UNIVERSITY OF WARSAW
FACULTY OF MATHEMATICS, INFORMATICS AND MECHANICS

HASSELT UNIVERSITY
DATA SCIENCE INSTITUTE

DOCTORAL DISSERTATION

Algorithms for computational mass spectrometry based on the optimal transport theory

Author:
Michał Aleksander Ciach

Supervisors:
prof. dr hab. Anna Gambin
prof. Dirk Valkenborg

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science

May 30, 2022

Supervisors' statement

Hereby, I confirm that the presented thesis was prepared under my supervision and that it fulfills the requirements for the degree of PhD in Computer Science.

Supervisor 1: prof. dr hab. Anna Gambin, Faculty of Mathematics, Informatics and Mechanics, University of Warsaw

---

Signature and date

Supervisor 2: prof. Dirk Valkenborg, Data Science Institute, Hasselt University

---

Signature and date

Author's statement

Hereby, I declare that the presented thesis was prepared by me and none of its contents was obtained by means that are against the law.
The thesis has never before been a subject of any procedure of obtaining an academic degree.
Moreover, I declare that the present version of the thesis is identical to the attached electronic version.

mgr Michał Aleksander Ciach

Signature and date

Abstract

Algorithms for computational mass spectrometry based on the optimal transport theory

In this dissertation, we present a novel approach to developing computational methods for the analysis of mass and nuclear magnetic resonance spectra. We start with the discussion of the state-of-the-art approaches, illustrated by a study of nuclear magnetic resonance spectra of St John's wort extracts. We describe the need for mathematical theory for comparison of spectra of different molecules and with different resolutions. We then describe how to use the notion of optimal transport of signal and the Wasserstein distance to develop algorithms for fitting a linear combination of reference spectra to a spectrum of a mixture of chemical compounds. The algorithm makes it possible to accurately estimate the amounts of compounds with overlapping spectra. We finish the dissertation with an application of our methods to the problem of segmentation of mass spectrometric images, where we show that they allow for obtaining biologically accurate and meaningful results when other common approaches fail. Our results are applicable for various types of spectrometry and spectroscopy, including NMR spectroscopy and mass spectrometry. The algorithms developed as a part of this thesis are available in an open-source Python 3 package `masserstein` available at https://github.com/mciach/masserstein.

Streszczenie

Algorytmy obliczeniowej spektrometrii mas oparte na teorii optymalnego transportu

W niniejszej rozprawie przedstawiamy nowe podejście do projektowania metod obliczeniowych do analizy widm masowych oraz widm magnetycznego rezonansu jądrowego (NMR). Rozprawę rozpoczynamy omówieniem obecnie stosowanych metod na przykładzie analizy widm NMR wyciągów z dziurawca. Uzasadniamy potrzebę opracowania aparatu matematycznego do porównywania widm różnych cząsteczek oraz o różnej rozdzielczości. Następnie opisujemy, w jaki sposób wykorzystać koncepcję optymalnego transportu sygnału i odległości Wassersteina do opracowania algorytmu dopasowującego kombinację liniową widm referencyjnych do widma mieszaniny związków chemicznych. Metoda ta pozwala na dokładną estymację zawartości związków o nakładających się widmach. Rozprawę kończymy zastosowaniem opracowanych metod do analizy obrazów spektrometrycznych, gdzie pokazujemy, że pozwalają one na otrzymanie biologicznie znaczących wyników nawet gdy inne metody zawodzą. Podejście do analizy widm zaprezentowane w niniejszej pracy ma zastosowanie do różnych typów spektrometrii i spektroskopii, wliczając w to spektroskopię magnetycznego rezonansu jądrowego oraz spektrometrię mas. Algorytmy opracowane w ramach ninejszej pracy zostały zaimplementowane w pakiecie `masserstein` języka programowania Python 3, dostępnym pod adresem https://github.com/mciach/masserstein.

Keywords

Wasserstein Regression, Mass Spectrometry, Optimal Transport, Wasserstein Distance, Computational Methods, Mass Spectrometric Image, Image Segmentation, Regression of Mass Spectra, NMR spectroscopy

Thesis domain (Socrates-Erasmus subject area codes)

11.3 Informatyka

Subject classification

I.6. Simulation and Modelling
J.3. Life and Medical Sciences

Tytuł pracy w języku polskim

Algorytmy obliczeniowej spektrometrii mas oparte na teorii optymalnego transportu

Acknowledgements

I would like to express my utmost gratitute to my supervisors, prof. dr hab. Anna Gambin and prof. Dirk Valkenborg, who have guided me through the ups and downs of PhD studies. You have offered me tremendous support, both scientific and personal. This dissertation would have never happened without you.

I would also like to thank dr (sc)hab[1]. Błażej Miasojedow, both as an unofficial scientific supervisor and mentor, as well as an official friend.

Furthermore, I would like to thank all my collaborators and colleagues: Grzegorz Skoraczyński, Michał Startek, Maciej Zielenkiewicz, Wanda Niemyska, Grzegorz Bokota, Barbara Poszewiecka, Szymon Majewski, Krzysztof Gogolewski, Aneta Manda-Handzlik, and Piotr Radziński. I had great fun working and memeing with you.

I would also like to express my gratitude towards my supervisors and collaborators from my other projects, not related to this disseration, who have expanded my academic horizons and abilities: prof. Paweł Górecki and dr Anna Muszewska.

Naturally, this work would not be possible without financial support. Let me thank the institutions that have provided me with the following funding during the research and preparation of this dissertation:

- The NCN OPUS grant 2018/29/B/ST6/00681 titled "Algorithmic challenges of mass spectrometry"
- The NCN OPUS grant 2021/41/B/ST6/03526 titled "Optimal-transport based algorithms for Mass Spectrometry and NMR"

I would also like to acknowledge the sources from which I have recieved financial support while working on my numerous side projects:

- The NCN OPUS 2015/19/B/ST6/00726 "Computational Genomics: Problems, Algorithms and Models"
- The NCN OPUS 2019/33/B/ST6/00737 "Biologically Meaningful Inference of Phylogenetic Networks"
- The Warsaw University — Medical University of Warsaw bilateral micro-grant 1WW/NUW1/18 "Modelowanie matematyczne odpowiedzi granulocyta na patogen – walidacja w oparciu o układ neutrofil – E. coli".

Last, but not least, I would like to thank my parents, Alina Ciach and Wojciech Góźdź. You have instilled the interest in science in me and taught me how to make good life choices, even though I may often seem too stubborn to take your advice.

[1] An unofficial academic title, where (sc) stands for (science)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Mass spectrometry is a laboratory technique that measures the mass-to-charge ratio of ionized chemical molecules [1]. Informally speaking, it's a technique of "weighing" individual molecules. The mass spectrometer, i.e. the instrument used to carry out such a measurement, separates ions in an electromagnetic field according to the Lorentz force:

$$F = z(E + v \times B),$$

where z is the charge of the ion, E is the electric field vector, v is the ion velocity vector, and B is the magnetic field vector. In conjunction with Newton's second law of motion, which says that the acceleration a of an object is equal to the value of the force acting on it divided by the object's mass, we get the relationship:

$$a = F/m = \frac{z}{m}(E + v \times B),$$

where the value of z/m depends on the properties of the analyzed ion, and the values of E and B depend on the instrument's setting. While the acceleration is proportional to the ratio z/m, in practice, for the convenience of data analysis, values of m/z are used. Since in many cases the ions measured in spectrometers are singly charged ($z = \pm 1$), in most of this Dissertation we write about m/z values as ion masses, meaning the mass divided by a unit charge. Because of this, we will express the m/z values in Dalton units, i.e. the units of atomic mass defined as $1/12$ of the mass of an unbound, neutral carbon atom. By extension, in mass spectrometry, the units of the m/z axis are referred to as Daltons even if the spectra contain multiply-charged molecules.

The result of a single measurement is not just the mass of a single ion, but an entire *mass spectrum*, i.e. a graph showing the dependence of the intensity of the measured signal on the value of m/z. The intensity is usually proportional to the numbers of ions with a given mass, but without a simple relation that could be used to compute one value using the other in general. Therefore, the units of signal intensity do not have a physical meaning, and usually the relationships between intensities at different mass values are analyzed. Accordingly, when plotting a mass spectrum, the y axis is often not labeled.

Mass spectrometry, in combination with other techniques, provides large amounts of information on the molecules under study. In the case of mixtures with a known chemical composition, it can be used to estimate the relative content of each component of the mixture (as long as these components can be ionized). In the case of unknown compounds, fragmenting them and then measuring the masses of the fragments often allows them to be identified. Consequently, mass spectrometry is used in many different fields of science, from archaeology to medicine, and in many branches of industry, from synthetic polymers to pharmacy.

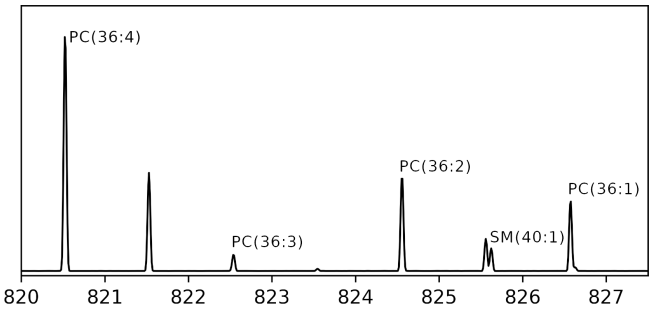

FIGURE 1.1: An example mass spectrum of a lipid mixture. The signals are labeled with the common names of the corresponding compounds. The signal with a mass of 821.6 Da belongs to the so-called isotopic envelope of the lipid PC(36:4). The lipids PC(36:4) and PC(36:3) have overlapping isotopic envelopes.

While each ion has a unique m/z value, suggesting that a mass spectrum should be a discrete function. This, however, is not the case. In practice, during a spectrometric measurement, each ion has its own initial velocity and position, which makes the measurement for each ion slightly different. Because of these measurement uncertainties, actual mass spectra are continuous functions - for each type of ion we get a certain distribution of signal intensity around its true mass. Such spectra are said to be in *profile mode*. An exemplary profile mode mass spectrum of a lipid mixture is shown in Fig. 1.1.

A spectrum in a profile mode can be converted to a discrete form by a procedure called *centroiding*, which consists of an identification of local signal maxima and a numerical integration of the signals around the maxima. This procedure results in discrete signals in locations corresponding to approximate m/z values of the analyzed ions. Such spectra are said to be in *centroid mode*. By extension, we apply this term to computationally simulated, theoretical mass spectra of ions, which are discrete functions as well.

Due to the presence of naturally occurring isotopes, each ion is observed in the spectrum as a series of signals collectively referred to as an *isotopic envelope*. The signal corresponding to the most common isotopes is called the *monoisotopic peak* of the ion. In the context of biological and most of the organic molecules, on which this Dissertation is focused, the monoisotopic peak is generally the signal with the lightest mass in a given envelope. The spectrum in Fig. 1.1 is composed of five such isotope envelopes, and the monoisotopic peak of each of them is labeled with the name of the corresponding lipid.

Envelopes of ions with similar masses often overlap, causing some of their signals to merge. Therefore, each signal in the spectrum is potentially a mixture of signals coming from different ions. The fact that a single ion corresponds to many signals and a single signal to many ions makes many aspects of the analysis of spectrometric data highly nontrivial. In particular, determining the relative abundances of ions in a spectrum requires the separation of overlapping signals. However, this fact is often ignored, and the integrals of the monoisotopic peaks are used as simple measures of the abundances. While this carries the risk of obtaining erroneous results, especially in the case of complex mixtures, the currently existing alternative methods for this task have not been widely adopted. This is caused, among others, by computational difficulties and the lack of appropriate mathematical tools, making those alternative approaches difficult to use and sometimes inaccurate.

In this Dissertation, we present a new approach to computational mass spectrometry, based on the mathematical theory of optimal transport. We start the Dissertation with a discussion of the currently used mathematical and statistical tools, illustrated by an analysis of nuclear magnetic resonance spectra of St. John's wort, carried out jointly with the Faculty of Pharmacy of the Medical University of Warsaw, published in the article entitled *Harvest time affects antioxidant capacity, total polyphenol and flavonoid content of Polish St John's wort's (Hypericum perforatum L.) flowers* [2]. While the laboratory technique used in this study is slightly different from mass spectrometry, the data analysis methods used in both types of spectrometry are the same.

We use the example analysis to highlight and discuss the imperfections of the currently available methods. One of them is the lack of mathematical tools which could be used to meaningfully compare spectra of different molecules, spectra of the same molecules obtained under different experimental conditions, or spectra acquired experimentally and predicted theoretically. In order to solve these problems, we treat mass spectra as probabilistic measures on the real line $\mathbb{R}$ — discrete or continuous depending on the type of spectrum — and use the optimal transport theory to compare them. Our main tool is the Wasserstein distance, equal to the minimum total distance on the m/z axis over which the signal needs to be moved in order to transform one spectrum into the other [3, 4].

We then use the approach based on optimal signal transport and Wasserstein distance to develop a new algorithm for regression of mass spectra, i.e. the problem of approximating a spectrum of a mixture of chemical compounds by a linear combination of spectra of the components of the mixture. We present a basic version of the algorithm, which solves a simple version of this problem. We assess the accuracy of the estimation and the practical computational complexity by conducting computer simulations. The results on the Wasserstein distance properties in the context of analysis of mass spectra and the basic Wasserstein regression algorithm were published at a peer-reviewed conference Workshop on Algorithms in Bioinformatics (WABI) as an article entitled *The wasserstein distance as a dissimilarity measure for mass spectra with application to spectral deconvolution* [5].

Next, we conduct further research on the practical applications of the spectral regression method and identify the main factors that reduce the accuracy of the estimation. We present an extension of the basic method that automatically detects and removes contaminating signals in the spectrum of the analyzed mixture during the fitting procedure. We also conduct further research on the computational aspects of the presented method and obtain a more efficient algorithm. The results were published in an article entitled *Masserstein: Linear regression of mass spectra by optimal transport* [6].

The Dissertation ends with a presentation of the application of the developed method to the analysis of mass images, i.e. images in which each pixel is associated with a mass spectrum. Together with a research group headed by prof. Olga Vitek from the Northeastern University in Boston, we conduct research on mass image segmentation, i.e. the problem of dividing an image into regions with characteristic chemical compositions. Using simulated data, we show that the currently used methods carry the risk of obtaining erroneous results due to overlapping signals from different compounds and pixel-to-pixel variance of signals. We then show that our method returns more accurate spatial distributions of compounds thanks to separating overlapping isotopic envelopes, and the segmentation method developed by

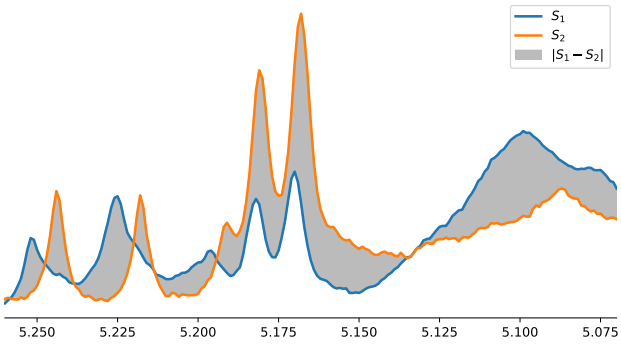

FIGURE 1.2: An illustration of the L^1 distance between two NMR spectra. The distance between spectra S^1 and S^2 is equal to the area of the shaded region.

the group of prof. Vitek returns improved segments thanks to mitigating the influence of signal variance on their shapes. We then verify our results on two mass images obtained experimentally. We detect cases analogous to the simulated ones and show that lipids with overlapping isotopic envelopes are ubiquitous in this kind of data. We then use our methods to obtain segments with a good correspondence to the actual anatomical regions. An additional conclusion from this study is that our methods of simulating mass spectrometric images provide a good representation of real data sets. We have described the results in an article entitled *Resolving overlapping isotopic envelopes improves segmentation of mass spectrometric images* (submitted for review) [7].

1.1 Computational Mass Spectrometry: the state of the art

Computational methods have been used in mass spectrometry since its very beginning [8, 9, 10]. Initially, they allowed for simple analyzes, such as possible elemental compositions of low-mass compounds. The development of computers and laboratory techniques made it possible to carry out increasingly complex calculations on increasingly accurate data, resulting in the continuous development of computational spectrometry.

One of the challenges of the modern computational mass spectrometry is the problem of regression of mass spectra (often referred to as *separating of isotopic envelopes* or *deconvolution of mass spectra*), in which an experimentally measured spectrum of a mixture of chemical compounds is explained by a linear combination of theoretically predicted spectra of the mixture's components [11]. Various spectral regression approaches look for coefficients of said linear combination that minimize a chosen measure of difference between it and the spectrum of the mixture. In this work, we use the terms *regression* and *deconvolution* of spectra interchangeably.

One of the frequently used measures of the difference between spectra is the L^1 distance, illustrated in Fig. 1.2. For spectra normalized so that their signal integrates to a unity, we can formally define it as

$$L^1(\mu, \nu) = \int_{\mathbb{R}} |\mu(x) - \nu(x)| dx.$$

This distance is used in many different kinds of spectrometry and spectroscopy, including mass spectrometry and nuclear magnetic resonance (NMR) spectroscopy.

The L^1 distance measures the cumulative difference of the intensities of signals with identical m/z values in both spectra, and therefore the total difference in the relative concentrations of molecules in the compared samples. Other approaches to comparing spectra are conceptually similar, and differ mainly in the way that the intensity at a given m/z value is compared between spectra. They include the L^2 distance, the correlation of signals, and multiple similar methods.

Measures of difference (or similarity) of spectra are used not only to compare spectra in the problem of regression, but also in the problem of compound identification (where, based on a spectrum of fragments of a chemical molecule, we want to figure out its structural formula). They are also used in data analysis to compare spectra of various mixtures of compounds, for example to assess the influence of experimental conditions on their content.

In order to put the Dissertation in the context of the current state of computational spectrometry, as well as to familiarize the reader with the use of known mathematical tools in this context, we start the Dissertation with an exemplary analysis of extracts from *Hypericum perforatum* — the popular St. John's wort — carried out in cooperation with the Faculty of Pharmacy of the Medical University of Warsaw [2]. The goal of this study is to compare the content of extracts obtained from plants collected at two different time points with the use of two different solvents. For this task, we use the L^1 distance to determine the overall differences between the spectra. Then, we identify the signals corresponding to selected compounds, determine the relative concentrations of these compounds by integrating their signals, and compare these contents between spectra. We also study the correlation of selected signals in spectra obtained on two different instruments. While we have used nuclear magnetic resonance spectrometry instead of mass spectrometry in this study, the approach to data analysis remains exactly the same for both experimental techniques.

The mathematical tools used in the analysis of St. John's wort extract make it possible to obtain large amounts of information, but also have significant limitations. Due to the properties and interpretation of the L^1 distance, it is applicable to experiments in which we want to compare two mixtures in terms of the concentration of their compounds. However, this distance does not allow for a comparison of how different the compounds in two mixtures are — two spectra containing different ions will generally have a unit L^1 distance, regardless of the chemical similarity between these ions. On the other hand, comparing the spectra in terms of the chemical similarity of their ions is needed in many types of experiments, for example in the identification of chemical molecules.

Another problem with the L^1 distance is that, in order for the distance to be chemically meaningful, the compared spectra must be obtained on similar instruments with similar settings. Consequently, it has a limited application when comparing experimentally measured mass spectra (where the signal is a continuous function) and theoretically predicted spectra (where the signal is a discrete function). For this to be possible, the continuous spectrum (called a *profile mode spectrum* in the spectrometric literature) must be converted to a discrete spectrum (called a *centroid mode spectrum*) by a *centroiding* procedure. Typically, centroiding consists of finding local signal maxima and numerical integration of the signal within a region around each of these maxima. Such procedures, like all data transformations of this type, inevitably lead to the loss of information. For example, we usually lose the information about the width of the signals in profile spectra. Moreover, closely located signals can merge into a single peak with a location inbetween the m/z values of two

ions, decreasing the mass accuracy of the measurement. Finally, they have a number of computational difficulties, such as correct and robust identification of signal maxima and proper integration radii.

In the spectrometric literature to date, there have been no methods that could meaningfully compare spectra of different chemical molecules or spectra obtained with different methods. This was particularly problematic in mass spectral regression, where we want to compare the combination of theoretical spectra with a given experimental spectrum, as well as in the identification of compounds based on the spectra of their fragments, where we want to compare fragmentation patterns of different ions. For these reasons, we have conducted research into alternative methods for comparing mass spectra.

1.2 Transporting signals — a new approach to mass spectra comparison

The inspiration for the development of a new approach to comparing mass spectra came from the field of biological sequence analysis, namely the method of *sequence alignment* [12]. In simple terms, it is a method based on introducing gaps in two sequences so as to match their characters as well as possible. The distance between the sequences is large if we need to insert many gaps and if their characters, paired through such gap insertion, are dissimilar. An analogy to the desired properties of a mass spectra comparison method arises, which should reflect the distances between the signals on the m/z axis, as well as the differences in the intensities of paired signals.

However, an approach directly based on this kind of alignment of signals in two spectra would have two major drawbacks. First, it would only be useful for the comparison of centroid spectra with discrete signals. Second, it would most likely be computationally expensive — comparing biological sequences has a quadratic computational complexity with respect to their length, so for mass spectra one can expect the complexity to be the same or higher.

It turned out that the achievements of the theory of optimal transport could be used to compare spectra in a way that is conceptually similar to sequence alignment, but at the same time applicable to profile spectra and less computationally complex. This can be accomplished by treating mass spectra as probabilistic measures on the real line $\mathbb{R}$ and comparing them using the *Wasserstein distance*.

Let μ and ν be the two probability distributions on the real line $\mathbb{R}$. Let us consider the space of their joint distributions Γ. Let d be some distance function on $\mathbb{R}$. The Wasserstein distance W_p^d is then defined as:

$$W_p^d(\mu,\nu) = \left(\min_{\gamma\in\Gamma}\int_{\mathbb{R}^2} d(x,y)^p\gamma(x,y)dxdy\right)^{1/p}.$$

We interpret it as the minimum total distance (in the sense of the distance d in $\mathbb{R}$) that the signal from one of the spectra needs to travel to convert it into the other. The joint distribution γ of measures μ and ν is interpreted as a transport plan: the amount of signal transferred from μ at x to ν at y is equal to $\gamma(x,y)$. The definition of the Wasserstein distance can be extended to other metric spaces than $(\mathbb{R},d)$, but we will not cover it in this Dissertation [3, 4].

In the context of mass spectrometry, the most interesting is the $W_1^{|\cdot|}$ Wasserstein distance, i.e. the one where $p = 1$ and $d(x,y) = |x-y|$. The usefulness of this

distance comes from the fact that it has a chemical interpretation as the total difference in the m/z values of signals in both spectra, and, in turn, the difference in the m/z values is a natural measure of the difference between chemical molecules. Such a Wasserstein distance can therefore be interpreted as a summarized difference between the ions in the compared spectra, as opposed to the summarized difference of their concentrations measured by L^1.

In the following part, we will refer to $W_1^{|\cdot|}$ simply as the Wasserstein distance, and denote it with the letter W:

$$W(\mu,\nu) = \min_{\gamma\in\Gamma} \int_{\mathbb{R}^2} |x-y|\gamma(x,y)dxdy.$$

In the article entitled *The wasserstein distance as a dissimilarity measure for mass spectra with application to spectral deconvolution* [5], published at a peer-reviewed conference Workshop on Algorithms in Bioinformatics (WABI), we showed that the Wasserstein distance reveals the differences in chemical structures of molecules better than the distances based on comparing the intensities of signals with identical positions on the m/z axis.

The Wasserstein distance makes it possible to rigorously compare profile spectra with their centroided counterparts. This allowed for the development of a new mass spectral regression algorithm, also published in *The wasserstein distance as a dissimilarity measure for mass spectra with application to spectral deconvolution* [5]. This algorithm allows for an easier and more accurate fitting of linear combinations of theoretically predicted spectra to experimentally measured ones than the previous approaches. Thanks to the properties of the Wasserstein distance, our algorithm does not require centroiding of the experimental spectra, nor the processing of theoretical spectra so as to reduce their accuracy to the level of the experimental spectrum, which is necessary for other methods. It is also more robust to measurement inaccuracies on the m/z axis compared to algorithms based on point-wise comparison of intensities.

Formally, the problem of Wasserstein regression of mass spectra is defined as follows. Let μ be a spectrum of a mixture (usually measured experimentally) and let $\nu_1,\ldots,\nu_k$ be the spectra of the components of the mixture (usually predicted theoretically). Assume that all the spectra are normalized so that their intensities sum up to unity. We define a model of the mixture as a linear combination $\nu_p = \nu_1 p_1 + \cdots + \nu_k p_k$, where $p = p_1 + \cdots + p_k = 1$. Wasserstein regression is a problem of finding a vector of proportions p^* that solves an optimization problem given by:

$$p^* = \operatorname*{arg\,min}_{p:p_1+\cdots+p_k=1} W(\mu,\nu_p). \tag{1.1}$$

In this Dissertation, we show that for normalized discrete spectra (i.e. ones with a finite number of m/z values which correspond to non-zero signal intensities that add up to unity) the above problem is equivalent to the following linear program. Let $(s_1,\ldots,s_n)$ be an ordered vector of m/z values for which any of the μ,ν_i spectra has a non-zero signal intensity. Let N_i denote a vector of length n containing the cumulative sums of intensities of ν_i in points s_1 to s_n, and let N denote a $k\times n$ matrix where the i'th row is the N_i vector, and let M denote a vector of the cumulative sums of signal intensities of μ. Let I_n denote an $n\times n$ unit matrix, and let J_n denote a matrix of dimensions $n-1\times n$ equal to an identity matrix without the last row. Let c be a vector of length $2n+k$ equal to $c=(M,-M,0_k)$, and let b be a vector of length

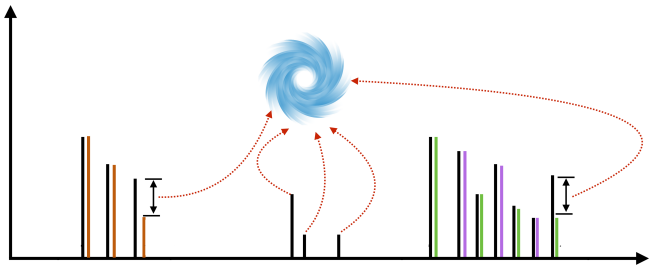

FIGURE 1.3: An illustration of the Wasserstein regression-denoising method with an experimental spectrum in black, theoretical spectra in colors, and the auxiliary spectrum ω presented as a vortex sucking in the noise signals.

$n-1+k$ equal to $b = (-d, 0_k)$, where $d_i = s_{i+1} - s_i$ for $i = 1.2, \ldots n-1$. Let's define

$$A = \begin{bmatrix} -J_n & -J_n & 0 \\ N & -N & -I_k \end{bmatrix}.$$

Then, the proportions optimizing the problem (1.1) correspond to the last k coordinates of a vector y that solves the following linear program:

$$\max_y \{y^T b \mid A^T y \leq c, y \in \mathbb{R}\} \tag{1.2}$$

In the Dissertation we present an efficient algorithm based on the Interior Point Method for solving the above linear program, using the structure of the matrix A to speed up the calculations. Experiments on simulated mass spectra showed that the Wasserstein regression method allows for an accurate estimation of the proportions of mixture components and is robust to measurement inaccuracies on both the m/z axis and, to a moderate degree, the signal intensity axis. The algorithm was implemented in the Python 3 programming language and published as the `masserstein` package, available at `https://github.com/mciach/masserstein`.

1.3 Further research on spectral regression — removal of contaminating signals

Good preliminary results in the comparison and regression of spectra using the Wasserstein metric resulted in further research into this approach. In an article entitled *Masserstein: Linear regression of mass spectra by optimal transport* [6], we have examined more properties of this metric in the context of comparing mass spectra. In particular, we give theoretical calculations of its values on pairs of spectra representing some common cases, such as the distance between spectra of the same ion in profile (continuous) and centroid (discrete) modes. We have also further developed the original Wasserstein regression algorithm to increase the accuracy of the estimation, its computational complexity, and the method's practical applicability. While the basic algorithm allowed for obtaining the correct proportions of the mixture components in well-prepared spectra, its application to typical mass spectra required further work.

At this stage of the research, the main problem that caused errors in the estimation of component proportions from real spectra was the presence of signals originating from sample impurities, spectrometer errors, and other sources. We refer to

such signals collectively as *noise*.

Removing noise during data pre-processing would carry the risk of removing some of the actual signal from the spectrum, and would also stand in conflict with the paradigm of the `masserstein` package, which aims to reduce the need for such preprocessing and provide algorithms that can analyze "raw" data. For this reason, we have modified the Wasserstein regression algorithm so that it detects and removes the noise simultaneously with the fitting of the model (i.e. the linear combination of theoretical spectra).

The main conceptual novelty in the new regression algorithm was the introduction of an auxiliary "artificial spectrum" ω to which we could send some signal from the spectrum of the mixture. The ω spectrum is a relatively unusual concept: it is a normalized spectrum (i.e. with a unitary total signal intensity) whose signal is concentrated at one point without a specified position, but which distance to any point on the line $\mathbb{R}$ is equal to a number κ which is a parameter of the model. A visualization of the proposed approach is shown in Fig. 1.3.

The Wasserstein regression-denoising procedure presented in the Dissertation consists in solving an optimization problem with $k+1$ variables:

$$p^* = \underset{p:p_0+\cdots+p_k=1}{\arg\min} W(\mu, p_0\omega + \nu_p), \tag{1.3}$$

where, as in the problem (1.1), μ is the spectrum of the analyzed mixture, and $\nu_p = \nu_1 p_1 + \ldots \nu_k p_k$ is the model of this spectrum, i.e. a linear combination of the spectra of the mixture components. Note that now we require $p_0 + p_1 + \cdots + p_k = 1$, so the signal of the spectrum ν_p may not sum up to unity (or integrate to unity in the case of profile spectra). This represents a situation when not all signal from μ can be represented by the model.

The unusual structure of the "spectrum" ω provides the regression-denoising method with several properties which are important from a practical point of view. We treat the signal transported from the μ spectrum to the ω spectrum as noise removed from μ. Due to the fact that ω is defined as equidistant from each point on $\mathbb{R}$, we get a constant cost κ of removing signal, in the sense that the cost does not depend on the signal's location on the m/z axis. The total amount of signal removed from μ is equal to the estimated proportion of spectrum ω, that is p_0. In addition, such an abstract definition of the regression-denoising problem facilitates its theoretical analysis. In the Dissertation we present an interpretation of the κ parameter as the maximum "feasible" distance of signal transport on the m/z axis. This interpretation is important from the point of view of practical applications of our method, because it allows for the initial selection of the value of the κ parameter. We also present the results of several special cases, including examples when transport can occur at distances greater than κ (showing that κ does not give a strict threshold of such distances, which adds some desired flexibility to the model), and we examine the uniqueness of the solution to the regression-denoising problem (namely, we construct a case for which the solution is non-unique).

While the regression-denoising procedure in the form of Equation (1.3) is well suited for theoretical analysis, it does not constitute in itself the basis for an implementation of an algorithm that estimates the proportions of reference spectra. In order to transform the problem (1.3) into a form suitable for implementation in a computer program, we apply an approach analogous to that used in the basic version of the algorithm: we show that this problem can be expressed as an L^1-regression on the cumulative sums of signals of the studied spectra, that can then be solved

by linear programming. In this case, however, we use a different method of transforming the L^1 regression problem into a linear program, and we carry out further transformations to simplify it. Finally, we show that solving problem (1.3) is equivalent to solving the following (dual) linear program, where V is the intensity vector of the spectrum μ (i.e. $V_i = \mu(s_i)$), and W is the intensity matrix of the ν_i spectra (i.e. $W_{ij} = \nu_j(s_i)$):

$$\begin{array}{lrlll} \text{maximize} & V^T z & \text{over} & z & \\ \text{subject to} & W^T z & \leq 0, & & \\ & z_i - z_{i+1} & \leq & s_{i+1} - s_i, & i = 1, 2, \ldots, n-1, \\ & z_i - z_{i+1} & \geq & s_i - s_{i+1}, & i = 1, 2, \ldots, n-1, \\ & & & z \leq \kappa, & \end{array} \tag{1.4}$$

The optimal proportions of the components of the mixture are equal to the dual variables for the constraints $W^T z \leq 0$ of the above program, and the amount of signal removed from the μ spectrum at point s_i is a dual variable for the constraint $z_i \leq \kappa$. To solve problem (1.4), we used the Simplex method implemented in the PuLP package of the Python 3 language. Note that, thanks to the aforementioned simplification of the linear program, we obtained an optimization problem over n variables, compared to $n - 1 + k$ variables in the algorithm presented in the previous section, even though the method itself is more complex and implicitly involves additional variables corresponding to amounts of noise removed from the points of μ.

1.4 Mass Image Segmentation

The Wasserstein regression-denoising method is not only interesting due to its mathematical and computational properties, but is also useful for chemical and biological research. Thanks to the separation of overlapping signals, it allows for a more precise determination of the concentrations of particular ions in the studied spectrum. This, in turn, allows us to obtain more accurate results and to perform more complex analyzes. One of the applications of our method is the analysis of mass spectrometric images, that is, images in which each pixel is associated with a mass spectrum. Such images are typically obtained for biological samples, e.g. for tissue sections, and allow for a visualization of the spatial distributions of hundreds of molecules in a single experiment.

Together with the group of prof. Olga Vitek from Northeastern University in Boston, we conducted a research on mass spectrometric image segmentation, where the regression-denoising method turned out to be crucial for obtaining biologically correct results. A segmentation of a mass spectrometric image is a problem of dividing it into areas with characteristic spectra. Generally, the purpose of the segmentation is to discover areas of the sample with characteristic chemical compositions, such as tissues, tumors, or other anatomical regions. A correct segmentation, i.e. one in which the segments correspond to actual anatomical regions, allows, among other things, for identification of *biomarkers*, i.e. compounds characteristic for those regions. Conversely, if the segments inaccurately or incorrectly correspond to anatomical regions, this leads to inaccurate or erroneous conclusions about the chemical characteristics of the studied tissues.

One of the basic and common approaches to the segmentation problem is to select an m/z value of interest (typically the monoisotopic mass of a selected molecule), determine the corresponding signal intensities in each pixel, and cluster these intensities with the K-means algorithm. In this case, we cluster one-dimensional data

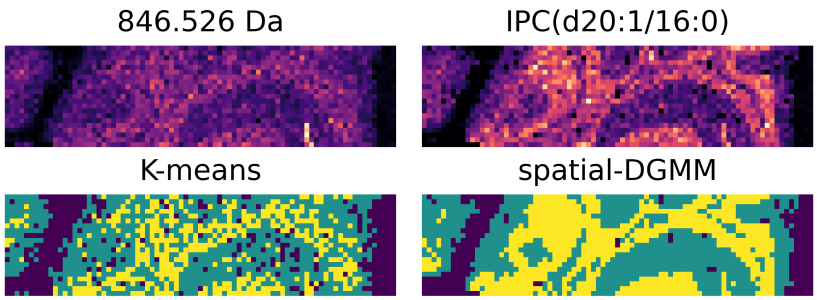

FIGURE 1.4: A comparison of mass spectrometric image segmentations based on the K-means clustering (bottom left) of the monoisotopic peak intensity of a selected lipid (top left) and the `spatial-DGMM` segmentation (bottom right) based on lipid signal estimated with the regression method proposed in this Dissertation (top right).

and ignore the spatial relationships between pixels. Naturally, such a simple approach turns out to be sensitive to the problem described in the introduction, namely overlapping isotopic envelopes, especially when the overlapping ions have different spatial distributions. It is also be sensitive to the naturally occurring variance of intensity between pixels (termed *pixel-to-pixel variance* or *variability*), which means that two pixels from different tissues may have similar intensities at a given m/z, while at the same time those intensities for pixels in a single tissue may differ.

In the article titled *Resolving overlapping isotopic envelopes improves segmentation of mass spectrometric images* (submitted for review) we show that overlapping isotopic envelopes and pixel-to-pixel variance do indeed carry the risk of obtaining segmentation that either has no biological significance or even misidentifies biomarkers [7]. We construct a simulated data set in which the K-means segmentation mixes pixels from different tissues and misidentifies biomarkers of healthy and diseased tissues. We then show that Wasserstein regression method, developed in this Dissertation and implemented in the `masserstein` package, allows for a proper assignment of biomarkers thanks to separating overlapping signals. However, due to the pixel-to-pixel variability of ion signals, the method is not sufficient to obtain segments with a high degree of agreement with the anatomical regions. To overcome the latter problem, we combined the `masserstein` package with the `spatial-DGMM` segmentation method, developed by the group of prof. Vitek, which accounts for the spatial relationships between pixels.

The segmentation based on the combination of `masserstein` and `spatial-DGMM` methods corresponded well to the anatomical regions on the simulated data. We then applied our methods to the analysis of two mass images, in which we have detected overlapping isotopic envelopes which caused the K-means method to return an incorrect segmentation. Our methods, on the other hand, produce segments that closely match actual anatomical regions. A comparison of segments, corresponding to different characteristic concentrations of a selected lipid, obtained by the K-means method and our approach is shown in Fig. 1.4. We have also determined that, in both images, about 50% of the lipids have monoisotopic peaks that fall within isotopic envelopes of other lipids, which results in the risk of erroneous estimation of their concentration.

1.5 Summary of the Chapter

In the dissertation titled *Algorithms for computational mass spectrometry based on the optimal transport theory*, we investigate the use of the Wasserstein distance for comparing mass spectra. We use this distance to develop an algorithm for the problem of regression of mass spectra, which we then extend so as to detect and remove contaminating signals and speed up the computations. We apply the developed algorithm to the problem of mass image segmentation and show that it outperforms currently used approaches in terms of the biological relevance of the results, allowing for a more reliable identification of tissue biomarkers.

This Dissertation presents interdisciplinary results, connecting problems and solutions from mathematics, computer science, chemistry, and biology. Accordingly, it is structured in a way to make it accessible for readers from different disciplines. In each chapter, we provide explanations of laboratory techniques and terminology from chemistry and biology, as well as intuitive explanations of mathematical definitions and results. In Chapters 4 and 5, which deal with algorithms for regression of mass spectra, we first provide brief and general overviews of the algorithms, and then proceed to demonstrate their applications. Formal proofs are relegated to separate sections for interested readers.

Acknowledgment of scientific collaboration

The work presented in this Dissertation has been conducted in collaboration with different co-workers and research groups.

The study of *Hypericum perforatum* extracts presented in Chapter 2 has been conducted in collaboration with the Faculty of Pharmacy With Laboratory Medicine Division of the Medical University of Warsaw, the Faculty of Biology of the University of Białystok, and the Magritek GmbH company. The experimental part of this study, including sample collection, obtaining the high-field NMR spectra, and their annotation based on available literature, was performed by Katerina Makarova, Joanna J. Sajkowska-Kozielewicz, Katarzyna Zawada, Ewa Olchowik-Grabarek, Natalia Dobros, Paulina Ciechowicz and Hélène Freichels. Hélène Freichels obtained the low-field benchtop NMR spectra.

The preliminary study of the applications of the Wasserstein distance to computational mass spectrometry, presented in Chapter 4, was done in close collaboration with my colleagues from the Faculty of Mathematics, Informatics and Mechanics of the University of Warsaw and Szymon Majewski, who, at that time, was a doctoral student in the Institute of Mathematics of the Polish Academy of Sciences. Błażej Miasojedow helped with the design of a measure that could meaninfully compare mass spectra of different ions, and realized that the Wasserstein distance has some properties similar to the alignment of biological sequences. Szymon Majewski provided a tremendous help with the mathematical derivations needed to obtain an efficient Interior Point Method algorithm for the problem of Wasserstein regression and with the analysis of its computational complexity. Michał Startek and Wanda Niemyska helped with the implementation of algorithms and computational experiments. Prof. Anna Gambin guided us throughout the study and coordinated our work.

The work presented in Chapters 3 and 5 was done in collaboration with my colleagues from the University of Warsaw and Szymon Majewski. Prof. Dirk Valkenborg provided a set of 200 spectra to test our methods. Błażej Miasojedow and Szymon Majewski have helped with the mathematical proofs and derivations. Grzegorz Skoraczyński, Michał Startek and Anna Gambin helped with the implementations of algorithms and the design of computational experiments to test the Wasserstein regression method. Prof. Anna Gambin coordinated our work.

The work presented in Chapter 6 was done in collaboration with the research group of prof. Olga Vitek from Khoury College of Computer Sciences at Northeastern University. In this Chapter, Dan Guo, who was at the time a PhD student under the supervision of prof. Vitek, performed the segmentation of images using `spatial-DGMM`. Prof. Anna Gambin coordinated our work and helped to design the study. Prof. Dirk Valkenborg provided suggestions about the study, suggested possible explanations of the apparent lack of ^{41}K isotope in the spectra, and how to further investigate this problem. Prof. Vitek provided valuable insights about the presentation of our results.

Publications with results from the dissertation

Majewski, Szymon, et al. "The wasserstein distance as a dissimilarity measure for mass spectra with application to spectral deconvolution." *18th International Workshop on Algorithms in Bioinformatics, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.* (2018).

Makarova, Katerina, et al. "Harvest time affects antioxidant capacity, total polyphenol and flavonoid content of Polish St John's wort's (*Hypericum perforatum* L.) flowers." *Scientific reports* 11.1 (2021).

Ciach, Michał Aleksander, et al. "Masserstein: Linear regression of mass spectra by optimal transport." *Rapid Communications in Mass Spectrometry* e8956 (2021).

Domżał, Barbara, et al. "Masserstein+: robust tool for separating overlapping signals in mass spectra". Manuscript submitted for review.

Ciach, Michał Aleksander, et al. "Resolving overlapping isotopic envelopes improves segmentation of mass spectrometric images". Manuscript submitted for review.

Other publications

Ciach, Michał Aleksander, et al. "Estimation of Rates of Reactions Triggered by Electron Transfer in Top-Down Mass Spectrometry." *Journal of Computational Biology* 25.3 (2018).

Ciach, Michał Aleksander, Anna Muszewska, and Pawel Górecki. "Detecting Locus Acquisition Events in Gene Trees." *17th International Workshop on Algorithms in Bioinformatics, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik* (2017).

Ciach, Michał Aleksander, Anna Muszewska, and Paweł Górecki. "Locus-aware decomposition of gene trees with respect to polytomous species trees." *Algorithms for Molecular Biology* 13.1 (2018).

Ciach, Michał, Julia Pawłowska, and Anna Muszewska. "Horizontal gene transfer in 44 early diverging fungi favors short, metabolic, extracellular proteins from associated bacteria." *bioRxiv* (2021).

Ciach, Michał Aleksander. "The Clade Displacement Index: how to detect horizontal gene transfers in unrooted gene trees." *bioRxiv* (2021).

Bokota, Grzegorz, et al. "Trapalyzer: Automatic feature detection and quantification for in-vitro NET release studies". Manuscript submitted for review.

Chapter 2

Current approaches to analysis of spectral data

Due to its wide range of well documented pharmacological activities, such as antidepressant, antiviral, and antibacterial effects, St. John's wort (*Hypericum perforatum* L.) is one of the most consumed medicinal plants in the world [13]. Its extracts are used as phytopharmaceuticals and nutraceuticals. St. John's wort antidepressant activity has been related to the synergetic effect of hypericin and phenolic compounds [14]. The latter modulate the key cellular processes such as redox, metabolic and energetic homeostasis, proteostasis, signaling and oxidative stress, thus decreasing the risk of cardiovascular, neurodegenerative and metabolic diseases, as well as of some forms of cancers [15].

Although, due to the presence of phenolic compounds, *H. perforatum* has antioxidant properties, there are only a few studies on this subject [14, 16, 17, 18, 19, 20, 21, 22]. The variation in total polyphenol (TP) content and antioxidant properties of St. John's wort from the Balcan peninsula [17], Lithuania [23, 24], Turkey [25] and China [26] was studied. The main factors which influenced the TP and antioxidant properties were geographical origin, whether the plant was wild or cultivated [27], individual chemotype [28], part of the plant studied (leaves, flowers, fruits, roots), harvesting stage (floral budding stage, blooming stage or fruit set stage) and the age of the plant (1-, 2- or 3- year plant) [26, 29, 30, 31]. However, many other factors, including the temperature and light intensity, also influence these properties [30, 32].

TP and antioxidant properties are very general characteristics of plant extracts, typically studied by chromatographic techniques. These techniques require reference compounds and usually do not reveal unknown metabolites that may contribute to the biological activity of the phytochemicals [33]. Nuclear Magnetic Resonance (NMR) spectroscopy makes it possible to overcome these limits, detecting both known and unknown constituents of complex mixtures. In particular, 1H NMR spectroscopy is widely used in studies of plant extracts to quantitatively and simultaneously analyze all proton-bearing compounds, and consequently all relevant substance classes in the extracts [33, 34].

The signals of main components of *H. perforatum* extracts were assigned previously using high-field NMR spectra, i.e. NMR spectra with a high resolution of signals [35]. However, handling a high-field NMR spectrometer is sophisticated and costly, and requires large quantities of not environmentally friendly liquid helium and liquid nitrogen. Thus, in the last decades, benchtop instruments are gaining popularity. A number of benchtop NMR spectrometers operating at 40-100MHz is available and used both in research and industry. Benchtop NMR spectrometers are significantly cheaper, smaller and much easier in operation, and do not require liquid helium nor liquid nitrogen. However, due to their low resolution, benchtop spectrometers are used mainly in chemical reaction monitoring, studies of synthetic

drugs, and other applications where well-separated signals can be obtained. On the other hand, typical NMR spectra of plant extracts consist of highly overlapping signals. Low operating frequencies increase the overlap even further, resulting in signals which are hard to identify or quantify. This limits the possibilities of direct applications of benchtop NMR for the studies of plant extracts.

The contents of this Chapter. In this Chapter, we present an application of statistical and mathematical methods to analyze spectra. We illustrate this application, as well as the drawbacks of the modern approaches, on a study of the compositional variation of *H. perforatum* extracts. We use spectra obtained on two types of spectrometers: a high-field one, with a high resolution on the ppm axis (corresponding to the m/z axis in mass spectrometry), and a benchtop one, with a low resolution and many overlapping signals. We apply basic statistical methods to check if the benchtop instrument accurately reflects the differences in samples as observed in the high-field spectra. We discuss problems arising due to overlapping signals that can be potentially solved by advancements in computational spectrometry. We illustrate the problems with comparing spectra obtained on the two instruments caused by differing resolutions.

2.1 Materials and methods

Plant material (flowering tops) was collected from its natural habitat in 2016 from the end of June till the end of August (26.06, 20.07, 8.08, 18.08, 28.08) in the vicinity of Radom (Mazovia Province), located in the east-central of Poland (GPS coordinates 51.317709, 21.259254). The place is 162m above sea level. According to Köppen-Geiger climate classification this climate is classified as Dfb (warm-summer humid continental climate). The plant material was compared with the botanical description key [36] and the botanical drawing from an atlas of plants[1]. The shape of the stem and leaves, the arrangement of the leaves and the inflorescence, the structure / type of flower, and the appearance of the fruit was compared. The characteristic feature of *H. perforatum*, i.e. translucent dots on flower petals and leaf blades, were identified. After identification of the plant, only flowers at the blooming stage were collected. Part of the collected plant material was air-dried in a dark place, at room temperature, for 7 days. A second part of the collected plant material was frozen and lyophilized at -25 °C for 96 h. The water content of dried plant material was determined by the oven-drying method: portions of dried and lyophilized plant materials were weighted, dried in an oven for 2 hours at 105 °C, and reweighed. Then, 50 ml of solvent (either ethanol 96% or an ethanol-water mixture (1:1)) was added to 1 g of mechanically ground plant material and sonicated at 30 °C for 15 min. Then, extracts were filtered and dried in vacuum at either 25 °C (ethanol) or 35 °C (ethanol-water). For analysis, reconstituted solutions with a concentration of 2 mg/ml were used. All samples were prepared in duplicates. The prepared extracts were kept at -18 °C until used.

For NMR analysis, the extracts (10 mg) were dissolved in 400 μl of MeOD (deuterized methanol). 1H NMR spectra were recorded at 301 K on a Magritek 60 MHz Ultra Spinsolve instrument (Magritek GmbH, Aachen, Germany) (256 scans, a repetition time of 15 s, an acquisition time of 6.4 s with the suppression of the water peak) or a Varian VNMRS 300 Oxford spectrometer (Agilent Technologies, Santa Clara, USA) operating at 299.61 (1H) MHz (128 scans, a repetition time of 1 s and an

[1] http://www.biolib.de/

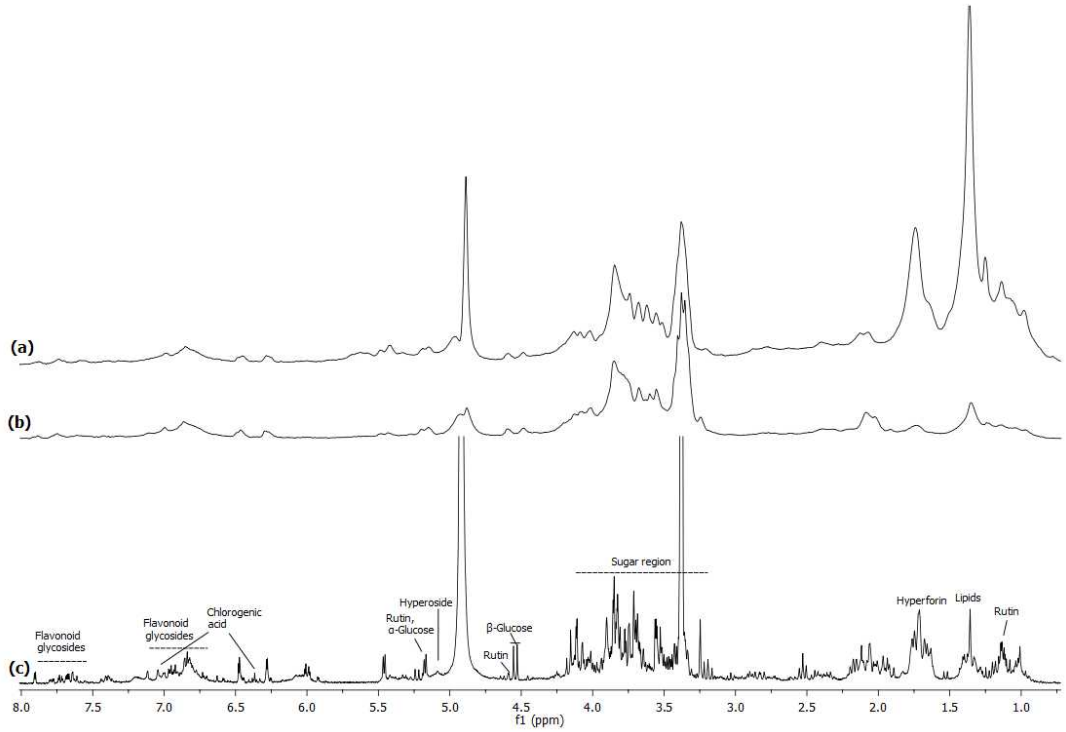

FIGURE 2.1: 1H NMR spectra of *H. perforatum* extracts. a) A benchtop NMR (60 MHz) spectrum of the ethanol extract; b) A benchtop NMR (60 MHz) spectrum of the ethanol-water extract; c) A high field NMR (300 MHz) spectrum annotated based on available literature.

acquisition time of 2 s). Each sample was prepared in duplicate. The spectra were normalized by their total signal calculated by numerical integration with the trapezoidal method. Spectra were recalibrated manually by shifting the signals in order to match intense signals at 4.92000 ppm and 1.36204 ppm between the spectra.

2.2 Results and discussion

We have used two NMR instruments to analyze samples from ethanol and ethanol-water based extraction for two collection dates, June 26 and August 18 (Fig. 2.1). The first instrument was a conventional spectrometer with 300 MHz frequency, while the second one was a benchtop instrument with 60 MHz frequency. Apart from the analysis of the spectra, we aimed to investigate the extent to which the low-resolution benchtop instrument can be used to analyze samples after prior identification of signals on the higher-resolution 300 MHz spectra.

The assignments of NMR signals of the main compounds, shown in Fig. 2.1, were based on the data published by Rasmussen et al. [35] and Bilia et al. [33]. The analysis showed that the studied extracts contain hyperforin, sugars, lipids, flavonoid glycosides (including such constituents of *H. perforatum* extracts such as hyperoside and rutin), and chlorogenic acid.

The spectra recorded with the benchtop NMR (60MHz) have a lower resolution and consist of highly overlapping signals. Nevertheless, they show noticeable differences in the ethanol and ethanol-water extracts (Fig. 2.1). We have also observed some differences in the extract composition for different collection dates.

TABLE 2.1: Signal fold changes for identified components computed from 300 MHz spectra. Increasing and decreasing signals highlighted in green and blue respectively. June 28 w.r.t. August 18: the signal area at August 18 divided by the corresponding signal area at June 28 per solvent and compound; Ethanol-Water w.r.t. Ethanol: the signal area in the spectrum of an ethanol-water extract divided by the corresponding signal area in the spectrum of an ethanol extract per date and compound.

		Signal area ratio			
		June 28 w.r.t. August 18		**Ethanol-Water w.r.t. Ethanol**	
Compound	**ppm range**	**Ethanol-Water**	**Ethanol**	**August 18**	**June 28**
Rutin	0.85-1.06	0.607	0.967	0.256	0.407
Lipids	1.07-1.3	0.678	1.224	0.187	0.338
Hyperforin	1.6-1.8	0.501	0.985	0.223	0.439
Beta-glucose	4.51-4.57	1.257	1.361	1.328	1.437
Hyperoside	5.04-5.14	0.654	0.993	0.38	0.577
Alpha-glucose	5.15-5.19	1.002	1.208	1.062	1.281
Chlorogenic acid	6.24-6.31	0.747	0.782	1.295	1.356
Flavonoid glycosides	6.55-7.26	0.66	0.773	1.271	1.49
	7.88-7.93	1.003	0.854	1.177	1.002

To compare the high-field spectra of the samples quantitatively, we have normalized the spectra by their total area, integrated the assigned signals, and computed the ratios of their areas. The results are shown in Table 2.1. Compared to the ethanol extracts, the ethanol-water ones contained less lipids, hyperforin, and hyperoside than the ethanol-water ones, but more sugars, flavonol glycosides and chlorogenic acid. The most pronounced changes from June 26 to August 18 were a decrease in the amount of chlorogenic acid, hyperoside, hyperforin and flavonol glycosides, as well as an increase in the sugar content, in particular the alpha- and beta-glucose. The ethanol extracts also showed a slight increase in the lipid content.

The ethanol-water extracts showed a decrease in the signal in the lipid region, contrary to the ethanol extracts. However, as the former contained a much smaller amount of lipids than the latter (between three to five times, see Table 2.1), which is likely to cause a less precise measurement, we can conclude that the lipid content has increased from June 26 to August 18. The increase in lipid content is in agreement with the results reported by Amira et al. [37], where the flowers harvested at the end of August presented higher values of lipids (25.48 %) than harvested at the end of June (18.56 %). Similar results were reported for the leaf extracts of *Ilex paraguariensis*, where a higher content of fatty acids was found in autumn and winter compared to spring and summer. This could be associated with biotic and abiotic stresses or plant hormones, especially jasmonic acid and its derivatives [38].

To check the overall differences between the high-field spectra, we have calculated their L^1 distances, defined as the area contained between their signals (i.e. the integral of the absolute difference of the signals). This distance measures the overall difference in signals of all compounds. In general, the choice of solvent had a much more pronounced effect on the sample composition than the collection date. The differences between samples obtained using different solvents were two- to four times larger than between different collection dates, as shown in Fig. 2.2. The ethanol spectra had only a minor difference between collection dates, with the L^1 distance approximately equal 0.17. On the other hand, the composition of samples was highly

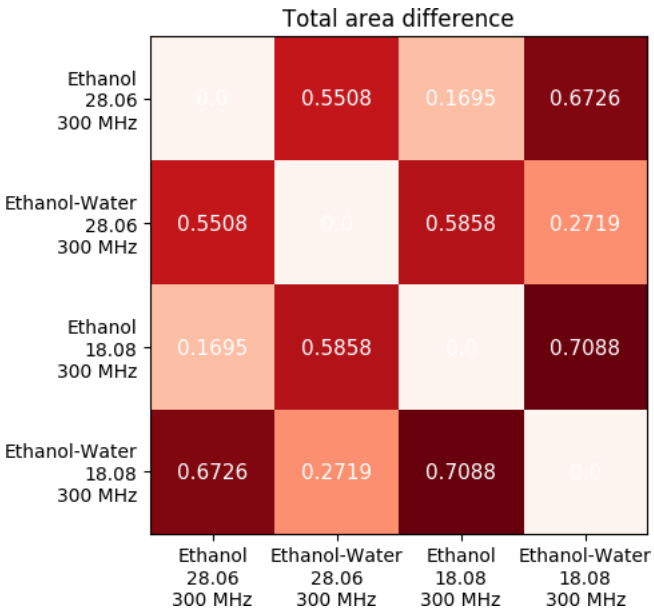

FIGURE 2.2: Comparison of NMR spectra obtained on the 300 MHz instrument for different solvents and collection dates. The L^1 distance (total area difference) measures the overall difference in signal intensities between the compared spectra. The results indicate that the solvent has a two to four times larger impact on the overall sample composition than the collection date.

dependent on the choice of the solvent, as the L1 distance between the corresponding spectra reached 0.70.

Our final goal was to check if similar results can be obtained on a 60 MHz benchtop instrument. We have focused on signals which could be visually identified based on the 300 MHz spectra. Note that, due to a much lower resolution, the regions identified in the 60 MHz spectra were much broader than in the case of 300 MHz spectra, and only a few of them did not overlap with other signals. Thus, the integration regions were adapted manually to correct for the difference in resolutions (see Table 2.2 and Figure 2.3).

The 60 MHz instrument has correctly detected the increases and decreases of the quantities of analyzed compounds. Moreover, we have detected a statistically significant positive correlation between the 300 MHz and 60 MHz results when different solvents and different dates were compared ($\rho = 0.9$, $p < 0.005$, Student's exact test). The results suggest that a low-field benchtop instrument can be used to detect changes in the sample composition when signals are identified on a high-field instrument and carefully adjusted for low-field spectra. Further development of statistical methodology, such as confidence intervals for area ratios, is also needed in order to obtain more reliable results with low-field benchtop instruments. If this is achieved, such instruments have the potential to greatly reduce the time and costs of preliminary analyses and screenings of samples.

2.3 Summary of the Chapter.

In this Chapter, we have presented an example analysis of NMR spectra, using methods which are commonly applied to nearly all kinds of spectrometry and spectroscopy. Using a basic mathematical and statistical concepts such as the L^1 distance

TABLE 2.2: Comparison of signal area ratio changes measured on 300 MHz and 60 MHz instruments. Increasing and decreasing signals highlighted in green and blue respectively. Ethanol-Water w.r.t. Ethanol: the signal area in the spectrum of an ethanol-water extract divided by the corresponding signal area in the spectrum of an ethanol extract per date and compound.

			Signal area ratios			
			Ethanol-Water w.r.t. Ethanol			
Compound	**ppm range**		**August 18**		**June 28**	
	300 MHz	**60 MHz**	**300 MHz**	**60 MHz**	**300 MHz**	**60 MHz**
Rutin and lipids	0.75-1.55	0.7-1.55	0.231	0.289	0.356	0.324
Hyperforin	1.6-1.8	1.57-1.92	0.223	0.307	0.439	0.286
Beta-glucose	4.51-4.57	4.44-4.66	1.328	1.602	1.437	1.853
Chlorogenic acid	6.24-6.31	6.16-6.34	1.295	1.627	1.356	1.780
Flavonoid glycosides	6.55-7.26	6.55-7.26	1.271	1.833	1.490	1.974
	7.88-7.93	7.85-7.96	1.177	2.302	1.002	2.091
Correlation			0.921988		0.905054	
p-value			<0.005		<0.005	

and the Pearson's correlation, we could infer that the collection date has an influence on the composition of the extracts of *Hypericum perforatum*, including a higher concentration of glucose in plants collected in late summer. We have also shown that a low-resolution benchtop NMR can be used to evaluate the composition of the extracts.

From this study, we can draw several major conclusions about the current approaches to the analysis of spectra and pinpoint areas in which computational and statistical methods can improve the reliability of the results. First, a notable amount of work needs to be done manually. To calculate the fold changes in the concentrations of different molecules in the high-resolution 300 MHz spectra, we had to manually select the regions of the molecules' signals based on available literature. Then, to compare the spectra obtained on different instruments, we had to manually adjust those regions to match the broader signals of the low-resolution spectra. Here, we did not have any other reference or a library of identified signals, so the adjustment was based on a visual comparison of signals in the high- and low-resultion spectra. Naturally, such approach to data analysis is highly subjective. What's more, in NMR spectra, just like in mass spectrometry, a single compound gives rise to multiple signals (see e.g. the regions of flavonoid glycosides in Table 2.1). Typically, only some of those signal are selected for integration, because some signal can be subject to interference from other molecules, and others can be too low to be accurately quantified. Again, this selection is often subjective. Advancements in computational spectrometry and spectroscopy have the potential to decrease the extent of subjectivity in the analysis of spectra. In particular, methods that could accurately compare spectra with different resolutions and identify matching signals are needed.

Low-field benchtop NMR instruments have the potential to greatly reduce the time and costs of preliminary analyses and screenings of samples. However, although they can detect major changes in the sample composition, they currently have limited applications in quantitative analyses of selected metabolites. This is caused, among others, by the lack of appropriate statistical and computational methods that could be applied to analyze spectra with a rather small signal-to-noise ratio and a considerable number of overlapping signals. The problem of separating overlapping signals in order to accurately quantify the concentrations of molecules is

ubiquitous in various kinds of spectrometry and spectroscopy. It is especially desirable to be able to quantify the amounts of molecules using their reference spectra obtained either on different instruments (such as high-resolution NMR spectroscopes) or by theoretical means (such as theoretical isotopic envelopes in mass spectrometry).

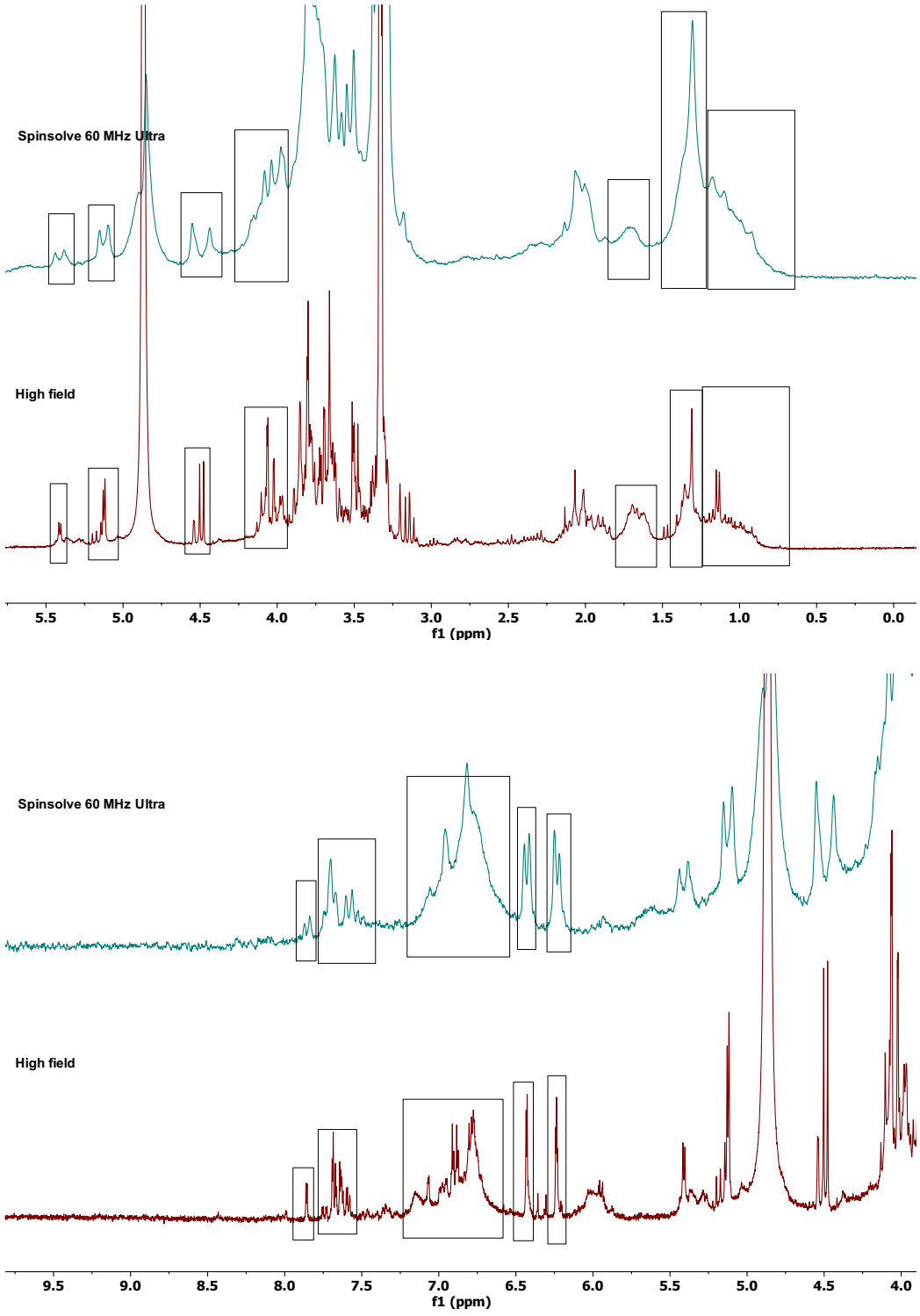

FIGURE 2.3: Regions selected for the area ratio analysis of low-resolution and high-resolution NMR spectra.

Chapter 3

Comparing spectra using the Optimal Transport Theory

Mass Spectrometry (MS) is one of the main analytical techniques of modern proteomics and metabolomics, which allows for identification and quantification of molecular compounds. In the first step, the particles are ionized; next, they are separated in an electromagnetic field according to their mass to charge ratio (m/z), and finally transferred to a detector. The detected signal, usually proportional to the number of ions, is plotted against the corresponding m/z value on a *mass spectrum*. A pair of detected m/z value and the corresponding signal intensity is called a *peak*. The signal intensity is often referred to as *ion current* [39, 10].

The m/z value can be used to infer the chemical composition of molecules (see e.g. [40]), but it does not give information about its chemical structure. To gain insight into the latter, several measurement steps are performed in a technique called Tandem Mass Spectrometry (Tandem MS). After each step, a range of m/z value is selected, and ions from that range are subjected to fragmentation before the next measurement. The mass spectrum obtained from the n-th measurement is referred to as an MS^n spectrum.

Even though the MS^1 spectrum is recorded prior to any fragmentation, a single compound can give rise to several peaks. This is due to the natural occurrence of *isotopes*, i.e. atoms with the same number of electrons and protons, but different numbers of neutrons. Molecules which differ only in their isotopic compositions are termed *isotopologues*. A group of peaks corresponding to isotopologues of a single molecule is referred to as an *isotopic envelope* (c.f. Fig. 3.1).

Tandem MS can be used to identify the molecule under study. There are two main approaches to this task: *de novo* sequencing and database search. The first one strives to identify the elemental composition and/or structure of the molecule purely based on the mass spectrum of fragments. The second one searches a database of mass spectra obtained from known molecules to find the most similar one [41, 42, 43].

To be able to search for a similar spectrum, either a similarity or a distance measure needs to be employed. There are two main groups of such measures. The first one relies on the number of *matching peaks*. Two peaks are said to match if their m/z values differ by less than a given threshold. An example of such measure is the Jaccard score, equal to the number of matching peaks divided by the number of distinct peaks in both spectra. The second group of measures takes into account both the location and the intensities of peaks. An example of such measure is the Euclidean distance or the correlation coefficient [41, 42].

Both groups are similar in the sense that they compare peaks with the same m/z value. As a consequence, they are highly sensitive to even the slightest differences in chemical formulas. For example, *apigenin* ($C_{15}H_{10}O_5$) and *quercetin* ($C_{15}H_{10}O_7$)

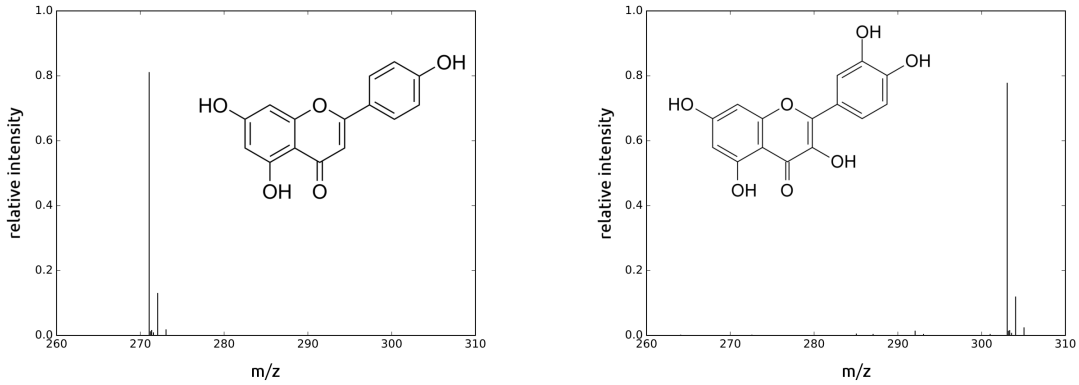

FIGURE 3.1: Molecular structures and MS^1 spectra of apigenin (left) and quercetin (right) showing their isotopic envelopes. Peak intensities have been normed to sum to 1. The mass spectra have been downloaded from the MassBank database (MassBank IDs: TY000164, TY000119).

are two molecules which differ by two oxygen atoms (see Fig. 3.1). Even though this difference is relatively small compared to the overall atom count, the MS^1 spectra contain no matching peaks. Consequently, the discussed measures do not detect any similarity between these molecules. Some approaches make a preprocessing of spectra to infer an optimal pairwise matching of peaks before computing the similarity [44]. This, however, requires an additional computational step, increasing the computational complexity as well as the risk of inaccurate results.

Apart from the differences in chemical compositions of molecules, naturally occurring measurement inaccuracies also hinder the capabilities of the common approaches to meaningfully compare mass spectra. In order to overcome these problems, in this Chapter we will investigate a spectrum dissimilarity measure based on the theory of optimal transport, with the aim to develop a method that allows us to compare spectra with different resolutions and is robust to measurement errors in the mass domain. The measure is based on the concept of transporting the ion current (i.e. the signal) between the spectra. The idea behind the measure is to transport the ion current from one spectrum onto the other and quantify the distance that the current needs to travel. The dissimilarity between the two spectra is equal to the minimal distance in m/z domain that needs to be traveled in order to fully transform one spectrum into the other. This makes it possible to express the distance between spectra in Daltons. The particular distance investigated in this Dissertation is known in the field of probability theory as the (first) Wasserstein distance [3], and in the field of image processing as the Earth Mover's distance [45]. Under certain assumptions, it can be computed in time linear in the number of distinct peaks in both spectra.

The Wasserstein distance makes it possible to more accurately reflect the differences in chemical compositions of the molecules. In particular, the Wasserstein distance between MS^1 spectra of two ions with the same charge is approximately equal to the molecules' mass difference. For example, the distance between the MS^1 spectra from Fig. 3.1 is equal to 31.48 Da, while the difference in their masses is equal to 32.19 Da. With this interpretation, we can consider as fairly similar those spectra for which the Wasserstein distance is less than one hydrogen mass. Apart from quantifying the dissimilarity, the computed transport of ion current allows for a matching of corresponding peaks in the compared spectra, which can aid in the detection of

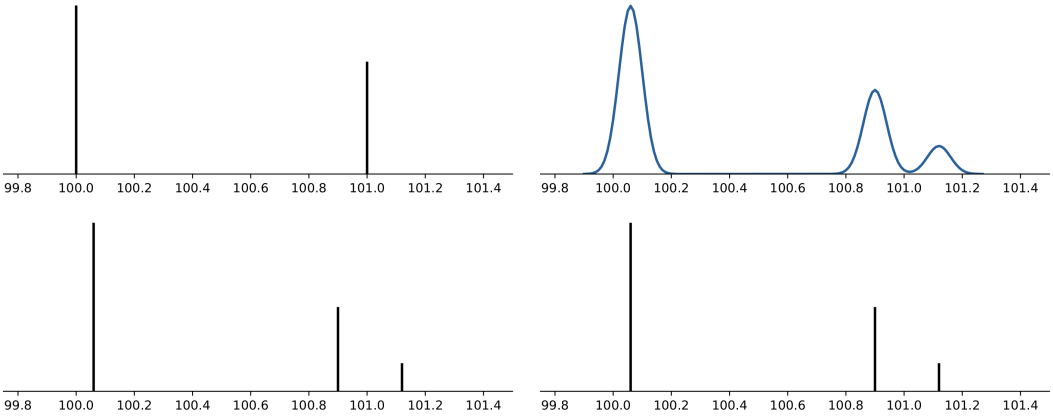

FIGURE 3.2: Example values of the Wasserstein distance between mass spectra. Left: Two spectra in centroid mode, with the Wasserstein distance between them being equal to 0.078 Da. Right: A spectrum in profile mode and a corresponding centroided one, with the Wasserstein distance equal to 0.032 Da. Both values, being less than 1 hydrogen mass, indicate a fairly high degree of similarity, even though no peaks match in the first example and the spectra are in different modes (profile vs centroid) in the second one.

differences in elemental composition and chemical structure (see Fig 3.3).

Some other examples of the values of the Wasserstein distance between pairs of spectra are shown in Fig. 3.2. The spectra are artificially constructed in order to provide simple and clear examples. Consequently, we have purposefully neglected several important phenomena occurring in actual spectra, such as the background noise, which will be dealt with later on in this Dissertation. Worked examples of how to compute this distance between pairs of spectra are also provided later on in this Chapter.

In the right panel of Fig. 3.2, we consider a pair of corresponding spectra in profile and centroid mode. Even though such spectra cannot be meaningfully compared using conventional measures, such as the Euclidean distance, the Wasserstein distance between them has a small value of 0.032 Da. As shown later on in this Chapter, this value reflects the different resolutions (i.e. peak widths) of those spectra (where the centroid-mode spectrum is assumed to have an infinite resolution).

The left panel of Fig. 3.2 presents a pair of spectra with no matching peaks. Pointwise measures, such as signal correlation, do not capture any similarity between them (note that, in certain applications, this may be a desirable phenomenon). The Wasserstein distance, on the other hand, again has a small value of 0.078 Da, indicating that those spectra are very similar in terms of the m/z differences of peak positions. The fact that the peaks do not match, however, is still reflected by the Wasserstein distance, since it is over twice as large as between the spectra in the right panel.

An alternative interpretation of the Wasserstein distance is the minimal amount of distortion such as shifting, broadening and narrowing of peaks that is required to transform one of the spectra into the other. This approach is naturally robust to small distortions in the m/z or intensity measurements, and has no requirements as to the accuracy or resolution of the measurement. In particular, to the best of our knowledge, it is the only similarity measure capable of an accurate comparison of profile and centroided spectra.

The contents of this Chapter. In the following Sections, we describe the mathematical formalism behind the Wasserstein distance and expand on its interpretation in the context of mass spectrometry. Then, we show worked examples of computation of this distance that further illustrate its properties related to mass spectra. Next, we quantitatively study the correlation between the structural similarity of chemical compounds and the Wasserstein distance between their MS^2 spectra. We finish this Chapter by discussing some practical considerations when using the Wasserstein distance to compare profile spectra. In particular, we focus on how such spectra are represented and stored on computers, leading to a risk of inaccurate distance values when a simple algorithm to compute them is used. We show how to circumvent this problem and how to better estimate the distance in practice by using resampling methods.

3.1 The Wasserstein distance

Let μ and ν be any two mass spectra to be compared. In order to simplify the description of the Wasserstein distance and the concept of transporting signal between spectra, we will assume for now that both spectra are centroided. The case of profile spectra will be discussed later on.

In order to compare μ and ν, we aim to transport all the signal from one spectrum to the other and quantify the minimal amount of distance in the mass domain over which the signal needs to be transported. An example of such transport is conceptually visualized in Fig. 3.3. We do not differentiate between a source and a target spectrum, so that the final distance is symmetrical—in other words, we will assume that transforming μ into ν inflicts the same cost as transforming ν into μ.

We assume that all the considered spectra are normalized by their total ion current, so that the peak intensities of each spectrum (including the peaks arising from background noise or contaminants) sum up to 1. Note that such normalization may be meaningless from a data analyst's point of view, because the normalizing factor includes the noise intensity. However, it is often used for technical reasons when developing computational methods, because it allows for treating mass spectra as probability measures and using the tools of probability theory to analyze them (see, e.g., [46]). Accordingly, throughout this Dissertation, unless stated otherwise, we will assume that spectra are normalized and we will treat (centroided) mass spectra as (discrete) probability distributions on the real line $\mathbb{R}$, with $\mu(x)$ denoting the intensity at the m/z value x in spectrum μ, and $x \in \mu$ denoting that x belongs to the support of μ, i.e. that $\mu(x) > 0$.

Let $\gamma(x, y)$ be the amount of signal transported between the point x in spectrum μ and the point y in spectrum ν. The function γ is referred to as a *transport plan*. Any transport plan needs to satisfy the following properties [3, 4]:

$$\sum_{y \in \nu} \gamma(x, y) = \mu(x), \qquad \sum_{x \in \mu,} \gamma(x, y) = \nu(y). \tag{3.1}$$

The first of the above properties means that all the signal intensity $\mu(x)$ needs to be transported somewhere into the spectrum ν. Similarly, the second property means that the intensity $\nu(y)$ is fully filled by the ion current coming from the spectrum μ. Naturally, the transport plan also needs to be non-negative, $\gamma(x, y) \geq 0$, as we cannot transport negative intensity. As a consequence, γ can be interpreted as a joint probability distribution (also referred to as a coupling) of the two probability

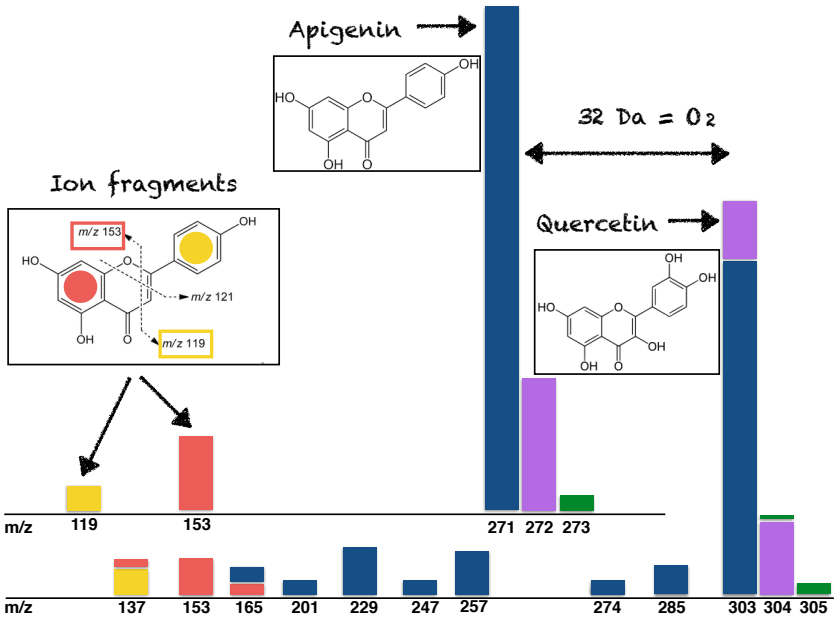

FIGURE 3.3: The optimal ion current transport plan for MS^2 spectra of apigenin (top) and quercetin (bottom), fragmented using 30 eV collision energy. The colors on the quercetin mass spectrum correspond to the origin of the transported ion current. The isotopic envelope of quercetin is shifted by 32 Da, i.e. the mass of two oxygen atoms.

distributions μ and ν. In turn, any function satisfying the above properties is a valid transport plan.

The cost of a given transport plan is the total distance traveled by the ion current. This is calculated by multiplying the distance between the points x and y by the amount of signal traveling between them, and summing over all peaks in both spectra:

$$\sum_{\substack{x\in\mu\\ y\in\nu}} |x-y|\gamma(x,y).$$

We assume that the transport is not directed, so that transporting the intensity in the direction of increasing mass inflicts the same cost as in the other direction. This is formally expressed by the absolute difference between the points in the above equation. Note that the dependence of the sum on the spectra μ and ν is only implicitly expressed through the function γ and Equations (3.1).

Denote by Γ the space of all possible transport plans (couplings) between μ and ν. The Wasserstein distance between two spectra, $W(\mu,\nu)$, is defined as the minimal cost of transport over all possible transport plans from the space Γ [3, 47, 4]:

$$W(\mu,\nu) = \min_{\gamma\in\Gamma} \sum_{\substack{x\in\mu\\ y\in\nu}} |x-y|\gamma(x,y), \tag{3.2}$$

The function W defined this way satisfies the mathematical properties of a distance function, that is, non-negativity $W(\mu,\nu) \geq 0$, symmetry $W(\mu,\nu) = W(\nu,\mu)$ and the triangle inequality $W(\mu,\nu) \leq W(\mu,\zeta) + W(\zeta,\nu)$. This can be easily understood intuitively by noting that we cannot transport the signal over negative distances, that the definition is symmetric with respect to μ an ν, and that optimal transport of the signal from μ to ν needs to be less costly than first transporting signal from μ to

ζ, and then from ζ to ν. For formal proofs, we refer the reader to a book by Villani [3] or Santambrogio [4].

Since the distance is symmetric, there is no designated *source* spectrum of the transported signal, nor a *target* spectrum to which the signal is transported. However, when depicting or discussing transport plans for $W(\mu, \nu)$, we adopt a convention that the signal is transported from one spectrum to the other.

Although the formulation of the distance may seem baffling, it turns out that under reasonable assumptions the algorithm to compute $W(\mu, \nu)$ has a linear time complexity [47]. The only requirement for this is that the input spectra are available as lists of sorted m/z values and the corresponding signal intensities. Since raw spectra, in both centroid and profile mode, are usually stored this way on computers, the Wasserstein metric is computationally equally efficient to the Jaccard score and the Euclidean distance. The algorithm to compute the Wasserstein distance between a pair of spectra is based on the following theorem [4, 47]:

Theorem 3.1. *Let μ and ν be two probability measures on the real line $\mathbb{R}$. Let M and N be the cumulative distribution functions (CDFs) of μ and ν respectively. Then,*

$$W(\mu, \nu) = \int_{\mathbb{R}} \Big| M(t) - N(t) \Big| dt. \tag{3.3}$$

In mathematical terms, Theorem 3.1 states that the cumulative distribution function is an isometry between the Wasserstein space (in our case, the space of discrete probability measures with finite supports, equipped with the W distance) and the L^1 space of functions. Theorem 3.1 can be applied to spectra normalized by their total ion current, since such spectra can be interpreted as probability measures. It captures both the case of centroided and profile spectra (in which case we normalize the spectrum by the integral of its signal). In the case of centroided spectra, the cumulative distribution functions are step functions, and the formula can be considerably simplified.

Theorem 3.2. *Let μ, ν be two centroided mass spectra normalized by the total ion current. Let $S = \{s_1, s_2, \ldots, s_n\}$ be an ordered list of all distinct masses in both spectra. Let M and N be the cumulative distribution functions (CDFs) of μ and ν, i.e. $M(t) = \sum_{x \leq t} \mu(x)$. Then,*

$$W(\mu, \nu) = \sum_{i=1}^{n-1} (s_{i+1} - s_i) \Big| M(s_i) - N(s_i) \Big|. \tag{3.4}$$

Formula (3.4) admits a simple interpretation. Observe that $M(s_i) - N(s_i)$ is the difference in the ion currents on the left hand side of point s_i in both spectra, which needs to be transported either to or from the point s_{i+1} in order to achieve balance. This amount of ion current is then transported over a distance equal to $s_{i+1} - s_i$, and such "partial costs" of transport are summed over all points of both spectra.

Based on Theorem 3.2, we can easily compute the distance $W(\mu, \nu)$ for centroided, normalized mass spectra. A common way of representing such a spectrum is a peak list, i.e. a list of pairs (x_i, p_i) such that x_i are in increasing order and represent m/z values of peaks with intensities p_i. Algorithm 1, adapted from [47], shows how to efficiently compute W given two such lists of peaks. It is based on the observation that the absolute difference between the cumulative distribution functions of mass spectra, $|M - N|$, is a step function, and therefore it is easily integrable numerically. However, the algorithm does not require an explicit calculation of the CDFs.

The runtime of Algorithm 1 is $\mathcal{O}(n + m)$, where n and m are the lengths of the peak lists of both spectra. This can be proved by noting that in each iteration of

the main loop either i or j is incremented, no index variable will ever exceed the length of the corresponding list, and the algorithm terminates when both indices have reached the end of their respective lists.

Algorithm 1: Computation of Wasserstein distance between two spectra

Data: Two lists, L_1, L_2, of pairs (x, p), containing the lists of peaks of respective spectra
Result: W distance between given spectra

$i \leftarrow 0$; $j \leftarrow 0$;
$ret \leftarrow 0.0$; $\gamma \leftarrow$ an empty transport scheme
$n \leftarrow \text{length}(L_1)$; $m \leftarrow \text{length}(L_2)$
while $i < n \vee j < m$ **do**
  $d \leftarrow \min(L_1[i].p, L_2[j].p)$
  $ret \leftarrow ret + d \cdot |L_1[i].x - L_2[j].x|$
  $L_1[i].p \leftarrow L_1[i].p - d$
  $L_2[j].p \leftarrow L_2[j].p - d$
  $\gamma(i, j) \leftarrow d$
  if $0 = L_1[i].p$ **then**
    $i \leftarrow i + 1$
  else
    $j \leftarrow j + 1$
  end
end
The variable ret contains the Wasserstein distance and γ the transport plan.

3.2 Worked examples.

In order to illustrate the computation of the Wasserstein distance and make the concept more intuitive to the reader, we present two worked examples below.

3.2.1 Example 1.

Consider two abstract spectra, μ and ν, such that μ is concentrated at 100.5 Da (i.e. $\mu(100.5) = 1$) and ν is distributed evenly over 98, 99, 100, 101 and 102 Da (i.e. $\nu(98) = \nu(99) = \nu(100) = \nu(101) = \nu(102) = 0.2$). Obviously, those spectra do not correspond to any actual ions. They only serve for an easy illustration of the computation of the optimal signal transport. The transport of the signal is illustrated in Fig. 3.4. The cumulative distribution functions M and N of μ and ν, respectively, are given by

$$M(t) = \begin{cases} 0.0 & : \quad t < 100.5, \\ 1.0 & : \quad 100.5 \leq t, \end{cases}$$

$$N(t) = \begin{cases} 0.0 & : \quad t < 98, \\ 0.2 & : \quad 98.0 \leq t < 99, \\ 0.4 & : \quad 99.0 \leq t < 100.0, \\ 0.6 & : \quad 100.0 \leq t < 101.0, \\ 0.8 & : \quad 101.0 \leq t < 102.0, \\ 1.0 & : \quad 102.0 \leq t. \end{cases}$$

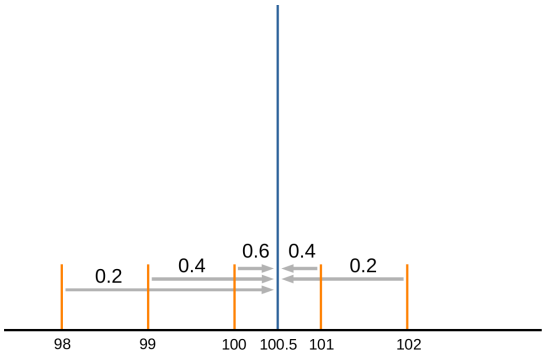

FIGURE 3.4: An example of an optimal transport scheme between two abstract spectra μ and ν, depicted on a single graph in blue and orange respectively. The signal of the spectrum μ is concentrated at the point 100.5, while ν is distributed evenly at five points from 98 to 102. Grey arrows depict the transport of the ion current. Numbers above arrows show the proportions of the ion current flowing between neighboring peaks, i.e. the absolute difference of the cumulative distribution functions.

The list of all distinct masses, S, is now equal to $(98, 99, 100, 100.5, 101, 102)$, and the difference between the cumulative distribution functions, $N(t) - M(t)$, indicating the ion current imbalance at point t, is equal to

$$N(t) - M(t) = \begin{cases} 0.0 & : \quad t < 98.0, \\ 0.2 & : \quad 98.0 \leq t < 99.0, \\ 0.4 & : \quad 99.0 \leq t < 100.0, \\ 0.6 & : \quad 100.0 \leq t < 100.5, \\ -0.4 & : \quad 100.5 \leq t < 101.0, \\ -0.2 & : \quad 101.0 \leq t < 102.0, \\ 0.0 & : \quad 102.0 \leq t. \end{cases}$$

From the above equation, we can read out that 0.2 of the signal is transported from the point 98.0 to 99.0; 0.4 of the signal is transported from 99.0 to 100.0; 0.6 from 100.0 to 100.5. Next, as the sign of the imbalance changes, so does the direction of transport between neighboring points, and so 0.4 of the signal is transported from 101.0 to 100.5, and 0.2 of the signal from 102 to 101. Finally, the ion currents of both spectra balance out at the point 102.

The final distance can be computed by taking the absolute values of the ion current imbalance and multiplying them by the distance travelled, so that $W(\mu, \nu) = 0.2 \cdot 1 + 0.4 \cdot 1 + 0.6 \cdot 0.5 + 0.4 \cdot 0.5 + 0.2 \cdot 1 = 1.3$ Da.

3.2.2 Example 2.

Consider a spectrum μ concentrated at 100 Da, and a profile spectrum ν consisting of a single Gaussian peak centered at 100 Da with a standard deviation σ. We therefore have $\mu(100) = 1$ and

$$\nu(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-100)^2}{2\sigma^2}}.$$

Let Φ be the cumulative distribution function of the standard Gaussian random variable:

$$\Phi(t) = \int_{x \leq t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

The CDF of ν is then given by $N(t) = \Phi((t-100)/\sigma)$. The exact value of the Wasserstein distance between μ and ν is then equal to

$$\begin{aligned} W(\mu,\nu) &= \int |M(t) - N(t)|dt = \int_{x\leq 100} |N(t)| + \int_{x\geq 100} |N(t) - 1| \\ &= \int_{x\leq 100} N(t)dt + \int_{x\geq 100} (1 - N(t))dt. \end{aligned}$$

After plugging in Φ in the above integrals and substituting the variables, one arrives at

$$W(\mu,\nu) = \sigma \int_{x\leq 0} \Phi(x)dx + \sigma \int_{x\geq 0} (1 - \Phi(x))dx.$$

Using the relation $\Phi(-x) = 1 - \Phi(x)$, we get

$$W(\mu,\nu) = \sigma \int_{x\leq 0} \Phi(x)dx + \sigma \int_{x\geq 0} \Phi(-x)dx = 2\sigma \int_{x\leq 0} \Phi(x)dx.$$

As the antiderivative of $\Phi(x)$ is $x\Phi(x) + \phi(x) + C$, where ϕ is the density of a standard Gaussian variable, we arrive at

$$W(\mu,\nu) = 2\sigma\phi(0) = 2\sigma\frac{1}{\sqrt{2\pi}} = \sigma\sqrt{\frac{2}{\pi}}.$$

Therefore, we obtain a simple formula for the distance between a centroided spectrum and a corresponding profile spectrum with Gaussian peaks of standard deviation σ. This result explains the small distance between spectra in the right panel of Fig. 3.2.

3.3 Quantitative properties of the Wasserstein distance between mass spectra

To quantitatively analyze the properties of the Wasserstein metric when applied to mass spectral data, we have analyzed two sets of spectra obtained from the MassBank database [48]. In both cases, we have compared the performance of the Wasserstein distance with two other popular approaches: the Euclidean distance and the Jaccard score (i.e. the ratio of matching peaks to the total number of different peaks in both spectra). When analyzing those two measures, the spectra were binned to 0.01 Da resolution to increase the number of matching peaks and decrease their sensitivity to small measurement errors. No binning was performed during the analysis of the Wasserstein metric.

The first test was based on 615 MS^1 ESI-QTOF spectra with positive ionization mode. The goal of comparing MS^1 spectra was to verify the correlation between distance values and the difference in mass of the molecules. The spectra have been compared pairwise, resulting in 188805 pairs. These pairs were then used to compute the Spearman's rank correlation between the distance and the absolute difference between masses of the corresponding molecules. The results are summarized in Table 3.1.

As expected, the metrics based on peak matching are sensitive to mass differences, and therefore less correlated than the Wasserstein distance. Note that for the Wasserstein and Euclidean distance the correlation is expected to be positive, while for the Jaccard similarity metric it is expected to be negative. Surprisingly, we have

found a negative correlation between the mass difference and the Euclidean distance.

MS^1 spectra

	M	**W**	**J**	**E**
M	1.00	0.89	-0.07	-0.37
W	—	1.00	-0.08	-0.22
J	—	—	1.00	-0.17
E	—	—	—	1.00

MS^2 spectra

	T	**RW**	**J**	**RE**
T	1.00	-0.41	0.22	-0.24
RW	—	1.00	-0.21	0.43
J	—	—	1.00	-0.11
RE	—	—	—	1.00

TABLE 3.1: Spearman's rank correlations between the Wasserstein distance, Jaccard score, Euclidean distance, and either the absolute mass difference or Tanimoto similarity of chemical structures. M, absolute mass difference; W, Wasserstein distance; J, Jaccard score; E, Euclidean distance; T, Tanimoto similarity; RW, relative Wasserstein distance; RE, relative Euclidean distance (see text).

The second test was based on MS^2 ESI-QTOF spectra with positive ionization mode. Here, the goal was to investigate the relationship between the distance values and the molecules' structural similarity. Note that the Wasserstein distance is particularly sensitive to the fragmentation intensity—two MS^2 spectra obtained for a given molecule can have a large distance if there is a significant difference in the intensity of fragments. To account for that, we have selected a subset of 473 MS^2 spectra for different molecules in which the precursor peak had around 10% relative intensity. This resulted in 111628 pairs of spectra. For each pair of spectra, we have computed the Wasserstein, Jaccard and Euclidean metrics. Next, we have computed the Tanimoto similarity between the structures of the corresponding molecules, based on the Morgan circular fingerprints [49, 50]. The fingerprints have been computed using the RDKit package (`http://www.rdkit.org`), with the radius of 2 and the default set of the feature-based invariants. The results are summarized in Table 3.1.

Note that the selected set of spectra comes from a diverse set of molecules. In particular, the mean mass is 310 Da, while the standard deviation is 160 Da. This poses a problem for the Wasserstein metric, as a pairs of small molecules will yield small distances regardless of the structural similarity. To account for this, we have divided the distance by the product of masses of the analyzed molecules. Without this correction, the correlation between the Wasserstein metric and the Tanimoto similarity drops to -0.22. This procedure also improved the correlation between the Tanimoto similarity and the Euclidean distance, but not the Jaccard score. We refer to the distances with this correction as *relative* distances.

The detailed relationship between the relative Wasserstein distance and the Tanimoto similarity is depicted in Fig. 3.5. For comparison, the Figure also shows the relationship between the Tanimoto similarity and the Jaccard score. Note that all compounds with high Tanimoto similarity have small relative Wasserstein distances. However, this relative distance is much more variable for compounds with low similarity. This frequent occurence of molecules with highly divergent structures but similar MS^2 spectra decreases the extent to which the Wasserstein distance correlates with the Tanimoto structural similarity.

The experiments show that the Wasserstein distance outperforms the Jaccard score and the Euclidean distance in terms of correlation with the molecules' mass difference in MS^1 spectra and their chemical structure similarity in MS^2 spectra. However, at this moment the Wasserstein distance should be applied only to MS^2 spectra with similar proportions of precursor molecules, and preferably obtained

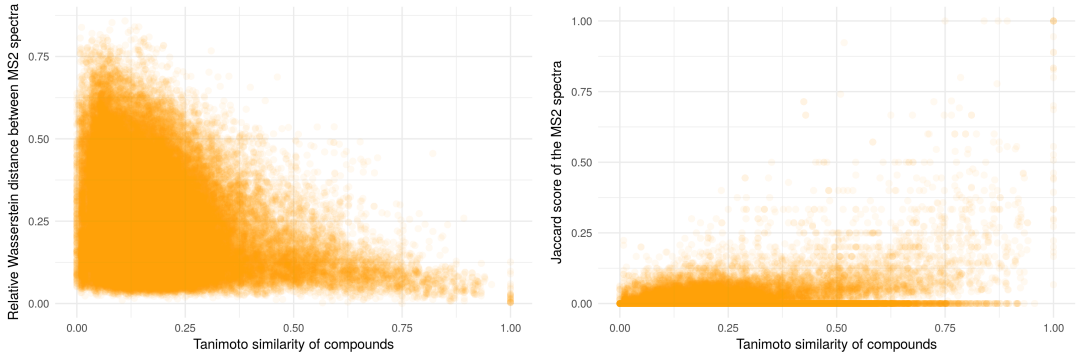

FIGURE 3.5: The relationship between MS^2 spectra and the structural similarity according to the relative Wasserstein distance (left) and the Jaccard score (right).

from compounds of similar mass. The results so far are optimistic, but more work needs to be done in order to generalize the Wasserstein distance so that it can be applied to a broader class of mass spectra.

3.4 Some qualitative properties of the Wasserstein distance between mass spectra

To give the reader some additional intuitions behind the Wasserstein distance, in this short Section we discuss some of its qualitative properties when applied to mass spectra. Some of the points discussed here reflect the computational experiments performed in the previous Section.

For MS^1 spectra of two molecules, the Wasserstein distance is approximately equal to the absolute mass difference of the molecules. The other main factor that influences the distance in this case is the presence of measurement inaccuracies. Note, however, that this influence remains small as long as the inaccuracies in intensity measurements are small compared to the corresponding peak intensities. In case of spectra of mixtures of compounds, the relation between the masses of molecules and the Wasserstein distance is more complex. However, the absolute difference of the centers of masses of two spectra always gives a lower bound for the distance.

Usually, some inaccuracy in both the intensity and the mass measurement is present. Naturally, the latter poses a major problem for measures based on peak matching. On the other hand, the Wasserstein distance is not significantly influenced by small mass measurement errors—instead, the imprecise measurement simply gets shifted to match its theoretical counterpart.

The implicit assumption of this metric, which may not be desirable in some applications, is that the mass difference reflects chemical difference. Therefore, two molecules differing by an OH group are assumed to be more similar to each other than two molecules differing by a C_2H_5 group. It is possible to relax this assumption by applying a different metric in the mass domain, say $c(x, y)$, in the definition of the Wasserstein distance:

$$W_c(\mu, \nu) = \min_{\gamma \in \Gamma(\mu,\nu)} \sum_{i=1}^{n} \sum_{j=1}^{m} \gamma(i, j) c(x_i, y_j).$$

An important caveat in this case is that treating all modifications as equivalent may lead to unexpected results—notably, a protein being treated as a single carbon atom with an extremely large modification. Furthermore, using other distances in the mass domain may lead to difficult optimization problems. The use of absolute difference, $|x_i - y_j|$, avoids costly optimizations in the space of all possible transport plans thanks to Theorem 3.2.

If the two molecules differ by a modification which does not change their fragmentation pattern, then the Wasserstein distance between their MS^2 spectra will not exceed the weight of the modification. This follows from the observation that the modification is present only in some of the fragments, which account to a fraction of the total intensity in the spectrum. Note that this is a highly idealized example, since modifications may significantly change the fragmentation patterns and inflict a greater influence on the Wasserstein distance. In general, however, this distance cannot exceed the mass of the heavier molecule.

Lastly, the structure of the optimal transport plan is highly sensitive to chemical noise, i.e. the presence of unexpected molecules. Recall that all the intensity from one spectrum needs to be used to explain all the intensity of the second spectrum. Therefore, if one of the analyzed spectra contains an additional peak, some of the intensity from the first spectrum needs to be used to explain it. This may lead to global changes in the structure of the optimal transport plan. It follows that in the presence of chemical noise, the Wasserstein distance may not reflect the similarity between the analyzed compounds.

3.5 Handling profile spectra in practice.

Although, in principle, profile spectra are continuous functions, they are usually represented as finite lists of mass and intensity pairs. In this Section, we show that, to compute the Wasserstein distance, such lists can be simply treated as centroid spectra. Under certain assumptions, Theorem 3.2 gives an accurate approximation of the cost of the optimal transport plan.

Assume we are given a finite list of mass and intensity pairs, (s_i, I_i) for $i = 1, 2, \ldots, n$, approximating a profile spectrum μ. Assume as well that we have a constant spacing between consecutive intensity measurements, so that $s_{i+1} - s_i = 1/n$ for $i = 1, 2, \ldots, n-1$. Treating μ in the same way as a centroided spectrum, we compute its cumulative distribution function as $\hat{M}(t) = \sum_{x_i \leq t} I_i / \sum I_i$, while its true cumulative distribution function is given by $M(t) = \int_{x \leq t} \mu(x)dx / \int \mu(x)dx$. Now, we have

$$\hat{M}(t) = \frac{\sum_{x_i \leq t} I_i}{\sum I_i} = \frac{\frac{1}{n}\sum_{x_i \leq t} I_i}{\frac{1}{n}\sum x_i I_i} \approx \frac{\int_{x \leq t} \mu(x)dx}{\int \mu(x)dx} = M(t),$$

where the approximation is based on the fact that the sums on the left hand side are Riemann sums of the integrals on the right hand side.

The assumption of an uniform signal sampling (i.e. a constant spacing between consecutive measurements) is crucial for the approximation to work. When this is not satisfied, a spectrum needs to be resampled prior to computing the distance. However, care needs to be taken to resample the spectrum in a way that does not lead to a loss of data or an introduction of additional noise signal. We propose to use a resampling algorithm based on a piecewise-linear interpolation of profile spectra, presented in Subsection 3.5.1. Note, however, that a resampled spectrum should always be checked at least visually for the presence of any introduced artifacts, and the total ion current of the original and the resampled spectra should be compared.

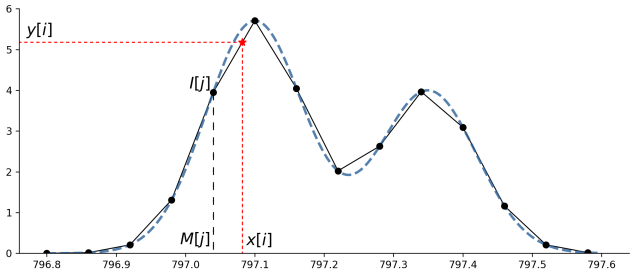

FIGURE 3.6: A graphical description of piecewise-linear interpolation of profile mass spectra. The blue dashed line shows the true, unobserved signal intensity. The black points show the intensity measurements $I[j]$, corresponding to masses $M[j]$, observed in a profile mass spectrum. The black lines show the interpolated intensities. The red dashed lines show a result of approximation of the true signal, $y[i]$, at the point $x[i]$. The indexes i and j correspond to the ones used in Algorithm 2. Note that in actual spectra the measured points are more densely spaced, resulting in a much better interpolation.

We also note that further research on the Wasserstein distance between profile spectra sampled on not uniform m/z arrays would likely allow for an easier and more accurate computation.

Naturally, an alternative way to analyze a profile spectrum with the means of optimal transport is to first convert it to the centroided mode. This can be especially useful when resampling is infeasible, e.g. due to increased data size. However, a proper centroiding needs to accurately reflect the intensities in a profile spectrum, which is a highly non-trivial task in practice due to different peak widths, presence of background noise, varying peak shapes (caused e.g. by merging of closely located peak) and non-zero baseline signal.

There are numerous approaches to signal centroiding, including methods based on simple numerical integration, as well as more sophisticated (and computationally costly) approaches based on wavelet transforms [51]. In Subsection 3.5.2, we present a simple centroiding algorithm, which is based on the integration approach but robust against multiple sources of error. We will use the techniques of resampling and centroiding to analyze profile spectra in Chapter 5, where we discuss the problem of regression of noisy mass spectra.

3.5.1 Piecewise-linear interpolation of spectra.

For certain types of mass spectrometers, the width of a measured signal increases with the m/z value. In order to decrease the volume of data, typically profile spectra are stored as lists of m/z and intensity pairs, where the distance between consecutive m/z values increases with mass. This increase is chosen so that each signal in the spectrum is represented by an approximately constant number of m/z, intensity pairs. However, such a non-uniform sampling of intensity distorts the value of the Wasserstein distance computed using Equation 3.4 from Theorem 3.2. In order to correct that, we can interpolate each spectrum by a piece-wise linear function, so that the interpolated intensity in a given point is a weighted average of the neighboring measured intensities.

Although the terminology may sound obscure, the idea behind the piecewise-linear interpolation is straightforward: we join pairs of consecutive points by straight lines and use those lines to approximate the signal intensities at any given set of points at

the mass axis. Algorithm 2 shows a numerically optimized way to achieve this. The idea behind the algorithm is visualized in Fig. 3.6.

In practical implementations of Algorithm 2, we add another constraint that if the distance between the m/z value at which we interpolate and any of it's neighboring points is further than a given threshold, the interpolated intensity is set to 0. This is done in order to handle broad regions with no recorded intensity present in many experimental spectra.

Algorithm 2: Piecewise-linear interpolation of profile mass spectra.

Data: A sorted list of m/z measurements M, a list of corresponding intensity measurements I, a sorted list of m/z values in which to interpolate the intensity x.

Result: A list of interpolated intensity values corresponding to x.

1 Initialize a zero-filled list y of the same length as x.
2 Set i as the index of the first mass from x that lies between $M[0]$ and $M[1]$.
3 Set $j = 0$.
4 **while** $j < length(M) - 1$ **do**
5 **while** $i < length(x)$ *and* $x[i] < M[j]$ **do**
6 Set $y[i] = I[j+1] - (M[j+1] - x[i])(I[j+1] - I[j])/(M[j+1] - M[j])$.
7 Set $i = i + 1$.
8 **end**
9 Set $j = j + 1$.
10 **end**
11 The variable y contains signal intensities corresponding to m/z values of x.

3.5.2 Centroiding the profile spectra.

There are numerous available algorithms for peak centroiding, both open-source and proprietary. We have decided not to use the latter in this work, as it is not possible to determine the way they work and therefore have confidence in their results. Instead, we have implemented a simple algorithm that detects local maxima of intensity and integrates peaks within regions delineated by an intensity threshold, expressed as a proportion of the apex intensity. Algorithm 3 describes the basic idea behind our approach in pseudo-code. Below, we explain the rationale behind it and describe additional details and constraints used in the practical implementation.

After peak location is determined, typically by detecting a local maximum of the signal, there are two main approaches to obtain peak intensity: either as the apex intensity, or as the peak area [10]. For many mass spectrometers, only the latter gives a correct result, as the peak width increases with the m/z value. On the other hand, approaches based on the apex intensity are simpler and less prone to errors. Therefore, they are useful for the analysis of spectra with a small range of m/z values, where peaks have a similar width.

In our implementation, we set additional constraint on maximal peak width. If the width of the region in which the intensity is to be integrated exceeds this threshold, the peak is discarded.

Note that, when two peaks overlap, they may share their integration region. In that case, the centroided m/z and intensities of such peaks are identical. Since we keep a set of peaks instead of a list (note the line 1 of Algorithm 3), such a peak cluster

is represented by one peak. This essentially merges highly overlapping peaks into one. Separating highly overlapping peak clusters requires much more sophisticated approaches (see e.g. [52]).

Algorithm 3: Centroiding of profile mass spectra.

Data: A sorted list of m/z measurements M, a list of corresponding intensity measurements I, apex intensity proportion t.
Result: A list of (m/z, intensity) pairs of peaks.

1 Initialize an empty set $\mathcal{L}$.
2 Set $\mathcal{P}$ as a list of m/z values of the local maxima of intensity from I.
3 **for** m_p *in* $\mathcal{P}$ **do**
4 Set ι as the signal intensity corresponding to the point m_p.
5 Set $t_p = t\iota$ as the intensity threshold.
6 Identify the nearest m/z values $m_1 \leq m_p \leq m_2$ that correspond to intensities below t_p.
7 Use linear interpolation to approximate m/z values $m_1^* \leq m_p \leq m_2^*$ that correspond exactly to the intensity t_p.
8 Use the trapezoid rule to integrate the intensity in the interval $[m_1^*, m_2^*]$.
9 Use the trapezoid rule to obtain the peak centroid m/z by integrating the intensity multiplied by the corresponding m/z values in the interval $[m_1^*, m_2^*]$ and dividing the result by the integrated intensity.
10 Add a tuple of peak centroid m/z and integrated intensity to the set $\mathcal{L}$.
11 **end**

In our implementation, when we identify the integration region in line 6 of Algorithm 3, we additionally require that the intensity is monotonically decreasing with the distance from the apex. When we detect that the signal intensity starts to increase, we discard the peak. This ensures that only the highest peak of any peak cluster is considered. It also allows us to discard numerous small peaks that occur due to background noise.

3.6 Summary of the Chapter

In this Chapter, we have presented and analyzed a new approach to the comparison of mass spectra based on the Wasserstein metric. This metric is a well-established and well-studied concept used in both probability theory and image processing field. Compared to the current approaches to comparison of spectra, it is more robust to measurement errors and better reflects the differences in chemical compositions of ions. In the MS^2 spectra of similar compounds, the Wasserstein distance reflects differences in both chemical structure and fragmentation intensities.

The extensive mathematical research on the topic of optimal transport has resulted in powerful theorems that express the Wasserstein distance as a computationally feasible integral. The optimal cost of transporting the signal between spectra can be computed in a straightforward manner without the need for a numerical optimization in the space of all possible transport plans.

Further research. The Wasserstein distance quantifies the similarity of spectra only in terms of the difference in m/z locations of signals. A drawback of this approach is that it is relatively sensitive to differences in peak intensities. This is further emphasized by the different nature of measurement uncertainties in the mass and the

intensity domain, caused by the fact that mass spectrometers measure the m/z and the number of ions differently. Although accounting for this phenomenon would be desirable, it is highly non-trivial to formalize it mathematically in a way that would be suitable for applications in mass spectrometry and linear regression of spectra (defined in Chapter 4) in particular. Especially in the latter, allowing peak intensities to vary may lead to instability of the results—when everything is variable, we can fit everything to anything.

Chapter 4

The Wasserstein regression of mass spectra

A frequently encountered task in various types of spectrometry and spectroscopy is the quantification of signal corresponding to a particular set of ions. From an algorithmic point of view, the most challenging parts of this problem are the separation of the signal from the noise and the separation of signals from overlapping spectra. In the context of mass spectrometry, numerous approaches have been developed in order to tackle this problem in the context of specific experiments. These include the estimation of reaction rates in ETD fragmentation [53]; quantification of polymer chain lengths and compositions [54, 55]; annotation of MS^2 spectra in data independent acquisition label-free quantification experiments [11]; studies of fragmentation of aliphatic diselenides and selenosulfenates [56, 57], and studies of protein deamidation and ^{18}O labelling [58].

Despite the apparent abundance of various algorithmic approaches and software tools, their common underlying theme is the approximation of an experimentally observed spectrum by a set of reference spectra. Therefore, from a mathematical point of view, all the described problems can be expressed with a single equation:

$$\mu = p_1\nu_1 + p_2\nu_2 + \cdots + p_k\nu_k. \tag{4.1}$$

Here, μ is the observed spectrum, ν_i's are the reference spectra of the ions in question, and p_i's are the unknown proportions of the latter. Since the reference spectra are often predicted using computational methods, we refer to them as the *theoretical* spectra. By analogy to the ordinary linear regression known from the field of statistics, in this work we call the problem of finding p_i the *linear regression of spectra*. The mathematical definition of this problem encompasses multiple kinds of spectrometry and spectroscopy, including the NMR spectroscopy. However, for the sake of clarity of exposition, throughout this Dissertation we will discuss it mainly in the context of mass spectrometry.

Equations similar to Equation 4.1 have appeared e.g. in [11], where the authors have used it to annotate data independent acquisition label-free quantification experiments. Figure 4.1 illustrates a linear regression of a mass spectrum consisting of overlapping isotopic envelopes of human hemoglobin subunits α and δ.

In the case when the reference spectra do not overlap, the background noise is small, and there is only a handful of molecules of interest, linear regression of spectra can easily be performed manually by integrating selected regions of the experimental spectrum. However, algorithmic approaches need to be employed when there is a considerable overlap of the reference spectra, the signal-to-noise ratio is small, or thousands of molecules need to be analyzed in a high-throughput setting. The ever-increasing popularity of high-throughput methods, such as mass spectrometry

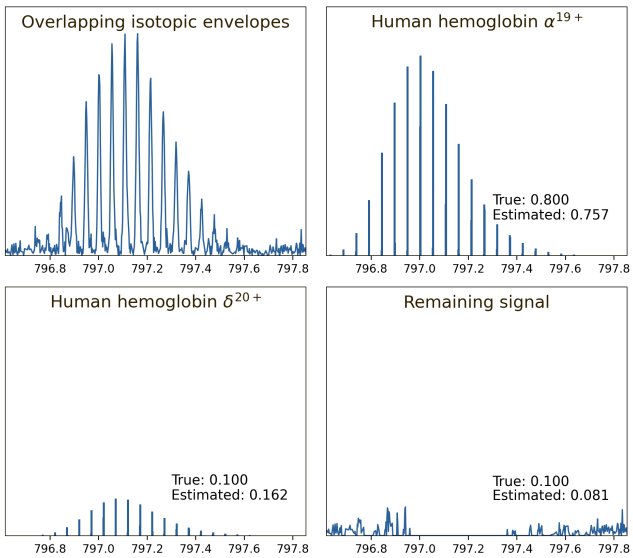

FIGURE 4.1: An example of a regression of a simulated human hemoglobin ESI MS[1] spectrum. Using the method presented in this Chapter, we have separated the signal into α 19+ and δ 20+ subunits and the remaining background noise. The proportions were estimated directly from the top-left spectrum, without any additional preprocessing such as peak picking or denoising.

imaging, calls for advanced algorithmic solutions which allow for rapid processing of massive data sets. In those types of experiments, and with software tools that sometimes neglect computational optimization, the time needed to process a single data set can take up to several days, needlessly extending the time needed to analyze the sample of interest.

Unfortunately, the currently used terminology is somewhat misleading. The problem of estimating abundances of overlapping reference spectra is known in the mass spectrometric literature under several names, most common ones being the *resolving of isobaric interferences* and *deconvolution*. In the context of NMR, Raman, and Infrared spectroscopy, this problem is also known as *curve fitting*. Since the experimental spectrum is expressed as a linear combination of the theoretical ones, this problem has also been referred to as the *linear deconvolution* [11].

On the other hand, in mass spectrometric literature, the term *deconvolution* is used to refer to several problems which usually deal with separating overlapping peaks and/or isotopic envelopes (but rarely have anything in common with the mathematical operation of convolution). The authors of [59] define deconvolution as inferring the relative quantities of molecules with overlapping isotopic envelopes. Similar problems have been described in [60, 61, 62]. However, the term deconvolution is sometimes used as a synonym for *deisotoping*, that is, conversion of isotopic envelopes into single peaks with average m/z value and joint intensity [63, 64]. Another common application of the term deconvolution is converting m/z values into masses, as exemplified by the popular MaxEnt algorithm [65], also referred to as the *charge deconvolution* [66, 67, 68]. In this Dissertation, we introduce the term *linear regression of mass spectra* in order to avoid confusion with other algorithmic problems of mass spectrometry.

The general problem of regression of mass spectra can be formally expressed as follows:

Problem 1 (Generalized Mass Spectral Regression)**.** *Let μ be a normalized mass spectrum, and let $\{\nu_i : 1 \leq i \leq k\}$ be a collection of normalized mass spectra. Let $d(\mu, \nu)$ be a distance measure between spectra. Let Δ_{k-1} be a $k-1$-dimentional probability simplex. Find a set of weights $p^* \in \Delta_{k-1}$ which minimizes the distance between μ and the convex combination of ν_i:*

$$p^* = \operatorname*{arg\,min}_{p \in \Delta_{k-1}} d\left(\mu, \sum_{i=1}^{k} p_i \nu_i\right) \tag{4.2}$$

In the above definition, μ is referred to as an *experimental* spectrum, and ν_i are referred to as *theoretical* or *reference* spectra. Note that neither the distance measure nor the origin of the theoretical spectra is specified in this general definition. Depending on their choice, this definition can be reduced to several of the aforementioned computational problems. For example, if the theoretical spectra correspond to isotopic envelopes, the solution to MSR can be used for deisotoping, in which case p_i corresponds to the joint intensity of the i-th envelope. On the other hand, if ν_i correspond to mass spectra of a single molecule with different charges, the problem reduces to conversion of m/z values to mass. Finally, if ν_i are mass spectra from a database, the problem can be reduced to the annotation of a mass spectrum. However, an underlying assumption is that the reference spectra are known. Therefore, from the methodological point of view, this problem is an exact opposite of molecule identification, where the task is to identify the identity of ions, but not their quantities.

A common approach to the problem of regression of mass spectra, found e.g. in the specter [11] and the `masstodon` [69] tools, is to perform an ordinary least squares regression where the experimental spectrum is treated as the dependent variable and the theoretical ones as independent variables. Mathematically, this technique minimizes the Euclidean distance between the experimental spectrum and a linear combination of the theoretical ones. A similar technique, the L1 regression (also known as the least absolute deviation regression), is sometimes used [55]. Another approach is a sequential subtraction of estimated signal from the experimental spectrum, as exemplified by the THRASH algorithm [70].

The performance of the currently available methods is hindered by a series of problems. Most of them arise from the fact that they are based on a point-wise comparison of spectra, that is, they compare peaks with the same m/z values. However, unlike the theoretically predicted spectra, experimental ones have limited resolution and accuracy. Because of this, peak picking of the experimental spectrum is required, which is often imperfect (such as the conventional peak centroiding) or computationally expensive (such as the continuous wavelet transform approach [51]). The experimentally obtained peaks never match the theoretical ones exactly due to accuracy limits of the instrument and numerical errors of peak-picking procedures. This limits the performance and applicability of most of the approaches to measure the similarity between spectra, like the Euclidean distance, correlation, spectral contrast angle [42], or the entropy function [46].

The contents of this Chapter. In this Chapter, we propose a solution to the problem of regression of mass spectra with the Wasserstein metric as the distance measure. We will refer to this computational problem as the *Wasserstein regression*, and we will use the terms Wasserstein regression and mass spectral regression (MSR) interchangeably throughout the rest of this Dissertation. Therefore, we will be dealing with the following problem:

Problem 2 (Mass Spectral Regression, MSR)**.** *Let μ be a normalized mass spectrum, and let $\{\nu_i : 1 \leq i \leq k\}$ be a collection of normalized mass spectra. Let Δ_{k-1} be a $k-1$-dimentional probability simplex. Find a set of weights $p^* \in \Delta_{k-1}$ which minimizes the Wasserstein distance between μ and the convex combination of ν_i:*

$$p^* = \underset{p \in \Delta_{k-1}}{\arg\min}\ W\left(\mu, \sum_{i=1}^{k} p_i \nu_i\right) \tag{4.3}$$

Apart from fixing the distance function, we tackle the problem of linear regression of mass spectra in its abstract form instead of focusing on a particular type of experiment. This way, our methods retain their full scope of applicability, from analytical chemistry to metabolomics to proteomics to synthetic polymer science. They are not limited to any single type of mass spectrometer or pre-processing software. The mathematical foundations presented in this Chapter stay the same regardless whether an FTICR, TOF, or quadrupole instrument is used, or whether applied to mass spectrometry or NMR spectroscopy.

In the next Sections, we present the main ideas behind the solution, state the most important results, and show the performance of the method assessed by computational simulations. The technical details and proofs are relegated to the final Sections of this Chapter. In Section 4.3, we show how to express the problem of (Wasserstein) regression of spectra as a linear program. In Section 4.4, we describe the details behind the algorithm for solving this program.

4.1 An overview of the solution to MSR

To solve the MSR problem with the Wasserstein distance, we can express it as a linear program. In this Section, we will give a brief overview of this approach, focusing on the general ideas and main results, and pointing the reader to specific Sections of this Chapter where we discuss the mathematical details behind particular results.

Let $\mu = \sum_{j=1}^{m_0} w_{0,j}\delta_{x_{0,j}}$ be the experimental spectrum with m_0 peaks and for $i = 1, 2, \ldots, k$ let $\nu_i = \sum_{j=1}^{m_i} w_{i,j}\delta_{x_{i,j}}$ be the i-th theoretical spectrum with m_i peaks. Denote the set of all support points from the empirical and theoretical spectra by $\mathcal{S} = \{x_{i,j} : 1 \leq j \leq m_i, 0 \leq i \leq k\}$ and let $n = |\mathcal{S}|$. Let $\boldsymbol{s} = s_1 < s_2 < \cdots < s_n$ be a vector of ordered elements of $\mathcal{S}$.

For $1 \leq j \leq n-1$, let $N_{i,j}$ and M_j be the values of the cumulative distribution functions of ν_i and μ respectively on the interval $[s_j, s_{j+1})$ (notice that those functions are constant on those intervals), and set $N_{i,n} = 1 = M_n$. Let $d_j = s_{j+1} - s_j$ be the length of the j-th interval for $1 \leq j \leq n-1$. Denote by I_k an identity matrix of size k, and by J_n an $(n-1) \times n$ matrix equal to the identity matrix of size n without the last row. Define a matrix:

$$A = \begin{bmatrix} -J_n & -J_n & 0 \\ N & -N & -I_k \end{bmatrix}.$$

Finally, let c be a vector of length $2n + k$, such that $c = (M, -M, 0_k)$, where 0_k is the vector of zeros of length k, and let $b = (-d, 0_k)$ be a vector of length $n - 1 + k$. The following Lemma, proved in Section 4.3, states that MSR can be reduced to linear programming.

Lemma 4.1. *The following dual linear programming problems:*

$$\begin{aligned} \min_x \quad & x^T c \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0 \end{aligned} \qquad\qquad \begin{aligned} \max_y \quad & y^T b \\ \text{s.t.} \quad & A^T y + z = c \\ & z \geq 0 \end{aligned} \tag{4.4}$$

are feasible. Furthermore, for any solution (x_*, y_*, z_*) of the above problem, the vector of the last k elements of y_* belongs to the set of solutions of MSR.

We propose to solve the above linear program using an Interior Point Method (IPM), while using the structure of our linear programming problem to significantly decrease the time and memory cost of each iteration. In general, IPM for linear programming solves both primal and dual problems simultaneously, by solving a cleverly chosen nonlinear approximation of those problems using the Newton's Method. For an overview of IPM we refer our reader to [71] and references therein. A notable advantage of this approach is that it does not require the starting point nor the subsequent iterates to be in the feasible region of primal and dual problems.

In Algorithm 4 we present a pseudocode for the general scheme of IPM for the dual problem (4.4). In what follows, x_t, y_t, z_t are t-th iterates of variables x, y, z from the problem, while $X_t = \operatorname{diag}(x_t)$, $Z_t = \operatorname{diag}(z_t)$ are diagonal matrices.

Algorithm 4: Solving MSR with a primal-dual IPM

Data: Matrix A and vectors b, c defining dual linear problems. A starting point (x_0, y_0, z_0) and error tolerance $\epsilon > 0$.
Result: an ϵ-feasible ϵ-solution of the linear problem

Set $t = 0$
repeat
  Compute centrality $\mu_t = \langle x_t, z_t \rangle / (2n + k)$
  Compute primal residual $r_p^t = b - Ax_t$ and dual residual $r_d^t = c - A^T y_t - z_t$
  Choose scaling factor σ_t (see Section 4.4)
  Find direction (d_x, d_y, d_z) by solving the system of linear equations:

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^T & I \\ Z_t & 0 & X_t \end{bmatrix} \begin{bmatrix} d_x \\ d_y \\ d_z \end{bmatrix} = \begin{bmatrix} r_p^t \\ r_d^t \\ \sigma_t \mu_t \mathbf{1} - X^t z_t \end{bmatrix} \tag{4.5}$$

  Find $\alpha_p \in (0, 1]$ such that $x_{t+1} = x_t + \alpha_p d_x > 0$
  Find $\alpha_d \in (0, 1]$ such that $z_{t+1} = z_t + \alpha_d d_z > 0$ and take $y_{t+1} = y_t + \alpha_d d_y$
  Set $t = t + 1$
until *triple (x_t, y_t, z_t) is an ϵ-feasible ϵ-solution;*

A triple (x_t, y_t, z_t) is called an ϵ-feasible ϵ-solution if it is both primal-dual feasible and optimal up to ϵ tolerance. Since the computational cost of each iteration of Algorithm 4 is dominated by finding the direction of the step by solving equation (4.5), we focus on this part and relegate to Section 4.4 the discussion about choosing the stopping condition, the starting point (x_0, y_0, z_0) and the scaling factors σ_t.

The solution of the equation (4.5) can be obtained by solving the following normal equation for d_y:

$$AZ_t^{-1}X_tA^Td_y = b + AZ_t^{-1}(X_tr_d^t - \sigma_t\mu_t\mathbf{1}).$$

After solving the normal equation for d_y, the remaining direction coordinates d_x, d_z can be obtained using formulas:

$$d_z = z = r_d^t - A^Td_y, \quad d_x = -x_t + Z_t^{-1}(\sigma_t\mu_t\mathbf{1} - X_td_z).$$

The computational cost of single step of IPM is dominated by solving the normal equation, which is of order $\mathcal{O}(k(n-1)^3)$. However, thanks to the specific structure of matrix A, for any vectors v, w we can compute Av, A^Tv and solve an equation $AZ_t^{-1}X_tA^Tv = w$ efficiently. The detailed derivation of an efficient algorithm, tailored to the MSR problem, is given in the Section 4.4. This allows us to perform one step of IPM efficiently, with the computational cost of a single step equal to $\mathcal{O}(k^3 + k\sum_{i=1}^k m_i + n)$ and the memory cost equal to $\mathcal{O}(k^2 + n)$. For the given error tolerance ϵ, the IPM needs $\mathcal{O}(\sqrt{2n + k}\log(\epsilon^{-1}))$ iterations to find an ϵ-solution, i.e. one for which the duality gap is less than ϵ (see the Section 4.4 for details).

4.2 Computational experiments on simulated data

We have performed several computational experiments to illustrate the performance of the proposed solution to MSR, and to analyze its robustness to various kinds of distorsions occuring in MS measurements. In contrast to the previous case studies, in this Section we use *in silico* generated spectra. This allows us to precisely control the signal-to-noise ratio, and to rigorously estimate the error of the method.

Our main goal is to demonstrate the applicability of the Wasserstein distance to MSR in case of noisy experimental spectra. There are several sources of noise in mass spectrometric measurements, among others: *(i)* precision of the intensity measurement, *(ii)* precision of the m/z measurement, *(iii)* resolving power, i.e. the ability to detect peaks with similar masses, *(iv)* chemical noise, i.e. presence of unexpected molecules in a spectrum [10]. In this Section, we focus mostly on the first three types of noises, i.e. low resolving power and/or precision. The first step of all our experiments was to generate the isotopic envelopes of selected molecules by the IsoSpec algorithm [72]. These envelopes form the set of the theoretical spectra. The experimental spectrum was obtained by taking a convex combination of the latter. Finally, the experimental spectrum was distorted in the following manner:

- Gaussian noise has been added to the logarithm of the peak intensity, and the result has been exponentiated (equivalent to multiplying the intensities by log-normally distributed random variables),
- Each peak has been replaced by the density function of the normal distribution,
- The resulting intensity distribution has been binned.

Both Gaussian noises had a standard deviation of 0.01. For binning of the mass spectrum, we have assumed two resolving powers of the spectrometer: 0.001 Da and 0.01 Da. An example of the result of this procedure is depicted in Fig. 4.2. A spectrum without these distortions is referred to as *clean*, while the distorted one as *noisy*.

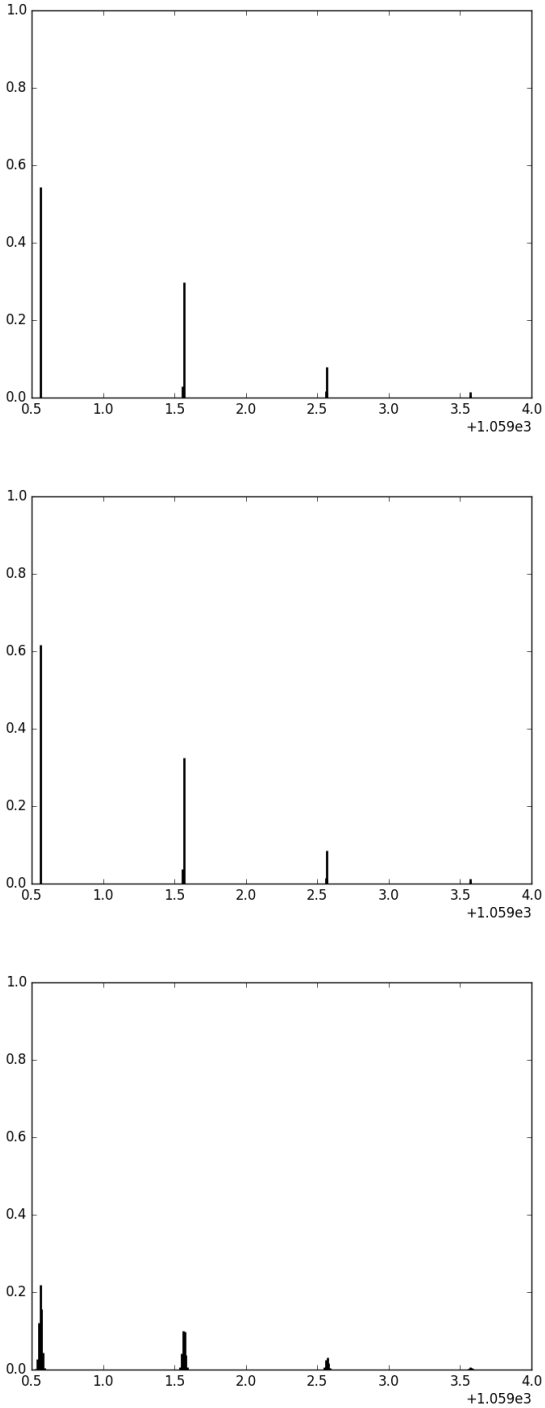

FIGURE 4.2: An illustration of the simulated measurement inaccuracies based on a theoretical spectrum of bradykinin ($C_{50}H_{73}N_{15}O_{11}$). Top: clean spectrum. Middle: noise in the intensity domain. Bottom: noise in the mass domain. The apparent change in intensity in the right spectrum is caused by a Gaussian blurring of the peaks and binning afterwards.

The performance of our approach to MSR has been quantified by the Root Mean Square Error (RMSE) between the original and inferred proportions of different isotopic envelopes, which approximates the average error made on each proportion. More formally, if the RMSE for regression of n isotopic envelopes is equal to ϵ, then the estimated proportions are on an $(n-1)$-dimentional hypersphere centered at the true proportions with a radius $\epsilon\sqrt{n}$. In particular, the absolute error on any single proportion is not larger than this radius.

We have performed three tests, inspecting the method's sensitivity to the number of overlapping envelopes, the molecular mass of deconvolved molecules, and the molecules' charge. The first test is based on random molecular formulas. The next two are based on simulated proteins composed of *averagine*—a model amino acid with a molecular formula $C_4 \cdot 9384\, H_7 \cdot 7583\, N_1 \cdot 3577\, O_1 \cdot 4773\, S_0 \cdot 0417$ and an average molecular mass of 111.1254 Da [73]. In the first test, we have inspected both clean and noisy spectra. In tests two and three, only noisy spectra were analyzed. To check the standard deviation of the prediction error, the tests were replicated, with noise added independently in each replicate. Below we present each test in detail.

Test no. 1 — increasing number of molecules. This test is based on 17 randomly chosen isobars (molecules with the same nominal masses, i.e. masses rounded to the nearest integer) consisting of carbon, oxygen, hydrogen, nitrogen and sulfur, each one with the nominal mass of 30 000 Da. The experimental spectra with a range of interfering isobars were constructed by gradually extending a subset of those molecules. This procedure was replicated 20 times, resulting in 340 experimental spectra. The results are shown on Fig. 4.3. The prediction error is very low and stable for less than 5 isobars, suggesting that the ratios of particular elements in the deconvolved molecules have no considerable influence on the method's performance.

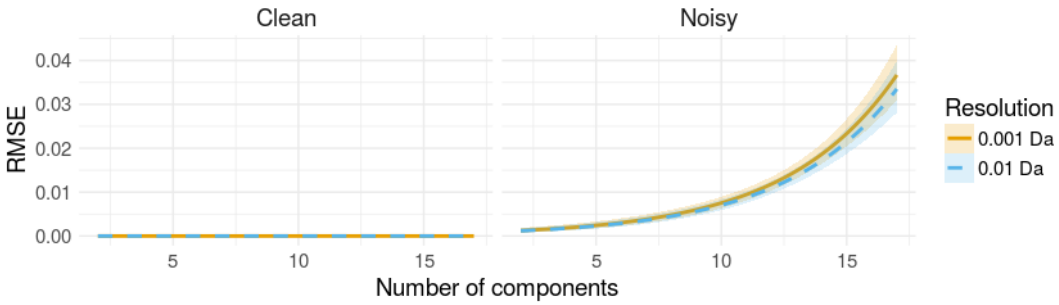

FIGURE 4.3: The performance of our MSR method for an increasing number of deconvolved molecules. The solid line represents the average RMSE over 20 repetitions. The ribbon represents the standard deviation of the error.

Test no. 2 — increasing mass of molecules. In this test, we consider overlapping isotopic envelopes of two types of proteins: singly charged protein consisting of n units of averagine, and doubly charged protein consisting of $2n$ units, where the values of n were selected so that the average m/z ratio of proteins spans the range from 1,500 Da to 45,000 Da. For any given n, the isotopic envelopes of the two corresponding proteins were mixed in proportions 0.8 and 0.2 respectively. The procedure was replicated 50 times. The outcome of this experiment is presented in Fig. 4.4.

Test no. 3 — increasing charge of molecules. In this test, we consider the following four proteins based on averagine: $C_{1482}H_{2328}N_{408}O_{444}S_{12}$, $C_{1482}H_{2329}N_{408}O_{444}S_{12}$,

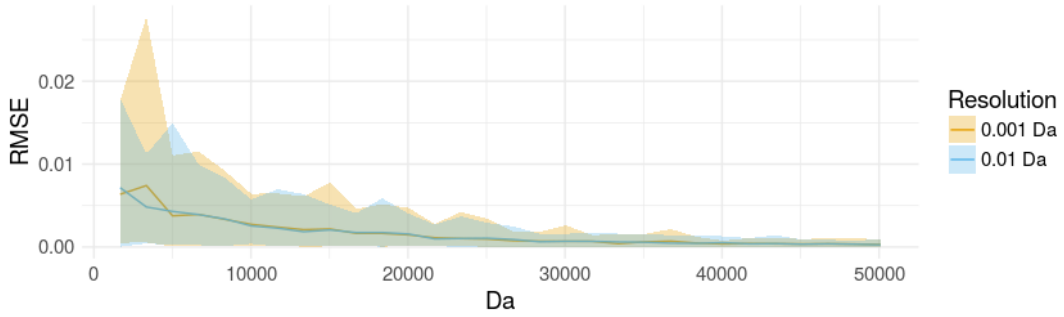

FIGURE 4.4: The performance of our MSR method for increasing mass of deconvolved molecules. The solid line represents the average RMSE over 50 repetitions. The ribbon represents the standard deviation of the error.

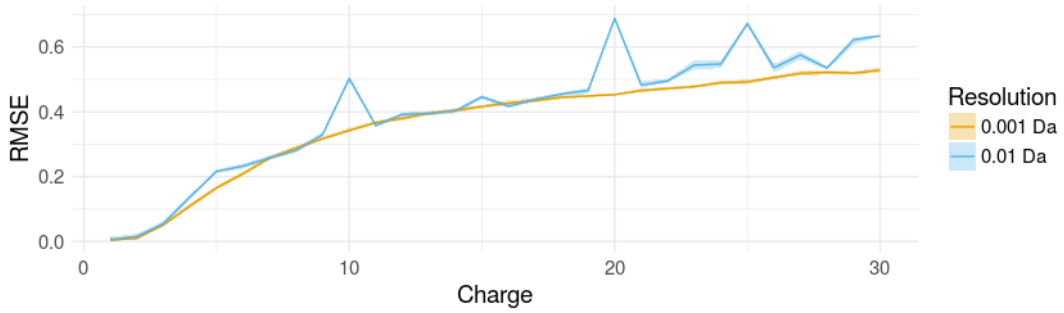

FIGURE 4.5: The performance of our MSR method for increasing charge of deconvolved molecules. The solid line represents the average RMSE over 50 repetitions. The ribbon represents the standard deviation of the error.

$C_{1482}H_{2330}N_{408}O_{444}S_{12}$ and $C_{1481}H_{2341}N_{408}O_{444}S_{12}$, mixed in proportions 0.3, 0.5, 0.1 and 0.1, respectively. The first three molecules differ by one hydrogen atom, resulting in partially overlapping isotopic envelopes. The fourth molecule is an isobar of the second one, as one carbon has been replaced by 12 hydrogens. All molecules have been equally charged, with the charge varying from 1 to 10. This yields a sequence of regression problems with increasing difficulty, because the peaks become more densely packed while the resolution stays constant. For each charge, 50 replicates were performed. The results are presented in Fig. 4.5.

Computational experiments show that our approach is able to deconvolve complex spectra in the presence of measurement inaccuracies. Even for 17 isobars, for which we obtain exceptionally complex spectra, the RMSE does not exceed 0.05. However, it must be noted that our approach to MSR is expected to be sensitive to chemical noise, because the Wasserstein metric requires that all the intensity of the experimental spectrum is explained. Therefore, the solution to the problem of regression of mass spectra presented in this Chapter should be applied to spectra of highly purified compounds. In Chapter 5, we will discuss an extension of our method to handle contaminating signals in the experimental spectra, as well as cases when the set of the reference spectra does not contain all the molecules in the experimental mixture.

4.3 Regression as a linear program

In this Section, we prove Lemma 4.1 by showing how the MSR problem defined in (4.2) can be reduced to linear programming in the case when the distance measure is the Wasserstein metric W_1. Recall that our problem is to find

$$p^* = \underset{p\in\Delta_{k-1}}{\arg\min}\ W_1\left(\mu, \sum_{i=1}^{k} p_i\nu_i\right), \tag{4.6}$$

where ν_i, μ are discrete probability measures with finite support.

First, we show that the MSR problem can be restated as a weighted L_1 regression on the probability simplex Δ_{k-1}. By denoting M and N_i as the cumulative distribution functions (CDFs) of μ and ν_i, and using Theorem 3.1, we can write:

$$\underset{p\in\Delta_{k-1}}{\arg\min}\, W_1\left(\mu, \sum_{i=1}^{k} p_i\nu_i\right) = \underset{p\in\Delta_{k-1}}{\arg\min} \int_{\mathbb{R}} \left|\sum_{i=1}^{k} M(x) - p_iN_i(x)\right| dx. \tag{4.7}$$

Recall that $\mathcal{S}$ denotes the set of points from theoretical and empirical spectra, and that $(s_i)_{i=1}^n$ are elements of $\mathcal{S}$ ordered increasingly. Note that for $x < s_1$ and $x \geq s_n$, the function under the integral on the right hand side of (4.7) is zero. At the same time, the function is constant on intervals $[s_i, s_{i+1})$.

For $1 \leq j \leq n-1$, let $N_{i,j}$ and M_j be the values of the cumulative distribution functions (CDFs) of ν_i and μ, respectively, on the interval $[s_j, s_{j+1})$, and set $N_{i,n} = M_n = 1$. For $1 \leq j \leq n-1$ let $d_j = s_{j+1} - s_j$ be the length of the j-th interval. We can now write

$$\int_{\mathbb{R}} \left|M(x) - \sum_{i=1}^{k} p_iN_i(x)\right| dx = \sum_{j=1}^{n-1} d_j \left|M_j - \sum_{i=1}^{k} p_iN_{i,j}\right|,$$

and we reduce the optimization problem (4.6) to a weighted L_1 regression on a probability simplex

$$p^* = \underset{p\in\Delta_{k-1}}{\arg\min} \sum_{j=1}^{n-1} d_j \left|M_j - \sum_{i=1}^{k} p_iN_{i,j}\right| \tag{4.8}$$

Now, we apply a well known technique of representing a weighted L_1 regression as a linear programming problem (see e.g. [74]). Let us introduce dummy variables t_j, such that $t_j \geq |M_j - \sum_{i=1}^k p_iN_{i,j}|$. With this notation, problem (4.8) is equivalent to minimizing a linear function $\sum_{i=1}^{n-1} d_jt_j$.

Now, for any j, the inequality $t_j \geq |M_j - \sum_{i=1}^k p_iN_{i,j}|$ can be represented by an inequality $t_j \geq \max(M_j - \sum_{i=1}^k p_iN_{i,j}, -M_j + \sum_{i=1}^k p_iN_{i,j})$. This, in turn, can be expressed as a conjunction of two linear inequalities, $t_j \geq M_j - \sum_{i=1}^k p_iN_{i,j}$ and $t_j \geq -M_j + \sum_{i=1}^k p_if_{i,j}$.

To take into account the fact that vector $(p_i)_{i=1}^k$ needs to belong to the probability simplex, we need to add inequality constraints $p_i \geq 0$ for $i = 1, \ldots, k$ and an equality constraint $\sum_{i=1}^k p_i = 1$. By rewriting the latter as two inequality constraints, we end

up with the following linear program:

$$\begin{aligned} \min_{p,t} \quad & d^T t \\ \text{s.t.} \quad & -t_j + \sum_{i=1}^{k} p_i N_{i,j} \leq M_j \\ & -t_j - \sum_{i=1}^{k} p_i N_{i,j} \leq -M_j \\ & 1 \leq \sum_{i=1}^{k} p_i \leq 1 \\ & p_i \geq 0. \end{aligned} \tag{4.9}$$

Note that the target function minimized in (4.9) does not itself depend on p, but the variable p appears in the constrains. From the construction of (4.9) it follows that, for any feasible pair (p, t), we have:

$$d^T t \geq \sum_{i=1}^{n-1} d_j \left| M_j - \sum_{i=1}^{k} p_i N_{i,j} \right|,$$

with equality holding if and only if $t_j = |M_j - \sum_{i=1}^{k} p_i N_{i,j}|$. Therefore, for any solution p^* of problem (4.8), there exists t^* such that (p^*, t^*) is a solution of (4.9). It also follows that, for any solution (p^*, t^*) of (4.9), the vector p^* is a solution of problem (4.8). Furthermore, since (4.8) is a problem of optimizing continuous convex function on a compact set, at least one solution exists.

We have thus reduced the MSR problem (4.6) to a linear programming problem (4.8). We will now express problem (4.8) in a concise matrix form. Let I_k be an identity matrix of size k, and J_n an $(n-1) \times n$ matrix equal to an identity matrix of size n without the last row (equivalently, an identity matrix of size $n-1$ with an appended column of zeroes). We can rewrite problem (4.9) as

$$\begin{aligned} & \min_{t} \quad d^T t \\ \text{s.t.} & \begin{bmatrix} -J_n^T & N^T \\ -J_n^T & -N^T \\ 0 & -I_k \end{bmatrix} \begin{pmatrix} t \\ p \end{pmatrix} \leq \begin{pmatrix} M \\ -M \\ 0_k, \end{pmatrix} \end{aligned}$$

where 0_k is a zero vector of length k. Note that the constraint $\sum_{i=1}^{k} p_i = 1$, split into two inequality constraints $\sum_{i=1}^{k} p_i \leq 1$ and $-\sum_{i=1}^{k} p_i \leq -1$, is included in the above program using the last row of matrix N^T and the last element of vector M, which are equal to 1.

Let us now define $c = (M, -M, 0_k)$, $b = (-d, 0_k)$ and

$$A = \begin{bmatrix} -J_n & -J_n & 0 \\ N & -N & -I_k \end{bmatrix}.$$

By using the above notation, adding the slack variables z to replace inequality constrains by equality constrains, and replacing a minimization problem with a maximization one, we can rewrite the problem (4.9) in a standard form:

$$\begin{aligned} \max_{y} \quad & y^T b \\ \text{s.t.} \quad & A^T y + z = c \\ & z \geq 0. \end{aligned} \tag{4.10}$$

This completes the reduction of the original minimization to a problem linear programming. We can summarize this Section with a proof of Lemma 4.1.

Proof of Lemma 4.1. From the discussion in this Section it follows that the feasible region of the problem (4.10) is non-empty. Furthermore, note that the problem given by (4.8) is bounded from below. This, combined with the inequality:

$$y^T b \leq -\sum_{j=1}^{n-1} d_j \left| M_j - \sum_{i=1}^{k} p_i N_{i,j} \right|,$$

means that $y^T b$ is bounded from above, and therefore the maximization problem from Lemma 4.1 has a solution. From the duality theory for linear programs it follows that the dual minimization problem is feasible as well, and the duality gap is zero.

Finally, from the structure of the A matrix and the discussion in this Section it follows as well that if (x_*, y_*, z_*) is a solution of dual problems in Lemma 4.1, then the last k elements of y_* form a vector in Δ_{k-1} that is a solution of the initial MSR problem. □

4.4 Solving MSR with Interior Point Method

In this Section, we present the details behind the algorithm for solving the dual linear problems from Lemma 4.1. In Section 4.1 we presented a general scheme for a primal-dual Interior Point Method for the case when primal problem has only equality constraints, as is the case for the problem of interest to us. In this Section, we discuss the details of Algorithm 4, and prove our claim about its time and memory complexity.

4.4.1 Starting point, stopping criterion and the scaling factor.

We first address the issues of initial conditions (x_0, y_0, z_0) for the primal-dual program, stopping criterion and choosing the scaling factor σ_t. The IPM does not require the starting point or the iterates to be in feasible region. The only requirement is that all elements of vectors x_t, z_t are positive for $t \geq 0$. Hence, we can choose the starting point almost arbitrarily, but in practice it is beneficial to choose x and z that are not too close to zero.

We propose the following choice of the starting point x_0 for the primal problem. For $1 \leq i \leq n-1$ we take $(x_0)_i = (x_0)_{n+i} = d_i/2$ (half the length of the interval $[s_i, s_{i+t})$). We also set $(x_0)_n = 2/3$, $(x_0)_{2n} = 1/3$ and $(x_0)_{2n+i} = 1/3$ for $1 \leq i \leq k$. Naturally, we have $x_0 > 0$. It is also straightforward to check that this choice of x_0 satisfies $Ax_0 = b$.

Finding a starting point (y_0, z_0) for the dual problem that is dual feasible can be done using the connection of the dual problem to the L_1 regression problem (4.8). We choose a uniform vector $p \in \Delta_{k-1}$ and set $t_j = |M_j - \sum_{i=1}^{k} p_i N_{i,j}| + 1$ for all $j \leq n-1$. By taking $y_0 = (t, p)$ and $z_0 = c - A^T y_0$, we get a dual feasible point with $z_0 \geq 1$. This starting point can be efficiently computed using subroutines for computing $N^T v$ and $A^T v$ which we describe later on in this Section.

The main loop of the IPM stops when the triple (x_t, y_t, z_t) becomes an ϵ-feasible ϵ-solution or when the algorithm reaches a maximum number of iterations. We say that a point is ϵ-feasible if the norms of the residuals r_p^t, r_d^t, given by $r_p^t = b - Ax_t$ and $r_d^t = c - A^T y_t - z_t$, are smaller than ϵ, i.e. $\|r_p^t\|_2 < \epsilon$ and $\|r_d^t\|_2 < \epsilon$. We say that a point is an ϵ-solution it the duality gap, given by a scalar product $\langle x_t, z_t \rangle$, is smaller than a given ϵ.

Choosing the scaling factor can be done in a variety of ways. The simplest method is to choose a single σ_k for each k as a constant in $(0, 1)$. However, there exist methods for choosing the scaling factor that are both practical and lead to theoretical guarantees on the number of iterations necessary for an IPM to converge. We will use the existence of such methods in our theoretical analysis of the computational complexity of Algorithm 4. One of such methods is the predictor-corrector method, which first finds the Newton direction $(\hat{d}_x, \hat{d}_y, \hat{d}_z)$ for the most optimistic $\hat{\sigma}_t = 0$. Then, the actual scaling factor σ_t is chosen based on how much reduction in duality gap could be achieved while going in the direction $(\hat{d}_x, \hat{d}_y, \hat{d}_z)$. Most importantly, for the state-of-the-art methods, the cost of computing the scaling factor is the same as the cost of solving the system of linear equations for (d_x, d_y, d_z), and is actually done by solving a system of equations with the same right hand side as 4.5. For more details on methods of choosing the scaling factor we refer the reader to [71] and references therein.

The cost of computing one step of Algorithm 4 is dominated by solving a system of equations for (d_x, d_y, d_z) given by

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^T & I \\ Z_t & 0 & X_t \end{bmatrix} \begin{bmatrix} d_x \\ d_y \\ d_z \end{bmatrix} = \begin{bmatrix} r_p^t \\ r_d^t \\ \sigma_t \mu_t \mathbf{1} - X^t z_t \end{bmatrix}, \tag{4.11}$$

A common technique for solving the system of equations (4.11) is to apply a block-wise Gaussian elimination to reduce it to the normal equation $\Sigma d_y = r$, where:

$$\Sigma = A Z_t^{-1} X_t A^T,$$
$$r = b + A(Z_t)^{-1}(X_t r_d^k - \sigma_k \mu_k \mathbf{1}).$$

Given a solution d_y of the normal equation, we can compute d_x, d_z using the fact that $d_z = r_d^k - A^T d_y$ and $d_x = -x_t + (Z_t)^{-1}(\sigma_k \mu_k \mathbf{1} - X_t d_z)$.

As we have noted in Section 4.1, to effectively perform a single step of the IPM procedure, we need to be able to efficiently compute Av, $A^T v$ and solve $\Sigma v = w$. We devote the rest of this Section to presenting and analyzing methods for those three computational problems that take advantage of the specific structure of matrix A. We first observe that matrix N is related to a sparse matrix.

Lemma 4.2. *Let $m = \sum_{i=1}^{k} m_i$, where m_i is the size of the i-th theoretical spectrum. Let U denote an upper-triangular $n \times n$ matrix with $U[i, j] = 1$ for $i \leq j$ and $U[i, j] = 0$ for $i > j$. Then, there exists a sparse $m \times n$ matrix W, with m non-zero entries, such that $N = WU$. Furthermore, the sparse representations of matrices W and W^T can be*

constructed in $\mathcal{O}((n+m)\log(n+m))$ time and $\mathcal{O}(m)$ memory complexity from a list of spectra represented as lists of pairs of m/z and intensity values.

Proof. Let $s_1 \leq \ldots \leq s_n$ be the ordered vector of point from the set $\mathcal{S}$ of all m/z values with non-zero intensity in any of the spectra μ, ν_i. For any theoretical spectrum ν_i, we can represent it as $\nu_i = \sum_{j=1}^{n} w_{i,j}\delta_{s_j}$, where only m_i of elements $w_{i,j}$ are non-zero.

Define matrix W as $W[i,j] = w_{i,j}$. Then, the matrix W is sparse and has $m = \sum_{i=1}^{k} m_i$ non-zero elements in total. We now have $N = WU$, because for any i, j we have $WU_{i,j} = \sum_{l \leq j} w_{i,l} = N_{i,j}$.

To finish the proof, we need to show how to efficiently construct sparse representations of W and W^T. In a classic sparse representation of a matrix, we represent each row of W as a list of nonzero elements, i.e. $(j, w_{i,j})$ for j such that $w_{i,j} > 0$, and we represent W as a list of rows. This is representation of a sparse matrix allows to compute Wv in time $m = \sum_{i=1}^{k} m_i$ for any vector $v \in \mathbb{R}^n$. We construct the representation of W^T in the same manner.

We will now give an efficient algorithm for construction of the sparse representations of W and W^T. Consider a peak-list representation of ν_i given by $(s_{i,j}, w_{i,j})_{j=1}^{m_i}$ (sorted with respect to $s_{i,j}$ for each i) and a peak-list representation of μ given by $(s_{0,j}, w_{0,j})_{j=1}^{m_0}$ (sorted with respect to $s_{0,j}$). Now, concatenate the peak-list representations of μ and ν_i, storing the index of each spectrum, to get a list of triples $L = (s_{i,j}, i, w_{i,j})$. Sort L with respect to the first element, i.e. the m/z values. Now, it is enough to pass through this sorted list once to construct both W and W^T. Note that each triple with $i > 0$ corresponds to exactly one element of the matrix W. Now, since we pass through the triples in an increasing order of the m/z values $s_{i,j}$, for a given triple (s, i, w) we have $W[i,l] = w$ if we have seen l different s values so far. We then simply add (l, w) to the i-th list in the sparse representation of W and (i, w) to the l-th list in the representation of W^T. The total cost of this construction is $\mathcal{O}((n+m)\log(n+m))$ due to the sorting L, and the memory needed for storing both representations is $\mathcal{O}(m)$. □

We can now analyze the time and memory complexity of computing Av and $A^T w$ during the IPM method. Note that constructing the sparse representations of W and W^T needs to be done only once before the first iteration of the IPM method. Therefore, in the following Lemma, we assume that those representations are already pre-computed.

Lemma 4.3. *For any vector $v \in \mathbb{R}^N$ and $w \in \mathbb{R}^{n+m-1}$, the products Av and $A^T w$ can be computed in $\mathcal{O}(n+m)$ time and using additional $\mathcal{O}(n+k)$ memory.*

Proof. The only nontrivial part of operations Av and $A^T w$ is computing Nx and $N^T y$ for vectors x, y of appropriate lengths. This can be done efficiently thanks to the representation $N = WU$ given by Lemma 4.2. To compute $h = Nx$, we first compute $u = Ux$ which can be done in $\mathcal{O}(n)$ time, without the need to explicitly store the matrix U, by computing suffix sums of vector x. We need $\mathcal{O}(n)$ memory to store the result of this operation. Next, we compute $h = Wu$. Thanks to the sparse representation of W, this multiplication can be done in $\mathcal{O}(m)$ time and we need $\mathcal{O}(k)$ memory for storing the result. This gives a total of $\mathcal{O}(n+m)$ time and $\mathcal{O}(n+k)$ memory complexity.

Similarly, to compute $g = N^T y$, we first multiply y by W^T and then multiply the resulting vector of length n by U^T, which corresponds to computing prefix sums, and does not require an explicit construction of U. In this case, we need $\mathcal{O}(n+m)$ time and $\mathcal{O}(n)$ memory. □

We are left with the task of solving a system of linear equations $\Sigma v = w$. Recall that $\Sigma = AZ_t^{-1}X_tA^T$ and note that Z_t^-1 and X_t are diagonal matrices with positive elements on their diagonals. Therefore, we can generalize our task and deal with solving systems of linear equations for matrices Σ of a form $\Sigma = AHA^T$, where H is a diagonal matrix with positive elements on the diagonal. To prove that this can be done efficiently, we first prove

Lemma 4.4. *For any diagonal matrix G of size $n \times n$, the matrix NGN^T can be computed in $\mathcal{O}(m(k+m)+n)$ time and using $\mathcal{O}(m^2)$ memory.*

Proof. We start with representing NGN^T as $WUGU^TW^T$. It is straightforward to check that for $i, j \leq n$ we have $(UGU^T)_{i,j} = \sum_{l=max_{i,j}}^{n} G[l,l]$. Therefore, if we compute the suffix sums of the diagonal of G in time and space $\mathcal{O}(n)$, we can retrieve $(UGU^T)_{i,j}$ for any i, j in constant time. Denote $\alpha_{i,j} = (UGU^T)_{i,j}$. We now have:

$$(NGN^T)_{i,j} = \sum_{p \leq n} \sum_{q \leq n} w_{i,p}\alpha_{p,q}w_{j,q}.$$

Using the sparse representations of rows W_i and W_j, the above sum can be computed in time $\mathcal{O}(m_i m_j)$. There is, however, a faster way. Notice that for $q \geq p$ we have $\alpha_{p,q} = \alpha_{q,q}$, and write:

$$\sum_{1 \leq p \leq q \leq n} w_{i,p}\alpha_{p,q}w_{j,q} = \sum_{1 \leq p \leq q \leq n} w_{i,p}\alpha_{q,q}w_{j,q} \tag{4.12}$$

The above sum can be computed in $\mathcal{O}(m_i + m_j)$ time. Let L_i, L_j be the lists containing the sparse representations of W_i, W_j. We first compute the suffix sums of $\alpha_{q,q}w_{q,j}$ for $(q, w_{j,q}) \in L_j$, which can be done in $\mathcal{O}(m_j)$ time since the list L_j is ordered by the column number q. Then, for any $(p, w_{i,p}) \in L_i$, we add to the result the suffix sum of $\alpha_{q,q}w_{q,j}$ for the smallest $q \geq p$ such that $(q, w_{j,q}) \in L_j$. Since the lists L_i and L_j are ordered by column numbers p, q, this can be done in $\mathcal{O}(m_i + m_j)$ time in standard way. Therefore, the sum in equation (4.12) can be computed in $\mathcal{O}(m_i + m_j)$ time. Similarly, the sum:

$$\sum_{1 \leq q < p \leq n} w_{i,p}\alpha_{p,q}w_{j,q} = \sum_{1 \leq q < p \leq n} w_{i,p}\alpha_{p,p}w_{j,q}$$

can be computed in $\mathcal{O}(m_i + m_j)$ time. Therefore, after we compute the suffix sums of the diagonal of G in time and memory $\mathcal{O}(n)$, the cell (i, j) of matrix NGN^T can be computed in $\mathcal{O}(m_i + m_j)$ time. We have

$$\sum_{1 \leq i,j \leq n} m_i + m_j = (2k-1)m$$

and therefore the whole matrix NGN^T can be computed in $\mathcal{O}(km + n)$ time and $\mathcal{O}(k^2 + n)$ memory. □

We can now prove that the normal equation $\Sigma d_y = r$ can be solved efficiently for any right hand side.

Lemma 4.5. *Let $r \in \mathbb{R}^{n+k-1}$ be a vector and let H be a $(2n+k) \times (2n+k)$ diagonal matrix with positive elements on the diagonal. Then, the matrix AHA^T has a full rank, and the equation $AHA^Tv = r$ can be solved with $\mathcal{O}(k^3 + km + n)$ time and $\mathcal{O}(k^2 + n)$ memory complexity.*

Proof. First, we present a useful decomposition of the matrix AHA^T. Let $H = \operatorname{diag}(H_1, H_2, H_3)$ where H_1, H_2, H_3 are diagonal matrices of sizes n, n, k respectively. Then, using the definition of A, we can write:

$$AHA^T = \begin{bmatrix} J_n(H_1 + H_2)J_n^T & J_n(H_2 - H_1)N^T \\ N(H_2 - H_1)J_n^T & N(H_1 + H_2)N^T + H_3 \end{bmatrix}. \tag{4.13}$$

We will now use a block-wise LDU decomposition of the right hand side of the above equation. For a given matrix $B = \begin{bmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{bmatrix}$ with an invertible block $B_{1,1}$, the block-wise LDU decomposition of B is:

$$B = \begin{bmatrix} I & 0 \\ B_{2,1}B_{1,1}^{-1} & I \end{bmatrix} \begin{bmatrix} B_{1,1} & 0 \\ 0 & B_{2,2} - B_{2,1}B_{1,1}^{-1}B_{1,2} \end{bmatrix} \begin{bmatrix} I & B_{1,1}^{-1}B_{1,2} \\ 0 & I \end{bmatrix}$$

Note that $J_n(H_1 + H_2)J_n^T$ is a diagonal $(n-1) \times (n-1)$ matrix with positive elements on the diagonal, and is therefore invertible and easy to invert numerically. Denote $K = J_n(H_1 + H_2)J_n^T$, $L = J_n(H_2 - H_1)$ and $G = H_1 + H_2 - L^T K^{-1} L$. Using the block LDU decomposition for (4.13), we get

$$AHA^T = \begin{bmatrix} I_{n-1} & 0 \\ NL^T K^{-1} & I_k \end{bmatrix} \begin{bmatrix} K & 0 \\ 0 & NGN^T + H_3 \end{bmatrix} \begin{bmatrix} I_{n-1} & K^{-1}LN^T \\ 0 & I_k \end{bmatrix} =: LMR$$

From this representation it follows that AHA^T has a full rank as a product of full-rank square matrices L, M, R. The L and R matrices have full rank as upper-triangular matrices with ones on their diaginals. To see that M also has a full rank, observe that G is a diagonal matrix with positive elements. This follows from the fact that, element-wise, we have $G \geq H_1 + H_2 - (H_2 - H_1)^2/(H_1 + H_2)$ with equality on all elements except the right bottom row. On the other hand, $H_1 + H_2 - (H_2 - H_1)^2/(H_1 + H_2) = 4H_1H_2/(H_1 + H_2)$, which is a diagonal matrix with positive elements. Therefore, NGN^T is non-negative definite, and since H_3 is positive definite, we conclude that $NGN^T + H_3$ is positive definite. Now, since K is positive definite as well, M has a full rank, and so does AHA^T.

For given vector r, we now need to solve the equation $LMRv = r$. The way to do this is to first solve the equation $Lv_1 = r$, then the equation $Mv_2 = v_1$ and lastly $Rv = v_2$ to obtain $LMRv = LMv_2 = Lv_1 = r$. We can therefore work with each matrix L, M, R separately.

To solve $Lv_1 = r$, write $r = (r_1, r_2)$, where r_1 is of length $n-1$ and r_2 has length k. Then,

$$v_1 = \begin{pmatrix} r_1 \\ r_2 - NL^T K^{-1} r_1 \end{pmatrix},$$

which can be computed efficiently since the cost of multiplying a vector by $L^T K^{-1}$ is $\mathcal{O}(n)$, and we can efficiently multiply vectors by N thanks to Lemma 4.3. We can solve the equation $Rv_3 = v_2$ analogically, with the same time complexity since we can efficiently multiply by N^T as well.

The only thing left is solving $Mv_2 = v_1$. Assume $v_1 = (u_1, u_2)$, where u_1 is of size $n-1$ and u_2 is of size k. Let $P = NGN^T + H_3$. Then, P is a positive definite matrix of size $k \times k$. Thanks to Lemma 4.4, we can compute it in $\mathcal{O}(km + n)$ time. Let v' be a solution of $Pv' = u_2$, which we find using standard methods in time $\mathcal{O}(k^3)$ and space $\mathcal{O}(k^2)$. Then, $v_2 = (u_1^T, v'^T)^T$ is the solution of equation $Mv_2 = v_1$.

Summing up, the cost of finding a solution to the equation $AHA^T v = r$ is $\mathcal{O}(k^3 +$

$km + n)$, where $\mathcal{O}(k^3)$ is the cost of solving a linear equation with a $k \times k$ matrix and $\mathcal{O}(km + n)$ is the cost of creating this matrix. All other operations have costs linear in n, k, m. We also need $\mathcal{O}(k^2)$ memory for computing the $k \times k$ matrix. All other operations can be done using additional memory linear in n, k. □

We can now state the computational complexity of Algorithm 4. For each iteration, we need to solve a finite, bounded number of normal equations (possibly more than one, depending on our mechanism of choosing scaling factor) and additionally perform a finite, bounded number of multiplications of type Av, $A^T v$, Nv and $N^T v$. Since all the other operations can be done in time and memory linear in n, k, m, we conclude that one iteration of our primal-dual Interior Point Method can be done in $\mathcal{O}(k^3 + km + n)$ time and $\mathcal{O}(k^2 + n)$ memory.

4.5 Summary of the Chapter

In this Chapter, we have proposed a formalization of the Mass Spectral Regression (MSR) problem, which encompasses separating overlapping isotopic envelopes, deisotoping, and decharging. We have shown that the Wasserstein distance can be used to effectively solve this problem in the presence of measurement inaccuracies.

The proposed solution for MSR works for a wide range of m/z values and multiply charged ions. Furthermore, it is not limited to a single class of compounds like peptides or metabolites. In principle, the theoretical isotopic envelopes can be either predicted *in silico* or measured experimentally. However, the main limitation of the proposed method comes from the fact that it needs to be used on spectra of highly purified samples, as it is sensitive to the presence of contaminating signals. We will deal with this limitation in Chapter 5.

Chapter 5

Regression of noisy spectra

If all the signal of an experimentally observed spectrum μ can be explained by a model spectrum $\nu_p = p_1\nu_1 + \cdots + p_k\nu_k$, the regression problem can be expressed as finding proportions p^* such that

$$p^* = \operatorname*{arg\,min}_{p \in \Delta_{k-1}} W(\mu, \nu_p).$$

In Chapter 4, we have shown that this approach yields accurate results when the only differences between μ and ν_p are caused by mass measurement errors and differing resolution. However, experimentally obtained spectra most often contain signals which are not theoretically predicted, like chemical contaminants or background noise. Such signals strongly disturb the optimal transport plan, leading to incorrect estimation of the proportions p^*. This is because the Wasserstein distance requires both spectra to have equal amounts of intensity, enforcing $p_1 + \cdots + p_k$ to be equal to 1.

In this Chapter, we assume that the observed spectrum can be approximated by the model spectrum with some additional chemical and/or background noise ε, so that for any m/z value x we have

$$\mu(x) = \nu_p(x) + \varepsilon(x). \tag{5.1}$$

The total signal of ν_p is now equal to the proportion of μ that is explained by the model. Therefore, in this Chapter, the model spectrum ν_p is not assumed to be normalized, and its total ion current may be less (but not greater) than one.

The contents of this Chapter. In this Chapter, we further investigate the problem of linear regression of mass spectra. We analyze the case of noisy spectra of mixtures of chemical compounds, which were identified as a potential source of estimation errors in Chapter 4. We develop a method that is robust against interfering signals coming from chemical impurities and background noise, at the same time being robust against measurement inaccuracies and numerical errors of peak picking algorithms thanks to using the Wasserstein distance.

We note that spectral regression methods based on the ordinary least squares regression are also naturally robust to contaminating signals. The main novelty in our approach to linear regression, compared to the existing methods, is the use of the Wasserstein distance, which is naturally robust to uncertainties in m/z measurements and different resolutions of the compared spectra. In particular, to our knowledge, this is the first algorithm that is capable of explaining an experimental spectrum in profile mode using a set of computationally generated, infinitely resolved theoretical spectra. On the other hand, when the experimental spectrum is analyzed in centroid mode, our algorithm does not require peak matching, and is therefore

capable of utilizing the full information in both the experimental and the theoretical spectra.

We test the performance of our Wasserstein regression algorithm on a custom-made data set consisting of 200 repeated measurements of the same set of compounds. This allows us to assess both the accuracy and variance of the estimation of ion proportions. We further confirm our results using simulated data sets, where we take into account several measurement inaccuracies occurring naturally in mass spectra.

5.1 Wasserstein regression of noisy spectra

To account for the additional signal in μ, we extend the m/z axis by adding an auxilliary point ω onto which such signal can be transported. All the signal transported onto ω is assumed to be unexplained by the model spectrum, and usually treated as noise that can be removed from μ. This applies regardless of the nature of the transported signal, i.e. whether it is a contaminant or background electronic noise. The idea behind this approach is visualized in Fig. 5.1.

We assume that the cost of transporting signal from μ to ω does not depend on the m/z value of the signal (equivalently, that ω is equidistant to all points on $\mathbb{R}$). We denote as κ the cost of transporting signal from μ to ω (equivalently, the distance between ω and any point on $\mathbb{R}$). Since κ is interpreted as the cost of removing noisy signal from μ, we refer to it as the *denoising penalty*.

Let $p = (p_1, p_2, \ldots, p_k)$ be a vector of non-negative *weights* or *proportions* such that $p_1 + \cdots + p_k \leq 1$ (note the inequality sign). Define ν_p to be a linear combination of the theoretical spectra with weights p, i.e.

$$\nu_p = p_1\nu_1 + \cdots + p_k\nu_k.$$

We will call ν_p a *model spectrum* in analogy to a model matrix used in an ordinary least squares linear regression. Note that the ν_p's intensity sums up to $p_1 + \cdots + p_k$, which may be less than 1. The inequality is allowed in order to account for the fact that ν_p may not explain all of μ's signal. We define the Wasserstein regression of mass spectra, accounting for signal unexplained by the model spectrum, as

$$p^* = \underset{p \in \Delta_k}{\arg\min}\, W(\mu, p_0\omega + \nu_p), \tag{5.2}$$

where p_0 is the amount of the unexplained signal in μ transported onto ω. Note that we now have $k+1$ weights, with $p_0 + p_1 + \cdots + p_k = 1$.

During the minimization of the Wasserstein distance in Equation 5.2, the observed signal that cannot be feasibly transported onto any theoretical spectrum gets transported onto ω. An important property of this approach is that the decision whether to remove a given signal is performed independently for each intensity measurement in the experimental spectrum. This allows it to work properly in the case when the number of the background noise peaks greatly exceeds the number of the peaks of the molecules of interest, which is usually the case in mass spectrometry. For example, the noise peaks in the middle of Fig. 5.1 are going to be removed (i.e. transported into ω) regardless of their number.

Note that the denoising penalty κ admits a physical interpretation in terms of m/z units. Transporting intensity from the experimental to the model spectrum over a distance larger than κ is more costly than removing it by transporting it to ω. The

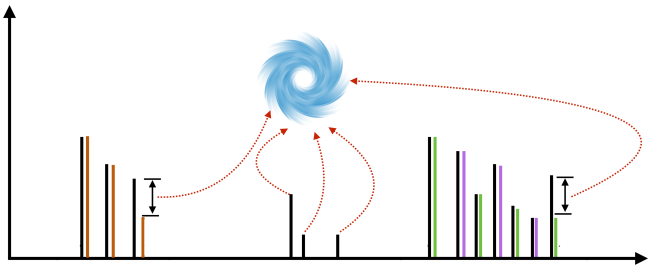

FIGURE 5.1: An illustration of the Wasserstein regression of mass spectra. An experimental spectrum (black) is explained by a set of three theoretical spectra (orange, green and magenta). Excess signal from the theoretical spectrum, occurring due to background noise or sample impurities, is transported onto an auxilliary point ω, represented as the vortex.

penalty can therefore be treated as a maximum distance over which the transport is feasible. This allows for some intuition behind the optimal values of this parameter: the maximum feasible transport distance should be set as the smallest value that allows to match corresponding theoretical and experimental peaks. The instrument accuracy (in Dalton units) is therefore an example of a reasonable value for the κ parameter. In practice, however, the choice of this parameter is more complicated. This is because, apart from the instrument accuracy, there are several factors that influence the distance between experimentally observed peaks and their theoretical counterparts, including, but not limited to, the resolving power. Therefore, usually the results for several different values of κ need to be inspected manually. This issue is discussed in more detail in the subsequent Sections.

A major advantage of our approach, as opposed to matching peaks by mass windows in methods based on least squares regression, is that κ does not set a hard threshold on the transport distance, allowing for more flexibility in the choice in this parameter. Furthermore, in some cases it may be beneficial to transport the signal over distances larger than κ, while in other cases interfering signal is removed regardless of its proximity to theoretical peaks. Specifically, whether a given signal is removed depends not only on it's distance from the nearest theoretical peaks, but also on the shapes of the theoretical isotopic envelopes—in contrast to methods based on linear regression, in which a signal is always incorporated when it's sufficiently close to any theoretical peak. This phenomenon is illustrated by the following example.

5.1.1 A worked example

Consider a theoretical spectrum ν consisting of n peaks, and let the i-th peak be at m/z $m_1 + i/q$ for some $m_i \geq 0$, and let it have intensity a_i, where $a_1 + a_2 + \cdots + a_n = 1$. This models a low-resolution theoretical spectrum of an ion with charge q and with the monoisotopic peak at m_1. Since the proportions a_i are arbitrary, this model encompasses all possible single-ion low-resolution spectra, allowing us to obtain a general result.

Let the experimental spectrum μ be equal to the theoretical one scaled by $1 - \epsilon$, with an additional noise peak in location $x \leq m_1$ with intensity ϵ (see Fig. 5.2). We will investigate which values of κ allow to correctly identify the additional peak as noise, and return the proportion of ν as $1 - \epsilon$ instead of 1. To this end, we will compare the Wasserstein distance $W(\mu, \nu)$ with the cost of denoising κ. Note that,

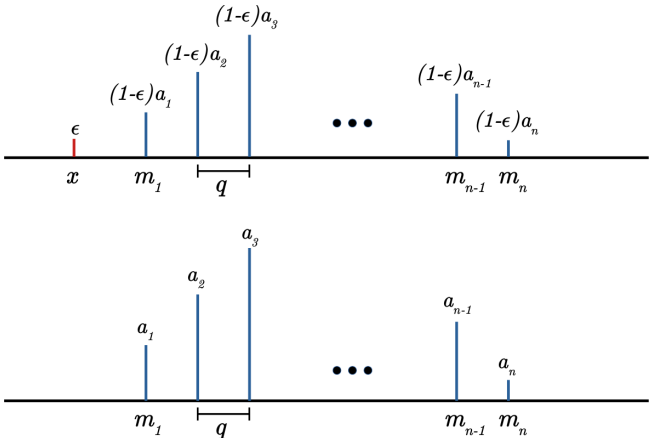

FIGURE 5.2: An example pair of experimental and theoretical spectrum, where the former (top) is equal to the latter (bottom) plus an additional noise peak.

since the spectra are identical save for the noise peak, if this peak gets removed then κ is the full cost of regression (i.e. we get $W(\mu, p_0\omega + \nu_p) = \kappa$).

Let A_i be the cumulative intensity of the theoretical spectrum ν. The Wasserstein distance $W(\mu, \nu)$ is equal to

$$\begin{aligned} W(\mu,\nu) &= \epsilon(m_1 - x) + \sum_{i=1}^{n-1} \frac{1}{q} \left| A_i - (1-\epsilon)A_i - \epsilon \right| \\ &= \epsilon(m_1 - x) + \sum_{i=1}^{n-1} \frac{1}{q} \left| \epsilon A_i - \epsilon \right| \\ &= \epsilon(m_1 - x) + \frac{\epsilon}{q} \sum_{i=1}^{n-1} (1 - A_i) \\ &= \epsilon(m_1 - x) + \frac{\epsilon}{q} (n - \sum_{i=1}^{n-1} A_i). \end{aligned}$$

Now, note that $\sum_{i=1}^{n-1} A_i = \sum_{i=1}^{n-1} (n-i)a_i = n(1-a_n) - \sum_{i=1}^{n-1} i a_i$, where the first equality comes from counting the number that each of the a_i variables occur in the sum, and the second one follows from the fact that the intensities sum to 1. This can be further simplified by noting that $n(1-a_n) - \sum_{i=1}^{n-1} i a_i = n - \sum_{i=1}^{n} i a_i$, where we included the term $n a_n$ in the sum.

Now, note that since the intensities are normalized,the average mass of the spectrum ν, denoted as $\bar{m}$, is equal to $\sum_{i=1}^{n} (m_1 + i/q) a_i$. This allows us to express $\sum_{i=1}^{n} i a_i$ simply as $q(\bar{m} - m_1)$, i.e. the distance between the average and the monoisotopic mass multiplied by the charge. Plugging this into the equation for $W(\mu, nu)$, we get a simple formula

$$W(\mu,\nu) = \epsilon(m_1 - x) + \epsilon(\bar{m} - m_1).$$

The cost of removing the noise peak will be lower than the cost of such transport whenever $\epsilon(m_1 - x) + \epsilon(\bar{m} - m_1) > \kappa$. It follows that the influence of the noise peak on the estimated proportions depends not only on it's proximity to the signal, $m_1 - x$, as is the case in methods based on linear regression, but also on the width of the isotopic envelope, $\bar{m} - m_1$. Note that the latter term is always positive for spectra

with more than one peak. Therefore, the shape of the isotopic envelope facilitates the detection of noise peaks by the linear regression procedure based on the Wasserstein distance.

The presence of the noise peak changes the proportion of ν from 1 to $1-\epsilon$, even though it is the only isotopic envelope in μ. This is because we estimate the amount of the total signal in μ that, under a given value of κ, can be feasibly explained by the model spectrum, and not the amount of ions. To obtain the latter, the proportions of the theoretical spectra returned by our method need to be normalized so that they sum up to 1.

5.1.2 Computation of the optimal proportions.

The formulation of the regression problem discussed in the previous Sections is well suited for theoretical analysis. However, it does not show how to obtain the optimal proportions in practice. Below, we show an equivalent formulation that is better suited for implementation and practical applications. The formal proof that the two formulations are equivalent is presented in Section 5.5.

Let $M(t)$, $N_j(t)$ be the cumulative distribution functions of μ and ν_j respectively. The cumulative distribution function of the model spectrum $\nu_p = \sum_{j=1}^{k} p_j \nu_j$ is equal to $N_p(t) = \sum_{j=1}^{k} p_j N_j(t)$.

Let $g(s_i)$ for $i = 1, 2, \ldots, n$ be the amount of μ's signal at the point s_i that is transported onto ω under a given transport plan. Note the distinction between g and ω: the latter is an auxilliary spectrum, therefore it's a concept analogous to μ and ν_p; the former denotes the amount of signal transported to ω, therefore being analogous to γ. Let $G(s_i) = g(s_1) + \cdots + g(s_i)$ be a cumulative distribution function of the unexplained signal. Note that $G(s_n) = g_{s_1} + \cdots + g_{s_n} = p_0$. Conceptually, G is a different construct than M or N, since it does not denote the amount of signal present in any given spectrum, but rather the amount of signal removed from μ by transporting it onto ω. In particular, we need to have $g(x) \leq \mu(x)$ for any point x.

For centroided spectra, the optimization problem (5.2) can be rephrased as a minimization over the variables p and g as follows:

$$p^*, g^* = \arg\min_{p,g} \left\{ \kappa p_0 + \sum_{i=1}^{n-1} (s_{i+1} - s_i) \Big| M(s_i) - G(s_i) - N_p(s_i) \Big| \right\}. \tag{5.3}$$

Note that the minimization above is performed over both p_j and $g(s_i)$ variables. Moreover, we require $p_0 + \sum_{j=1}^{k} p_j = \sum_{i=1}^{n} g(s_i) + \sum_{j=1}^{k} p_j = 1$, as all the signal in the observed spectrum needs to either be explained by the theoretical spectra or labeled as unexplained. The sum of p_j variables denotes the proportion of signal explained by the model spectrum.

Note that the optimization problem (5.3) is similar to formula (3.4) from Theorem 3.2, expressing the Wasserstein distance between a pair of spectra. It is equivalent to computing the Wasserstein distance between the model and the experimental spectrum after removing the unexplained signal from the latter, and additionally penalizing for the amount of signal removed. However, it does not give any clear physical interpretation of the denoising penalty κ.

The minimization problem (5.3) is an example of the Least Absolute Deviation (LAD) regression [75]. One of the common approaches to solving such problems is by using the technique of linear programming, which optimizes a linear function under a set of linear constraints [74]. However, the function to be minimized in

problem (5.3) is not linear due to the absolute value, and the problem needs to be converted before using linear programming to solve it. There are several ways to convert a Least Absolute Deviation regression problem into a linear program, and choosing the proper one in each particular application is a crucial factor in obtaining a computationally efficient solution. We investigated several approaches and found that, in the case of linear regression of mass spectra, the approach based on the ideas described in [76] seems to be the most efficient.

The derivation of the linear program equivalent to the minimization problem (5.3) is shown in Section 5.5. The main idea is to show that solving the regression problem is equivalent to solving the following linear program for the vector of variables z, and analyzing the differences between left- and right-hand sides of the inequalities:

$$\begin{array}{lrcll} \text{maximize} & V^T z & \text{over} & z & \\ \text{subject to} & W^T z & \leq 0, & & \\ & z_i - z_{i+1} & \leq & s_{i+1} - s_i, & i = 1, 2, \ldots, n-1, \\ & z_i - z_{i+1} & \geq & s_i - s_{i+1}, & i = 1, 2, \ldots, n-1, \\ & & & z \leq \kappa, & \end{array} \tag{5.4}$$

Above, V is a vector of the observed spectrum intensities, such that $V_i = \mu(s_i)$, and W is a matrix of theoretical intensities, $W_{ij} = \nu_j(s_i)$. Note that V and the columns of W may represent spectra in either centroided or profile mode, because both types of spectra are represented as finite lists of m/z and intensity measurements when stored on computers. Moreover, since we only use the vectors of intensities and m/z values, we do not need to compute any additional values prior to solving the program (such as the cumulative sums used in the Wasserstein distance).

The minimization problem above can be solved using a number of algorithms for linear programming, such as the Simplex or the Interior Point methods. In our implementation, available in our Python 3 package `masserstein`, we have used the Simplex algorithm.

As in the case of computing the Wasserstein distance, the regression algorithm can be applied to profile experimental spectra provided that the sampling of intensity values is uniform over the m/z axis. This is how the results shown in Fig. 4.1 were obtained, with $\kappa = 0.02$. According to our knowledge, the Wasserstein distance is the first solution that allows for this kind of processing without the requirement for peak detection or centroiding. In the next Section, we validate the performance of this approach on a set of experimentally obtained spectra as well as on simulated data.

5.1.3 The choice of the denoising penalty.

Our method requires the user to specify a single parameter: the cost of denoising κ, interpreted as the maximum feasible transport distance in the m/z domain. This interpretation of κ allows to treat it analogously to the radius of a mass window. However, a major difference between the two approaches is that κ acts as a *soft threshold*: during regression, whether some signal is transported between two peaks depends not only on their distance relative to κ, but also on the shape of the theoretical isotopic envelopes. This allows for some flexibility in setting the value of κ and makes the results more stable when the value is sub-optimal.

Similarly to the width of a mass window, in practical applications the choice of κ is not straightforward. The performance of the regression methods for a given parameter value is influenced by several factors, including the instrument accuracy

and resolving power. Usually, the results for several different values need to be inspected manually in order to make a final decision. The half-base widths of peaks in profile spectra and the instrument accuracy serve as a convenient reference for the reasonable range of values of this parameters. Setting κ higher than 1 is usually inadvisable, as it allows to transport the signal between peaks of ions with different atomic compositions.

The choice of κ resembles the classical variance-bias trade-off known from the field of machine learning. Small values of κ lead to down-estimation of ion proportions, as insufficient signal is available to be transported onto their theoretical isotopic envelopes. On the other hand, while high values of κ allow to transport all the required signal, they lead to incorporation of noise in the estimated proportions, therefore increasing the variance of the estimation. Additionally, if systematic sample impurities are present, high values of κ may lead to an over-estimation of the amounts of the molecules of interest. Further studies are needed in order to give precise guidance as to the optimal value of this parameter.

5.1.4 A note about experimental data.

In real mass spectra, there are several different sources of unexpected signal, including chemical contaminants, random baseline electronic noise etc. In this work we do not distinguish between them, and broadly classify the experimental signal into two classes: the signal of the molecules of interest, referred to as the explained or expected signal, and all the other signal, referred to as the unexplained signal or noise.

Naturally, in noisy spectra, it is often far from clear whether a given signal is expected or not. Accordingly, any numerical method may mistake actual signal of interest for noise, and the other way around. Therefore, the signal transported from μ onto ω is the noise estimated by our method, which, as for any computational method, may differ from the actual noise signal.

One of the most difficult types of unexpected signal is the chemical contaminants with isotopic envelopes that highly overlap with the ones of the molecules of interest. In this case, the contaminating signal might be transported onto the theoretical envelope of the molecule of interest instead of ω. One of the possible ways of handling this problem is to use a database of common contaminants, such as the cRAP database[1], and include their theoretical isotopic envelopes in the model spectrum.

There are multiple factors other than the noise signal that influence the similarity of the observed spectra to their theoretical counterparts and the similarity between experimental spectra from replicate experiments. These include the variation of peak m/z position and the variation of peak intensities caused by random isotopologue sampling and measurement inaccuracies. Naturally, with any computational method, including ours, a change in the input will result in a change in the output. However, the variations in m/z and intensity values have different effects on the results obtained.

Due to the use of the optimal transport theory, our approach is inherently robust to the variation in the mass domain, as long as the parameter κ is properly adjusted and the variation is not excessive, as may be caused e.g. by an improper calibration of the instrument. Natural variations of m/z occurring in replicate experiments have close to no influence on the estimated proportions of molecules.

On the other hand, the variation in intensity has a pronounced effect on the obtained results, because it influences the amount of signal in the observed isotopic

[1] https://www.thegpm.org/crap/

envelope of a given molecule. The variability of the signal intensity is therefore reflected in the variability of the estimated proportions. In particular, low ion counts lead to a highly unstable estimation. An example of this phenomenon is discussed in the next Section.

Although our method is fairly robust against measurement inaccuracies and sample impurities, any computational method will give improper results if the quality of the data is insufficient. Therefore, it is always the responsibility of the spectrometrist and the data analyst to first inspect the spectrum visually in order to determine its quality.

5.2 Validation on experimental data

In this Section, we verify whether the regression algorithm based on the Wasserstein metric can accurately filter out the background noise and estimate ion proportions in experimental spectra. We also compare the performance of the method on centroid and profile spectra.

As our test dataset, we take 200 repeated measurements of Pierce® LTQ Velos ESI Positive Ion Calibration Solution (Thermo Scientific). This calibration mixture is composed of caffeine, a short peptide with the sequence MRFA, and Ultramark 1621, a compound shown in Fig. 5.3. Note that due to varying side group lengths, Ultramark 1621 is in fact a mixture of 13 different compounds. In this dataset, all the isotopic envelopes of the compounds of interest are disjoint. We simulate the effect of overlapping envelopes by superposing shifted spectra.

Due to the different ionization rates of different compounds, it is difficult to predict the correct proportions of ion signals from their concentrations in samples. Therefore, we have assumed that our ground truth to which we compare our estimates are the true signal areas of compounds. To obtain the latter, we have manually selected the signal regions of all compounds, taking into account their monoisotopic peaks and peaks of isotopologues containing one additional neutron (i.e. first isotopic peaks). The region selection was performed on an average spectrum based on the 200 profile spectra. To ensure that the selected regions contain whole peaks in all the spectra, we have additionally taken into account the standard deviation of intensity at each point. The selected regions are shown in Table 5.1, and a fragment of the average spectrum used to select them in Fig. 5.3. Next, for each spectrum we have integrated the signal within the selected regions using the trapezoidal rule. For each compound, the signals of its monoisotopic and first isotopic peaks were summed to yield the total signal of the compound.

We have inspected the results of our Wasserstein regression method for several manually selected values of κ. The selection was based on the observed peak widths in profile spectra, which ranged from 0.013 at 195 Da to 0.38 at 1725 Da. In principle, setting κ to half base width of the broadest observed peak allows the method to use all the necessary experimental signal. However, we have observed that the performance was best for $\kappa = 0.4$, i.e. the full base width of the broadest peak. This value allowed for a more flexible transport of signal between peaks of an isotopic envelope, effectively countering the effect of the variance of experimental peak intensity. The effect is mostly pronounced for low intensity peaks with small signal to noise ratio, in which case the particularly high variance of signal intensity has a detrimental effect on the estimation unless κ is sufficiently high. The reason behind this effect is explained in more detail in the following paragraphs.

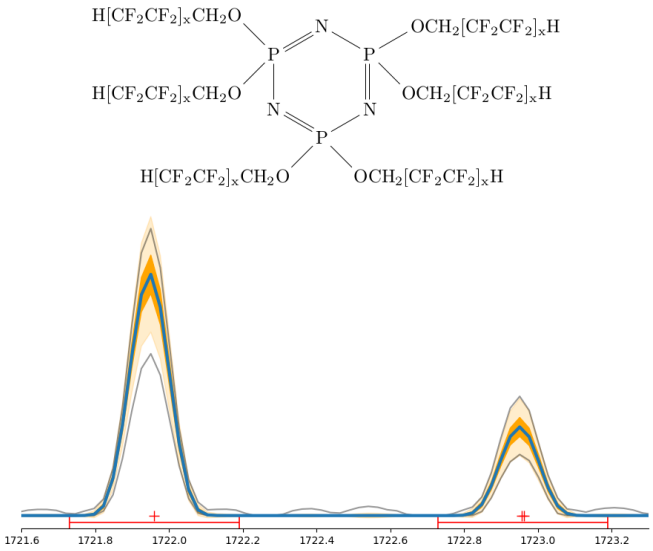

FIGURE 5.3: Top: Ultramark 1621, one of the compounds analyzed in this study; x=1,2,3. Bottom: A fragment of an average spectrum used to define m/z regions of ion signals, showing the first two peak of an isotopic envelope of Ultramark 1621 with 14 CF_2CF_2 groups. Blue line shows the average signal. Dark orange and light orange ribbons show $\pm\sigma$ and $\pm 3\sigma$ regions, where σ is the standard deviation of signal intensity. Grey lines show the maximum and minimum signal over the 200 spectra. Red dots show the m/z values of theoretically predicted peaks. As the theoretically predicted masses agree well with the signal apexes, we infer that the spectrum is properly calibrated. As the maximum and minimum lines are in proximity of the $\pm 3\sigma$ ribbon, we infer that there are no outlying measurements of intensity. Random increases of maximum signal between the peaks indicate the presence of background noise.

Name	Region
Caffeine	195.075 - 195.100 Da
Caffeine	196.073 - 196.103 Da
MRFA 2+	262.614 - 262.655 Da
MRFA 2+	263.118 - 263.154 Da
MRFA 1+	524.216 - 524.306 Da
MRFA 1+	525.217 - 525.320 Da
Ultramark 8x	1121.85 - 1122.13 Da
Ultramark 8x	1122.84 - 1123.15 Da
Ultramark 9x	1221.84 - 1222.12 Da
Ultramark 9x	1222.84 - 1223.15 Da
Ultramark 10x	1321.82 - 1322.15 Da
Ultramark 10x	1322.82 - 1323.15 Da
Ultramark 11x	1421.77 - 1422.17 Da
Ultramark 11x	1422.78 - 1423.16 Da
Ultramark 12x	1521.77 - 1522.16 Da
Ultramark 12x	1522.76 - 1523.17 Da
Ultramark 13x	1621.75 - 1622.19 Da
Ultramark 13x	1622.74 - 1623.19 Da
Ultramark 14x	1721.73 - 1722.19 Da
Ultramark 14x	1722.73 - 1723.19 Da
Ultramark 15x	1821.70 - 1822.17 Da
Ultramark 15x	1822.76 - 1823.17 Da

TABLE 5.1: Regions of the m/z axis used to compute ion signal intensities for the validation of the Wasserstein regression. Ultramark 8x denotes Ultramark 1621 with 8 CF_2CF_2 groups, etc.

5.2.1 Analysis of centroided spectra.

The first goal of this study is to verify whether the Wasserstein regression algorithm implemented in the `masserstein` package returns accurate results when applied to a centroided experimental spectrum. In order to perform the centroiding in a controlled manner, we have implemented our own peak-picking procedure. Briefly, the spectra are centroided by integrating the signals within regions delimited by a fraction of 0.2 of the apex intensity. The pseudo-code of the full procedure is shown in Algorithm 3 in Section 3.5.2 of Chapter 3. We have validated our implementation by comparing the centroided peak intensities with the manually integrated signal areas and found a good agreement.

We have generated the theoretical spectra of the compounds of interest using IsoSpecPy [72], assuming a proton adduct. In each spectrum we have retained the monoisotopic and the first isotopic peaks, so that the theoretical spectra correspond to the manually identified regions. After that, we have used `masserstein` to regress each of the 200 experimental spectra against our set of theoretical spectra. We have inspected several values of the denoising penalty κ ranging from 0.05 to 0.6. For $\kappa = 0.4$, the regression of all spectra took around 98 seconds on a single core of an Intel(R) Core(TM) i5-6300HQ CPU @ 2.30GHz processor

The `masserstein` package requires that all the spectra are normalized. Prior to normalization we have computed the total ion currents of the experimental spectra. After obtaining the estimated proportions of ions, they were multiplied by the total ion current of the experimental spectrum. This way we obtain the total signal (as opposed to the proportion of the total signal) explained by each ion, which we then compare to the manually calculated signal area.

The results for $\kappa = 0.4$ are shown in Fig. 5.4. The overall correlation between the

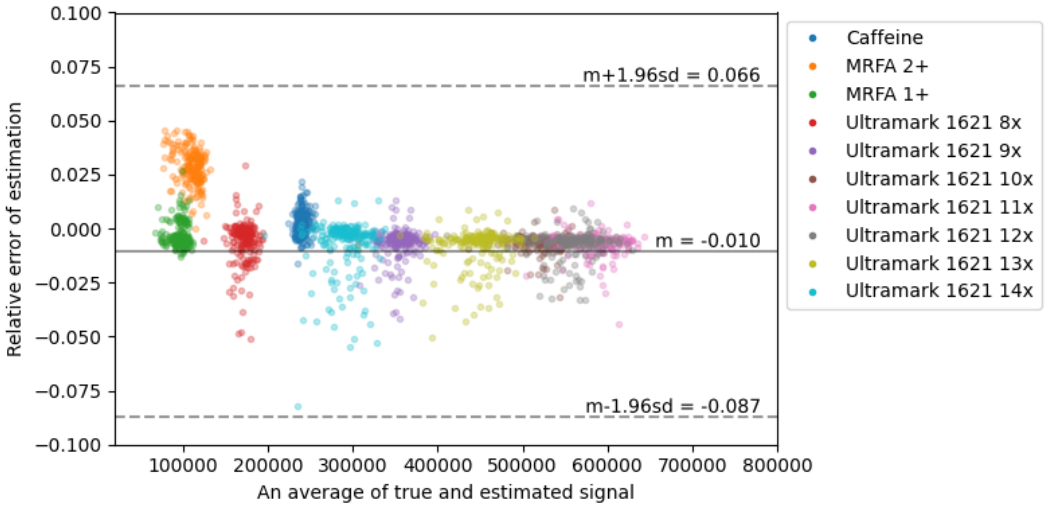

FIGURE 5.4: A Bland-Altman plot summarizing the Wasserstein regression results for the calibration mix on centroided spectra. Each point corresponds to an estimate of an ion intensity in one of the 200 spectra analyzed, with colors corresponding to different ions. The Y coordinate of each point corresponds to the estimated minus the true signal divided by the latter. Ultramark 1621 8x denotes Ultramark 1621 with 8 CF_2CF_2 groups, etc.

estimated and the manually computed signal was equal to $\rho = 0.9998$. The mean difference between the estimated and manually integrated signal relative to the latter was equal to 1%, indicating a slight downward bias for this value of denoising penalty. The mean absolute relative difference was equal to 0.017, meaning that the average error of the estimation is equal to 1.7% of the true value. The detailed results for each ion are shown in Fig. 5.5.

For $\kappa = 0.3$ the results were similar for most ions, and the overall correlation of the estimated and true signals was equal to $\rho = 0.9994$. However, we have found a strong down-estimation of the signal of Ultramark 1621 with 15 CF_2CF_2 side groups, which caused the drop in the correlation. We have found out that the bias is caused by a large variability of the relative peak intensities of this ion. In some spectra, the first isotopic peak was up to two times lower than on average. Such a large variability is most likely caused by a small number of ions of this compound.

Highly variable peak heights, combined with low denoising penalties, are detrimental to our current implementation of the Wasserstein regression. When the maximum feasible transport distance is low, the procedure necessarily fits to the smallest matching peak of the experimental spectrum. This is because, after all the signal of such peak is transported to a theoretical spectrum, there is no neighboring signal left that can be feasibly transported.

Increasing the denoising penalty allows to distribute the experimental signal more evenly over the theoretical spectrum, therefore increasing the accuracy of the estimation. However, when the penalty is too high, the background noise may also be transported to the theoretical spectrum, leading to an overestimation of an ion's signal.

The results demonstrate that the optimal transport theory applied to the problem of linear regression of mass spectra is capable of giving very accurate estimates of

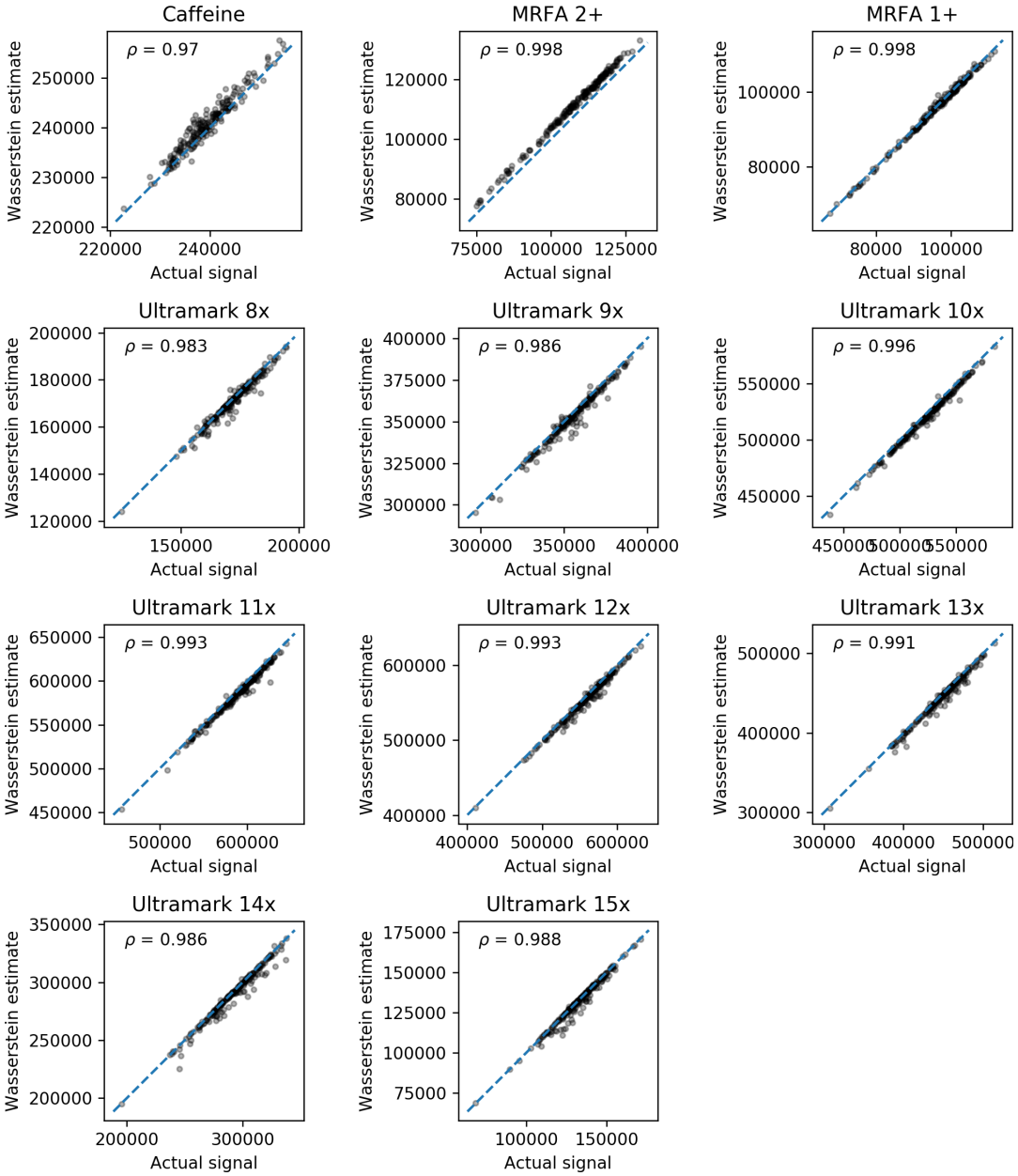

FIGURE 5.5: The Wasserstein regression results for 200 centroided spectra of Pierce® LTQ Velos ESI Positive Ion Calibration Solution (Thermo Scientific). The plots show ion intensities estimated by the linear regression based on the Wasserstein distance versus manually integrated peak areas. Each point corresponds to a mass spectrum. Numbers in top-left corners represent the Pearson correlation. Note the different scales in plots, due to different average signal intensities of different molecules.

ion signals, provided that all the considered signals are over the limit of quantification to ensure small variability of peak areas. As we show below, it also opens the possibility of directly analyzing the profile spectra without the peak-picking step.

5.2.2 Analysis of profile spectra.

In the current implementation of the Wasserstein regression algorithm, we treat profile spectra in the same way as the centroided ones. As discussed in the previous Section, this approach gives a good approximation when the signal sampling is uniform over the m/z axis. However, this is often not the case for spectrometers which have a non-constant resolving power. As peaks get broader with the increasing m/z value, less data points are needed to reflect the signal shape. This phenomenon is often exploited in order to decrease the data size. One of the way to circumvent this problem in order to use the current implementation of our method is to resample the signal intensities.

We have resampled our spectra using a piecewise linear interpolation, in which the signal intensity in each point is approximated by a weighted average of the neighbouring intensities. The full procedure is shown in pseudo-code in Algorithm 2 in Section 3.5.1 of Chapter 3. For each spectrum, we have interpolated its signal such that the spacing between neighboring m/z values was 0.001.

The downside of the resampling strategy is a large increase of the data complexity and, consequently, the computational time. For the regression of the 200 profile spectra with $\kappa = 0.4$, the computations took 35 minutes, compared to 98 seconds for centroided spectra.

The penalty $\kappa = 0.4$ yielded an overall correlation of 0.9998. The results for all compounds are shown jointly in Fig. 5.4, and for each ion in detail in Fig. 5.6. We have noticed a systematic slight overestimation of the caffeine signal, most likely due to incorporation of a small unidentified peak at 195.72 Da which was present in most spectra. For $\kappa = 0.3$ we have observed a bias in the estimation of Ultramark 1621 with 15 CF_2CF_2 groups, similar as for the centroided spectra.

Further information about the results of the regression can be obtained by inspecting the spectrum of the remaining signal (i.e. the signal not explained by the theoretical spectra). An example of such spectrum, obtained for one of the 200 mass spectra, is shown in Fig. 5.7. The remaining signal corresponds to approximately 45% of the total ion current. By inspecting the spectrum of the remaining signal we conclude that all the molecules of interest were properly detected. However, not all of the signal of interest was used for regression, most likely due to the peak height variability discussed in the previous paragraph. We also detect contaminating ions, one of which is visible at 1395 Da in the right panel of Fig. 5.7, and numerous apparently random peaks. On the other hand, inspecting the fragment with caffeine (not shown) confirmed our assumption that a contaminant with a highly overlapping isotopic envelope caused an overestimation of the signal. Further research into the application of optimal transport to the processing of mass spectra should allow for a better separation of the signal of interest from the contaminants.

In general, the results were similar to the ones obtained on centroided spectra. This shows that `masserstein` allows for the processing of profile spectra without the need of peak peaking. Further research into applications of the optimal transport theory to the processing of mass spectra has the potential to increase both the computational efficiency and the accuracy of the results.

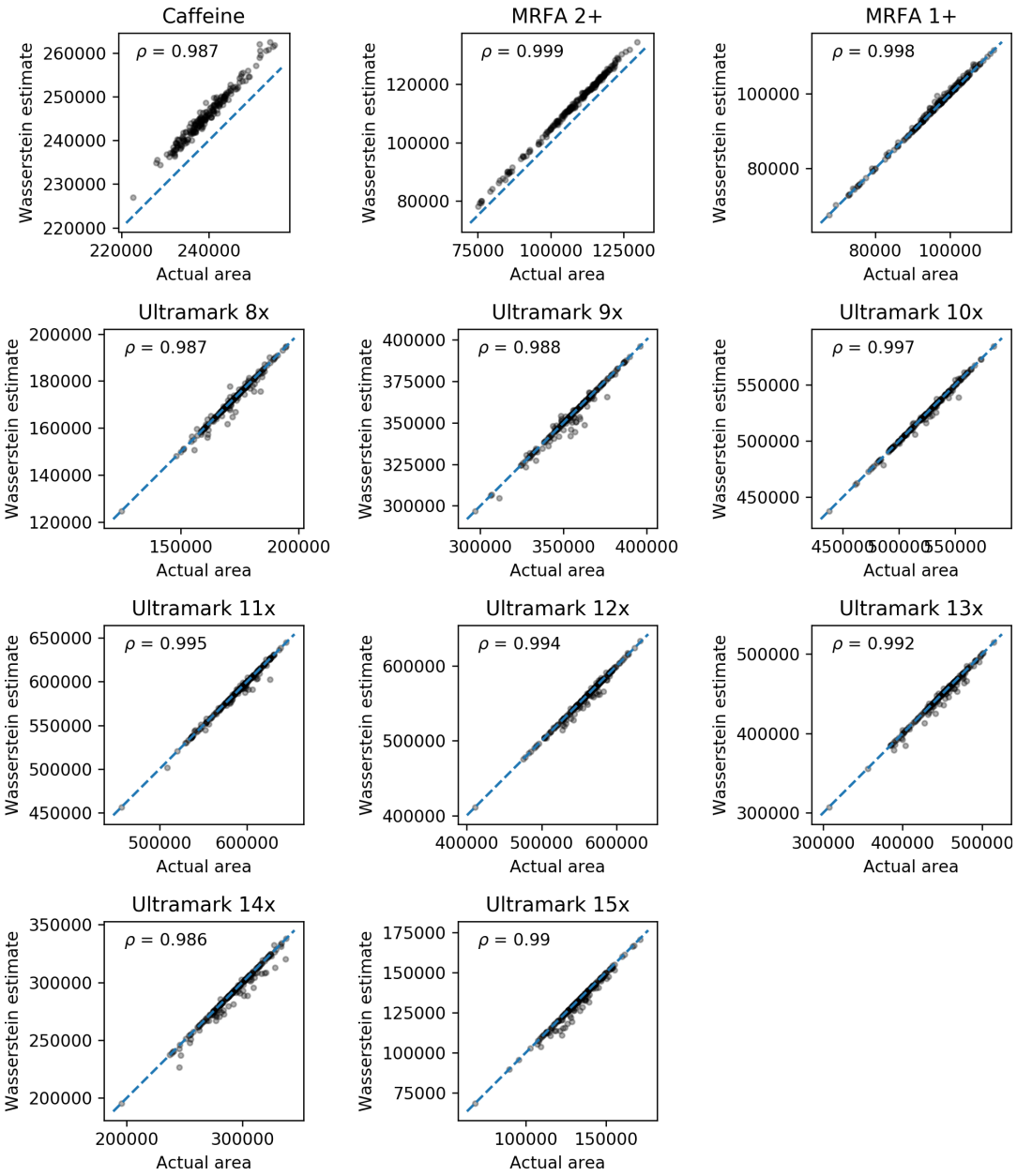

FIGURE 5.6: The Wasserstein regression results for 200 profile spectra of Pierce® LTQ Velos ESI Positive Ion Calibration Solution (Thermo Scientific). The plots show ion intensities estimated by the linear regression based on the Wasserstein distance versus manually integrated peak areas. Each point corresponds to a mass spectrum. Numbers in top-left corners represent the Pearson correlation. Note the different scales in plots, due to different average signal intensities of different molecules.

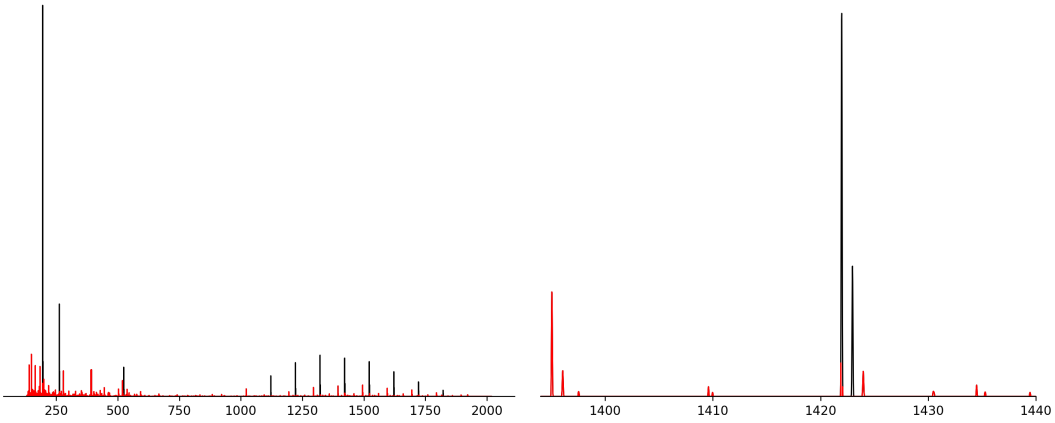

FIGURE 5.7: One of the 200 mass spectra used in this study with the signal remaining after regression with $\kappa = 0.4$ highlighted in red. Left: Full spectrum. Black peaks, corresponding to no remaining signal, indicate that the molecules of interest were properly detected. Right: A zoomed in fragment containing the isotopic envelope of Ultramark with 11 CF_2CF_2 groups at 1421.77 Da. The signal from the first isotopic peak was fully used for regression, while there is still some remaining signal of the monoisotopic peak. As expected, the signal from the second isotopic peak was not used, because this peak was discarded from the theoretical spectrum.

5.2.3 Overlapping isotopic envelopes.

In our final experiment on the calibration mix data, we investigate the influence of overlapping envelopes on the accuracy of the estimation. In the case of disjoint isotopic envelopes, like in the previous experiments, it is easy to obtain the ground truth by manually selecting peak regions for integration. However, the task is complicated when the peaks of the envelopes overlap, because manual integration does not allow to separate them and compute their individual signals. Therefore, we have decided to simulate this effect using the calibration mix data.

For each spectrum in profile mode, we have created its copy shifted by one hydrogen mass. Each spectrum was mixed with its shifted copy in proportion 0.7 of the original and 0.3 of the copy. The spectra were subsequently centroided as in the previous examples.

To generate a model spectrum, we have taken the formulas from the previous experiments, and the same formulas with one additional hydrogen. When adding a hydrogen atom to the sum formula, we only modified the ion's mass, while keeping its charge unchanged. Note that, from a computational perspective, it does not matter whether the formulas obtained this way correspond to any actual chemical compound.

From the computational point of view, this dataset is much more difficult than the previous ones, and we expect a decrease in estimation accuracy. One of the reasons is that when overlapping signals merge, the apex positions shift relative to the original ones. Therefore, peak picking a profile spectrum with overlapping signals returns distorted peak positions, and for a sufficiently large overlap one gets a single peak instead of two. This poses a major difficulty for approaches based on pointwise comparison of spectra, while the Wasserstein distance is more robust to such changes.

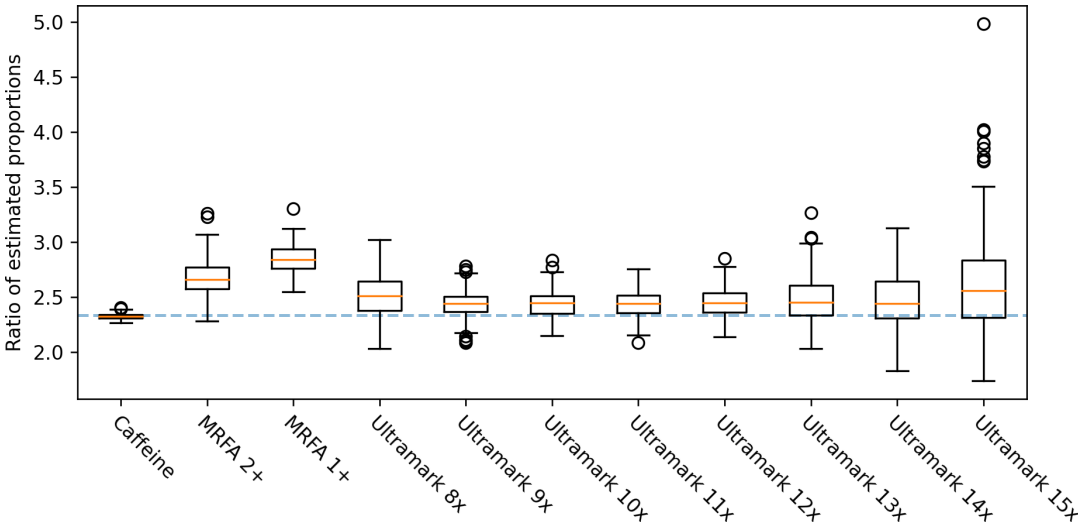

FIGURE 5.8: The ratio of estimated signals of ions with overlapping isotopic envelopes. Each boxplot corresponds to a pair of ions with sum formulas differing by one hydrogen atom. The dashed line represents the true ratio.

Using `masserstein`, we have regressed the mixed spectra with denoising penalty $\kappa = 0.4$. This time, the regression of all the spectra took approximately 3 minutes (not including the time needed for centroiding). The results were compared with the signal areas integrated in the previous experiment, rescaled either by 0.7 or 0.3 to accommodate for the mixing proportions. The correlation between the estimates of `masserstein` and the true signals was equal to $\rho = 0.9985$, only slightly smaller than for the previous datasets. The detailed results are shown in Fig. 5.9.

Additionally, we have calculated the ratios of estimated proportions of corresponding ions. In each spectrum we have compared the estimated signal of an original ion to the estimated signal of it's counterpart with one additional hydrogen atom. We have compared the ratio obtained this way to the reference value of $0.7/0.3 \approx 2.33$. The result is shown in Fig. 5.8.

We have observed that the ratio is overestimated for MRFA 1+. Comparing the ratio with the results shown in Fig. 5.9, we conclude that this is caused by an overestimation of the signal of the original MRFA 1+ ion (i.e. without the added hydrogen atom). For all the other ions, the true ratio was within the 95% confidence interval of the estimation. The estimated ratio showed a high variance for Ultramark 15x, which is likely caused by the lower signal to noise ratio of this ion compared to the other ions. On the other hand, the estimated signal ratio was the most accurate for caffeine, likely due to low amount of background noise in the neighbourhood of its isotopic envelope and a sufficiently high signal intensity.

The results of the three experiments presented in this Subsection show that our Wasserstein regression algorithm implemented in `masserstein` is capable of accurate estimation of signal intensity based on experimental spectra in both centroided and profile mode, and also in the presence of overlapping isotopic envelopes. However, the dataset presented in this Section contained only a handful of molecular formulas. In order to verify the results for a broader range of molecules, and to demonstrate that `masserstein` is not limited to any particular type of ions (be it lipids, peptides or metabolites), in the next Subsection we perform an extensive analysis on simulated datasets consisting of purely random molecular formulas.

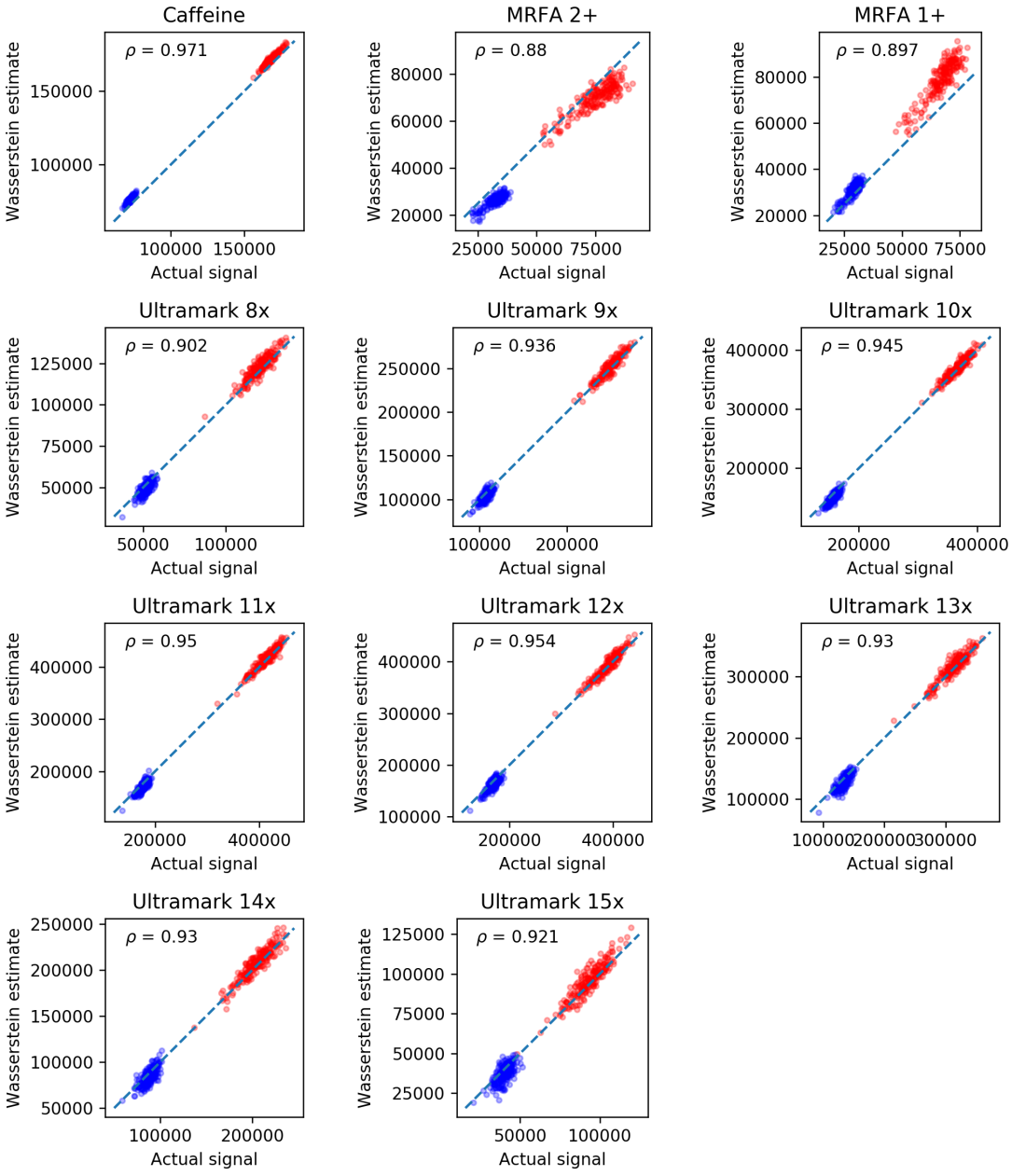

FIGURE 5.9: The Wasserstein regression results on 200 spectra of Pierce® LTQ Velos ESI Positive Ion Calibration Solution (Thermo Scientific) after introducing overlapping isotopic envelopes. Red: signals from original spectra; Blue: signals from spectra shifted by one hydrogen mass. Note the different scales on the X and Y axes.

5.3 Computational experiments on simulated data

In this Section, we evaluate the accuracy and bias of the estimation performed by `masserstein` using simulated mass spectra. We introduce a series of measurement distortions into the observed spectrum in order to reflect the variability of peak heights and limited resolving power and accuracy observed in real spectra.

We have created several simulated datasets by computing spectra of mixtures of randomly generated molecules. In each dataset, all the molecules had the same nominal mass in order to ensure high overlap of their isotopic envelopes. The molecules were simulated by subsequently sampling elements in the order of C, O, N, S, and P. Note that any user-supplied formula can be used in `masserstein`, as no mathematical procedure used in this work depends on the chemical properties of molecules. Therefore, in order to simplify the simulation procedure, we did not restrict the sampled molecules to ones which are chemically possible.

We have simulated datasets consisting of isobaric molecular formulas for a range of nominal masses from 60 to 12 000 Da and from 1 to 8 isotopic envelopes. Based on the formulas, we have generated theoretical spectra using IsoSpecPy [72]. To obtain simulated experimental spectra, for each dataset we have mixed the theoretical ones in random proportions. We have generated both centroid and profile experimental spectra. In the latter case, we have assumed a Gaussian shape of peaks.

To each experimental spectrum, in both profile and centroid mode, we have introduced extensive distortions to simulate the effects of a finite number of molecules, electronic noise, measurement inaccuracy in mass and intensity domain and limited resolving power (100 000 at 600 Da in the case of profile spectra). The procedure is described in detail in Section 5.6.

To assess the quality of regression results, we calculate the mean absolute deviation (MAD) between true and estimated signal contributions of the theoretical spectra:

$$\mathrm{MAD} = \frac{1}{k}\sum_{i=1}^{k} |\hat{p}_- p_i|$$

In the above formula, k is the number of theoretical isotopic envelopes, p_i is the true proportion of the i-th envelope, and $\hat{p}_i$ is its estimated proportion. We do not directly compare the amount of unexplained signal, p_0, as this information is implicitly included in the sum of the estimated proportions.

In the case of profile observed spectra, we have inspected two denoising penalties, equal 0.0075 (half peak base width) and 0.08. The results are shown in Fig. 5.10. For the lower penalty, the MAD was mostly between 10^{-2} and 10^{-1}, indicating at least one accurate decimal digit in an average estimate. We have observed some bias in the estimation, as the estimated proportions were usually lower than the true ones. The mean of the residues was equal to -0.0197, while their standard deviation was 0.025. The bias increased with increasing true proportion. This can be explained by the fact that the isotopic envelopes of abundant ions have a larger variance of their peak heights due to isotopologue sampling. For the 0.08 penalty, the MAD was smaller and usually around 10^{-2}. The decrease in MAD was especially pronounced for low numbers of isotopic envelopes. One of the reasons for better performance in this case is a lower bias, as the mean of the residues was -0.008. However, at the same time, their variance increased to 0.028. In all cases, the processing of a single spectrum took less than one second on a laptop computer with Intel(R) Core(TM) i5-6300HQ CPU @ 2.30GHz processor.

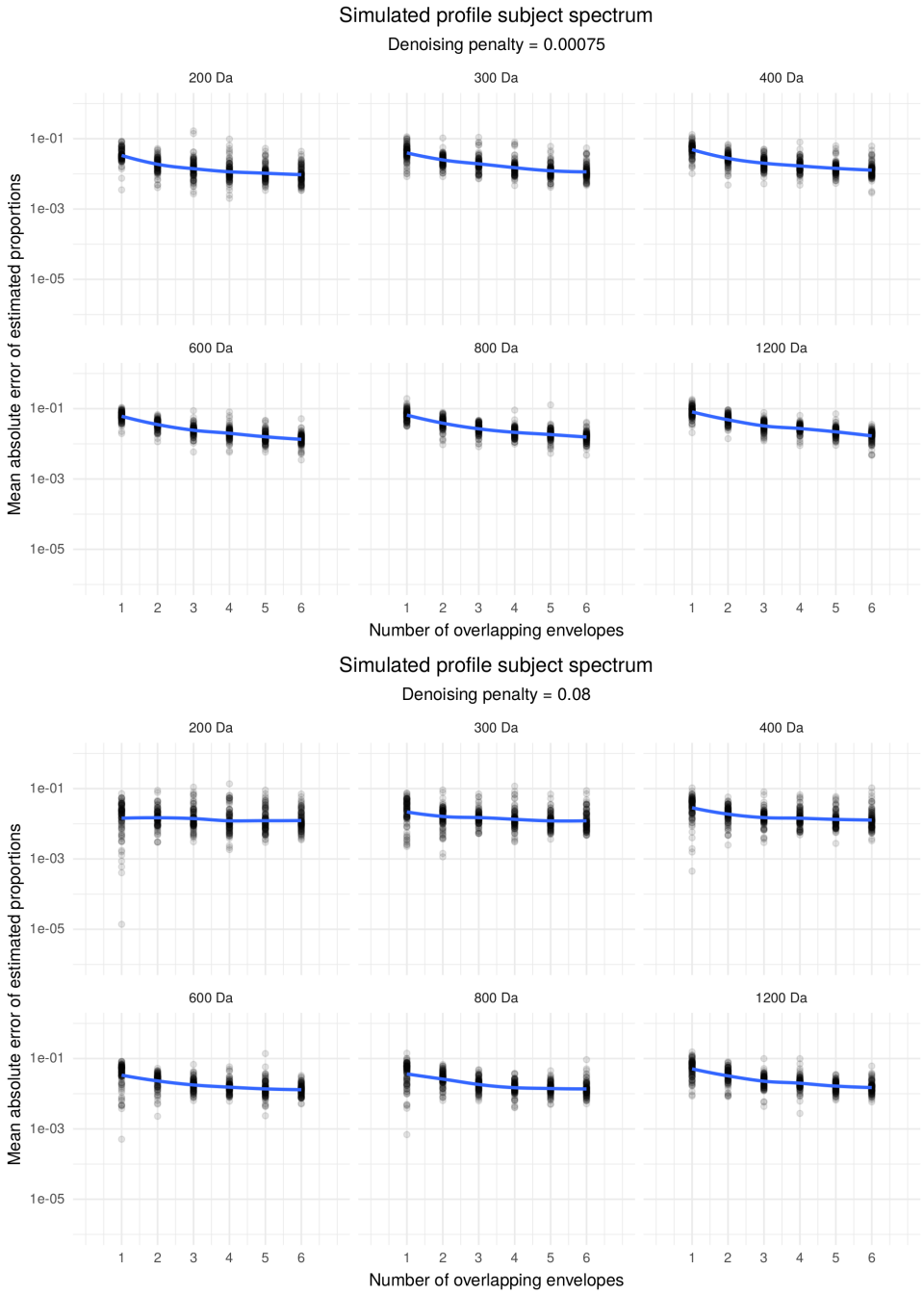

FIGURE 5.10: Mean absolute deviation of estimation of molecule proportions on profile experimental spectra for two denoising penalties. Note the logarithmic scale of the plots.

In the case of centroided spectra, the denoising penalty was set to 0.02 (ten times the standard deviation of the simulated m/z measurement error). The results are shown in Fig. 5.11. The MAD was mostly between 10^{-3} to 10^{-2}, indicating at least two accurate decimal digits an average estimation. There were no clear differences between low and large masses, indicating that isotopic envelopes of 200 Da ions have enough information for accurate regression using our method. In most cases, the best estimates per observed spectrum had at least three accurate decimal digits. We did not observe any bias of the estimation. The mean of the residues was one order of magnitude lower than in the case of profile spectra.

The results show that our method is able to perform accurate estimation of ion proportions even in the case of several isobaric interferences and additional chemical noise. Moreover, it is applicable to both profile and centroided spectra. In the case of profile spectra, we have observed a tradeoff between the estimation bias and variance for different values of the denoising penalty κ. The results presented above were better for centroided spectra. However, the two different types of distortions used to simulate both types of observed spectra are not directly comparable, and their magnitudes differ.

Notably, we have obtained accurate results for up to 6 isobaric molecules of 200 Da in the presence of 50 additional interfering peaks, even though the molecules themselves have only a few peaks in their isotopic envelopes. In this case, the largest absolute error of estimation per centroided spectrum, averaged over 100 replicates, was 0.026 (see Fig. 5.11). This indicates that, on average, all the estimates had at least one accurate decimal digit.

5.4 Reduction to LAD regression on CDFs

In Section 5.1.2, we have presented two formulations of an optimization problem that allows for an estimation of ion proportions accounting for the presence of additional signal, not present in the theoretical spectra (Eqs. (5.2), (5.3) in the main text):

$$p^* = \arg\min_p W(\mu, p_0\omega + \nu_p),$$

$$p^*, g^* = \arg\min_{p,g} \{\kappa p_0 + \sum_{i=1}^{n-1} (s_{i+1} - s_i) \Big| M(s_i) - G(s_i) - N_p(s_i) \Big| \}.$$

The bottom equation is an example of a Least Absolute Deviations (LAD) regression problem, also known as the L^1 regression [75, 74]. In this Section, we show a proof of the equivalence of the formulas. We also give further examples which illustrate several properties of our Wasserstein regression procedure, such as the relationship between the denoising penalty κ and the maximum transport distance. Those examples further motivate the interpretation of κ as a kind of *soft threshold* on the transport distance.

Consider a set of theoretical spectra ν_j and an observed spectrum μ. Recall that the latter is assumed to be composed of theoretical spectra and possibly some remaining signal. Formally, we assume that there are some *true proportions* of spectra ν_j, denoted p_j, satisfying

$$\mu = p_1\nu_1 + \cdots + p_k\nu_k + \varepsilon, \tag{5.5}$$

where ε is a spectrum representing the remaining signal. We assume that $p_1 + \cdots + p_k \leq 1$, allowing some of them to be zero. Furthermore, recall that we assume that

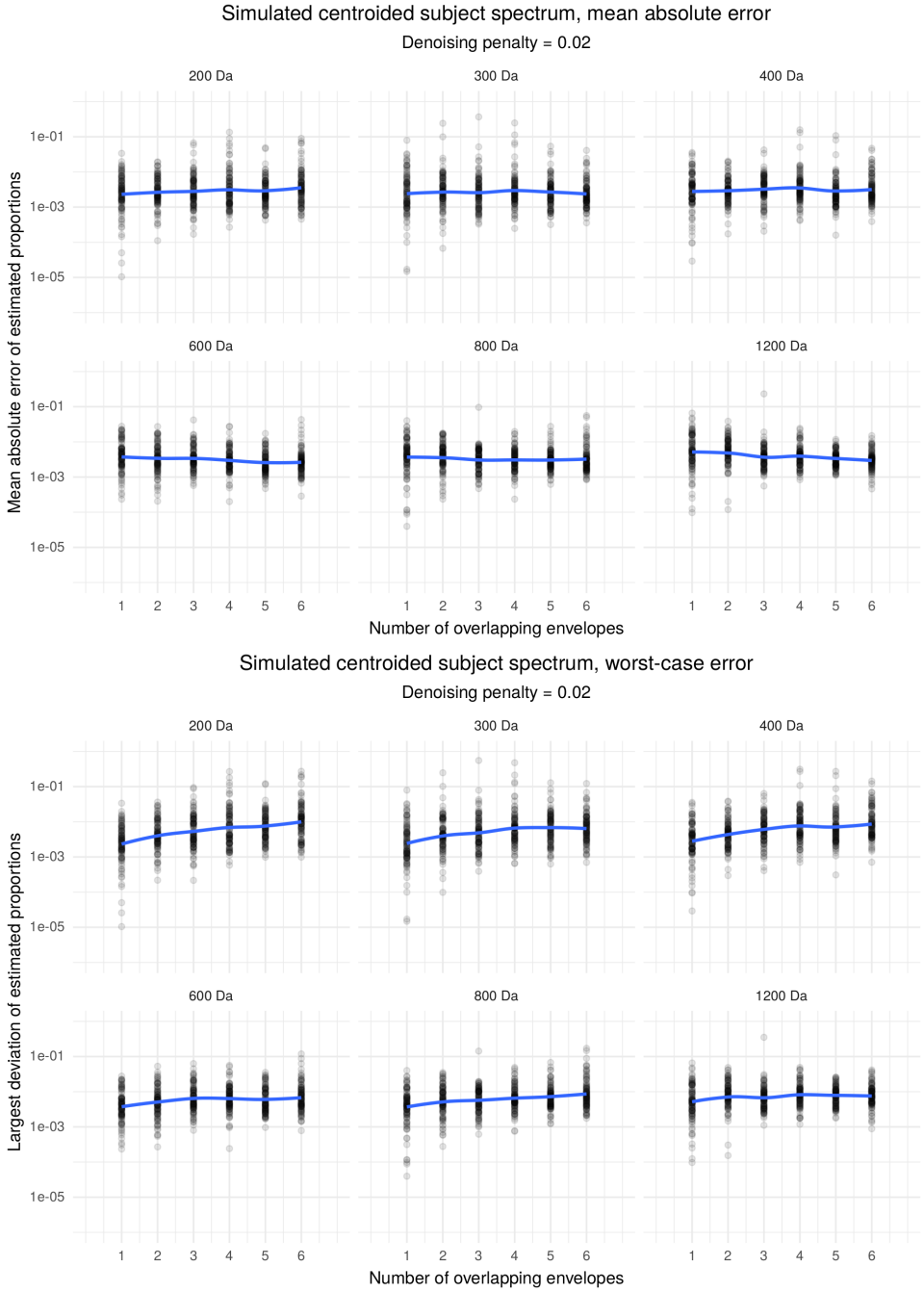

FIGURE 5.11: Errors of estimation of molecule proportions on centroided experimental spectra. Top: mean absolute deviation of estimation per spectrum. Bottom: largest deviation of estimation per spectrum. Note the logarithmic scale of the plots.

all the spectra μ and ν_j are normalized by their total ion current. It follows that the total signal intensity in ε is equal to $1 - p_1 - p_2 - \cdots - p_k$, which will be denoted as p_0.

Under the assumption that ε is empty, the optimal proportions p_j^* were found by minimizing the Wasserstein distance $W(\mu, p_1\nu_1 + \cdots + p_k\nu_k)$ between μ and a linear combination of ν_j [5]. With non-empty ε, we need to incorporate it's removal to the estimation procedure. In order to do that, we introduce an auxiliary spectrum ω and transport this signal from μ to ω. That is, the amount of signal transported from $\mu(s_i)$ onto ω is interpreted as the amount of the remaining signal at s_i after transporting the rest of the signal onto the theoretical spectra.

The auxiliary spectrum ω may be a somewhat non-intuitive concept. First, we assume its total signal sums up to one, but we do not explicitly assume this signal to have any particular m/z value. Second, we assume that there is a constant cost of transporting signal from μ to ω, denoted as κ and referred to as the *denoising penalty*. In a way, ω can be thought of as being equidistant to all signals in μ. The reason for defining ω in such way is to formally define a cost of denoising that does not depend on the m/z value of the removed signal.

Now, we look for optimal proportions as

$$p^* = \underset{p_0+p_1+\cdots+p_k=1}{\arg\min} W(\mu, p_0\omega + p_1\nu_1 + \cdots + p_k\nu_k). \tag{5.6}$$

That is, we look for proportions that allow for the optimal transport of the signal from μ onto ν_j's and ω.

The Wasserstein distance between two measures μ and ν defined on a space $\mathcal{X}$ is given by

$$W(\mu, \nu) = \min_{\gamma \in \Gamma} \int_{\mathcal{X}} \int_{\mathcal{X}} \rho(x, y)\gamma(x, y)dxdy,$$

where Γ is the space of all joint distributions of μ and ν and $\rho(x, y)$ is a distance function between points $x \in X$ and $y \in Y$. In general, the definition works for any distance function ρ. In our case, we define $\rho(x, y)$ between two m/z values x and y as follows:

$$\rho(x, y) = \begin{cases} |x - y| & \text{if } y \in \mathbb{R}, \\ \kappa & \text{if } y = \omega. \end{cases}$$

Based on the above definitions, we have

$$W(\mu, p_0\omega + p_1\nu_1 + \cdots + p_k\nu_k) = \min_{\gamma \in \Gamma} \int_{\mathbb{R}} \int_{\mathbb{R} \cup \{\omega\}} \rho(x, y)\gamma(x, y)dxdy, \tag{5.7}$$

where the minimization is over all joint distributions γ of μ and $p_0\omega + p_1\nu_1 + \cdots + p_k\nu_k$. It follows that γ satisfies the following properties:

$$\begin{aligned} \int_{x \in \mathbb{R}} \gamma(x, y)dx &= p_1\nu_1(y) + \cdots + p_k\nu_k(y) \text{ if } y \in \mathbb{R}, \\ \int_{x \in \mathbb{R}} \gamma(x, \omega)dx &= p_0, \\ \int_{y \in \mathbb{R}} \gamma(x, y)dx &= \mu(x) - \gamma(x, \omega), \\ \int_{x \in \mathbb{R}} \int_{y \in \mathbb{R}} \gamma(x, y) &= p_1 + \cdots + p_k = 1 - p_0 \end{aligned}$$

We now proceed to convert the optimization problem (5.7) to a form that is computationally feasible. We start by splitting the integral over $\mathbb{R} \cup \omega$ into two summands and simplifying the resulting terms:

$$\begin{aligned}\int_{\mathbb{R}} \int_{\mathbb{R}\cup\{\omega\}} \rho(x,y)\gamma(x,y)dxdy &= \int_{x\in\mathbb{R}} \left(\kappa\gamma(x,\omega) + \int_{y\in\mathbb{R}} |x-y|\gamma(x,y)dy \right) dx \\ &= \int_{x\in\mathbb{R}} \kappa\gamma(x,\omega)dx + \int_{(x,y)\in\mathbb{R}^2} |x-y|\gamma(x,y)dydx \\ &= \kappa p_0 + \int_{(x,y)\in\mathbb{R}^2} |x-y|\gamma(x,y)dydx\end{aligned}$$

In the last line, we arrive at the total cost inflicted by signal removal, $p_0\kappa$, and a double integral that is strikingly similar to the definition of the Wasserstein distance between two spectra. Namely, we integrate the transport distance $|x-y|$ multiplied by $\gamma(x,y)$, the amount of signal transported between x and y. However, unlike in the definition of the Wasserstein distance, now Γ is a set of joint distributions over $\mathbb{R}$ and $\mathbb{R} \cup \omega$, so γ function may not be a joint distribution of two probabilistic measures defined on the real line. It means that we cannot yet use the formula that joins the Wasserstein distance to the cumulative distribution functions of the compared measures.

To circumvent the above problem, we proceed as follows. Define a measure $g(x) = \gamma(x,\omega)$ and denote it's cumulative distribution function as $G(t)$. Observe that the total signal in g is equal to p_0. It follows that

$$\int_{x\in\mathbb{R}} \int_{y\in\mathbb{R}} \gamma(x,y)/(1-p_0) = \int_{x\in\mathbb{R}} (\mu(x) - g(x))/(1-p_0) = (1-p_0)/(1-p_0) = 1.$$

We write $\gamma_{|\mathbb{R}^2}$ to explicitly denote γ function restricted to $\mathbb{R}^2$. From the above integrals it follows that $\gamma_{|\mathbb{R}^2}(x,y)/(1-p_0)$ is a two-dimentional probabilistic measure on $\mathbb{R}^2$. Its marginal measures are

$$\begin{aligned}\int_{y\in\mathbb{R}} \gamma_{|\mathbb{R}^2}(x,y)/(1-p_0)dy &= (\mu(x) - g(x))/(1-p_0), \\ \int_{x\in\mathbb{R}} \gamma_{|\mathbb{R}^2}(x,y)/(1-p_0)dx &= (p_1\nu_1(y) + \cdots + p_k\nu_k(y))/(1-p_0),\end{aligned}$$

which are both probabilistic measures on $\mathbb{R}$. Note that in both equations above we consider $\gamma_{|\mathbb{R}^2}$, meaning that we assume $y \neq \omega$. We now can use Theorem 3.2 which expresses the Wasserstein distance between two centroided spectra, μ and ν, in terms of their cumulative distribution functions, M and N. In the first step, we split the

minimization over Γ into two steps: first, minimization of $\gamma_{|\mathbb{R}^2}$, and then minimization of g.

$$\begin{aligned}
\min_{\gamma\in\Gamma}\int_{x\in\mathbb{R}}\int_{y\in\mathbb{R}}|x-y|\gamma(x,y)dydx &= \min_{g}\min_{\gamma_{|\mathbb{R}^2}}\int_{x\in\mathbb{R}}\int_{y\in\mathbb{R}}|x-y|\gamma(x,y)dydx \\
&= (1-p_0)\min_{g}\min_{\gamma_{|\mathbb{R}^2}}\int_{x\in\mathbb{R}}\int_{y\in\mathbb{R}}|x-y|\gamma(x,y)/(1-p_0)dydx \\
&= (1-p_0)\min_{g}\int_{\mathbb{R}}\Big|(M(t)-G(t))/(1-p_0)-\sum_{j=1}^{k}p_jN_j(t)/(1-p_0)\Big|dt \\
&= \min_{g}\int_{\mathbb{R}}\Big|M(t)-G(t)-\sum_{j=1}^{k}p_jN_j(t)\Big|dt.
\end{aligned}$$

Putting it all together, we arrive at

$$W(\mu, p_0\omega + p_1\nu_1 + \cdots + p_k\nu_k) = \kappa p_0 + \min_{g}\int_{\mathbb{R}}\Big|M(t)-G(t)-\sum_{j=1}^{k}p_jN_j(t)\Big|dt.$$

The κp_0 term in the above equation is the penalty for removing p_0 of the signal from the observed spectrum. The minimized integral can be interpreted as the Wasserstein distance between the observed spectrum without the additional signal and the combination of expected spectra under the optimal signal removal plan described by g, on the condition that we remove p_0 of the signal.

To obtain the optimal proportions, we minimize the equation over the proportions p. Since the term κp_0 does not depend on the signal removal plan g, as long as p_0 of the signal is removed, we can minimize over both p and g together and write the following formula for optimal proportions:

$$\min_{p} W(\mu, p_0\omega + p_1\nu_1 + \cdots + p_k\nu_k) = \min_{p,g}\{\kappa p_0 + \int_{\mathbb{R}}\Big|M(t)-G(t)-\sum_{j=1}^{k}p_jN_j(t)\Big|dt\}.$$

Formally, we minimize over all proportions p_0 to p_k and over the signal removal plan such that g_i sum up to p_0. This is, however, equivalent to minimization over p_1 to p_k (i.e. without p_0) and g_i such that $\sum_{j=1}^{k}p_j + \sum_{i=1}^{n}g_i = 1$. This formulation of the minimization problem is used in the next Section.

5.4.1 Some more worked examples and properties

Before we proceed to describe the algorithm for solving the minimization problem, we give some additional remarks about the problem itself. First, the denoising penalty κ is interpreted as the distance between any peak from the observed spectrum to the auxiliary spectrum ω. If, for a given observed peak, all the theoretical peaks are further away than κ, then ω is the closest spectrum to this observed peak. Therefore, κ can be interpreted as the maximum *feasible* transport distance. This interpretation is helpful in estimating reasonable denoising penalties. However, it should be treated as an intuition or a rule of thumb rather than a formal property, as transport for distances greater than κ might occur in some cases.

An example where a long distance transport is beneficial is shown in Fig. 5.12. In this example, we regress an artificially constructed experimental spectrum against

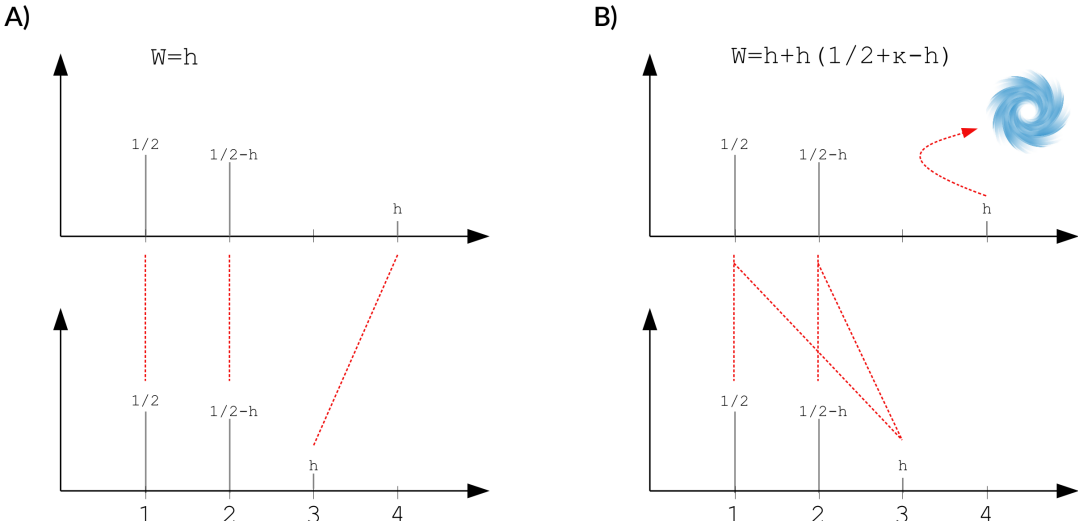

FIGURE 5.12: An example of a long-distance transport. The observed spectrum is shown at the top, the theoretical one at the bottom. The spectrum ω is represented as a vortex. Two transport plans are compared: A, no signal removed, and B, one removed peak. Scenario A is less costly than B regardless of the denoising penalty, and in fact optimal in this example. Long distance transport occurs, because removing the rightmost peak from the experimental spectrum highly disturbs the transport plan.

one theoretical spectrum. Both spectra are abstract examples which serve for a simple illustration of the properties of our method, and do not correspond to any actual molecule.

The theoretical spectrum is composed of three peaks with m/z values $1, 2$ and 3 Da, and intensities $1/2, 1/2 - h, h$ with $0 < h < 1/2$. The experimental spectrum is identical to the theoretical one, except that the peak at 3 Da is shifted 1 Da to the right. We analyze two limiting scenarios that may occur in this situation: either the shifted peak is removed, or it's not.

In the first scenario, no signal is removed, and therefore the proportion of the theoretical spectrum is equal to 1 and the shifted peak is transported onto its theoretical counterpart. The Wasserstein distance in this case is therefore equal to the height of the shifted peak, denoted h.

In the second scenario, the shifted peak is removed. The proportion of the theoretical spectrum is equal to the amount of the remaining experimental signal, that is $1 - h$. In order to compute the cost of the signal transport in this case, we remove the shifted peak from the experimental spectrum, multiply the theoretical peak intensities by $1 - h$, and compute the cost of the optimal transport between the resulting spectra.

The CDF of the experimental spectrum after this procedure becomes equal to

$$\tilde{M}(t) = \begin{cases} 0.0 & : \quad t < 1, \\ 1/2 & : \quad 1 \leq t < 2, \\ 1 - h & : \quad 2 \leq t, \end{cases}$$

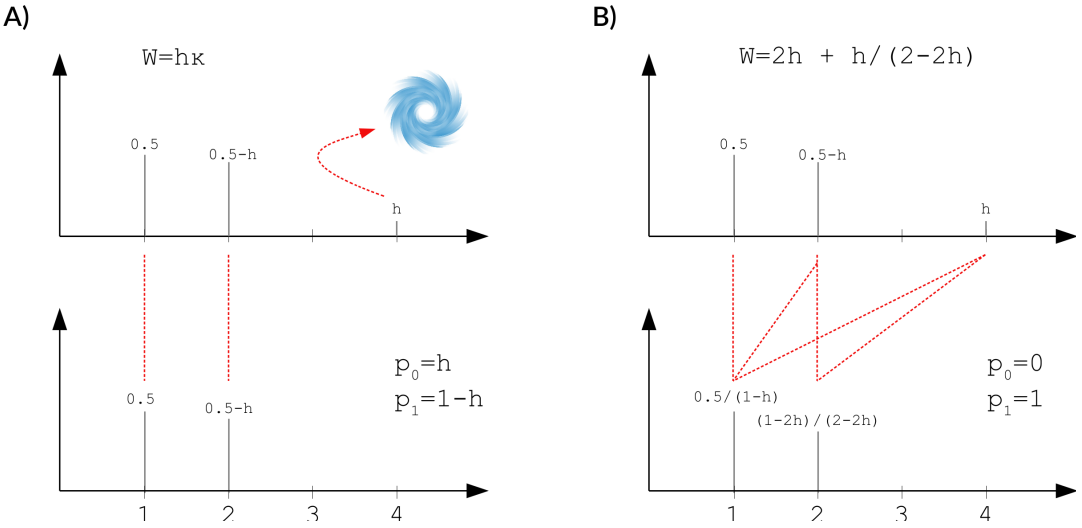

FIGURE 5.13: An example of regression when κ controls the maximum transport distance. In this example, setting $\kappa < 2$ makes the scenario A less costly than scenario B, and prohibits the transport of the small experimental peak at 4 Da. The transport of this peak over a distance of 2 Da is permitted when $\kappa > 2$.

and the one of the theoretical spectrum becomes equal to

$$\tilde{N}(t) = \begin{cases} 0.0 & : \quad t < 1, \\ (1-h)/2 & : \quad 1 \leq t < 2, \\ (1-h) - h(1-h) = (1-h)^2 & : \quad 2 \leq t < 3, \\ 1-h & : \quad 3 \leq t. \end{cases}$$

The cost of signal transport between the two spectra is therefore equal to

$$\left|\frac{1}{2} - \frac{1-h}{2}\right| + \left|1 - h - (1-h)^2\right| = \left(\frac{3}{2} - h\right),$$

and the total cost of this scenario, obtained by adding the cost of the peak removal to the above cost of transport, is equal to $h\kappa + h(3/2 - h) = h(\kappa + 3/2 - h)$, and is higher than the cost of the first scenario whenever $h < \kappa + 1/2$. However, since $0 < h < 1/2$ and $\kappa \geq 0$, the first scenario is always less costly than the second one. Therefore, regardless of the value of the denoising penalty, it is always beneficial to transport the rightmost experimental peak to its theoretical counterpart rather than remove it.

The phenomenon described above occurs because removing a peak induced a large disturbance in the optimal transport plan, depicted in Fig. 5.12. In the first of the described scenarios, the peaks are matched one to one, and only the rightmost experimental peak needs to be transported. On the other hand, in the second scenario, some portion of the signal needs to be transported from each experimental peak, because their intensities exceed the ones of their theoretical counterparts.

In general, κ sets a threshold on the maximum transport distance for those peaks of the experimental spectrum which can be removed without causing major distortions in the optimal transport plan or the optimal proportions of theoretical spectra. An example of such case is shown in Fig. 5.13. In this example, κ does indeed define a strict limit on the maximum transport distance. The computation of the costs of the scenarios is done in the same way as in the previous examples.

Another property of our method that should be noted is that the solution to the minimization problem may not be unique in some cases (see Fig. 5.14 for an example). In the current implementation, we do not attempt to make it unique. Instead, when there are several equally good solutions, we simply pick one at random.

5.5 Reduction to linear programming

In the previous Section, we have derived the formula that needs to be minimized in order to obtain optimal proportions of the theoretical spectra within the observed spectrum,

$$\min_{p,g}\{\kappa p_0 + \int_{\mathbb{R}} \Big|M(t) - G(t) - \sum_{j=1}^{k} p_j N_j(t)\Big| dt\}. \tag{5.8}$$

In this Section, we show a computational procedure of finding the proportions p_j and the amounts noise $g(s_i)$. In principle, the above formula could be used to treat profile and centroided spectra differently. The cumulative distribution functions of the profile spectra are continuous, while the ones of centroided spectra are step functions.

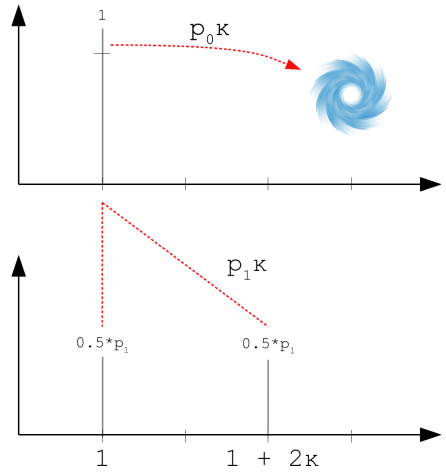

FIGURE 5.14: An example of a non-unique solution to the Wasserstein regression problem. The observed spectrum is shown at the top, the theoretical one at the bottom. The Wasserstein distance is equal to $W(\mu,\nu) = p_0\kappa + p_1\kappa = \kappa$ and does not depend on the proportion of the expected spectrum ν.

However, in the current implementation, we treat both types of spectra in the same manner. A profile spectrum is treated simply as a particularly long peak list, or a series of intensity measurements at discrete m/z values. This is consistent with the way profile spectra are stored on computers, i.e. as finite lists of m/z values and corresponding signal intensities. It follows that all our CDFs are step functions.

In order to solve the problem (5.8) under the above assumptions, we will convert it to a linear program. A linear program is a problem of finding a minimum of a linear function under a set of linear constraints [74]. We follow the ideas of converting a Least Absolute Deviations (LAD) regression problem into a linear program outlined in [76].

Recall that we write $S = (s_1, s_2, \ldots, s_n)$ for a sorted list of all observed m/z values. Since the CDFs of the spectra are step functions, the integral in the problem (5.8) is equal to a simple sum:

$$\int_{\mathbb{R}} \Big|M(t) - G(t) - \sum_{j=1}^{k} p_j N_j(t)\Big| dt = \sum_{i=1}^{n} (s_{i+1} - s_i)\Big|M(s_i) - G(s_i) - \sum_{j=1}^{k} p_j N_j(s_i)\Big| \tag{5.9}$$

As discussed in the main text, equations of the above form admit a natural interpretation in terms of optimal transport. The difference of CDFs at the point s_i is the difference in the amount of ion current between the compared spectra on the left hand side of this point. This difference needs to be balanced by transporting the ion

current either to or from the next point, s_{i+1}. Therefore, the summands can be interpreted as the amount of ion current that flows between points s_i and s_{i+1} multiplied by the interval length.

Let $M_i = M(s_i)$, $N_{ij} = N_j(s_i)$ and $G_i = G(s_i)$. Let $l_i = s_{i+1} - s_i$ be the i-th interval length, and let $\epsilon_i = M_i - G_i - \sum_{j=1}^{k} p_j N_{ij}$ be the ion current flow between s_i and s_{i+1}. The task now is to minimize ϵ_i over proportions p_j and amounts of removed signal g_i. Note that, since the spectra are normalized, the ion current balances out at s_n, so that $\epsilon_n = 0$. Therefore, in the optimization problems below, we consider ϵ_i only for $i = 1, 2, \ldots, n-1$.

Using the above notation notation, we reformulate our optimization problem as follows:

$$
\begin{array}{lrcl}
\text{minimize} & \sum_{i=1}^{n-1} l_i|\epsilon_i| + \sum_{i=1}^{n} \kappa g_i & \text{over } \epsilon, p, g & \\
\text{subject to} & \epsilon_i + \sum_{j=1}^{i} g_j + \sum_{j=1}^{k} N_{ij} p_j & = M_i, & i = 1, 2, \ldots, n-1 \\
& \sum_{i=1}^{n} g_i + \sum_{j=1}^{k} p_j & = 1, & \\
& g_i, p_j & \geq 0. &
\end{array}
$$

Note that while we enforce a constraint that the total amount of removed signal does not exceed the leftover signal in the observed spectrum, we do not enforce such a constraint peak-wise. That is, during the course of numerical optimization, it may happen that the proportion of signal removed from the i-th experimental peak will temporarily exceed the intensity of that peak. However, such a peak-wise constraint is fulfilled automatically in the optimized signal removal scheme.

The objective function, $\sum_{i=1}^{n-1} l_i|\epsilon_i|$, is not yet linear. However, by splitting the error ϵ_i into a positive part, ϵ_i^+, and a negative part, ϵ_i^-, so that $\epsilon_i = \epsilon_i^+ - \epsilon_i^-$, we can rewrite the above optimization problem as a linear program:

$$
\begin{array}{lrcl}
\text{minimize} & \sum_{i=1}^{n-1} l_i \epsilon_i^+ + \sum_{i=1}^{n-1} l_i \epsilon_i^- + \sum_{i=1}^{n} \kappa g_i & \text{over} & \epsilon^+, \epsilon^-, p, g \\
\text{subject to} & \epsilon_i^+ - \epsilon_i^- + \sum_{l=1}^{i} g_l + \sum_{j=1}^{k} N_{ij} p_j & = M_i, & i = 1, 2, \ldots, n-1, \\
& \sum_{i=1}^{n} g_i + \sum_{j=1}^{k} p_j & = 1, & \\
& \epsilon_i^+, \epsilon_i^-, g_i, p_j & \geq 0. &
\end{array}
\tag{5.10}
$$

The above linear program has $3n - 2 + k$ variables and $n + 1$ constraints. For large spectra, where n can be of the order of tens of thousands peaks, this leads to a computationally intensitve optimization problem. In order to obtain a more efficient algorithm, we consider a *dual* problem. A comprehensive treatment of the duality theory in linear programming can be found in [74]. In short, each linear program admits a so-called dual program which is equivalent in the sense that it has the same optimal value of the optimized function. Furthermore, after solving the dual program, we can easily reconstruct the optimal values of the variables of the *primal* program (5.10).

The dual formulation of the program (5.10) is as follows:

$$\begin{array}{lrcll} \text{maximize} & \sum_{i=1}^{n} M_i y_i \text{ over } y & & & \\ \text{subject to} & \sum_{l=1}^{n} N_{ij} y_i & \leq & 0, & j = 1, 2, \ldots, k, \\ & \sum_{i=l}^{n} y_i & \leq & \kappa, & l = 1, 2, \ldots, n, \\ & y_i & \leq & l_i, & i = 1, 2, \ldots, n-1, \\ & y_i & \geq & -l_i & i = 1, 2, \ldots, n-1, \\ & & & & y \in \mathbb{R}^n. \end{array} \tag{5.11}$$

In the above dual program, we have n variables, $n + k$ constraints and $2n - 2$ bounds. We now proceed to further simplify the program.

Let $U = ([i \geq j])_{i,j=1,2,\ldots,n}$ be an $n \times n$ square, lower-triangular binary matrix with ones on and below the diagonal and zeros above it. We use it to re-write the above dual program in matrix notation:

$$\begin{array}{lrcl} \text{maximize} & M^T y & & \\ \text{subject to} & N^T y & \leq & 0, \\ & U^T y & \leq & \kappa, \\ & & & -l \leq y_{1:(n-1)} \leq l, \\ & & & y \in \mathbb{R}^n \end{array}$$

Let $W = (\nu_j(s_i))$ be the matrix of intensities of the theoretical spectra on the points s_i for $i = 1, 2, \ldots, n$, i.e. the theoretical spectra stacked column-wise. Similarly, let $V = (\mu(s_i))$ be a vector of intensities of experimental spectrum for $i = 1, 2, \ldots, n$. Note that we have $N = UW$ and $M = UV$, and substitute that into the program formulation:

$$\begin{array}{lrcl} \text{maximize} & V^T U^T y & & \\ \text{subject to} & W^T U^T y & \leq & 0, \\ & U^T y & \leq & \kappa, \\ & y_{1:(n-1)} & \leq & l, \\ & y_{1:(n-1)} & \geq & -l, \\ & & & y \in \mathbb{R}^n. \end{array}$$

Since the matrix U is full-rank, substitution $z = U^T y$ is a valid variable change. Note that

$$U^{-1} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix}$$

and since $y = U^{-T} z$, we have $y_i = z_i - z_{i+1}$ for $i = 1, 2, \ldots, n-1$ and $y_n = z_n$. After substituting and rearranging rows of the linear program we obtain

$$\begin{array}{lrcll} \text{maximize} & V^T z & & & \\ \text{subject to} & W^T z & \leq 0, & & \\ & z_i - z_{i+1} & \leq & l_i, & i = 1, 2, \ldots, n-1, \\ & z_i - z_{i+1} & \geq & -l_i, & i = 1, 2, \ldots, n-1, \\ & & & z \leq \kappa, & \end{array} \tag{5.12}$$

which is a program with n upper-bounded variables and $k + 2n - 2$ constraints.

Program (5.12) is the dual linear program in its final form. Its major advantage over the first representation of the dual program is the sparsity of matrix W as opposed to matrix N, meaning that many of its values are zero. Modern implementations of the Simplex algorithm, one of the algorithms used to solve linear programs, take advantage of matrix sparsity to speed up the computations.

Since the linear program (5.12) was obtained by rearranging the lefthand side terms in Program (5.11), the optimal target function value stays unchanged, just as the values of the primal program variables corresponding to the constraints. The latter can be obtained from the solution, as the modern Simplex implementations implicitly solve both the dual and the primal problem. The probability values, p_j, can be obtained as the duals to the constraints $W^T z \leq 0$, while the noise amounts g_i can be obtained as the duals to the constraints $z_i \leq \kappa$.

On a final note, note that we have allowed for non-zero noise value g_i on non-experimental peaks. In the dual linear program (5.12), this is visible as upper bounds for all z_i variables, instead of just the ones that correspond to points s_i with a non-zero experimental signal. This does not influence the results, as non-zero g_i at a point without any experimental signal is always sub-optimal. If one would like to explicitly forbid non-zero values of g_i at masses without experimental signal, it would suffice to remove upper bounds for z_i corresponding to purely theoretical peaks. Removing bounds leads to a simplification of the feasible region, which speeds up the Simplex algorithm. However, we have found out that the speedup obtained this way is negligible in practice.

5.6 Simulation of mass spectra

In this Section, we describe the details behind the simulation procedures used in Section 5.3. To simulate random molecular formulas, we use Algortihm 5 for $\mathcal{E} = (C, O, N, S, P)$ and $\mathcal{W} = (12, 16, 14, 32, 31)$. For each element, the number of atoms is sampled uniformly from 0 to the maximum number allowed by the remaining mass, and the remaining mass is filled with hydrogen atoms. Note that, in this algorithm, the order of elements influences their abundance. Elements which are closer to the beginning of the list $\mathcal{E}$ tend to be more abundant that those at the end of the list.

We have simulated 100 sets of molecular formulas (referred to as replicates) for each combination of the following parameters:

- Nominal mass of molecules: 60, 120, 600, 1200, 6000, 12000,
- Number of overlapping isotopic envelopes: 1, 2, 3, 4, 5, 6, 7, 8.

After the molecules were simulated, theoretical spectra were computed using the IsoSpecPy package [72]. To construct observed spectra, we have simulated several sources of measurement distortions. An example result of the simulation is shown in Fig. 5.15.

First, we have simulated the effect of a finite number of molecules, which causes the peak heights to be variable due to random numbers of isotopologues. We assume that each observed spectrum is formed by $N = 10000$ ions to obtain a moderate to high variability of peak heights. For each set of molecular formulas, we sample their proportions $p_1, \ldots, p_k$ uniformly from a unit simplex $\Delta = \{p \in \mathbb{R}^k \colon \sum_{i=1}^{k} p_i = 1, \forall_j p_j \geq 0\}$. The number of ions of the j-th molecular formula is then equal to Np_j. Each ion is then assigned to an isotopic composition according to the probabilities

computed by IsoSpec, under the assumption of standard isotopic compositions of elements. Next, we assume that each ion contributes a random amount of signal intensity to its spectrum, with a Gaussian distribution with mean 1 and standard deviation 0.001. A mass spectrum is then obtained by summing the signal contributions of all ions.

Next, to simulate chemical noise, we have added 50 random peaks, with uniformly distributed m/z values and Gamma-distributed intensities (shape=scale=2). Those peaks were then scaled so that the total amount of noise had a Beta distribution (alpha=1.444, beta=5). The parameters were selected so that, on average, the noise peaks amount for 10% of the total signal intensity in the spectrum. Further distortions were different for centroided and profile observed spectrum.

In the case of simulated centroided spectra, we additionally simulate the effects of limited resolution and accuracy and errors introduced during the preprocessing procedures such as peak picking. To each m/z value we add a Gaussian random variable with mean 0 and standard deviation of 0.002. The parameters were chosen to obtain only two accurate decimal digits in the m/z values. To simulate a limited resolving power, we round the m/z to three decimal digits and merge peaks with equal masses.

In the case of simulated profile spectra, we simulate the effect of limited resolving power and electronic noise. We use a Gaussian filter with a standard deviation of 0.0025 to obtain a resolving power of 100 000 at 600 Da. Next, to each intensity measurement we add a Gaussian random variable with mean zero and standard deviation of 0.0001.

Algorithm 5: Simulation of random molecules.

Data: Set of non-hydrogen elements $\mathcal{E}$, mass numbers $\mathcal{W}$, integer number N.
Result: A random molecule $\mathcal{M}$ with total mass number N.

Initialize an empty list $\mathcal{M}$
for *i in* $1, 2, \ldots, |\mathcal{E}|$ **do**
  Let $w \leftarrow \mathcal{W}[i]$
  Sample U from $\{0, 1, \ldots, \lfloor \frac{N}{w} \rfloor\}$
  Let $\mathcal{M}[i] \leftarrow U$
  Let $N \leftarrow N - wU$
end
Let $\mathcal{M}[|\mathcal{E}| + 1] \leftarrow N$
Variable $\mathcal{M}$ contains the numbers of sampled elements with hydrogen as the last entry.

5.7 Summary of the Chapter

In this Chapter, we present advances in theoretical studies on the problem of linear regression of mass spectra, defined as fitting a linear combination of theoretical spectra to an experimental one. The problem is ubiquitous in mass spectrometry, appearing either implicitly or explicitly in areas as diverse as metabolomics, proteomics and polymer science. The theoretical foundations of the methods studied in this Chapter are not limited to any particular type of experiment. Furthermore, the method can be readily applied to other types of spectra, such as the nuclear magnetic resonance ones, as long as reliable reference spectra are available. The broad

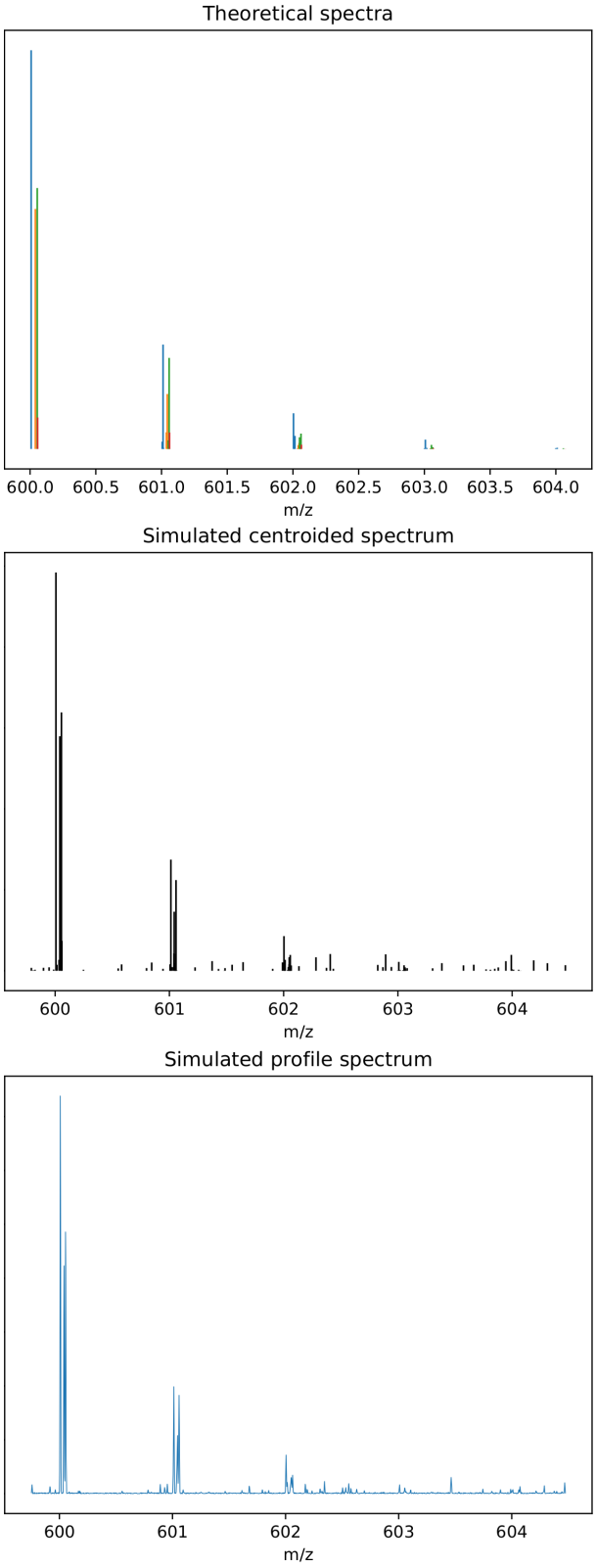

FIGURE 5.15: Simulated mass spectra with four overlapping isotopic envelopes. The top spectrum was used to construct a centroided and a profile observed spectrum with measurement distortions. Isotopic envelopes of different ions are highlighted with different colors.

range of applications is achieved by focusing on an abstract problem of fitting a linear combination of reference signals to an experimentally measured one, as opposed to developing a method restricted to a particular application.

One of the major factors that hinder the performance of currently available linear regression methods is the fact that the locations of experimentally measured peaks never match the theoretically predicted ones perfectly. This is caused, among others, by measurement inaccuracies, which, even if seemingly small, are unavoidable. Although they may be negligible when the spectra are analyzed manually, an m/z difference as small as a millionth of a Dalton means that a computer program sees the peak locations as different.

In order to circumvent this problem, the currently available software matches peaks within so called mass windows of predefined width [11, 53]. The width needs to be specified by the user, which makes it more difficult to apply this kind of approach in practice. The choice of a mass window width is further complicated by the occurrence of highly overlapping peaks in profile spectra. In some cases, the individual apexes of such peaks are no longer visible. Instead, we obtain a single apex located between the "true" apexes of the component peaks. This kind of merging of peaks, occurring especially in complex profile spectra, leads to an increased difference between theoretical and observed peak locations in centroided spectra. Therefore, especially in complex or low-resolution spectra, the optimal window width may be considerably greater than the nominal accuracy of the instrument.

Even if the user knows the optimal width of a mass window, this kind of approach has several intrinsic drawbacks. Due to the infinite resolution of a theoretically predicted spectrum, a mass window usually contains several theoretical peaks. Within the window, those peaks are effectively treated as one. Therefore, this procedure effectively limits the resolution of the *in silico* predicted spectrum, leading to an unnecessary loss of information. On the other hand, it tends to merge closely positioned signal and noise peaks in the experimental spectrum, which influences the estimated proportions.

In order to alleviate those difficulties, we have investigated the application of the Wasserstein distance to the problem of linear regression of mass spectra. Methods based on quantifying the distance in the m/z domain needed to transform one spectrum into the other are naturally robust to limited resolving power and accuracy of instruments. This robustness makes them a promising tool for methods based on comparing experimentally acquired spectra.

Further research. A practical consequence of allowing the signal to only be removed from the experimental spectrum is that, when simulating the theoretical spectra, we need to discard peaks that are under the level of quantification. The measured intensities of such peaks are unreliable, leading to erroneous results for the whole ion. In an extreme case, when a peak is missing in the experimental spectrum but is present in the theoretical one, it forces the estimated proportion to be zero. The lack of experimental intensity that can be transported onto this theoretical peak means that the whole isotopic envelope needs to be discarded. Allowing for some flexibility in the intensities of theoretical peaks would allow to alleviate this difficulty. However, in order to accurately reflect the observed variability of peak intensities, the regression procedure needs to be coupled with a mathematical model of the shape of experimentally measured isotopic envelopes. Whether such coupling is mathematically and computationally feasible remains an open question.

Implementation. We have implemented the discussed algorithms in a Python 3 package called
`masserstein`. Our implementation is designed to be applicable in larger data processing pipelines. Efficient development of pipelines requires freely available modular tools, which perform specific tasks and can be easily combined. Therefore, our implementation does not perform any additional pre- or postprocessing of the results, such as peak-picking or correcting for proton affinity of molecules. Such procedures can be performed separately using designated tools, available e.g. in the OpenMS package [77].

On a final note, we reiterate that, from a methodological point of view, molecule identification and quantification are two separate tasks. Accordingly, in this Chapter, we have assumed that the chemical formulas of the molecules to be quantified are known a priori. These may come either from the scientific question at hand (such as the frequency of a given posttranslational protein modification), from the knowledge about the experimental setup (such as whether lipid extraction was performed), or from an identification study performed prior to quantitative analysis (e.g. using the SIRIUS program [78]). Another approach, applicable to proteomics, is to use the averagine model of amino acid, as exemplified by the MasSPIKE program [79]. In Chapter 6, we show an example application of our methods in a case when only the general class of molecules is known (in this case, phosphoglycerolipids).

Chapter 6

Improved segmentation of mass spectrometric images

Mass spectrometry imaging (MSI) has established its place as a valuable technique in numerous fields of studies. The possibility to characterize the spatial distribution of hundreds of molecules in a single experiment offers great opportunities in disciplines as diverse as medical research [80], fundamental and applied biology [81], food science [82, 83], and synthetic polymer research [84]. In order to fully benefit from the vast amounts of information contained in a single mass spectrometric image, mathematical, statistical and computational tools are routinely used. New algorithms and software are constantly being developed, making analyses easier, faster, and opening new possibilities to harness the complexity of the data for new discoveries [85, 86, 87].

Among the most popular methods of MSI data analysis is the image segmentation, used to identify regions with characteristic chemical compositions [85, 88]. Ideally, such regions correspond to physically distinct parts of the sample, such as tissues, lesions, tumors etc. Accurate segmentation methods offer a simple and reliable approach to identify novel biomarkers [89, 90].

Segmentation methods can be roughly divided into two types: univariate and multivariate. The former segment the image based on a single selected feature (therefore being more of a "targeted" approach), while the latter use multiple features (being more "untargeted"). While multivariate methods may be easier to use and offer a more accurate segmentation, univariate methods allow for a greater degree of control over the segmentation process, as the user may directly select a feature of interest to be analyzed. They also allow for an easier interpretation of the segments as regions with characteristic concentrations of a given chemical compound [91].

A common approach to univariate mass spectrometric image segmentation is to select a peak of interest and simply cluster its intensities from all pixels using algorithms such as (1-dimensional) K-means or Gaussian Mixture Models [89]. This approach disregards the spatial relations between pixels, and because of that, it suffers from two issues caused by pixel-to-pixel variability: first, pixels from different anatomical regions can have similar peak intensities purely by chance, and second, closely located pixels from the same regions can have differing intensities [92]. This results in spatially inhomogeneous segments with only partial correspondence to anatomical regions. Although it has already been noted and addressed by some authors, methods that ignore spatial relationships are still widely used and developed. A recent article has addressed this problem by developing a spatially-informed segmentation method called `spatial-DGMM`, based on a Bayesian approach to Gaussian Mixture Models [91].

Another problem is that the aforementioned approach implicitly assumes that a single peak is an independent, autonomous feature. However, isotopic envelopes

of ions with similar masses can overlap, causing some peaks to be composed of signals from more than one ion. On the other hand, typical ions have more than one peak in their isotopic envelope. A single mass spectral feature (a peak) is therefore a complex combination of parts of chemical features (ionized chemical compounds), the analysis of which is the goal of MSI. This problem has been studied, among others, in the context of proteomics, lipidomics and polymer science [11, 54, 55, 93, 94, 95], but seems to have gained less attention in MSI literature.

One of the solutions to this problem is to calculate the theoretical spectra of ions of interest and fit them to the observed spectrum to estimate their proportions. One of the tools developed for this task is `masserstein`, which is based on the optimal transport paradigm of the analysis of mass spectra, making it robust to moderate measurement inaccuracies and model misspecifications [5, 6, 96, 97].

The contents of this Chapter. In this Chapter, we study the possible consequences of and the interplay between two challenges in mass spectrometry image segmentation: the pixel-to-pixel variability and the overlapping of isotopic envelopes. First, we develop a simulation scheme which mimics real images in terms of pixel-to-pixel variability and shapes of isotopic envelopes. We use it to construct a simulated image which shows that, without taking the two challenges into account, the resulting segmentation can be not only inaccurate but downright misleading, resulting in apparent ion concentration regions contrary to the actual ones. Then, we demonstrate how to solve this problem by using two recently developed tools, `masserstein` and `spatial-DGMM`. We show that this combination is capable of detecting regions of actual concentration of molecules in complex MSI data. *Ipso facto* we show that the two problems described in the introduction constitute the main obstacles in obtaining a biologically meaningful segmentation. We validate the conclusions drawn from the simulations by analyzing similar situations in a mass spectrometry image of a mouse bladder. We show that the combination of `masserstein` and `spatial-DGMM` improves the qualitative and quantitative agreement between segments and anatomical regions compared to the basic approach exemplified by the K-means clustering of peak intensities. We quantify the prevalence of overlapping isotopic envelopes of lipids in the image and show that the discrepancy between mass spectral and chemical features is ubiquitous.

6.1 Materials and methods

Data sets. An MS image of a tissue section of a mouse bladder [98] was downloaded from the PRIDE database [99] (ID PXD001283). An additional simulated data set was prepared as a part of this work as described later in this section. For the experimental data set, all spectra were normalized by their total ion current calculated by numerical integration of intensities in profile mode. For the mouse bladder MSI, due to an extensive background area that influenced the average spectrum of the image, an approximate mask image of the tissue sample was constructed manually based on the accompanying microscopic image and selected ion images using the GNU Image Manipulation Program (GIMP; https://www.gimp.org).

Average spectra. Average spectra of MS images were computed in order to inspect their overall composition and to detect overlapping isotopic envelopes. A common mass axis was fixed with a uniform distribution of m/z values (from 600 to 1100 Da, spaced by 0.01 Da). For each pixel, signal intensities in points of the mass axis were approximated by a piecewise-linear interpolation as described in

previous works [6]. The interpolated spectra were then summed over all pixels and normalized by their total ion currents calculated by numerical integration (trapezoid method). For the mouse bladder data set, pixels corresponding to the background were ignored. The average spectra were centroided using a procedure implemented in the `masserstein` package [6].

Theoretical spectra. All chemical formulas of glycerolipids, glycerophospholipids and sphingolipids were obtained from the LIPID MAPS database [100] on March 28, 2022 (4235 formulas of 22239 different lipids). Formulas containing elements other than CHNOP were discarded. Theoretical spectra of lipid ions with potassium adducts were computed using IsoSpec. Theoretical spectra were truncated to contain only the first two peaks, because in experimental spectra the heavier peaks tended to be below the limit of detection. Spectra with the monoisotopic mass lower than 600 Da or greater than 1100 Da were discarded. This has resulted in 2460 theoretical spectra.

Detection of overlapping isotopic envelopes. Linear combinations of truncated theoretical spectra were fitted to the normalized and centroided average spectra of MS images using `masserstein` (MTD=0.02 for mouse bladder, MTD=0.01 for mouse cerebellum, selected based on a visual assessment of the quality of the model fit). Ions with estimated signal less than 100 ppm were discarded. Remaining ions were assigned to a single cluster if the difference of their monoisotopic masses was smaller than 2.2 Da.

Ion images and signal images. Ion images of selected lipids were generated by taking the apex intensities (i.e. heights) of their monoisotopic peaks in each normalized pixel spectrum in profile mode. Using peak heights instead of areas is justified in the case of this study, because we compare lipid intensities on a per-cluster basis. Each cluster spans only a small mass region, therefore containing peaks of similar width. For clusters of overlapping isotopic envelopes, lipid signal images were obtained by using the `masserstein` package to fit a combination of truncated theoretical spectra of lipids from the cluster to centroided and normalized pixel spectra of the MS image.

Choice of algorithm parameters. To obtain proper lipid signal images, the MTD parameter of `masserstein` was estimated by comparing the lipid signal images to the ion images of selected lipids with no evidence for interference in the average spectra of MS images and selecting the lowest value that gave a sufficient visual agreement.

Simulation of a mass spectrometric image. A 40x40 pixel reference image for the simulated mass spectrometric image, containing four distinct regions, was drawn manually in the GNU Image Manipulation Program. Three lipid formulas were used to simulate an MS image, assuming a potassium adduct: PC(38:1), $C_{46}H_{90}NO_8PK$, 854.603 Da; PA(44:0), [$C_{47}H_{93}O_8PK$, 855.624 Da; and PC(38:0), $C_{46}H_{92}NO_8PK$, 856.619 Da. The first lipid was concentrated in the top half of the image; the second in the bottom half; and the third in a 20x20 center square (see Fig. 6.1).

In each pixel, the number of ions of each lipid species was drawn from a negative binomial distribution with the average value given by Table 6.1 and a coefficient of variance equal to 20%. Next, simulated isotopic envelopes for the three lipids were generated by drawing samples from a multinomial distribution, with the numbers of trials equal to the drawn numbers of lipid ions in each pixel, and probability vectors corresponding to the theoretical isotopic envelopes of the lipids generated with IsoSpec [72]. In each pixel, the simulated isotopic envelopes of the three lipids were added together. Then, to simulate chemical contaminants and other signals, 10 randomly located peaks, jointly accounting for 10% of the intensity of the spectrum,

TABLE 6.1: Average numbers of lipid ions in regions of the simulated mass spectrometric image.

	PC(38:1)	PA(44:0)	PC(38:0)
Region 1	10 000	2 000	1 000
Region 2	1 000	4 000	1 000
Region 3	10 000	2 000	2 000
Region 4	1 000	4 000	2 000

were added. A Gaussian filter was then applied to the pixel spectra to simulate a limited resolving power (FWHM=0.12 at 854.6 Da).

6.1.1 Analysis of the simulated image.

To generate ion images of lipids, in each pixel the spectrum was integrated in the following m/z ranges: 854.4 Da to 854.8 Da for PC(38:1); 855.4 Da to 855.8 Da for PA(44:0); 856.4 Da to 856.8 Da for PC(38:0). Trapezoid method implemented in the `numpy` package was used for integration. To obtain proportions of lipids, theoretical spectra were fitted to each pixel spectrum using `masserstein` with MTD parameter equal 0.2. To obtain signal images of lipids, the lipid proportions in each pixel were multiplied by the total ion current in that pixel, obtained through numerical integration of the associated spectrum. Ion images and signal images were segmented with the K-means algorithm using the `scikit-learn` package [101] and with `spatial-DGMM` [91].

6.2 Results and discussion

We have designed a simulated mass spectrometry image to illustrate the potential pitfalls of common image segmentation approaches. The image is composed of three lipid ions with distinct spatial distributions, each with a single region of high concentration. The goal of segmentation is to discover these regions. However, a simple K-means clustering of peak intensities leads to surprising results, including an inversion of the apparent enrichment region. True distributions can only be discovered when the spatial structure of the segments and the composition of spectra are taken into account. Next, we identify similar situations in a real mass spectrometry image, demonstrating that our simulations reveal problems encountered in actual data analysis. Along the way, we discover an unexpected problem with the common assumption of potassium adducts in lipid ions in MSI.

6.2.1 Ion images can be misleading.

In order to give a clear illustration of how the interference between isotopic envelopes makes it difficult to analyze the mass spectrometry imaging data, we have generated a simulated image containing three lipid ions with overlapping isotopic envelopes and characteristic spatial distributions. This image also serves as a proof of concept that `masserstein` and spatialDGMM make it possible to discover the true distributions of clustered ions, a claim that will be further supported by experiments on real data later in this Chapter.

Peaks are combinations of molecular features. The three lipids used in the simulations are PC(38:1), PA(44:0) and PC(38:0), each one Dalton heavier than the previous one. The first lipid is concentrated in the top half of the image; the second in

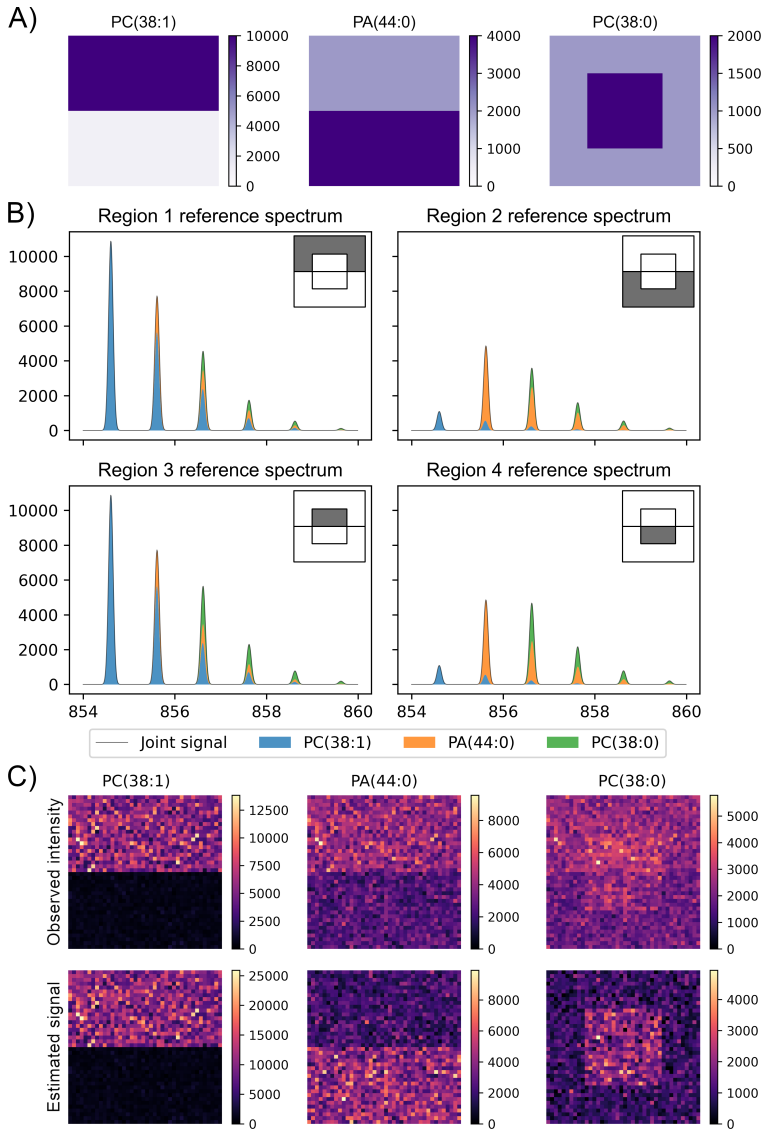

FIGURE 6.1: Overlapping isotopic envelopes distorted observed ion images. A) The reference images of lipid concentrations used to simulate a mass spectrometry image; colors represent the numbers of ions. B) The spectra of distinct regions of the image, shown before simulating pixel-to-pixel and isotopic envelope variability. C) The ion images obtained by integrating monoisotopic peaks and the signal images estimated by fitting spectra with `masserstein`. The peak intensity of PA(44:0) became inverted with respect to its concentration due to interference from PC(38:1).

the bottom half; and the third in a center square (Fig. 6.1). The goal of mass spectrometric image segmentation is to reconstruct those regions.

Due to a high overlap of the spectra, the ion images of the last two lipids do not reflect their true spatial distributions (Fig. 6.1). The intensity of the monoisotopic peak of PA(44:0) is higher in the top half of the image, despite the lipid being less concentrated in this region. This "swap" of regions is caused by the large intensity of PC(38:1) in the top half, and the fact that the isotopic envelope of this lipid overlaps with the monoisotopic peak of PA(44:0). Therefore, the intensity of PC(38:1) contributes to the apparent intensity of PA(44:0).

While the apparent distribution of PA(44:0) is reversed with respect to the true one, a different kind of effect is exhibited by the monoisotopic peak intensity of PC(38:0). This peak is influenced by the isotopic envelopes of both PC(38:1) and PA(44:0). As a consequence, four different regions can be seen in its ion image, despite the lipid being concentrated in the middle square region of the image.

K-means segments are mixtures of biological segments. The results of K-means (K=2) clustering of the monoisotopic peak intensity follow the patterns visible on the ion images (Fig. 6.2). The clustering of the signal of PC(38:1) monoisotopic peak follows the spatial distribution of this lipid, with a high-intensity segment in the upper half of the image. The signal of PA(44:0) also segments the image into the upper and the lower half. However, the average intensities in the segments suggest that this lipid is concentrated in the top half, contrary to its true spatial distribution. This is caused by the high intensity of the n+1 peak of PC(38:1), which overlaps with the monoisotopic peak of PA(44:0). The segmentation of the peak intensity of the third lipid, PC(38:0), shows little resemblance to its true spatial distribution, because of the interference of the two lighter lipids. No segment corresponds to the central square in which this lipid is concentrated.

Ion intensity estimation recovers true molecular features. The lipid intensities estimated with `masserstein` follow their true spatial distributions (Fig. 6.2). This is because fitting the whole isotopic envelopes simultaneously to the cluster makes it possible to separate their signals and remove interferences. Accordingly, the K-means segmentation of estimated lipid intensities shows a better qualitative agreement with their spatial distribution (Fig. 6.2). The high-intensity cluster of PA(44:0) roughly corresponds to the bottom half of the image, and the central square is visible as the high-intensity cluster of PC(38:0).

Although `masserstein` makes it possible to achieve a qualitative agreement between the segmentation and the true lipid enrichment regions, the quantitative agreement is still far from perfect, especially for a simulated data set. The percentage of pixels from high-concentration regions correctly identified as such was 56% for PA(44:0) and 68% for PC(38:0). Disregarding the spatial relationships between pixels leads to rugged and dispersed clusters, as pixels from different segments can have similar lipid intensities due to the variability of the ion count.

Spatially-aware segmentation recovers true biological segments. Using the `spatial-DGMM` algorithm, a spatially-aware segmentation method developed specifically for mass spectrometry imaging data, returns more spatially homogeneous clusters. The percentage of correctly classified high-concentration pixels increased to 98% for PA(44:0) and 80% for PC(38:0). The combination of both algorithms makes it possible to obtain image segmentation with high quantitative agreement with the ground truth.

Overlapping isotopic envelopes are ubiquitous. In order to see whether overlapping isotopic envelopes pose a substantial challenge in the analysis of real mass spectrometric images, and whether they can lead to misleading ion images, we have

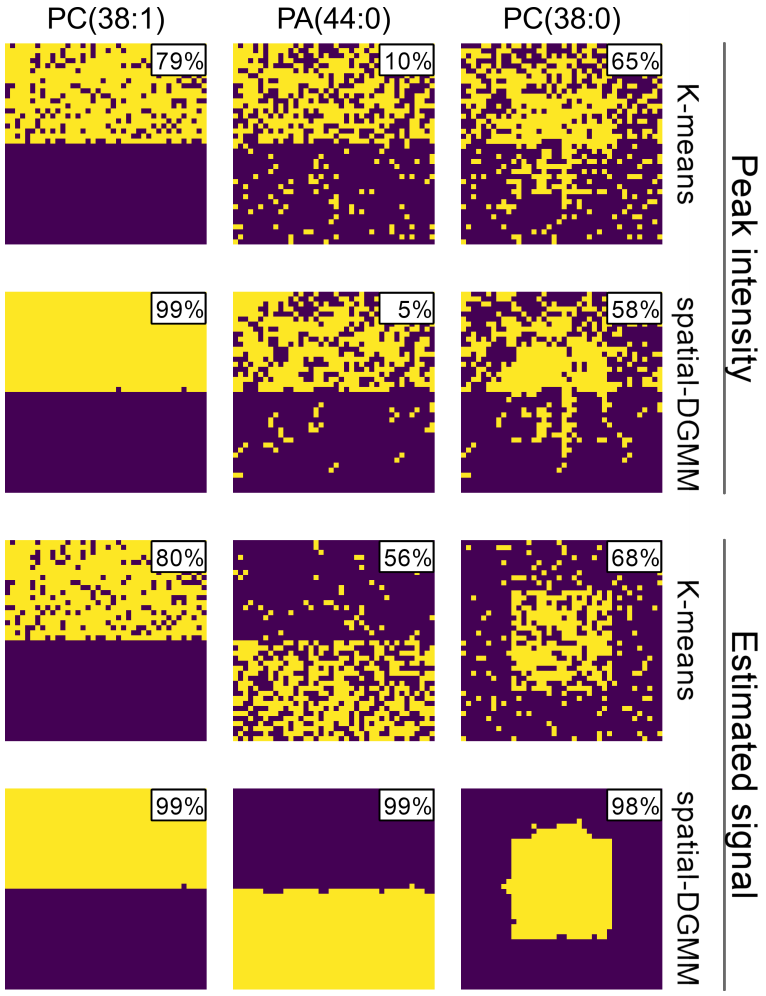

FIGURE 6.2: K-means clustering of peak intensities wrongly suggests that PA(44:0) is concentrated in the top half of the image and produces noisy segments. Estimating ion signals with `masserstein` corrects the lipid spatial distribution by separating overlapping isotopic envelopes. Segmenting the estimated signals with `spatial-DGMM` produces spatially homogeneous clusters by modeling the image's spatial structure. Numbers in the top-right corners show the percentages of correctly classified pixels.

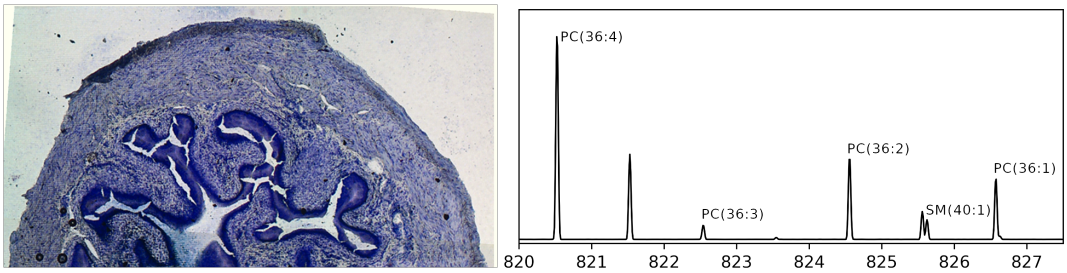

FIGURE 6.3: Left: An optical image of the tissue section of a mouse bladder used to obtain a mass spectrometric image analyzed in this Chapter. The image was published alongside the mass spectrometric image in the PRIDE database (ID PXD001283). Right: Using `masserstein` to fit a linear combination of theoretical isotopic envelopes to the average spectrum of the mass spectrometric image provides a putative annotation of a cluster of isotopic envelopes.

analyzed a previously published image of a mouse bladder tissue section (Fig. 6.3) [98]. We have used chemical formulas of 2460 glycerolipids from the LIPID MAPS database and calculated their theoretical isotopic envelopes. According to the original article, we have assumed a potassium adduct.

Out of the 2460 lipid ions, `masserstein` detected 78 in the average spectrum of the tissue. Only 33 of them were not subject to an interference from a lower-mass lipid with an overlapping isotopic envelope. We have detected 17 clusters of at least two overlapping isotopic envelopes, jointly accounting for 62 lipid ions.

Fitting theoretical spectra provided a putative annotation of peaks. We have selected a cluster of 5 overlapping isotopic envelopes for further analysis (Fig. 6.3). The peaks of the cluster were annotated with lipids detected by `masserstein`, further verified by accurate mass matching against the LIPID MAPS database. The annotated lipids are phosphatidylcholines and a sphingomyelin, classes commonly discovered in lipidomics MSI experiments. Other lipids from those classes were discovered in this image and verified through MS/MS in the original work.

Estimating lipids' signal was equivalent to increasing the spectrometer's resolving power. Due to the lack of ground truth about the lipid locations, in order to compare a K-means segmentation based on peak intensity and a `spatial-DGMM` segmentation based on lipid signals, we have first performed a computational experiment in which we have artificially lowered the mass resolutions of the pixel spectra by applying a Gaussian filter (Fig. 6.4).

This resulted in merging of the monoisotopic peak of SM(40:1), a lipid located in the muscle tissue, with the second peak of PC(36:2), a lipid located mostly in the urothelium. As a consequence, the ion images made from low-resolution spectra suggest that SM(40:1) is located throughout the whole tissue (Fig. 6.4). However, `masserstein` was still able to return the correct spatial distribution, with minimal changes compared to the image generated from full-resolution spectra. This shows that `masserstein` is able to correctly separate overlapping signals, and therefore it should give more accurate results than ion images in full-resolution spectra as well.

Spatially-aware segmentation of estimated ion signals recovered underlying tissues. We have estimated the spatial distributions of the lipids using three methods: by taking their ion images, by taking ion images at the $n+1$ peaks (i.e. peaks 1 Da heavier than the monoisotopic ones, which are typically less affected by interferences from lighter lipids [94]) and by fitting their theoretical spectra to each pixel spectrum. The results are compared in Fig. 6.5. The estimation using `masserstein`

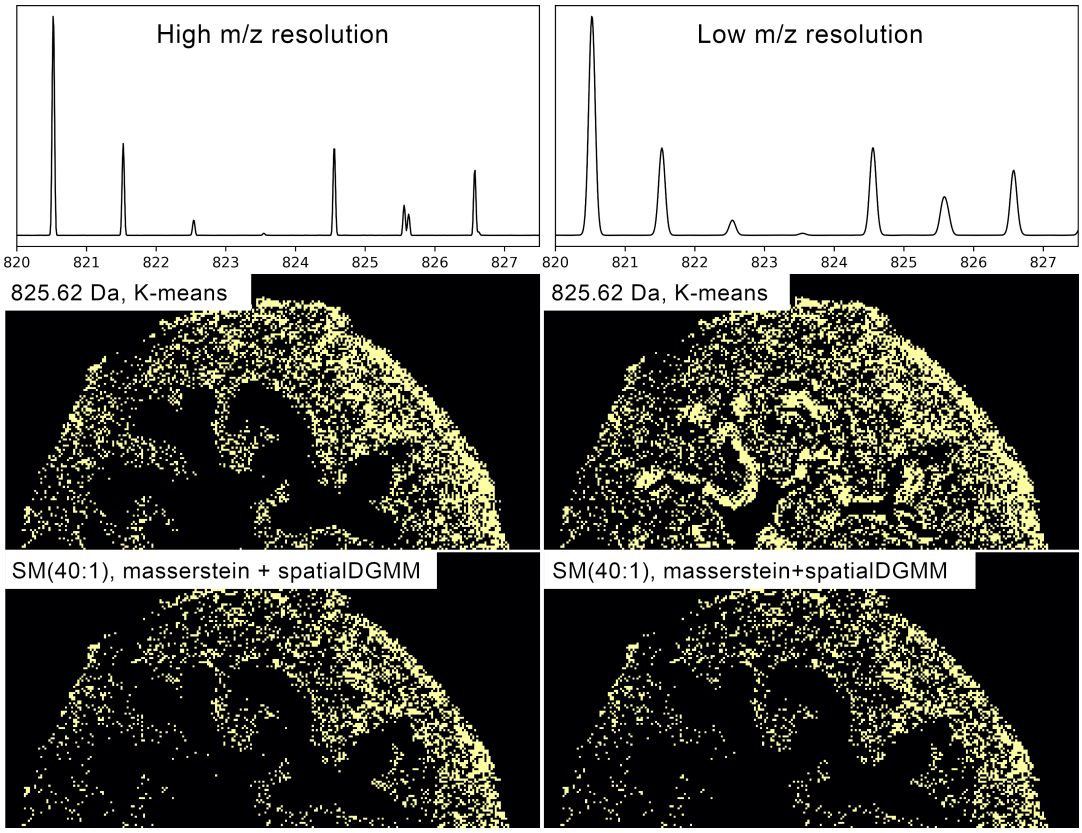

FIGURE 6.4: Estimation with `masserstein` has the same qualitative effect as increasing the resolving power of the spectrometer. In low-resolution spectra, the monoisotopic peak of SM(40:1) at 825.62 Da merges with the $n+1$ peak of PC(36:2). The ion image is then a mixture of the distributions of both lipids. In consequence, the K-means segment occupies the whole sample. Fitting theoretical isotopic envelopes of the lipids recovers the true spatial distribution visible in ion images from high-resolution spectra. Accordingly, the K-means segment corresponds to the muscle tissue.

typically resembled the $n+1$ ion images more than the monoisotopic ones, seemingly due to filtering of interferences by separation of overlapping signals.

The distributions obtained with `masserstein` and monoisotopic peak ion image are highly divergent for PC(36:3). While `masserstein` shows that this lipid is located mostly in the umbrella cells (a subregion of the urothelium), its ion image suggests that it is located in the whole urothelium, and the enrichment in the umbrella cells is not clearly visible. The ion image also shows a considerable signal of PC(36:3) in the muscle cells, a region in which this lipid is absent according to `masserstein`. The differences are seemingly caused by an interference from PC(36:4), which isotopic envelope overlaps with the monoisotopic peak of PC(36:3). As the lower-mass lipid is highly abundant, it has a significant impact on the ion image of the heavier one.

As in the case of the simulated images, the K-means segmentation of lipid signals returns highly dispersed clusters, resulting in a poor quantitative agreement with the actual locations of tissues in the mouse bladder. Segmentation with `spatial-DGMM` results in more spatially homogeneous clusters, which have a higher agreement with the true anatomical regions (Fig. 6.6). In particular, for PC(36:4), we can see a clear segmentation of the image into a segment consisting of urothelium and the adventitial layer (Fig. 6.6, top row, left, yellow) and a segment corresponding to the muscle

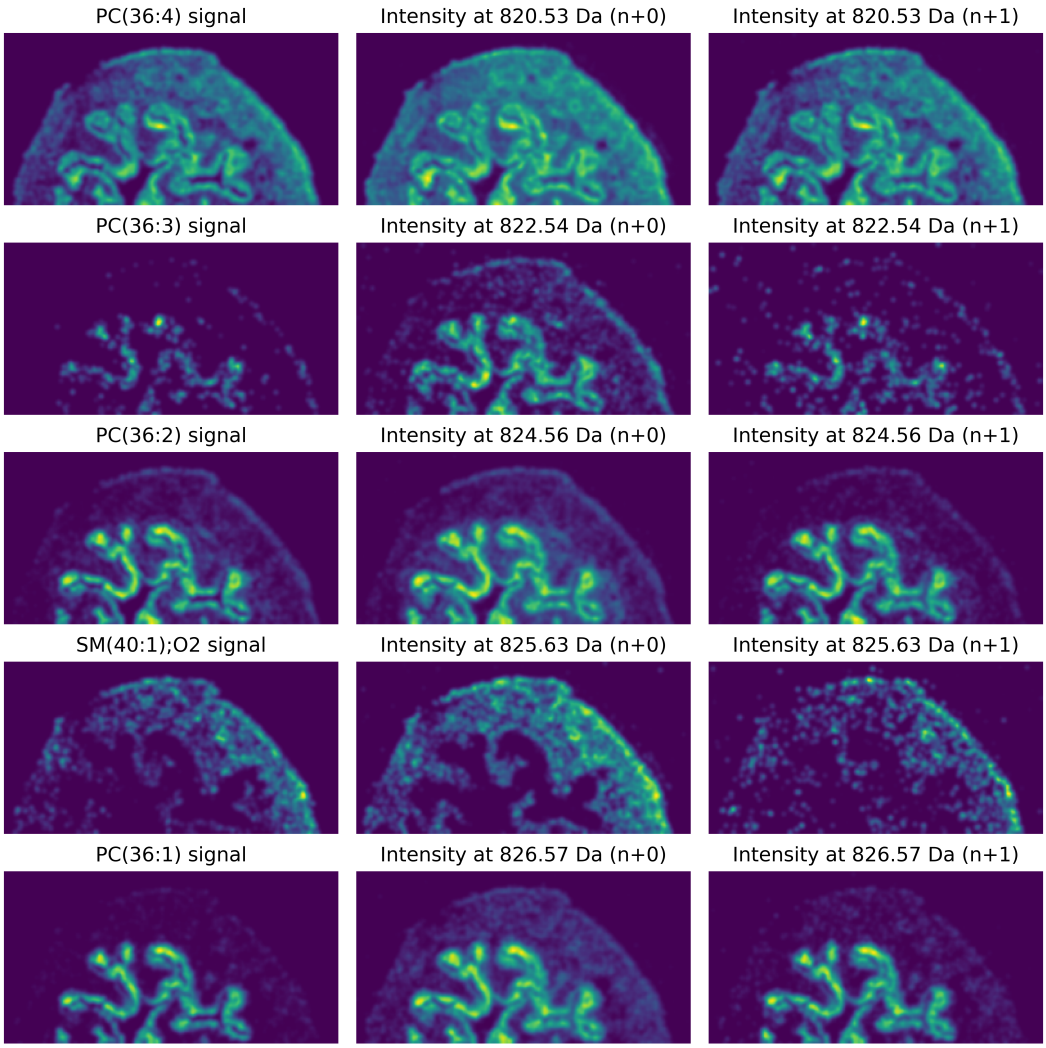

FIGURE 6.5: Signal of lipids estimated with `masserstein` is less prone to interferences from lighter lipids than the monoisotopic peak intensity. Estimation with `masserstein` resembles the intensity of $n+1$ peak, showing that `masserstein` uses information from the whole isotopic envelope. In case of lipids, the $n+1$ peak was previously reported as less prone to interferences.

tissue (Fig. 6.6, top row, left, aquamarine), while the segments obtained with K-means clustering of peak intensities for k=2 and k=4 mix all tissues (Fig. 6.6, top row, middle and right panel). Similarly, for PC(36:2), we can see clear regions of concentration characteristic for the umbrella cells (Fig. 6.6, third row, left, yellow) and the rest of the urothelium (Fig. 6.6, third row, left, aquamarine), while the K-means segmentation results in much less clearly delineated regions (Fig. 6.6, third row, middle and right).

6.3 Summary of the Chapter

Overlapping isotopic envelopes in mass spectrometric images seem to be the norm rather than an exception. In the mouse bladder data set, we have found that the monoisotopic peaks of at least 60% of detected lipids are, to some extent, influenced by isotopic envelopes of other ions. Since the image contained multiple molecules other than lipids, this figure is an underestimation of the true frequency of this phenomenon.

As demonstrated with simulated data sets and confirmed on an experimental one, the interfering signal from an overlapping isotopic envelope can dramatically influence an ion image. In the case of low-intensity ions influenced by high-intensity ones, this can lead to wrong conclusions about their spatial distributions. The common practice to segment an image disregarding the spatial relationships between pixels is subject to another type of overlap, where different regions of interest can have pixels with similar ion intensities due to pixel-to-pixel variability. This leads to spatially dispersed clusters, with each cluster containing pixels from parts of several different tissues. Both phenomena have negative implications for tissue characterization and biomarker detection, making the results less reliable. At the same time, they currently seem to be the main challenges in obtaining accurate image segmentation.

The two challenges can be overcome using recent developments in computational mass spectrometry designed to resolve overlapping isotopic envelopes and mitigate the effect of pixel-to-pixel variability. The `masserstein` tool returns a correct spatial distribution of the signal of a lipid when ion images generated from monoisotopic peak intensity fail due to interferences from lighter lipids. The `spatial-DGMM` tool combines closely located pixels into both spatially and chemically homogeneous segments. Combining both approaches makes it possible to obtain a more biologically meaningful univariate segmentation of mass spectrometry images.

Arguably, the downside of the approach presented in this paper is that more advanced tools require more effort in tuning their parameters. Often, multiple values need to be inspected in order to obtain a segmentation that matches the expectations based e.g. on histological staining. Further research in the methodology of applying software tools and diagnosing their results is needed to give precise procedures of parameter tuning.

In this Chapter, we have analyzed a mass spectrometry image of a mouse bladder assuming that lipids are ionized with a potassium adduct. This is a prevalent assumption in MSI data analysis, including the original work with which the data set was published. However, comparing the average spectrum of the image to the theoretically predicted lipid spectra and inspecting the fine isotopic distribution seems to contradict this assumption (Fig. 6.7).

Although `masserstein` is robust to moderate model misspecifications, the lack of 41K peak in the pixel spectra had a negative effect on the software's performance.

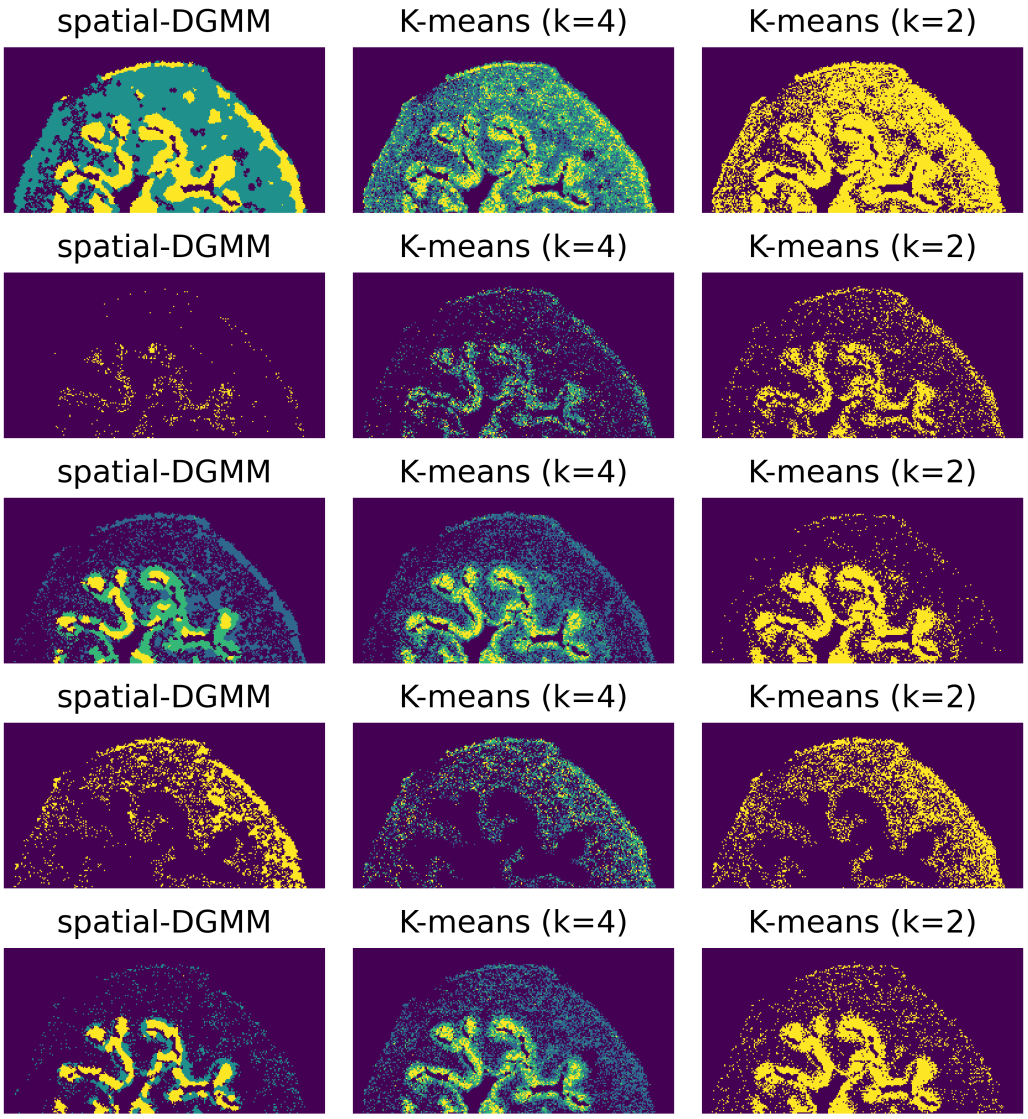

FIGURE 6.6: Segmentation with `spatial-DGMM` increases the spatial homogeneity of segments compared to K-means, leading to a better agreement with the underlying anatomical regions. Rows correspond to lipids in the order of Fig. 6.5

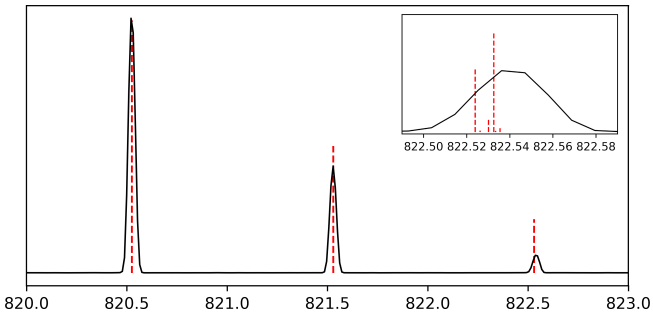

FIGURE 6.7: A fragment of the average spectrum of the image overlaid with a theoretical isotopic envelope of PC(36:4). Although the first two peaks match almost exactly, the third theoretical peak is over two times too high and shifted to the left. Inspecting its fine isotopic distribution shows that this peak splits into two, of which the lighter one corresponds to 41K isotope, which seems to be missing in the experimental spectrum and causes the apparent shift of intensity and location.

For this reason, we had to truncate the theoretical spectra to only contain the first two peaks. Although we still could obtain a good separation of the overlapping isotopic envelopes, solving this problem — either by software development or further studies in mass spectrometry — will most likely improve the results even further. The ability to use full theoretical isotopic envelopes would make it possible to resolve more complex clusters with an even higher accuracy.

Simulated data sets are often met with skepticism, as many people believe that they do not reflect many features of real mass spectra. However, a properly simulated mass spectrometric image can be hardly distinguishable from an experimentally acquired one. Such data sets prove useful in theoretical studies on the properties of mass spectra and spectrometric images, revealing potential problems in data analysis and their solutions.

Chapter 7

Conclusions

In this dissertation, we have presented an approach to computational spectrometry and spectroscopy based on the notion of an optimal transport of signal between spectra. Our main mathematical tool is the Wasserstein distance, which quantifies the difference between two spectra as the minimal distance in the m/z axis needed to match their signals. The Wasserstein distance was the basis for the development of a regression-denoising algorithm for fitting a linear combination of reference spectra to an experimental spectrum of a mixture of chemical compounds. The regression-denoising algorithm was implemented as a Python 3 package called `masserstein`, available at https://github.com/mciach/masserstein. We have demonstrated the practical applicability of our approach by improving the methods of segmentation of mass spectrometric images.

Our approach is not the first attempt at the problem of linear regression of mass spectra. In fact, this problem has been tackled multiple times, usually in the context of very specific experiments. Many of the previous solutions are very crude, without mathematical formalism that would allow for their theoretical analysis. One of such examples is to estimate the intensity of the lightest ion by integrating its monoisotopic peak, and then to subtract the isotopic envelope of this ion from the analyzed cluster of overlapping envelopes. Such procedures are not only difficult to analyze, but also to generalize to other kinds of experiments and experimental methods.

Our main conceptual contribution is to treat spectra as probabilistic measures to encompass both profile and centroid spectra in a single mathematical formalism, and to use the notion of optimal transport to compare different spectra to each other. Thanks to this approach, we were able to develop a method of fitting a linear combination of discrete theoretical spectra to a continuous experimental one. Our approach is the first that is capable of such fitting. Further research has the potential to improve the quality of estimation in such setting.

This work is a result of an interdisciplinary collaboration between mathematicians, biologists, statisticians, chemists, and computer scientists. Through our work, we have identified practical difficulties encountered in experimental research with mass spectrometry or NMR spectroscopy. We have expressed those difficulties as mathematical problems and solved them using optimization algorithms. Finally, we have shown that our approach can be generalized to other types of data analysis and makes it possible to improve the biological accuracy of the results. Starting from natural sciences, we have moved to abstract mathematical definitions, solved mathematical problems, and returned back to natural sciences to apply our solutions.

Although, for most of this thesis, we presented `masserstein` as a tool for the analysis of mass spectra, this was done mostly for the sake of consistency and simplicity of the exposition of the method. The optimal transport paradigm and the regression-denoising algorithm are applicable to multiple kinds of spectrometry and

spectroscopy, including the nuclear magnetic resonance (NMR) spectroscopy. Preliminary experiments, which were not included in this dissertation, show very promising results for the analysis of NMR spectra, and show that `masserstein` is capable of an accurate estimation of concentrations of molecules from particularly complex spectra which are either difficult or impossible to analyze by hand.

Bibliography

[1] Fred W McLafferty, František Tureček, and Frantisek Turecek. *Interpretation of mass spectra*. University science books, 1993.

[2] Katerina Makarova, Joanna J Sajkowska-Kozielewicz, Katarzyna Zawada, Ewa Olchowik-Grabarek, Michał Aleksander Ciach, Krzysztof Gogolewski, Natalia Dobros, Paulina Ciechowicz, Hélène Freichels, and Anna Gambin. Harvest time affects antioxidant capacity, total polyphenol and flavonoid content of polish st john's wort's (hypericum perforatum l.) flowers. *Scientific reports*, 11(1):1–12, 2021.

[3] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[4] F. Santambrogio. *Optimal transport for applied mathematicians*. Springer, 2015.

[5] Szymon Majewski, Michal Aleksander Ciach, Michal Startek, Wanda Niemyska, Blazej Miasojedow, and Anna Gambin. The wasserstein distance as a dissimilarity measure for mass spectra with application to spectral deconvolution. In *18th International Workshop on Algorithms in Bioinformatics (WABI 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[6] Michał Aleksander Ciach, Błażej Miasojedow, Grzegorz Skoraczyński, Szymon Majewski, Michał Startek, Dirk Valkenborg, and Anna Gambin. Masserstein: Linear regression of mass spectra by optimal transport. *Rapid Communications in Mass Spectrometry*, page e8956, 2021.

[7] Michał Aleksander Ciach, Dan Guo, Olga Vitek, and Anna Gambin. Resolving overlapping isotopic envelopes improves segmentation of mass spectrometric images. W przygotowaniu.

[8] Juris Meija. Mathematical tools in analytical mass spectrometry. *Analytical and Bioanalytical Chemistry*, 385(3):486–499, 2006.

[9] Tobias Kind and Oliver Fiehn. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics*, 8(1):1–20, 2007.

[10] Ingvar Eidhammer, Kristian Flikka, Lennart Martens, and Svein-Ole Mikalsen. *Computational methods for mass spectrometry proteomics*. John Wiley & Sons, 2008.

[11] Ryan Peckner, Samuel A Myers, Alvaro Sebastian Vaca Jacome, Jarrett D Egertson, Jennifer G Abelin, Michael J MacCoss, Steven A Carr, and Jacob D Jaffe. Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics. *Nature methods*, 15(5):371, 2018.

[12] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[13] Klaus Linde, Gilbert Ramirez, Cynthia D Mulrow, Andrej Pauls, Wolfgang Weidenhammer, and Dieter Melchart. St john's wort for depression—an overview and meta-analysis of randomised clinical trials. *Bmj*, 313(7052):253–258, 1996.

[14] Veronika Butterweck, Guido Jürgenliemk, Adolf Nahrstedt, and Hilke Winterhoff. Flavonoids from hypericum perforatum show antidepressant activity in the forced swimming test. *Planta medica*, 66(01):3–6, 2000.

[15] Manuela Leri, Maria Scuto, Maria Laura Ontario, Vittorio Calabrese, Edward J Calabrese, Monica Bucciantini, and Massimo Stefani. Healthy effects of plant polyphenols: molecular mechanisms. *International journal of molecular sciences*, 21(4):1250, 2020.

[16] Juana Benedí, Rocio Arroyo, Carmen Romero, Sagrario Martín-Aragón, and Angel M Villar. Antioxidant properties and protective effects of a standardized extract of hypericum perforatum on hydrogen peroxide-induced oxidative damage in pc12 cells. *Life Sciences*, 75(10):1263–1276, 2004.

[17] Biljana Božin, Nebojša Kladar, Nevena Grujić, Goran Anačkov, Isidora Samojlik, Neda Gavarić, and Branislava Srđenović Čonić. Impact of origin and biological source on chemical composition, anticholinesterase and antioxidant properties of some st. john's wort species (hypericum spp., hypericaceae) from the central balkans. *Molecules*, 18(10):11733–11750, 2013.

[18] Emily J Hunt, Cynthia E Lester, Elizabeth A Lester, and Randall L Tackett. Effect of st. john's wort on free radical production. *Life sciences*, 69(2):181–190, 2001.

[19] Edoardo Napoli, Laura Siracusa, Giuseppe Ruberto, Alessandra Carrubba, Silvia Lazzara, Antonio Speciale, Francesco Cimino, Antonella Saija, and Mariateresa Cristani. Phytochemical profiles, phototoxic and antioxidant properties of eleven hypericum species–a comparative study. *Phytochemistry*, 152:162–173, 2018.

[20] Dejan Z Orčić, Neda M Mimica-Dukić, Marina M Francišković, Slobodan S Petrović, and Emilija Đ Jovin. Antioxidant activity relationship of phenolic compounds in hypericum perforatuml. *Chemistry Central Journal*, 5(1):1–8, 2011.

[21] Bruno A Silva, Federico Ferreres, Joao O Malva, and Alberto CP Dias. Phytochemical and antioxidant characterization of hypericum perforatum alcoholic extracts. *Food chemistry*, 90(1-2):157–167, 2005.

[22] Yanping Zou, Yanhua Lu, and Dongzhi Wei. Antioxidant activity of a flavonoid-rich extract of hypericum perforatum l. in vitro. *Journal of Agricultural and Food Chemistry*, 52(16):5032–5039, 2004.

[23] Edita Bagdonaitė, Valdimaras Janulis, Liudas Ivanauskas, and Juozas Labokas. Ex situ studies on chemical and morphological variability of hypericum perforatum l. in lithuania. *Biologija*, 53(3), 2007.

[24] Edita Bagdonaitė, Pavol Mártonfi, Miroslav Repčák, and Juozas Labokas. Variation in concentrations of major bioactive compounds in hypericum perforatum l. from lithuania. *Industrial Crops and Products*, 35(1):302–308, 2012.

[25] Ali Kemal Ayan, Cüneyt Çirak, and Oguzhan Yanar. Variations in total phenolics during ontogenetic, morphogenetic, and diurnal cycles inhypericum species from turkey. *Journal of Plant Biology*, 49(6):432–439, 2006.

[26] Ping Sun, Tianlan Kang, Hua Xing, Zhen Zhang, Delong Yang, Jinlin Zhang, Paul W Paré, and Mengfei Li. Phytochemical changes in aerial parts of hypericum perforatum at different harvest stages. *Records of Natural Products*, 13(1), 2019.

[27] Renato Bruni and Gianni Sacchetti. Factors affecting polyphenol biosynthesis in wild and field grown st. john's wort (hypericum perforatum l. hypericaceae/guttiferae). *Molecules*, 14(2):682–725, 2009.

[28] Eirini Sarrou, Lefki-Pavlina Giassafaki, Domenico Masuero, Daniele Perenzoni, Ioannis S Vizirianakis, Maria Irakli, Paschalina Chatzopoulou, and Stefan Martens. Metabolomics assisted fingerprint of hypericum perforatum chemotypes and assessment of their cytotoxic activity. *Food and Chemical Toxicology*, 114:325–333, 2018.

[29] Raffaella Filippini, Anna Piovan, Anna Borsarini, and Rosy Caniato. Study of dynamic accumulation of secondary metabolites in three subspecies of hypericum perforatum. *Fitoterapia*, 81(2):115–119, 2010.

[30] Mehmet Serhat Odabas, Necdet Camas, Cuneyt Cirak, Jolita Radušiene, Valdimaras Janulis, and Liudas Ivanauskas. The quantitative effects of temperature and light intensity on phenolics accumulation in st. john's wort (hypericum perforatum). *Natural Product Communications*, 5(4):1934578X1000500408, 2010.

[31] Gordana Zdunic, Dejan Godjevac, Katarina Savikin, and Silvana Petrovic. Comparative analysis of phenolic compounds in seven hypericum species and their antioxidant properties. *Natural product communications*, 12(11):1934578X1701201140, 2017.

[32] Cuneyt Cirak and Jolita Radusiene. Factors affecting the variation of bioactive compounds in hypericum species. *Biologia futura*, 70(3):198–209, 2019.

[33] Anna Rita Bilia, Maria Camilla Bergonzi, Giovanni Mazzi, and Franco Francesco Vincieri. Analysis of plant complex matrices by use of nuclear magnetic resonance spectroscopy: St. john's wort extract. *Journal of agricultural and food chemistry*, 49(5):2115–2124, 2001.

[34] Teresa W-M Fan. Metabolite profiling by one-and two-dimensional nmr analysis of complex mixtures. *Progress in nuclear magnetic resonance spectroscopy*, 28(2):161–219, 1996.

[35] Bonnie Rasmussen, Olivier Cloarec, Huiru Tang, Dan Stærk, and Jerzy W Jaroszewski. Multivariate analysis of integrated and full-resolution 1h-nmr spectral data from complex pharmaceutical preparations: St. john's wort. *Planta medica*, 72(06):556–563, 2006.

[36] B Broda and J Mowszowicz. Guide to determination of medicinal, poisonous and usable plants. *Warszawa: Wyd. Lekarskie PZWL*, 2000.

[37] Amal Ben Amira, Christophe Blecker, Aurore Richel, Anthony Argüelles Arias, Patrick Fickers, Frédéric Francis, Souhail Besbes, and Hamadi Attia. Influence of the ripening stage and the lyophilization of wild cardoon flowers on their chemical composition, enzymatic activities of extracts and technological properties of cheese curds. *Food chemistry*, 245:919–925, 2018.

[38] Deisy dos Santos Freitas, Wilian da Silva Nunes, Rafael do Prado Apparecido, Thiago Inácio Barros Lopes, and Glaucia Braz Alcantara. Nmr-based approach reveals seasonal metabolic changes in mate (ilex paraguariensis a. st.-hil.). *Magnetic Resonance in Chemistry*, 56(5):311–320, 2018.

[39] C. E. Housecroft and E. C. Constable. *Chemistry: An introduction to organic, inorganic and physical chemistry*. Pearson education, 2010.

[40] S. Böcker and K. Dührkop. Fragmentation trees reloaded. *Journal of Cheminformatics*, 8(1):5, 2016.

[41] S. Neumann and S. Böcker. Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. *Analytical and Bioanalytical Chemistry*, 398(7-8):2779–2788, 2010.

[42] K. X. Wan, I. Vidavsky, and M. L. Gross. Comparing similar spectra: from similarity index to spectral contrast angle. *J. of Am. Soc. Mass Spectrom.*, 13(1):85–88, 2002.

[43] Ş. Yilmaz, E. Vandermarliere, and L. Martens. *Methods to Calculate Spectrum Similarity*, pages 75–100. Springer New York, New York, NY, 2017.

[44] M. E. Hansen and J. Smedsgaard. A new matching algorithm for high resolution mass spectra. *J. of Am. Soc. Mass Spectrom.*, 15(8):1173 – 1180, 2004.

[45] Y. Rubner, C. Tomasi, and L.J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, Nov 2000.

[46] Bruce B Reinhold and Vernon N Reinhold. Electrospray ionization mass spectrometry: Deconvolution by an entropy-based algorithm. *Journal of the American Society for Mass Spectrometry*, 3(3):207–215, 1992.

[47] Jedrzej Jablonski and Anna Marciniak-Czochra. Efficient algorithms computing distances between radon measures on r. *arXiv preprint arXiv:1304.3501*, 2013.

[48] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, et al. Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry*, 45(7):703–714, 2010.

[49] Nina Nikolova and Joanna Jaworska. Approaches to measure chemical similarity–a review. *Molecular Informatics*, 22(9-10):1006–1026, 2003.

[50] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

[51] Pan Du, Warren A Kibbe, and Simon M Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *bioinformatics*, 22(17):2059–2065, 2006.

[52] Jonathan Lu, Michael J Trnka, Soung-Hun Roh, Philip JJ Robinson, Carrie Shiau, Danica Galonic Fujimori, Wah Chiu, Alma L Burlingame, and Shenheng Guan. Improved peak detection and deconvolution of native electrospray mass spectra from large protein complexes. *Journal of the American Society for Mass Spectrometry*, 26(12):2141–2151, 2015.

[53] Frederik Lermyte, Mateusz Krzysztof Łącki, Dirk Valkenborg, Geert Baggerman, Anna Gambin, and Frank Sobott. Understanding reaction pathways in top-down etd by dissecting isotope distributions: A mammoth task. *International Journal of Mass Spectrometry*, 390:146–154, 2015.

[54] Kevin De Bruycker, Tim Krappitz, and Christopher Barner-Kowollik. High performance quantification of complex high resolution polymer mass spectra. *ACS Macro Letters*, 7(12):1443–1447, 2018.

[55] Martin S Engler, Sarah Crotty, Markus J Barthel, Christian Pietsch, Katrin Knop, Ulrich S Schubert, and Sebastian Böcker. Coconut - an efficient tool for estimating copolymer compositions from mass spectra. *Analytical chemistry*, 87(10):5223–5231, 2015.

[56] Juris Meija and Joseph A Caruso. Deconvolution of isobaric interferences in mass spectra. *Journal of the American Society for Mass Spectrometry*, 15(5):654–658, 2004.

[57] Juris Meija, Thomas L Beck, and Joseph A Caruso. Interpretation of alkyl diselenide and selenosulfenate mass spectra. *Journal of the American Society for Mass Spectrometry*, 15(9):1325–1332, 2004.

[58] Surendra Dasari, Phillip A Wilmarth, Ashok P Reddy, Lucinda JG Robertson, Srinivasa R Nagalla, and Larry L David. Quantification of isotopically overlapping deamidated and 18o-labeled peptides using isotopic envelope mixture modeling. *Journal of proteome research*, 8(3):1263–1270, 2009.

[59] F. Lermyte et al. Conformational Space and Stability of ETD Charge Reduction Products of Ubiquitin. *J. Am. Soc. Mass Spectrom.*, Aug 2016.

[60] M.M Koek et al. Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics*, 7(3):307–328, Sep 2011.

[61] K. Xiao et al. Accurate and Efficient Resolution of Overlapping Isotopic Envelopes in Protein Tandem Mass Spectra. *Sci Rep*, 5:14755, Oct 2015.

[62] J. Meija and J.A. Caruso. Deconvolution of isobaric interferences in mass spectra. *J. of Am. Soc. Mass Spectrom.*, 15(5):654 – 658, 2004.

[63] N. Jaitly et al. Decon2ls: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinformatics*, 10(1):87, 2009.

[64] Q. Kou, L. Xun, and X. Liu. Toppic: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics*, 32(22):3495–3497, 2016.

[65] AG Ferrige, MJ Seddon, S Jarvis, John Skilling, and Robert Aplin. Maximum entropy deconvolution in electrospray mass spectrometry. *Rapid communications in mass spectrometry*, 5(8):374–377, 1991.

[66] M. Mann, C. K. Meng, and J. B. Fenn. Interpreting mass spectra of multiply charged ions. *Analytical Chemistry*, 61(15):1702–1708, 1989.

[67] B. B. Reinhold and V. N. Reinhold. Electrospray ionization mass spectrometry: Deconvolution by an entropy-based algorithm. *J. of Am. Soc. Mass Spectrom.*, 3(3):207 – 215, 1992.

[68] Salvatore Cappadona, Peter R Baker, Pedro R Cutillas, Albert JR Heck, and Bas van Breukelen. Current challenges in software solutions for mass spectrometry-based quantitative proteomics. *Amino acids*, 43(3):1087–1108, 2012.

[69] Mateusz K Łącki, Frederik Lermyte, Błażej Miasojedow, Michał P Startek, Frank Sobott, Dirk Valkenborg, and Anna Gambin. masstodon: A tool for assigning peaks and modeling electron transfer reactions in top-down mass spectrometry. *Analytical chemistry*, 91(3):1801–1807, 2019.

[70] David M Horn, Roman A Zubarev, and Fred W McLafferty. Automated reduction and interpretation of. *Journal of the American Society for Mass Spectrometry*, 11(4):320–332, 2000.

[71] J. Gondzio. Interior point methods 25 years later. *European Journal of Operational Research*, 218(3):587–601, 2012.

[72] Mateusz K Łącki, Michał Startek, Dirk Valkenborg, and Anna Gambin. Isospec: Hyperfast fine structure calculator. *Analytical chemistry*, 89(6):3272–3277, 2017.

[73] Michael W. Senko, Steven C. Beu, and Fred W. McLafferty. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 6(4):229 – 233, 1995.

[74] R. J. Vanderbei et al. *Linear programming*. Springer, 2015.

[75] Yadolah Dodge. *Least Absolute Deviation Regression*, pages 299–302. Springer New York, New York, NY, 2008.

[76] Harvey M Wagner. Linear programming techniques for regression analysis. *Journal of the American Statistical Association*, 54(285):206–212, 1959.

[77] Hannes L Röst, Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aicheler, Sandro Andreotti, Hans-Christian Ehrlich, Petra Gutenbrunner, Erhan Kenar, et al. Openms: a flexible open-source software platform for mass spectrometry data analysis. *Nature methods*, 13(9):741, 2016.

[78] Kai Dührkop, Markus Fleischauer, Marcus Ludwig, Alexander A Aksenov, Alexey V Melnik, Marvin Meusel, Pieter C Dorrestein, Juho Rousu, and Sebastian Böcker. Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature methods*, 16(4):299, 2019.

[79] Parminder Kaur and Peter B O'Connor. Algorithms for automatic interpretation of high resolution mass spectra. *Journal of the American Society for Mass Spectrometry*, 17(3):459–468, 2006.

[80] Klara Scupakova, Zita Soons, Gokhan Ertaylan, Keely A Pierzchalski, Gert B Eijkel, Shane R Ellis, Jan W Greve, Ann Driessen, Joanne Verheij, Theo M De Kok, et al. Spatial systems lipidomics reveals nonalcoholic fatty liver disease heterogeneity. *Analytical chemistry*, 90(8):5130–5138, 2018.

[81] Sangwon Cha, Hui Zhang, Hilal I Ilarslan, Eve Syrkin Wurtele, Libuse Brachova, Basil J Nikolau, and Edward S Yeung. Direct profiling and imaging of plant metabolites in intact tissues by using colloidal graphite-assisted laser desorption ionization mass spectrometry. *The Plant Journal*, 55(2):348–360, 2008.

[82] Tae Hun Hahm, Mitsuru Tanaka, Huu-Nghi Nguyen, Ayaka Tsutsumi, Koichi Aizawa, and Toshiro Matsui. Matrix-assisted laser desorption/ionization mass spectrometry-guided visualization analysis of intestinal absorption of acylated anthocyanins in sprague-dawley rats. *Food Chemistry*, 334:127586, 2021.

[83] Hanna Bednarz, Nils Roloff, and Karsten Niehaus. Mass spectrometry imaging of the spatial and temporal localization of alkaloids in nightshades. *Journal of agricultural and food chemistry*, 67(49):13470–13477, 2019.

[84] Katharina Krueger, Cindy Terne, Carsten Werner, Uwe Freudenberg, Vera Jankowski, Walter Zidek, and Joachim Jankowski. Characterization of polymer membranes by maldi mass-spectrometric imaging techniques. *Analytical chemistry*, 85(10):4998–5004, 2013.

[85] Amanda Rae Buchberger, Kellen DeLaney, Jillian Johnson, and Lingjun Li. Mass spectrometry imaging: a review of emerging advancements and future insights. *Analytical chemistry*, 90(1):240, 2018.

[86] Theodore Alexandrov. Maldi imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC bioinformatics*, 13(16):1–13, 2012.

[87] Nico Verbeeck, Richard M Caprioli, and Raf Van de Plas. Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. *Mass spectrometry reviews*, 39(3):245–291, 2020.

[88] Hang Hu, Ruichuan Yin, Hilary M Brown, and Julia Laskin. Spatial segmentation of mass spectrometry imaging data by combining multivariate clustering and univariate thresholding. *Analytical chemistry*, 93(7):3477–3485, 2021.

[89] Emrys A Jones, Sören-Oliver Deininger, Pancras CW Hogendoorn, André M Deelder, and Liam A McDonnell. Imaging mass spectrometry statistical analysis. *Journal of proteomics*, 75(16):4962–4989, 2012.

[90] Pengyi Yang, Zili Zhang, Bing B Zhou, and Albert Y Zomaya. A clustering based hybrid system for biomarker selection and sample classification of mass spectrometry data. *Neurocomputing*, 73(13-15):2317–2331, 2010.

[91] Dan Guo, Kylie Bemis, Catherine Rawlins, Jeffrey Agar, and Olga Vitek. Unsupervised segmentation of mass spectrometric ion images characterizes morphology of tissues. *Bioinformatics*, 35(14):i208–i217, 2019.

[92] Theodore Alexandrov and Jan Hendrik Kobarg. Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. *Bioinformatics*, 27(13):i230–i238, 2011.

[93] Kaijie Xiao, Fan Yu, Houqin Fang, Bingbing Xue, Yan Liu, and Zhixin Tian. Accurate and efficient resolution of overlapping isotopic envelopes in protein tandem mass spectra. *Scientific reports*, 5:14755, 2015.

[94] Marcus Höring, Christer S Ejsing, Sabrina Krautbauer, Verena M Ertl, Ralph Burkhardt, and Gerhard Liebisch. Accurate quantification of lipid species affected by isobaric overlap in fourier-transform mass spectrometry. *Journal of lipid research*, 62, 2021.

[95] Miao Wang, Yingying Huang, and Xianlin Han. Accurate mass searching of individual lipid species candidates from high-resolution mass spectra for shotgun lipidomics. *Rapid Communications in Mass Spectrometry*, 28(20):2201–2210, 2014.

[96] Nathan A Seifert, Kirill Prozument, and Michael J Davis. Computational optimal transport for molecular spectra: The fully discrete case. *The Journal of Chemical Physics*, 155(18):184101, 2021.

[97] Nathan A Seifert, Kirill Prozument, and Michael J Davis. Computational optimal transport for molecular spectra: The semi-discrete case. *The Journal of Chemical Physics*, 156(13):134117, 2022.

[98] Andreas Römpp, Sabine Guenther, Yvonne Schober, Oliver Schulz, Zoltan Takats, Wolfgang Kummer, and Bernhard Spengler. Histology by mass spectrometry: label-free tissue characterization obtained from high-accuracy bioanalytical imaging. *Angewandte chemie international edition*, 49(22):3834–3838, 2010.

[99] Yasset Perez-Riverol, Jingwen Bai, Chakradhar Bandla, David García-Seisdedos, Suresh Hewapathirana, Selvakumar Kamatchinathan, Deepti J Kundu, Ananth Prakash, Anika Frericks-Zipper, Martin Eisenacher, et al. The pride database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic acids research*, 50(D1):D543–D552, 2022.

[100] Manish Sud, Eoin Fahy, Dawn Cotter, Alex Brown, Edward A Dennis, Christopher K Glass, Alfred H Merrill Jr, Robert C Murphy, Christian RH Raetz, David W Russell, et al. Lmsd: Lipid maps structure database. *Nucleic acids research*, 35(suppl_1):D527–D532, 2007.

[101] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.