light, helps ionize the biological molecules in the sample. The mass spectrometer can then detect these biological ions for the irradiated pixel in question. This process is repeated pixel by pixel until a full image of the sample is obtained. A schematic overview of the process can be found in figure 1-1.

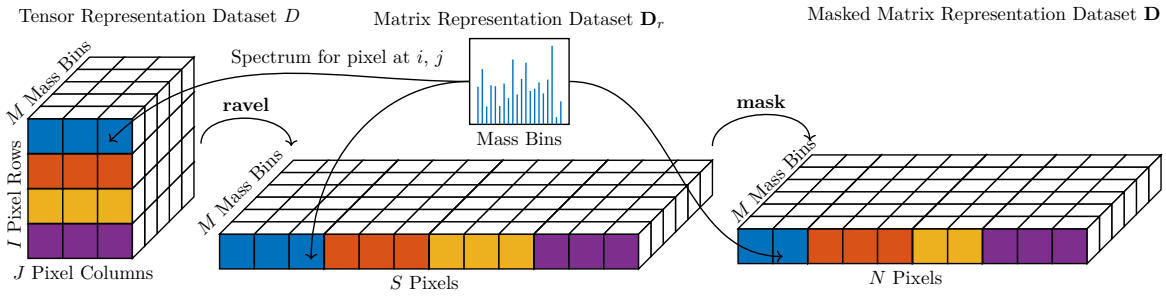


Figure 2-2: Schematic overview of the IMS dataset tensor and the matrix formulation used in this thesis.

Dataset Format

The pixel spectra measured during an IMS experiment are commonly organized as a 3-mode array, or tensor, $D \in \mathbb{R}^{I \times J \times M}$ with two spatial dimensions of respectively I pixels lines and J pixel columns and one spectral m/z dimension with M mass-bins. Each scalar value d_{ijm} in the tensor depicts the intensity in arbitrary units of a particular mass peak at a pixel-row position i and pixel-column position j in mass-bin m. A visualization of the data format is shown in figure 2-2.

We focus on dimensionality reduction based on matrix factorization and decomposition and leave the higher-order variants, such as tensor factorization, for future research. For this reason, we ravel or fold the 3-mode tensor D into a 2-mode array, or matrix, $\mathbf{D}_r \in \mathbb{R}^{S \times M}$ by reordering the pixel associated with spatial positions i and j into a long vector of length $S = I \times J$. The matrix $\mathbf{D}_r$ is a stacked representation with the S spectra, as row obtained from the I pixel lines and J pixel columns, and the M m/z bins as columns. This operation is reversible, and we can obtain the original ion images by the reverse unfolding operation.

Not all pixels in the ion images cover the tissue sample, and for this reason, we mask these empty pixels on the raveled array $\mathbf{D}_r$ into $\mathbf{D}$. The matrix $\mathbf{D} \in \mathbb{R}^{N \times M}$ only contains the N pixel spectra that report the molecular composition of the sample. Throughout the rest of this thesis, we switch between the masked representation $\mathbf{D} \in \mathbb{R}^{N \times M}$ for dimensionality reduction and the tensor representation $D \in \mathbb{R}^{I \times J \times M}$ for visualization of the distribution of molecules for a particular mass-bin. We use the notation $\mathbf{d}_{n\cdot}$ to denote the spectrum at pixel n, $\mathbf{d}_{\cdot m}$ to denote the intensities in a particular mass-bin j and d_{nm} to denote an individual intensity at pixel n in mass-bin m.

Motivation behind the ion intensity threshold

In this thesis, the ion intensity threshold defines a minimum intensity level above which the difference in the concentration between two different molecules can be reliably discerned in the mass spectral signal. The ion intensity threshold is connected to the concepts of Level of Blank (LOB), Level of Detection (LOD), and Level of Quantitation (LOQ), which together describe the smallest abundance of a measurand that can reliably be detected by an analytical procedure [43]. The ion intensity threshold is linked to LOQ, as it describes the lowest abundance at which the analyte can be reliably detected and which meets some predefined constraints for bias and imprecision [43].

In the context of IMS, the LOQ can be dependent on mass analyzer type, detector type, specimen preparation, experiment design, as well as chemical properties such as LOD and chemical noise in the specimen. For this reason, this intensity threshold is as a general guide determined by the mass spectrometer operator and falls therefore under expert and case-study-specific knowledge. The intensities below this intensity threshold are considered unreliable, biased, or imprecise, and these intensities can usually not be used to support biological conclusions. Therefore, it stands to argue that these low-intensity values are redundant for manual and computational analysis and hence the focus of this thesis on using this intensity margin to further improve rank estimation.

Coronal Rat Brain dataset preparation

In this study, we use the same IMS dataset as was used by Verbeeck et al. for connecting medical atlases to IMS measurements [44]. The study of Verbeeck et al. stated the following sample preparation steps:

> *"The brain tissue sections were collected from a rat PD model in which the adult male Sprague Dawley rats were anesthetized with isoflurane, pretreated with desmethylimipramine (12.5 mg/kg, ip) and placed in a stereotaxic frame. After an incision of the dorsal surface of the skull and placement of a burr hole, animals were again injected with despramine. 10 min later 1.5:L of 6-hydroxydopamine HBr (6-OHDA; 4.0:g/uL, free base) was unilaterally injected into the substantia nigra (AP:–5.4; L:2.3; DV:–8.4) to selectively destroy nigrostriatal dopaminergic neurons. Although there is a crossed-nigrostriatal pathway, it is quite small, contributing well under 5% of the dopamine content to the contralateral striatum, and thus the contralateral striatum is usually referred to as the intact (control) side. All in-house animal experiments were performed with approval by the Vanderbilt Institutional Animal Care and Use Committee. Brain tissue was harvested, snap frozen using liquid nitrogen, and stored at* $-80\,^{\circ}\mathrm{C}$ *until use. Frozen brain tissue was sectioned in the coronal plane at* $10\,\mu\mathrm{m}$ *using a cryostat (*$-20\,^{\circ}\mathrm{C}$*, Leica CM3050S; Buffalo Grove, IL, USA) and thaw mounted onto conductive Indium-tin-oxide coated glass slides (Delta Technologies, Loveland, CO, USA). Samples were washed to remove interfering lipids and salts in sequential washes of 70% ethanol (30 s), 100% ethanol (30 s), Carnoy's fluid (6:3:1 ethanol:chloroform: acetic acid) (2 min), 100% ethanol (30 s), water with 0.2% TFA (30 s), and 100%*

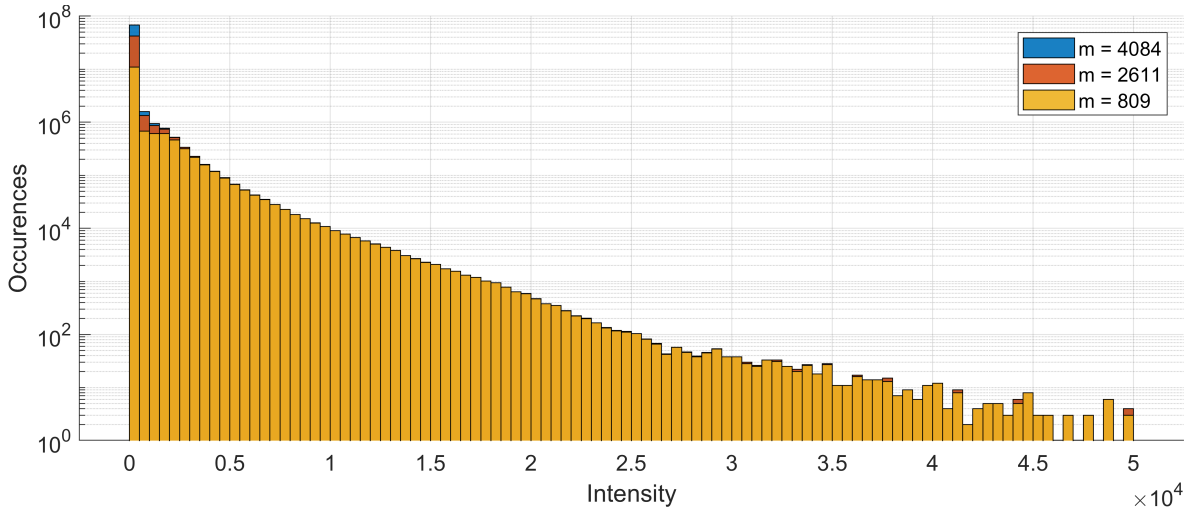


Figure 2-3: The intensity histograms of the three datasets with different number of mass-bins associated with different peak-picking thresholds of respectively 10 ($M = 4048$), 20 ($M = 2611$), and 100 ($M = 809$) in bins with width 500 truncated at intensity 5×10^4.

ethanol (30 s) [45]. The MALDI matrix 2,5-dihydroxyacetphenone (DHA, Sigma-Aldrich Chemical CO., St. Louis, MO, USA) was applied using a TM Sprayer (HTX Technologies, Carrboro, NC, USA) and rehydrated as described previously [46, 47, 45]. MALDI IMS images were collected using a 15 T Fourier transform ion cyclotron resonance (FTICR) mass spectrometer (Bruker Daltonics, Billerica, MA, USA) with a spatial sampling resolution of $75\,\mu\text{m}$ *(laser spot size* $\sim 50\,\mu\text{m}$*) and a mass resolving power of 50,000 (m/FWHM) at m/z 5000. The molecular images focused on a m/z range of 1300 to 23,000 with a total of* $\sim 20,000$ *pixels. The instrument was tuned for protein imaging as described previously [45, 46]. After IMS acquisition, matrix was removed using 100% ethanol and the tissue was stained with hematoxylin and eosin (H&E stain) for histological analysis."*

This data was then imported into MATLAB 2015b (TheMathworks Inc., Natick, MA) in which they were normalized to Total Ion Current (TIC) and peak picked with a range of thresholds 100, 20, and 10, resulting in dataset sizes with respectively 809, 2611 and 4084 mass-bins. We use the different peak picking thresholds to show the influence of the number of mass-bins on the intensity-aware rank estimate. The histograms of this dataset with the different peak picking thresholds are depicted in figure 2-3.

2-2-2 Synthetic Dataset

We use a synthetic dataset to obtain a deeper insight in, evaluate, and make a fair comparison of the intensity-aware rank estimation methods and standard rank estimation. The synthetic dataset has a known rank and ion intensity threshold and as such allows for validation of sufficient capturing in the low-rank approximation of the threshold-exceeding intensities in the intensity-aware methods. With this synthetic dataset, we aim to provide a toy-example for investigation of the influences of noise and the choice of threshold to simplify analysis.

The toy-example approach requires this synthetic dataset to be simple, neglecting many of the effects present in IMS. We intend with this synthetic dataset to only model a difference in information-content above and below the threshold, because capturing the majority of the threshold-exceeding information is the focus of this thesis. The distribution of the intensities over the intensity range affects the rank estimation as it changes the number of intensities below and above the threshold. For this reason, we use an approximation to an exponential distribution similar to the intensities in the IMS dataset, while manually introducing a region with low-reliability. Contrary to IMS, this dataset does not contain spatial autocorrelation [48]. Furthermore, for simplicity, we assume the noise is Gaussian distributed, instead of Poisson distributed as in the case of an IMS dataset [49]. These simplifications make the synthetic dataset less representative for the IMS, but we prioritize validation of the mechanism capturing the majority of the threshold-exceeding information.

A standard way of constructing a synthetic dataset for validation of rank estimation in dimensionality reduction is to define a matrix with known rank K, which we consider the signal, and let noise distort this signal matrix. As a result of this distortion, the final measurement matrix becomes full rank. Then with rank estimation, we intend to find the underlying signal rank K again from the distorted measurements. In the case of intensity-aware rank estimation, we extend this procedure to specifically model the difference in information content in above versus below-threshold intensities in the underlying signal matrix. For this reason, we have formulated a synthetic dataset directly in matrix form using:

$$\mathbf{D}_{\text{syn}}^{\text{orig}} = f_{\text{perturbation}}(\mathbf{X}_{\text{signal}}) + \sigma_{\text{noise}}\mathbf{X}_{\text{noise}} \tag{2-3}$$

in which $\mathbf{X}_{\text{signal}}$ denotes the low-rank matrix representing the biological signal, $f_{\text{perturbation}}(\bullet)$ denotes a function to model intensity dependent perturbations, $\mathbf{X}_{\text{noise}}$ denotes Gaussian distributed noise, and σ_{noise} defines the variance of this noise.

In this formulation, we intend the effect of the threshold to be dominant over the noise intensity. To assert every intensity is positive after application of the Gaussian noise we set any negative value for pixel n and mass-bin m in $\mathbf{D}_{\text{syn}}^{\text{orig}}$ to zero:

$$d_{\text{syn},nm} = \begin{cases} d_{\text{syn},nm}^{\text{orig}} & d_{\text{syn},nm}^{\text{orig}} \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2-4}$$

In the next paragraphs, we discuss the specific choices on the signal construction, noise perturbations and the dataset size to construct the dataset with the histogram shown in figure 2-4.

Signal The spectral signal is represented by a rank-K matrix and contains the principal patterns that we aim to unravel with dimensionality reduction. The histograms in figure 2-3 show that IMS datasets are generally sparse, containing many low-intensity values and a limited number of high intensities. To obtain a similar distribution and sparsity, we have modeled the low-rank signal as the multiplication of two exponentially distributed matrices $\mathbf{W}$ and $\mathbf{H}$. The exponential distributions ensure that the dataset can have a limited number of threshold-exceeding intensities and a large quantity of below-threshold intensities, depending on the choice of intensity threshold.

$$\mathbf{X}_{\text{signal}} = \alpha\mathbf{W}\mathbf{H} \tag{2-5}$$

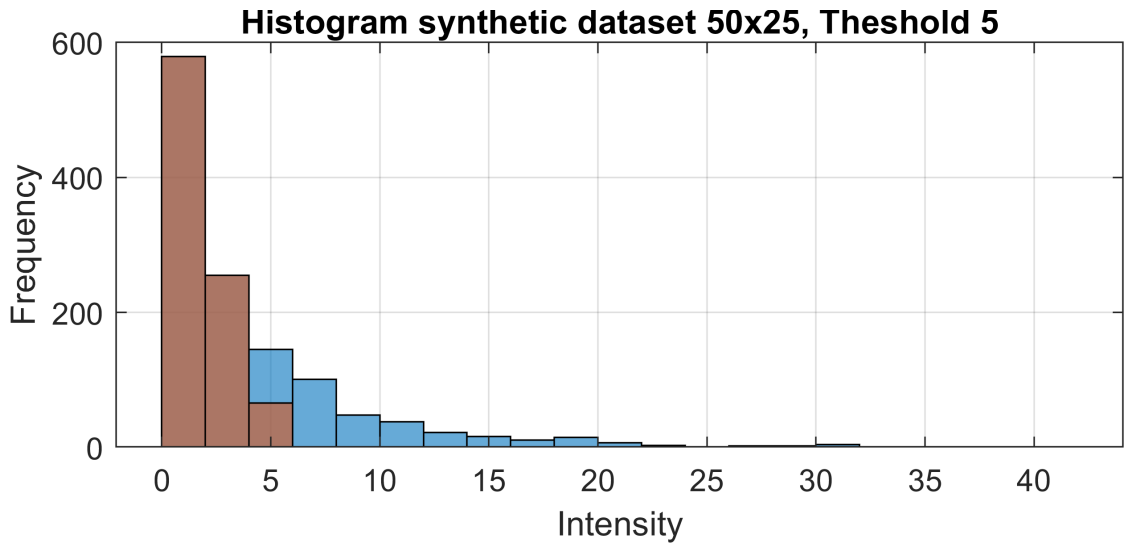


Figure 2-4: The histogram of a synthetic dataset $\mathbf{D}_{\text{syn}} \in \mathbb{R}^{50\times 25}$ with $\text{rank}\,(\mathbf{D}_{\text{syn}}) = 5$, SNR $= \infty$, $\alpha = 10$ and 10% sparsity. The intensities below the ion intensity threshold $\tau = 5$ are shown in red.

with $\mathbf{X}_{\text{signal}} \in \mathbb{R}^{N\times M}$, $\text{rank}\,(\mathbf{X}_{\text{signal}}) = K$, $\mathbf{W} \in \mathbb{R}^{N\times K}$, $\mathbf{W} \sim \mathcal{E}(1)$, $\mathbf{H} \in \mathbb{R}^{K\times M}$ and $\mathbf{H} \sim \mathcal{E}(1)$ in which $\sim \mathcal{E}(\bullet)$ denotes an exponentially distributed signal. We use α to scale the intensities, such that a distribution of threshold-exceeding and below-threshold relative to the selected threshold can be chosen. Additional sparsity is added by randomly setting a percentage of entries in either the $\mathbf{W}$ and $\mathbf{H}$ matrix to zero. The multiplication of two exponential distributions is not exponential. However, figure 2-4 demonstrates the manually imposed sparsity on the matrices $\mathbf{W}$ and $\mathbf{H}$ results in a close approximation of the exponential distribution.

Perturbations The function $f_{\text{perturbation}}$ is intended to distort the below-threshold intensities and break the patterns apparent in the signal matrix $\mathbf{X}_{\text{signal}}$ in the below-threshold region. The threshold-exceeding intensities are left undistorted. To model the distortion, we have chosen to randomly permute the entries in the below-threshold region. The permutations break the patterns apparent in the dataset, while keeping the exponential distribution in the total dataset $\mathbf{X}_{\text{signal}}$ intact.

$$f_{\text{perturbation}}(x) = \begin{cases} \text{Random sample of } T & x \leq \tau \\ x & \text{otherwise} \end{cases} \tag{2-6}$$

where τ is the intensity threshold, and $T = \{X_{\text{signal}} \leq \tau\}$ is the set of elements of X_{signal} below threshold τ.

This perturbation model ensures that the below-threshold intensities become unstructured, while the threshold-exceeding intensities remain intact. The model assumes no structure in the below threshold region. We are aware that this perturbation model is not realistic for a real IMS dataset, which we expect to contain a continuous transition from more unreliable to more reliable intensities. However, we have chosen this approach to demonstrate intensity-aware methods at work in this simple case and simplify the analysis of the toy example.

Extensions could be made to this perturbation model to support an intermediate situation in which below-threshold intensities still contain some structure.

An example of such an extension is to do only permutations of intensities within a similar intensity range. These permutations limit the randomness of the below-threshold intensity dependent on what intensity range the intensity is in. The creation of several below-threshold intensity ranges in decreasing size with increasing intensity allows a more gradual increase in reliability, as a result of the size of the intensity range. However, verification of the broken patterns in below-threshold intensities as a result of this perturbation is not straightforward. It is nontrivial to choose the intensity ranges in such a manner that randomness of the below-threshold is still dominant over the randomness as a result of the added noise. For this reason, we have chosen the simplified perturbation model and see more natural modeling of below-threshold perturbations as a topic for future research.

Noise The addition of intensity independent noise ensures that also threshold-exceeding intensities are perturbed and that the to be captured pattern is somewhat distorted. The additive noise allows independent modification of the strength of the intensity-independent noise and the introduced intensity-dependent perturbations. The noise strength is chosen on a fixed signal-to-noise ratio based on the original signal in equation (2-5) with:

$$\sigma_{\text{noise}} = \frac{\|f_{\text{pertubation}}(X_{\text{signal}})\|_F}{\sqrt{\text{SNR}}\,\|X_{\text{noise}}\|_F} \tag{2-7}$$

with $\|\bullet\|_F$ representing the Frobenius norm.

Size Dimensionality reduction requires sufficient support for the patterns in the threshold-exceeding intensities. As a result, we have chosen the synthetic dataset size based on sufficient sampling of the threshold-exceeding region, while at the same time enabling thresholds to be larger than the noise intensity originating from the Signal-to-Noise Ratio (SNR). High thresholds for the synthetic dataset, and as such little proof for the patterns we aim to capture cause LDR to fail to capture the majority of the pattern. However, low thresholds are potentially in the noise region and hamper objective validation of the threshold-aware LDR techniques.

Empirically, we have found that a dataset with $N = 50$ rows, $M = 25$ columns, and rank $K = 5$ provides sufficient sampling above the threshold, choices of threshold, and choices of SNR. The histogram of an example dataset is shown in figure 2-4. Throughout the rest of this thesis we investigate effects of the variation of SNR and threshold choice on rank estimation. Other parameters are fixed at $\alpha = 10$ and 10% sparsity.

Chapter 3

Intensity-aware Rank Estimation

In this chapter, we focus on intensity-aware rank estimation, which intends to estimate the rank using unmodified Linear Dimensionality Reduction (LDR), while capturing the majority of the intensity-threshold-exceeding information. It attempts to use the intensity threshold to pick a set of components the preferentially represent above-threshold structure. This approach uses the intensity-related information post-LDR whereas intensity-aware dimensionality reduction, discussed in the next chapter, uses this information for the LDR itself.

This chapter introduces traditional LDR techniques Principal Component Analysis (PCA) and Nonnegative Matrix Factorization (NMF) and shows that rank estimation solely based on a choice of intensity threshold for PCA and NMF can be ambiguous.

3-1 Linear dimensionality reduction in Imaging Mass Spectrometry (IMS)

3-1-1 Principal Component Analysis (PCA)

PCA is a popular multivariate analysis technique used for dimensionality reduction, based on capturing the majority of variation in datasets consisting of possibly correlated variables. The dimensionality reduction aspect is based on the transformation of these original variables to a new and limited set of uncorrelated principal components, ordered by the amount of variance they represent [40].

In the context of IMS, PCA is used to reduce the dimensionality of IMS datasets [18], in which the principal components can be viewed as a set of uncorrelated biochemical trends within the specimen [2]. PCA is also used for denoising, the identification of linearly dependent molecules, and detection of spatial regions of interest [19, 15, 18].

In PCA, dimensionality reduction is achieved by a projection of the dataset onto a lower dimensional orthogonal basis. The vectors in this basis, spanning a lower-dimensional latent

subspace, are the consecutive axes of maximum variance. As a result, the maximum variance of the measurements possibly represented in the K-dimensional space is obtained by a projection of the data on the first K latent basis vectors.

This basis aligned with the axes of maximum variance is identical to the change of basis matrix $\mathbf{C}$ used to diagonalize the covariance matrix $\mathbf{S}$ of the measurements, or in other words, the eigenvectors of the covariance matrix $\mathbf{S}$. This covariance matrix is defined as:

$$\mathbf{S} = \frac{1}{N-1}(\mathbf{D} - \overline{\mathbf{D}})^T(\mathbf{D} - \overline{\mathbf{D}}) \quad \text{or} \quad \mathbf{s}_{pq} = \frac{1}{N-1}\sum_{n}^{N}(d_{np} - \overline{d_p})(d_{nq} - \overline{d_q}) \tag{3-1}$$

in which the matrix $\mathbf{S}$ denotes the covariance matrix, s_{pq} denotes the covariance between variables or mass-bins p and q, $\mathbf{D} \in \mathbb{R}^{N\times M}$ denotes the IMS data matrix with N observations or pixels and M variables or mass-bins, $\overline{\mathbf{D}}$ denotes the column or mass-bin average, $\overline{d_p}$ and $\overline{d_q}$ denote respectively the average of the columns or mass-bins p and q, and d_{np} and d_{nq} denote the intensity at pixel j for respectively mass-bin p and mass-bin q. As a result, the individual entries of the covariance matrix s_{pq} are dependent on the combination of the deviations of the mean from mass-bin p and mass-bin q.

Overall, the procedure of PCA-based dimensionality reduction is to construct the covariance matrix $\mathbf{S}$, obtain the basis aligned with the axes of maximum variance $\mathbf{C} \in \mathbb{R}^{M\times M}$ and project the deviations from the mean $\mathbf{D}^* = \mathbf{D} - \overline{\mathbf{D}}$ on a subset of basis vectors. PCA is usually implemented as a Singular Value Decomposition (SVD) of the mean centered data $\mathbf{D}^*$ matrix or eigenvalue decomposition of the covariance matrix $\mathbf{S}$ [40].

Dimensionality Reduction and Rank Estimation

This section extends the mathematical concept of rank as introduced in section 2-1 for PCA and introduces the terminology used in the rest of the thesis. Similar to the definition of rank in section 2-1, PCA decomposes the IMS measurement matrix $\mathbf{D}^* \in \mathbb{R}^{N\times M}$ for rank M into M rank-one matrices or factors $\mathbf{F}$ so that they sum up to the original matrix $\mathbf{D}^*$:

$$\mathbf{D} = \sum_{i=1}^{M}\mathbf{F}_i = \sum_{i=1}^{M} w_i h_i = \mathbf{WH} \tag{3-2}$$

Generally, in PCA, the column vectors w_i and the row vectors h_i are commonly named the score vectors and the loading vectors respectively. The scores and loadings for a rank-K reduction are constructed by taking K basis vectors $\mathbf{C}_K \in \mathbb{R}^{M\times K}$ from the basis $\mathbf{C}$ obtained from the covariance matrix $\mathbf{S}$. The loading vectors are the first K row vectors $\mathbf{H} = \mathbf{C}_K^T$ of the basis matrix $\mathbf{C}^T$. The score vectors are the projection of the deviations from the mean $\mathbf{D}^*$ on the K basis vectors $\mathbf{C}_K$ so that we obtain $\mathbf{W} = \mathbf{D}^*\mathbf{C}_K$. Utilizing the matrix notation, the decomposition of the matrix $\mathbf{D}^*$ into K principal components can be represented as:

$$\mathbf{D}^* = \mathbf{D}_K^* + \mathbf{E} = \mathbf{WH} + \mathbf{E} = \mathbf{D}^*\mathbf{C}_K\mathbf{C}_K^T + \mathbf{E} = \sum_{i=1}^{K}\mathbf{D}^* c_i c_i^T + \mathbf{E} \tag{3-3}$$

in which $\mathbf{E}$ denotes the residuals between the rank-K approximation $\mathbf{D}_K^*$ and $\mathbf{D}^*$. The rank-K approximation $\hat{\mathbf{D}}_K$ of the IMS matrix $\mathbf{D}$ can be obtained by addition of the means $\hat{\mathbf{D}}_K = \hat{\mathbf{D}}_K^* + \overline{\mathbf{D}}$.

In PCA, each column vector c_i in $\mathbf{C}$ is commonly referred to as a principal component of the matrix $\mathbf{D}$ and is associated with a coefficient λ_i indicating the explained variance by this component. These coefficients stem from the eigenvalues of the covariance matrix $\mathbf{S}$ and order the individual principal components c_i based on the represented variance. As a result, the current principal component always represents more variance than the next one.

As introduced before, we are interested in selecting the number of vector pairs w_i or h_i, or rank-one matrix $\mathbf{F}_i$, to obtain a lower dimensional representation of the matrix $\mathbf{D}$ to capture the majority of the threshold-exceeding information. In the context of PCA, this means an estimate of the number of basis vectors, columns of $\mathbf{C}$, as a result of the relation of the vector pairs w_i or h_i with the basis vectors c_i shown in equation (3-5). Due to the association of every component $\mathbf{F}_i$ with a coefficient λ_i in PCA, selecting a rank implies capturing a percentage of the total variance due to the associated explained variance coefficient with each principal component. This total captured variance is inversely related to the magnitude of the total residual between the low-rank approximation and the original dataset by definition of PCA [40]. For this reason, we investigate in section 3-2 the relation of the intensity and rank.

Variable Weights

The covariance matrix used for constructing the basis $\mathbf{C}$ is not scaling invariant, meaning the different magnitudes of the individual entries d_{np} and their deviations from the mean $\overline{d_p}$ determine the magnitude of the covariance matrix entries c_{pq}. An alternative is using the correlation matrix $\mathbf{R}$ instead of the covariance matrix $\mathbf{S}$. The correlation matrix is a normalization of the covariance matrix such that covariances s_{pq} are normalized by the standard deviations of the variables p and q to correlations $r_{pq} \in [-1, 1]$. As pointed out in [2], this normalization makes sense in the case of mostly heterogeneous data with different amplitudes, but not in the case of spectrometry where the data has a physical relationship and are directly comparable. In this thesis especially, we are interested in capturing the threshold-exceeding peaks and as a result the relative scale of the individual variables.

The two approaches towards intensity-aware rank estimation outlined in section 2-1 can be extended for PCA. In this chapter, we assume the variable weights in the covariance matrix as-is and intend to estimate the rank without modification of the variable weights. In chapter 4 we explicitly modify the variable weights to emphasize threshold-exceeding intensities.

Optimization Formulation

To obtain a similar problem formulation as NMF, PCA can also be formulated as an optimization problem. The result obtained with the SVD of mean-centered matrix $\mathbf{D}^*$ or the eigenvalue decomposition of the covariance matrix $\mathbf{S}$ is the guaranteed global optimum of this optimization problem [50]. The approach of finding the linear subspace that maximizes the variance is equivalent to finding the linear subspace that minimizes the least-squares divergence of the projection $\mathbf{C}$ [40].

$$\begin{aligned} \underset{C}{\text{minimize}} \quad & \left\| \mathbf{D}^* - \mathbf{D}^* \mathbf{C} \mathbf{C}^T \right\|_F^2 \\ \text{subject to} \quad & \mathbf{C}^T \mathbf{C} = \mathbf{I} \end{aligned} \tag{3-4}$$

3-1-2 Nonnegative Matrix Factorization (NMF)

NMF was first developed by Paatero and Tapper [51] in the form of Positive Matrix Factorization and later extended as NMF by Lee and Seung [42]. NMF factors the IMS measurement matrix $\mathbf{D}$ in two strictly nonnegative components $\mathbf{W} \in \mathbb{R}^{N\times K}$ and $\mathbf{H} \in \mathbb{R}^{K\times M}$. In a similar sense to, equation (2-2), NMF can reduce the dimensionality by choosing rank K lower than the respective dimensions M and N, resulting in a lossy compression of the original matrix according to:

$$\mathbf{D} \approx \hat{\mathbf{D}} = \sum_{i=1}^{K} \mathbf{F}_i = \sum_{i=1}^{K} w_i h_i = \mathbf{WH} \quad \text{with} \quad \mathbf{W} \geq 0, \mathbf{H} \geq 0 \tag{3-5}$$

As a result of the nonnegativity, the approximation to the IMS measurement matrix $\hat{\mathbf{D}}$ is composed by additions of parts [42]. In the context of IMS dataset $\mathbf{D}$, NMF can be interpreted as unmixing of pixel spectra as combinations of spars spectral sub-signatures h_i. In traditional NMF, these sub-signatures are typically ion spectra describing one or a set features appearing in one or more pixels. These pixels then show a membership w_i to these signatures, defining the pixel as a composition of sub-signatures. The nonnegativity constraint allows a more biologically valid representation of these sub-signatures than PCA. Due to this grouping and the nonnegativity constraint, NMF is a relatively interpretable decomposition for nonnegative measurement data contrary to PCA, which can yield negative-valued signatures.

NMF constructs factors $\mathbf{W}$ and $\mathbf{H}$ from a matrix $\mathbf{V}$ by solving the optimization problem:

$$\begin{aligned} \underset{\mathbf{W},\mathbf{H}}{\text{minimize}} \quad & D\left(\mathbf{V}, \mathbf{WH}\right) + R(\mathbf{V}, \mathbf{W}, \mathbf{H}, \dots) \\ \text{subject to} \quad & \mathbf{W} \geq 0 \\ & \mathbf{H} \geq 0 \end{aligned} \tag{3-6}$$

in which $D(\mathbf{A}, \mathbf{B})$ denotes the divergence function between $\mathbf{A}$ and $\mathbf{B}$ and $R(\mathbf{V}, \mathbf{W}, \mathbf{H}, \dots)$ denotes the additional regularization terms. $\mathbf{V} \in \mathbb{R}^{N\times M}$ is the matrix to be factorized matrix, $\mathbf{W} \in \mathbb{R}^{N\times K}$ and $\mathbf{H} \in \mathbb{R}^{K\times M}$ and $K \leq \min(N, M)$ are the factors. In the default case, the divergence function is commonly chosen as the Euclidian distance between $\mathbf{V}$ and $\mathbf{WH}$ and no regularization terms $R(\mathbf{V}, \mathbf{W}, \mathbf{H}, \dots) = 0$ are included, resulting in:

$$D\left(\mathbf{V}, \mathbf{WH}\right) = \|\mathbf{V} - \mathbf{WH}\|_F^2 = \sum_{n=1}^{N}\sum_{m=1}^{M} \left([\mathbf{V}]_{nm} - [\mathbf{WH}]_{nm}\right)^2 \tag{3-7}$$

The Euclidian Distance NMF optimization problem can be solved via Nonnegative Least Squares (NNLS) with a multiplicative update algorithm [52] or alternatively via Alternating Least Squares (ALS) [53], Hierarchical Alternating Least Squares (HALS), or Block Coordinate Descent (BCD) algorithms [53]. For the results in this thesis, we have used the multiplicative update algorithm:

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{(\mathbf{VH}^T)}{(\mathbf{WHH}^T)} \qquad \mathbf{H} \leftarrow \mathbf{H} \odot \frac{(\mathbf{W}^T\mathbf{V})}{(\mathbf{W}^T\mathbf{WH})} \tag{3-8}$$

in which $\odot$ denotes the element-wise product.

3-2 Linear dimensionality reduction: intensity-dependent capturing

In this section, we investigate how PCA and NMF capture different intensity ranges in a dataset, spread over the different ranks. We propose that a rank estimate based on the prior defined intensity threshold is required to report a rank for which the threshold-exceeding intensities are sufficiently captured in the corresponding low-rank approximation. In linear dimensionality reduction, the discrepancy between the original dataset and the low-rank approximation is the reconstruction error or the residuals. In the case of IMS, these residuals depict quantitatively what intensities or part of intensities have been discarded from the spectra by reducing the dimensionality. In this section, we use the residuals of the rank-K approximation to quantify what intensities have been captured in the corresponding low rank-approximation:

$$\mathbf{E}_K = \mathbf{D} - \hat{\mathbf{D}}_K \tag{3-9}$$

in which $\mathbf{E}_K \in \mathbb{R}^{N \times M}$ denotes the residuals for the rank-K approximation, $\mathbf{D} \in \mathbb{R}^{N \times M}$ is the matrix representation of the IMS measurements and $\hat{\mathbf{D}}_K \in \mathbb{R}^{N \times M}$ is the rank-K approximation of these measurements.

In the case of intensity-aware rank estimation, we prioritize the reduction of residuals associated with threshold-exceeding intensities. For this reason, we bin these residuals $\mathbf{e}_{nm}$ in $\mathbf{E}_K$ into intensity windows $\epsilon_{b-\delta,b+\delta}$ of width 200 ($\delta = 100$) by the corresponding intensity value found in the original matrix d_{nm} where n denotes the pixel and m the mass-bin. $\epsilon_{b-\delta,b+\delta}$ denotes the set of residuals in the window in intensity range from $b - \delta$ to $b + \delta$, in which δ defines the window width and b the intensity bin. These intensity windows $\epsilon_{b-\delta,b+\delta}$ quantify how well the low-rank approximation captures peaks in the spectra with respect to their intensity. Moreover, these intensity windows describe how the distribution of these residuals over the intensity range changes with a change of the chosen rank. The distribution of these residuals allows us to determine if there is a relation between the rank-K approximation and the intensity of the captured peaks. Consequently, we can observe how well traditional PCA and NMF capture the threshold-exceeding peaks in the spectra.

Throughout this section we use the notion of bin and window in two ways:

Mass-bin The intensity values, or ion count, in one or more mass spectra in the dataset $\mathbf{D}$ associated with a particular ion mass.

Intensity window A set values, in this case residuals, associated with individual intensities d_{nm} in one or more mass spectra in the dataset $\mathbf{D}$, grouped by their original intensity d_{nm}.

We use the per-intensity window Root Mean Squared Residual (RMSR), $\text{rms}(\epsilon_{b-\delta,b+\delta})$, and Median Absolute Residual (MAR), $\text{median}(|\epsilon_{b-\delta,b+\delta}|)$, to construct a scalar value, representing the magnitude of the residuals in an intensity window. The RMSR is a notion for the dispersion of the residuals in an intensity window, while the MAR reflects for the magnitude of the residuals. We investigate both the RMSR with the MAR of the intensity windows to gain more insight into the distribution of the residuals in the intensity windows. For example, a large discrepancy between RMSR and MAR could demonstrate a few large residuals

dominate the residuals in the intensity window. Alternatively, a smaller discrepancy between RMSR and MAR could suggest more equally distributed residuals.

Another reason to combine MAR and RMSR is a comparison between intensity windows independent of the number of peaks in the respective windows. All intensity windows $\epsilon_{b-\delta,b+\delta}$ have width 2δ. Consequently, the number of peaks in all intensity windows is not equal. The majority of the peaks in the IMS dataset are in the low-intensity windows, as shown in the histograms in figure 2-3. An unequal peak-count in the window could possibly distort the comparison of the residuals via RMSR over the different intensities. For example, the low-intensity windows containing many peaks are less sensitive to outliers than the high-intensity windows, containing fewer peaks. The opposite is also valid. A substantial group of large residuals could vanish in the total residual, due to the number of peaks in the intensity windows. With the comparison of both measures, we aim to mitigate these issues.

In the context of intensity-dependent capturing of the data structure by PCA and NMF, we are most interested in the differences in the residuals in the region around the threshold. For this reason, we have focused on the intensity region $[0, 10^4]$ and intensity windows of equal width $\delta = 200$.

3-2-1 PCA intensity-dependent capturing

Figure 3-1 displays the RMSR and MAR of the binned residuals $\epsilon_{b-\delta,b+\delta}$ for the three IMS datasets with respectively 809, 2611, and 4048 mass-bins. The following paragraphs describe several observations on the residuals in these intensity windows to understand the mechanics of intensity capturing in PCA. Consecutively, we discuss the implications of these observations on intensity-aware rank estimation.

High RMSR and MAR for high intensities in low-rank approximations. All graphs in figure 3-1 demonstrate that both the RMSR and MAR for high intensities ($d_{nm} \geq 4000$) are relatively high compared to their low-intensity ($d_{nm} < 4000$) counterparts for low-rank approximations of $K \leq 10$, $K \leq 10$, $K < 16$ for respectively the datasets with 809, 2611, 4084 mass-bins.

The high RMSR and MAR for high intensities for low-rank approximations originate from the nature of PCA. In the strictly nonnegative IMS dataset, particular ion bins with high intensities and associated high dispersion contribute more variance than lower intensities. As PCA obtains a low-rank approximation by projecting the dataset on the consecutive axes of maximum variance, PCA is expected to mainly describe intensities contributing large portions of variance. In the case of the low-rank approximations for high RMSR and MAR for high intensities, we expect that the majority of the variance has not been captured. As such, a portion of the high-intensity peaks in the spectra have not been fully captured, resulting in large residuals for these high-intensity peaks.

RMSR and MAR decreasing more for high intensities with increasing rank. For high intensities ($d_{nm} \geq 2000$), both the binned RMSR and binned MAR decrease at a higher rate for increasing rank when compared to the low intensities ($d_{nm} < 2000$)

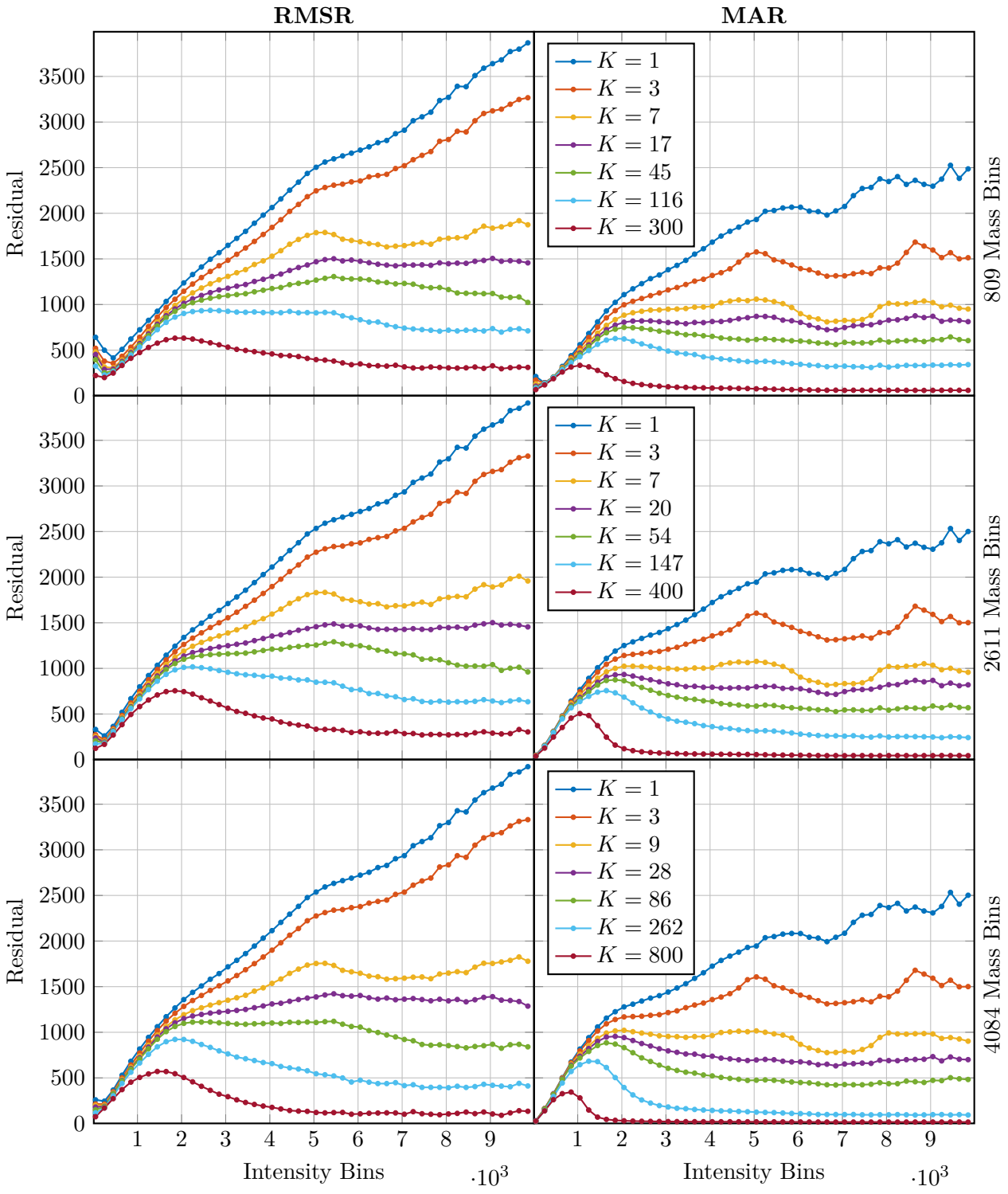


Figure 3-1: The binned Root Mean Squared Residual (RMSR) and Median Absolute Residual (MAR) obtained from the residuals $\mathbf{E}_K = \mathbf{D} - \hat{\mathbf{D}}_K$ between the original dataset $\mathbf{Y}$ and the rank-K reconstruction $\hat{\mathbf{D}}_K$ obtained via PCA for three different dataset sizes. These residuals are binned by the original intensity in $\mathbf{D}$ in 50 bins with equivalent width in the range of $[0, 10^4]$. The graphs display RMSR and MAR per bin to give an idea about how the reconstruction of different intensities in the spectrum evolve for different ranks.

We believe the rate of decrease in the binned RMSR and binned MAR can be partially attributed to the different amount of peaks in respectively the high and low-intensity windows, as discussed in the previous section.

Furthermore, in a similar sense to the previous paragraph, we expect a mass-bin with high-intensity peaks to individually contribute more variance when compared to a mass-bin with low-intensity peaks. Consequently, we expect that these values are reconstructed in the first set of PCA components, representing large quantities of variance. As such, RMSR could be initially high and reducing at a high rate in the first set of components.

Near-zero RMSR and MAR for the near-zero intensity values. All graphs demonstrate in figure 3-1 that both the RMSR and MAR for near-zero intensity values ($d_{nm} < 1000$) in the spectra are close to zero. The near-zero residuals suggest near-perfect capturing for the near-zero intensity values independently of the chosen rank, but there are few caveats.

First, the magnitude of the residuals itself has a relation to the associated intensity. A missing high-intensity peak results in a larger residual than a low-intensity peak. The near-zero intensity windows contain predominantly low-intensity peaks. As such, when PCA insufficiently captures intensities in a particular mass-bin, we expect the residuals from the near-zero intensities to be small.

Second, the near-zero intensity windows $\epsilon_{b-\delta,b+\delta}$ for $b \leq 1000$ contain the majority of the intensities the IMS dataset, as shown in figure 2-3. Due to the number of intensities in the near-zero windows, the contribution of these intensity windows to the total sum of squared residuals can be substantial when compared to the more high-intensity windows ($d_{nm} > 1000$), even when the RMSR and MAR are small. In other words, these near zero intensity windows could still account for the majority of the total discrepancy between the original dataset and the low-rank estimate.

Third, we expect the influence on the RMSR and MAR of few large residuals in a large pool of low residuals to be limited, due to a large number of peaks in the low-intensity windows. As such, additional components, decreasing these large residuals, reduce the RMSR and MAR for the low-intensity only little when compared to intensity windows containing a low number of peaks.

Bump in the RMSR and MAR for high ranks. For low intensities ($d_{nm} < 2000$) and high rank, respectively $K \geq 31$, $K \geq 34$, $K \geq 63$, both the binned RMSR and binned MAR have decreased at a higher rate for increasing rank when compared to the low intensities ($d_{nm} < 2000$). As a result, a bump appears in both the binned RMSR and the binned MAR for low-intensity windows. This bump in residual fraction becomes more visible in the larger dataset sizes. For even higher ranks, respectively $K \geq 96$, $K \geq 108$, $K \geq 251$, the bump in both the binned RMSR and binned MAR for low-intensity peaks moves towards the lowest intensity-window and decreases slowly. At the same time, the high-intensity values have reached very low residuals, which we can interpret as near perfect capturing.

We hypothesize that these intensities are largely noncovariant and related to the low-reliability of intensities below the ion intensity threshold. We expect that these incoherent patterns do not appear in a large number of mass-bins or other more high-intensity mass-bins. Consequently, these low-intensity peaks contribute little in terms of covariance compared to more

coherent patterns. As a result, PCA requires a significant number of principal components to capture these patterns in the first set of principal components.

3-2-2 NMF intensity-dependent capturing

Figure 3-2 displays the RMSR and MAR of the binned residuals $\epsilon_{b-\delta,b+\alpha}$ in 50 bins of equal width with $\delta = 100$ in the range $[0, 10^4]$ for a set of different ranks and for the two IMS datasets with respectively 809 and 2611 mass-bins. This section zooms in on the differences in these intensity-binned residuals with PCA between figure 3-2 and figure 3-1 to understand the mechanics of intensity capturing in NMF.

Similar to PCA, we see for NMF initially high residuals for high-intensity windows as the result of insufficient rank for NMF to capture these high intensities. However, at higher ranks (8+) we see the MAR become approximately flat over the 1000+ intensities, whereas PCA shows a higher reduction in residuals for these intensities. At the same time, the RMSR shows a slightly positive relation with intensity, meaning high intensities have a slightly higher RMSR than the lower intensities. Furthermore, for even higher ranks (24+) the decrease in RMSR and MAR is limited compared to the initial decrease. PCA shows a more significant reduction in residuals.

Two potential causes for these effects are the less-overfitting behavior of NMF, due to the nonnegativity constraint and associated sparsity [54], and the non-convexity and increased ill-posedness of NMF due to the enlarged parameter space associated with the higher rank [24]. As a result of the less-overfitting behavior, the additional components do not necessarily capture noise, but instead, improve the capturing of distinct patterns by for example partitioning these patterns spatially or spectrally. As a result of the increased ill-posedness, NMF is more likely to converge to local minima at higher ranks. A clear example of this non-convexity is rank $K = 300$ resulting in higher residuals than rank $K = 200$.

3-2-3 Residuals and intensity-aware rank estimation

The RMSR and MAR for PCA in figure 3-1 and NMF in figure 3-2 both demonstrate that a rank resulting in a reasonable reduction of dimensionality for which we obtain zero residuals of the threshold-exceeding intensities, in other words, perfect capturing of these intensities, is non-existent for PCA and NMF for an arbitrary choice of threshold. Consequently, traditional dimensionality reduction and estimating rank in an intensity-aware manner requires an approximation of the threshold-exceeding intensities.

At the same time, the residuals show no clear relationship between an arbitrary reliability threshold on the intensity and sufficiently small residuals for all threshold-exceeding intensities. We observe a substantial decrease in the residuals in the high-intensity windows, but the rank at which we have sufficiently small residuals for the threshold-exceeding intensities is uncertain. The absence of this relation suggests another measure is required to identify sufficient capturing of the threshold-exceeding intensities in the low-dimensional abstract subspace for both NMF and PCA.

The observations on the residuals in the intensity windows in the previous paragraph lead to the following questions. The seemingly asymptotically decreasing rate in reduction of the

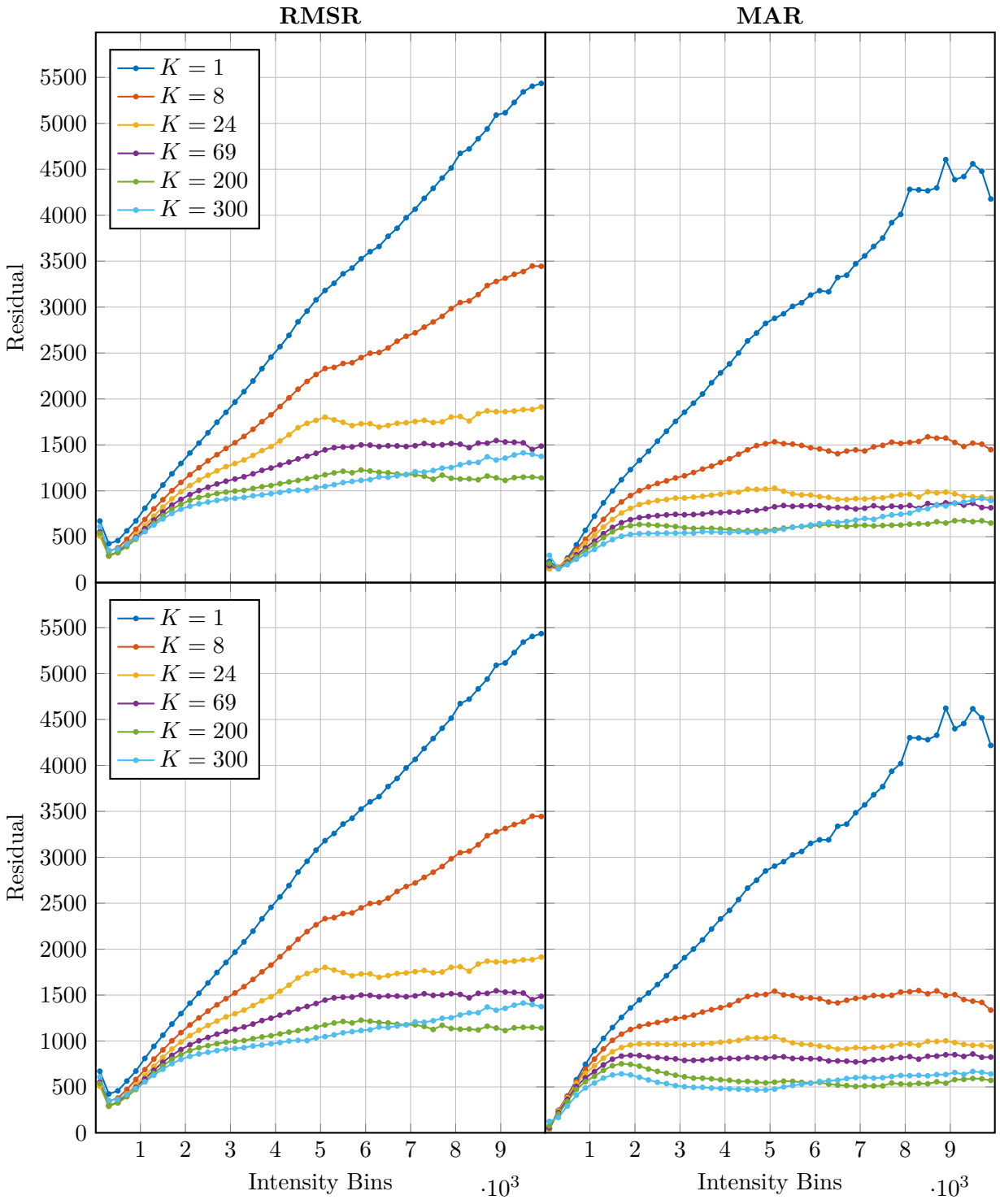


Figure 3-2: The binned Root Mean Squared Residual (RMSR) and Median Absolute Residual (MAR) obtained from the residuals $\mathbf{E}_K = \mathbf{D} - \hat{\mathbf{D}}_K$ between the original dataset $\mathbf{Y}$ and the rank-K reconstruction $\hat{\mathbf{D}}_K$ obtained via NMF for two different dataset sizes. These residuals are binned by the original intensity in $\mathbf{D}$ in 50 bins with equivalent width in the range of $[0, 10^4]$. The graphs display RMSR and Median Absolute Residual (MAR) per bin to give an idea about how the reconstruction of different intensities in the spectrum evolves for different ranks.

residuals associated with threshold-exceeding intensities extends the question of intensity-aware rank estimation with at what rank is the capturing of the threshold-exceeding peaks by traditional PCA sufficient and the residuals sufficiently small? This hypothesis can be summarized in the following question.

- *When is the low-rank capturing of the threshold-exceeding intensities sufficient to select rank K over rank $K+1$ to capture the majority of the threshold-exceeding information?*

In the case of PCA, we observe a bump in the residuals at a fixed low-intensity region, for which the RMSR and MAR have reduced more slowly than for the high-intensities. The appearance of this bump in RMSR and MAR raises the question if the rank could be estimated in a data-driven manner, if the low-reliability of the below-threshold intensities reveals itself as the bump in the intensity windows. This hypothesis can be summarized in the following question.

- *Gives the difference of the above-threshold residuals over the below-threshold residuals a reason to select rank K over rank $K+1$ to obtain low-rank approximate containing maximal information of threshold-exceeding intensities?*

To answer these questions based on these observations and considerations we have formulated two approaches to relate rank defined in the abstract latent space, constructed by PCA, to the ion intensity threshold existing in the original measurement space.

Data-data driven approach In the data-driven approach we continue with the principal components as constructed by unmodified PCA. In this approach, we assume the existence of the threshold is based on its presence in the form of an unstructured region in the measurement space of these datasets. As a result of this non-structure, we assume these intensities to not appear in covariant patterns because of the low-reliability of the below-threshold intensities. This assumption is based on the appearance of the bump in the residuals for the higher ranks. For this reason, we propose to compare the residuals associated with above- and below-threshold intensities and expect that these exhibit different behavior in terms of PCA capturing these intensities over the considered ranks. Section 3-3 continues on this premise.

Threshold-driven approach In the threshold-driven approach we step towards intensity-aware dimensionality reduction in which we construct principal components emphasizing threshold-exceeding intensities. We assume the existence of the threshold is not based on a definite presence in the properties of these datasets, by i.e. incoherence or non-structure. Instead, we assume that the threshold is an external parameter and explicitly classify below-threshold intensities as irrelevant. Therefore, we attenuate the influence of the below-threshold intensities in the dimensionality reduction and expect a reduction of the rank estimate when only describing the majority of the above-threshold intensities. Chapter 4 continues on this reasoning.

We have chosen to continue with PCA to obtain an initial starting point for intensity-aware rank estimation and leave intensity-aware rank estimation in NMF as a topic for future research. PCA provides several simplifications over NMF. The non-convexity and ill-posedness

of NMF hamper conclusive analysis of the rank estimation results. PCA, on the other hand, allows for obtaining a globally optimal solution. Moreover, contrary to NMF, PCA has an incorporated rank estimation method in the form of explained variance.

3-3 Residual-fraction rank estimation

In this section, we investigate a comparison of the residuals associated with respectively the above- and below-threshold intensities as a method for rank estimation. This comparison is based on the premise that the ion intensity threshold is present in the form of an incoherent or non-structured region in these datasets. We expect the below-threshold intensities to appear in noncovariant patterns because of the low-reliability of these intensities. We demonstrate that in the case of PCA an optimum can be found in the capturing of above-threshold intensities relative to the below-threshold intensities. We do note this is a heuristic method and not all (IMS) datasets exhibit similar behavior, but it shows merit for the considered IMS dataset.

To recap, our reasoning behind residual-fraction rank estimation is based on figure 3-1 and the premise on the ion intensity threshold associated with residuals in the above- and below-threshold groups exhibiting different behavior with increasing rank, as discussed in section 3-2. The above-threshold intensities should contain physical signal plus some additional random variation. The below-threshold intensities would be less reliable, and we expect them to exhibit more random behavior. This increased randomness of the below-threshold intensities causes these low-intensities to be mostly incoherent and non-structured. As such, we expect the contribution to the covariance between mass-bins of these low-intensities is small. Consequently, as we observe in section 3-2, PCA requires a large number of components to reduce the residuals for these intensities. For this reason, we suspect the residuals for the below-threshold intensities to decrease more slowly for increasing rank in the case of PCA as a result of this randomness, when compared to the above-threshold intensities representing more biological signal.

We expect that at sufficient rank the majority of the intensities above the threshold, representing a biological signal, are captured. As such, consecutive components capture an increasing part of additional random variation or noise, which is also incoherent and random. At this rank, we expect a shift in the share of residuals associated with above-threshold and below-threshold intensities at the rank for which we expect LDR to start capturing noise.

We propose the residual fraction of the below-threshold Residual Sum of Squares (RSSQ) and above-threshold RSSQ, equation (3-11), as a measure on how PCA captures features with respect to the threshold τ for different choices of rank K. The residuals in below- and above-threshold τ groups are given by:

$$\begin{aligned} \mathbf{E}_{<\tau,K} &= \mathbf{D} - \hat{\mathbf{D}}_K \qquad \text{for} \qquad d_{nm} < \tau \\ \mathbf{E}_{\geq\tau,K} &= \mathbf{D} - \hat{\mathbf{D}}_K \qquad \text{for} \qquad d_{nm} \geq \tau \end{aligned} \tag{3-10}$$

The residual fraction $\rho(K)$ describes the ratio between the below-threshold RSSQ and above-threshold RSSQ.

$$\rho(K) = \frac{e^{\text{rssq}}_{<\tau,K}}{e^{\text{rssq}}_{\geq\tau,K}} = \sqrt{\frac{\sum_n^N \sum_m^M e^2_{<\tau,K,nm}}{\sum_n^N \sum_m^M e^2_{\geq\tau,K,nm}}} \tag{3-11}$$

in which $e_{x,K}$ is an element of the residual matrices $\mathbf{E}_{\geq\tau,K}$ and $\mathbf{E}_{<\tau,k}$ respectively satisfying the condition $d_{nm} \geq \tau$ and $d_{nm} < \tau$ and τ denotes the threshold.

In a similar manner to section 3-2, dependent on the threshold τ the number of below-threshold intensities is significantly higher than the number of above-threshold intensities in the residual fraction. Furthermore, the intensities itself influences the residual, as high and low intensities result in respectively high and low residuals for insufficient rank. However in this case, these differences are not a problem, since in this comparison we are solely interested in the development of the residuals over different ranks. Over different ranks the number and the intensity of the below- and above-threshold groups remain constant.

Nonetheless, to construct trends in the residuals in the same order of magnitude for visualization purposes, we scale the individual values of residual ratios with the constant $\gamma = \sqrt{\frac{N_{\geq\tau}}{N_{<\tau}}}$ in which $N_{\geq\tau}$ the number of intensities d_{ij} larger than τ and $N_{<\tau}$ the number of intensities d_{nm} smaller than τ. This scaling is independent of the rank K and as such does not influence the rank estimate and is, as such, only used to scale the residual $\rho(K)$ to the same order of magnitude.

$$\rho(K) = \frac{e^{\text{rssq}}_{<\tau,K}}{e^{\text{rssq}}_{\geq\tau,K}}\sqrt{\frac{N_{\geq\tau}}{N_{<\tau}}} = \sqrt{\frac{N_{\geq\tau}\sum_n^N\sum_m^M(e_{<\tau,K,nm})^2}{N_{<\tau}\sum_n^N\sum_m^M(e_{\geq\tau,K,nm})^2}} = \frac{e^{\text{rms}}_{<\tau,K}}{e^{\text{rms}}_{\geq\tau,K}} \tag{3-12}$$

This residual fraction describes how the distribution of residuals of the respective below- and above-threshold intensities change with the rank. An increase in the fraction indicates either lower residuals and components capturing more of the above-threshold intensities or larger residuals and worse capturing of the below-threshold intensities. Similarly, a decrease of the fraction marks either components capturing more and associated lower residuals of the below-threshold intensities or larger residuals and worse capturing of the above-threshold intensities.

Consequently the optimum, meaning minimal residuals above the threshold and maximal residuals below, is either caused by an increase or decrease of the residuals of respectively the below-threshold or above-threshold intensities. An increase in the ratio, due to an increase of the residuals of the below-threshold intensities, is unwanted. Nonetheless, we expect this not to be a problem. The increase of RSSQ in one of the intensity windows is compensated by an extra decrease in the other, as the squared sum of the residuals has to be minimal and monotonically decreasing with rank by definition of PCA [40]. Consequently, the decrease of either the residuals of either the above- or below-threshold intensities to be always more significant than the increase of the counterpart. Nonetheless, the individual sums of squared residuals for the above- and below-threshold intensity windows are not necessarily minimal and therefore not strictly monotonically decreasing.

The results of residual-fraction rank estimation for PCA based on the residual fraction applied on a synthetic and an IMS dataset are discussed in chapter 5.

Chapter 4

Intensity-aware Dimensionality Reduction

Intensity-aware dimensionality reduction approaches capturing the majority of the intensity-threshold-exceeding information in the original Imaging Mass Spectrometry (IMS) dataset by making the dimensionality reduction algorithm itself intensity-aware. We focus with intensity-aware dimensionality reduction on improved capturing of the threshold-exceeding intensities and argue that this approach could lead to a reduction in rank in the process. We propose Threshold-Aware Principal Component Analysis (TAPCA), an extension of Peak Intensity Weighted Principal Component Analysis (PIWPCA), as a method to make dimensionality reduction intensity-aware. Furthermore, we propose several transformation schemes for TAPCA, for which the effects are evaluated in chapter 5. Additionally, we discuss the problems we encountered with alternative methods to make dimensionality reduction intensity-aware and possibilities for further research.

4-1 Peak Intensity Weighted Principal Component Analysis (PIWPCA)

PIWPCA is an unsupervised decomposition technique for an IMS-measured organic tissue section with a focus on unraveling the underlying biochemical trends [2]. PIWPCA transforms the ion counts in the measurement space with a histogram transformation function $T(x)$ to manipulate the implicit weights of the covariance matrix, as discussed in section 3-1-1. The element-wise application of the transformation $T(x)$ results in the following definition of the covariance matrix:

$$\mathbf{S}_T = (\mathbf{D}_T - \overline{\mathbf{D}_T})^T(\mathbf{D}_T - \overline{\mathbf{D}_T}) \quad \text{with} \quad \mathbf{D}_T = T(\mathbf{D}) \tag{4-1}$$

As a result of this manipulated covariance matrix based on the application of the transformation function $T(x)$ on the original dataset, the basis $\mathbf{C}_T$ can be adjusted towards the objective defined in the transformation function. An example is a gray-scale stretching or contrast enhancing transformation in which they stretch out a particular range of intensities [2].

4-2 Threshold-Aware Principal Component Analysis (TAPCA)

We propose TAPCA as a method to make dimensionality reduction by Principal Component Analysis (PCA) aware of the intensity threshold and associated reliability of the intensities in the spectrum. TAPCA is an extension of PIWPCA [2] and explicitly manipulates the covariance weights in line with the ion intensity threshold. With this manipulation, we specifically aim to reduce the importance of the sub-threshold measurement intensities relative to the threshold-exceeding intensities in the decomposition

With the lowered relative importance of the sub-threshold intensities, we hope to reduce the number of required components to describe the majority of the threshold-exceeding intensities dependent on the threshold. With this threshold-dependent transformation, we relate the rank as defined in the abstract latent space, as constructed by PCA, to the threshold defined in the original measurement space. This relation is further discussed in section 4-2-2.

For TAPCA, we use a specific threshold-dependent ion count transformation $d'_{nm} = T_\tau(d_{nm})$ to lower the relative importance of the sub-threshold intensities.

To achieve the envisioned adjustment of the covariance weights, we propose a transformation decreasing the variance contribution associated with the below-threshold part of the intensities. Variance is defined as the squared distance from the mean. Consequently, this contribution is reduced by lowering the distance to the mean of the below-threshold intensities while leaving the above threshold intensities intact. An example of such a transformation is the downwards shift of all intensities, while setting all sub-zero intensities as a result of the downwards shift to zero. This transformation is demonstrated figure 4-1. More transformations are discussed in section 4-2-4.

This reduction of the below-threshold variance contribution is motivated by the multiplication of standard deviations of the individual mass-bins n and m, $\sigma_n\sigma_m$ as default weights for covariance entry s_{nm}. For example, lowering the standard deviation, or related variance, results in de-emphasizing a particular mass-bin. For this reason, with this transformation, we focus on the reduction of the deviation from the mass-bin mean of the below-threshold intensities. Consequently, we reduce the covariance contribution associated with below-threshold variance contribution and, as a result, these implicit weights are increasingly based on the threshold-exceeding intensities. With this particular class of ion count transformations, we aim to specifically manipulate the deviations from the mean of the sub-threshold intensities. We leave the relative distances in the rest of the dataset intact to obtain optimal capturing of the threshold-exceeding intensities for a minimal rank.

4-2-1 Clipping Threshold-Aware Principal Component Analysis (CTAPCA)

This section proposes clip-shifting as a simple example of how these transformations can reduce variance contributed by the below-threshold intensities. The idea of this simple example is to demonstrate the influence of this transformation on the rank estimate obtained via this technique. We demonstrate and discuss more complicated choices of transformation functions in section 4-2-4 and show them in figure 4-2. In the case of the clip-shifting transformation, the intensity values in the spectra are first shifted down by the intensity threshold τ, and subsequently we set all elements in the spectra below zero to zero.

This operation can be accomplished by element-wise application of the following transformation function:

$$T_{\text{clip}}(x) = \begin{cases} x - \tau & x \geq \tau \\ 0 & x < \tau \end{cases} \tag{4-2}$$

in which τ denotes the supplied threshold. This operation creates the shifted measurement matrix $\mathbf{D}_{\text{clip}} = T_{\text{clip}}(\mathbf{D})$ from the measurement matrix $\mathbf{D}$. Subsequently, the new covariance matrix $\mathbf{S}_{\text{clip}}$ can be calculated using the mean-centered shifted measurement matrix $\mathbf{D}^*_{\text{clip}} = \mathbf{D}_{\text{clip}} - \overline{\mathbf{D}_{\text{clip}}}$:

$$\mathbf{S}_{\text{clip}} = (\mathbf{D}_{\text{clip}} - \overline{\mathbf{D}_{\text{clip}}})^T(\mathbf{D}_{\text{clip}} - \overline{\mathbf{D}_{\text{clip}}}) \tag{4-3}$$

in which $\overline{\mathbf{D}_{\text{clip}}}$ denotes the column-wise mean of the shifted measurement matrix $\mathbf{D}_{\text{clip}}$. From this new covariance matrix $\mathbf{S}_{\text{clip}}$, the new axes of maximum variance $\mathbf{C}_{\text{clip}}$, or principal components can be obtained. These principal components reflect only the variance originating from intensities in the spectrum above τ. Finally, the projection of the original matrix $\mathbf{D}$ or the shifted matrix $\mathbf{D}_{\text{clip}}$ on the first K axes, $\mathbf{C}_{\text{clip},K}$ constructs a low dimensional representation. Similarly to traditional PCA, a rank-K reconstruction can be constructed by taking the first K axes resulting in $\hat{\mathbf{D}}_{\text{clip},K} = \mathbf{D}^*_{\text{clip}}\mathbf{C}_{\text{clip},K}\mathbf{C}^T_{\text{clip},K} + \overline{\mathbf{D}_{\text{clip}}}$. The considerations and choices for the projection step are explained in detail in section 4-2-3.

4-2-2 Threshold-shifted rank estimation with TAPCA

We propose using the explained variance for rank estimation analogous to PCA. However, in the case of TAPCA, we use the explained variance of transformed intensities as a method for rank estimation. The per-component explained variance is obtained by using the eigenvalues of the covariance matrix $\mathbf{S}_{\text{clip}}$. These M eigenvalues in descending order are normalized so that they represent the cumulative fraction of the per-component contributed variance. Consequently, cumulative fraction of the per-component variance at the K-th eigenvalue is reflected by:

$$\lambda'_K = \frac{\sum_i^K \lambda_i}{\sum_i^M \lambda_i} \tag{4-4}$$

in which M denotes the number of mass-bins and K the component number.

By computation of the normalized eigenvalues per TAPCA covariance matrix $\mathbf{S}_{\text{clip}}$ for a range of clip-shifting thresholds τ, we can construct a relation between threshold and rank. By stacking these sets of normalized eigenvalues ordered by the corresponding threshold, we can demonstrate how the fractions of explained variance evolve for various thresholds and choices of rank. Contour lines for common fractional explained variance truncation can show trends in the relationship between threshold and rank for TAPCA. A visualization of this representation of the per-threshold covariance eigenvalues for an IMS dataset is given in figure 5-15.

In contrast to the covariance matrix $\mathbf{S}$ in traditional PCA, the covariance matrix $\mathbf{S}_{\text{clip}}$ does not capture any variance originating from intensity values in the range $[0, \tau)$. Consequently, we expect the covariance matrix $\mathbf{S}_{\text{clip}}$ to be sparser than the one obtained from traditional PCA. The main reason is, that as a result of the transformation the contribution to the covariance from intensities in the range $[0, \tau)$, is sharply reduced compared to mass-bins consisting to intensities in the range $[\tau, \infty)$.

As a result of the sparser covariance matrix, a reduced number of principal components is required to capture the majority of the variance associated with the threshold-exceeding part of the intensities. As such, TAPCA should require a lower rank to describe the majority of information above τ.

4-2-3 Projection on a low-rank basis

The projection of our measurement matrix on the first K axes $\mathbf{C}_{\text{clip},K}$ of the basis $\mathbf{C}_{\text{clip}}$ constructs a K-dimensional representation.

We have two choices on which measurement matrix to project. First, we could project the transformed centered matrix $\mathbf{D}_{\text{clip}} - \overline{\mathbf{D}_{\text{clip}}}$ on basis $\mathbf{C}_{\text{clip},K}$ in a similar way as [55]. This projection results in a reconstruction error matrix:

$$\mathbf{E}_{\text{clip}} = (\mathbf{D}_{\text{clip}} - \overline{\mathbf{D}_{\text{clip}}}) - (\mathbf{D}_{\text{clip}} - \overline{\mathbf{D}_{\text{clip}}})\mathbf{C}_{\text{clip},K}\mathbf{C}_{\text{clip},K}^T \tag{4-5}$$

By definition of PCA the reconstruction error $\|\mathbf{E}_\tau\|_F^2$ is minimal for a rank-K approximation [40]. This implies this solution is the best reconstruction in terms of lower residuals for all intensity in the spectra above τ. However, the clip-shifting disables capturing anything below τ, even if it is coherent with patterns above τ. Furthermore, the reconstruction $\hat{\mathbf{D}}_{\text{clip},K} = (\mathbf{D}_{\text{clip}} - \overline{\mathbf{D}_{\text{clip}}})\mathbf{C}_{\text{clip},K}\mathbf{C}_{\text{clip},K}^T + \overline{\mathbf{D}_{\text{clip}}}$ reconstructs the shifted dataset and not the original. Consequently, this complicates computing the residuals, as we can no longer compare them to the original measurements.

Alternatively, we could project the centered matrix $\mathbf{D} - \overline{\mathbf{D}}$ unaffected by the transformation on basis $\mathbf{C}_{\text{clip},K}$ resulting in a residual matrix:

$$\mathbf{E} = (\mathbf{D} - \overline{\mathbf{D}}) - (\mathbf{D} - \overline{\mathbf{D}})\mathbf{C}_{\text{clip},K}\mathbf{C}_{\text{clip},K}^T \tag{4-6}$$

This projection has the advantage that intensities below the threshold for a mass-bin are still captured, assuming the mass-bin represents a reasonable amount of variance above the threshold. Furthermore, the rank-K reconstruction $\hat{\mathbf{D}}_K = (\mathbf{D} - \overline{\mathbf{D}})\mathbf{C}_{\text{clip},K}\mathbf{C}_{\text{clip},K}^T + \overline{\mathbf{D}}$ approximates the original measurement matrix $\mathbf{D}$ instead of the transformed variant $\mathbf{D}_{\text{clip}}$, leaving the intensities of the original data intact. The disadvantage is that intensities not represented in the covariance matrix could lead to a larger reconstruction error. Furthermore, the residual $\|\mathbf{E}\|_F^2$ is not necessarily minimal anymore by definition of PCA, because the proof for minimal reconstruction error in [40] is no longer valid.

However, we expected a relatively better reconstruction for intensities above the threshold than their below-threshold counterparts as we obtain a basis reflecting the variance of these intensities. Furthermore, we expected the variance contribution of the below-threshold intensities to be limited as a result of their low-intensity. Consequently, our hypothesis was that projection of the original measurement matrix on the basis obtained from the shifted measurement matrix would result in lower residuals of intensities above the threshold, but a higher total residual caused by the low intensities when compared to traditional PCA. Section 5-1-3 evaluates the residuals per intensity and shows that this is not the case.

For this thesis, we have chosen to evaluate both methods and as such to project both the original data and shifted data on the new τ-aware, low-rank basis $\mathbf{C}_{\text{clip},K}$. This choice is based on that we are mostly interested to see how we can create a low-rank approximation taking into account the intensity threshold.

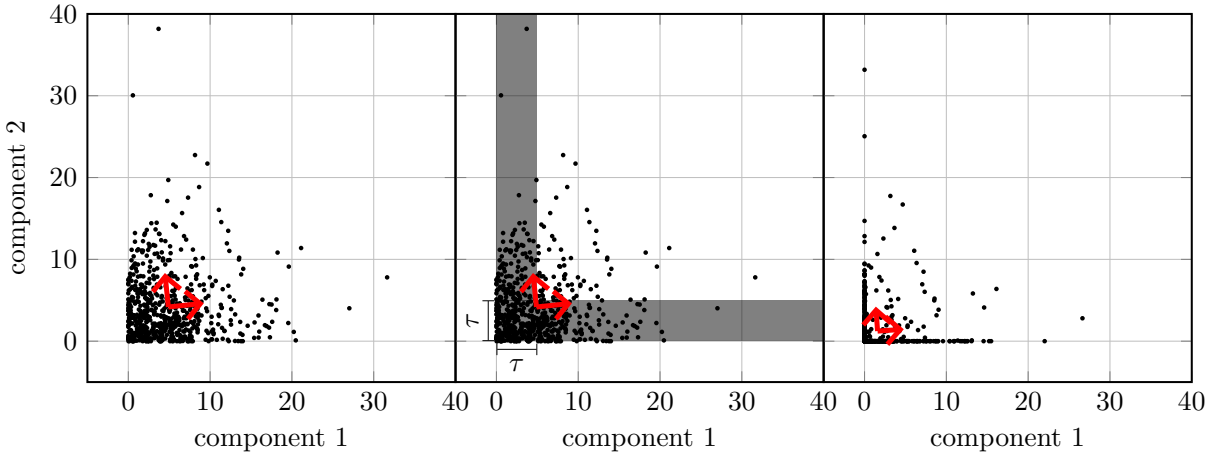


Figure 4-1: A 2D visualization of TAPCA. The left plot displays the original data with corresponding axes of maximum variance. The length of the arrow corresponds to the accounted variance for this particular axes. The middle plot demonstrates what measurements are considered unreliable by intensity threshold τ. The right plot displays the shifted dataset and corresponding axes of maximum variance. These axes are slightly rotated compared to the original axes in the left plot. Furthermore the lengths of the arrows are smaller, meaning both axes account for less reliable variance compared to the original variance.

4-2-4 Alternative transformations for TAPCA

The hard cut-off for CTAPCA removes all fluctuations in the intensities in the range $[0, \tau)$. Consequently, TAPCA equalizes the variance contributed by individual intensities under the threshold. A pixel even slightly above the intensity threshold, on the other hand, still reflect these small fluctuations and relative differences. As a result, stark differences in contribution to covariance can occur between just below and just above the threshold. Consequently, the hard threshold in the clip-shifting potentially creates significant differences in capturing intensities just below and just above the threshold.

To avoid the effects of the clip-shift transformation, we propose a set of alternative transformation functions. The transformation focuses on especially de-emphasizing the sub-threshold intensities in a more continuous manner, yet aim to keep the relative differences in threshold-exceeding intensities intact similar to the clip-shift transformation. The more continuous weighting of below-threshold intensities enables us to de-emphasize the difference between two intensities in the sub-threshold region in a more continuous manner.

Examples of such histogram shift transformations are the piecewise-linear, equation (4-7), and a quadratic transformation function, equation (4-8), but other functions custom to the objective are possible and could be a topic for future research.

The linear function, equation (4-7), maps the domain $[0, \tau + d]$ in a linear manner to $[0, d]$. Equation (4-8) maps the domain $[0, 2\tau]$ with a quadratic function to $[0, \tau]$. The quadratic transformation reduces the difference between two values, $|T(a) - T(b)| < |a - b|$ in the dataset in the domain $[0, 2\tau]$ nonlinearly, effectively placing two intensity values within this range closer together, while leaving the difference in the domain $[\tau, \infty)$ intact.

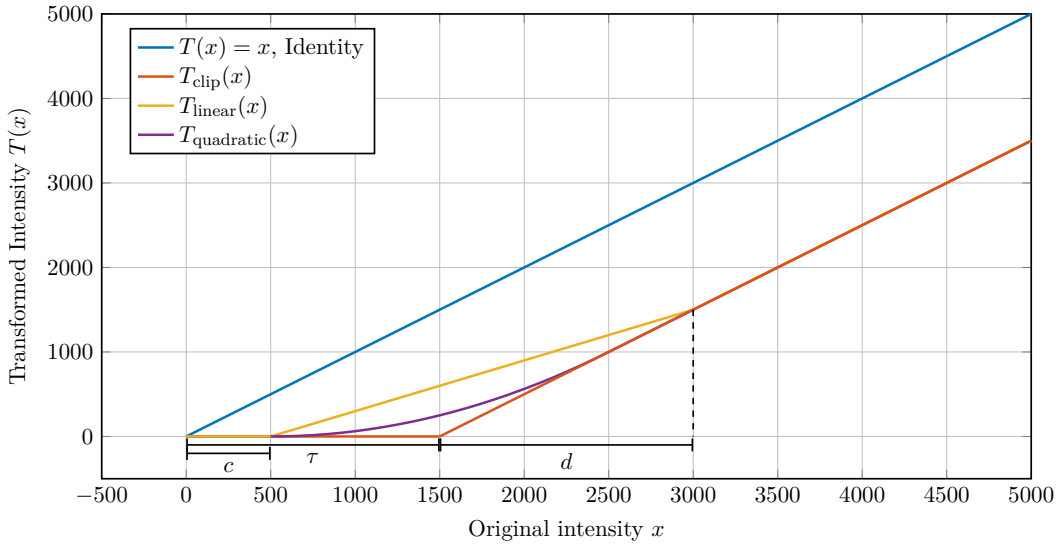


Figure 4-2: A visualization of linear-shift $T_{\text{linear}}(x)$ and quadratic shift $T_{\text{quadratic}}(x)$ histogram transformation functions for two choices of c compared to clip-shifting function $T_{\text{clip}}(x)$ and the identity histogram transformation for $\tau = 1500, c = 500, d = \tau$. The original values are displayed on the horizontal axis and the output after transformation on the vertical axis. In the linear case, the low-intensity values in the range $[c, \tau + d]$ are mapped by a linear function to the domain $[c, \frac{\tau+d}{2}]$. In the quadratic case, the low-intensity values in the range $[c, 2\tau - c]$ are mapped by a quadratic function to the domain $[c, \tau - c]$. In both cases, the values in the domain $[0, c]$ are set to zero. This mapping reduces the differences $|T(a) - T(b)|$ for values closer to zero more than for values close to the threshold with equal difference $|a - b|$ similar to the clip-shift $T_{\text{clip}}(x)$.

$$T_{\text{linear}}(x) = \begin{cases} \frac{d}{\tau+d-c} & c \leq x \leq \tau + d \\ x - \tau & x > \tau + d \end{cases} \tag{4-7}$$

$$T_{quadratic}(x) = \begin{cases} \frac{l(x-c)^2}{2\tau} & l\frac{x-c}{\tau} \leq 1 \\ x - c - \frac{\tau}{2l} & l\frac{x-c}{\tau} > 1 \end{cases} \quad \text{with} \quad l = \frac{\tau}{2(\tau - c)} \tag{4-8}$$

in which c defines the domain $[0, c]$ for which the relative distance: $|T(a) - T(b)| = 0$ with $a, b \in [0, c]$ and the variance contribution is completely discarded. Figure 4-2 shows the differences between these linear-shifting and clip-shifting function (equation (4-2)).

In section 5-1, we evaluate the influence of the different transformations on respectively the reconstruction and capturing of patterns and rank estimation for an equal choice of the shift point, in these formulas denote as τ. In section 5-2 and section 5-3, we evaluate the influence of the clip-shift transformation and linear-shift transformation on the rank estimation. Through the rest of this thesis, we use CTAPCA to denote the version of TAPCA utilizing the clip-shift transformation, equation (4-2). We use Linear Threshold-Aware Principal Component Analysis (LTAPCA) and Quadratic Threshold-Aware Principal Component Analysis (QTAPCA) to denote the version of TAPCA transformation utilizing respectively the linear histogram shift transformation, equation (4-7) and the quadratic histogram shift transformation equation (4-8).

4-3 Other Methods

We have also evaluated incorporating the threshold into dimensionality reduction based on other methods than TAPCA, but these proved either unsuccessful or require more research beyond the scope of this thesis.

4-3-1 Weighted Covariance Principal Component Analysis (WCPCA)

WCPCA is a specific variant of PCA, which allows emphasizing entries in a dataset over others by the application of weights [55]. These weights can be applied column- or row-wise or on individual entries. WCPCA differs from the PIWPCA and the approach taken in this thesis, as the weights are applied in the covariance matrix instead of a preprocessing step before constructing the covariance matrix as is the case of PIWPCA [2].

In the case of intensity-aware dimensionality reduction, we aim to respectively de-emphasize below-threshold and emphasize above-threshold intensities. As such, we can obtain a low-rank representation with maximal capturing of the above threshold intensities. We have tried several weighting schemes based on WCPCA with a focus on down-weighting below-threshold intensities.

WCPCA constructs the covariance matrix by the application of a weight to the individual intensities by element-wise application of a weighting matrix:

$$\mathbf{S}_\tau = \frac{(\mathbf{W} \odot (\mathbf{D} - \overline{\mathbf{D}_\mathbf{W}}))^T (\mathbf{W} \odot (\mathbf{D} - \overline{\mathbf{D}_\mathbf{W}})}{\mathbf{W}^T\mathbf{W}} \qquad \text{with} \qquad \overline{\mathbf{d}_{\mathbf{W},m}} = \frac{\sum_n^N w_{nm} d_{nm}}{\sum_n^N w_{nm}} \tag{4-9}$$

We conclude that threshold-driven rank estimation based on WCPCA is not straightforward and further research is needed. As an example, we demonstrate a binary scheme, for which the results are shown in appendix A. We think this scheme is close to the worst case scenario, but believe similar problems hold for other weighting schemes as well. In this scheme, we set all weights of respectively the below-threshold and above-threshold intensities to zero and one.

$$w_{nm} = \begin{cases} 1 & d_{nm} \geq \tau \\ 0 & d_{nm} < \tau \end{cases} \tag{4-10}$$

As a result, in WCPCA the covariance matrix reflects none of the below-threshold intensities and only the above-threshold intensities. The application of this weighting function has a few side effects on the covariance. The weighting itself significantly changes the mean $\overline{\mathbf{d}_{\mathbf{W},m}}$ of mass-bin m. The mean of a mass-bin consisting of a small set of high intensities and the majority below-threshold intensities become close to these high intensities as a result of the weighting. As a result, we observed several problems in the binary weighting scheme. First, patterns considered covariant by traditional PCA are not necessarily covariant in the weighted version as a result of the changed mean due to weighting matrix $\mathbf{W}$. Second, the covariance contribution of specific high intensity patterns plummeted due to the shifted mean. As a result, the weighted version fails to capture these patterns. We have considered other weighting functions, such as linear and sigmoid, but selecting a weighting function and

associated parameters to capture all high intensities proved to be not straightforward. We recommend that searching for a weighting scheme that achieves sufficient emphasis yet does not have these side effects as a suitable topic for future research.

Concerning the TAPCA, we are aware that the shifting transformation in TAPCA also changes the mass-bin means. However, we observed that the effect of this downwards movement is significantly smaller than the changes in the mean due to the weighting scheme in WCPCA.

4-3-2 Threshold PIWPCA

As an alternative to the shifting functions proposed in section 4-2-4, we intended to use a threshold function as proposed for PIWPCA [2] for intensity-aware rank estimation. The threshold function sets all below-threshold intensities to zero, while keeping the threshold-exceeding intensities intact according to the following function:

$$T_{\text{threshold}}(x) = \begin{cases} x & x \geq \tau \\ 0 & x < \tau \end{cases} \tag{4-11}$$

However, in the context of intensity-aware dimensionality reduction, this does not achieve the desired effect. Intensity-aware dimensionality reduction aims to reduce the influence of below-threshold intensities, which in the case of PCA means a reduction in the variance contribution of the below-threshold intensities. However, in some cases, the application of threshold PCA can have an adverse effect. For example, the variance contribution of a mass-bin consisting of intensities partly above, partly below the threshold is amplified. In this case, the thresholding operation places the intensities further away from the mean, resulting in a boosted mass-bin variance and covariance contribution. Due to the variable weighting in PCA, as discussed in section 3-1-1, the boosted covariance-contribution causes unwanted increased significance for mass-bins consisting of partly threshold-exceeding intensities and partly below-threshold intensities.

4-3-3 Weighted Nonnegative Matrix Factorization (WNMF)

WNMF is an extension of Nonnegative Matrix Factorization (NMF) analogous to WCPCA versus PCA. In the context IMS, it allows putting additional emphasis on specific intensities in the mass spectra $\mathbf{D}$ [56]. The memberships $\mathbf{W}$ and latent subspectra $\mathbf{H}$ can be constructed by solving the following optimization problem:

$$\begin{aligned} \underset{\mathbf{W},\mathbf{H}}{\text{minimize}} \quad & F(\mathbf{S} \odot d\left(\mathbf{V}, \mathbf{WH}\right)) \\ \text{subject to} \quad & \mathbf{W} > 0 \\ & \mathbf{H} > 0 \end{aligned} \tag{4-12}$$

where $F(\mathbf{X})$ denotes a reduction operator i.e, the Frobenius norm, $d(\mathbf{V}, \mathbf{WH})$ denotes the element-wise divergence function, $\odot$ denotes the Hadamard element-wise product, $\mathbf{S} \in \mathbb{R}^{N \times M}$ is the weight matrix, $\mathbf{D} \in \mathbb{R}^{N \times M}$ is the IMS mass spectra in matrix form.

We envisioned a rank estimate by emphasizing threshold-exceeding intensities or de-emphasizing below-threshold intensities by choosing a weighting function in a similar sense to WCPCA.

However, we encountered similar issues as for WCPCA in the context of IMS. These issues were that the binary weighting scheme had undesirable side effects and the selection of an intermediate scheme was non-trivial. In the case of WNMF, the binary weighting scheme had even more unwanted effects than for WCPCA. As a result of this weighting, low intensities could be significantly overestimated as their residuals were not taken into account due to weighting $\mathbf{S}$ in equation (4-12). For similar reasons as WCPCA, selecting a less rigid weighting function resulting in insufficient emphasis on all threshold-exceeding intensities was not straightforward.

Next to the selection of a weighting scheme, we require a rank estimation methodology for NMF. In the literature, only a limited number of rank estimation methods is available [27, 37, 57], as listed in section 1-2 and one is not readily available as is the case with explained variance rank estimation for WCPCA and TAPCA. Combining one of the available rank estimation methods with WNMF in a weighting scheme that emphasizes threshold-exceeding intensities to obtain an intensity-aware rank estimation method proved non-trivial and more research in this area is needed.

Chapter 5

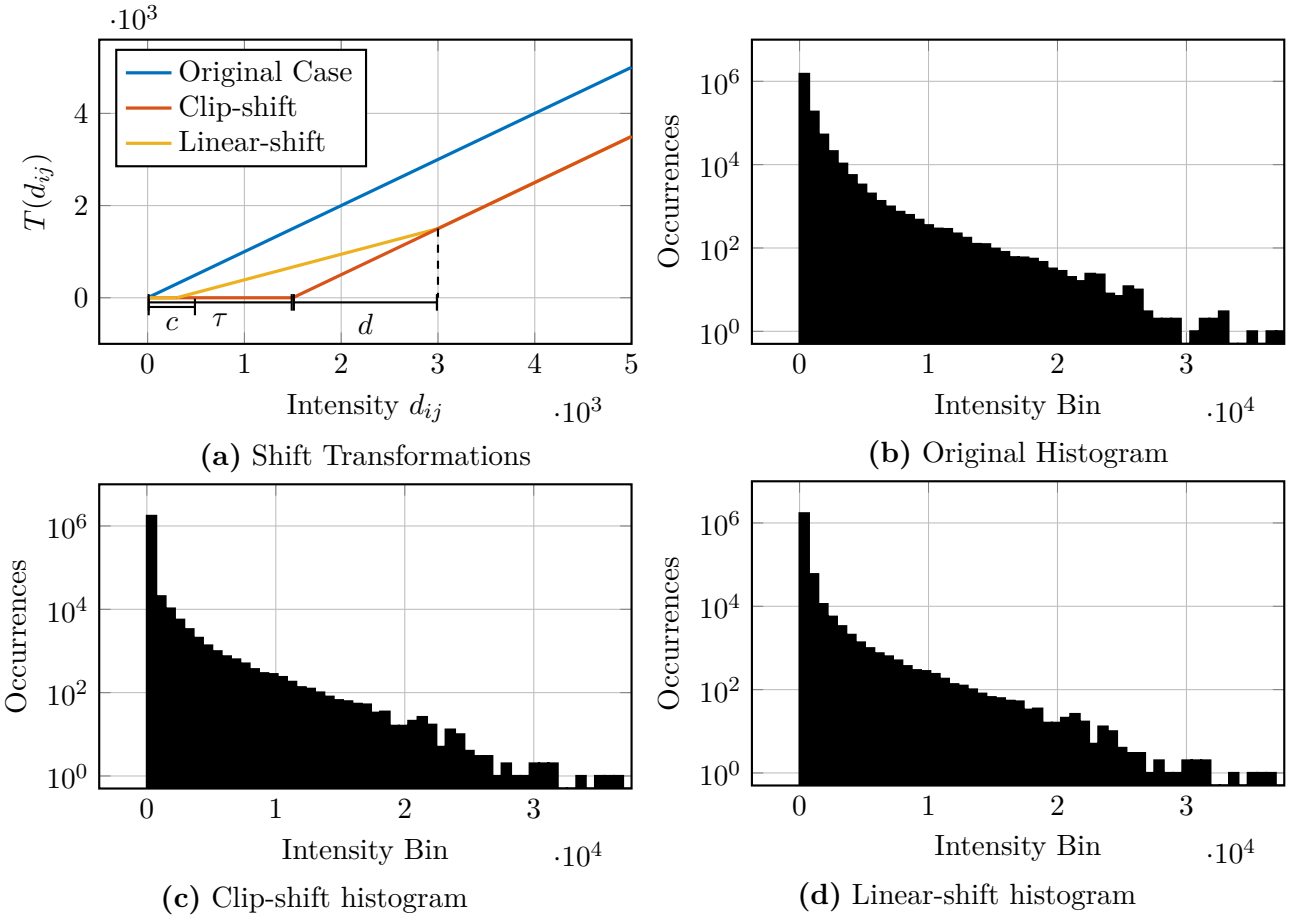
Evaluation of Intensity-Aware Rank Estimation Methods

This chapter starts with an evaluation of Threshold-Aware Principal Component Analysis (TAPCA) method by comparing these to traditional Principal Component Analysis (PCA). We compare TAPCA and PCA both qualitative, by comparing the Principal Components (PCs), the identified relevant spatial patterns and the Root Mean Squared (RMS) intensities of the mass-bins, as well as quantitative, by a comparison of the binned residuals. Furthermore, we demonstrate the attenuation of the influence of particular mass-bins on the PCs, in function of a a changing intensity threshold and we evaluate the choice of the projection of deviations with or without transformation, as proposed in chapter 4. Subsequently, we demonstrate the distribution of these mass-bins of interest to show the effect of these transformations. Afterwards, we evaluate both intensity-aware rank estimation methods with the help of a synthetic dataset with known rank and threshold, followed by an application to a real Imaging Mass Spectrometry (IMS) dataset. In the case of the IMS dataset, we make a comparison to Cross-Validation and percentage-of-explained-variance rank estimation methods for respectively residual-fraction rank estimation and threshold-shifted rank estimation.

5-1 Comparison of LDR with PCA and TAPCA

In this section, we demonstrate and analyze the differences between TAPCA and PCA when applied to the IMS dataset, as introduced section 2-2-1. To simplify this analysis and clarify the visualization, we select only the first 100 mass-bins in some cases (new $M = 100$) of this dataset with ion masses in the range of $[1500, 2800]$. Figure 5-1 shows the applied clip T_{clip} and linear T_{linear} shift transformations for threshold $\tau = 1500$, $c = \frac{\tau}{5}$ and $d = \tau$ and the effect on the histogram of this dataset.

Empirically, we find that the effect of the linear or quadratic functions is limited compared to different choices of the shift parameter τ in equation (4-2) and equation (4-7) for these shifting transformations. For this reason, we fix τ, which facilitates a clear understanding of the effects



(a) Shift Transformations

(b) Original Histogram

(c) Clip-shift histogram

(d) Linear-shift histogram

Figure 5-1: Visualization of **(a)** the linear and clip-shift histogram transformations for $\tau = 1500$, $c = \frac{\tau}{5}$ and $d = \tau$ compared to the untransformed case. The histograms for the original **(b)**, clip-shifted **(c)**, and linear-shifted **(d)** cases show respectively the different effects of the histogram transformation.

of Linear Threshold-Aware Principal Component Analysis (LTAPCA) and its differences with Clipping Threshold-Aware Principal Component Analysis (CTAPCA). Furthermore, a slight increase in the parameter c, defining the clipped region, showed a significant reduction in the differences hampering this comparison. However, we did want to include a clipped region in the linear-shift transformation and settled for $c = \frac{\tau}{5}$. As a consequence and due to the limited influence of the type of transformation, we chose a linear-shift transformation affecting the increased intensity region $[0, 2\tau)$ for LTAPCA to obtain sufficiently clear differences.

The histogram associated with the clip-shift transformation in figure 5-1c is shifted left compared to the original in figure 5-1b. The histogram in figure 5-1c shows a steep strong peak in the low-intensity region $[0, \leq 1.5e3)$, which is the result of the linear-shift mapping the intensities $[0, 2\tau)$ into $[0, \tau)$.

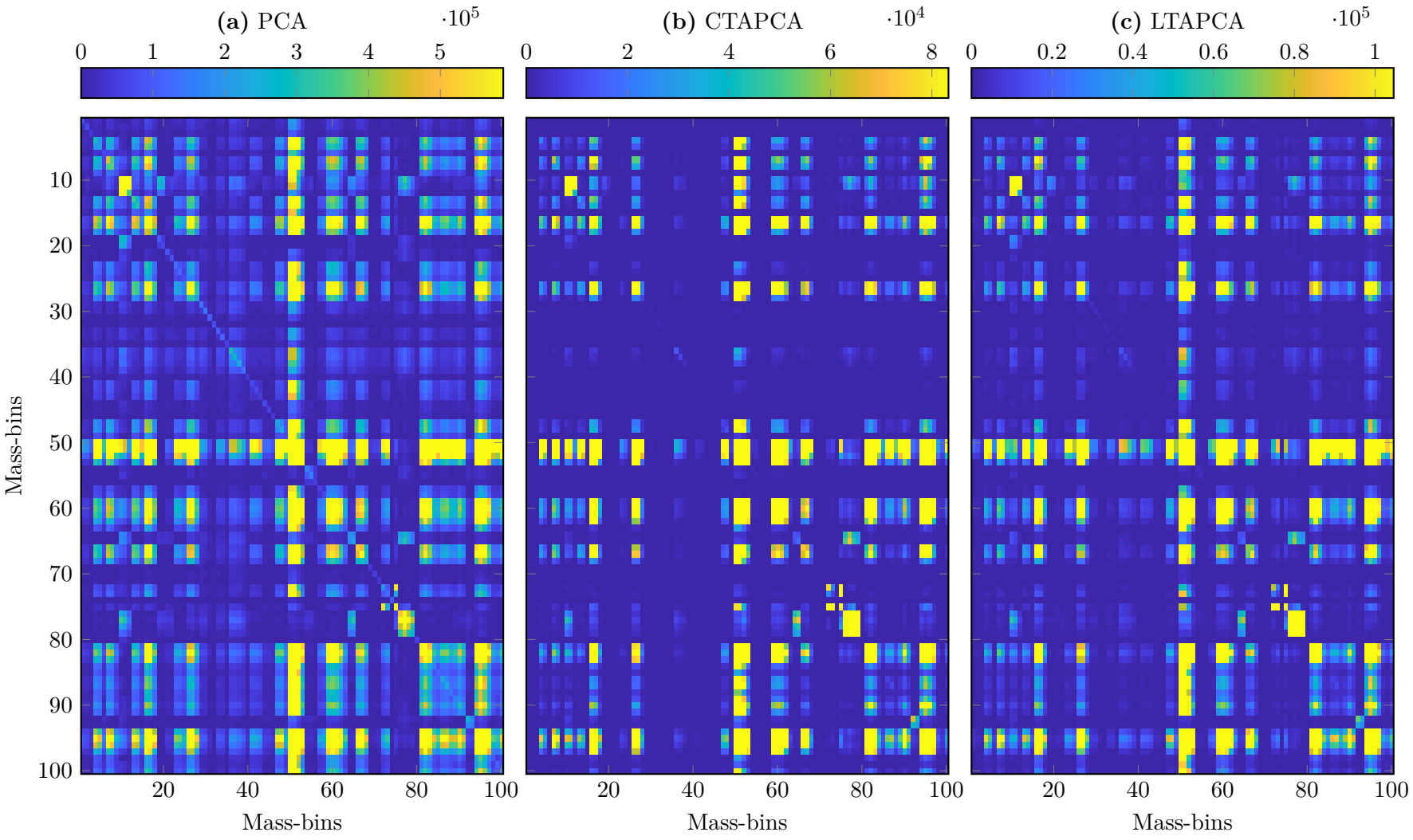


Figure 5-2: The different absolute covariance $|s_{pq}|$, equation (3-1) in the original case for PCA, after the application of the clip-shift for CTAPCA, and linear-shift transformation for LTAPCA for threshold $\tau = 1500$, $c = \frac{\tau}{5}$ and $d = \tau$ on the first 100 mass-bins of the Coronal Rat Brain dataset. The color scale in this plot aligns with the $[5\%, 95\%]$ percentiles in the absolute covariance to highlight the differences in the intermediate intensity region as a result of the transformation.

5-1-1 Covariance matrix and principal components

In this section, we evaluate the influence of the imposed transformation on the covariance matrix and the PCs obtained by application of both PCA, CTAPCA, and LTAPCA. We are interested in the effects of the imposed shift transformation on the spatial distribution in the scores s_i and the density of the characteristic subspectra or loadings c_i for the various components. The first four PCs are visualized in figure 5-3 and figure 5-4 and constructed with the aforementioned shift transformations based upon the covariance matrices in figure 5-2. We believe four components give a sufficiently clear understanding of the differences between the non-shifted and shifted case. As introduced in section 4-2, we intend to discard the unreliable peaks below the threshold while capturing the intensities above and with these figures show how this affects the scores and loadings. To evaluate if the shift histogram transformation has the aforementioned effect, we analyze in the following paragraphs the differences between traditional PCA with CTAPCA, the clip-shift transformation, and the CTAPCA with LTAPCA, a linear-shifting transformation.

Clipping Threshold-Aware Principal Component Analysis (CTAPCA)

Principal Component (PC) one in figure 5-3 shows that after transformation the loadings (bottom) of the first PC are more centered around a couple of dominant peaks, while the loadings without prior transformation contain more small peaks. In the associated scores (top), we observe less high-intensity (red) pixels, which we attribute to the downward shift of all intensities. We also observe a reduction of clutter in the low and negative values in blue surrounding the high-intensity patterns, and as a result clearer spatial delineation. Furthermore, we see in figure 5-2 that the modified covariance matrices are significantly sparser than the original covariance matrix. We suspect the reduced clutter and the sparser spectra are related to the reduction in the covariance of the below-threshold intensities due to the imposed transformation.

To explain this, in equation (3-1) we zoom on the covariance contribution for pixel j between mass-bins p and q:

$$s_{pqj} = (d_{jp} - \overline{d_p})(d_{jq} - \overline{d_q}) \tag{5-1}$$

In this equation, the absolute contribution of mass-bin p to the covariance s_{pq} of an arbitrary pixel j originates from the deviation from the mean $d_{jp} - \overline{d_p}$. The imposed shifting transformation reduces the deviation from the mean $d_{jp} - \overline{d_p}$ of pixel j for the below-threshold-intensities $d_{jp} \in [0, \tau)$ by placing d_{jp} closer to the mean $\overline{d_p}$. The magnitude of this reduction is dependent on the mean $\overline{d_p}$ of the mass-bin originating from the other pixels in the mass-bin. Consequently, if the previously covariant pixels j in either mass-bin p or q are below the threshold, the contribution of these pixels to the covariance s_{pq} is reduced. For this reason, we suspect that as a result of the shift this feature is no longer covariant with some of the below-threshold intensity peaks. Besides the increased sparsity, another aspect to note is the difference in variance of the individual mass-bins reflected by the entries on the diagonal of the covariance matrix in figure 5-2. The original covariance matrix shows a diagonal line, signifying relatively high variances for mass-bins even when generally non-covariant with other mass-bins. After transformation, this line disappears which suggest that this variance mostly originated from below-threshold intensities.

Figure 5-3 and figure 5-4 show the permutation in order of PCs two and three. The order in the PCs originates from the reflected covariance by each particular PC. Consequently, the imposed clip-shift reduced the covariance of the original PC two more than that of the original PC three. We observe that the covariance reflected by PC two originates from a significant number of small intensity peaks, whereas the covariance of PC three originates from a limited set of high-intensity peaks. As such, we suspect the permutation is the result of the lowered contribution to the covariance matrix of the low-intensity peaks. The lowered contribution affects PC two more, due to the covariance associated with a significant number of small intensity peaks.

The scores associated with PCs two, three, and four in figure 5-3 and figure 5-4 demonstrate that CTAPCA acquires a cleaner delineation of particular high-intensity patterns, whereas traditional PCA acquires a more blurry score. PCA seems to classify these patterns as covariant as a result of low intensities. We believe that the blurry scores are the result of PCA considering the low-intensity peaks covariant together with the high-intensity patterns. In line with the aforementioned reasoning, the imposed shift results in a lowered contribution to the covariance of the below-threshold intensities, and as a result, the below-threshold intensities are no longer considered covariant.

Linear Threshold-Aware Principal Component Analysis (LTAPCA)

For PC one, the LTAPCA decomposition in figure 5-3 and figure 5-4 shows a less clear delineation and a more blurry background compared to CTAPCA. Furthermore, the characteristic subspectra are denser than the CTAPCA and more similar to the spectra in traditional PCA. Compared to the CTAPCA, PC three and four in the LTAPCA case show more overlap in the form of the red dot in the right bottom corner most likely associated with m/z around 2500. We believe both are the result of the linear transformation not completely discarding the variance contribution of the below-threshold intensities, but only reducing it. Consequently, below-threshold peaks are still considered covariant with high-intensity features. We see traditional PCA considers the mass bins m/z 1750 and m/z 2500 covariant, while in the CTAPCA case this relation is significantly reduced. The LTAPCA version ends up somewhere in the middle between CTAPCA and LTAPCA since part of the below-threshold intensities are still taken into account.

5-1-2 Captured spatial patterns

This section qualitatively assesses the differences between TAPCA and traditional PCA in the captured patterns after dimensionality reduction by a comparison to the original patterns and the patterns captured by traditional PCA. We assess this again using the truncated dataset with the same 100 mass-bins in the mass range $[1500, 2800]$. We use the rank-K reconstruction $\hat{\mathbf{D}}_k$ from equation (3-5) to evaluate if the various dimensionality reduction methods capture the patterns. Empirically, we have found that a rank of 11 provides sufficient PCs to capture most of the high-intensity patterns after application of traditional PCA to the truncated dataset. We use a similar rank for the reconstruction of TAPCA, so a clear view on the differences can be obtained.

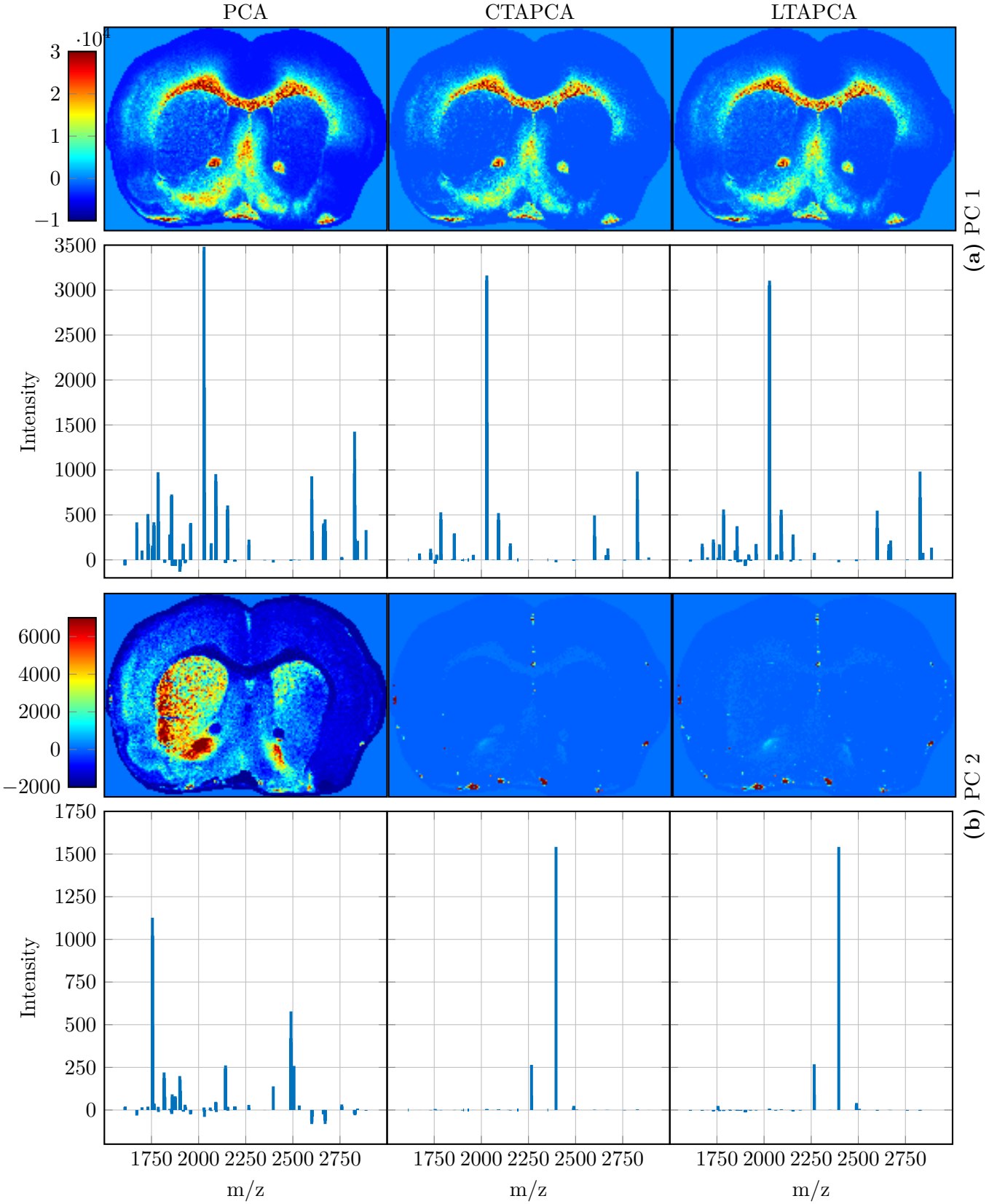


Figure 5-3: The scores (top) and loadings (bottom) of the first **(a)** and second **(b)** principal components constructed with PCA, CTAPCA, and LTAPCA for threshold $\tau = 1500$, $c = \frac{\tau}{5}$ and $d = \tau$ on the first 100 mass-bins of the Coronal Rat Brain dataset.

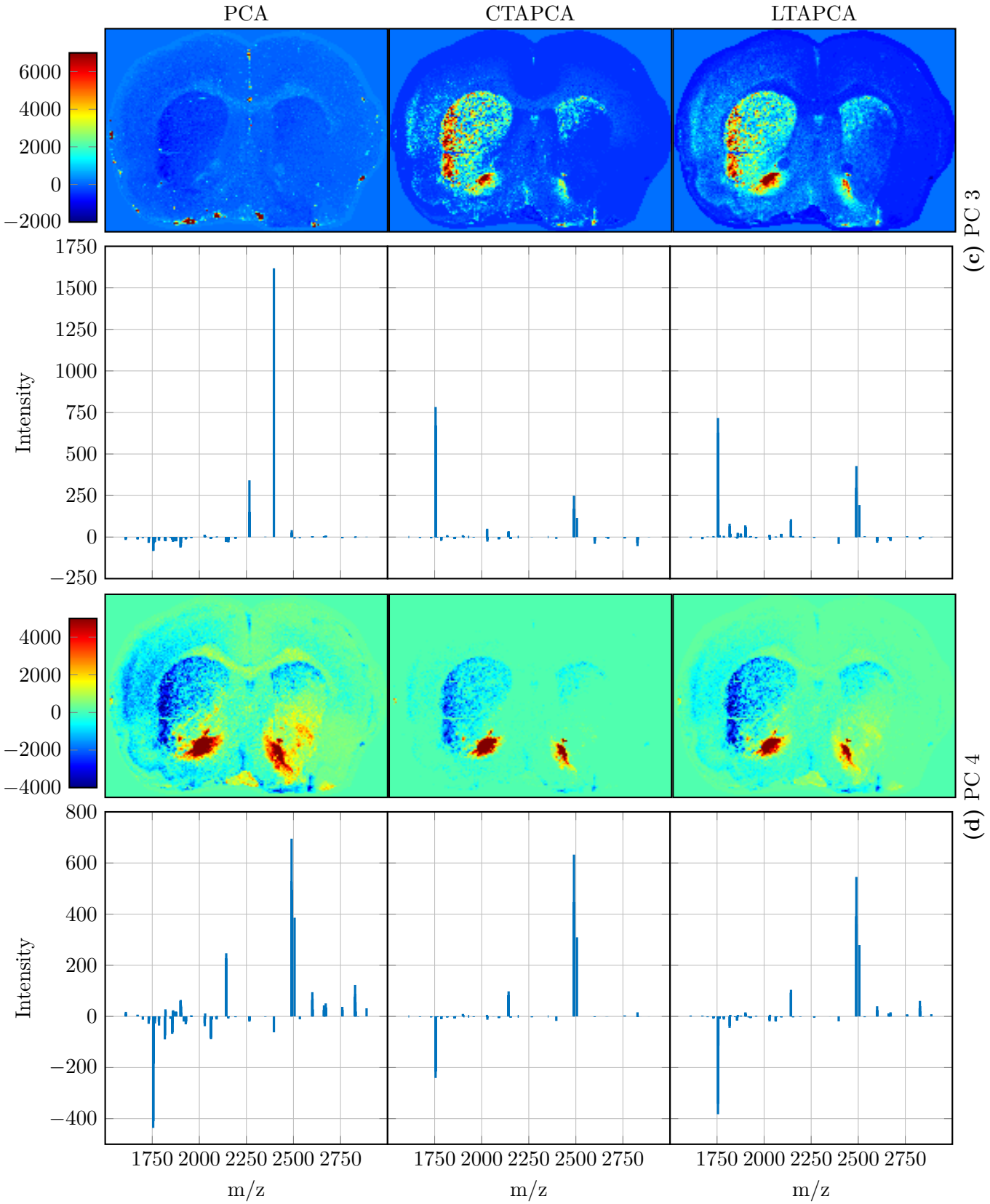


Figure 5-4: The scores (top) and loadings (bottom) of the third **(a)** and fourth **(b)** principal components constructed with PCA, CTAPCA, and LTAPCA for threshold $\tau = 1500$, $c = \frac{\tau}{5}$ and $d = \tau$ on the first 100 mass-bins of the Coronal Rat Brain dataset.

In this dataset we distinguish three distinct cases, in which the feature in the particular ion image is mainly described:

- a set of below-threshold intensities.
- a set of partly above-threshold partly below-threshold intensities.
- a set of intensities predominantly but not all above the threshold.

Below-threshold patterns Figure 5-5 demonstrates, as desired, that both CTAPCA and LTAPCA do not capture any predominantly below-threshold intensities in this ion image. PCA, on the other hand, captures a rough cross-section. The absence of captured information for TAPCA versions can be attributed to the minimal contribution to the covariance of the below-threshold intensities in this mass-bin, as a result of the imposed shift transformation. Consequently, CTAPCA does not describe the variance in this mass-bin in the first set of PCs. Similarly, LTAPCA also seems to capture only little of the predominantly below-threshold intensities present in this particular image. Moreover, as a result of the linear transformation, the intensities are lowered, but still present. The difference in the two projection procedures is the result of basing it on the mass-bin mean after transformations or the original mass-bin mean.

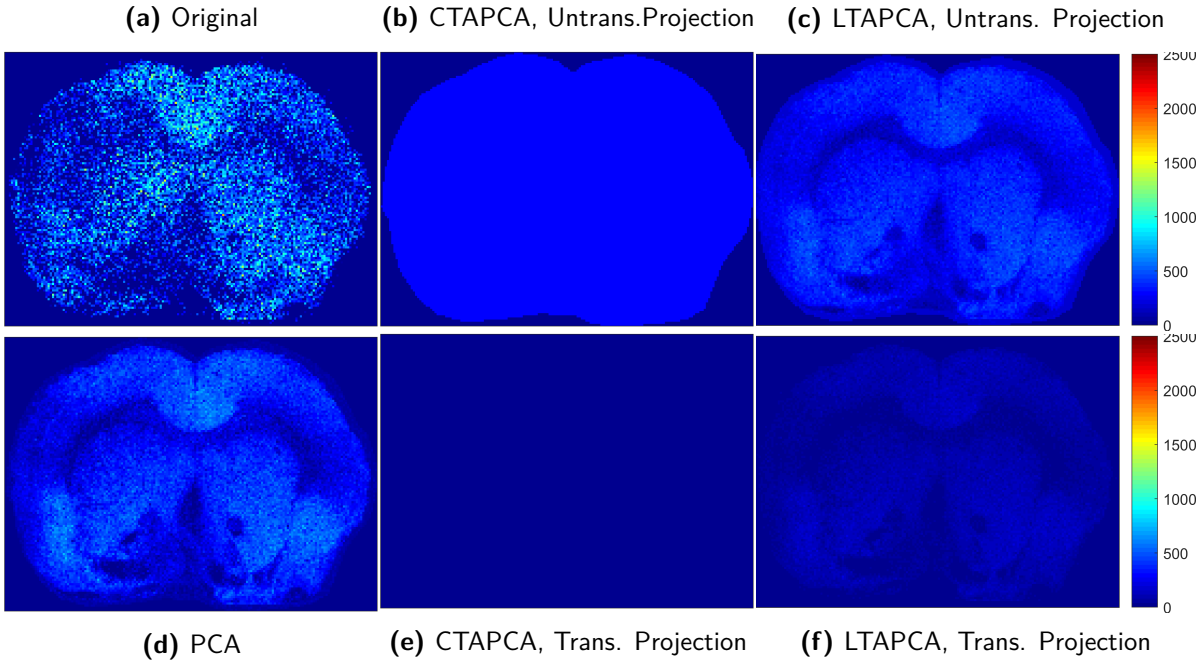


Figure 5-5: Comparison of an ion image associated with mass-bin m/z 1608.81 with predominantly below-threshold intensities and its capturing by traditional PCA, CTAPCA and LTAPCA with threshold 1500 and rank 11 on the first 100 columns of the IMS dataset. Figures **a** and **b** show the capturing when the mean-deviations of the original intensities prior to transformation are projected on the rank 11 basis, whereas **e** and **f** show the capturing when the mean-deviations of the post-transformation intensities are projected as discussed before in section 4-2-3.

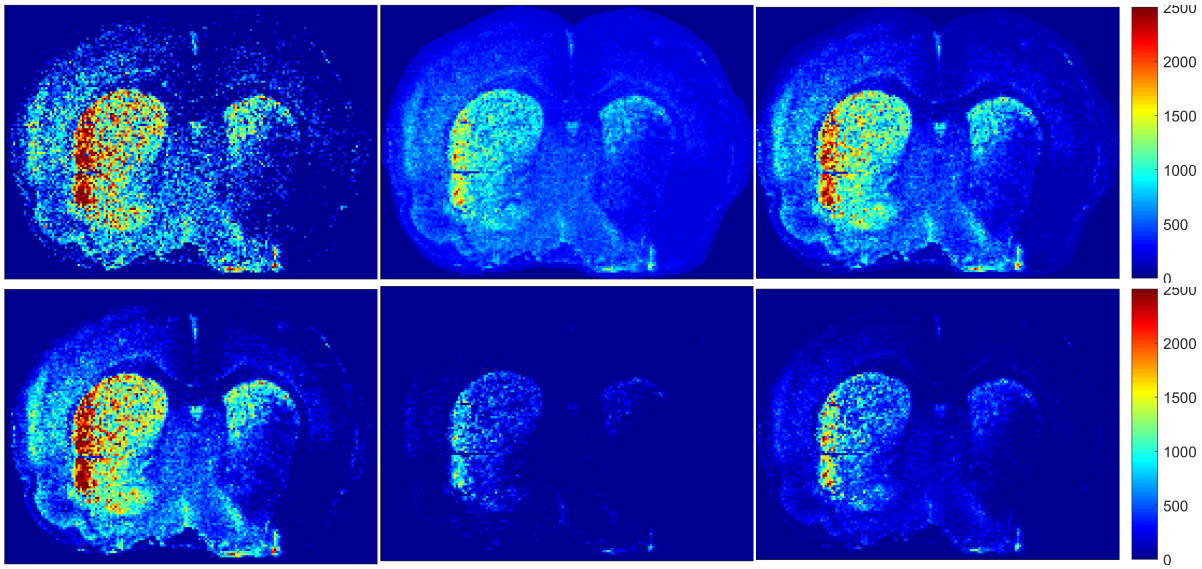


Figure 5-6: Comparison of an ion image associated with mass-bin m/z 1756.97 with partly-above and partly-below threshold intensities for intensity threshold 1500 and its capturing by traditional PCA, CTAPCA, and LTAPCA with rank 11 on the first 100 columns of the IMS dataset. Figures **a** and **b** show the capturing when the mean-deviations of the original intensities prior to transformation are projected on the rank 11 basis, whereas **e** and **f** show the capturing when the mean-deviations of the post-transformation intensities are projected as discussed before in section 4-2-3.

Partly above-threshold partly below-threshold patterns Figure 5-6 shows that, in contrary to PCA, both CTAPCA figure 5-6e and LTAPCA figure 5-6f methods neglect parts of the feature described in this ion image, consisting of partly above and partly below-threshold intensities. After inspection of the covariance matrix and covariant high-intensity mass-bins 1755.0 and 1756.0, we believe this can be attributed to the mostly low-intensity values and few high-intensity spikes describing this pattern. As a result of the imposed shift, the covariance contribution of these low-intensity values is reduced in such a way that other mass-bins contribute more covariance the covariance matrix. Consequently, we expect this feature to not be described in the first set of PCs.

This observation reveals the limitation of TAPCA. TAPCA can miss patterns supported by a small number of intensities above the threshold if their total contribution of covariance associated with above-threshold intensities is small, for example, sparse patterns with a limited total intensity above the threshold. The reconstruction after application of LTAPCA shows an improvement over CTAPCA by demonstration of slightly higher intensities. These higher intensities can be attributed to the linear transformation of the below-threshold intensities, causing these intensities in this ion image to contribute more to the covariance matrix than in the case of CTAPCA.

Figure 5-6b shows that the projection of the original deviations also fails to capture the ma-

jority of the high intensities similar to the projections of the deviations after transformations, figure 5-6e. The equalized background is also the effect or projection on the mass-bin mean as before the transformation.

Figure 5-8 shows the distribution of residuals for this particular mass-bin. CTAPCA in figure 5-8d shows more near-zero residuals but at the same time also a few more substantial outliers in line with the ion images where CTAPCA misses most of the spatial pattern. LTAPCA in figure 5-8e shows an increase in the spread in the residuals, which is comparable with partial capturing of the spatial pattern in the ion image. The projection of the original deviations in figure 5-8b and figure 5-8c show overall an increased spread in residuals, which affects CTAPCA more. Two potential causes are that TAPCA misses the majority of the spatial pattern in this mass-bin and that in the case of this projection we compare with the original measurements. The higher intensities in the original measurements could cause an increase in the residual between a partially captured intensity peak and the original intensity.

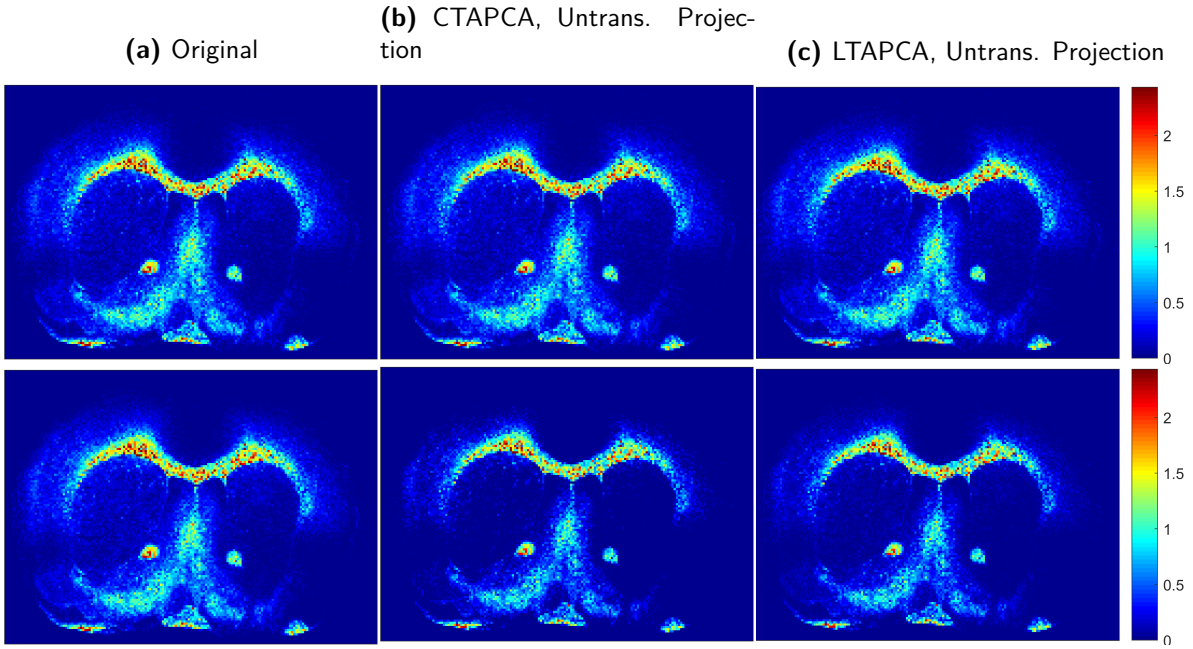


Figure 5-7: Comparison of an ion image associated with mass-bin m/z 2029.07 with predominantly intensities above the intensity threshold 1500 and its capturing by traditional PCA, CTAPCA and LTAPCA with rank 11 on the first 100 columns of the IMS dataset. Figures **a** and **b** show the capturing when the mean-deviations of the original intensities prior to transformation are projected on the rank 11 basis, whereas **e** and **f** show the capturing when the mean-deviations of the post-transformation intensities are projected as discussed before in section 4-2-3.

Dense threshold-exceeding patterns Figure 5-7 demonstrates that patterns supported by a large quantity of threshold-exceeding intensities are captured in a similar manner in both PCA and TAPCA. TAPCA in figure 5-7e and figure 5-7f show less clutter in the low-intensity background of this particular mass-bin compared PCA. However, in the case the original deviations from the mean are projected for TAPCA we observe similar background clutter

again. Figure 5-9 shows the distribution of residuals for this particular mass-bin, in which we can see that overall TAPCA results in significantly lower residuals for this mass-bin compared to PCA.

The per mass-bin RMS values in figure 5-10 show the effect of the transformation and different projection on the reconstruction from the low-dimensional latent representation. Between the original and the PCA reconstruction, figure 5-10a and figure 5-10d, most of the differences are in the slightly lowered RMS values for the originally small RMS intensity peaks in the original spectrum.

Between the original RMS spectrum and the TAPCA reconstruction, figure 5-10e and figure 5-10f, we see all RMS intensities are significantly lowered as a result of the clip-shift and linear-shift transformation. In the CTAPCA case, we see some the RMS intensities of some mass-bins have almost completely vanished, resulting in a sparse spectrum. In the case of LTAPCA, we see an intermediate situation, in which already low RMS intensities have been diminished but not completely vanished yet.

Between the original RMS spectrum and the TAPCA reconstruction with the projection of the original deviations, figure 5-10b and figure 5-10c, we see that overall the spectra are similar. Interesting is that some individual RMS intensities are amplified while other attenuated. A potential cause is that the minimal residual proof for PCA [40] is no longer valid in this projection. Another potential cause is the method we use to project the deviations on the mean. As a result of this projection, the RMS value could be amplified if below-mean intensities are insufficiently described and as such are considered larger than in reality.

Implications on intensity-aware rank estimation

To summarize, we see TAPCA overall producing sparser spectra and more clear delineation of spatial patterns in the principal components. LTAPCA is producing more dense spectra than CTAPCA and results in a less crisp delineation of spatial patterns for equal thresholds. However, TAPCA may fail to sufficiently capture certain patterns partly above, partly below the ion intensity threshold. Overall, the residuals for high-intensity mass-bins seem to become smaller. We want to note that these significantly smaller residuals could potentially be a sign of overfitting high-intensity patterns. More research on other datasets is required to validate these effects. The mass-bin consisting of partly above, partly below the ion intensity threshold showed an overall decrease in residuals but an increased number of large outliers.

In the context of intensity-aware rank estimation, we see that TAPCA is able to capture the threshold-exceeding intensities, but mass-bins consisting of many near-threshold intensities might be an issue.

Regarding the projection of the original deviations from the mean on the basis obtained with TAPCA, we see it has similar issues with partly above-threshold partly-below threshold patterns as the default projection method TAPCA. Nonetheless, it offers the benefit that we end up with spectra in similar magnitude. However, for the evaluated mass-bins the increased spread in residual histograms demonstrate worsened reconstruction of the particular mass-bins. The next section tries to quantitatively evaluate whether this projection method is desired by a comparison of the residuals with respect to their associated intensity.

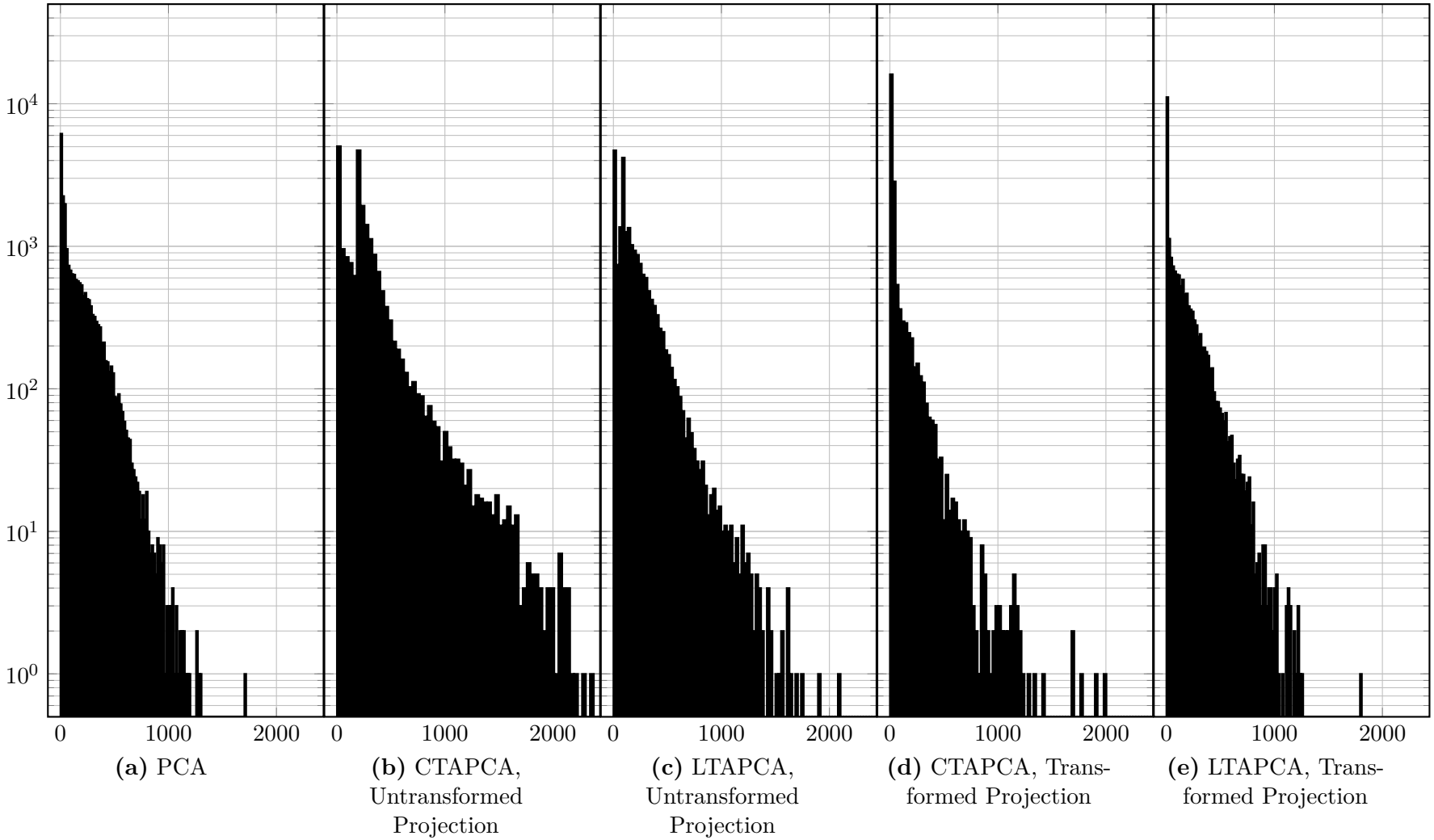
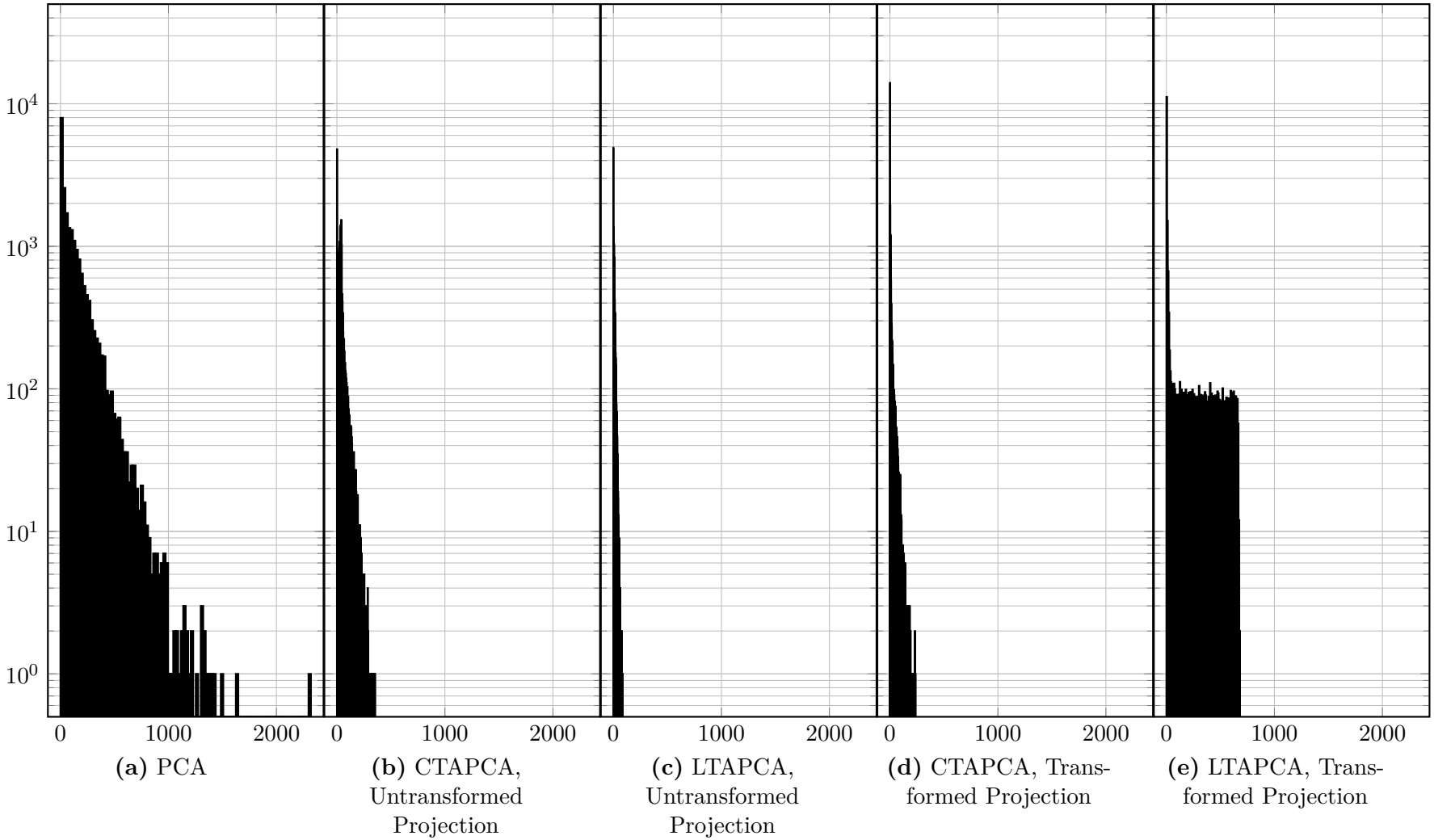


Figure 5-8: Histogram of the absolute residuals associated with partly above-threshold and partly below-threshold intensities between original intensities in the mass-bin and the intensities in the rank-11 reconstructed mass-bin $\left|\mathbf{d}_{\cdot j} - \hat{\mathbf{d}}_{\cdot j}\right|$ in the case of **(a)**, **(b)**, **(c)** and between the transformed mass-bin and the reconstructed mass-bin, $\left|T(\mathbf{d}_{\cdot j}) - \hat{\mathbf{d}}_{T,\cdot j}\right|$ in the case of **(d)**, **(e)**. In this equation j denotes the mass-bin 1756.97 associated with the ion images in the figure 5-6.



(a) PCA **(b)** CTAPCA, Untransformed Projection **(c)** LTAPCA, Untransformed Projection **(d)** CTAPCA, Transformed Projection **(e)** LTAPCA, Transformed Projection

Figure 5-9: Histogram of the absolute residuals associated with predominantly threshold-exceeding intensities between original intensities in mass-bin and the intensities in the rank-11 reconstructed mass-bin $\left|\mathbf{d}_{\cdot j} - \hat{\mathbf{d}}_{\cdot j}\right|$ in the case of **(a)**, **(b)**, **(c)** and between the transformed mass-bin and the reconstructed mass-bin, $\left|T(\mathbf{d}_{\cdot j}) - \hat{\mathbf{d}}_{T,\cdot j}\right|$ in the case of **(d)**, **(e)**. In this equation j denotes the mass-bin 2029.07 associated with the ion images in the figure 5-7.

5-1-3 Comparison of the intensity-dependent capturing

In this section, we quantitatively assess the differences between TAPCA and traditional PCA by a comparison of the residuals between the original data and a low-rank reconstruction obtained by PCA and CTAPCA. We assess both the application of TAPCA with the projection of the deviations from the mean after application of the transformation (transformed) and the projection of the deviations from the mean as before application of the transformation (untransformed), as discussed in section 4-2-3. In a similar manner to section 3-2, we bin the residuals again by their original intensity into intensity windows. Figure 5-11 demonstrates how the residuals for different intensities evolve with increasing rank for traditional PCA (line with dots) and CTAPCA (dots). In the case of the untransformed projection (top), we obtain the residuals by comparison to the original measurements. In the case of the transformed projection (bottom), we obtain the residuals by comparison to the transformed measurements. As a result of this difference in comparison, we are interested in the low-intensity windows ($[0, \tau]$) in the transformed projection case, while in the untransformed we are not. In the case of the untransformed projection, we make the following observations:

- The untransformed projection in figure 5-11a shows predominantly higher Root Mean Squared Residual (RMSR) and Median Absolute Residual (MAR) values for TAPCA than for PCA for relatively low, yet threshold-exceeding intensities. We see predominantly higher RMSR values for TAPCA for all of the reconstructions with $K \leq 7$. These higher residuals combined with predominately increased MAR for these ranks suggest that the capturing of both the threshold-exceeding and the below-threshold patterns worsened for these ranks.
- The untransformed projection in figure 5-11a shows an increase in the bump appearing for high ranks in residuals for low-intensity windows for TAPCA when compared PCA.
- Figure 5-11 a show generally lower RMSR and MAR values for TAPCA for the high intensities for the ranks $K \geq 17$. Furthermore, we see the intersection point between the RMSR and MAR of respectively PCA and TAPCA is moving towards lower intensities for increasing rank.

We hypothesize that the high intensities are adequately described in the bases obtained by TAPCA, due to their contribution to the covariance being less affected by the shift transformation. As a consequence, we see lower residuals of the high intensities, compared to the lower intensities.

For the transformed case, we make the following observations:

- The transformed projection in figure 5-11b shows predominantly lower RMSR and MAR values for TAPCA than for PCA for relatively low, yet thresholding exceeding intensities for ranks $K \leq 7$. For lower ranks, we do not observe conclusively lower residuals.
- The transformed projection in figure 5-11b shows predominantly higher RMSR and MAR values for the TAPCA than for PCA for low intensities (≤ 1000).

Based on these results we question whether dimensionality reduction based on CTAPCA with the projection of the original data is preferred over PCA. Quantitatively, this projection

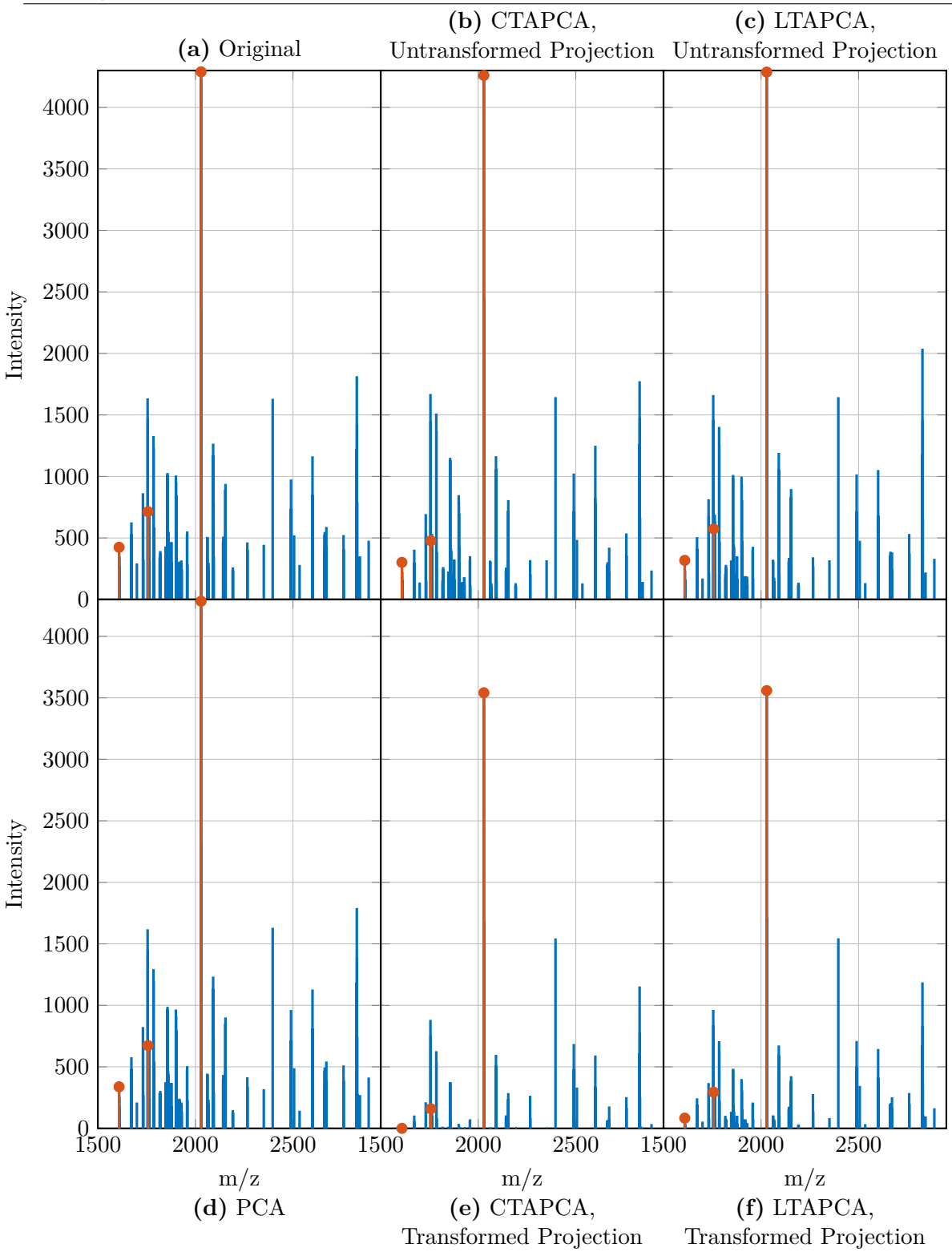


Figure 5-10: Comparison of the RMS intensity per mass-bin of the original dataset, the rank-11 reconstruction of PCA, the rank-11 reconstruction of CTAPCA and LTAPCA with both projection mechanisms. The ion peaks are highlighted corresponding to from left to right respectively predominantly below-threshold intensities (m/z 1608.81), partly below-threshold partly above-threshold intensities, (m/z 1756.97), and predominantly above threshold intensities (m/z 2029.07).

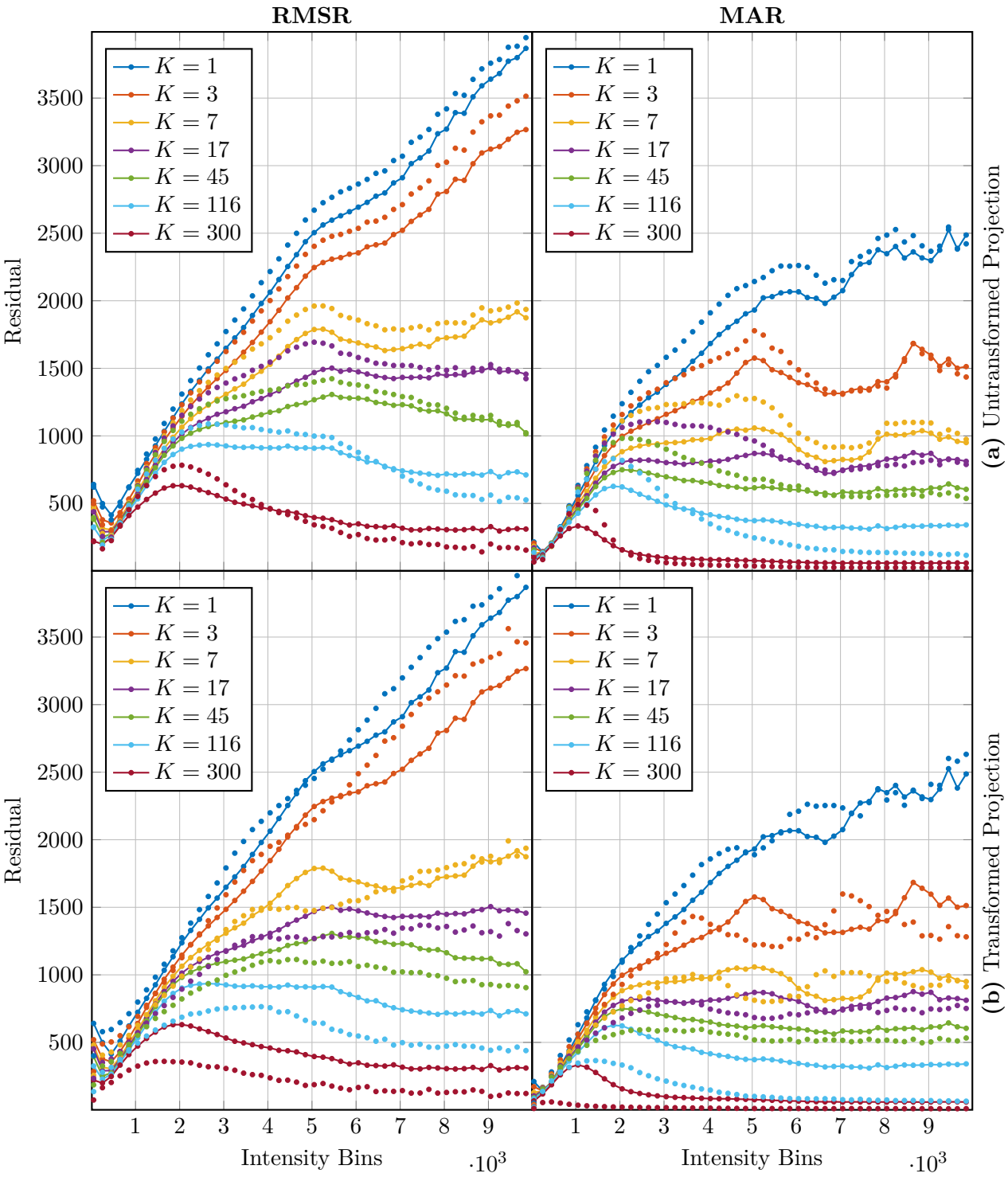


Figure 5-11: The RMSR and MAR of the residuals binned by their original intensity for the IMS dataset with 809 mass-bins for traditional PCA (line with dots) and CTAPCA (dotted) with a threshold $\tau = 1500$ and intensity-bin-width of 200.

shows higher residuals for the threshold-exceeding intensities for the low ranks, which are primarily the ranks of interest. Consequently, this projection does not seem to necessarily improve captured information above the threshold for the ranks of interest. This observation was in line with the expectations due to the nature of PCA. However, we would like to note that the unexpected lower residuals for the high intensities suggest that sufficiently threshold-exceeding features are still better described in this representation.

On the other hand, for CTAPCA with the projection of the transformed deviations, we do see predominantly lower residuals for the threshold-exceeding intensities for all ranks above rank 10, compared to PCA. These results suggest for capturing the threshold-exceeding intensities CTAPCA is preferred over PCA.

5-1-4 Mass-bin contribution

Another way of quantifying the effects of TAPCA is the observation of the relevance of individual mass-bins with respect to the threshold. We expect the shift transformation in TAPCA to reduce the relevance of particular mass-bins, as their variance contribution mainly originates from below-threshold intensities. Similarly, we expect an increase in relevance for mass-bins mainly contributing variance above the threshold.

The introduction of the ion intensity threshold into dimensionality reduction enables determining the relevance of the particular mass-bins with respect to the threshold. One way of specifying the relevance of a particular mass-bin with respect to the threshold is the observation of the coefficient matrix $\mathbf{C}_\tau$, obtained by TAPCA. The coefficients in the individual rows describe the importance of a particular mass-bins and its coherence with other mass-bins in the total low-rank approximation. A higher squared sum of the coefficients in the rows associated with a particular mass-bin means more of the original mass-bin is captured. The row-wise sum of the squared coefficients is maximized at one, as the coefficient matrix $\mathbf{C}_\tau$ is orthonormal. Equation (5-2) shows the formula used to determine the relevance of a particular mass-bin i, for rank-K, in which $c_{\tau,i,j}$ denotes an element at row i and column j of the coefficient matrix $\mathbf{C}_\tau$ obtained after application of TAPCA with threshold τ.

$$r_{\tau,i} = \sqrt{\sum_{j}^{K} (c_{\tau,i,j})^2} \tag{5-2}$$

For visualization purposes, we reduce the dataset again to the first 100 columns of the original dataset with 809 mass-bins. Figure 5-12 shows the individual contribution as a relevance measure for the reduced dataset for various thresholds at rank 15. We require sufficient rank to asses the effect of the threshold on the contribution of the individual mass-bins in a transparent manner. Insufficient rank introduces sudden changes in the individual contribution. The covariant components obtained by TAPCA are possibly ordered differently for some thresholds due to the change in the variance contribution due to the transformation. As such, insufficient rank discards components depending on the threshold and related variance contribution. Rank 15 showed sufficiently stable behavior to make an equal assessment on the influence of the threshold.

In relation to the qualitative assessment done in section 5-1-2, we would like to identify three different cases which are closely related to the captured features.

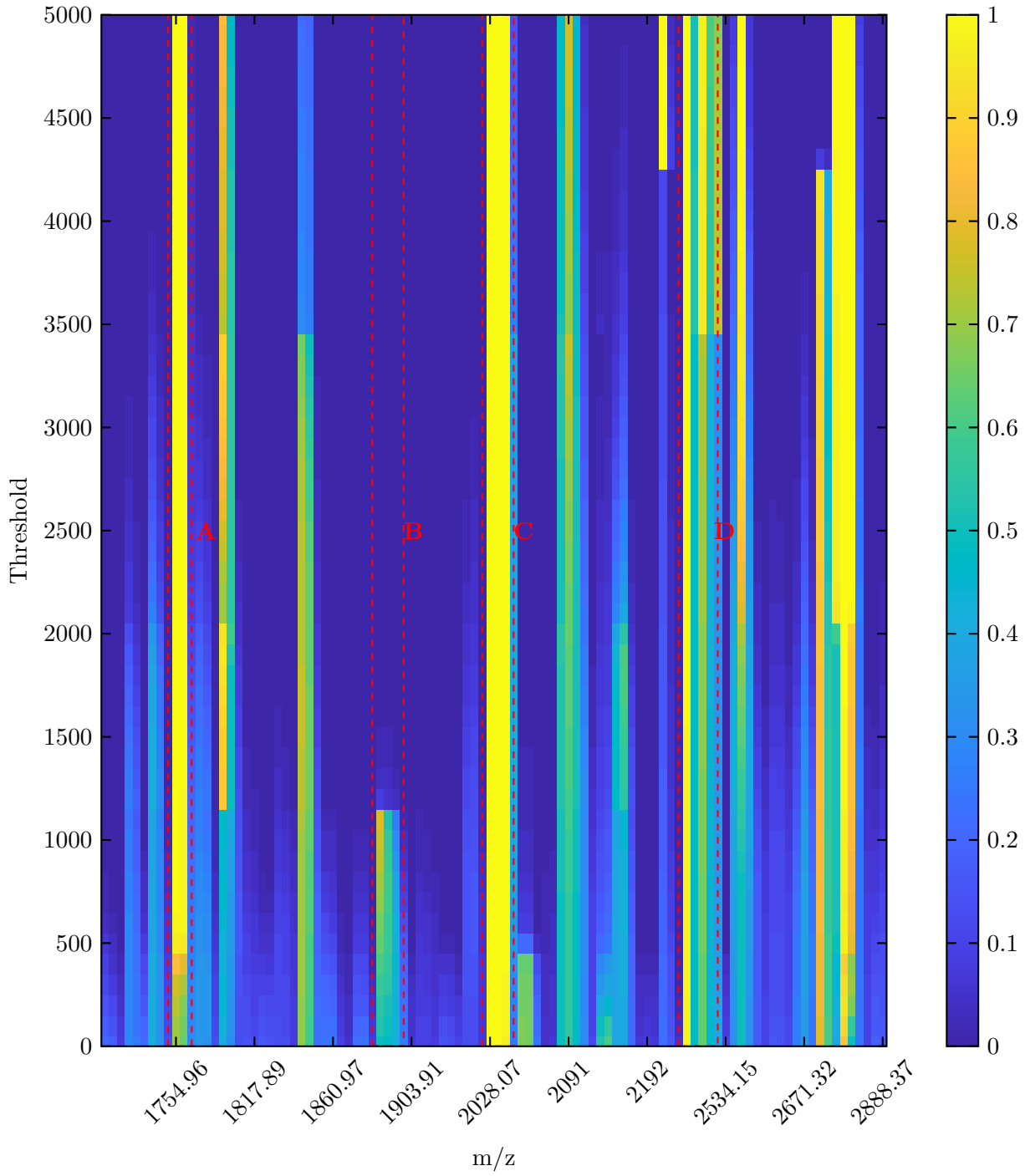


Figure 5-12: The contributions of individual mass-bins at different thresholds for rank 15 after application of TAPCA on the first 100 mass-bins of the IMS dataset peak-picked with threshold 100 and containing 809 mass-bins in total.

Mass-bin contribution attenuated for higher thresholds. These mass-bins consist mostly of intensities in the range $[0-3000]$. As such, for a sufficiently high threshold, the variance contribution of these mass-bins attenuates due to the transformation. Examples are the mass-bins labeled with **B** in figure 5-12.

Mass-bin contribution amplified for higher thresholds. The contribution of the mass-bins representing a sufficient amount of variance associated with a set of intensities above the threshold is amplified. We attribute this amplification on lower variance contribution of the below-threshold intensity bins. As such, the variance contribution of the above-threshold relative to the below-threshold intensity bins is raised. Examples are the mass-bins labeled with **A** in figure 5-12.

Mass-bin contribution remains approximately constant. For the constant contribution cases we can identify two situations. First, the mass-bins containing many high intensities already make a relatively large contribution. As such, the increase in contribution of these bins is not visible on the color scale. Examples are the mass-bins labeled with **C** in figure 5-12. The other situation is mass-bins consisting of a few relatively high intensities. The removal of a set of low intensities reduces the variance contribution of these mass-bins. At the same time, we expect the lower total variance contribution due to the shift transformation counters the effect of the lowered variance contribution of this particular bin. As such, this particular mass-bin contribution remains approximately constant. Examples are the mass-bins labeled with **D** in figure 5-12.

5-2 Rank estimation in the synthetic dataset

We use the synthetic dataset, outlined in section 2-2-2, for the validation of the intensity-aware rank estimation method. We consider here the residual-fraction rank estimation method based on the residual fraction and the threshold-shifted method based on intensity-aware dimensionality reduction as derived in the previous chapters. The synthetic dataset simulates a clear cut in information-carrying threshold-exceeding intensities and randomly permuted below-threshold intensities.

5-2-1 Residual-fraction rank estimation

Figure 5-13 displays the residual fraction, equation (3-12), for two synthetic datasets, with and without additional noise and with rank 5. Both plots display a sharp peak in the residual ratio at rank 5. At this rank, the reconstruction of threshold-exceeding intensities is optimal compared to the below-threshold intensities as their respective total residuals are minimal. This optimal reconstruction implies that the majority of the threshold-exceeding information is captured, while the below-threshold is mostly discarded. This peak in the residual ratio confirms the hypothesis that in the dataset, for which only the threshold-exceeding intensities reflect reliable signal, the rank, for which the majority of the threshold-exceeding information is captured, is close to the rank of the underlying reliable signal.

Furthermore, the residual ratio demonstrates sensitivity to the threshold choice, as the choices of the threshold below and above the actual threshold of the synthetic dataset exhibit different

behavior. The residual ratio for a too-low threshold choice demonstrates no optimum for the above threshold reconstruction. The residual ratio for a too-high threshold choice increases the height of the peak. We have observed, that the synthetic datasets are more likely to slightly overestimate actual rank based on the residual ratio dependent on the choice of threshold.

5-2-2 Threshold-shifted rank estimation

Figure 5-14 displays the cumulative explained variance based on TAPCA, section 4-2, per choice of threshold for the same synthetic datasets with and without additional noise. We use the no-noise case to discuss the behavior we observe taking into account the threshold in rank estimation. Then, we demonstrate that the behavior in the added noise case is similar. The contour lines in these plots show the rank required for a percentage of explained variance in the original dataset with respect to the threshold. At threshold 0 no threshold is taken into account, and this situation corresponds to a percentage of variance based rank estimation in traditional PCA.

Figure 5-14a shows that in the no-noise case we would require rank 16 to capture 99% of the variance, not taking into account the threshold. The underlying rank of the synthetic dataset is 5, but the random permutations of the below threshold intensities construct a full rank (25) matrix. We attribute the increased rank to capture the majority of the variance to the randomness of the below-threshold intensities. For a too-low threshold, TAPCA considers the variance of the sub-threshold intensities relevant. As such, a significant amount of components is required to capture the majority of the variance.

We see a decrease in required rank for equivalent percentages of variance captured, for an increase of the threshold used for TAPCA, as the below-threshold variance contribution is lowered. For a threshold close to the synthetic dataset threshold, we see with TAPCA for approximately 98% we would need rank 5. For this threshold, TAPCA discards the majority of variance contributions by below-threshold intensities, yet we do not obtain 100% captured variance. The shifting transformation does not preserve the original factors $\mathbf{W}$ and $\mathbf{H}$ in Equation 2-2 constructing the dataset. Consequently, this method is unable to discover the exact original factors and always result in some information loss.

In the noise case, with 10% added Gaussian noise we require capture of approximately 90% of the variance in the synthetic dataset. Figure 5-14b displays the synthetic dataset rank 5 for threshold 5 and 90% explained variance. In the presence of noise, it seems that the effect of the noise is dominant over the shifting not preserving the factors $\mathbf{W}$ and $\mathbf{H}$.

Furthermore, the noise case shows a more smooth reduction in explained variance compared to the no-noise case. We attribute to the noise also distortion of the threshold-exceeding intensities. Independent of the threshold, the variance in the signal becomes less coherent due to this distortion, requiring more components to capture the majority of the variance. For an increasing threshold, the variance-contribution of the random below-threshold intensities becomes less, causing an increase in the explained variance for the same rank.

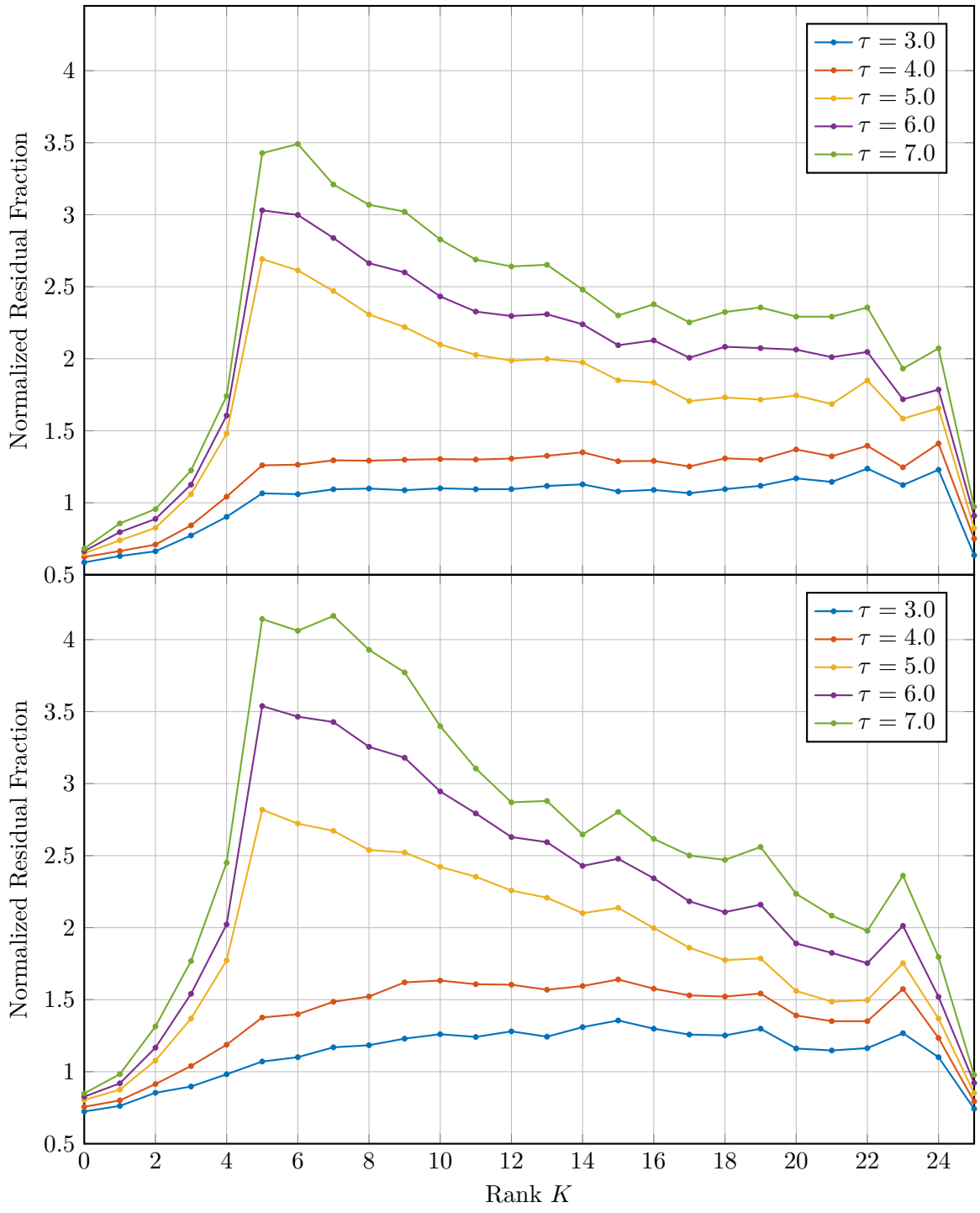
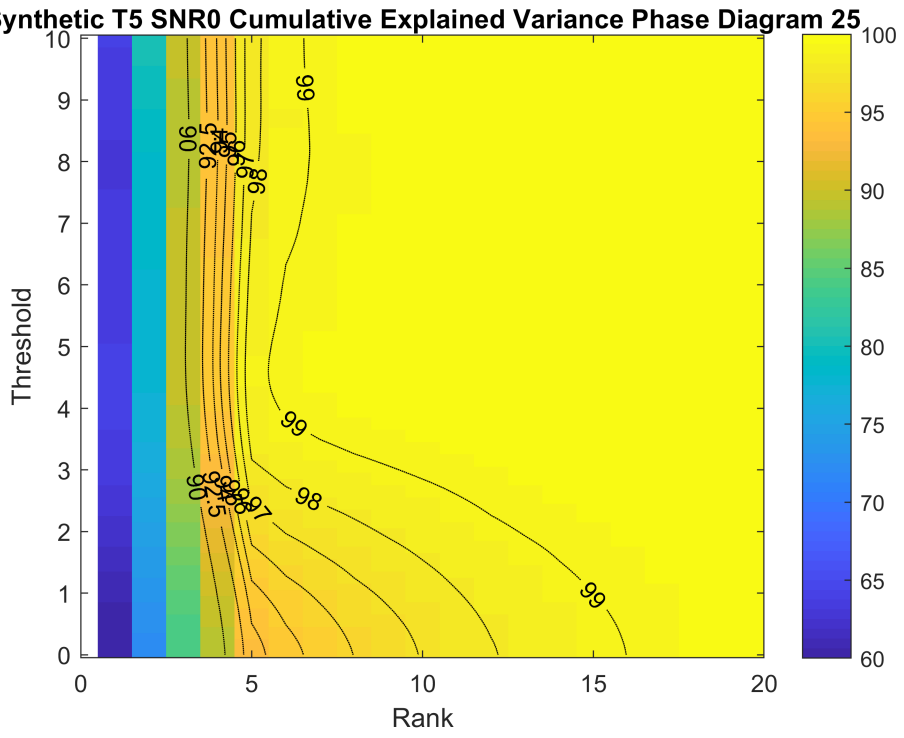
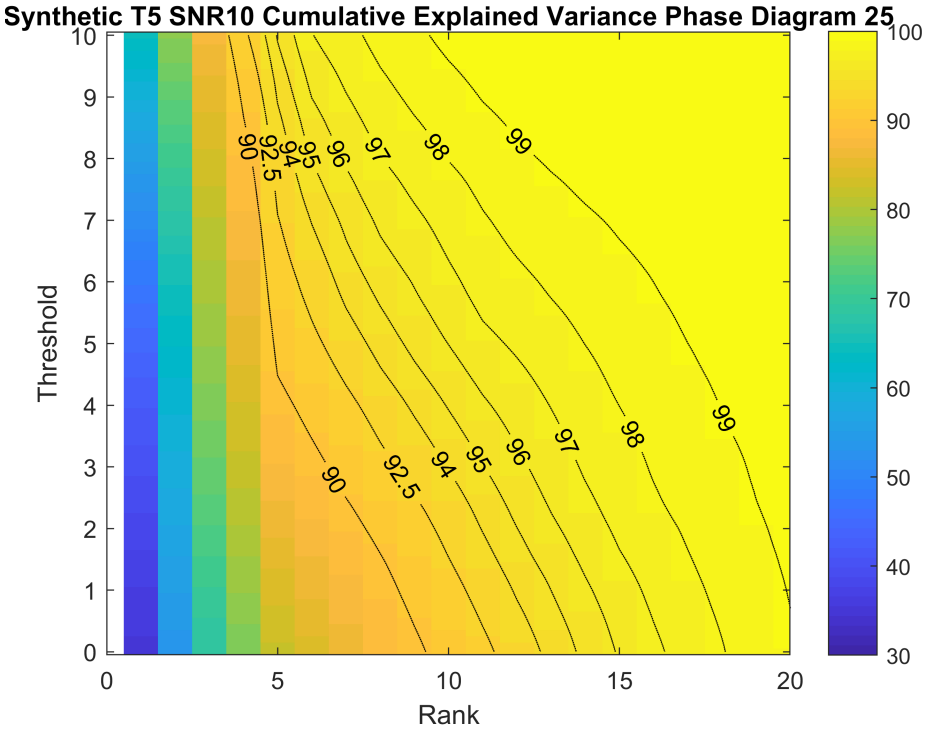


Figure 5-13: The residual fraction for the synthetic datasets in the case of no noise and 10% noise with dimensions 50×25, rank 5. The below-threshold intensities are randomized for a threshold of 5.



(a) SNR $= \infty$



(b) SNR $= 10$

Figure 5-14: The cumulative explained variance obtained with TAPCA per threshold and rank for two synthetic datasets with dimensions 50×25, rank 5. The below-threshold intensities are randomized for a threshold of 5

5-3 Rank estimation for the IMS Dataset

This section compares intensity-aware rank estimation using residual-fraction rank estimation and threshold-shifted rank estimation with the percentage of cumulative explained variance based on TAPCA to a baseline of rank estimation methods available in the literature. The rank estimation methods selected from the literature for reference are Cross-Validation (CV) [28], and the Cumulative Percentage of Explained Variance [40].

5-3-1 Threshold-shifted rank estimation

In this section, we compare threshold-aware rank estimation based explained variance for TAPCA with threshold-unaware rank estimation based on explained variance for normal PCA when applied to the Coronal Rat Brain dataset. Figure 5-15 shows the percentage of cumulative explained variance in function of the threshold and a specific choice of rank, for two dataset sizes obtained via TAPCA. The percentage of cumulative explained variance at intensity threshold zero is identical to rank estimation via cumulative explained variance of traditional PCA. In the context of IMS and TAPCA we can do the following observations:

Higher thresholds require decreasing rank for same cumulative explained variance. All plots in figure 5-15 show for an increasing threshold a decrease in the required rank to obtain a similar percentage of cumulative explained variance. The reduction in rank is in line with the expectations outlined in section 4-2. The below-threshold variance contribution is discarded. At the same time, this variance is expected to be mostly incoherent, due to low physical reliability of the intensities. Consequently, this supports our hypothesis that intensity-aware rank estimation via explained variance of TAPCA results in a lower rank estimate.

The rate of change of rank versus threshold increases for larger datasets. The comparison of Figure 5-15b to figure 5-15a shows that larger dataset sizes allow more reduction in the required rank to obtain a similar percentage of cumulative explained variance. We attribute this to the difference in peak picking thresholds for these datasets. A lower peak picking threshold results in more mass-bins with a significant amount of intensities below the threshold as shown in the histograms in figure 2-3. Consequently, the shift transformation, as applied in TAPCA, results in a considerable reduction in considered variance, due to the increased number of below-threshold intensities for the larger datasets.

Higher rank estimates for LTAPCA versus the CTAPCA. The linear-shift transformation, as introduced in section 4-2-4, suppresses the variance contribution below the threshold, whereas the clip-shifting removes the below threshold variance contribution altogether. Consequently, we observe in figure 5-16 that the rank for an equal percentage of explained variance of the linear-shifting variant is higher. Furthermore, a larger part of this variance is contributed by intensities that we consider unreliable in light of the threshold. As such, we expect a lower rank estimate for the clip-shifted version when compared to the soft shifted variant.

Significantly higher rank estimates when compared to CV Rank estimation via explained variance for both TAPCA and PCA give significantly higher estimates for variance thresholds common in literature (90%+) at low thresholds ($\tau \leq 1000$) than CV. We ascribe this discrepancy to, one, a fundamental difference between cross validation and explained variance rank estimation and, two, a substantial variance contribution of unreliable below-threshold intensities for low thresholds. CV intends to describe

5-3-2 Rank estimation based on Cross-Validation (CV)

This section introduces the CV method we have selected to compare with the intensity-aware methods when applied to a real IMS dataset. In CV for dimensionality reduction a part of the dataset is held out, and a low-rank approximation is obtained from the left-over data. Subsequently, the low rank approximation constructed without held-out data is ranked on how well it describes the held-out data for a particular rank by the calculation of the Predicted Residual Sum of Squares (PRESS). By iterative evaluation of this prediction error for different hold-outs an estimate of the optimal rank can be obtained.

We chose PLS Eigenvector$^{\text{TM}}$ as the preferred CV method, based on the review of Bro et al. [28], for to its simplicity, its resilience to overfitting, and computational advantages over Bi-Cross-Validation [27] and computation of PCA with missing values [28, 40]. Furthermore, we chose a L-fold leave-p-out CV [28, 26] approach over leave-one-out, due to two reasons. One, other methods, such as leave-one-out, would require a large number of iterations, due to the high dimensionality of IMS datasets and associated computational load. Two, we expect leave-p-out approach to perform better in IMS datasets, since a significant amount of mass-bins contain spatially similar patterns. Similar patterns would still be present in the training dataset after application of leave-one-out CV, enabling near-perfect reconstruction of the holdout. As a result, we expect leave-one-out to be more prone to overfitting than leave-p-out strategy.

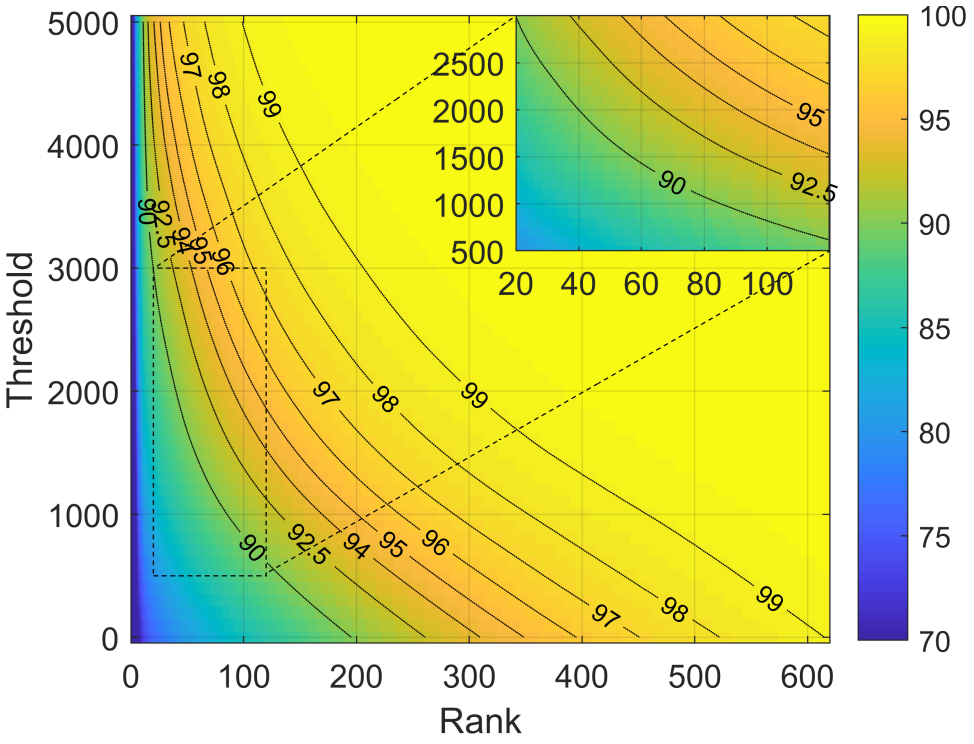
The matrix version of the PRESS for rank K and associated hold out i for the selected CV method is:

$$\text{PRESS}_{\text{K, i}} = \left\| \left(\mathbf{I} - \mathbf{C}_{(-i)} \mathbf{C}_{(-\mathbf{i})}{}^{T} + \text{diag}\{ \mathbf{C}_{(-i)} \mathbf{C}_{(-\mathbf{i})}{}^{T} \} \right) \mathbf{D}_{(i)} \right\|^2, \tag{5-3}$$

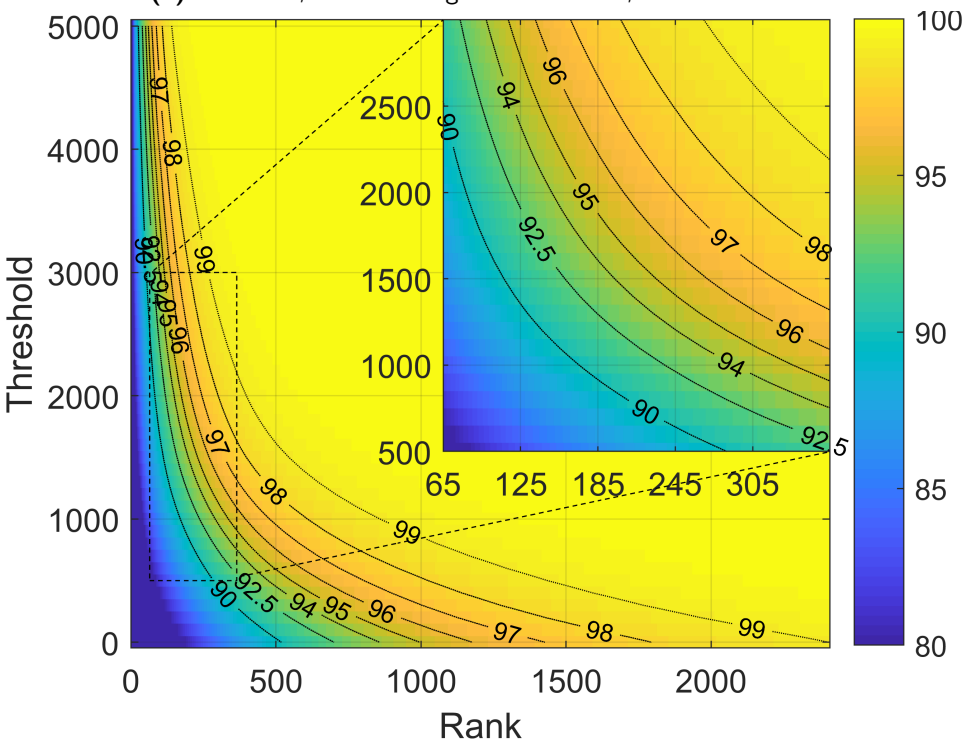
in which $\mathbf{I} \in \mathbb{R}^{N \times M}$ denotes the identity matrix, $\mathbf{C}_{(-i)} \in \mathbb{R}^{M \times K}$ the rank-K coefficient matrix as obtained from the training set $\mathbf{D}_{(-i)}$, $\mathbf{Y}$ without the rows i, test set $\mathbf{D}_{(i)}$ denotes the rows i in the matrix $\mathbf{D}$ and diag$\bullet$ denotes a matrix in which all off-diagonal elements are set to zero. The average PRESS for rank K can then be calculated via:

$$\text{PRESS}_{\text{K}} = \frac{\sum_{i=1}^{L} \text{PRESS}_{\text{K, i}}}{L} \tag{5-4}$$

Figure 5-17 shows the average PRESS per rank for CV with $L = 100$ different random hold-outs for three different dataset sizes. The rows of $\mathbf{Y}$ are held out randomly, and we chose the number of held-out rows empirically on 180, which is approximately one percent of the dataset. We see in figure 5-17 for the three different dataset sizes yield three different ranks, respectively 15, 16, and 24, for which the average PRESS is minimal. Generally, in rank estimation based on CV, the rank is chosen where the average PRESS is minimal. However, we note that the relative differences in the domain $[14, 24]$ are small.

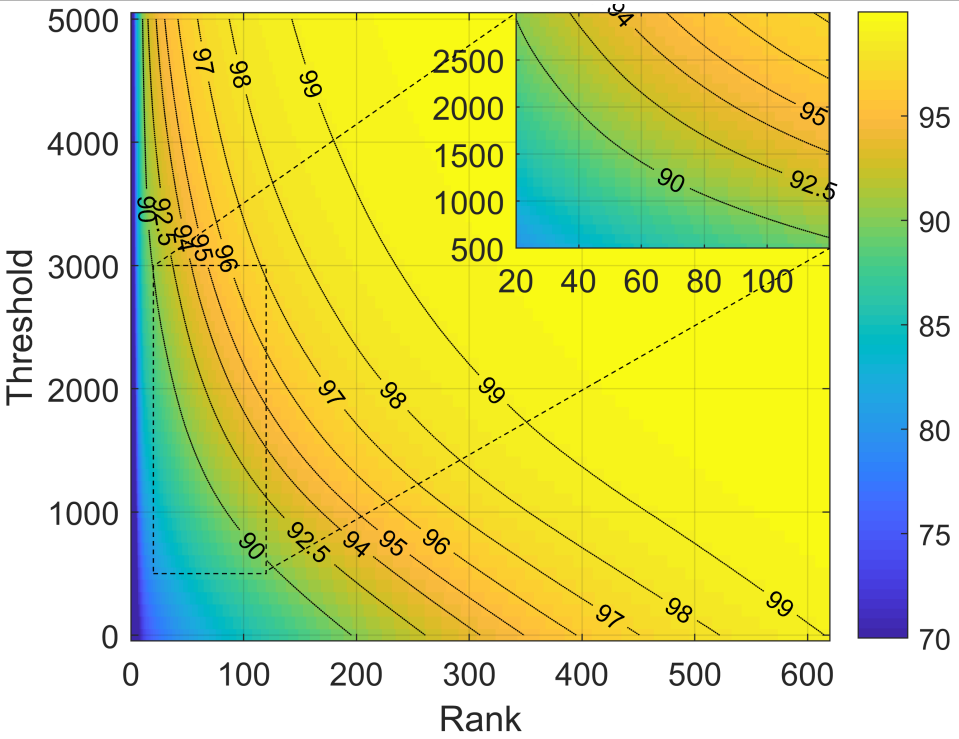


(a) CTAPCA, Peak Picking Threshold 100, 809 mass-bins

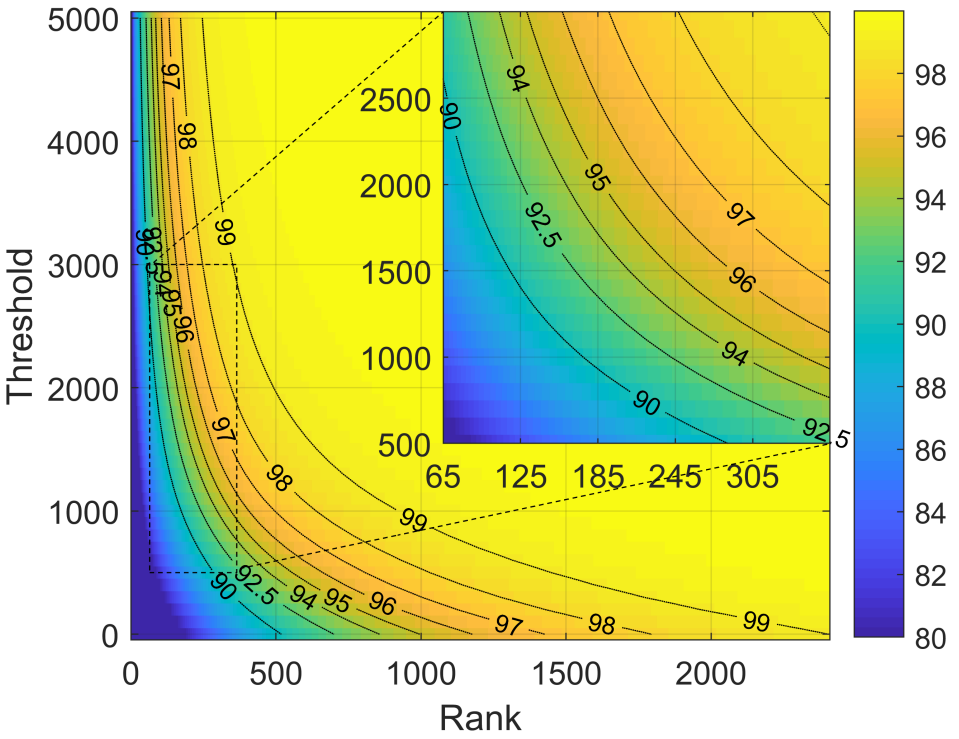


(b) CTAPCA, Peak Picking Threshold 10, 4084 mass-bins

Figure 5-15: The percentage of cumulative explained variance per rank plotted against the used intensity threshold τ for CTAPCA for two different dataset sizes. The contour lines show the variance-percentage truncations often chosen in the literature.



(a) LTAPCA, Peak Picking Threshold 100, 809 mass-bins



(b) LTAPCA, Peak Picking Threshold 10, 4084 mass-bins

Figure 5-16: The percentage of cumulative explained variance per rank plotted against the used intensity threshold τ for LTAPCA for two different dataset sizes. The contour lines show the variance-percentage truncations often chosen in the literature. In the linear-shift transformation the parameter c and d are chosen in relation with the threshold as $c = \frac{\tau}{5}$, $d = \frac{\tau}{2}$.

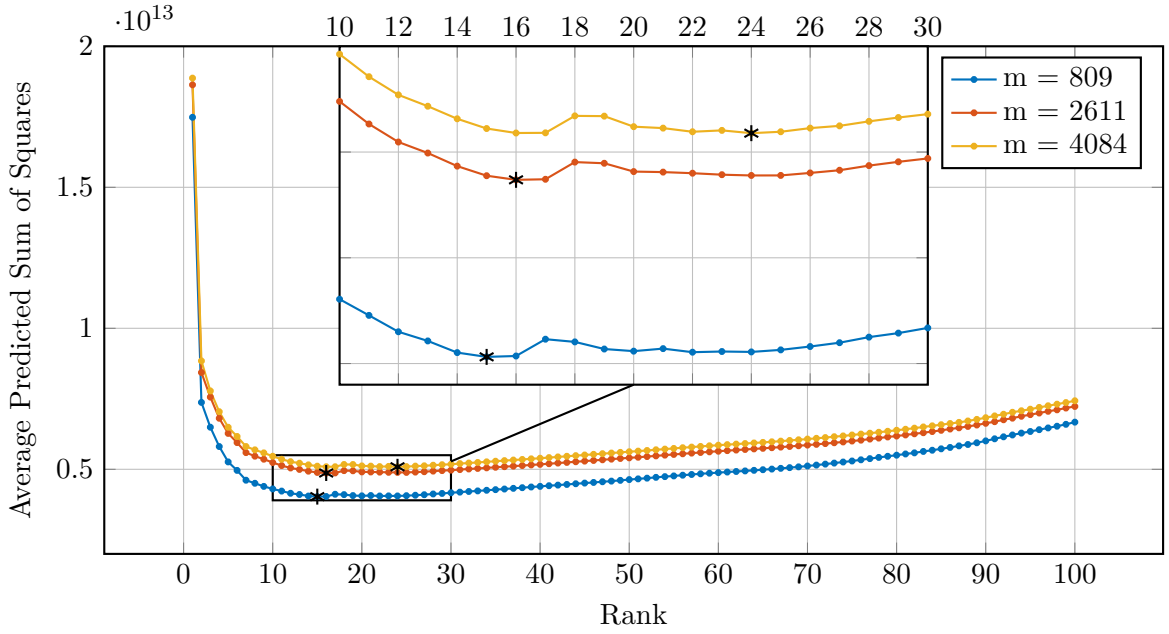


Figure 5-17: Total average residuals for CV using the PLS Eigenvector [28] method for the three different dataset sizes with respectively 809, 2611 and 4084 mass-bins averaged over 100 hold-outs.

5-3-3 Residual-fraction rank estimation

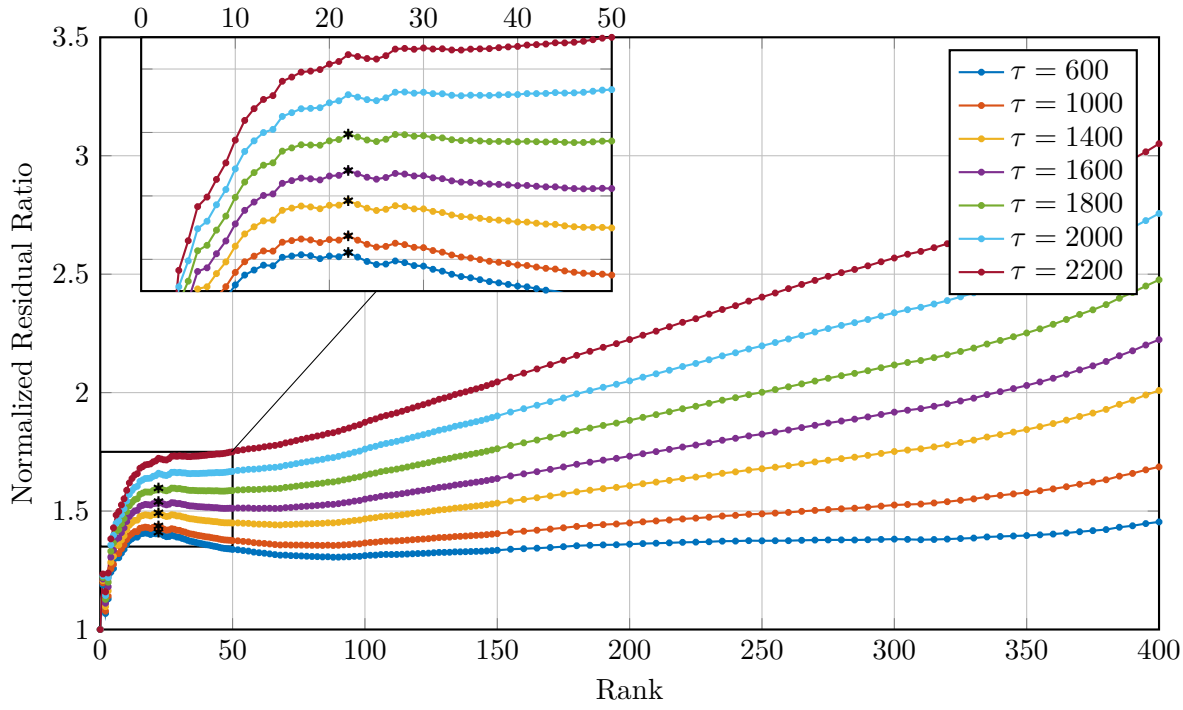
In this section, we compare residual-fraction rank with the threshold-unaware methods like cross-validation and explained-variance for the Coronal Rat Brain dataset.

Figure 5-18 plots the above-threshold and below-threshold residual ratio against the rank for a real IMS dataset. The plots can be divided into three ranges:

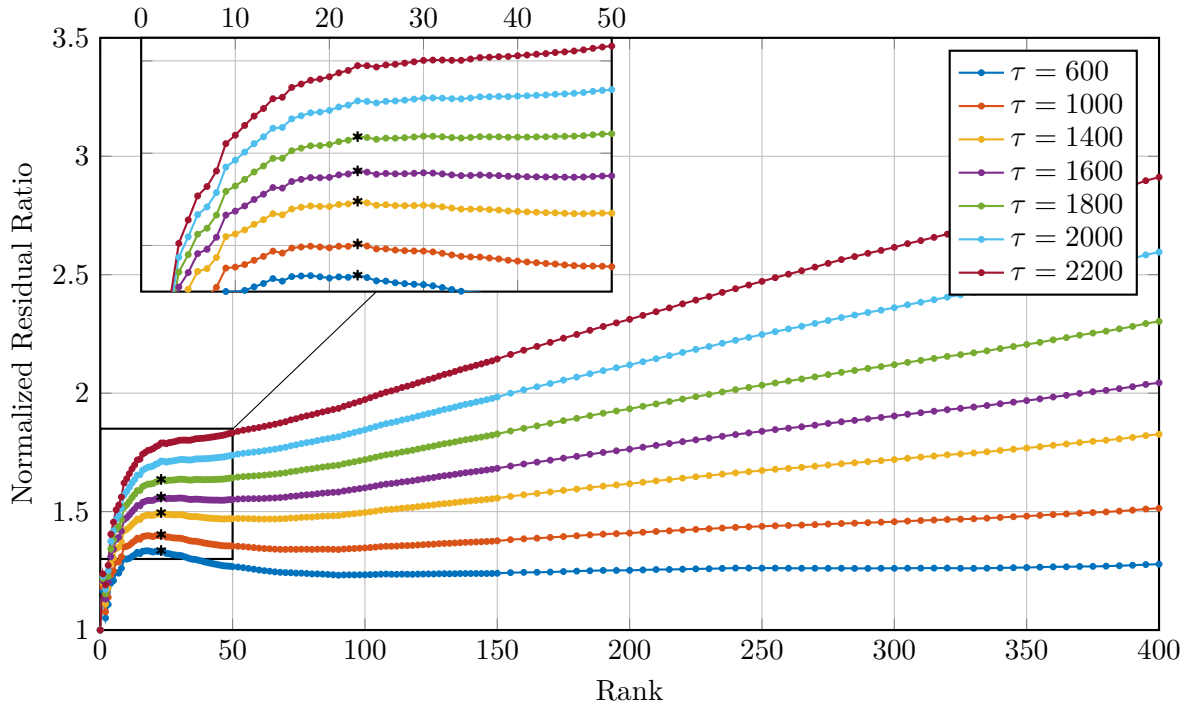
Low-rank (0-20) The residual ratio increases steadily, meaning the above-threshold peaks are captured better than the below-threshold counterpart. In this region, PCA captures a large part of the threshold-exceeding information. This observation is supported by the decrease in the residuals for the high-intensity windows for these ranks shown in figure 3-1.

Medium Rank (20-45) The residual ratio starts to stabilize. We interpret this stabilization as the additional components capture similar amounts from the above-threshold and the below-threshold intensities. In figure 3-1 we see a slower reduction of the residuals for the high-intensity bins. For thresholds $\tau < 1600$, the ratio at the high-end of this region reduces, meaning more below-threshold intensities are captured.

High Rank (45+) The ratio develops approximately in a linear way for the thresholds. For the thresholds $\tau > 1000$, the ratio increases steadily at a rate dependent on the height of the threshold, meaning the above-threshold intensities are captured in a better way than the below-threshold components. For the thresholds $\tau < 1000$ the ratio remains



(a) Dataset 809 bins



(b) Dataset 2611 bins

Figure 5-18: Ratio between below-threshold RSSQ and above threshold RSSQ as defined in equation (3-12) for IMS. An increase in the ratio indicates components, which capture more of the above-threshold intensities. Similarly, a reduction of the ratio signals components capturing more of the below-threshold intensities.

approximately constant, meaning additional components capture roughly equal amounts of below- and above-threshold intensities.

Figure 5-18 shows a maximum in the ratio around approximately rank 22 or 23 dependent of the dataset size. As we discussed in section 3-3, this maximum implies minimal above-threshold residuals relative to below-threshold residuals, or in other words, maximal capturing for the threshold-exceeding patterns when compared to the below-threshold ones.

Figure 5-17 shows that CV finds a similar rank as maximization of the residual ratio for the 4084 mass-bin dataset. We see that CV of 809 and 2611 mass-bin datasets result in significantly lower ranks of respectively 15 and 17, when compared to the residual ratio method. However, we can question the reliability of these rank estimates. We see that for CV the differences in PRESS are relatively small in the domain $[15, 25]$. Furthermore, these differences in PRESS are also influenced by the dataset size and hyperparameters such as the holdout size and the number of iterations. Consequently, the small differences in PRESS show that rank estimation based on CV does not always provide a conclusive answer. At threshold 0 in figure 5-15 we see the results of rank estimation based on explained variance. These rank estimates for variance percentages commonly chosen in the literature, respectively $[90, 95, 97.5]$ are significantly higher than the ones obtained by the residual ratio or CV. As discussed in chapter 1, the differing rank estimates are an effect of different motivations and associated goals of these rank estimation methods. CV considers intensities that have predictive relevance for other intensities relevant and the rest noise. The percentage of explained variance, on the other hand, focuses on capturing the majority of the variance in a dataset and as a result, also captures intensity patterns without this predictive relevance.

5-3-4 Residual-fraction and threshold-shifted rank estimation

Section 5-3 shows a large discrepancy in rank estimates between the residual-fraction and threshold-shifted version. Furthermore, we also observe a large difference in the intensity-unaware explained variance and CV methods. These differences in rank estimates do not mean one of the two methods is wrong. Residual-fraction and threshold-shifted rank estimation represent different motivations behind rank estimation and dimensionality reduction. The context and associated goal of the dimensionality reduction generally determine which motivation is better suited. We believe that the maximization of the residual-fraction is closer related to minimization of the capturing of noise. We expect a certain amount of randomness or noise in the intensities below the threshold if these measurements are unreliable for any biological conclusions. At the same time, all peaks regardless of their intensity with respect to the threshold contain some noise. Consequently, this suggests we capture an increasing amount of noise at the ranks for which we capture an increasing amount of the below threshold intensities. In other words, we argue that the residual-fraction is a measure influenced by the noise, under the premise that the dataset contains a certain amount of incoherence or non-structure in the below-threshold intensities. For this reason, we consider that the motivation behind residual-fraction rank estimation shows similarities to CV, which could explain the similar rank estimates.

The threshold-shifted rank estimation based on the percentage of explained variance originating from threshold-exceeding intensities in TAPCA has a different motivation altogether.

In the case of threshold-shifted rank estimation, we truncate at a rank to capture the majority of the above-threshold variance contribution. For this reason, we obtain compressed measurements maximally explaining the threshold-exceeding variance in the original measurements. Consequently, the motivation is to describe the majority of the variance in the threshold-exceeding intensities instead of minimizing the amount of captured noise. Due to the differences in objective, threshold-shifted rank estimation requires a significantly larger number of components than residual-fraction and cross-validation-based rank estimation.

We propose that the preferred method for rank estimation depends on the application in mind and the motivation behind the dimensionality reduction. First of all, residual-fraction rank estimation is a heuristic and might not generalize to all IMS datasets. Second, if the context is to apply clustering, classification, or doing manual analysis, post-reduction only, using a minimal amount of components maximally describing threshold-exceeding intensities, the residual-fraction approach might be preferred. Furthermore, residual-fraction rank estimation is advantageous as it provides principal components based on the original unmodified spectra, whereas threshold-shifted rank estimation does modify the spectra. For other objectives, such as reducing storage or computational load, the threshold-shifted rank estimation might be preferred.

Chapter 6

Conclusions and Recommendations

6-1 Conclusions

In this section, we summarize the conclusions found during the thesis. We start with the main conclusions and then zoom in on conclusions per sub-topic.

We described two methods to estimate the rank linear dimensionality reduction with PCA taking the reliability of measurement intensities in the considered IMS dataset into account. In line with the issues and challenges of dimensionality reduction for IMS, we can conclude the following.

- A rank for which all threshold-exceeding intensities are perfectly captured, for which PCA does not also capture a large part of the irrelevant and more unreliable below-threshold intensity region, does not seem to discernible. As a consequence, intensity-aware rank estimation requires a measure for sufficient capture of the information above the threshold, given that we cannot achieve perfect above-threshold capture with for a rank leading to substantial reduction.
- We have suggested a heuristic residual-fraction rank estimation for PCA as a potential measure of optimality for above-threshold intensity capture. Residual-fraction rank estimation is based on the maximization of the above-threshold capturing relative to the below-threshold capturing based on the residuals. Residual-fraction rank estimation assumes that a dataset exhibiting little structure or coherence in below-threshold intensities will be expressed in the residuals.
- We have proposed a threshold-shifted rank estimation method based on the explained variance in TAPCA, as an approach to estimate rank while explicitly de-emphasizing any below-threshold intensities.
- We have shown that threshold-shifted rank estimation enables imposing an explicit relationship between the intensity threshold in the measurement space and rank in the abstract subspace obtained by dimensionality reduction, based on the explained variance above the threshold.

The following sections summarize the conclusions related to the proposed rank estimation methods and TAPCA.

6-1-1 Residual-fraction rank estimation

- We have shown that application of residual-fraction rank estimation is capable of finding the correct rank for the synthetic dataset in both the noise and no noise case. We do note that the synthetic dataset exhibits a clean cut between the reliable threshold-exceeding and unreliable below-threshold peaks, which is not realistic for a real-world case study. Nonetheless, the application of PCA to the Imaging Mass Spectrometry (IMS) dataset showed similar behavior as with the synthetic experiment.

- We have shown that residual-fraction rank estimation obtains similar rank estimates to cross-validation, but significantly lower ranks than explained variance based rank estimation. We have suggested a potential cause might be the possibly different motivation behind these methods.

6-1-2 Threshold-shifted rank estimation

- We have shown that threshold-shifted rank estimation based on explained variance in TAPCA could result in a significant decrease in rank when compared to the threshold-unaware rank estimation based on explained variance.

- We have shown that higher thresholds for TAPCA result in lower rank estimates for both the synthetic dataset and IMS dataset. However, the no-noise synthetic dataset shows no reduction after the TAPCA threshold has superseded the synthetic dataset threshold. These effects are not so clearly visible in the real or the synthetic noise case.

- We have shown, in line with the expectations, that larger datasets containing more mass-bins and peak picked with higher thresholds show a higher reduction in rank. A dataset with a lower peak picking threshold contains more relatively low-intensity mass-bins and as such the threshold-aware rank estimation has bigger impact.

6-1-3 Pattern capturing of TAPCA

- We have shown that application of TAPCA to an IMS dataset can lead to sparser principal component spectral loadings and cleaner delineation of spatial scores, depending on the choice of the threshold and shifting transformation. Furthermore, the RMS spectrum contained the per-mass-bin RMS intensity has become sparser, which we believe is related to the overall sparser spectral loadings.

- We have shown that TAPCA is able to especially capture threshold-exceeding patterns and that it discards below-threshold patterns. We have also demonstrated that TAPCA potentially fails to fully capture patterns partially above- and partially below-threshold regardless of the projection method. Furthermore, the application of the linear-shift transformation in LTAPCA does not fully mend this shortcoming.

- TAPCA seems to attenuate the influence of particular mass-bins in line with the magnitude of the threshold-exceeding intensities present in these bins, depending on the threshold-choice.
- We have shown that projection of the original deviations from the mean in TAPCA allows a low dimensional representation on the scale of the original data based on the basis obtained with TAPCA. However, application to the IMS dataset shows that this projection also leads to predominantly increased residuals for the lower ranks of interest, and a reduction in residuals for high-intensities at the higher ranks independent of the imposed threshold. Consequently, in the context of the ion intensity threshold, the projection of deviations from the mean after transformation seems preferred.
- We have shown that projection of the transformed deviations from the mean in the IMS dataset on the basis obtained by both CTAPCA and LTAPCA leads to respectively no and reduced capturing of the below-threshold intensities in the lower-dimensional representation. We expected that the difference with LTAPCA on the capturing of the below-threshold intensities would have been more significant.

6-1-4 Relevance to dimensionality reduction in IMS

In the context of IMS, the residual-fraction and threshold-shifted rank estimation have made an initial connection between rank as defined in the lower dimensional abstract subspace constructed by PCA and the ion intensity threshold as defined in the spectra in the original measurement space.

This connection allows the mass spectrometrist, the mass spectrometer operator, to take into account the LOQ for PCA-based dimensionality reduction resulting in a low-dimensional representation in line with the instrument properties. Furthermore, intensity-aware rank estimation achieves a higher compression or lower rank estimates than intensity-unaware rank estimation. For the mass spectrometrist, the higher compression opens the door to more advanced experiments producing more data with similar computational resources, such as 3D IMS [11] and hyphenation with ion mobility detection [12].

For the data-scientist, this connection allows heeding of the LOQ during dimensionality reduction. The data-scientist can construct a lower dimensional representation maximally reflecting reliable threshold-exceeding intensities and discarding the low reliability below-threshold intensities, and then use this representation for subsequent analysis.

6-2 Recommendations for future work

6-2-1 Alternatives for intensity-aware rank estimation for PCA

We have briefly investigated the use of WCPCA for intensity-aware rank estimation and recommend this approach as a possibility for further research. WCPCA is potentially a more hybrid solution in between TAPCA and PCA, in which the intensities in the measurement-space seem unaffected, yet allows to emphasize threshold-exceeding intensities in the abstract subspace. We foresee a challenge in the selection of an appropriate weighting scheme that enables consistent emphasis of threshold-exceeding intensities.

We recommend investigation of an alternative to residual-fraction based rank estimation as a heuristic for sufficient capturing threshold-exceeding intensities. The residual-fraction depends on the underlying structure of the dataset and consequently we expect it will not be applicable to every dataset. For this reason, this approach towards intensity-aware rank estimation would benefit from another measure for sufficient capturing of the threshold-exceeding intensities in the low-dimensional abstract subspace.

6-2-2 Extension for intensity-aware rank estimation to NMF

We recommend more research into the extension of threshold-shifted rank estimation to NMF. De-emphasizing below-threshold intensities can be achieved in a similar manner as with TAPCA by transformation of the measurements in the measurement-space. We foresee a challenge in the subsequent rank estimation for NMF, as limited standardized rank estimation methods nor a relevance parameter such as the explained variance are available in the literature.

In the context of residual-fraction based rank estimation, we do not recommend more research into extension to NMF. Preliminary results showed patterns in the residuals dependent on the intensity, similar to PCA, did not occur for NMF. We attributed this to the less-overfitting property of NMF, and consequently, expect this method to not work as well for NMF.

Finally, we see in WNMF for NMF a similar solution as in WCPCA for PCA. WNMF allows the intensities in the measurement-space to be unaffected, yet allows to emphasize threshold-exceeding intensities in the abstract subspace. We foresee the consecutive rank estimation WNMF will be challenging, due to the limited availability of rank estimation in the literature.

6-2-3 Improvements for TAPCA

We recommend more research into mitigation of the issues arising for partially below- partially above-threshold patterns. A potential solution could be putting more emphasis on specific cases of interest. One approach could be more complex shifting transformations, but we expect that this problem will not to be solved by a one-size-fits-all solution. An alternative would be non-uniform shifting transformations. Tailor-made shifting transformations, focused on a particular spatial or spectral pattern, could allow specific emphasis on these patterns. The combination of the shifting transformations with a weighting scheme, such as WCPCA, to emphasize these patterns could also be a possibility.

Appendix A

Rank Estimation Covariance Weighted PCA

We have tested rank estimation with Weighted Covariance Principal Component Analysis (WCPCA) [55] by using a element-wise binary weighting scheme $w_{nm} \in \{0, 1\}$ as described in section 4-3-1. We focus on the effect of this weighting on the rank estimate and neglect the impact on the extracted principal components as a result of this possibly overly harsh weighting function. Figure A-1 shows the effect on the rank estimation by displaying the explained variance in relation to the threshold used for this binary weighting scheme. In this figure, we see a sharp reduction initially for the first nonzero threshold, which corresponds to the case for which we neglect all zero entries. However, after this point, the effect of the increasing threshold is limited for the intensity range of interest $[0, 3000]$. We attributed this to the normalization of the mass-bins weight done in WCPCA, such that every bin has equal weight. As a result, the few threshold-exceeding intensities in generally below-threshold mass-bins weigh relatively heavy compared to intensities in mass-bins containing many threshold-exceeding intensities.

However, disabling this normalization caused new effects as shown in figure A-2. In this case, we see increasing percentages of explained variance for increasing thresholds, which is the adverse effect of what we expected and hoped to achieve. At the point of writing it is unclear what causes this effect and more research would be required. We have run some preliminary experiments with simple affine $w_{nm}(x) = \min(ax + b, 1)$ and sigmoid weighting $w_{nm}(x) = \frac{1}{1+\exp^{-ax+\tau}}$ schemes, but these do not seem to make consequent differences in the results.

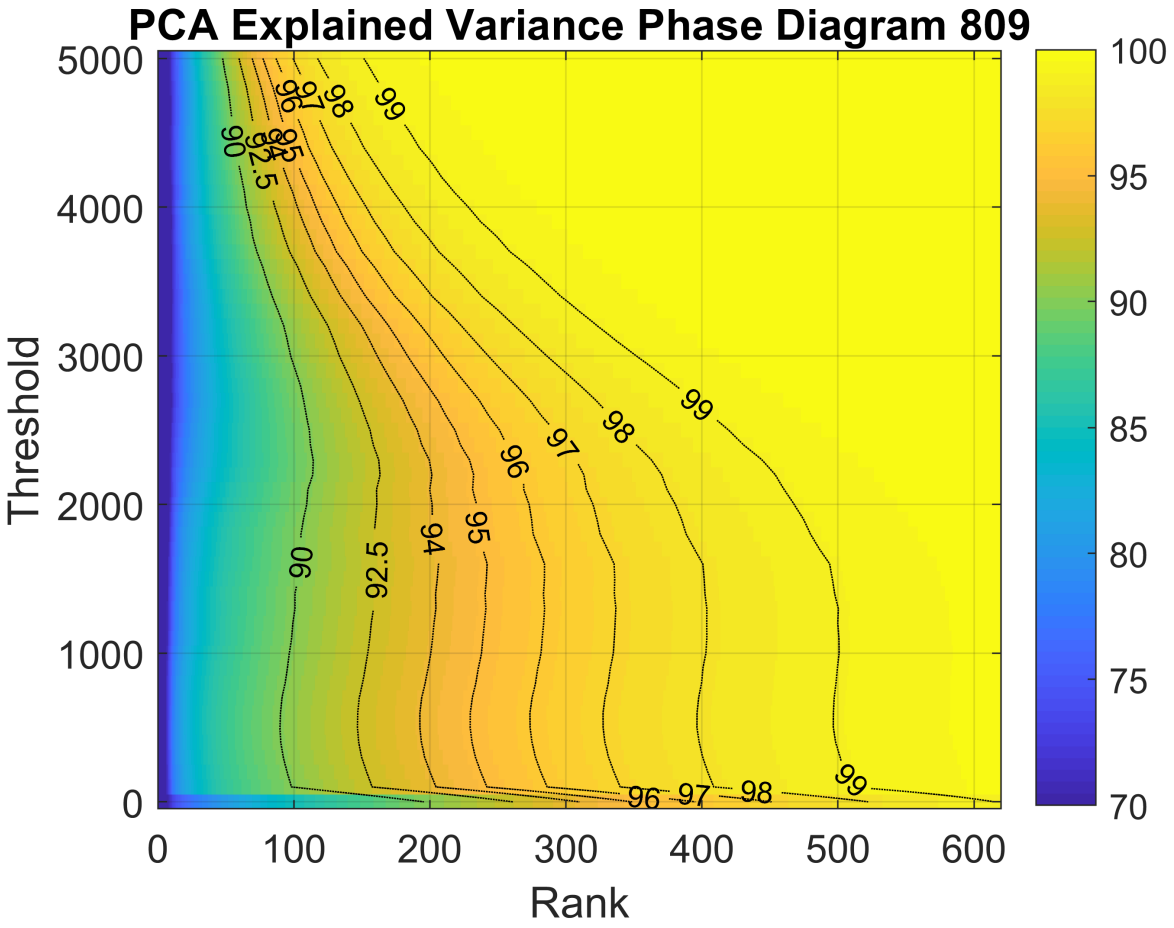


Figure A-1: The percentage explained variance in relation to the used threshold for the binary weighting scheme of WCPCA for the IMS dataset with 809 mass-bins. In this figure, the zero threshold corresponds to traditional PCA

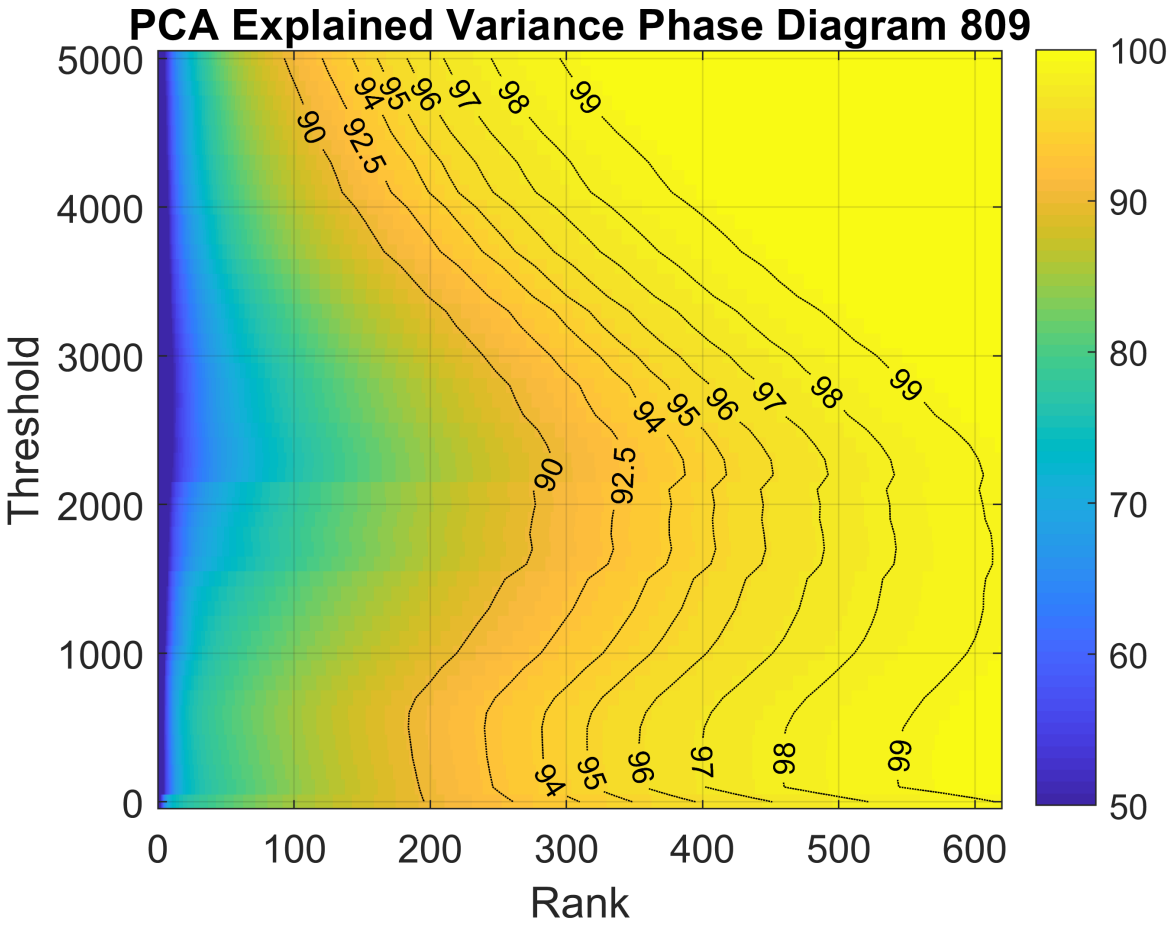


Figure A-2: The percentage explained variance in relation to the used threshold for the binary weighting scheme of WCPCA without normalization of the weights for the IMS dataset with 809 mass-bins. In this figure, the zero threshold corresponds to traditional PCA

Bibliography

[1] 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," *Nature Medicine*, vol. 7, no. 4, pp. 493–496, 2001.

[2] R. Van De Plas, B. De Moor, and E. Waelkens, "Imaging mass spectrometry based exploration of biochemical tissue composition using peak intensity weighted PCA," in *2007 IEEE/NIH Life Science Systems and Applications Workshop, LISA*, pp. 209–212, IEEE, nov 2008.

[3] L. A. McDonnell and R. M. A. Heeren, "Imaging Mass Spectrometry," *Mass Spectrom. Rev.*, vol. 26, pp. 606–643, 2007.

[4] J. Hanrieder, N. T. Phan, M. E. Kurczy, and A. G. Ewing, "Imaging mass spectrometry in neuroscience," 2013.

[5] P. Inglese, J. S. McKenzie, A. Mroz, J. Kinross, K. Veselkov, E. Holmes, Z. Takats, J. K. Nicholson, and R. C. Glen, "Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer," *Chemical Science*, vol. 8, no. 5, pp. 3500–3511, 2017.

[6] A. M. Race, R. T. Steven, A. D. Palmer, I. B. Styles, and J. Bunch, "Memory efficient principal component analysis for the dimensionality reduction of large mass spectrometry imaging data sets," *Analytical Chemistry*, vol. 85, no. 6, pp. 3071–3078, 2013.

[7] S. A. Thomas, A. M. Race, R. T. Steven, I. S. Gilmore, and J. Bunch, "Dimensionality reduction of mass spectrometry imaging data using autoencoders," in *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*, pp. 1–7, IEEE, 2017.

[8] A. Palmer, E. Molecular, and J. Bunch, "The Use of Random Projections for the Analysis of Mass Spectrometry Imaging Data," *Journal of the American Society for Mass Spectrometry*, vol. 26, no. February 2015, pp. 315–322, 2014.

[9] R. E. Bellman, "Adaptive control processes: A guided tour," *Princeton University Press*, 1961.

[10] Y. Zhang and X. Liu, "Machine learning techniques for mass spectrometry imaging data analysis and applications," *Bioanalysis*, vol. 10, no. 8, pp. 519–522, 2018.

[11] E. H. Seeley and R. M. Caprioli, "3D Imaging by Mass Spectrometry: A New Frontier," *Analytical Chemistry*, vol. 84, no. 5, pp. 2105–2110, 2012.

[12] R. M. A. Heeren, D. F. Smith, J. Stauber, B. Kükrer-Kaletas, and L. MacAleese, "Imaging Mass Spectrometry: Hype or Hope?," *Journal of the American Society for Mass Spectrometry*, vol. 20, no. 6, pp. 1006–1014, 2009.

[13] D. Trede, J. H. Kobarg, J. Oetjen, H. Thiele, P. Maass, and T. Alexandrov, "On the importance of mathematical methods for analysis of MALDI-imaging mass spectrometry data.," *Journal of integrative bioinformatics*, vol. 9, no. 1, p. 189, 2012.

[14] 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*, pp. 1–6, IEEE, oct 2008.

[15] E. R. Muir, I. J. Ndiour, N. A. Le Goasduff, R. A. Moffitt, Y. Liu, M. C. Sullards, A. H. Merrill, Y. Chen, and M. D. Wang, "Multivariate analysis of imaging mass spectrometry data," in *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE*, vol. 20, pp. 472–479, 2007.

[16] Y. Li and A. Ngom, "The non-negative matrix factorization toolbox for biological data mining Source Code for Biology and Medicine The non-negative matrix factorization toolbox for biological data mining," *Source Code for Biology andMedicine*, vol. 8, no. 10, 2011.

[17] R. Dubroca, C. Junor, and A. Souloumiac, "Weighted Nmf for High-Resolution Mass Spectrometry Analysis," *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, no. Eusipco, pp. 1806–1810, 2012.

[18] G. McCombie, D. Staab, M. Stoeckli, and R. Knochenmuss, "Spatial and spectral correlations in MALDI mass spectrometry images by clustering and multivariate analysis," *Analytical Chemistry*, vol. 77, no. 19, pp. 6118–6124, 2005.

[19] R. Van De Plas, F. Ojeda, M. Dewil, L. Van, D. Bosch, B. De Moor, and E. Waelkens, "Prospective Exploration of Biochemical Tissue Composition Via Imaging Mass Spectrometry Guided By Principal Component Analysis," *Pacific Symposium on Biocomputing*, vol. 12, pp. 458–469, 2007.

[20] L. A. Klerk, A. Broersen, I. W. Fletcher, R. van Liere, and R. M. Heeren, "Extended data analysis strategies for high resolution imaging MS: New methods to deal with extremely large image hyperspectral datasets," *International Journal of Mass Spectrometry*, vol. 260, no. 2-3, pp. 222–236, 2007.

[21] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[22] M. Hanselmann, M. Kirchner, B. Y. Renard, E. R. Amstalden, K. Glunde, R. M. A. Heeren, and F. A. Hamprecht, "Concise representation of mass spectrometry images by probabilistic latent semantic analysis," *Analytical Chemistry*, vol. 80, no. 24, pp. 9649–9658, 2008.

[23] Y. C. Harn, M. J. Powers, E. A. Shank, and V. Jojic, "Deconvolving molecular signatures of interactions between microbial colonies," *Bioinformatics*, vol. 31, no. 12, pp. i142–i150, 2015.

[24] N. Gillis, "The Why and How of Nonnegative Matrix Factorization," in *Regularization, Optimization, Kernels, and Support Vector Machines, J.A.K. Suykens, M. Signoretto and A. Argyriou (eds), Chapman & Hall/CRC, Machine Learning and Pattern Recognition Series,*, pp. 257–291, 2014.

[25] C.-Y. Liou and K.-D. Yang, "Unsupervised classification of remote sensing imagery with non-negative matrix factorization," in *Proc. ICONIP*, (Taipei, Taiwan), pp. 280–285, 2005.

[26] S. Wold, "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Model," *Technometrics*, vol. 20, no. 4, pp. 397–405, 1978.

[27] A. B. Owen and P. O. Perry, "Bi-cross-validation of the SVD and the nonnegative matrix factorization," *Annals of Applied Statistics*, vol. 3, no. 2, pp. 564–594, 2009.

[28] R. Bro, K. Kjeldahl, A. K. Smilde, and H. A. L. Kiers, "Cross-validation of component models: A critical look at current methods," *Analytical and Bioanalytical Chemistry*, vol. 390, no. 5, pp. 1241–1251, 2008.

[29] P. R. Peres-Neto, D. A. Jackson, and K. M. Somers, "How many principal components? stopping rules for determining the number of non-trivial axes revisited," *Computational Statistics and Data Analysis*, vol. 49, no. 4, pp. 974–997, 2005.

[30] A. Fisher, B. Caffo, B. Schwartz, and V. Zipunnikov, "Fast, Exact Bootstrap Principal Component Analysis for p>1 million," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 846–860, 2014.

[31] J. J. Daudin, C. Duby, and P. Trécourt, "Pca stability studied by the bootstrap and the infinitesimal jackknife method," *Statistics*, vol. 20, pp. 255–270, jan 1989.

[32] T. P. Minka, "Automatic choice of dimensionality for PCA," *M.I.T. Media Laboratory Perceptual Computing Section*, no. 514, pp. 1–16, 2000.

[33] A. Asensio Ramos, "The Minimum Description Length Principle and Model Selection in Spectropolarimetry," *The Astrophysical Journal*, vol. 646, no. 2, pp. 1445–1451, 2006.

[34] I. Ramírez and M. Tepper, "Bi-clustering via MDL-Based Matrix Factorization," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (J. Ruiz-Shulcloper and G. Sanniti di Baja, eds.), (Berlin, Heidelberg), pp. 230–237, Springer Berlin Heidelberg, 2013.

[35] M. O. Ulfarsson and V. Solo, "Tuning parameter selection for nonnegative matrix factorization," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, (Vancouver), pp. 6590–6594, IEEE, 2013.

[36] M. Schmidt, O. Winther, and L. K. Hansen, "Bayesian non-negative matrix factorization," *International Conference on Independent Component Analysis and Signal Separation*, pp. 540—-547, 2009.

[37] V. Y. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization," in *Signal Processing with Adaptive Sparse Structured Representations*, (Saint Malo, United Kingdom.), 2009.

[38] V. Y. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization with the (β)-divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1592–1605, jul 2013.

[39] Z. Yang, Z. Zhu, and E. Oja, "Automatic rank determination in projective nonnegative matrix factorization," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6365 LNCS, pp. 514–521, 2010.

[40] I. Jolliffe, *Principal component analysis*. Springer Verlag, 2 ed., 2002.

[41] S. Ubaru and Y. Saad, "Fast methods for estimating the Numerical rank of large matrices," *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, pp. 468–477, 2016.

[42] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[43] D. A. Armbruster and T. Pry, "Limit of Blank, Limit of Detection and Limit of Quantitation," *The Clinical biochemist. Reviews*, vol. 29, no. Suppl 1, pp. 49–52, 2008.

[44] N. Verbeeck, J. M. Spraggins, M. J. Murphy, H. dong Wang, A. Y. Deutch, R. M. Caprioli, and R. Van de Plas, "Connecting imaging mass spectrometry and magnetic resonance imaging-based anatomical atlases for automated anatomical interpretation and differential analysis," *Biochimica et Biophysica Acta - Proteins and Proteomics*, vol. 1865, no. 7, pp. 967–977, 2017.

[45] J. M. Spraggins, D. G. Rizzo, J. L. Moore, M. J. Noto, E. P. Skaar, and R. M. Caprioli, "Next-generation technologies for spatial proteomics: Integrating ultra-high speed MALDI-TOF and high mass resolution MALDI FTICR imaging mass spectrometry for protein analysis," *Proteomics*, 2016.

[46] J. M. Spraggins, D. G. Rizzo, J. L. Moore, K. L. Rose, N. D. Hammer, E. P. Skaar, and R. M. Caprioli, "MALDI FTICR IMS of intact proteins: Using mass accuracy to link protein images with proteomics data," *Journal of the American Society for Mass Spectrometry*, 2015.

[47] J. Yang and R. M. Caprioli, "Matrix sublimation/recrystallization for imaging proteins by mass spectrometry at high spatial resolution," *Analytical Chemistry*, 2011.

[48] A. Cassese, S. R. Ellis, N. Ogrinc Potočnik, E. Burgermeister, M. Ebert, A. Walch, A. M. Van Den Maagdenberg, L. A. McDonnell, R. M. Heeren, and B. Balluff, "Spatial Autocorrelation in Mass Spectrometry Imaging," *Analytical Chemistry*, vol. 88, no. 11, pp. 5871–5878, 2016.

[49] M. R. Keenan and P. G. Kotula, "Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images," *Surface and Interface Analysis*, vol. 36, no. 3, pp. 203–212, 2004.

[50] M. Udell, *Generalized Low Rank Models*. Dissertation, Stanford University, 2015.

[51] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[52] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, no. 1, pp. 556–562, 2001.

[53] J. Kim, Y. He, and H. Park, *Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework*, vol. 58. 2014.

[54] B. Rén, L. Pueyo, G. B. Zhu, J. Debes, and G. Duchêne, "Non-negative Matrix Factorization: Robust Extraction of Extended Structures," 2017.

[55] L. Delchambre, "Weighted principal component analysis: A weighted covariance eigendecomposition approach," *Monthly Notices of the Royal Astronomical Society*, vol. 446, no. 4, pp. 3545–3555, 2014.

[56] N. Ho, P. V. Dooren, and V. Blondel, "Weighted nonnegative matrix factorization and face feature extraction," *Image and Vision Computing*, no. March 2008, pp. 1–17, 2008.

[57] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: A resampling based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, vol. 52, no. i, pp. 91–118, 2003.

Glossary

List of Acronyms

IMS	Imaging Mass Spectrometry
MALDI	Matrix Assisted Laser Desorption Ionization
DESI	Desorption Electrospray Ionization
SIMS	Secondary Ion Mass Spectrometry Imaging
HSI	Hyperspectral Imaging
LDR	Linear Dimensionality Reduction
PCA	Principal Component Analysis
PC	Principal Component
SVD	Singular Value Decomposition
NMF	Nonnegative Matrix Factorization
TAPCA	Threshold-Aware Principal Component Analysis
CTAPCA	Clipping Threshold-Aware Principal Component Analysis
LTAPCA	Linear Threshold-Aware Principal Component Analysis
QTAPCA	Quadratic Threshold-Aware Principal Component Analysis
PIWPCA	Peak Intensity Weighted Principal Component Analysis
WCPCA	Weighted Covariance Principal Component Analysis
WNMF	Weighted Nonnegative Matrix Factorization
RMS	Root Mean Squared
RSSQ	Residual Sum of Squares

RMSR Root Mean Squared Residual

MAR Median Absolute Residual

SNR Signal-to-Noise Ratio

CV Cross-Validation

PLSA Probabilistic Latent Semantic Analysis

LDA Linear Discriminant Analysis

NN-PARAFAC Nonnegative Parallel Factor Analysis

t-SNE t-Distributed Stochastic Neighbors Embedding

AIC Akaike Information Criterion

MLE Maximum Likelihood Estimation

MDL Minimum Description Length

SURE Stein's Unbiased Estimator

ARD Automatic Relevance Determination

PRESS Predicted Residual Sum of Squares

LOB Level of Blank

LOD Level of Detection

LOQ Level of Quantitation

NNLS Nonnegative Least Squares

ALS Alternating Least Squares

HALS Hierarchical Alternating Least Squares

BCD Block Coordinate Descent

TIC Total Ion Current