Inaugural-Dissertation

zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich–Mathematischen Gesamtfakultät
der
Ruprecht–Karls–Universität
Heidelberg

vorgelegt von
Dipl.–Inform. Karl Michael Stefan Hanselmann
aus Schwäbisch Hall

Tag der mündlichen Prüfung: 18. Oktober 2010

Computational Methods for the Analysis of Mass Spectrometry Images

Gutachter: Prof. Dr. Fred A. Hamprecht
Prof. Dr. Michael Gertz

Abstract

Mass spectrometry imaging (MSI) is an emerging, promising technology that combines mass spectrometry (MS) with microscopic imaging. This allows the simultaneous analysis of the spatial distribution of thousands of different (bio-)molecules. However, the amount of data acquired renders a direct manual analysis infeasible. This thesis introduces novel methods for the computational analysis of MSI data and thereby contributes to reaching the full potential of this new technology.

Exploratory Data Analysis. In exploratory data analysis, MS images are decomposed into characteristic component spectra and abundance maps. This thesis proposes the application of *probabilistic latent semantic analysis (pLSA)* for non-negative decomposition. It outperforms currently used techniques that do not model the non-negativity of mass spectra, and yields improved interpretability. pLSA is further combined with a statistical complexity estimation scheme to automatically estimate the number of characteristic spectra.

Automated Classification. In many studies, prior knowledge on the composition of a sample is available, e.g., in the form of spatially resolved labels, e.g., for cancerous tissue. This thesis presents a novel workflow for the *automated pixel-based classification* of MS images. It relies on a random forest classifier, which is particularly suitable for such high-dimensional data. Motivated by a local homogeneity assumption, a spatially regularized solution is developed, which is robust to spatial noise. Application to diverse MSI data sets demonstrates that this "digital staining" is a powerful complement to conventional chemical staining techniques.

The Role of Preprocessing. A large-scale study analyzes the dependency of automated classification approaches on the choice and parameterization of the methods used for preprocessing the raw spectra. Furthermore, it is discussed how the analysis may be complicated by high technical and biological variability between measurements.

Efficient Labeling. Labeling of MSI data is time-consuming. To amend this problem, this thesis introduces a novel *multi-class active learning strategy*, which significantly reduces labeling time without compromising on classification accuracy.

Segmentation. The Segmentation of multivariate data is challenging, especially if some of the channels are uncorrelated such as in MS images. This thesis proposes to combine a random forest classifier with a watershed segmentation, and establishes three novel methods to obtain scalar boundary indicator maps from class probability maps. It further introduces the *multivariate watershed* as a generalization of the classic watershed approach.

Zusammenfassung

Die bildgebende Massenspektrometrie ist eine aufstrebende und vielversprechende Technologie, die die Massenspektrometrie (MS) mit der mikroskopischen Bildgebung kombiniert. In einem einzigen Experiment kann die räumliche Verteilung von vielen tausend (Bio-)Molekülen simultan untersucht werden. Aufgrund ihrer Größe können die aufgenommenen Daten jedoch nicht manuell ausgewertet werden. In dieser Arbeit werden neue Methoden für die computergestützte Analyse vorgestellt, die dazu beitragen, das Potential der bildgebenden Massenspektrometrie besser auszuschöpfen.

Explorative Analyse. Bei der explorativen Datenanalyse werden die aufgenommenen MS Bilder in charakteristische Spektren und Verteilungskarten zerlegt. In dieser Arbeit wird die Anwendung der *Probabilistic Latent Semantic Analysis (pLSA)* vorgeschlagen, mit der eine nicht-negative Zerlegung erreicht wird. Diese Methode führt zu besseren und leichter interpretierbaren Ergebnissen als herkömmliche Verfahren, die die Nicht-Negativität von Massenspektren nicht modellieren. Weiterhin wird ein statistisches Kriterium für die Schätzung der Anzahl der charakteristischen Spektren eingeführt.

Automatische Klassifikation. In vielen aktuellen Studien ist ein gewisses Vorwissen über die Zusammensetzung der betrachteten Gewebeprobe vorhanden - z.B. in Form von räumlich aufgelösten Annotationen für Krebs- und Normalgewebe. In dieser Arbeit wird ein neuer Workflow für die *automatische, pixelbasierte Klassifikation* von massenspektrometrischen Bildern präsentiert. Dieser basiert auf einem Random Forest Klassifikator, der sich besonders gut für die Klassifikation von hochdimensionalen Daten (wie MS Bildern) eignet. Unter Annahme von lokaler Homogenität des Gewebes wird das Klassifikationsergebnis räumlich geglättet, um den Einfluss von Rauschen zu vermindern. Experimente mit verschiedenartigen MS Bildern zeigen, dass dieses "digitale Anfärben" eine leistungsstarke Ergänzung zu herkömmlichen chemischen Färbemethoden darstellt.

Einfluss der Vorverarbeitung. In einer umfangreichen Studie wird die Abhängigkeit automatischer Klassifikationsansätze von der Wahl und der Parametrisierung der verwendeten Vorverarbeitungsmethoden untersucht. Weiterhin wird dargelegt, inwiefern die Analyse durch hohe technische und biologische Variabilität zwischen verschiedenen Messungen erschwert werden kann.

Effiziente Annotation. Das Annotieren von massenspektrometrischen Bilden ist zeitaufwändig. In dieser Arbeit wird eine neue *aktive Lernstrategie für Mehrklassenprobleme* vorgestellt. Damit kann die Anzahl der benötigten Annotationen stark vermindert werden, ohne dass Einbußen bei der Klassifikationsgüte hingenommen werden müssen.

Segmentierung. Die Segmentierung von multivariaten Daten ist ein schwieriges Problem - insbesondere, wenn nicht alle Spektralkanäle korreliert sind. In dieser Arbeit wird der Random Forest Klassifikator mit der Wasserscheidensegmentierung kombiniert. Es werden drei neue Methoden vorgeschlagen, mit denen skalare Kantenindikatoren aus Wahrscheinlichkeitskarten gewonnen werden können. Weiterhin wird eine *multivariate Verallgemeinerung des klassischen Wasserscheidenverfahrens* eingeführt.

Acknowledgments

First of all, I would like to thank my advisor Prof. Dr. Fred Hamprecht for providing a stimulating and inspiring scientific work environment. It was him who first introduced me to the challenges of analyzing mass spectrometry images and enthused me with the topic of this thesis. I am very grateful for his constant support and guidance and that he provided me with manifold opportunities for interdisciplinary collaborations, computing resources and financial support. I also greatly benefited from many fruitful and inspiring discussions with my colleagues in the multidimensional image processing group (MIP) at the University of Heidelberg; in particular Dr. habil. Ullrich Köthe, Jens Röder, and my fellow researchers in the mass spectrometry group ("the mass spec conspiracy"), Dr. Marc Kirchner, Dr. Bernhard Renard, Xinghua Lou, Anna Kreshuk and Bernhard Kausler. They were always willing to discuss scientific problems (not necessarily related to mass spectrometry), share their wisdom, contribute valuable advice, and in general are great people to work and spend time with. It was also a pleasure to collaborate with Sebastian Boppel, Björn Voss, Thorben Kröger, Martin Lindner, Martin Riedl, and Buote Xu who all spent some (or more!) time in the mass spectrometry group. Further thanks go to all other members of the MIP group, in particular Christoph Sommer, Frederik Kaster, Björn Andres, Rahul Nair, Christian Scheelen, Christoph Straehle, Nathan Hüsken, Dr. Martin Schmidt and Barbara Werner.

I am also in debt of Prof. Dr. Ron Heeren from the FOM-Institute for Atomic and Molecular Physics (AMOLF) at the University of Amsterdam for granting me access to cutting-edge data, providing the unique opportunity to closely collaborate in highly interdisciplinary projects, and granting me insight into a state-of-the-art mass spectrometry laboratory and its instruments. In my several research visits to AMOLF, the members of his group have introduced me to practical and theoretical aspects of developing novel mass spectrometry (imaging) instruments as well as the challenges of sample preparation and data acquisition. This knowledge proved to be highly valuable in tackling the projects that were part of my thesis. I would especially like to thank Erika Amstalden, Dr. Don Smith, Ivo Klinkert, Dr. Leendert Klerk, and Dr. Andriy Kharchenko for data provision and their efforts to teach a computer scientist basics in biochemistry, physics

and electro engineering. Further thanks go to Dr. Kristine Glunde and Tiffany Greenwood from the Russell H. Morgan Department of Radiology and Radiological Science at the Johns Hopkins University School of Medicine in Baltimore, USA and Dr. Nathalia Giese from the European Pancreas Center, Heidelberg for providing the tissue samples of which data was acquired at AMOLF.

Let me further express my thanks to Prof. Dr. Axel Walch and Benjamin Balluff from the Institute of Pathology at the Helmholtz center in Munich as well as to Prof. Dr. mult. Manfred Schmitt from the Department of Obstetrics and Gynecology at the Technical University of Munich for data provision, support and dedication in our collaboration.

I would also like to thank Prof. Michael Gertz for taking on the role of my second advisor, as well as Prof. Klaus Ambos-Spies and Prof. Reinhard Männer for their participation in my oral exam.

Moreover, I am very grateful for financial support through the Deutsche Forschungsgesellschaft (DFG) and the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences (HGS).

Finally, I would like to thank my family, especially my parents, my brother, my grandparents, and of course my wife Melanie for their constant love and support throughout the time I worked on my thesis and all the years before. One of the first things my parents did when I was born, was to save some money to ensure that I would be able to get a good education. They never questioned that they would always support me, both morally and financially, in my studies of computer science and my PhD thesis, including the possibility to spend some time abroad. It is great to have such a backing! Last but not least, I would like to express my deepest gratitude to my wife Melanie.

For just everything.

To my family.

Contents

Chapter 1

Introduction to this Thesis

In 1990, the National Institutes of Health (NIH) and the US Department of Energy (DOE) launched the human genome project (HGP) to identify and map the approximately 20,000 to 25,000 genes of the human genome. Merely ten years later, the project was completed and the draft of the human genome was announced [220, 51]. By comparing DNA in healthy and diseased tissue, scientists hope to find causes of disease as well as cures and novel treatments. In a related field of research, human DNA is compared to DNA of other creatures. For instance, chimpanzees share about 98% of their DNA with humans but react differently or are even immune to certain diseases like Alzheimer [194, 167]. In comparison to humans, they have a more favorable reponse to the hepatitis B virus (HBV) and an apparently low incidence of epithelial malignancy [153][1].

Today, the human genome project and genomics in general are essential tools for gaining insight into biological processes. In recent years, *proteomics* was established as a complementary field of research. In proteomics, the abundance of proteins and their fragments in different parts of an organism is analyzed and monitored over time, both under constant and varying conditions [66]. Its enormous potential can be demonstrated with a simple example: In figure 1.1 a European peacock butterfly and its corresponding caterpillar are shown. Even though the two creatures appear highly dissimilar, they share the same DNA. It has been understood that genome activation plays a central role in the transformation, and that the different appearances actually result from stark differences in the *proteome* [50], the collectivity of all proteins that the genome produces.

Over- or underexpression of certain proteins, and dysfunctions of protein interaction networks are major causes for disease [50, 7], and proteomics starts to play an important role in medical research and biochemistry. Current research fields include cancer grading [197, 233], Alzheimer's [180] and Parkinson's [199] disease studies, studying

[1] Furthermore, it has long been assumed that chimpanzees are immune to the acquired immune deficiency syndrome (AIDS). This is, however, questioned by recent studies [117].

Figure 1.1.: European peacock *(inachis io)* in two stages of its life cycle: as a butterfly (left) and as a caterpillar (right). Both share the same genome but have a different proteome.

drug distributions and effects (e.g., does a drug actually reach its target), metabolism analysis [15, 53], as well as doping testing [211].

Common methods for scrutinizing the structure and function of proteins are immunohistochemistry, mutagenesis, and–with increasing importance–mass spectrometry [4]. Mass spectrometry proteomics is a relatively young but rapidly evolving field of research. According to the web of science database [174], the number of publications per year has increased from less than 50 papers in 1996 to over 3,000 in 2007 [97]. Mass spectrometry is capable of simultaneously monitoring hundreds or thousands of molecules. In recent years, mass spectrometry imaging (MSI) (also imaging mass spectrometry) [34, 142, 206, 46], a descendant that additionally performs a detailed analysis of the spatial distributions of (bio-)molecules, has evolved into a promising technology. However, the enormous size of datasets acquired with state-of-the-art instrumentation–often up to multiple gigabytes [67]–renders a direct manual analysis infeasible and makes mass spectrometry (imaging) proteomics highly dependent on bioinformatics [66].

Currently, we observe an increasing demand for novel computational methods [66, 32] that consider the particular properties of MSI data and that are tailored to the special needs of the field. This thesis develops and deploys novel algorithms for the automated analysis of MSI data (cf. figure 1.2):

- In chapter 3, we propose the application of *probabilistic latent semantic analysis (pLSA)* for non-negative decomposition of MSI datasets in exploratory settings. In contrast to principal component analysis (PCA), which is the de-facto standard in the field [214, 234, 219], it enforces a non-negativity constraint for the elucidation of interpretable component spectra and abundance maps. We further combine pLSA decomposition with a statistical complexity estimation scheme based on

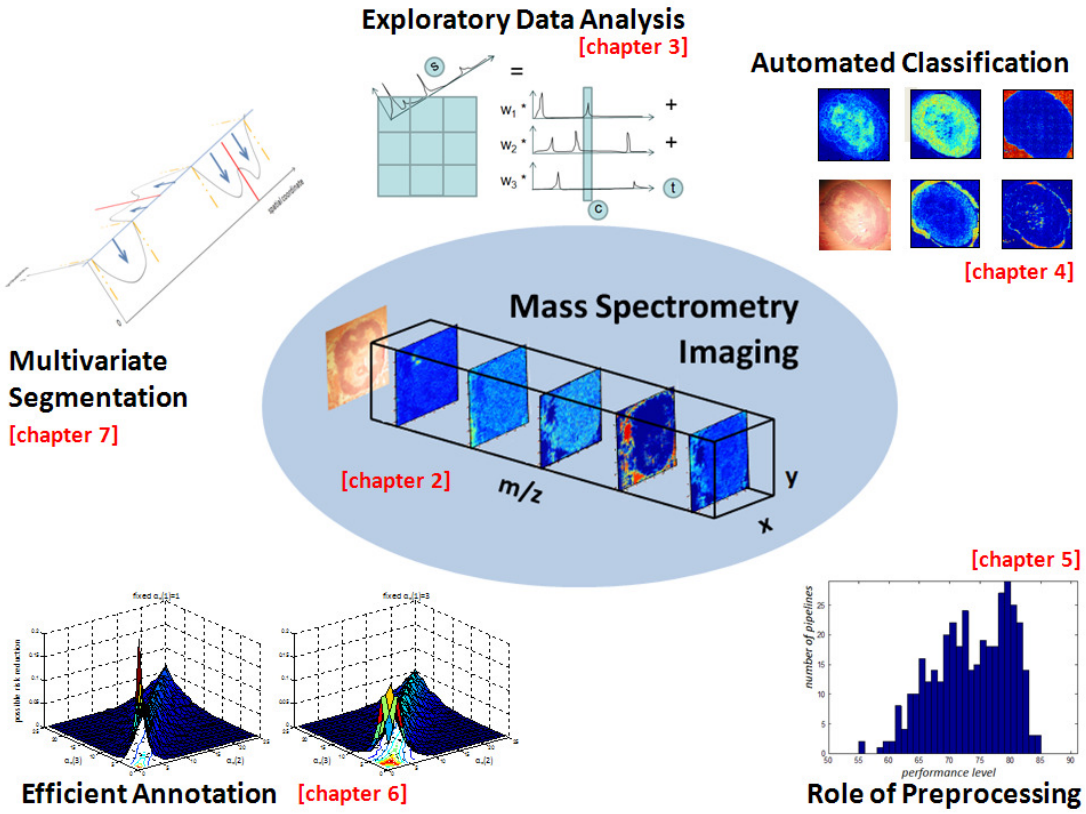

Figure 1.2.: Overview of the topics discussed in this thesis.

the Akaike information criterion (AIC) to automatically estimate the number of components present in a tissue sample dataset.

- In chapter 4, we demonstrate that "digital staining" by means of MSI constitutes a promising complement to chemical staining techniques. We suggest the random forest classifier for nonlinear *automated tissue classification*, which yields high sensitivities and positive predictive values–even when technical variability is present in the data. To reduce noise effects and further improve the classification, Markov random fields smoothing and vector-valued median filtering are employed.

- In chapter 5, we analyze the influence of the choice and parameterization of the algorithms used for *preprocessing* the acquired spectra on the outcome of automated classification approaches. We furthermore discuss how high technical and biological variability between measurements may complicate the analysis of state-of-the-art MS images.

- Training of supervised classifiers requires labeled example data. However, labeling MSI data is both costly and time-consuming, especially if the tissue is heterogeneous or the classifier has to be trained anew on each newly acquired set. Chapter 6 introduces a novel active learning scheme for multi-class problems. In comparison to random sampling, our method reduces the number of required expert labels without compromising on classification accuracy. It is thus suitable for the *efficient annotation* of MSI data.

- In chapter 7, we propose the multivariate watershed, a multivariate generalization of the classic watershed transform, as well as three novel boundary indicators for *multivariate segmentation* of compositional data. Such data arises, for instance, when classifying MSI data with random forests (chapter 4). We show on both synthetic and real world MSI data, that our segmentation methods compete well with existing approaches and are superior in some scenarios.

If not stated otherwise, the mass spectrometry images analyzed in this thesis have been acquired and kindly made available by our collaboration partners:

- Prof. Dr. Ron M. A. Heeren's group, FOM-AMOLF, FOM-Institute for Atomic and Molecular Physics, Amsterdam, The Netherlands (data used in chapters 3, 4, 6, 7).

- Dr. Kristine Glunde's group, Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, Baltimore, USA (data used in chapters 3, 4, 6, 7).

- Prof. Dr. Axel Walch's group, Institute of Pathology, Helmholtz Zentrum München, Germany (data used in chapter 5).

Chapter 2

Introduction to Mass Spectrometry Imaging

In the following chapter we shortly introduce basic biological and chemical background that is instrumental in understanding mass spectrometry proteomics. We furthermore review the basic principles of mass spectrometry and mass spectrometry imaging and describe the properties of the raw data that serves as input for further analysis.

2.1. Proteins, Peptides, and Amino Acids

Proteins have been described as the executive molecules in the cells [66]. They are composed of amino acids, which are joined by peptide bounds, arranged in a linear chain, and folded into a globular form. The genetic code defines 20 standard amino acids. Proteins with short amino acid chains (that is less than 50–100 amino acids [17]) are called *peptides*. Five elements make up the vast majority of all natural elements incorporated into proteins, namely hydrogen (H), carbon (C), nitrogen (N), oxygen (O), and sulfur (S). For each element, a set of isotopes with differing mass and abundance exist, which are mostly unstable. The most common (stable) carbon isotope is ^{12}C with an abundance of 98.91% and a mass of 12.0000, followed by ^{13}C with 1.09% abundance and a mass of 13.0034 (cf. table 2.1) [66].

Proteins take part in all kind of processes that act within and between cells of an organism: Some serve as enzymes that catalyze biochemical reactions and are vital to metabolism, others control the expression level of the genome, some have structural or mechanical functions, and others play important roles in immune response, cell signaling and the cell cycle. Post-translational modifications (PTMs) like phosphorylations can heavily influence a protein's function. Whereas the genome is identical in all cells of an organism (although it plays an important role in activation), the proteome largely differs

element		abundance in %	mass
hydrogen	^{1}H	99.99	1.00783
	^{2}H	0.01	2.01410
carbon	^{12}C	98.91	12.0000
	^{13}C	0.01	13.0034
nitrogen	^{14}N	99.6	14.0031
	^{15}N	0.4	15.0001
oxygen	^{16}O	99.76	15.9949
	^{17}O	0.04	16.9991
	^{18}O	0.20	17.9992
sulfur	^{32}S	95.02	31.9721
	^{33}S	0.76	32.9715
	^{34}S	4.22	33.9676

Table 2.1.: Abundances and masses of stable isotopes for the five naturally most abundant elements [66].

between individual cells and can even vary over time [66].

2.2. Mass Spectrometry Proteomics and Mass Spectrometry Imaging

A mass spectrometer can be pictured as a scale for simultaneously weighing hundreds or thousand of molecules [218] to determine the elemental composition of a sample (see section 2.3 for a detailed description). Researchers have employed this powerful technique in many different contexts, ranging from proteomics [66] to quality control of oil [3], analysis of art and archaeological materials [48], monitoring of climate change [85], and even to the analysis of molecular ions in extraterrestrial space [166]. In proteomics, mass spectrometers are used to study the three-dimensional structure of proteins [198, 133], their relative and absolute abundance [155, 226, 119, 125], their interactions with other proteins [63, 45, 70], or their spatial distribution in an organism. Especially the latter research question is predominantly investigated with *mass spectrometry imaging* (MSI) [34, 142, 206, 10, 46, 223][1] which emerged in 1999, when matrix assisted laser desorption/ionization mass spectrometry imaging (MALDI MSI, see section 2.3.1) was proposed [43]. Since then, MSI has evolved into a promising technology that has established itself in many fields of application, especially in medical research.

[1] Mass spectrometry imaging (MSI) is also known as imaging mass spectrometry. However, the corresponding abbreviation IMS is ambiguous since it is also used for ion mobility spectrometry that was recently coupled with mass spectrometry [113].

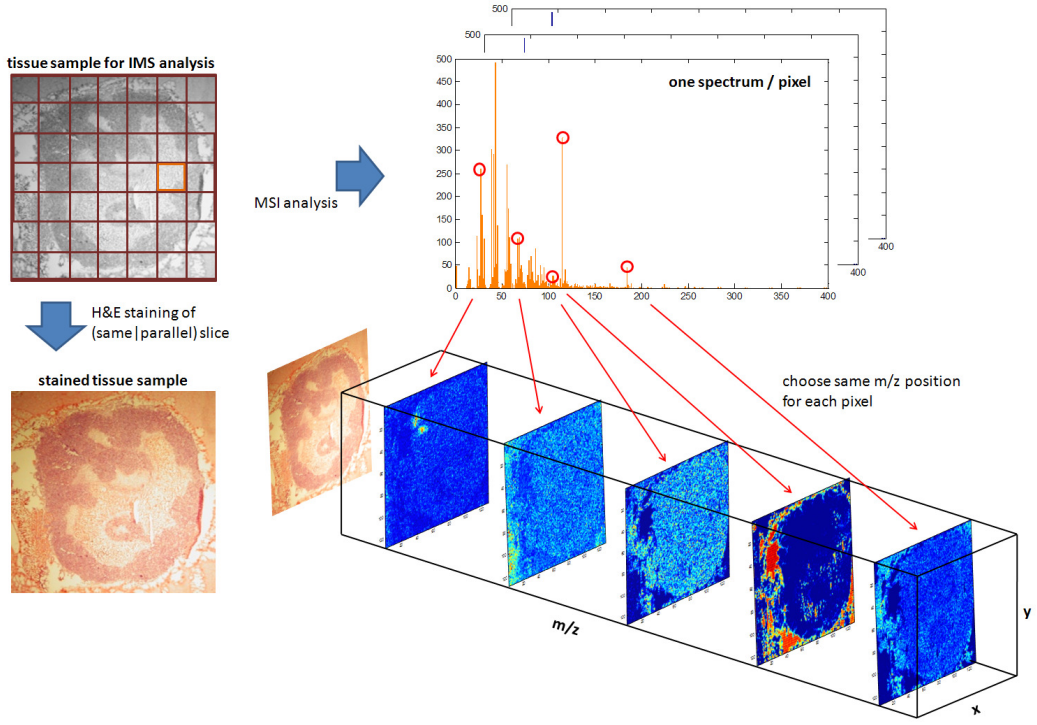

Figure 2.1.: The figure shows the workflow of a mass spectrometry imaging experiment [170]. First, a grid is superimposed on the tissue sample, and at each pixel location a mass spectrum is acquired. This yields a 3-dimensional data cube with two spatial (x and y) and one spectral dimension (m/z). By fixing the m/z dimension, a 2-dimensional image is obtained for each mass spectral peak (see lower right). Labels are typically assigned based on chemical staining of the very same or a parallel tissue slice. Note that the grid in the upper left is only a coarse sketch. In reality, the grid is much more fine-grained.

Today, wet lab staining techniques are still the method of choice for visualizing the spatial distribution of specific biomolecules or cell compartments. However, the high specificity of a wet lab stain is also its principal limitation: one slice of tissue can be treated with a small number of stains only (for exceptions, see [216]) and cannot be reanalyzed at will when different biomolecules take center stage. MSI, in contrast, is not handicapped by these restrictions and combines the capabilities of mass spectrometry with microscopic imaging in a single experiment [142, 135]. Without requiring labels, it permits a detailed analysis of spatial distributions of (bio-)molecules (such as proteins, peptides, lipids, or metabolites [42]), while simultaneously screening a large spectral range covering hundreds or thousands of different molecules [138].

With this richness of information at hand, classification of tissue types (e.g., healthy vs. tumorous or different tumors) may become more reliable. Influential mass spectral peaks or groups of mass spectral peaks with their corresponding molecules may in turn be identified as *biomarkers*. The strength of the information richness is also a major burden–automated and reliable computational analysis of MSI data becomes indispensable, giving rise to an increasing demand for novel algorithms in bioinformatics.

Current methods for the computational analysis of mass spectrometry images fall into two categories. If no or little prior knowledge on the composition of a sample is available (e.g., in exploratory data analysis), *unsupervised* methods can be applied to decompose the data into characteristic component spectra and corresponding abundance maps, visualizing spectral and spatial structure (see chapter 3 and figure 2.2). In cases where prior knowledge exists, for instance in form of spatially resolved labels, *supervised* classifiers can be trained to discriminate different tissue types in an automated way (see chapters 4,5,6).

With the further development of mass spectrometry and mass spectrometry imaging, new challenges in bioinformatics arise continually. For instance, a recent development that requires tailored computational analysis is 3D mass spectrometry imaging [55, 11, 72]. The idea of 3D MSI is to aggregate MSI results from a set of tissue slices (which are cut in parallel to each other) to reconstruct a volume and visualize 3D distributions of peptides and proteins in an organism.

2.3. Principles of a Mass Spectrometer

A mass spectrometer consists of three integral parts: an ion source, a mass analyzer measuring the mass over charge ratio (m/z, in Dalton) of ionized analytes, and a detector that counts the number of ions for all m/z values under consideration (see figure 2.3) [4].

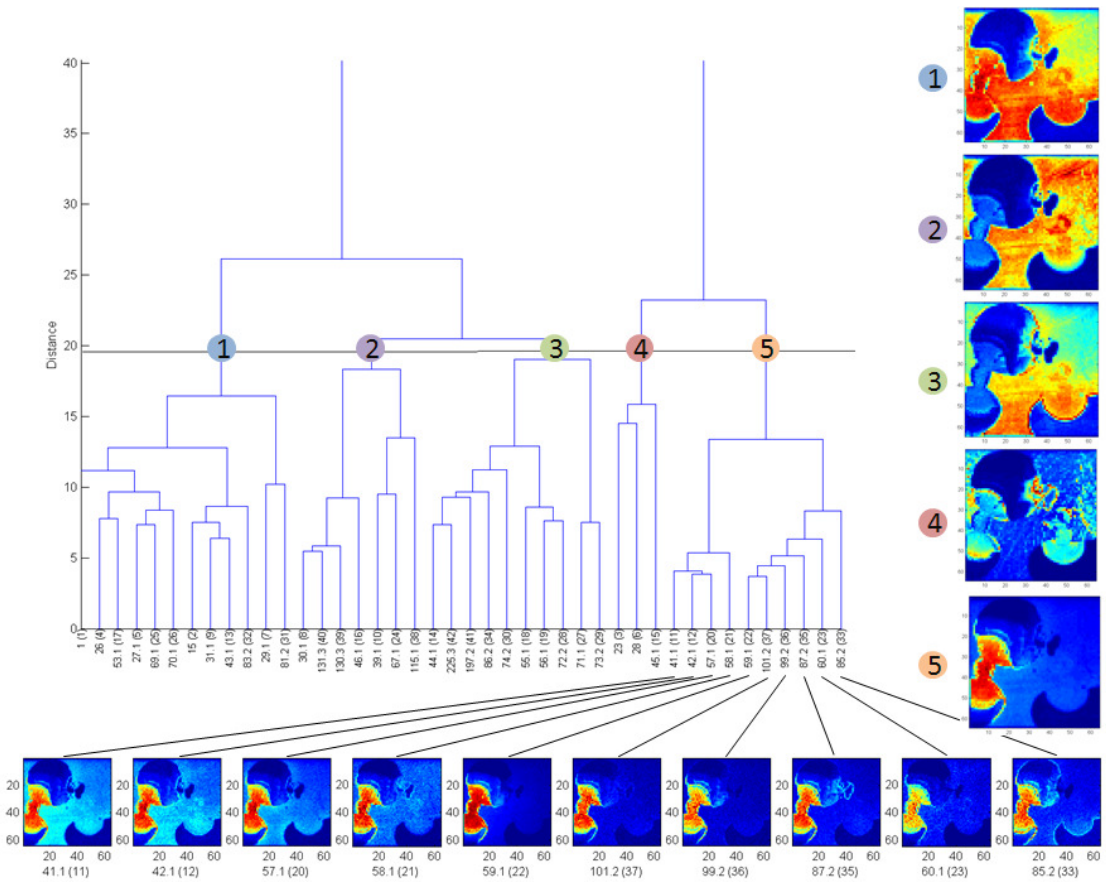

Figure 2.2.: Unsupervised analysis of MSI data: here we used hierarchical clustering [96, 57] to decompose a SIMS dataset of polymers (see [121] for a detailed data description) where we clustered molecules that exhibit similar spatial distributions with respect to a distance metric. The figure shows the dendrogram, the average abundance maps corresponding to the five most prominent clusters as well as the abundance maps of all molecules within cluster five.

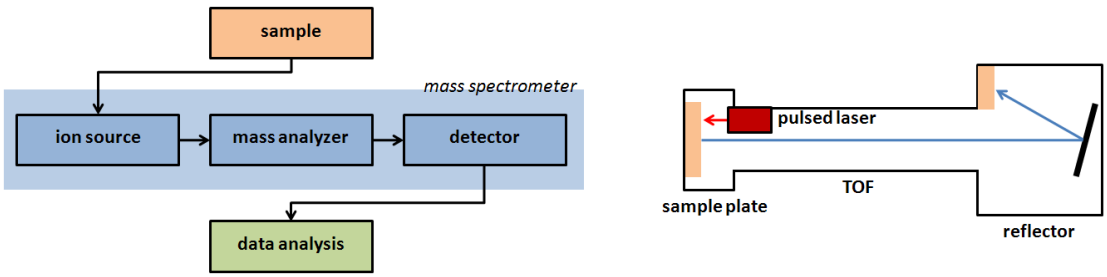

Figure 2.3.: On the left, the schematic setup of a mass spectrometry experiment is shown: The sample is analyzed with a mass spectrometer that consists of an ion source, a mass analyzer and a detector. Data acquisition is then followed by manual or automated analysis. On the right, a time of flight (TOF) instrument, one of the most common mass analyzers available, is shown. The mass of a molecule is determined by its time of flight: The heavier the molecule, the longer the time between laser pulse and time of arrival. By using a reflector, the time of travel and thus, the accuracy of the instrument is increased [4].

2.3.1. Ion Source

Electrospray ionization (ESI) [69][2], Surface enhanced laser desorption/ionization (SELDI) [106], matrix assisted laser desorption/ionization (MALDI) [99, 114], and secondary ion mass spectrometry (SIMS) [131, 37] are the most popular ionization techniques. The latter two are the most common ones in mass spectrometry imaging. Whereas ESI ionizes the analytes out of a solution making it unsuitable for MSI, MALDI ionizes the samples out of a dry, crystalline matrix using laser pulses. MALDI is usually used for relatively simple mixtures [4] but can be employed for analyzing extremely large molecules up to 200,000 Da [218, 54] and has thus been established as "the method of choice for recording intact protein distributions" [142]. In contrast to MALDI, which is laser based, SIMS uses a highly focused, continuous beam of primary ions to bombard the analyte and release molecules and *secondary* ions from the surface [142]. Both, MALDI and SIMS offer unique advantages and, in a way, are complementary: While MALDI is suitable for analyzing high mass molecules, its spatial resolution is limited to approximately 25–200μm if the common microprobe mode (see section 2.3.3) is employed [98]. SIMS, in contrast, can be used to obtain MS images with extremely high spatial resolutions[3] but offers a very limited mass range (such as 0–1,000 or 0–500 Da), since the number of ion counts decreases rapidly with increasing mass [142]. Hybrid approaches like matrix-

[2] In 2002, the Nobel Prize in Chemistry was awarded to John B. Fenn for his work in mass spectrometry and the development of ESI.

[3] In theory, the primary ion beams used in SIMS can be focused to spot sizes as small as 50nm, but for this setting the number of obtainable ions per pixel is too low to result in useful molecular information [98].

enhanced SIMS (ME-SIMS) [231] have also been proposed. Recent work covers MSI based on ambient ionization methods that allow for the in situ examination of samples, that is outside the vacuum of the mass spectrometer [229, 138].

2.3.2. Mass Analyzer

A variety of different mass analyzers exist that can be used as standalone or in combination. The most commonly used instruments include the ion trap, time of flight (TOF) [204, 224], quadrupole [164], Fourier transform ion cyclotron resonance (FT-ICR) [141], and the Orbitrap [136, 104]. These approaches differ in sensitivity, resolution, mass accuracy, and their potential to generate tandem MS (also MS/MS or MS^2, see section 2.6) spectra from peptide fragments [4].

MALDI and SIMS are often combined with TOF mass analyzers yielding instruments with high sensitivity, excellent spectral resolution and high mass accuracy [142, 4]. In a TOF instrument, the ions are accelerated by an electric field of known strength and travel in vacuum along a path of known length (see figure 2.3). The time between ionization (e.g., with a pulsed laser) and arrival at the detector is measured. It holds, that the heavier the ion, the longer the travel time is. By identifying the potential energy of the charged particle with its kinetic energy, it can be derived that the time of flight of an ion is proportional to the square root of its mass over charge ratio (m/z) [204, 116]. By using TOF in reflection mode, the travel time of the ions is prolonged, and the mass accuracy of the instrument can be increased.

2.3.3. Detector

Most modern day mass spectrometers use microchannel plate detectors [64]. Many variants exist, which we do not describe in detail here. As an exception, we introduce the concept of position sensitive detectors [169]. The basic idea is to stabilize the ion cloud while it is traveling through, e.g., a TOF mass analyzer. By determining the relative position of each ion within the ion cloud, the spatial resolution of the resulting MSI dataset can be increased. For instance, the SIMS TRIFT II instrument that was used for acquiring several datasets that are analyzed in this thesis (cf. chapters 3, 4, 6) can be run in a so-called mosaic mode to acquire a series of mosaic tiles [7]. Typically, a single mosaic tile covers $150\mu m \times 150\mu m$ of sample area. By using a position sensitive detector, the resolution can be increased such that each mosaic tile is subdivided into 256×256 square pixels [4].

Two techniques for performing MALDI MSI experiments exist: the microprobe and the microscope mode [142, 120]. The former is the standard approach, which is method-

[4] However, in practice these high resolution datasets are often characterized by low ion counts per pixels such that "useful spatial resolution detection below $1\mu m$ is almost impossible" [when using SIMS with liquid metal sources] [71].

ologically simpler and features lower resolution. In short, a focused laser beam is used to ionize and analyze a localized region. The obtained mass spectrum is stored along with its corresponding spatial coordinates, and the laser spot is moved to the next acquisition spot. The spatial resolution of the resulting imaging dataset is directly related to the laser spot diameter since all spatial resolution that is higher than the laser spot size is lost [7]. In contrast, microscope MALDI MSI uses a stigmatic mass spectrometric microscope [134] with a position sensitive detector to retain the spatial resolution within the laser spot such that much higher lateral resolutions (below $1\mu m$) can be achieved. At the same time, microscope imaging is typically faster [122].

2.4. Sample Preparation and Matrix Deposition

Before an MS or MSI experiment can be performed, the sample has to be prepared for analysis. In MSI, standard sample preparation steps include a) tissue collection and embedding, b) tissue sectioning and mounting, c) washing and d) surface modifications [98]. If MALDI or ME-SIMS is employed as ionization technique, the last step includes the deposition of a matrix solution which improves the ionization capabilities of the molecules and which can be added manually or with spray robots [223]. Different matrix solutions exist, and the actual choice depends on the biological research question. Another (optional) surface modification step is enzymatic on tissue digestion [203], which decomposes a protein into a large number of peptide products with lower masses, making MS respectively MSI analysis feasible. A detailed overview of sample preparation protocols for SIMS and MALDI can be found in [8].

2.5. Preprocessing of Mass Spectra

The most common output of a mass spectrometer is a mass spectrum which essentially is a histogram over molecule weights in the analyte. Since ion counts cannot be negative, a mass spectrum is all positive[5]. Its spectral resolution and mass accuracy is determined by instrument settings and properties, and it may be affected by a varying amount of chemical and/or electronical noise and baseline effects. The x-axis of a mass spectrum reports mass over charge (m/z), usually measured in Dalton (Da), and the y-axis gives the corresponding abundances. However, these intensities are usually not suitable for absolute quantitation since their order of magnitude depends on many factors such as the acquisition time per spot or the cutting thickness of the sample. A mass spectrum should thus be considered a relative measurement only. Figure 2.4 depicts three examples of mass spectra from different instruments.

[5] At least in theory; due to measurement artifacts, it can sometimes happen that negative ion counts are reported.

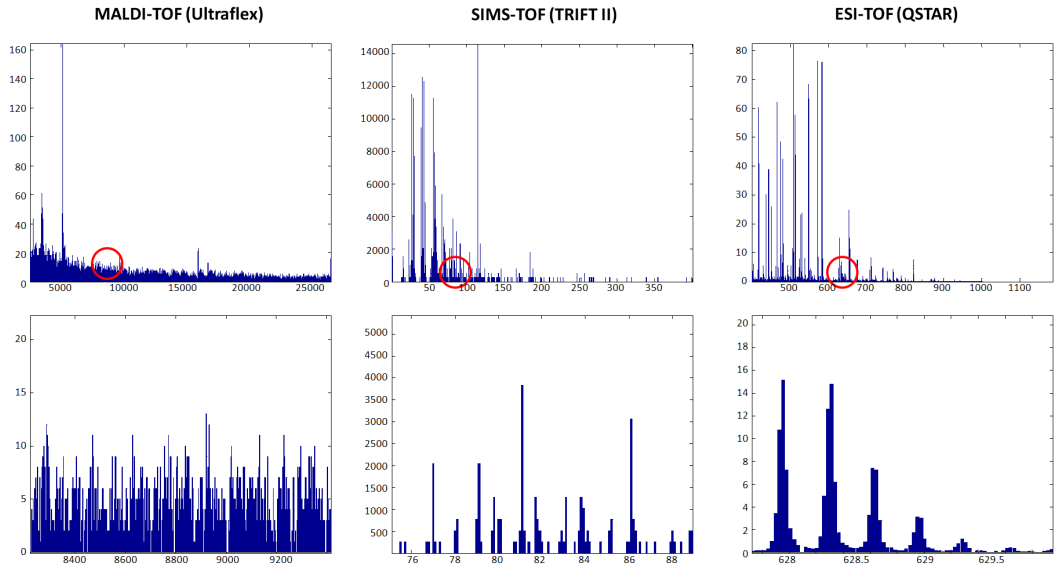

Figure 2.4.: This figure shows example mass spectra which have been acquired with three different mass spectrometers (see headings). In the top row, the full spectral range is displayed, and the bottom row shows the zoomed-in areas corresponding to the red circles. The MALDI and SIMS spectra stem from imaging datasets that are analyzed in this thesis, and the ESI spectrum is publicly available [173]. Whereas the former two do not show isotope patterns, the zoomed-in area of the ESI-TOF spectrum reveals the isotope distribution of a triply charged peptide, i.e., the individual peaks are spaced by approximately 1/3 Da. Also note that the MALDI spectrum features a very dominant baseline.

Many instrument vendors have developed proprietary file formats to store the raw output of MS or MSI experiments. Consequently, the data typically has to be converted to XML or another easily accessible data format before further processing steps can be applied. Several tools such as the spatial image composer (SIC) imaging software [122] are available.

2.5.1. Calibration, Baseline Correction, and Normalization

Standard preprocessing steps for MS and MSI data include spectral calibration, baseline correction and normalization. First, the measurement is calibrated by aligning it to several reference peaks. This can be achieved by adding well-studied reference compounds to the sample (internal calibration) or spotting known standards next to the sample onto the sample plate (external calibration) [66]. Baseline correction is applied to remove baseline effects (see figures 2.4 and 2.5) which predominantly originate from chemical

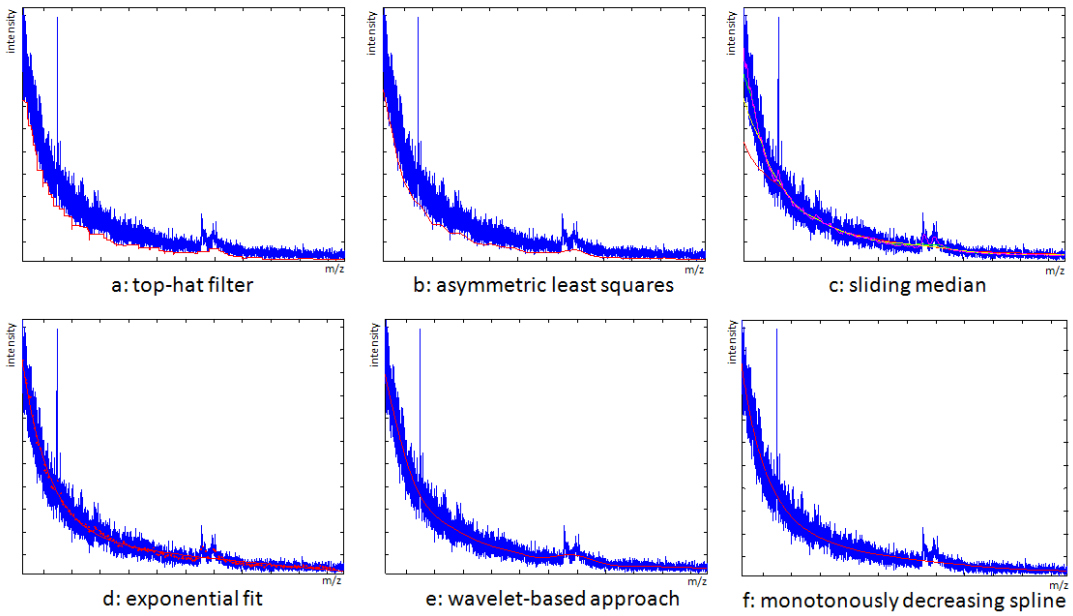

Figure 2.5.: The figure shows results from a selection of baseline correction approaches that have been applied to a MALDI-TOF example spectrum (blue, taken from the data set described in section 5.3.1). The estimated baselines are plotted in red: a) top-hat filtering [195, 189], b) asymmetric least-squares fitting by Eilers [68], c) median value within a sliding window (plotted for different window sizes), d) exponential fit to a preselected set of points (red dots) as proposed by Williams [227], e) wavelet-based removal of high frequency content (see [52] for related work), and f) spline fitting with a monotonicity constraint using the SLM toolbox [59]. Often, no ground truth information (i.e., the true baseline) is available, and the efficiency of a specific algorithm can only be judged by its potential to remove the overall trend while keeping the main peaks. No universally best method exists.

noise. Several different methods have been proposed [68, 67, 52, 227, 195]. Since mass spectral count data is usually not absolute quantitative, normalization is necessary to make different datasets comparable. The most common normalization techniques include normalization by total ion count (TIC) and normalization by the intensity of the base peak. The former method divides a measured spectrum by the sum of all measured ions within a spectrum and the latter divides the spectrum by the intensity of its most prominent peak, the "base peak". Again, many competing methods are available [146]. In cases where the spectrum is jagged, additional smoothing steps might be necessary [66].

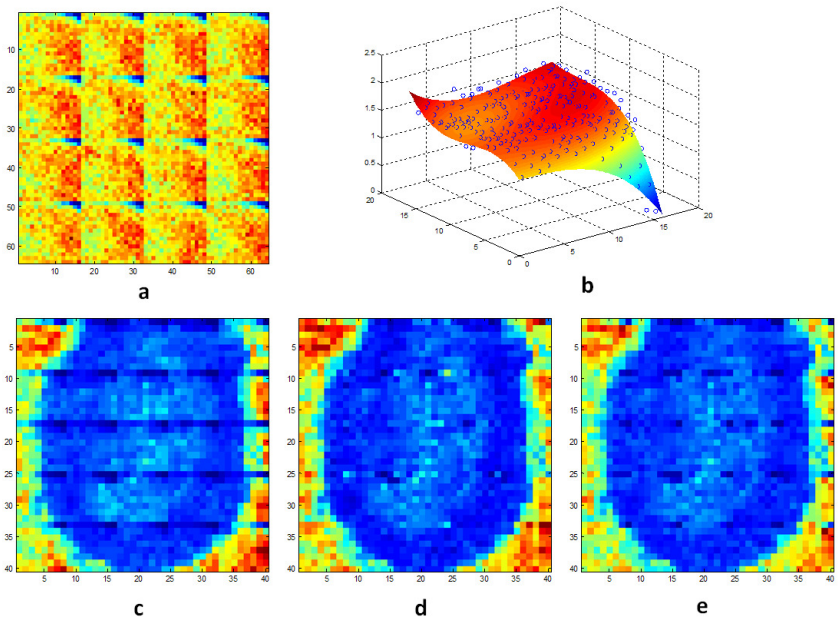

Figure 2.6.: Removal of detector artifacts: The total ion count (TIC) image of a homogeneous gelatin sample measured with SIMS (a) exhibits a clearly visible, repetitive pattern. In the following, we assume that the average intensities of the (16×16) positions within a mosaic tile are approximately the same when integrated over the whole image (that is all 16 tiles). We can then, e.g., attenuate the artifacts with the following two methods which are demonstrated on a SIMS dataset (with 8×8-sized tiles) of a pancreatic lesion showing a similar artifact (c): The TIC image in (d) was created by rescaling the measured spectra with the average intensity for each of the 8×8 pixel positions. The result in (e) was obtained by fitting a polynomial model (b) to the observed data. Both methods can be applied to the TIC image as well as to individual channels. In the latter case, the fitting to a polynomial model can be made robust by regularizing along the spectral dimension.

2.5.2. Detector Artifacts

Detector artifacts can occur if, e.g., the mass spectrometry image is acquired in mosaic mode (see section 2.3.3). Inhomogeneous detectors or ion beams can lead to repetitive patterns of intensity differences on the individual mosaic tiles. Figure 2.6a shows the total ion count image of a homogeneous gelatin sample acquired in a SIMS experiment. The intensity differences for the individual mosaic pixels are clearly visible. Similar patterns can be observed when imaging tissue sections (see figure 2.6c). A number of different methods can be applied to attenuate this effect.

2.5.3. Peak Picking

The relevant information of a mass spectrum is typically contained in its mass spectral peaks which can, e.g., correspond to proteins, peptides or lipids. However, the observed peaks can also be caused by matrix effects, electronic and chemical noise, or artifacts introduced in sample preparation. Furthermore, the mass spectral peaks are not sticks as the theory suggests but appear with a Gaussian, Lorentzian or other peak shape caused by the point spread function (PSF, also: peak shape function) of the instrument.

The information contained in a mass spectrum usually exhibits a high degree of redundancy. Depending on the employed ionization technique, ions corresponding to the same peptide can occur with different charge states and thus are contained in the mass spectrum multiple times. Given sufficient spectral resolution, we observe that the mass spectral peaks (e.g., corresponding to peptides) actually are groups of peaks, which form so-called isotope patterns. These patterns originate from the different natural occurrences of isotopes (as described in section 2.1). As a consequence, peptides and proteins occur with slight variations in mass, depending on their actual composition. These variations are reflected in the isotope pattern, which can be modeled by a multinomial distribution [36]. For low masses, the pattern has more mass on the left, but with increasing mass it converges to a Gaussian shape. The elements in table 2.1 all have masses that are close to integer numbers and, consequently, the spacing between the individual peaks is approximately 1.003 Dalton for singly charged and half a Dalton for doubly charged molecules (and so on). The monoisotopic mass of a molecule corresponds to its theoretical mass that can be calculated from restricting to the most abundant isotope of each element [66].

Since the mass spectral peaks hold the relevant information, data analysis of high resolution spectra typically relies on a feature extraction step that picks the monoisotopic peaks in a spectrum. In case of complex samples where the individual peaks show intermediate or high overlap, powerful peak picking algorithms are needed that can deconvolve the signal and reconstruct the individual contributions to a peak. Currently, NITPICK [173] and THRASH [102] are among the best-performing methods. Possible ways to reduce the complexity of the data and thus simplify the peak picking task are the use of gas chromatography (GC), liquid chromatography (LC) or ion mobility to separate compounds before applying mass spectrometry. For instance, in LC/MS the individual peptides elude at different times, allowing even to discriminate peptides that share the same mass over charge rate. An extension of NITPICK for LC/MS data has been proposed [22].

Many datasets that we analyze in this thesis, however, do either not feature sufficient spectral resolution or monitor a mass range in which no peptides can be observed (e.g., below 500 Da), such that NITPICK could not be applied. Consequently, we restrict to local maximum detection based peak pickers [66] to extract the relevant information.

2.6. Protein Identification

Methods for the identification of proteins subdivide into two categories: In the *top-down* approach, intact proteins are ionized with MALDI or ESI and subsequently analyzed with mass spectrometry. In contrast to *bottom-up* approaches, which form the second group of methods, no enzymatic digestion with proteases such as trypsin or pepsin is used to break down the protein into a peptide product prior to ionization.

The identification of proteins in complex samples is generally based on MS^2 (also tandem MS or MS/MS) spectra since the MS^1 alone cannot distinguish two peptides sharing the same amino acids (and thus same mass), but in differing orders. In contrast, in an MS^2 run, one of the many peptides entering the mass spectrometer is selected using a first mass analyzer, stabilized and fractionated yielding so-called parent ions. These ions are inducted into another mass analyzer. When fractionated, the peptide only breaks at well-known locations of the peptide bonds resulting in fragment ions. This knowledge can be exploited, and the characteristic patterns in the MS^2 spectrum can, e.g., be used to discriminate peptides sharing the same mass [66]. Two common methods exist for generating peptide interpretation: Sequence database searches use the characteristic fragment patterns in the MS^2 spectra of the peptide product to query a sequence database for candidates with similar structures. Commonly used search algorithms include Mascot, Sequest and ProteinPilot [172]. Note that database searches can only be applied if a database for the species under inspection exists, that is the species has been sequenced. In a subsequent step, the identified peptides are used to infer the proteins most likely to have resulted in these sequences. An alternative is *de novo* sequencing [75], which directly uses the characteristic fragment patterns to derive likely amino acids. Since both database searches as well as *de novo* sequencing have different benefits and drawbacks, hybrid approaches such as [144] have recently been proposed.

We note that first efforts for combining MS^2 with mass spectrometry imaging exist [135]. However, the analysis of such data is beyond the scope of this thesis.

2.7. Protein Quantitation

Another active field of research is protein quantitation where researchers study the differential protein expression in complex biological samples. Methods can be subdivided into label-free quantitation [237] and labeling approaches such as isotope labeling [226] and fluorescent labeling [149]. Due to the complex interplay of desorption and ionization in multi-component surfaces, local quantitation is not yet possible with MSI [98].

2.8. Analysis of Mass Spectrometry Images

The fundamentals of mass spectrometry and mass spectrometry imaging presented in this chapter are the starting point for MSI data analysis, which is the topic of this thesis. Many of the data properties discussed here need to be considered in the development and application of automated methods to obtain reliable and meaningful results. In the remaining chapters we introduce novel methods for the computational analysis of MSI data. We consider scenarios in which no prior knowledge on the composition of a sample is available (unsupervised analysis) as well as settings where spatially resolved labels exist (supervised analysis). The current state-of-the-art is reviewed at the beginning of the respective chapters.

Chapter 3

Concise Representation of Mass Spectrometry Images by Probabilistic Latent Semantic Analysis

3.1. Introduction

Many real-world applications of mass spectrometry imaging (MSI) require exploratory data analysis, where little or no prior information on the composition of a sample is available. This is particularly true when MSI is used as a discovery technique. In such an unsupervised setting it is useful to decompose the spectral image into a small number of characteristic component spectra and corresponding abundance maps to visualize the spectral and spatial structures in the data. These structures are not directly accessible, since manual inspection of individual m/z channel abundance maps is not only time-consuming but also tends to neglect information regarding the interaction of biomolecules. Nonetheless, low-dimensional representations that capture these correlations and make fast and efficient analyses of huge datasets possible are desirable.

Conventional techniques such as principal component analysis (PCA) [96, 219] and independent component analysis (ICA) [139] have successfully been applied in such settings, but they suffer from a number of drawbacks. The component spectra found by PCA are mutually orthogonal and hence feature negative counts. This implies that PCA cannot recover the true mass spectra of the tissue components. A rotation of the coordinate system as performed by VARIMAX-enhanced PCA [29] does not solve this problem. ICA suffers from similar non-negativity problems since its objective function tries to minimize the mutual information of the reconstructed components rather than making use

of the prior knowledge, that mass spectra are positive. Although physical interpretation of negative ion abundance rates is difficult, PCA is still widely used [232, 214, 219, 109]. Broersen [29] considered this problem and applied a non-negative PARAFAC (here abbreviated NN-PARAFAC) to MSI data, Smentkowski [201] applied a multivariate curve resolution method using constrained least squares algorithms. Mathematically, all these techniques perform a bilinear factor analysis, though based on different constraints and objective functions. An inherent problem for all those methods is that the number of components (here: tissue types) has to be specified–either prior to the decomposition process or in a post-processing step that selects the number of components heuristically, e.g., based on the percentage of variance represented by a given number of components. We therefore propose probabilistic latent semantic analysis (pLSA) for non-negative decomposition and the elucidation of interpretable component spectra and abundance maps and combine it with a statistical model selection scheme based on the Akaike information criterion (AIC) [6] that allows for an automated estimation of the number of components in the dataset.

We next revisit PCA, ICA as well as NN-PARAFAC and address their shortcomings that motivate the application of pLSA. We then introduce the AICc-corrected pLSA. The performance of the presented techniques is evaluated on simulated and real-world data.

3.2. Materials and Methods

3.2.1. Principal Component Analysis

Principal component analysis (PCA) [96] is a well-known and widely used technique for unsupervised data analysis and dimensionality reduction. PCA essentially performs a linear orthogonal transformation of the data domain. Let $x_l \in \mathbb{R}^{|C|}, l = 1, ..., |S|$ be a set of $|S|$ observed spectra each comprising $|C|$ m/z channels and $X = (x_1, ..., x_{|S|})$ the data matrix where each column holds an observed spectrum x_l. Further define $\tilde{X}$ as its corresponding mean centered version. PCA finds the principal components by diagonalizing the estimated data covariance matrix $\tilde{X}\tilde{X}'$ yielding the principal components (i.e., axes of the new coordinate system) ordered by decreasing non-negative eigenvalues (i.e., the observed variance along the PC axes). PCA projects onto the first k principal components keeping the linear subspace with the largest variance and yielding the best linear k-rank approximation to the data in the least-squares sense. In most applications, one is only interested in the first few components, assuming they hold the most relevant information. Usually the total percentage of variance retained after projection is used as an indicator of how many components should be used. The rationale behind this approach is that in many cases high variance along a direction will allow for good class separation. However, this assumption need not hold.

Mathematically, PCA can also be formulated as a factor analysis model, decomposing

the data into two matrices $\tilde{X}' = AB$ where $A'A$ is diagonal and $B'B = I$ [27]. Alternatively, PCA also follows from the singular value decomposition (SVD) [96] of $\tilde{X}'$, i.e., $\tilde{X}' = UDV'$ where U and V are orthogonal, D is diagonal, the columns of UD are the principal components (or "scores"), and the columns of V are the "loadings".

3.2.2. Independent Component Analysis

Independent component analysis (ICA) is a factor analysis model that was introduced in the context of blind-source-separation [96]. The key assumption of ICA is that the source signals, that is, in our case, the characteristic component spectra, are statistically independent with a non-Gaussian distribution. Typical preprocessing steps include centering and whitening, which lead to zero mean, unit variance, and zero correlation [96] as well as dimensionality reduction. The latter two are often solved with PCA [107, 80]. Unlike PCA, which is restricted to second degree cross-moments, ICA uses all cross-moments.

Let $Y = (y_1, ..., y_{|T|})$ be the matrix of independent components which are represented by the set T. The task is to determine the mixing matrix $A = W^{-1}$ and find

$$Y = W\tilde{X} \tag{3.1}$$

such that the components y_k become maximally independent. This can, e.g., be achieved by minimizing mutual information, which is a measure of the mutual dependence of two random variables [96], or maximizing non-Gaussianity [107].

3.2.3. Non-Negative PARAFAC

PARAFAC (PARAllel FACtors analysis) [95], also known as CANDECOMP (CANonical DECOMPosition) [35], is a multi-way decomposition method. As indicated by Kiers [118] it can also be seen as a constrained two-way PCA-model. In conjunction with non-negativity constraints for the modes this essentially corresponds to a standard non-negative matrix factorization [47]. PARAFAC does not approximate probability densities by marginals and lacks a proper probabilistic foundation. The solution is normally found by alternating least squares [28, 29], minimizing the squared reconstruction error, implicitly assuming Gaussian noise.

3.2.4. Probabilistic Latent Semantic Analysis

Probabilistic latent semantic analysis (pLSA) has been proposed in the realm of automated text analysis [100] and has become a standard method for structure and similarity identification in semantic document analysis. It can be described as a linear model with one latent variable t:

$$p(s, c) = \sum_{t \in T} p(t)p(s|t)p(c|t) \tag{3.2}$$

where s is a document, c a word and t a topic, the hidden variable. In the case of MSI, a spectrum can be considered a document, an m/z channel corresponds to a word, and a given tissue type corresponds to a topic. In the proposed model, each single tissue type is characterized by a distinct distribution, and each acquired spectrum is regarded as a specific mixture of these structures (cf. figure 3.1). The decomposition problem is solved by an expectation maximization (EM) procedure [100, 101] where expectation (E) and maximization (M) steps are alternated. In the E-step we assume the model parameters to be fixed and calculate the posterior probabilities for the latent variables t. This gives

$$p(t|s,c) = \frac{p(t)p(s|t)p(c|t)}{\sum_{\tilde{t}\in T} p(\tilde{t})p(s|\tilde{t})p(c|\tilde{t})}, \tag{3.3}$$

the probability that a count in channel c of a particular spectrum s is explained by the factor corresponding to t. Based on the posterior probabilities calculated in the E-step, the model parameters are updated in the subsequent M-step where

$$p(c|t) = \frac{\sum_{s\in S} X(c,s)p(t|s,c)}{\sum_{s\in S}\sum_{\tilde{c}\in C} X(\tilde{c},s)p(t|s,\tilde{c})} \tag{3.4}$$

$$p(s|t) = \frac{\sum_{c\in C} X(c,s)p(t|s,c)}{\sum_{\tilde{s}\in S}\sum_{c\in C} X(c,\tilde{s})p(t|\tilde{s},c)} \tag{3.5}$$

$$p(t) = \frac{\sum_{s\in S}\sum_{c\in C} X(c,s)p(t|s,c)}{\sum_{s\in S}\sum_{c\in C} X(c,s)}. \tag{3.6}$$

Here, $X(c,s)$ is the abundance of channel c in spectrum s. This can be reformulated as an SVD-like decomposition [100] by

$$P = \widehat{U}\widehat{D}\widehat{V}' \tag{3.7}$$

where we define $\widehat{U} = p(s|t)$, $\widehat{V} = p(c|t)$ and $\widehat{D} = diag(p(t))$ and where the matrix P of joint probabilities corresponds to the spectra-wise normalized data matrix X. Matrix formulations for the update rules in the E- and M-steps allow for an efficient implementation [79]:

$$R = norm(R \odot (W' \cdot [X \oslash (W \cdot R + eps)])) \tag{3.8}$$

$$W = norm(W \odot ([X \oslash (W \cdot R + eps)] \cdot R')) \tag{3.9}$$

where R is a $T\times S$ matrix with $R(t,s) = p(t|s)$, W is a $C\times T$ matrix with $W(c,t) = p(c|t)$, $norm$ is the column-wise normalization operator, and $\odot$ and $\oslash$ are the element-wise multiplication and division operators. Usually, R and W are initialized at random.

pLSA provides a probability distribution over the spectral dimension for each tissue type as the resulting components are normalized and non-negative. Unlike PCA, ICA and NN-PARAFAC, it has a sound statistical foundation and defines a proper generative

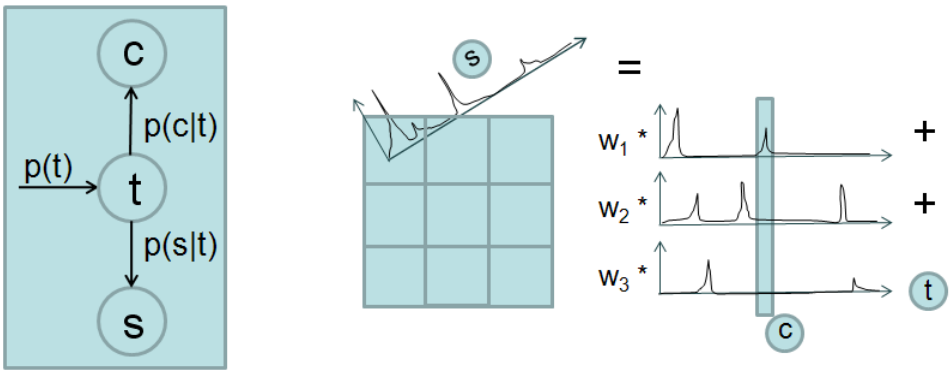

Figure 3.1.: On the left, the symmetric graphical model representation of pLSA is shown [100]. On the right, the decomposition principle is sketched: An observed spectrum s is considered a specific mixture (factors w_i) of the characteristic non-negative spectra for tissue types t comprising channels c.

model of the observed data [100][1]. This provides physical interpretability and allows to identify the discriminating peaks for a specific tissue type within a spectrum. pLSA is equivalent to non-negative matrix factorization with a Kullback-Leibler (KL) divergence measure [79] that is defined by

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{3.10}$$

and that quantifies the difference between two probability distributions P and Q.

Since peak intensities are ion counts, they follow a Poisson distribution. It can be shown that the joint distribution of a set of independent (not necessarily identically distributed) Poisson variables is a multinomial distribution conditional on their sum [235]. Indeed, pLSA assumes a likelihood function of multinomial sampling. Maximizing the predictive power of the pLSA model is equivalent to minimizing the KL divergence between the model and the empirical distribution [100]. It follows that for data like ours using the KL divergence is appropriate [16].

3.2.5. AICc-Controlled pLSA

A common feature of decomposition methods like NN-PARAFAC and pLSA is that the number of components $k = |T|$ has to be specified. This property is beneficial if such prior knowledge is available. For scenarios where this is not the case, we propose a method that is capable of automatically estimating k by using a statistical complexity

[1] It is, however, usually not considered a generative model for unseen data [21].

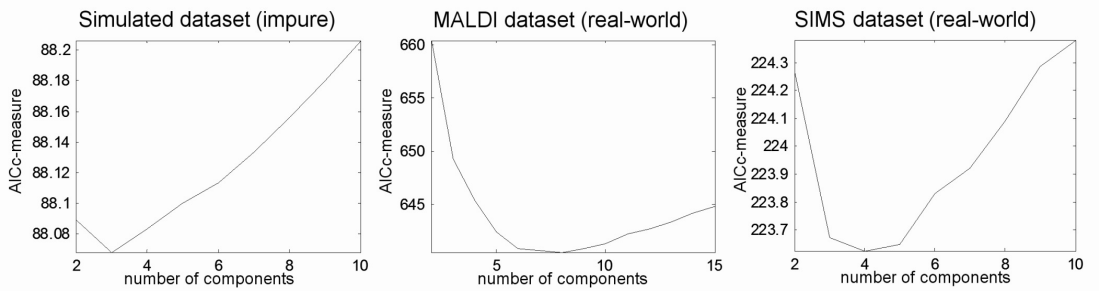

Figure 3.2.: The AICc curves for the datasets used in this study: The AICc criterion correctly identifies the number of mixture components for the simulated dataset (left) that is a mixture of three tissue types and results in reasonable estimates for the real-world data sets (middle, right).

estimation scheme. It is possible to select from a variety of different model selection criteria [96]. Compared to alternatives such as the Bayesian information criterion (BIC) or minimum description length (MDL), AIC-type criteria for feature selection tend to select slightly more features if the model specification does not reflect the true generating process of the data [96, 205]. It is therefore more "conservative" and hence preferable in our application in which one would rather estimate slightly too many components than loose subtle differences by penalizing model complexity too heavily.

Starting with $k = 2$, we run the pLSA-algorithm repeatedly with increasing k. We then apply a corrected Akaike information criterion (AICc) [6, 31] (which is based on the KL divergence) to automatically select the correct model. In our case, the AICc-type criterion is defined as

$$AICc(k) = \underbrace{-\frac{2}{N}\mathcal{L}(k)}_{data\ likelihood} + \underbrace{\frac{2}{N}M\sigma^2}_{penalty\ term} + \underbrace{\frac{1}{N}\frac{2M(M+1)}{N-M-1}}_{correction\ term} \tag{3.11}$$

where N is the number of observations ($|S| \cdot |C|$), $\mathcal{L}(k)$ is the data log-likelihood

$$\mathcal{L}(k) = \sum_{s \in S} \sum_{c \in C} X(c,s) \log \sum_{t=1}^{k} p(c|t)p(t|s), \tag{3.12}$$

and we rely on a second-order correction term for small sample sizes. The number of free model parameters $M = k \cdot (|S| + |C|)$ is equivalent to the number of elements in the two solution matrices representing $p(c|t)$ and $p(s|t)$. The first term in equation (3.11) measures how good a model fits to the observed data, the second term penalizes complexity to prevent overfitting, and the last term corrects the AIC for small sample sizes. To robustly estimate the noise variance σ^2 we calculate the median of the squared residuals

of the difference between the original spectra with the observed data and their spatially smoothed version (rectangular 3×3 mean filter on each m/z slice), assuming that neighboring tissue consists of similar tissue mixtures. The idea behind AICc-controlled pLSA is to stop the iterations as soon as we can be sure that decompositions with a higher number of components will not yield a lower AICc value than the current minimum. We calculate equation (3.11) for increasing k until a stopping criterion is met and then take the value of k for which the minimum of the resulting AICc-curve is attained as an estimate for the optimal model complexity (see figure 3.2). The early stopping criterion is defined as follows [173]. It holds that $\mathcal{L}(k) \leq 0 \ \forall k$ and thus that we can abort the calculations if the (strictly increasing) penalty term for the current k (cf. eq. (3.11)) is higher than any previous value in the AICc-curve. However, this bound is not very tight and therefore not practical. We therefore set a reasonable upper bound $\tilde{k}$ for the number of components, for example 100. Since $-2\mathcal{L}(k)/N$ is monotonously decreasing, we can stop as soon as

$$-\frac{2}{N}\mathcal{L}(\tilde{k}) + \frac{2}{N}M\sigma^2 + \frac{1}{N}\frac{2M(M+1)}{N-M-1} > \min_{2 \leq k \leq k_{curr}} AICc(k) \tag{3.13}$$

is met for the current number of components k_{curr} (where $M = k_{curr} \cdot (|S| + |C|)$). With this criterion we are guaranteed to find the minimum of the AICc curve within the given bounds $[2; \tilde{k}]$ with low computational overhead.

3.2.6. Sparsity

After decomposing the data it is often of interest to identify decisive peaks that allow for a discrimination of different tissue types. We call a peak decisive if it is only present in one or few of the k component spectra. Candidate peaks are identified with Hoyer's sparsity measure [103] which is defined as

$$sparsity(u) = \frac{\sqrt{k} - (\sum |u_i|)/\sqrt{\sum u_i^2}}{\sqrt{k} - 1} \tag{3.14}$$

where u' is a $1 \times k$ row vector of the matrix that holds $p(c|t)$. The measure is based on the ratio between $\sum |u_i|$ (the L_1-norm) and $\sqrt{\sum u_i^2}$ (the L_2-norm), assigning a high sparsity value to a vector u if the intensity distribution over the components u_i is sparse and the respective channel can therefore be used to discriminate between components or tissue types.

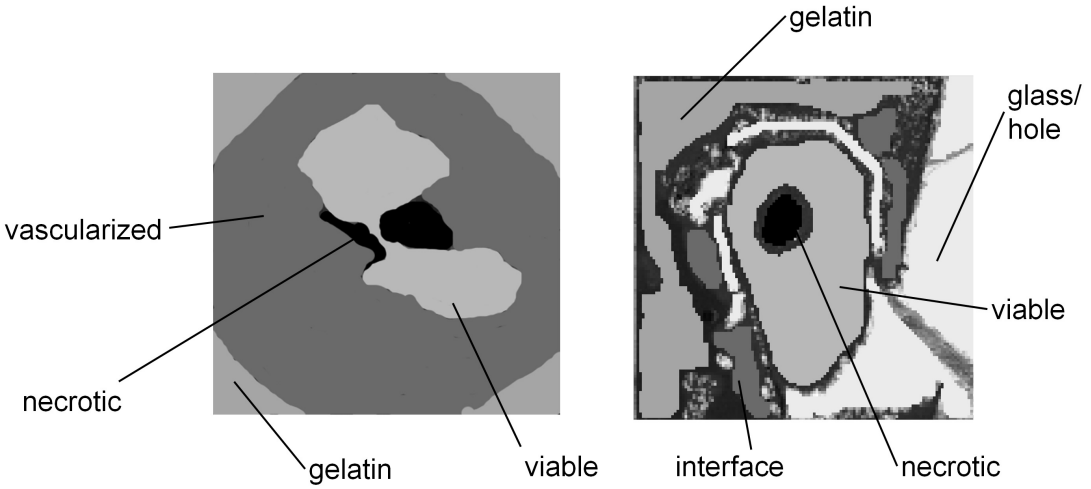

Figure 3.3.: Labeling of the matrix assisted laser desorption/ionization (MALDI) (left) and secondary ion mass spectrometry (SIMS) dataset (right) based on total ion count images and histologically stained parallel slices.

3.3. Experiments

3.3.1. Simulated Data

We first tested the algorithms on two simulated datasets which were generated in the following manner: Three regions of interest corresponding to different classes (as assigned by an expert) were selected from a real-world MSI dataset of breast cancer tissue (see figure 3.3). We treated the average spectra of those regions as characteristic spectra and mixed them according to the two mixture maps shown in figures 3.4 and 3.5 (left panel) yielding a "pure" and an "impure" ground truth dataset. Two of the characteristic spectra featured considerable spectral overlap which complicated the decomposition process. Furthermore, Poisson noise was added prior to decomposition.

3.3.2. Real-World Data

Two real-world datasets were used to evaluate the methods described above. These sets were acquired with different instrumentation and feature two kinds of orthotopic human breast cancer xenografts grown in mice. Prior to the application of mass spectrometry imaging, the tissue samples were embedded in gelatin, flash-frozen, cryo-sectioned to $\approx 10\mu m$ and thaw-mounted on a cold indium tin oxide-coated glass slide. Using the spatial image composer (SIC) tools by Klinkert et al. [122], the obtained MSI data was converted from the instrument vendors' raw formats to an accessible format and further

processed within MATLAB.

MALDI Dataset. The first dataset was acquired on a modified matrix assisted laser desorption/ionization (MALDI) TRIFT II instrument coupled with a time-of-flight (TOF) mass analyzer. For MALDI MSI, a 2.5-dihydroxybenzoic acid (DHB) 30mg/mL in 50% acetonitrile/0,1% trifluoroacetic acid was applied using an air driven thin liquid chromatography spray. After drying, the tissue was additionally sputter coated with 1nm gold. The spectral resolution of the resulting dataset was rebinned to 0.1 Da. One pixel spans $150 \times 150 \mu m$. The tissue contains MDA-MB-231, a highly metastatic breast cancer tumor.

SIMS Dataset. The second dataset was acquired as part of a large-scale study using secondary ion mass spectrometry (SIMS). Therefore, a Physical Electronics TRIFT II TOF SIMS that was equipped with an Au+ liquid metal ion gun and run in mosaic mode was employed. Due to the high number of tissues that had to be processed, short acquisition times of 2 seconds per spot were used. To guarantee a reasonable number of ion counts in each mass spectrum, the spatial resolution of the data had to be rebinned to $35 \times 35 \mu m$ [71]. The analysis was confined to a mass range of 0–2,000 Da. After data acquisition, the spectral resolution was rebinned to 0.1 Da, and the mass range between 0–400 Da was selected (resulting in 4,009 mass channels). The tissue features MCF7, a weakly metastatic and estrogen-sensitive breast cancer tumor.

Gold Standard Labels. For both datasets, a hematoxylin-eosin-(HE)-stained parallel slice is available. Despite some topological differences between the stained and MSI-subjected slices, the stained slices can be used as gold standards enabling further evaluation of the decomposition quality of the four methods (cf. figure 3.3). However, the label information was not used in the decomposition process.

3.3.3. Evaluation Criteria

Since two different tumor types have been used for MALDI and SIMS analysis, we refrain from comparing the resulting datasets against each other but rather concentrate on the relative performance of the decomposition techniques presented above. One of the main motivations for the application of decomposition methods like PCA or pLSA lies in their potential to achieve dimensionality reduction while keeping the relevant information. When PCA is applied in practical applications, often only a few principal components are considered since the analysis of hundreds of components is simply impractical. Thus, we decided to compare the described methods in a scenario where the number of components is limited to only a few. We based our evaluation on four criteria: reconstruction quality, complementarity of the resulting components, the ability to reconstruct major peaks, and visual inspection.

Reconstruction Accuracy. The first criterion measures how well the original data can be reconstructed after factorization favoring decompositions that explain the observed data with high accuracy. We calculated the reconstruction error between the

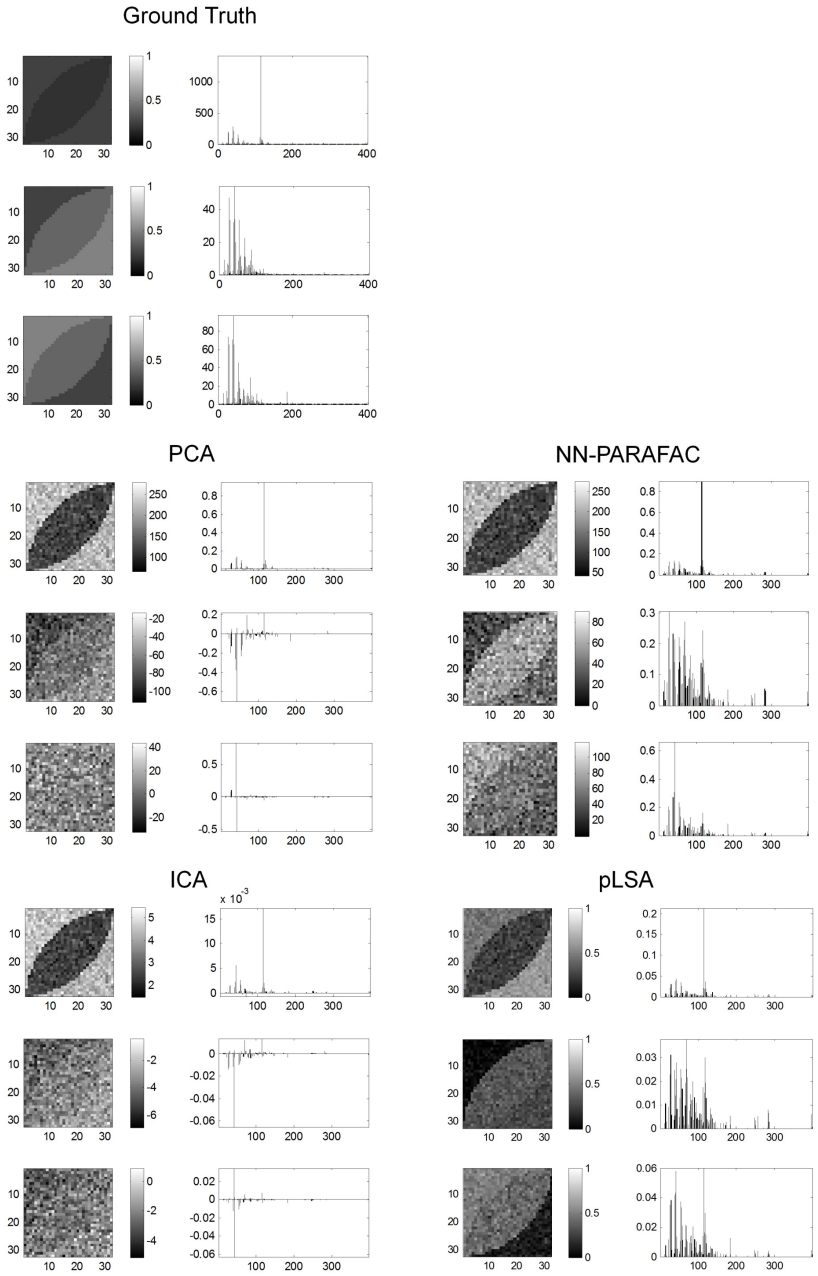

Figure 3.4.: Decomposition results on the simulated dataset with impure mixtures: The PCA component abundance maps are ordered from top to bottom according to the eigenvalue associated with the principal components. For ICA, NN-PARAFAC and pLSA the order of components is arbitrary and has been permuted such that the observed abundance maps coincide in their ordering with the ground truth. NN-PARAFAC and, especially, pLSA clearly outperform PCA and ICA (see also text and figure 3.5).

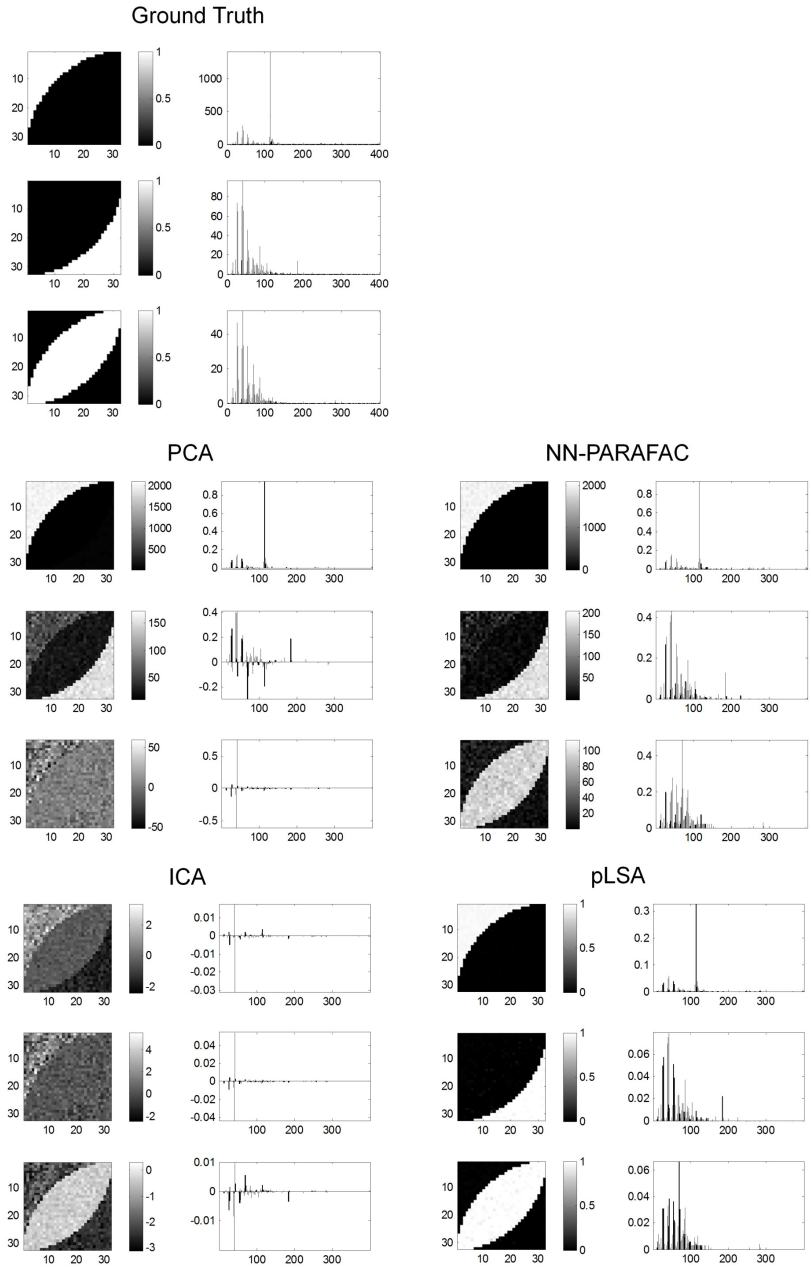

Figure 3.5.: Decomposition results on the simulated dataset with pure mixtures: As for the impure mixtures, NN-PARAFAC and pLSA clearly outperform PCA and ICA. Taking additional principal components into account was neither beneficial in the pure nor the impure mixture experiment since the higher components mainly contained noise.

observed and reconstructed spectra with respect to three different measures to avoid bias towards one of the methods under consideration: the L_1-norm, the L_2-norm and KL divergence. The latter is defined in equation (3.10) and the L_p-norm of a vector a is defined as

$$\|a\|_p = (\sum_{i=1}^{n} |a_i|^p)^{1/p}. \tag{3.15}$$

First, we created two vectors by concatenating all original (observed) spectra of a dataset and all reconstructed spectra. We then calculated the L_1- and L_2-norm of the difference and divided the result by the number of spectra in the dataset. Calculation of the KL divergence requires that the two concatenated spectra are probability distributions, that is are all-positive and sum up to one. However, non-negativity is not guaranteed in the case of PCA and ICA. Especially if only a few components are used in the reconstruction, it is possible that some channels of the reconstructed spectra exhibit negative values. We therefore set all channels in the reconstructed spectra that had negative intensities to zero (which was in favor of the PCA/ICA error estimates). Irrespective of the method used for the decomposition, we further added a small constant to all values in the two concatenated vectors to avoid division by zero and finally normalized the two vectors before applying equation (3.10).

If the number of components used for reconstruction is limited (as in many real-world scenarios), PCA and ICA can no longer perfectly reconstruct the data. The reconstruction errors obtained with the four methods under inspection are affected by both the magnitude of the reduction as well as the selected stopping criterion of the iterative methods (e.g., pLSA). Since we applied PCA for dimensionality reduction as a preprocessing step for ICA (see section 3.2.2), the estimates for those two methods were the same.

Complementarity of the Estimated Components. To simplify interpretation, it is often desirable to decompose the data into clearly distinguishable components. The second criterion therefore quantifies the complementarity of the resulting abundance maps. For a given number of components k and a given decomposition method, the complementarity was measured as follows: first, the k estimated abundance maps were thresholded at various quantiles between 95% and 50%. Then, we estimated the complementarity for each quantile by calculating the percentage of area covered after overlaying the k thresholded abundance maps, i.e., the percentage of pixels with an intensity value greater than zero (see figure 3.6). To allow for a fair comparison, for PCA and ICA we also calculated the corresponding lower quantiles in which we only retained the pixels with the most negative contributions. This was necessary since some structures were better reflected by areas with negative intensities. We then used the best out of the 2^k possible combinations of upper/lower thresholded abundance maps for each method and quantile. The higher the coverage index, the more complementary the analyzed

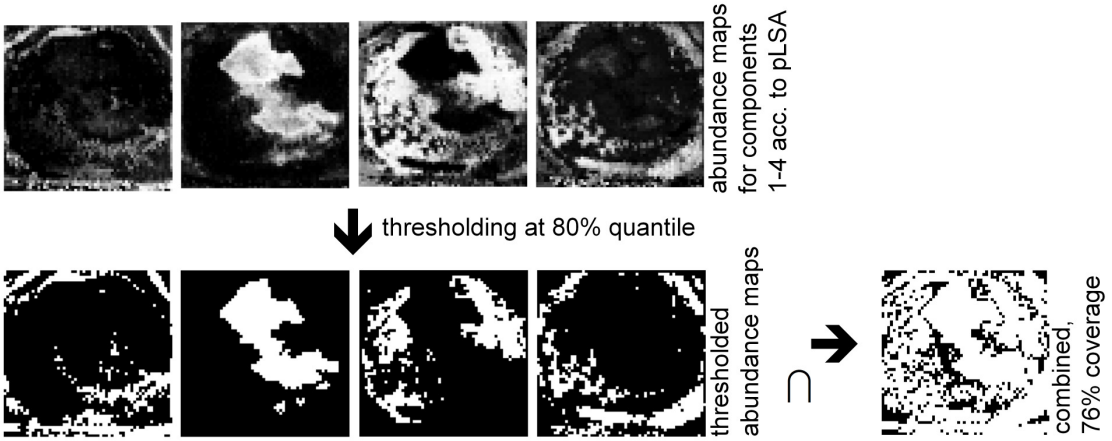

Figure 3.6.: Complementarity estimation example for the pLSA decomposition of the MALDI set at the 80% quantile: The abundance maps estimated by pLSA (top row) are thresholded by setting the top 20 percent of the most intense pixels to one and the remaining pixels to zero (bottom row). The percentage of white pixels in the combined image (boolean *OR*, right) is an indicator for the complementarity of the components (see also table 3.2).

components are.

Reconstruction of Major Peaks. The third criterion quantifies the ability of PCA, ICA, NN-PARAFAC, and pLSA to reconstruct the major peaks within the characteristic component spectra. Since the ground truth for the real-world datasets is not directly available, we restricted our comparison to the simulated data described above. We first calculated several quantile spectra (80%, 85%, 90%, 95%) of the three (known) ground truth spectra in which we only retained the most intense peaks.[2] Then, the corresponding quantile spectra were calculated for each decomposition method and each of the three reconstructed components. For a given quantile, we compared the reduced (quantile) ground truth spectrum to the corresponding reduced characteristic spectrum as estimated with one of the decomposition methods. The overlap of the peak positions contained in the two quantile spectra is a measure of how well the major peaks were reconstructed by a decomposition method. A value of 1.00 is only reached if the positions of the most intense peaks in the ground truth spectrum, and the reconstructed spectrum are completely identical.

Since the sign of the principal and independent components is arbitrary, in case of PCA and ICA we calculated the upper quantiles as well as the lower quantiles. The

[2]For instance, the 95% quantile spectrum only contains the top five percent of the major peaks in a spectrum.

latter were obtained by first inverting the sign of the components and then extracting the major peaks as described above. We then used the quantile spectrum (upper/lower) that showed more overlap with the respective ground truth spectrum.

Visual Inspection. In addition to the three quantitative measures we compared the decomposition methods by visual inspection of the resulting components.

3.3.4. Data Processing

Prior to decomposition, the datasets were preprocessed as follows: Baseline-correction was performed by channel-wise subtraction of the minimum (with respect to all measured spectra in a dataset) and spectra were normalized by their total ion count. Features were extracted with an in-house implementation of a threshold-based peak picker (see section 4.3.4 for details) to keep the relevant information and simultaneously decrease computation time. For ICA and NN-PARAFAC calculations, the freely available MATLAB toolboxes FastICA [80] and N-way [28] were used. We defined a relative change in the fit of 10^{-6} as stopping criterion for both NN-PARAFAC and pLSA.

3.4. Results

We compared the proposed methods on simulated data and confirmed the simulation results on real-world data. To minimize random effects, all non-deterministic methods were restarted five times and for each method, the best result was checked.

The simulated data was created from three tissue types, and we performed the decompositions with the respective number of components. For the (real-world) MALDI set we expected four different regions in the sample: viable or active tumor, necrotic tissue, vascularized region and embedding gelatin. Based on this prior knowledge (and not in favor of one of the methods), we performed the decompositions with four components for which PCA kept 91% of the total variance. Additionally, decompositions with 8 components and varying numbers of components were calculated. For the (real-world) SIMS set we were interested in the three tissue types and the gelatin region described above, but we furthermore assumed a fifth area since parts of the tissue were torn prior to analysis. As can be seen from the label map in figure 3.3, in those areas the indium tin oxide-coated glass slide is exposed and thus, we expected an indium peak at 115 Da.

The computation time needed for NN-PARAFAC and pLSA is higher than for PCA and ICA. Whereas the PCA and ICA decompositions were completed in less than one second for the SIMS-set, pLSA required ≈30–60 seconds and NN-PARAFAC up to 240 seconds. The AICc-type-enhanced pLSA needs several passings since decompositions with an increasing number of components have to be calculated. However, often only a few iterations (5–15) are necessary, and the calculations finish in reasonable time. We indicate that these measures are highly dependent on the parameter settings like the

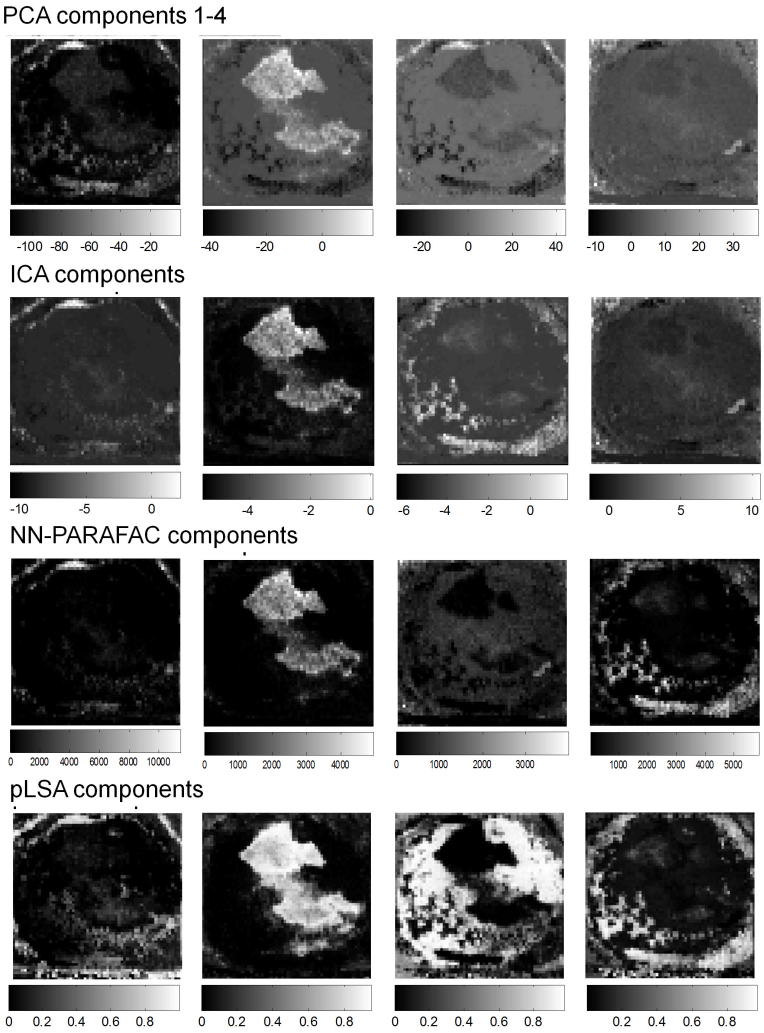

Figure 3.7.: Decomposition of the MALDI set with four components. Again, the ICA, NN-PARAFAC and pLSA components have been reordered to match the PCA components for which the ordering is unique. We further inverted some of the PCA and ICA components for better visual comparison. Only NN-PARAFAC and pLSA have entirely non-negative abundance maps, and only pLSA components are normalized and hence interpretable as tissue probability by definition. The more black-and-white a suite of components, the better their complementarity, cf. table 3.2. Refer to the text for interpretations.

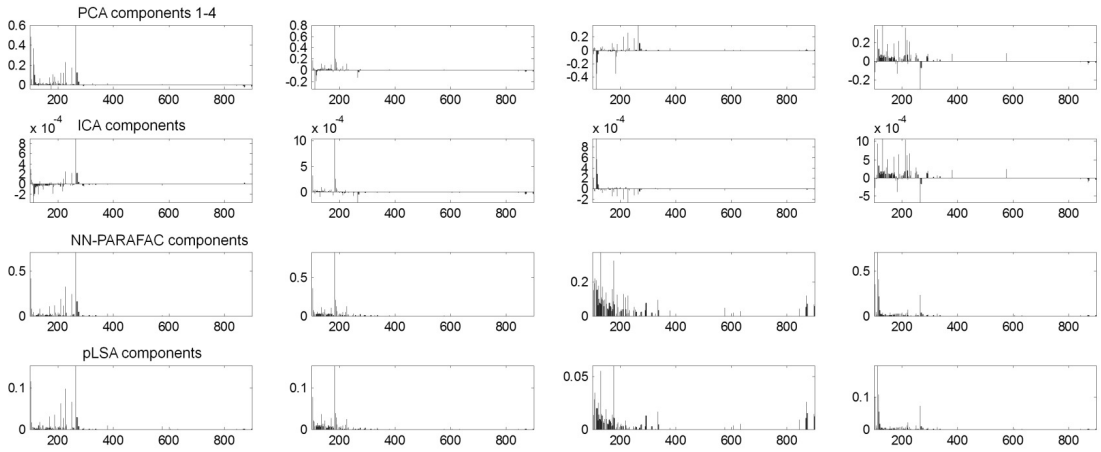

Figure 3.8.: The characteristic spectra of the four tissue types from the MALDI sample (see figures 3.3 and 3.7) as reconstructed by PCA, ICA, NN-PARAFAC and pLSA. The latter two result in all positive spectra and therefore allow direct physical interpretability. The most prominent peak in component spectrum two–which seems to correspond to the viable tumor area–lies at 184.5 Da (see also figures 3.7 and 3.10). This corresponds to recent findings that show that phosphocholine–which appears at that mass position–seems to play an important role in the discrimination of necrotic and viable tumor tissue (personal communication).

maximum number of iterations as well as on the random initialization. Furthermore, we used interpreted MATLAB code, and computational efficiency was not emphasized.

3.5. Discussion

3.5.1. Simulated Data

The decomposition results on our simulated datasets are illustrated in figures 3.2, 3.4 and 3.5.

Impure Mixtures. The abundance map and the characteristic spectrum of the first component are well reconstructed by all methods; the dominant indium peak at 115 Da is well visible. Nevertheless, PCA and ICA completely fail to recover the remaining two components which feature considerable spectral overlap. pLSA is able to recover significantly more structure which can be seen from the abundance maps in figures 3.4 and 3.5. NN-PARAFAC performs nearly as well as pLSA, but does worse for the third component.

Pure Mixtures. As expected, the decomposition task for pure mixtures is simpler and yields better results. PCA exhibits difficulty in extracting the second and third

component but is able to reconstruct the first component very well. ICA shows problems with representing all three components and even fails to extract the indium peak. Possible explanations are given below. In contrast, NN-PARAFAC and pLSA are able to deliver a convincing result and clearly outperform PCA and ICA. In spite of noise being present in the data, the reconstruction of the abundance maps is highly accurate for all three components.

The spectral components estimated by NN-PARAFAC and pLSA are much closer to the ground truth spectra than their PCA and ICA counterparts and outperform PCA and ICA with respect to reconstruction of major peaks (see table 3.3). The AICc-type criterion correctly estimated the number of components in the dataset to be three.

	Reconstruction Error					
	MALDI (4 components)			SIMS (5 components)		
Norm	PCA/ICA	NN-P.	pLSA	PCA/ICA	NN-P.	pLSA
L_1	$3.5 \cdot 10^{1}$	$2.4 \cdot 10^{1}$	$\mathbf{2.3 \cdot 10^{1}}$	$9.4 \cdot 10^{0}$	$\mathbf{3.7 \cdot 10^{0}}$	$3.9 \cdot 10^{0}$
L_2	$1.0 \cdot 10^{-1}$	$\mathbf{8.0 \cdot 10^{-2}}$	$9.5 \cdot 10^{-2}$	$1.4 \cdot 10^{-2}$	$\mathbf{5.5 \cdot 10^{-3}}$	$6.6 \cdot 10^{-3}$
KL	$2.2 \cdot 10^{0}$	$1.7 \cdot 10^{-1}$	$\mathbf{1.3 \cdot 10^{-1}}$	$1.2 \cdot 10^{0}$	$2.9 \cdot 10^{-2}$	$\mathbf{2.5 \cdot 10^{-2}}$

Table 3.1.: Reconstruction error for the two real-world datasets (MALDI, SIMS). The NN-PARAFAC (NN-P.) and pLSA reconstructions outperform the other methods by a large margin. PCA/ICA do only well with respect to the L_2-norm. pLSA does best for KL divergence, but it also performs well with respect to the other norms. Furthermore, NN-PARAFAC and pLSA grant increased interpretability.

3.5.2. Real-World Data

MALDI Dataset. Results are shown in figures 3.7 and 3.8. The reconstruction accuracy of both NN-PARAFAC and pLSA is high (see table 3.1). pLSA delivers the best result with respect to KL divergence, but it does also well regarding the other norms. NN-PARAFAC ranks highest with respect to the L_2-norm that is the only metric for which PCA and ICA also give a good result. We observe, that in a setting where only a few components are analyzed, NN-PARAFAC and pLSA outperform PCA and ICA with respect to reconstruction accuracy. For PCA and ICA, the complementarity of the estimated components is low (see table 3.2) and the assignments of reconstructed components to tissue types is not obvious, especially for components three and four. In contrast, the complementarity of the abundance maps calculated by pLSA is very high, and we get a clear spatial separation of the four types, which simplifies interpretation. Component two seems to represent the viable part of the tumor, component three the vascularized region, and components one and four seem to stand for gelatin. NN-PARAFAC performs better than PCA and ICA but does not pick up the vascularized part very well. The contrast of the abundance maps is lower than that of the respective

	Complementarity									
	MALDI (4 components)					SIMS (5 components)				
Quant.	PCA	ICA	NN-P.	pLSA	tm	PCA	ICA	NN-P.	pLSA	tm
95	0.20	0.19	0.19	**0.20**	0.20	**0.25**	0.24	**0.25**	**0.25**	0.25
90	0.37	0.35	0.35	**0.40**	0.40	0.47	0.46	0.49	**0.50**	0.50
85	0.54	0.50	0.50	**0.59**	0.60	0.63	0.64	0.68	**0.73**	0.75
80	0.67	0.63	0.61	**0.76**	0.80	0.75	0.77	0.82	**0.86**	1.00
75	0.76	0.74	0.69	**0.89**	1.00	0.84	0.86	0.91	**0.95**	1.00
70	0.85	0.84	0.75	**0.98**	1.00	0.92	0.93	0.97	**0.99**	1.00
65	0.94	0.91	0.80	**1.00**	1.00	0.97	0.96	0.99	**1.00**	1.00
60	0.98	0.95	0.84	**1.00**	1.00	0.99	0.98	**1.00**	**1.00**	1.00
55	0.99	0.98	0.88	**1.00**	1.00	**1.00**	0.99	**1.00**	**1.00**	1.00
50	0.99	0.99	0.91	**1.00**	1.00	**1.00**	**1.00**	**1.00**	**1.00**	1.00

Table 3.2.: Complementarity estimation for the two real-world datasets (MALDI, SIMS). The reported numbers correspond to the percentage of the region of interest that is covered after combining the four respectively five thresholded component abundance maps at various quantiles. The theoretical maximum (tm) of 1.00 is only reached for perfectly complementary abundance maps. In both scenarios, the pLSA solution is significantly more complementary than the PCA and ICA counterparts and more complementary than the NN-PARAFAC (NN-P.) solution.

pLSA components, which mirrors in the lower complementarity estimation values (see table 3.2). All methods have problems to separate the necrotic from the viable part in the four component decomposition.

The AICc criterion opted for a total of eight components. Manual inspection showed that this basically leads to splitting of the four components of interest described above for which we are currently evaluating biological reason. Indeed, the necrotic part of the tumor is much better represented in the eight component solution as shown in figure 3.9.[3] The performance estimates for a varying number of components is given in Appendix 3B. In short, the more components we took into account, the better the obtained reconstructions turned out, and when considering all PCA components, we ended up with a perfect reconstruction. However, in real-world applications we are often interested in reducing the dimensionality of the data and examine only few components. In this scenario, NN-PARAFAC and pLSA clearly outperformed PCA and ICA. Concerning the complementarity estimates, we noticed that taking more components into account was not necessarily beneficial for PCA with respect to its relative performance (cf. Appendix

[3]The complete decomposition results for the MALDI set and 8 components are given in Appendix 3A. Again, the abundance maps estimated by NN-PARAFAC and pLSA are much more complementary than their PCA/ICA counterparts.

		Peak reconstruction							
		Simulated set (impure mixtures)				Simulated set (pure mixtures)			
Quant.	Comp.	PCA	ICA	NN-P.	pLSA	PCA	ICA	NN-P.	pLSA
95	1	0.43	0.57	0.86	0.86	0.63	0.63	1.00	1.00
	2	0.71	0.71	0.71	0.71	0.63	0.63	0.88	0.88
	3	0.71	0.71	0.71	0.58	0.75	0.50	0.75	0.88
90	1	0.46	0.69	0.85	0.85	0.50	0.63	0.81	0.81
	2	0.77	0.85	0.77	0.77	0.38	0.50	0.94	0.94
	3	0.54	0.54	0.54	0.62	0.63	0.38	0.75	0.75
85	1	0.47	0.63	0.79	0.58	0.50	0.63	0.88	0.88
	2	0.79	0.79	0.84	0.84	0.42	0.46	0.92	0.92
	3	0.42	0.42	0.68	0.68	0.46	0.33	0.83	0.83
80	1	0.50	0.58	0.69	0.65	0.44	0.59	0.88	0.88
	2	0.69	0.73	0.88	0.88	0.38	0.44	0.88	0.88
	3	0.38	0.42	0.69	0.73	0.50	0.44	0.88	0.88
⊘ 80-95	1-3	0.57	0.64	**0.75**	0.73	0.50	0.51	0.86	**0.88**

Table 3.3.: Simulated dataset: the table quantifies which fraction of the most intense peaks in the (known) characteristic spectra of the three tissue types is also among the major peaks in the corresponding spectral components estimated by the four methods. For each method and each estimated spectral component we calculated various quantile spectra containing only the most intense peaks. The resulting peak lists were then compared with the corresponding quantile spectrum of the respective characteristic spectrum. The overlap of the two peak lists is expressed by a number in $[0; 1]$. A value of 1.00 is only reached if the positions of the major peaks in the ground truth spectrum and the reconstructed spectrum are identical. In the case of PCA and ICA, the signum of the reconstructed spectral components is arbitrary. Therefore, we calculated both the upper and lower quantile and used the quantile spectrum (upper/lower) that resulted in a higher percentage of overlap. The last row gives the average values of overlap obtained with the four methods. For pure and impure mixtures, the percentage of overlap is highest for NN-PARAFAC (NN-P.) and pLSA indicating that those methods are able to better reconstruct the major peaks of the characteristic spectra than PCA or ICA. Furthermore, ICA outperforms PCA, especially in the impure setting.

3B).

The spectral components estimated by PCA and ICA (cf. figure 3.8) are, as expected, partly negative. The phosphocholine-peak at 184.5 Da is detected as the most dominant peak by all methods, but physical interpretation of the negative parts of the PCA and ICA component spectra is difficult. This does not render the distribution of major peaks in the PCA and ICA components uninformative, but the characteristic spectra of the

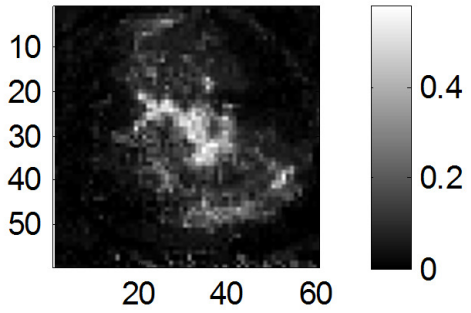

Figure 3.9.: This component of an eight component decomposition of the MALDI set with pLSA is convincingly correlated with the necrotic area (cf. the label map in figure 3.3). The full decomposition results are given in Appendix 3A.

underlying tissue types are not directly revealed. Figure 3.10 illustrates how the sparsity criterion described earlier can be used to automatically identify discriminating peaks.

SIMS Dataset. Decomposition results are given in figure 3.11. The reconstruction error for PCA and ICA is significantly higher than for NN-PARAFAC and pLSA, and the complementarity of the estimated components is not well expressed (cf. tables 3.1 and 3.2). The contrast in the PCA and ICA abundance maps is very low, and assigning these components to regions in the label map is difficult. In comparison, the NN-PARAFAC and pLSA solutions show higher contrast and complementarity. In the NN-PARAFAC and pLSA decompositions, the most prominent peak in the spectra corresponding to the first and third component (see figure 3.12) lies at 115 Da (indium). Thus, components one and three seem to correspond to background/holes. Component two and four seem to represent tumor and interface region, and component five seems to correspond to the gelatin region. Necrotic and viable part can be distinguished as well. The NN-PARAFAC result is similar to the pLSA solution; the reconstruction error is slightly reduced whereas the complementarity estimation shows that the components estimated with pLSA are slightly more sparse.

The AICc-type-controlled pLSA was capable of automatically estimating the number of components and only slightly preferred the decomposition with four components over the five component solution (see figure 3.2).

3.5.3. Interpretations and Method's Properties

Constraints and their Effects. Possible explanations for the inferior performance of PCA and ICA in our experiments lie in the constraints used by these methods. ICA relies on the assumption that the reconstructed sources have minimal mutual information, i.e., their statistical independence is maximal. The effect of this assumption is data-

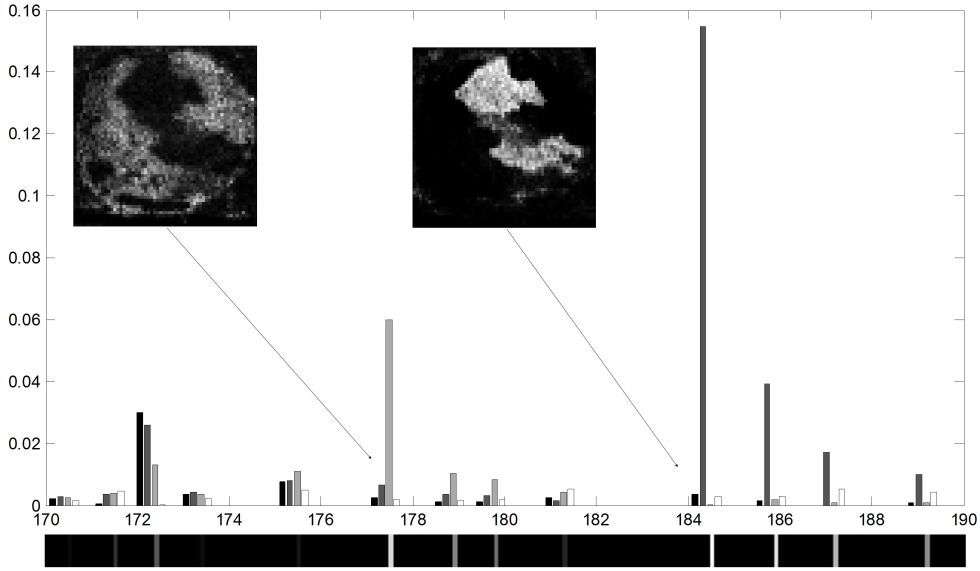

Figure 3.10.: An excerpt of the m/z range of the MALDI dataset between 170 and 190 Da. The bar plot shows the channel-wise intensities for the four components obtained with pLSA (from left to right). The bar on the bottom color-codes the level of sparsity for the intensity distribution for each m/z channel. Light colors indicate high sparsity (suggesting discriminative peaks) and dark colors indicate low sparsity. We also give the abundance maps corresponding to those m/z channels with the highest sparsity value. At 177.5 Da, we see a highly intense peak in the characteristic spectrum of tissue type three and low intensities for the characteristic spectra of the other components. Apparently, this peak is typical for tissue type three and can be used to distinguish this component from the others. In contrast, if a peak appears with equal intensity in all four component spectra, it has no discriminatory power.

dependent and may be beneficial in some scenarios and harmful in others. In situations where we want to unmix data containing tissue types that differ only slightly and thus feature similar spectral signatures, ICA may not be the right choice as a decomposition method. It is unlikely to yield two similar characteristic spectra since these would feature high mutual information. However, if this assumption holds we can hope for a good performance of ICA.

PCA is handicapped in the detection of differences in tissue composition that manifest themselves merely in small spectral changes. Such subtleties contribute little to the overall variance of the data and are hence relegated to higher principal components which are easily overlooked in the routine visual inspection of the first few components. In our

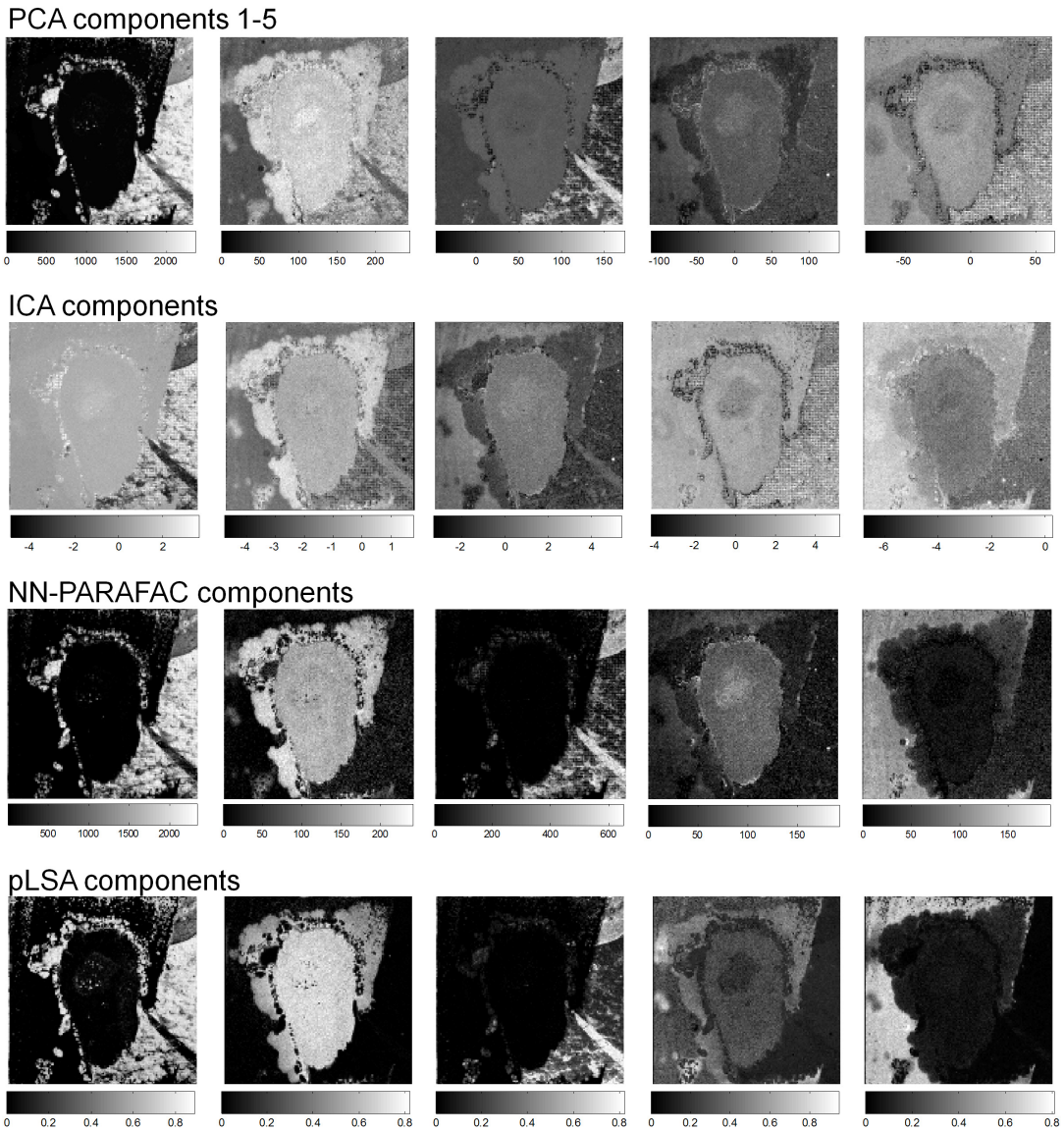

Figure 3.11.: Decomposition of the SIMS set with five components. The corresponding spectral components can be found in figure 3.12. Also see the caption of figure 3.7.

experiments, NN-PARAFAC and pLSA were better able to detect minor differences in unmixing both, the pure and impure mixtures (cf. figures 3.4 and 3.5). In contrast to the constraints used by PCA and ICA, the non-negativity constraint in NN-PARAFAC and pLSA is well motivated by physical properties and valid for all count datasets.

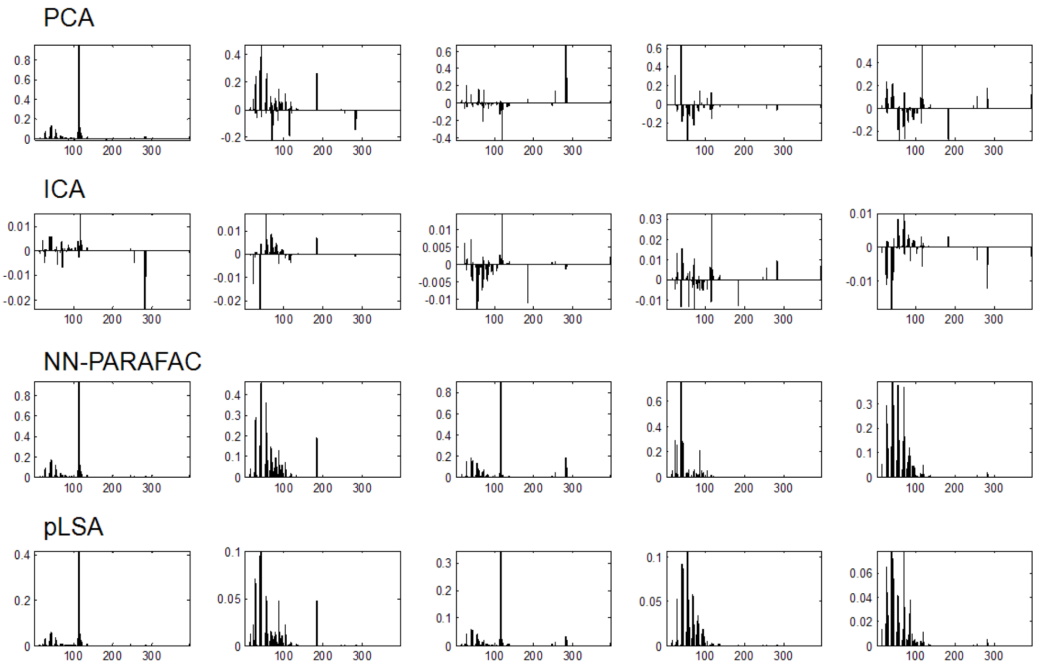

Figure 3.12.: Unsupervised decomposition of the SIMS set with five components. The components have been rearranged according to the ordering of the abundance maps. The most prominent peak in component one corresponds to indium (115 Da). Indeed, the tissue is torn in the respective areas, and the indium tin oxide-coated glass slide is exposed.

Number of Components and Reconstruction Accuracy. In contrast to PCA, NN-PARAFAC and pLSA require the number of components k to be specified in advance. This means that all observed spectral intensity and variability is distributed among the k component spectra. This behavior is desirable if k equals the correct number of components. If k underestimates the correct number of components, such an approach does not fully exploit information on further tissue types (for instance, see figure 3.7 where the four component decomposition of the MALDI set does not reveal the necrotic part). For PCA, the number of components is not specified in advance, and the full decomposition is always computed, even in cases where prior knowledge is available. Especially in such situations, PCA results in way too many components since the number of tissue types that one is interested in is normally much smaller than the number of mass channels, that is the number of principal components estimated. In typical situations, one can only examine the first few of possibly hundreds of principal components. This leads to a loss of information and reconstruction accuracy. We have demonstrated that

in scenarios where the number of components is limited, NN-PARAFAC and pLSA are superior to PCA and ICA. As demonstrated in a separate experiment, this also applies to the ability to reconstruct major peaks.

Complementarity. In our experiments on both simulated and real-world data, pLSA clearly outperformed PCA, ICA and NN-PARAFAC with respect to the complementarity of the estimated components (cf. table 3.2). It is not surprising that PCA results in low complementary estimates since it always performs the full decomposition. For pLSA, very often the theoretical maximum of the complementarity measure is reached or almost reached. The numbers presented are backed by visual inspection that also suggests that the NN-PARAFAC and pLSA partitionings are more sparse than the PCA and ICA solutions, even though sparsity is not explicitly enforced. Sparsity simplifies interpretation and should be recovered if present in the data, that is if most of the tissue predominantly belongs to one (but not necessarily the same) class. The results obtained on the simulated datasets show that NN-PARAFAC and pLSA not only perform well in the pure mixture case, but also for heterogeneous tissue (cf. figures 3.4 and 3.5). The different mixture areas were better reconstructed than with PCA or ICA. Furthermore, the NN-PARAFAC and pLSA decomposition maps corresponding to the real-world data were convincingly correlated with the structures that are visible in the label maps (cf. figures 3.3, 3.7 and 3.11).

Computation Time. The computation time needed for NN-PARAFAC and pLSA is higher than for PCA and ICA. However, in relation to the time required for data acquisition, it seems safe to say that all methods are sufficiently fast to be applied in practice.

3.6. Conclusion

We have shown on simulated and real-world data that pLSA is a suitable approach for the unsupervised analysis of mass spectrometry images. Both pLSA and NN-PARAFAC outperform PCA and ICA in terms of the quality of the decomposition maps as they use an additive model that correctly mirrors the physical properties of the data. In addition, they offer superior physical interpretability as they produce normalized and non-negative components which can directly be interpreted as peak intensity lists. They also lead to more complementary components and retain high reconstruction accuracy. In contrast to non-negative PARAFAC, pLSA is based on a sound probabilistic model.

We have further introduced the AICc-controlled pLSA, providing the methodology necessary to automatically estimate the number of tissue types, thus significantly decreasing the dependency on sample-specific prior knowledge. A sparsity measure was used to automatically identify those m/z channels that are relevant for the discrimination of tissue types and may give further valuable insights in exploratory data analysis.

Appendix

3a. Decomposition of the MALDI Set with 8 Components

Figures 3.13 and 3.14 as well as tables 3.4 and 3.5 illustrate the unsupervised decomposition results for the MALDI set with eight components. The abundance maps estimated by NN-PARAFAC and pLSA are much more distinct than their PCA/ICA counterparts. Most of them can directly be assigned to one of the expected tissue types like viable or necrotic tumor. Refer to the figure captions for further details.

	Reconstruction Error MALDI (8 components)		
Norm	PCA/ICA	NN-PARAFAC	pLSA
L_1	$1.8 \cdot 10^{1}$	$1.5 \cdot 10^{1}$	$\mathbf{1.3 \cdot 10^{1}}$
L_2	$4.9 \cdot 10^{-2}$	$\mathbf{4.8 \cdot 10^{-2}}$	$5.4 \cdot 10^{-2}$
KL	$2.7 \cdot 10^{-1}$	$7.5 \cdot 10^{-2}$	$\mathbf{4.7 \cdot 10^{-2}}$

Table 3.4.: Reconstruction error for the MALDI set using an eight component decomposition. Again, the NN-PARAFAC and pLSA reconstructions perform best.

	Complementarity MALDI (8 components)				
Quant.	PCA	ICA	NN-PARAFAC	pLSA	th.max.
95	0.35	0.36	0.32	**0.39**	0.40
90	0.61	0.61	0.57	**0.72**	0.80
85	0.79	0.79	0.74	**0.92**	1.00
80	0.90	0.89	0.84	**0.99**	1.00
75	0.96	0.96	0.91	**1.00**	1.00
70	0.99	0.99	0.94	**1.00**	1.00
65	**1.00**	0.99	0.96	**1.00**	1.00
60	**1.00**	**1.00**	0.96	**1.00**	1.00
55	**1.00**	**1.00**	0.97	**1.00**	1.00
50	**1.00**	**1.00**	0.98	**1.00**	1.00

Table 3.5.: Complementarity estimation for the MALDI set using an eight component decomposition: The reported numbers correspond to the percentage of the region of interest that is covered after combining the thresholded component abundance maps at various quantiles. The theoretical maximum (tm) of 1.00 is only reached for perfectly complementary abundance maps. Again, the pLSA solution is significantly more complementary than the PCA, ICA and NN-PARAFAC results.

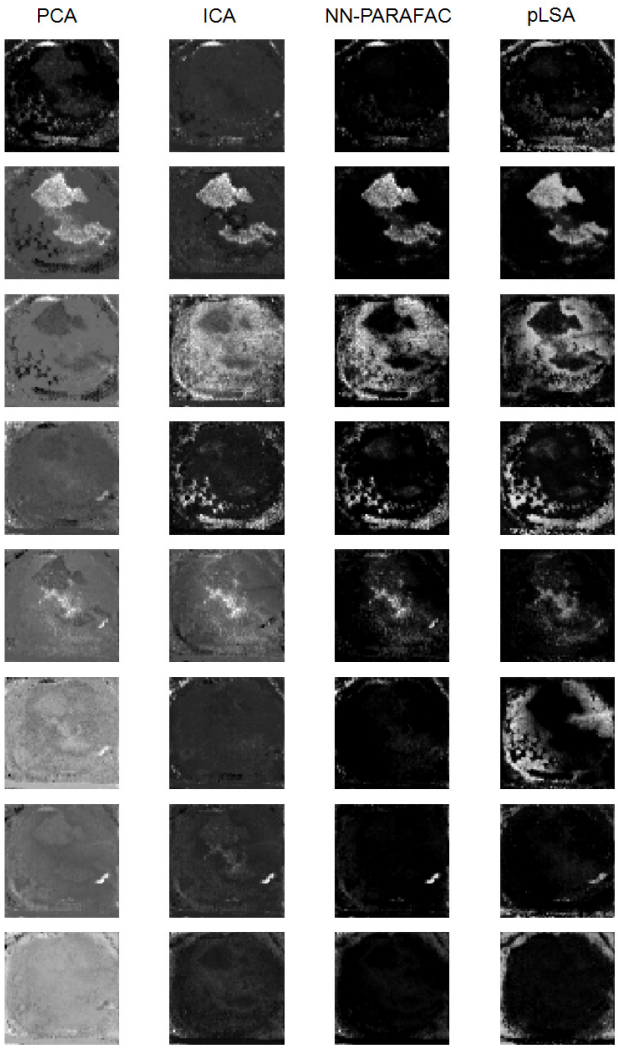

Figure 3.13.: Unsupervised decomposition of the MALDI set with eight components: The components obtained by NN-PARAFAC and pLSA are much more distinct than their PCA/ICA counterparts–especially with respect to component three (vascularized region). Components two (viable tumor) and five (necrotic tumor) also have a clear localization. pLSA splits up the third component into two parts, yielding components three and six. All components estimated by pLSA show spatial coherence. This goes well with the assumption that tissue normally has a spatial extent. ICA mostly outperforms PCA (see for example component four). The corresponding spectral components are shown in figure 3.14.

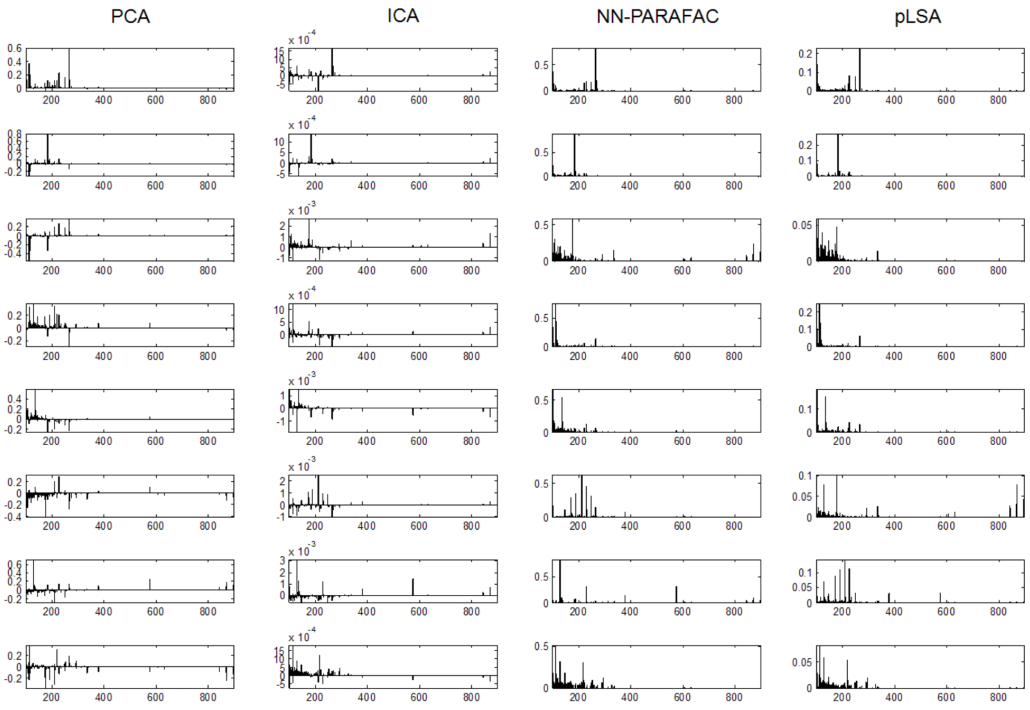

Figure 3.14.: Unsupervised decomposition of the MALDI set with eight components. All reconstructed spectral components have been arranged according to figure 3.13.

3b. Unsupervised Decomposition with a Varying Number of Components

The results obtained for the four and eight component decompositions of the MALDI set suggest that NN-PARAFAC and pLSA are superior to PCA and ICA in terms of reconstruction accuracy and complementarity. In the following experiment we quantified the reconstruction error as well as the complementarity for a varying number of components. Results are given in tables 3.6 and 3.7.

Since we always selected the best combination of upper/lower quantile abundance maps for PCA/ICA (see section 3.3) we only calculated decompositions where the combinatorial complexity was still manageable. The more components we took into account, the more accurate the reconstructions were, and when considering all PCA components, we ended up with a perfect reconstruction. However, in real-world applications we are often interested in reducing the dimensionality of the data and examine only few components. In this scenario, NN-PARAFAC and pLSA clearly outperformed PCA and ICA (cf. table 3.6).

		Reconstruction Error		
Comp.	Norm	PCA/ICA	NN-PARAFAC	pLSA
2	L_1	$5.1 \cdot 10^{1}$	$3.9 \cdot 10^{1}$	$\mathbf{3.7 \cdot 10^{1}}$
	L_2	$1.5 \cdot 10^{-1}$	$\mathbf{1.4 \cdot 10^{-1}}$	$1.8 \cdot 10^{-1}$
	KL	$3.8 \cdot 10^{1}$	$5.0 \cdot 10^{-1}$	$\mathbf{2.8 \cdot 10^{-1}}$
3	L_1	$4.4 \cdot 10^{1}$	$3.3 \cdot 10^{1}$	$\mathbf{2.8 \cdot 10^{1}}$
	L_2	$1.2 \cdot 10^{-1}$	$\mathbf{1.0 \cdot 1.0^{-1}}$	$1.3 \cdot 10^{-1}$
	KL	$3.6 \cdot 10^{0}$	$3.8 \cdot 10^{-1}$	$\mathbf{1.7 \cdot 10^{-1}}$
4	L_1	$3.5 \cdot 10^{1}$	$\mathbf{2.4 \cdot 10^{1}}$	$2.5 \cdot 10^{1}$
	L_2	$1.0 \cdot 10^{-1}$	$\mathbf{8.0 \cdot 10^{-2}}$	$1.2 \cdot 10^{-1}$
	KL	$2.2 \cdot 10^{0}$	$1.7 \cdot 10^{-1}$	$\mathbf{1.4 \cdot 10^{-1}}$
5	L_1	$3.0 \cdot 10^{1}$	$2.2 \cdot 10^{1}$	$\mathbf{2.0 \cdot 10^{1}}$
	L_2	$8.6 \cdot 10^{-2}$	$\mathbf{7.0 \cdot 10^{-2}}$	$8.3 \cdot 10^{-2}$
	KL	$1.6 \cdot 10^{0}$	$1.5 \cdot 10^{-1}$	$\mathbf{9.1 \cdot 10^{-2}}$
6	L_1	$2.8 \cdot 10^{1}$	$1.8 \cdot 10^{1}$	$\mathbf{1.6 \cdot 10^{1}}$
	L_2	$6.6 \cdot 10^{-2}$	$\mathbf{5.8 \cdot 10^{-2}}$	$7.1 \cdot 10^{-2}$
	KL	$4.9 \cdot 10^{-1}$	$9.7 \cdot 10^{-2}$	$\mathbf{6.8 \cdot 10^{-2}}$
7	L_1	$2.1 \cdot 10^{1}$	$1.6 \cdot 10^{1}$	$\mathbf{1.5 \cdot 10^{1}}$
	L_2	$5.9 \cdot 10^{-2}$	$\mathbf{5.1 \cdot 10^{-2}}$	$6.2 \cdot 10^{-2}$
	KL	$3.8 \cdot 10^{-1}$	$8.0 \cdot 10^{-2}$	$\mathbf{5.7 \cdot 10^{-2}}$
8	L_1	$1.8 \cdot 10^{1}$	$1.5 \cdot 10^{1}$	$\mathbf{1.3 \cdot 10^{1}}$
	L_2	$4.9 \cdot 10^{-2}$	$\mathbf{4.5 \cdot 10^{-2}}$	$5.4 \cdot 10^{-2}$
	KL	$2.7 \cdot 10^{-1}$	$7.1 \cdot 10^{-2}$	$\mathbf{4.7 \cdot 10^{-2}}$
9	L_1	$1.5 \cdot 10^{1}$	$\mathbf{1.3 \cdot 10^{1}}$	$\mathbf{1.3 \cdot 10^{1}}$
	L_2	$4.1 \cdot 10^{-2}$	$\mathbf{4.0 \cdot 10^{-2}}$	$5.2 \cdot 10^{-2}$
	KL	$1.4 \cdot 10^{-1}$	$5.6 \cdot 10^{-2}$	$\mathbf{4.3 \cdot 10^{-2}}$
10	L_1	$1.5 \cdot 10^{1}$	$\mathbf{1.2 \cdot 10^{1}}$	$\mathbf{1.2 \cdot 10^{1}}$
	L_2	$3.8 \cdot 10^{-2}$	$\mathbf{3.6 \cdot 10^{-2}}$	$4.7 \cdot 10^{-2}$
	KL	$1.3 \cdot 10^{-1}$	$5.2 \cdot 10^{-2}$	$\mathbf{3.9 \cdot 10^{-2}}$

Table 3.6.: Reconstruction error for the MALDI dataset: a varying number of components (first column) was used to confirm the results obtained for the four and eight component decompositions. Since we always selected the best combination of upper/lower quantile abundance maps for PCA/ICA (see section 3.3) we only calculated decompositions where the combinatorial complexity was still manageable. Naturally, the more components we used, the more accurate the PCA reconstructions were. However, in a real-world scenario we are often interested in reducing the dimensionality of the data and normally consider only a few components. For the MALDI-set, the AICc-type criterion suggested to use eight components for which the reconstructions of NN-PARAFAC and pLSA were more precise.

		Complementarity				
Comp.	Quantile	PCA	ICA	NN-PARAFAC	pLSA	th.max.
2	95	**0.10**	**0.10**	**0.10**	**0.10**	0.10
	75	**0.50**	**0.50**	0.46	**0.50**	0.50
	55	0.77	0.77	0.69	**0.90**	0.90
3	95	**0.15**	0.14	0.14	**0.15**	0.15
	75	0.65	0.66	0.55	**0.75**	0.75
	55	0.97	0.96	0.75	**1.00**	1.00
4	95	**0.20**	0.19	0.19	**0.20**	0.20
	75	0.76	0.74	0.69	**0.89**	1.00
	55	0.99	0.98	0.88	**1.00**	1.00
5	95	0.24	0.23	0.22	**0.25**	0.25
	75	0.88	0.84	0.77	**0.98**	1.00
	55	**1.00**	0.99	0.94	**1.00**	1.00
6	95	0.28	0.28	0.26	**0.30**	0.30
	75	0.92	0.90	0.83	**1.00**	1.00
	55	**1.00**	**1.00**	0.95	**1.00**	1.00
7	95	0.31	0.32	0.29	**0.34**	0.35
	75	0.95	0.95	0.86	**1.00**	1.00
	55	**1.00**	**1.00**	0.96	**1.00**	1.00
8	95	0.35	0.36	0.32	**0.39**	0.40
	75	0.96	0.96	0.91	**1.00**	1.00
	55	**1.00**	1.00	0.97	**1.00**	1.00
9	95	0.37	0.39	0.36	**0.43**	0.45
	75	0.97	0.97	0.91	**1.00**	1.00
	55	**1.00**	1.00	0.97	**1.00**	1.00
10	95	0.39	0.42	0.36	**0.45**	0.50
	75	0.98	0.98	0.92	**1.00**	1.00
	55	**1.00**	**1.00**	0.97	**1.00**	1.00

Table 3.7.: Complementarity estimation for the MALDI dataset: a varying number of components (first column) was used to check if the results obtained from the four and eight component decompositions also hold here (see also figure 3.6). Again, pLSA outperformed the competing methods.

Chapter 4

Toward Digital Staining using Mass Spectrometry Imaging and Random Forests

In the previous chapter, we have proposed the application of probabilistic latent semantic analysis (pLSA) for the analysis of mass spectrometry images in an *unsupervised* setting, that is where no prior knowledge on the composition of a tissue sample is available. In an increasing number of pre-clinical studies, some prior knowledge on the spatial composition of a tissue sample is available and spatially resolved labels exist (i.e., training examples where the observed data points are annotated with labels from a given set, see section 4.3.1). In such a scenario, the unsupervised methods can be replaced by more powerful *supervised methods* such as support vector machines (SVM) [188], random forests [25] or other classifiers to automatically distinguish between the classes of interest. These algorithms can constitute a valuable tool for pathologists or medical doctors that have to analyze large numbers of tissue samples. In that case, reliable classifiers can help minimize the risk for false negative diagnoses.

Currently, wet lab staining is the state-of-the-art method for highlighting structures in biological tissues. However, a tissue sample can be treated with a limited number of stains only (cf. section 2.2). Mass spectrometry imaging, in contrast, simultaneously monitors the spatial distribution of several hundreds or even thousands of molecules. We show that combining MSI with random forest classification ("digital staining") yields a powerful complement to chemical staining techniques.

4.1. Introduction

Several supervised classifiers have been applied in the field of mass spectrometry: k-nearest neighbors (knn) [183, 209], SVMs [190, 83, 92] and other approaches [191, 233]. Random forests have also successfully been used [14, 81, 217, 56, 94, 84, 110], but to our knowledge not on MSI data and often only for binary classification tasks. Random forests have many favorable properties. Empirically, the algorithm is robust to overfitting, the method has high prediction accuracy, is capable of dealing with a large number of input variables [132, 26], is rather robust to label noise [25, 182], allows easy and fast training (i.e., only a few seconds in the scenario described below), and performance is robust with respect to the exact choice of the two hyperparameters: the number of trees, and the size of the random feature subset evaluated at a node [162]. Random forests have successfully been applied to various kinds of spectral data, including remote sensing [86], astrophysics [78], and magnetic resonance spectroscopic imaging (MRSI) [145]. Previous studies have compared the performance of random forests to SVMs and other state-of-the-art methods in many fields of application and have concluded that they deliver comparable [60, 2, 160, 56, 162] or even superior [217] performance. The fact that there is no clear benefit of using SVMs or random forests but that random forests are "more" non-parametric, i.e., require less tuning, and that their training is fast (featuring a complexity of $O(N \cdot \log N \cdot \sqrt{M})$ for N training examples in M dimensions) makes this algorithm a natural choice for our study. We first demonstrate that random forests are a highly suitable automated approach for classifying MSI data. In spite of technical variability being present in our data, the classifier results in predictions with high sensitivity and high positive predictive values.

Owing to noise or instabilities in the data acquisition process, the classification maps obtained with random forests (or other classifiers) can have a "noisy" appearance. Single pixels that have been classified differently than their surrounding area can frequently be observed (see sections 4.4 and 4.5). Typically, the "gold standard" label maps provided by human experts tend to be much more homogeneous (see section 4.3.1). We therefore apply a post-hoc smoothing method that removes these "outliers" from the classification maps. To this end, we compare a Markov random field (MRF) [82, 20] approach to a vector-valued median filter [225] and show that post-hoc smoothing significantly improves the sensitivities and positive predictive values.

In this study, we analyze human breast cancer cells grown as tumor xenografts in mice (see also section 3.3.2). Within these tumor xenografts, five different regions (necrotic tissue, viable tumor, gelatin, tumor interface, glass/hole) can be identified (see section 4.3.1). The data comprises a total of 7 slices from two tumors (same cell line, no genetic variation). We describe and discuss in detail how the random forest algorithm combined with post-hoc smoothing techniques can be used for automated MSI data classification and show that it results in predictions with high sensitivity estimates and positive predictive values of about 90%.

4.2. Materials and Methods

4.2.1. Random Forest

The random forest classifier [25] is a decision tree based ensemble method. In contrast to single decision trees, a randomized tree ensemble is robust to overfitting [25]. In a typical training setup, a few hundred decision trees are constructed which in their entirety constitute the "forest". The single trees are generated by the following algorithm:

Let $S = \{(x^1, y^1), \ldots, (x^N, y^N)\}$ be the set of available M-dimensional training samples, that is mass spectra $x^k \in \mathbb{R}^M$ with M channels and corresponding class labels $y^k \in \mathcal{L}$–e.g., cancerous or healthy tissue[1]. First, a bootstrap sample (in-bag sample) of size N is chosen by randomly sampling N times from S with replacement. This training set is used for building the tree, whereas the rest of the samples (out-of-bag sample) is used for estimating the out-of-bag error. Construction starts from the root that contains all training samples. At each node, a subset of size $\tilde{M}$ of the M features is chosen at random (by default, $\tilde{M} = \sqrt{M}$ [25]), and that feature in the subset that allows for the best separation of the classes in the sample set in that node is determined. In other words, in each node, a set of $\tilde{M}$ random tests of the form $g(x) > \Theta$ with threshold Θ is created, and the best test according to a quality criterion is selected [182]. The node is subsequently split into two child nodes, and the process is repeated until the tree is fully grown, i.e., all leaf nodes contain samples from one class only (see figure 4.1). Finally, the respective labels are assigned to the leaf nodes. Note that the trees are *not* pruned [96]. For each tree, the training error is estimated by predicting the classes of the out-of-bag samples and comparing the results with the true class memberships y^k.

The importance of the different features (in our case mass channels) for the classification can be estimated with the *permutation accuracy criterion* [25]. In short, for each tree the prediction accuracy on the out-of-bag sample is calculated in a first step. Then this process is repeated $\tilde{M}$ times where in each run the values of the l-th feature are randomly permuted ($l = 1, \ldots, \tilde{M}$). For each feature the difference between the two accuracies (non-permuted and permuted) is averaged over all trees. This *mean decrease in accuracy* is used as feature importance measure. In addition to the overall feature importance, a class-wise version can be calculated. Intuitively, a feature is considered unimportant if permuting its values does not (or only marginally) affect prediction accuracy.

After training of the forest, a sample (in our case spectrum) is classified by putting it down on each of the trees in the ensemble. Each tree constitutes a crisp classifier and returns the label corresponding to the leaf node in which the sample ends up. The random forest averages over all trees, and the classification result is represented by a probability vector that reflects how many trees have voted for one specific class. This output can be interpreted as posterior probability that an object belongs to a particular

[1] note that whenever we refer to a physical sample we use the term tissue sample

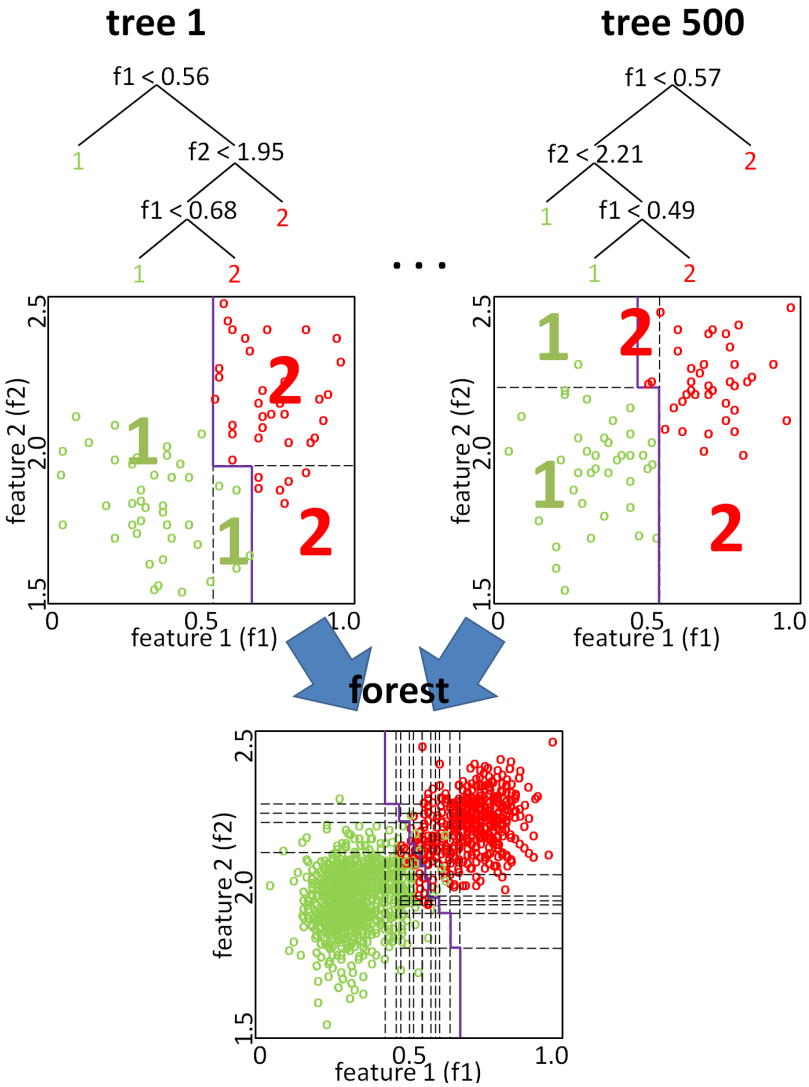

Figure 4.1.: The random forest classifier is an ensemble of decision trees where the single trees are constructed from bootstrap samples. On the top, two trees of the forest are shown in detail: At each node, the feature that allows for the best class separation is chosen (with respect to the subset of features selected for that node). The corresponding partitioning of the feature space is shown below with the decision boundary plotted in purple. On the bottom, the decision boundary of the random forest is displayed. It is based on the majority votes of the individual trees.

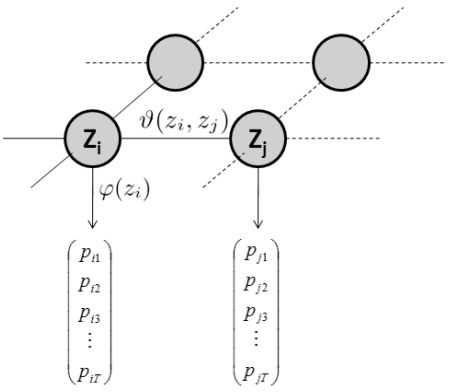

Figure 4.2.: Markov random field (MRF) model: Each node represents a pixel in the classification map. The optimal label assignment depends on both single site potential (φ) and pair potential (ϑ). The former is defined by the random forest output that assigns class probabilities to pixels, and the latter penalizes the assignment of different labels to neighboring pixels.

class, given the features of that object. In short, the output of a random forest can be seen as an estimate for class probabilities. A crisp classification can be obtained by taking the mode of this distribution.

4.2.2. Smoothing

In order to remove salt and pepper noise structures from the classification maps (i.e., images of the predicted class probabilities), we apply a post-hoc smoothing by means of Markov random fields (MRF) and a vector-valued median algorithm.

Markov Random Fields. In the MRF, each pixel of the classification map is represented by a node with 4-connectivity (see figure 4.2). In brief, the MRF employed is defined as follows (for an extensive introduction, see [228]). Each node takes a value in the set of labels (here $\mathcal{L} = \{$necrotic tissue, viable tumor, gelatin, tumor interface, glass/hole$\}$, see section 4.3.1). Motivated by a local homogeneity assumption and the available label maps, a regularized solution is sought that is a good compromise of two factors: the single site potentials (SSP) that encourage the agreement of each label with the local classification result (data term) and the pair potentials (PP) that call for the consistency of each label with the labels of the surrounding pixels and therefore encourage smoothness of the label map. According to this model, the optimum compromise or map of labels for all pixels, Z, is found as the maximizer of the log probability [208]

$$\log(p(Z|S)) = \sum_{i=1}^{N} \log(\varphi(z_i)) + \lambda \sum_{i=1}^{N} \sum_{j \in neigh(i)} \log(\vartheta(z_i, z_j)) + \text{const.} \qquad (4.1)$$

Here, *neigh(i)* identifies the set of neighboring nodes for node i and z_i denotes the label assigned to node i. The first term represents the single site potential $\varphi(z_i)$ for node i, and $\vartheta(z_i, z_j)$ is the pair potential for the neighboring nodes i and j weighted by the scalar λ. We use the random forest output as single site potential and define the pair potential $\vartheta(z_i, z_j)$ as follows: a (maximum) fit of 1 is assigned if i and j share the same label and $0 < c \ll 1$ otherwise. Subsequently, the pair potential matrix is normalized such that the sum of each column (and row) equals one.

An approximation maximizer of equation (4.1) can be found efficiently with loopy belief propagation (BP) in its max-sum version [165, 20]. BP is an iterative algorithm that tries to find a maximum a-posteriori estimate of the label distribution by repeatedly passing local messages between neighboring nodes of the MRF graph. These messages build on the potentials defined above and quantify the local fit of the labels to the data and the prior assumptions. After a stopping criterion is met, so-called *"beliefs"* are calculated for each node that express the probability of assigning a certain label to a node. Finally, the label with maximum belief is selected for each node. Since the defined graph contains cycles, BP is not guaranteed to converge. However, it usually results in good approximations of the optimum solution [124].

Vector-valued Median Filter. The scalar median filter [111] is known to efficiently remove salt-and-pepper noise in gray-valued images. Welk [225] introduced a vector-valued version of the median that Lerch [128] later enhanced by weighting factors. The weighted vector-valued median μ of a set $\tilde{S} = \{\tilde{x}^1, \ldots, \tilde{x}^K\}$ of K vectors in a M-dimensional feature space is given by

$$\mu(\tilde{S}) = \underset{a \in \mathbb{R}^M}{\operatorname{argmin}} \left(\sum_{k=1}^{K} w_k \left\| \tilde{x}^k - a \right\|_2 \right) \tag{4.2}$$

where $\|.\|_2$ denotes the Euclidean distance and w_k the weight of the k-th spectrum in $\tilde{S}$. The weights w_k were chosen from a Gaussian kernel centered at the respective pixels and with variance $\sigma^2 = 1.5$. The convex optimization problem in eq. (4.2) can be solved with a gradient descent approach [225]. Note that the resulting median μ is not necessarily a member of $\tilde{S}$.

4.3. Experiments

4.3.1. Data

Experimental data was acquired from orthotopic human breast cancer xenografts (MCF-7) grown in mice. For this study, six parallel tissue slices of the same tumor (entitled S3, S4, S5, S7, S9, S11; the S-slices) were subjected to secondary ion mass spectrometry (SIMS) MSI analysis using a Physical Electronics TRIFT II TOF SIMS instrument. One of the six obtained datasets (S9) was already used to evaluate the pLSA method in the

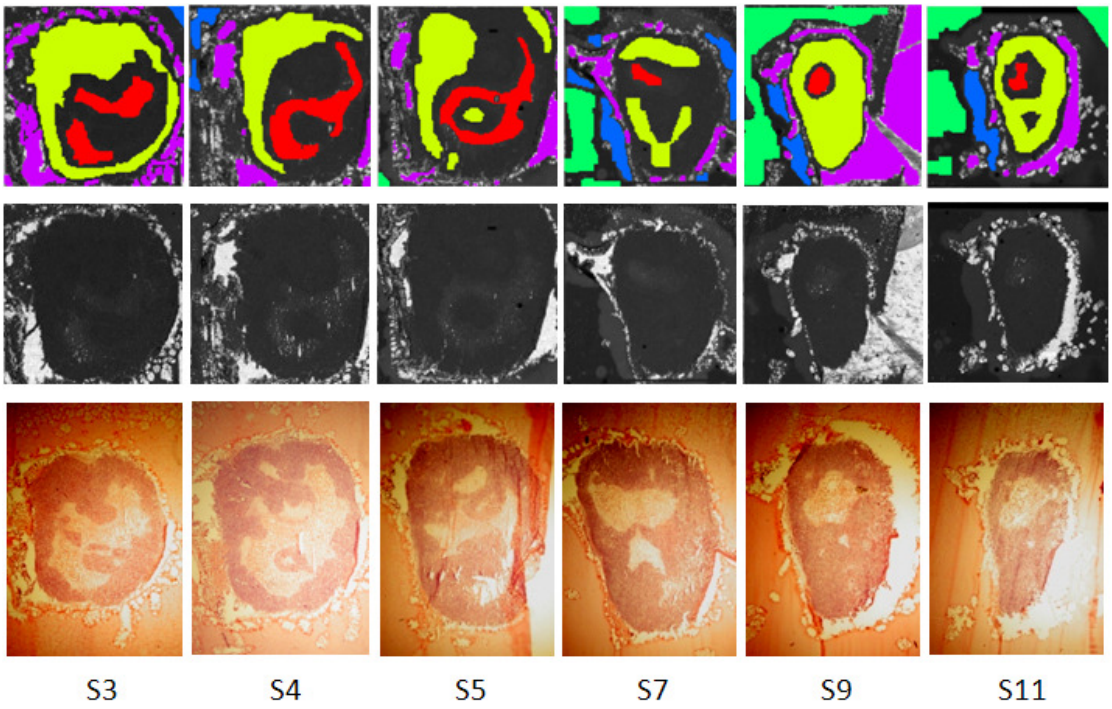

Figure 4.3.: Labels for the six slices S3, S4, S5, S7, S9, and S11: the top row shows the label maps that have been obtained by closely investigating the corresponding stained parallel slices (bottom row) as well as the total ion count images (TIC, middle row). Five regions can be observed: necrotic (red), viable (yellow), gelatin (green), interface (blue), and glass/hole (violet). Black/white indicates that no label is available. Note that after normalization (see Data processing section) the structure in the TIC images is no longer visible.

preceding chapter (see section 3.3.2 for details). Slices with odd numbers are equispaced, and the distance between S3 and S5 equals $\approx 500\mu m$. Additionally, one slice (entitled T1) from a second tumor of the same cell line, grown in a genetically identical mouse was analyzed. No genetic variation was present in the data. For each tumor slice, an additional hematoxylin-eosin (HE) stained parallel slice is available.

Despite some topological differences between the HE-stained and MSI-subjected slices, we again used the stained images as gold standards in the labeling process. Note, however, that this is an approximative ground truth only. As can be seen from the label maps in figure 4.3, five different regions are present in the tissue samples: necrotic tissue (overall 4,844 labeled spectra), viable/active tumor (16,663), embedding gelatin (6,340), tumor interface region (3,114), and glass/holes (10,373). In the glass/holes area, the

tissue was torn in the freeze-drying/microtome cutting process, and the glass surface is exposed to the analytical ion beam.

4.3.2. Research Questions

We addressed five research questions in our experiments. Experiments 1 to 3 concerned the performance of the random forest algorithm on real-world MSI data, experiment 4 analyzed the effect of post-hoc smoothing, and experiment 5 dealt with identifying important features for the classification:

- Experiment 1: We evaluated if random forests are capable of distinguishing different tissue types at all; in this restricted setting, a cross validation over pixels was performed, using all slices from the first mouse.
- Experiment 2: We used samples from all but one S-slice for training and evaluated the classifier's performance by predicting the classes for the (labeled) samples of the remaining S-slice ("leave one-slice-out cross validation"). This experiment shows if the classifier generalizes well to different parts of the *same* tumor in the *same* individual that were acquired in separate experiments.
- Experiment 3: We trained the algorithm on the six S-slices and tested it on the T1 slice. This experiment shows if the classifier generalizes well if the *same* tumor is studied in *different* individuals. This experiment is still limited in that both the tumors and the individuals are clones, that is potential variability between different cell lines or individuals is not covered.
- Experiment 4: The next research question analyzed the effect of smoothing of the classification results with the Markov random field and the vector-valued median filter defined above.
- Experiment 5: Finally, we were interested which features are decisive in the classification of the tissue samples.

4.3.3. Evaluation Criteria

We used balanced training sets [44] for random forest training, that is we trained the classifier with the same number of training samples for each tissue class. To ensure that a sufficient number of training samples was available from each slice, we further balanced the datasets by training the forest with the same number of labeled samples per class and slice. Ten-fold-cross-validation over pixels was used to evaluate the classifier's performance by means of sensitivity (SE) and positive predictive value (PPV) (see experiment 1 below). For a given class k, the sensitivity (also termed "true positive rate" and "recall") measures the ratio of samples correctly classified as k to all samples

that really belong to class k, that is $SE = \frac{TP}{TP+FN}$ where TP is the number of true positives, and FN is the number of false negatives. The PPV (also termed "precision") estimates the ratio of samples correctly classified as k among all samples classified as k: $PPV = \frac{TP}{TP+FP}$ where FP is the number of false positives. To ensure a fair comparison, we averaged the results over five training runs to minimize the influence of the random sampling of training points.

4.3.4. Data Processing

In our study, we compare different datasets, each of which corresponds to one mass spectral image. Mass spectral count data are not quantitative due to matrix and ionization effects and possible other variations. Moreover, depending on the acquisition time per spot and other instrument settings, the average intensity can vary, even if the same kind of tissue is imaged. We normalize the datasets by dividing each spectrum by its total ion count (TIC). This way, intensity differences visible in the total ion count images (cf. figure 4.3) are removed. It is ensured that the classification is based on different spectral distributions and not on intensity differences. The scaled spectra are baseline corrected by subtraction of the channel-wise minimum with respect to all spectra in a dataset. We detect the local maxima in the mean spectrum of each dataset and keep the mass channels that correspond to maxima that exceed a given threshold in order to extract features and obtain "feature lists" (also cf. section 5.2.4). To increase the robustness of the quantitation, the pertaining intensity value is calculated as the integral over the whole peak width. In our experiments, peak picking on the average spectrum was more robust to noise than peak picking on individual spectra, which is in line with the findings of Morris [152]. Moreover, the described procedure is faster.

Further, when training of the classifier is performed with samples from multiple images, the feature sets of the individual images have to be combined. First, the datasets have to be recalibrated to correct for potential peak shifts between sets. This is done by performing hierarchical clustering on the peak positions [212, 96] using an optimized cutoff value of 0.2 Da (for more details refer to section 5.2.5). After correction, the single feature sets are merged. Whenever a member of the merged feature list does not occur in the feature list of a dataset, we assume the corresponding intensity value to be zero [140].

Experiments were conducted using an in-house implementation of the random forest classifier and the vector-valued median algorithm; the in-house belief propagation code is based on Murphy's Bayes Net Toolbox [154]. Iterations were stopped as soon as the maximum change in local beliefs was below a given threshold. The regularization parameter was optimized manually.

4.4. Results

- Experiment 1: We randomly chose (labeled) samples from all S-slices for training and testing of the random forest. Ten-fold cross validation was used, i.e., nine out of ten subsets were used for training, and the remaining one was used for testing. Results can be found in table 4.1.

- Experiment 2: We trained the classifier with samples from all but one S-slice, and the labeled samples of the remaining S-slice were used for testing. Tissue sample preparation for MSI analysis was done individually for each slice, potentially decreasing the classifier's performance. Sensitivity and PPV estimates were calculated with respect to the label maps, and results are shown in table 4.2 and figure 4.4.

- Experiment 3: Random forest performance was tested on the T1 slice after training with samples from all six S-slices. Results are displayed in table 4.3 and figure 4.5.

- Experiment 4: Results concerning the last research question are given in table 4.4 and figure 4.6. Both smoothing methods required approximately the same computation time and resulted in similar SE and PPV estimates. One advantage of the vector-valued median is that in contrast to an MRF with Potts potential as described here, it can be used to directly smooth the probability maps instead of the crisp classification maps. Thus, smoothed versions of the soft and the crisp maps can be obtained.

- Experiment 5: Results concerning the feature importance are given in table 4.5 and figure 4.7.

4.5. Discussion

- Experiment 1: We first used 198 samples per class corresponding to 33 samples per class and slice, for which we obtained estimates for sensitivity and positive predictive value slightly above 90% (see table 4.1). With a total of 1,188 samples per class (i.e., 198 per class and slice), we gained a further increase of roughly 1.5%. Note that especially for the gelatin and interface regions, only few labeled samples are available (see section 4.3.1 and label maps in figure 4.3) and that these samples might have to be replicated in order to obtain a balanced training set when training is done with many samples [44]. In this scenario, the classifier is prone to overfitting to the training data. However, the results obtained after training with only 33 samples per class and slice (which is significantly below the number of available samples) and the results for classes for which hundreds of

		tissue class					
samples/class	measure	necrotic	viable	gelatin	interface	glass/hole	mean
198	SE	88.3	92.5	95.4	97.1	94.2	**93.5**
	PPV	94.0	90.0	97.1	94.0	95.1	**94.0**
1,188	SE	91.8	94.5	97.1	98.5	94.5	**95.3**
	PPV	93.0	92.5	98.8	95.5	96.8	**95.3**

Table 4.1.: Experiment 1: Training and testing has been performed with samples from all six slices and results are shown for different numbers of samples per class. We conducted two experiments with 198 and 1,188 training samples per class, respectively. No post-hoc smoothing was performed. SE and PPV values were estimated by ten-fold cross validation. High SE and PPV values are obtained for 198 samples. More samples clearly increase both the sensitivity and positive predictive values.

labeled points exist (see for example viable, glass/hole) are reasonably good. Most problems occurred in separating necrotic and viable tissue, probably because these classes are more similar to each other than to gelatin or glass/hole. Nevertheless, we still obtained good classification results for these tissue types.

- Experiment 2: We note that the sensitivity and PPV rates were slightly lower than in experiment 1 where also some samples from the test slice had been used in the training. However, the sensitivity was still at about 90% and the PPV at $\approx 85\%$. This experiment shows that random forests can be robust with respect to experimental variability witnessed when different slices are considered which were experimentally prepared and processed separately ("technical repeats").

  The soft classification maps obtained with the random forest classifier (see figure 4.4) provide further insight into the composition of a tissue sample. Crisp classification maps fail to give insight on the information regarding the ambiguity or amount of uncertainty of a prediction at a pixel. In situations where a tissue class is dominated by another tissue class, the information on the distribution of the weaker class(es) is lost in the crisp classification map. In some areas, the (crisp) class assignment is obvious, e.g., in regions where the glass slide is visible. In contrast, the class decision in the lower right part of slice S3 is less clear since both necrotic and viable tumor have high probabilities. Since the spatial transition between different stages of tumor development can be continuous, this information can be highly relevant in the treatment of the cancer. Imposing a hard threshold would suppress the subtleties of the true distribution, and a small shift of the threshold might lead to widely differing crisp classification maps which is undesirable.

- Experiment 3: We trained the classifier with 198 samples per class and slice from

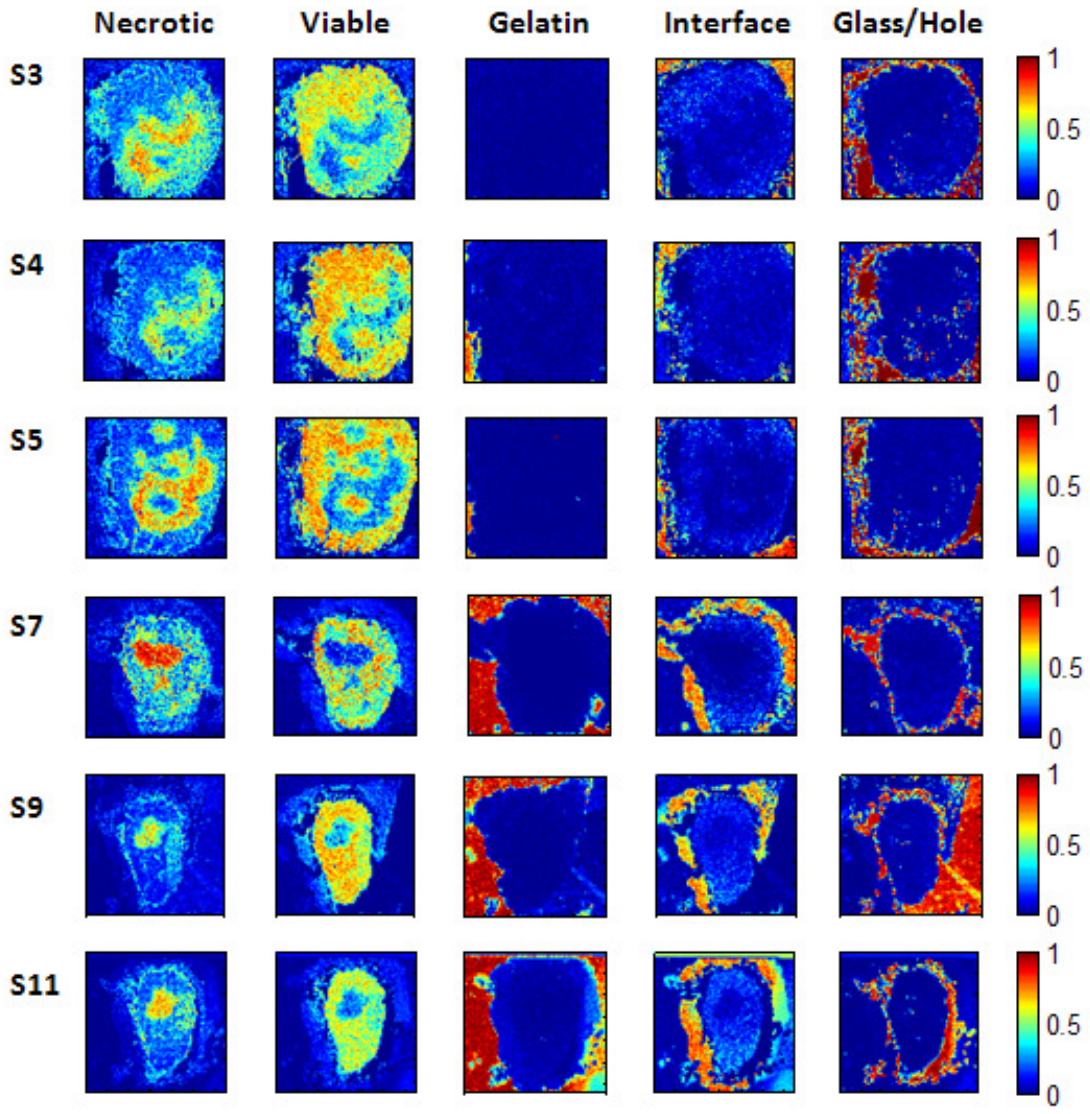

Figure 4.4.: Experiment 2: Soft classification maps for the S-slices. The probability maps (128×128 pixels in size) show the distribution of the five different tissue classes/regions and are nicely correlated with the label maps in figure 4.3. See text for details.

slice	measure	tissue class necrotic	viable	gelatin	interface	glass/hole	mean
S3	SE	92.8	82.7	-	89.7	89.5	**88.7**
	PPV	57.6	96.8	-	41.9	98.6	**73.7**
S4	SE	64.7	98.3	-	89.9	97.1	**87.4**
	PPV	95.6	85.7	-	98.3	86.3	**91.4**
S5	SE	94.7	91.3	86.7	-	98.1	**92.7**
	PPV	86.7	96.8	100	-	99.4	**95.8**
S7	SE	99.4	75.4	99.1	99.6	89.1	**92.5**
	PPV	30.4	99.7	99.5	94.8	99.5	**84.8**
S9	SE	81.0	96.2	84.0	96.4	97.2	**90.1**
	PPV	54.0	97.3	99.4	60.0	97.0	**81.5**
S11	SE	96.4	90.4	87.4	99.1	91.0	**92.8**
	PPV	40.0	98.7	99.4	68.3	99.7	**81.2**

Table 4.2.: Experiment 2: Training has been performed with samples from five slices, and the remaining one was used for testing. We trained the forest with 198 samples per class and slice, and no post-hoc smoothing was performed. The dash indicates that no labels are available for a certain slice/class combination. Average SE and PPV values are mostly between 80 and 95 percent. We see that a random forest trained with samples from five out of the six slices generalizes reasonably well and can be used to classify the test slice.

all six S-slices. When classifying T1, we obtained reasonably accurate SE and PPV estimates close to 90% (see table 4.3), indicating that the classifier is reliable even in situations where training and testing is performed with samples from different tumors. The slightly reduced sensitivity value for the necrotic class results from the fact that the classification result for this class is speckled. It is unclear if this has biological reasons or has to be considered a misclassification. Apart from that, the visual comparison between the stained slice and the classification result in figure 4.5 clearly underlines the good performance of the algorithm.

- Experiment 4: With respect to the gold standard, post-hoc smoothing with the proposed MRF and the vector-valued median algorithm significantly improved the classification results. We note however that in general, manual labeling builds on an underlying assumption of homogeneity; in addition, overly smooth labels may be obtained in situations where it is challenging for a human expert to mark every single differentiation that can be seen in a tissue sample. In our comparison, this favors the post-processing step. The salt-and-pepper noise was efficiently removed leading to a better highlighting of the shapes of the different tissue types (see figure 4.6). This observation is confirmed by the sensitivity and PPV estimates displayed

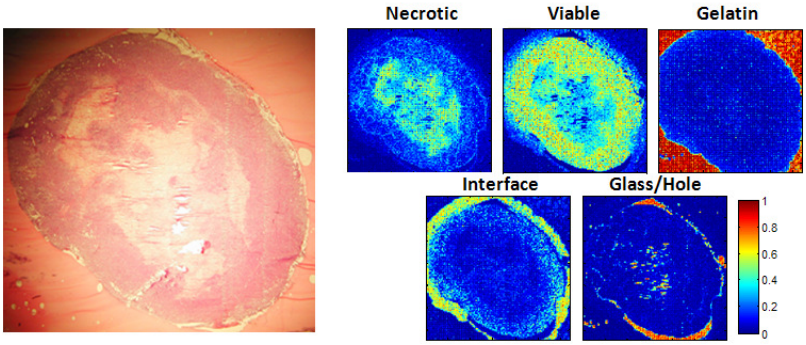

Figure 4.5.: Experiment 3: Classification of the T1 set after training on the S-slices. The similarity between the (parallel) stained slice (left) and the classification maps (right) is apparent. The necrotic part in the middle is well detected by the classifier. As expected, the interface part surrounds the viable part and is itself surrounded by gelatin. The holes are also visible in the stained slice.

		tissue class					
slice	measure	necrotic	viable	gelatin	interface	glass/hole	mean
T1	SE	71.9	91.6	99.9	90.3	99.6	**90.4**
	PPV	94.4	84.0	97.6	93.4	93.4	**92.5**

Table 4.3.: Experiment 3: Training has been performed with samples from all six slices of the first tumor (S3,S4,S5,S7,S9,S11), whereas testing was done on the T1-slice of the second tumor. We chose 198 samples per class and slice for training, and no post-hoc smoothing was performed. The results indicate that the classifier is reasonably accurate even if training and testing is performed on the same tumor type in different individuals.

in table 4.2. The classification results in figure 4.6 are well correlated with the stained images and label maps in figure 4.3.

Removing salt-and-pepper noise is beneficial if these structures arise from artifacts of the acquisition process like low signal to noise ratios at certain spatial locations. In these cases, isolated misclassifications can be efficiently corrected by the proposed methods leading to a clearer visualization of the main structures in the data. However, if these structures do indeed have a biological significance, i.e., different content in the tissue sample, oversmoothing is a problem as it may lead to a loss of valuable information. The decision of whether smoothing should be applied depends on the research question as well as on the quality of the data. In our experiments, smoothing led to increased sensitivity and PPVs with respect to our gold standard label maps.

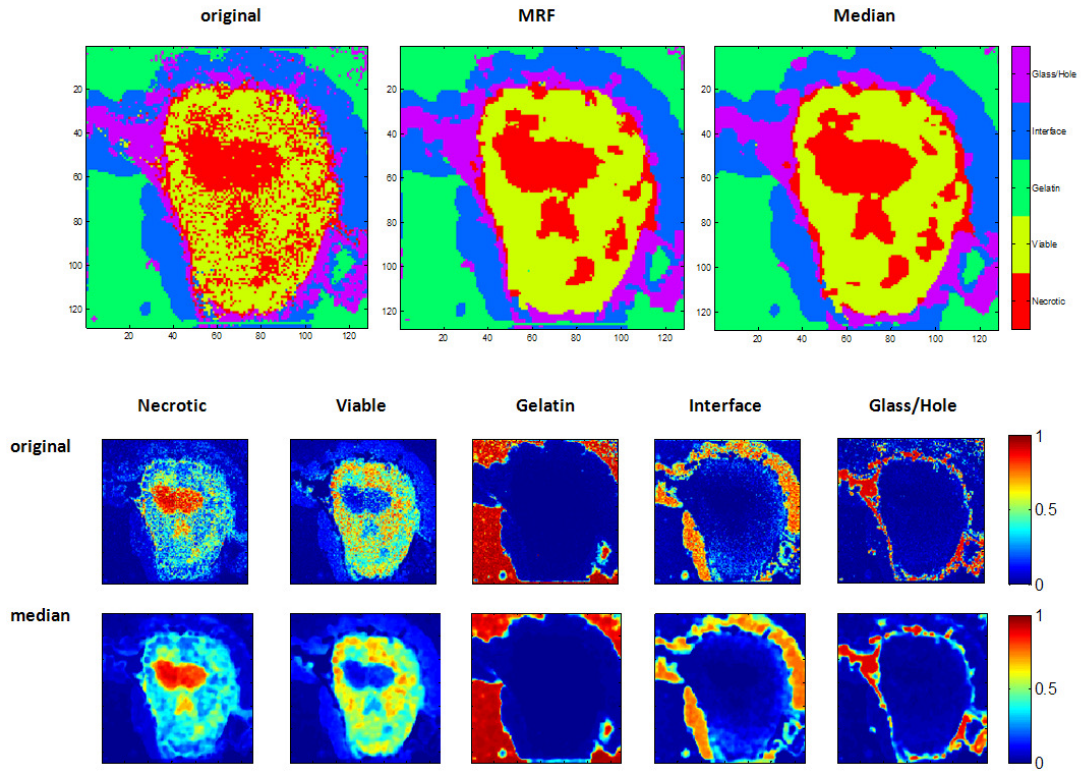

Figure 4.6.: Experiment 4: classification results for slice S7 after training with all *other* slices. The top row shows the crisp classification results: on the left the random forest result is shown, in the middle the result after post-hoc smoothing with MRFs ($\lambda = 0.01$) and on the right the result after applying the vector-valued median. The classification results are very close to the respective label maps, shown in figure 4.3. The smoothing efficiently removes the salt-and-pepper noise. Smoothed versions of the soft classification maps can be obtained with vector-valued median-filtering, and results are shown in the bottom rows.

slice	smoothing	measure	tissue class necrotic	viable	gelatin	interface	glass	mean
S4	none	SE	64.7	98.3	-	89.9	97.1	87.4
		PPV	95.6	85.7	-	98.3	86.3	91.4
	MRF	SE	83.6	100	-	93.6	98.8	**94.0**
		PPV	100	93.4	-	100	92.5	**96.4**
	VVM	SE	84.5	100	-	94.4	99.0	**94.5**
		PPV	100	95.1	-	100	96.6	**97.2**
S7	none	SE	99.4	75.4	99.1	99.6	89.1	92.5
		PPV	30.4	99.7	99.5	94.8	99.5	84.8
	MRF	SE	100	92.7	99.7	99.8	94.8	**97.4**
		PPV	58.9	99.7	99.5	97.8	100	**91.2**
	VVM	SE	100	91.3	100	100	95.8	**97.4**
		PPV	55.2	99.9	99.9	98.4	100	**90.1**

Table 4.4.: Experiment 4: The table shows the classification results obtained for slices S4 and S7 after training of the random forest with samples from all other S-slices. Post-hoc smoothing of the classification maps with Markov random fields (MRF) and vector-valued median filtering (VVM) clearly improves the classification result by more than 5% in sensitivity and positive predictive value. However, great care has to be taken when smoothing is applied, see text for details.

- Experiment 5: The random forest was trained as described in experiment 3. Among the most important features (mass channels) that were identified by the permutation accuracy criterion are $m/z = 114.9Da$ (indium) and $m/z = 184.0Da$ (phosphocholine [71]). The former is identified to be important for classifying samples of the glass region. This is plausible since the glass slides were indium-coated prior to MSI analysis (see Data section) and indium is dominant in these areas. The latter has previously been found to play an important role in the discrimination of necrotic and viable tissue and the interface region (see section 3.5.2) as well as in the metabolism of breast cancer cells in general [88, 87]. Note that a high score for a given feature and class combination does not necessarily imply that the respective mass channel shows high intensities for that class. A particularly high (or low) score rather indicates that the respective feature is important for the classification of samples belonging to that class (i.e., not belonging to other classes). As a consequence, some markers occur in the top lists of multiple tissue types. As can be seen from the channel images in figure 4.7 the mass channels with high importance scores show informative spatial distributions which are nicely correlated with the different areas in the label maps (cf. figure 4.3).

  Note however that these results should only be considered as first indicators for

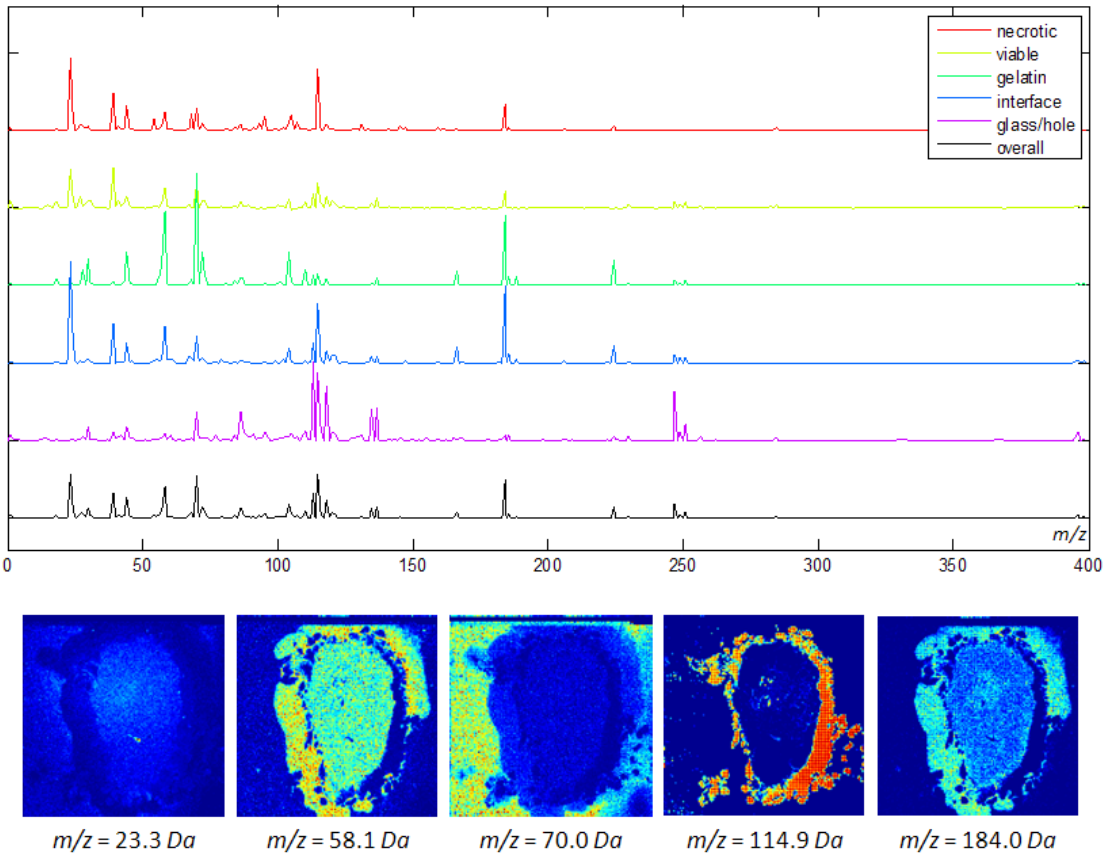

Figure 4.7.: Experiment 5: The upper diagram shows the permutation accuracy feature importance scores for the five tissue classes as well as the overall importance score. The five most important features are listed in table 4.6 and include indium ($m/z = 114.9Da$) and phosphocholine ($m/z = 184.0Da$). The corresponding channels for tissue sample S11 are plotted in the bottom row.

	tissue class				
rank	necrotic	viable	gelatin	interface	glass/hole
1	23.3 (9.2)	39.0 (5.2)	70.0 (14.4)	23.3 (13.1)	112.9 (10.2)
2	114.9 (8.0)	23.3 (5.0)	58.1 (9.7)	184.0 (9.9)	114.9 (8.7)
3	39.0 (4.9)	114.9 (3.2)	184.0 (9.1)	114.9 (7.8)	118.0 (7.0)
4	184.0 (3.5)	70.0 (3.1)	44.1 (4.5)	39.0 (5.1)	246.8 (6.4)
5	44.1 (3.2)	58.1 (2.7)	104.1 (4.4)	58.1 (4.8)	136.8 (4.2)

Table 4.5.: Experiments 5: For each tissue class the five most important features with respect to the permutation accuracy criterion are listed by their m/z-position in Da and their corresponding importance score (given in brackets, scaled by 10^2). The overall most important features and their interpretation (if available) are given in table 4.6 (also cf. figure 4.7).

rank	m/z	interpretation
1	114.9 (5.8)	indium
2	23.3 (5.6)	sodium
3	70.0 (5.6)	
4	184.0 (5.0)	phosphocholine
5	58.1 (4.1)	

Table 4.6.: Experiments 5: The overall most important features (score given in brackets) and their interpretation (if available). Also cf. figure 4.7.

potential biomarkers. Besides standardized tissue sample preparation and stable acquisition conditions, a robust detection of biomarkers also requires a more diverse dataset that comprises genetic variability.

The methodology described in this chapter can be applied to MSI data of different resolution and quality. Given sufficiently high resolved data and an adequate number of training examples, the random forest classifier could also be used on the cell level. This would enable classification, i.e., digital staining, of single cells, providing even further insight into the composition of tissue samples.

4.6. Conclusion

Our study gives clear evidence that digital staining may be a powerful complement to chemical staining techniques. We have introduced post-processed random forests for the classification of MSI data. Experiments on an animal model of human breast cancer grown in mice suggest that this classifier is well suited for automated annotation of MSI data. With the proposed methodology, we were able to separate necrotic tissue, viable

tumor, gelatin, tumor interface and glass/hole areas under the following experimental conditions: High sensitivity rates ($\approx 90\%$) and positive predictive values ($\approx 85\%$) have been obtained when training and testing was performed with samples from different slices of a single tumor. Similar performance was observed when samples from two different tumors of the same cell line were used for training and testing. Further experiments are required in which the presented methods are evaluated on data featuring genetic variation (see chapter 5).

The soft classification output of the random forest classifier can give valuable insight into the composition of tissue samples, and the permutation accuracy criterion yields discriminative features for the classification. We have demonstrated that spatially smoothing the crisp and soft classification maps with Markov random fields and vector-valued median filtering significantly improves the classification result, increasing sensitivity by approximately 3% in the examples shown.

Chapter 5

Differential Diagnostics of Breast Cancer using MALDI MSI: The Role of Preprocessing and Technical Variability

In the preceding chapter, we have demonstrated that digital staining with mass spectrometry imaging (MSI) is a promising complement to traditional staining techniques. However, our study was limited since no genetic variability between patients was considered. In contrast, the data that is analyzed in pre-clinical applications of MSI typically stems from different patients and features both technical *and* biological variability[1] [73, 193, 192, 223].

In the following chapter we enhance and adapt the methods for pixel-based classification presented in chapter 4 and consider MSI data from a pre-clinical study with 30 patients that suffer from two different kinds of breast cancer. The mass spectrometric analysis is targeted at high mass ranges up to 25,000 Da. Thus, matrix assisted laser desorption/ionization (MALDI) MSI is used instead of secondary ion mass spectrometry (SIMS) MSI. Since MALDI data tends to be affected by a higher level of noise and much stronger baseline effects than SIMS data, the main focus lies on the selection of the preprocessing methods for baseline correction, normalization, peak picking and spectral alignment. In our study we compare almost 400 different preprocessing pipelines and demonstrate that

- the exact choice of the preprocessing methods and their parameterizations heavily influences the obtainable classification performance,

[1] that is biological variability between samples from different classes *as well as* between samples from the same class that stem from different patients

- highly accurate pixel-based classification of MALDI MSI data is still a difficult task, although the obtained classification accuracies of 80–85% are promising,
- the within class variance (attributable to both technical and biological variability) can be very high in (pre-)clinical MSI studies.

We begin this chapter with an introduction to the biological research question that motivates our study.

5.1. Introduction

With one million breast cancer cases newly diagnosed and over 400,000 deaths a year, breast cancer is still the leading cause for cancer deaths among women [163]. A clinically important subgroup of breast cancer patients is termed triple negative and shows underexpression of the oncoprotein HER2 (human epidermal growth factor receptor 2). This subgroup is associated with a more aggressive tumor growth and therefore a poorer prognosis. Special targeted therapies are able to significantly improve prognosis of those patients. Thus, the determination of the HER2 status plays a key role for the further course of the therapy [230]. Recently, Rauser et al. showed [171] that MALDI MSI signatures can be used to predict the HER2 status of breast cancer tissue. Using histochemistry, they identified cancer regions with and without HER2 underexpression and trained supervised classifiers on the *mean spectra* corresponding to those regions. Accuracies of 85–90% were reported and a set of biomarker candidates that significantly correlate with HER2 expression was identified. However, the classification of individual pixels of an MS image was not considered. As described earlier (cf. chapter 4), pixel-based classification of MSI data has several advantages. Most importantly, in our application it can provide better stratification of HER2 positivity, e.g., by predicting the fraction of positive tumor cells in a sample (cf. IHC guidelines [230]). Whereas the method by Rauser et al. results in a global classification (i.e., the complete dataset is HER2 positive or negative), pixel-wise classification yields a spatially resolved result (see figure 5.1 for an example). This is advantageous if the tissue is heterogeneous and various tissue types are present. Moreover, once the classifier is trained it can directly be applied to newly acquired data. In contrast, if the classification is based on mean spectra over areas, the expert is required to define regions of interest that determine which spectra have to be averaged prior to classification.

5.2. Materials and Methods

As described in chapters 2 and 4, mass spectra require preprocessing before meaningful statistical analysis can be applied. Typical steps include baseline correction, normalization, peak picking, and spectral alignment with which one tries to remove or at least

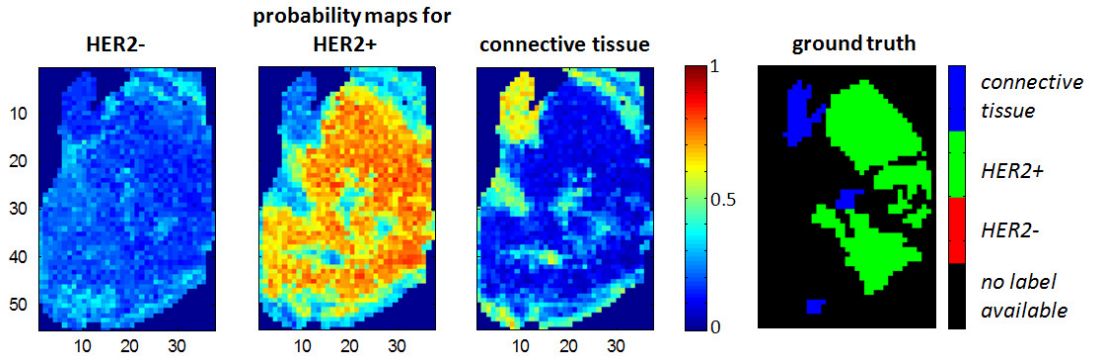

Figure 5.1.: Classification of HER2 positive vs. HER2 negative vs. connective tissue. Training of the classifier was performed with all datasets (cf. table 5.1) but the one used for testing. On the left, probability maps for the three classes are shown. The ground truth label map on the right was obtained using histochemistry.

mitigate the effects caused by technical variability in sample preparation and/or data acquisition [33]. Depending on the selected mass spectrometer and ionization technique, the importance of the individual preprocessing steps can vary. For instance, the SIMS data used in our previous study (cf. chapter 4) exhibits nearly no baseline effects and is affected by a low level of noise only. The choice of the baseline correction method thus had no significant influence on the classification performance. We will see that this is fundamentally different for the MALDI data analyzed next (see section 5.3.1 and cf. figure 5.5).

5.2.1. Choice of Preprocessing Methods

Obviously, many different methods can be applied to realize the individual steps of the preprocessing "pipeline". Unfortunately, no standard workflow exists and it is difficult to select the "right" algorithms and parameters since usually no ground truth (e.g., the true baseline) is available for evaluating *individual* steps (e.g., the baseline correction method). Only a compound ground truth in the form of labels of the individual pixels is given. Thus, we can only evaluate the performance of complete pipelines, e.g., by comparing their classification accuracies. This is especially true since the individual preprocessing methods may also interact. For instance, the choice of the baseline correction approach might influence the performance of the normalization methods.

Rauser et al. [171] employed ClinProTools 2.2 (Bruker Daltonik GmbH, Bremen, Germany) to preprocess the MSI data used in their study. Unfortunately, only a limited number of spectra can jointly be analyzed with this software. Furthermore, the choice of preprocessing methods is restricted to the built-in algorithms. Therefore, we decided

to export the raw MSI data to text files using FlexImaging 2.1 (Bruker Daltonik GmbH, Bremen, Germany). All subsequent statistical analysis was performed with algorithms that were implemented in MATLAB and C++. To identify promising combinations of preprocessing methods, a multitude of different pipelines was compared with respect to the resulting classification accuracy. Details for the methods used in the individual pipeline steps are provided below.

5.2.2. Baseline Correction

Three baseline correction approaches were compared to the default strategy: (1) no baseline correction. (2) Top-hat filtering [195, 189] is a popular approach that performs a morphological opening on a local neighborhood. Tuning of its only parameter–the half-width w of the local neighborhood–is not trivial. If the neighborhood size parameter is chosen too large, the baseline cannot be removed completely, especially if the baseline exhibits strong decay. On the contrary, a small neighborhood size might result in feature loss [227][2]. (3) Asymmetric least squares fitting for MALDI spectra as proposed by Eilers [68] is an approach that considers the asymmetry caused by peaks in fitting a smooth baseline to the observed data in a least squares sense. The level of asymmetry α and the smoothness λ of the baseline are controlled by user-defined parameters. For instance, a high value for α corresponds to a high penalty for the difference between peaks and the baseline. (4) Spline fitting fits a monotonously decreasing spline to an observed spectrum to estimate its baseline [59]. This method proved to be rather insensitive to its only parameter, the number of knots κ. In all cases, after baseline correction all negative intensities were considered noise and set to zero.

Many more baseline correction methods have been proposed in the literature (e.g., [227, 140, 52]) that cannot be considered here. For instance, Williams et al. [227] suggest to estimate the baseline by fitting an exponential function to an observed spectrum. To decrease the influence of peaks that may bias the estimated baseline toward higher intensities, they first subsample the spectrum in such a way that regions with less peaks are more densely sampled. The exponential is then fitted to the reduced set of points. However, in our experiments on example data from the HER2 study the estimated baselines still largely exceeded the observed spectra and we encountered instabilities when fitting the exponentials to some of our data. We thus decided to exclude this method from our comparison.

5.2.3. Normalization

Arguably, the most popular normalization method is to (2) divide each spectrum by its total ion count (TIC). Alternatively, we normalized each spectrum with the (3) global mean standard deviation method (mean-sd), which previously provided good results

[2] A similar effect can be observed for the sliding median approach, see figure 2.5c.

[146]. In this method, the spectrum-wise mean is subtracted from each channel and the result is divided by the standard deviation of the spectrum. Both approaches were compared to the default, that is (1) no normalization.

5.2.4. Peak Picking

The MALDI MSI data analyzed in this study is not isotope-resolved and the (spectral) bin sizes increase from 1.2 Da (low mass range) up to 3.6 Da (high mass range). Thus, advanced peak pickers like NITPICK [173] or THRASH [102] could not be applied, and peak picking was based on local maximum detection instead. Visual inspection suggested that severe peak shifts only occur between datasets but not within individual datasets[3]. To extract peaks, we first baseline corrected and normalized all spectra within a dataset S_i individually. Using a Savitzky-Golay filter [184] (4-degree polynomial, window size of 10, 3 cycles) we then calculated a smoothed version of the mean spectrum over all acquired spectra in S_i [152]. Next, local maxima M_j of the smoothed mean spectrum were identified from which the algorithm descended to the left and right until it hit the two neighboring local minima $m_{j,1}$ and $m_{j,2}$. M_j was added to the peak list P_i of dataset S_i if and only if the maximum height distance of M_j to its two neighboring minima exceeded a given threshold T_i. We then either used (1) intensities or (2) areas under peaks (that is the integral from $m_{j,1}$ to $m_{j,2}$ over the measured spectrum), to obtain the intensities corresponding to the selected peak positions.

5.2.5. Spectral Alignment

Spectral alignment was necessary since peak shifts between datasets were observed. Our alignment method consisted of two steps: First, we identified the position of the highest peak in the peak lists of all datasets (that happened to be the matrix peak, which is present in all spectra) and performed a linear shift alignment to ensure that the base peak lies at the same m/z-value in each spectrum. We then applied hierarchical clustering [212, 96] to the unified peak list $\tilde{P} = \{P_1, \ldots, P_N\}$ using the complete linkage scheme: First, a matrix of m/z distances between peaks was constructed. Then, the algorithm iteratively merged those peaks (or peak clusters) that after merging had the smallest maximum within cluster distance between elements[4]. To prevent merging of very distant peaks, the iterations were stopped as soon as a predefined distance threshold was exceeded. We used an adaptive threshold D that was defined in parts per million (ppm) to account for the fact that the mass accuracy of the data decreases with increasing mass over charge. Note that in its natural form, the dimension of the distance matrix

[3] Theoretically, small shifts up to 300 ppm (parts per million) may occur in a single set (according to the instrument manufacturer).

[4] Note that in the alignment step peaks from different datasets as well as peaks that are close and stem from the same dataset may be clustered.

equals the number of peaks in $\tilde{P}$ squared. Constructing such a matrix in memory is usually not possible. Instead, we propose to employ a sparse distance matrix that only contains information on peak pairs which are closer than D ppm.

After clustering of the peaks in $\tilde{P}$, a *master peak list* was estimated that contained the centers of mass for all peak groups/clusters. We discarded peak groups that occurred in less than a user-defined fraction of datasets to remove random noise peaks. This occurrence cutoff value has to be chosen carefully to avoid the loss of relevant peaks.

5.2.6. Classification

In contrast to support vector machines (SVMs) [188], random forests [25] (cf. section 4.2.1) are rather robust with respect to the choice of their hyperparameters and require only few parameter tuning. Thus, we restricted the following systematic analysis of preprocessing methods to the random forest classifier. We note, however, that in additional experiments, the SVM classifier performed approximately equally well, given that a suitable feature selection was used (see also table 5.4)[5].

5.3. Experiments

5.3.1. Data

The analyzed tissue stems from primary breast cancer patients with invasive ductal carcinoma. After resection, samples were snap-frozen and stored in liquid nitrogen. Prior to MALDI MSI analysis, the samples were cryo-sectioned, mounted on indium-tin-oxide coated glass slides, washed in 70% and 100% ethanol, dried under vacuum, and matrix-covered [171]. Data was acquired with a Bruker Ultraflex III MALDI-TOF/TOF in linear mode with a mass range of 2,400–25,000 Da. The lateral resolution was set to $200\mu m$. In total, 30 datasets from 30 different patients were acquired (known as the "discovery set" in [171]). All tumorous samples from a single dataset share the HER2 status: 12 datasets feature cancer tissue with a negative and 14 with a positive status. 4 datasets contain no cancer regions but only connective tissue[6]. The HER2 status was assessed using immunohistochemistry (IHC) and in situ hybridization (FISH) [230] (see [171] for details). 15 of the 26 sets featuring cancer also contain connective tissue. At the same time, the individual sets show different levels of progesterone (PR) and estrogen (ER) expression [171]. An overview over all sets is given in table 5.1.

[5] Empirically, we observed that the random forest classifier merely improved with feature selection. In contrast, performance of the SVM was highly dependent on adequate feature selection. Filtering of irrelevant features led to significant performance boosts.

[6] The HER2-positive or -negative samples from these sets [171] were removed in our study since their labels were less reliable than those of the other samples (personal communication).

Set ID	HER2 status	number of labels HER2-	HER2+	CT	Set ID	HER2 status	number of labels HER2-	HER2+	CT
1	-	100	0	0	16	-	594	0	0
2	+	0	30	0	17	+	0	69	11
3	+	0	27	12	18	+	0	111	126
4	+	0	244	0	19	+	0	506	89
5	+	0	84	45	20	-	425	0	21
6	-	23	0	0	21	-	351	0	0
7	+	0	458	28	22	+	0	161	11
8	+	0	198	0	23	+	0	15	10
9	N/A	0	0	151	24	-	163	0	24
10	-	257	0	53	25	+	0	324	21
11	-	138	0	0	26	+	0	198	14
12	N/A	0	0	106	27	N/A	0	0	5
13	+	0	65	5	28	-	219	0	96
14	-	119	0	0	29	N/A	0	0	33
15	-	45	0	0	30	-	51	0	0

Table 5.1.: Number of available samples for the different datasets (1–30) and classes (HER2-, HER2+, connective tissue).

5.3.2. Experiment 1: Comparison of Different Pipelines

In the first experiment, we identified which combinations of preprocessing methods yield the best results. Since no ground truth for evaluating the individual preprocessing steps was available, performance was evaluated by comparing the classification accuracies obtained for different preprocessing pipelines using a random forest classifier with 400 trees and 6,000 class-balanced samples (cf. chapter 4).

Evaluation Criteria. We chose the F1 score as a unified measure that integrates the popular metrics precision and recall into a single number. Recall (also termed sensitivity) is defined as $rec = \frac{TP}{TP+FN}$ where TP is the number of true positives and FN is the number of false negatives. Precision (also positive predictive value) gives the ratio of samples correctly classified as c_1 among all samples classified as c_1 and $prec = \frac{TP}{TP+FP}$ where FP is the number of false positives (see also section 4.3.3). The F1-score F is the harmonic mean of these measures:

$$F = 2\frac{prec \cdot rec}{prec + rec}. \tag{5.1}$$

Each dataset exclusively contains tumor tissue with a positive *or* negative HER2 status where some sets additionally contain connective tissue. For each combination of methods, that is for each pipeline, we performed 100 repeats of a leave-$(3+3)$-out cross

validation. Therefore, in each repeat we randomly selected 3 datasets with positive and 3 datasets with negative HER2 status as test sets, and trained the classifier on all remaining sets. The random generator was seeded to ensure that all combinations/pipelines were compared on exactly the same partition. The performance of a pipeline was measured by classifying all pixels in the test sets. The obtained results were averaged over all repeats. Note that since not all datasets feature the same number of labeled samples, some sets have higher influence on the averaged results than others.

We furthermore calculated an estimate DS for the dataset-wise classification rate. A dataset was considered HER2 positive if the majority of the cancer pixels was classified as positive and negative otherwise. DS was estimated as the fraction of datasets that were classified correctly.

Selected Methods and their Parameterizations. The evaluated preprocessing pipelines consisted of four consecutive steps. To make the comparison feasible, we had to restrict to a limited number of parameterizations of the selected algorithms. We used a total of sixteen baseline correction strategies (none, top-hat filter with window half-width $w \{50, 100, 250, 500, 1, 000\}$ bins, spline-approach with $k = 5$ knots, method by Eilers with (all combinations of) smoothness parameter $\lambda = \left\{10^7, 10^8, 10^9\right\}$ and asymmetry parameter $\alpha = \{0.001, 0.01, 0.1\}$), three normalization schemes (none, TIC, mean-sd), two peak picking strategies (intensities, areas) where we set the threshold such that the 250 highest peaks in each dataset were picked, and four alignment parameterizations for the hierarchical clustering (500, 1,000, 2,000, 4,000 ppm), and combined them with a random forest classifier (parameterization see above). In total, 384 different pipelines were compared.

5.3.3. Experiment 2: Results on Individual Datasets

Whereas the first experiment was performed to get *global* performance estimates for each pipeline, the second experiment investigated the classification rates that were obtained on the *individual* datasets. We were especially interested which datasets were classified well and which were classified poorly. For this experiment, we selected one of the best performing pipelines from the first experiment (Eilers' baseline correction with $\lambda = 10^9$ and $\alpha = 0.01$, TIC normalization, areas under peaks peak picking and an alignment window of 4,000 ppm; see table 5.2) and inspected the average confusion matrices of the 30 sets that were obtained in the first experiment.

5.4. Results and Discussion

5.4.1. Experiment 1: Comparison of Different Pipelines

The best performing pipelines resulted in mean recall values of around 82–85% and precision values of 80–82%, yielding F1 scores between 81 and 83% (all values averaged

methods				recall			precision			∅			
bl	norm	a/i	align	-	+	CT	-	+	CT	DS	rec	prec	F1
15	TIC	int	4,000	85.6	91.9	76.9	95.3	87.4	62.3	82.2	84.8	81.7	83.2
16	TIC	int	4,000	86.1	90.1	76.4	92.4	87.4	64.9	79.8	84.2	81.6	82.8
14	TIC	int	4,000	84.7	89.8	78.7	93.2	86.2	64.1	80.3	84.4	81.2	82.8
7	TIC	int	2,000	88.9	86.5	74.1	89.3	89.7	62.9	79.3	83.2	80.7	81.9
9	TIC	int	4,000	80.9	90.4	78.3	93.9	82.7	64.0	78.5	83.2	80.2	81.7
16	TIC	ar	4,000	84.3	91.7	71.6	91.8	86.8	63.8	81.2	82.5	80.8	81.7
15	TIC	ar	4,000	90.1	92.0	63.7	92.7	90.5	60.5	86.5	81.9	81.2	81.6
13	TIC	ar	4,000	84.4	89.0	74.2	89.8	86.4	65.1	81.5	82.5	80.4	81.5
13	TIC	int	4,000	83.4	87.8	77.8	90.2	85.3	64.4	78.3	83.0	79.9	81.5
4	TIC	int	4,000	81.9	90.4	76.2	93.5	83.7	62.6	77.2	82.8	79.9	81.4

Table 5.2.: Class-wise and averaged precision (prec), recall (rec) and F1-score values for the best 10 combinations of methods (that is pipelines) with respect to their F1-score. In our experiments, the best pipelines relied on TIC normalization and rather large windows for alignment, achieving precision, recall and F1 scores of 80–85%. The first four columns of the table encode the selected methods: Baseline correction (bl): 1 none, 2–6 Top-hat ($w = 50|100|250|500|1,000$), 7 Spline ($\kappa = 5$), 8–16 Eilers ($\lambda = 10^7$ with $p = 0.001|0.01|0.1$, $\lambda = 10^8$ with $p = 0.001|0.01|0.1$, $\lambda = 10^9$ with $p = 0.001|0.01|0.1$); normalization (norm) by total ion count (TIC) or mean standard deviation (mean-sd); peak-picking (a/i) using intensities (int) or areas under peaks (ar), and spectral alignment (align) with different window sizes given in parts per million (ppm). The last ten columns report the classification accuracies that were obtained with a random forest classifier (averaged over 100 repeats) where + and - are short for HER2+ and HER2-, and DS is the dataset wise classification rate. Note that the differences of the best 10 methods regarding their F1-score are not significant at the $p = 0.05$ level.

over the three classes, see table 5.2), and dataset-wise classification accuracies of around 84% were achieved. Thus, the case-wise classification rates were approximately 5% lower than the ones reported by Rauser et al. [171]. Before we discuss this performance loss in detail (see section 5.4.3), we focus on other aspects: First of all, we note that the choice of the preprocessing pipeline highly affected the classification performance on both the pixel and casewise level (cf. figure 5.2). In extreme cases, differences of up to 30% in precision, recall and F1-score were observed (cf. figure 5.2 top). Even if only pipelines that performed both baseline correction and normalization were considered, differences were as high as 15–20%. Often, even small changes in the parameterization of a method highly affected the classifier's performance. To further illustrate this, we now discuss a selection of twelve of the 384 pipelines (cf. table 5.3). In the following, whenever we use the term "significant", we refer to a significance level of $p = 0.05$.

pipeline	baseline correction	norm	a/i	align	rec	prec	F1
1	**Eilers**($\lambda = \mathbf{10^9}, \alpha = \mathbf{0.01}$)	**TIC**	int	**4,000**	**84.8**	**81.7**	**83.2**
2	**Top − hat**($\mathbf{w = 50}$)	**mean-sd**	int	**500**	71.5	71.2	71.4
3	Eilers ($\lambda = \mathbf{10^7}, \alpha = \mathbf{0.01}$)	mean-sd	int	2,000	**79.6**	**75.6**	**77.5**
4	Eilers ($\lambda = \mathbf{10^7}, \alpha = \mathbf{0.1}$)	mean-sd	int	2,000	76.0	72.6	74.3
5	Eilers ($\lambda = \mathbf{10^8}, \alpha = \mathbf{0.1}$)	mean-sd	int	2,000	**79.4**	**76.3**	**77.8**
6	Top-hat ($w = 1,000$)	**mean-sd**	int	4,000	78.6	76.7	77.6
7	Top-hat ($w = 1,000$)	**TIC**	int	4,000	**81.1**	**79.0**	**80**
8	Top-hat ($w = 1,000$)	**none**	int	4,000	69.7	71.1	70.4
9	**Top − hat**($\mathbf{w = 500}$)	**mean-sd**	areas	2,000	**79.7**	**77.5**	**78.6**
10	**Top − hat**($\mathbf{w = 500}$)	**TIC**	areas	2,000	75.3	75.6	75.5
11	**Spline**($\kappa = \mathbf{5}$)	**mean-sd**	areas	2,000	77.7	74.8	76.2
12	**Spline**($\kappa = \mathbf{5}$)	**TIC**	areas	2,000	**80.4**	**78.9**	**79.6**

Table 5.3.: A selection of 12 of the 384 analyzed preprocessing pipelines divided into four groups. Our experiments show that changing individual preprocessing steps (changes are highlighted in bold) can have a significant impact on the quality of the result (the best results are marked in bold). For instance, given pipeline 3, increasing parameter α leads to a significant performance decrease (pipeline 4). Refer to the text for further interpretations and to table 5.2 for definitions of the abbreviations.

Pipelines 1 and 2. Although both preprocessing pipelines seem to consist of a reasonable combination of methods, significant differences in classification accuracy can be observed, and pipeline 1 outperforms pipeline 2 by more than 10 percent in precision and recall.

Pipelines 3 to 5. Pipelines 3 to 5 only differ in the parameterization of the baseline correction approach. Recall that Eiler's method features two parameters that control the level of smoothness of the baseline (λ) and the asymmetry in the least squares fit (α) where the latter is used to weight the influence of the peaks. Whereas pipeline 3 yields comparably high precision and recall, pipeline 4 results in significantly lower values. Increasing α corresponds to shifting the baseline to a higher level which in this case might lead to some feature loss where peak intensities are low. Increasing the smoothness parameter λ seems to attenuate this effect, such that precision and recall rise again (pipeline 5)[7].

Pipelines 6 to 8. Pipelines 6 to 8 are identical except for the normalization methods employed. Here, TIC normalization significantly outperforms mean-sd normalization, which itself yields significantly better results than if no normalization is used.

Pipeline 9 to 12. However, although TIC normalization mostly outperformed mean-sd in our experiments, we also observed counter examples: Pipelines 9 to 12 demonstrate

[7]We note that the performance difference of pipelines 3 and 5 is not significant.

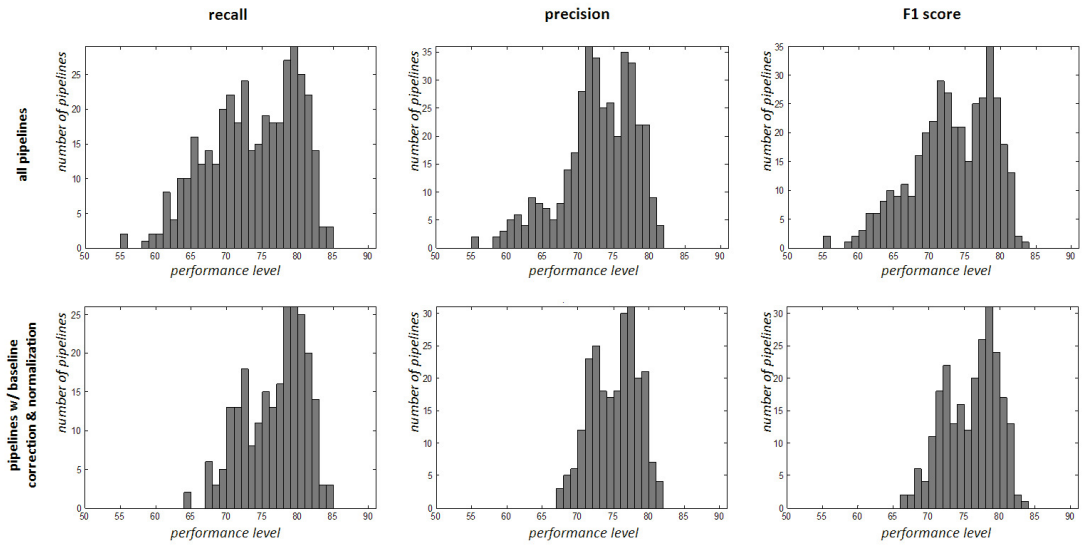

Figure 5.2.: Histograms of the performance distributions for the preprocessing pipelines that were compared in this study: The top row holds the histograms corresponding to the performances of all 384 pipelines. In the histograms in the bottom row pipelines that perform no baseline correction or normalization are ignored. We conclude that the classification result highly depends on the selection of the preprocessing algorithms.

how the selection of the baseline correction algorithm and the normalization scheme may interact. In this example, the spline baseline correction methods works significantly better if it is combined with TIC normalization than if it is combined with mean-sd. The opposite holds, if we use the top-hat baseline correction with parameter $w = 500$ (note that all other preprocessing steps were fixed). Slight changes to the individual preprocessing steps can again invert this behavior (see pipelines 6-8).

Obviously, one has to act with caution when generalizing these results to other MALDI MSI data. Nevertheless, some general trends became apparent in our experiments: The best performing pipelines *often* relied on methods that produced rather smooth baselines, used TIC normalization, employed peak picking using intensities, and relied on rather large ppm windows for aligning peaks. Our study further shows that the overall classification result can be extremely sensitive to the choice, combination and parameterization of the preprocessing methods. Selecting the "wrong" parameters or combination of methods can easily lead to a performance loss of 10% or more. This might have fatal consequences if the technique was applied in a clinical setting. Thus, we argue that simply using a popular standard combination of methods, e.g., top-hat filtering and TIC normalization (pipeline 10), is insufficient since it does not guarantee the best possible result. Instead, different pipelines should be evaluated in each study.

1 – 10017 (-)	-	+	CT
-	98.6	0	1.4
+	0	0	0
CT	0	0	0

2 – 1027 (+)	-	+	CT
-	0	0	0
+	0.4	29.6	0
CT	0	0	0

3 – 10639 (+)	-	+	CT
-	0	0	0
+	0.1	26.9	0
CT	0.5	2.9	8.6

4 – 10736 (+)	-	+	CT
-	0	0	0
+	0	244.0	0
CT	0	0	0

5 – 11593 (+)	-	+	CT
-	0	0	0
+	26.0	8.0	50.0
CT	0.7	0	44.3

6 – 12194 (-)	-	+	CT
-	0.1	6.3	16.7
+	0	0	0
CT	0	0	0

7 – 12298 (+)	-	+	CT
-	0	0	0
+	0.1	457.8	0.1
CT	0.1	23.5	4.4

8 – 12597 (+)	-	+	CT
-	0	0	0
+	1.0	135.3	61.8
CT	0	0	0

9 – 1260 (+)	-	+	CT
-	only in training set contains only ct		
+			
CT			

10 – 12909 (-)	-	+	CT
-	257.0	0	0
+	0	0	0
CT	41.7	0	11.3

11 – 13332 (-)	-	+	CT
-	117.9	20.0	0.1
+	0	0	0
CT	0	0	0

12 – 1377 (-)	-	+	CT
-	only in training set contains only ct		
+			
CT			

13 – 1437 (+)	-	+	CT
-	0	0	0
+	0	63.9	1.1
CT	0	0	5.0

14 – 16775 (-)	-	+	CT
-	108.8	0	10.2
+	0	0	0
CT	0	0	0

15 – 2008...3375 (-)	-	+	CT
-	7.2	37.8	0
+	0	0	0
CT	0	0	0

16 – 2094 (-)	-	+	CT
-	594	0	0
+	0	0	0
CT	0	0	0

17 – 21115 (+)	-	+	CT
-	0	0	0
+	0	56.0	13.0
CT	0	0.4	10.6

18 – 21147 (+)	-	+	CT
-	0	0	0
+	4.8	104.2	2.0
CT	21.8	9.8	94.3

19 – 2174 (+)	-	+	CT
-	0	0	0
+	0.6	504.8	0.6
CT	0.9	24.1	64.0

20 – 2632 (-)	-	+	CT
-	425.0	0	0
+	0	0	0
CT	21.0	0	0

21 – 4089 (-)	-	+	CT
-	349.9	0	1.1
+	0	0	0
CT	0	0	0

22 – 4152 (+)	-	+	CT
-	0	0	0
+	0	160.9	0.1
CT	0	1.1	9.9

23 – 4156 (+)	-	+	CT
-	0	0	0
+	0.8	13.9	0.3
CT	0.3	0.5	9.2

24 – 4247 (-)	-	+	CT
-	162.4	0	0.6
+	0	0	0
CT	11.9	0	12.1

25 – 4466 (+)	-	+	CT
-	0	0	0
+	0	324.0	0
CT	0	20.3	0.7

26 – 5559 (+)	-	+	CT
-	0	0	0
+	30.1	163.6	4.3
CT	9.8	1.5	2.7

27 – 6575 (-)	-	+	CT
-	only in training set contains only ct		
+			
CT			

28 – 6718 (-)	-	+	CT
-	118.2	77.7	23.1
+	0	0	0
CT	6.3	1.3	88.6

29 – 6985 (-)	-	+	CT
-	only in training set contains only ct		
+			
CT			

30 – 8286 (-)	-	+	CT
-	10.0	1.0	40.0
+	0	0	0
CT	0	0	0

Figure 5.3.: The figure shows the casewise results that were obtained for the 30 datasets. The coloring (green, yellow, red) can be seen as a rough indicator for the performance level ranging from green (good performance) over yellow to red (bad performance).

5.4.2. Experiment 2: Results on Individual Datasets

Figure 5.3 reveals that the classification accuracies largely vary between the different datasets. Whereas the datasets of two thirds of all patients were classified well (indicated in green), problems were observed for nine sets (marked yellow and red). Most misclassifications were due to confusions of connective tissue and tumor (e.g. for dataset 20), although one should expect that discriminating tumor from healthy tissue is simpler than discriminating different tumor types. However, since most of the connective tissue training samples stem from regions that are close to tumor areas, the tissue might have already undergone molecular changes. Re-inspection of the stained image corresponding to dataset 20 revealed that some of the areas labeled as connective tissue indeed did not contain pure connective tissue, but were rather close to tumor stroma. Similar observations were made for dataset 25. Thus, some confusion of connective tissue with tumor tissue can be explained, especially since wrong labels in one set may also influence the classification of other sets.

On the other hand, we cannot explain the confusions of connective tissue and cancer pixels in datasets 10, 25 and 26. Moreover, for unknown reasons three datasets (5, 6, 15) were classified wrongly regarding their HER2 status.

5.4.3. Challenges that Might Prevent Higher Classification Rates

In our experiments, classification rates of 80–85% were obtained. On the one hand, this shows that MALDI MSI can indeed be used to discriminate HER2 positive from HER2 negative tumor and connective tissue on a pixel level. On the other hand, this means that the classifier is wrong in considerably many cases. Since the correct determination of the HER2 status is decisive for the further course of the treatment of the patient, this is not satisfactory and higher classification rates are required. The rest of this chapter discusses properties of the data that we believe currently prevent higher classification rates. We focus on two aspects: 1) potential difficulties shared by our study (working on the pixel level) and the Rauser study (using mean spectra) and 2) additional problems that only arise in our study and that might explain the performance loss of around 5% in casewise classification.

The Data is Characterized by High Technical and Biological Variability. In contrast to the SIMS data that was used in our previous study (cf. chapter 4), the MALDI data used in this chapter and by Rauser et al. [171] features a significantly lower signal to noise ratio and much stronger baseline effects. In addition, the tissue stems from different patients such that both the technical and biological variability is significantly higher. We note that it is sometimes difficult to distinguish between these two types of variability. When observing differences between spectra that have been assigned to the same class, it is not obvious if the differences arise from instabilities in the data acquisition process or if the differences can be explained by different biological content

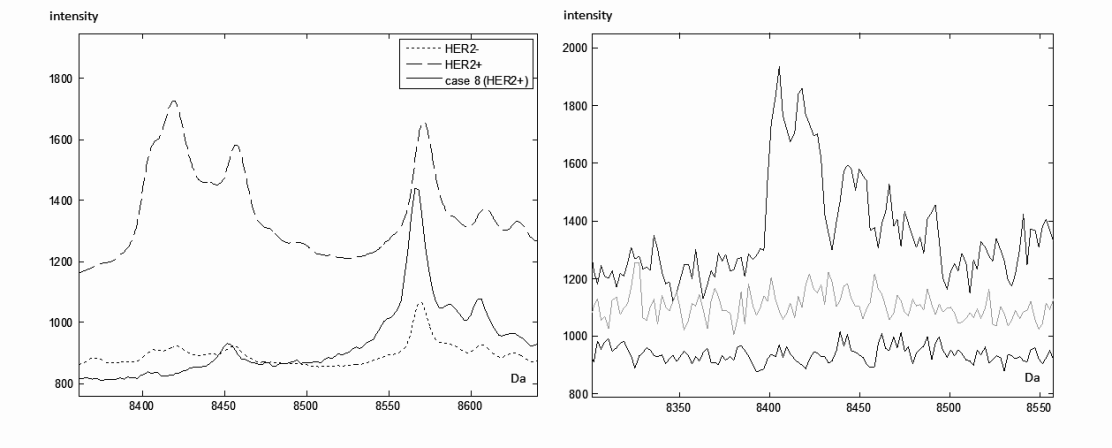

Figure 5.4.: Left: Mean spectra over all available cancer samples for the HER2 negative and positive classes around 8,500 Da (see also [171]). The pattern of two low peaks followed by a very high peak seems to be characteristic for the negative class. However, e.g., the mean spectrum of dataset 8 (labeled as HER2 positive) also features this pattern. Right: Three different HER2 positive spectra that were taken from the same dataset (case 5), where the upper two spectra even stem from pixels that belong to the same tumor region. We observe that even the within dataset (and within tumor) variance is very high–e.g., for the peak at 8,404 Da.

which is simply not reflected in the labels (and maybe not even in the histochemical stain). Generally, we note that, independent from their nature, these variabilities may be very large. In the following analysis, we mainly focus on the 8,400–8,600 Da area, which was previously shown to contain discriminative peaks [171].

We first focus on the variability that is even manifested on the level of the mean spectra and that also affected the analysis in the Rauser study: In figure 5.4 (left, see also [171]) the overall mean spectra for the two cancer classes are plotted. They seem to express characteristic patterns for the two classes in the 8,400–8,600 Da mass range. However, although most of the mean spectra indeed follow these patterns, we also observed outliers that feature inverted patterns (such as dataset 8). Consequently, discriminating between the two cancer types is very difficult, especially since the training set comprises only few datasets. Indeed, our classification procedure manages to correctly classify set 8 (cf. figure 5.3), but this may be just because other datasets in the training set also feature such "inverted" patterns. In figure 5.5 the mean spectra over HER2 negative tumor regions of two representative datasets (6 and 28) are shown over their full mass range. Whereas it is difficult to identify peaks in the left spectrum, the right spectrum comprises many clearly distinguishable peaks, and is characterized by a significantly higher signal

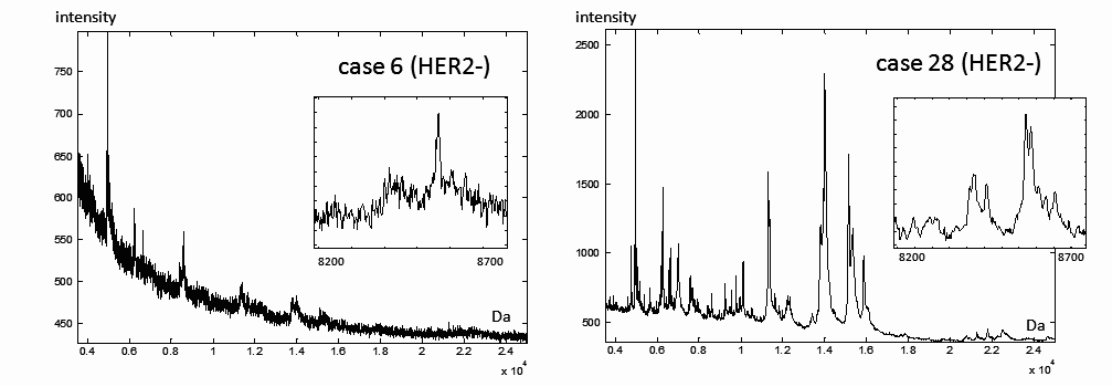

Figure 5.5.: The figure shows the mean raw spectra over HER2 negative cancer areas from different datasets. Note that the left mean spectrum (case 6) is an average over 23 spectra, and the right one (case 28) is an average over 219 spectra. As a consequence, the signal to noise ratio of the left spectrum is significantly lower and only few peaks can be identified. The 8,400–8,600 area that was previously shown to contain discriminative peaks [171] is very prominent in the left spectrum but is overshadowed by many other peaks in the right spectrum.

to noise ratio[8]. Large differences in the relative peak heights can be observed, for instance for the 8,400–8,600 Da mass range, which is very prominent in the left mean spectrum but is overshadowed by several higher peaks in the right spectrum. We thus note that the within class variability can be high, at least if the spectra stem from different datasets. This could arise from instabilities of the instrument, or may have biological reasons (recall that the individual sets not only differ in their HER2 status but, e.g., also feature different levels of estrogen and progesterone expression).

Although the high variability described above certainly complicates the analysis, both, Rauser's and our HER2 study had to face the same challenges such that these effects do not explain the relative performance loss of approximately 5% in casewise classification accuracy in our study. However, the variabilities on the mean spectra level just seem to be a consequence of the even higher variability that can be observed on the pixel level and only affects our study. For instance, we noted large differences between equivalently labeled spectra taken from the same dataset and even from the same tumor region within a set (cf. figure 5.4 for examples regarding dataset 5): Whereas some spectra of the tumor feature high peaks between 8,400 and 8,600 Da, others show only noise in the respective mass range. It is thus not surprising that we indeed observed several misclassified pixels for dataset 5 (cf. figure 5.3). By averaging over tissue areas like done by Rauser et al., some of this variability may be removed. This behavior is desired in

[8]Note that the individual datasets contain highly different numbers of tumor samples.

case of technical variability but might also camouflage biological variability that might be interesting in itself. As a consequence of the large variabilities between individual samples, high signal to noise ratios were only obtained if many spectra were averaged (cf. figure 5.5).

The Mean Spectra of the Rauser Study are "Good-Natured". The mean spectra that were analyzed in Rauser's study [171] were obtained by randomly sampling spectra from tumor regions, processing them with ClinProTools 2.2, and averaging over the tumor spectra of each dataset (personal communication and [171]). Comparing the resulting mean spectra with the mean spectra that were obtained after applying our preprocessing pipelines revealed significant differences for some datasets. Figure 5.6 shows the mean spectrum over all (HER2 positive) cancer samples of dataset 5 in the mass range between 8,400 and 8,600 Da. The mean spectrum of the raw data shows two peaks at around 8,400 and 8,570 Da where the second one is higher than the first one. The same pattern appears in the peak picked mean spectrum obtained with our preprocessing method (middle left). In contrast, the mean spectrum taken from the Rauser study shows an inverted pattern (middle right). The first peak is significantly higher than the second one, which is rather the characteristic of the HER2 negative samples than of the positive ones (cf. figure 5.4). Interestingly, dataset 5 is among the wrongly classified sets in our study (cf. figure 5.3) while it did not cause problems in Rauser's study. Since one dataset affects the dataset-wise classification accuracy by about 4%, this might explain a large fraction of the performance loss of 5%.

To further analyze this effect, we conducted the following experiment: We used ClinProTools 2.2 to preprocess all datasets a second time (with exactly the same settings), but this time using all labeled samples instead of randomly selected ones. We then averaged all cancer spectra of each dataset to get "new" mean spectra and compared the classification performance obtained with the new mean spectra with the performance obtained using the "old" mean spectra taken from Rauser's study[9]. Therefore, we used the same feature selection approach as Rauser et al., employed both a random forest classifier and an SVM (using the same parameters as in [171]), and estimated the classifiers' performance using ten times ten folds cross validations. Table 5.4 shows that the new mean spectra indeed yield significantly lower performance rates. While we were able to reconstruct the results from the Rauser study using the old mean spectra[10], the performance with the new mean spectra dropped by more than 5% in accuracy, that is to approximately the performance level that was obtained in our experiments described above (with the notable difference that we used a pixel-based classification!).

The performance difference on the new and old mean spectra is obvious. However, as shown in figure 5.6 it is at least worthy of discussion which mean spectra are actually

[9] Note that *both* of these sets were preprocessed with ClinProTools.

[10] The small differences arise since Rauser et al. used 30 sets whereas in our study 4 sets did not contain samples labeled as tumor and were left out.

mean spectra	criterion	SVM without	SVM w/ feature selection	random forest without	random forest w/ feature selection
old	recall	**70.7**	**88.6**	**85.8**	**91.8**
	precision	**66.0**	**84.9**	90.7	87.9
new	recall	67.9	83.6	77.7	80.4
	precision	62.5	76.5	**92.1**	**90.7**

Table 5.4.: Comparison of the classification rates obtained on the mean spectra of the Rauser study ("old" mean spectra) and the mean spectra that were obtained by preprocessing the data with ClinProTools for a second time, but this time using all samples ("new" mean spectra, see text for details). Clearly, the classification performance is better on the old mean spectra.

more correct/closer to the truth. The new mean spectrum of dataset 5 (cf. figure 5.6, right) is very close to the one that was obtained with our preprocessing pipeline and the raw data. This indicates that rather the old mean spectrum than our mean spectrum constitutes an outlier. Potentially, some of the differences between the old and new mean spectra can be explained by the fact that Rauser et al. used a random sampling scheme for determining the mean spectra. Since the tumor samples from dataset 5 (and also other sets) are rather heterogeneous (see above), a bias can be introduced if the number of random samples is set too low. In that sense, the old mean spectra might just be "good-natured". We note, however, that this question could not ultimately be answered.

The Size of the Training Set is Too Small. Obviously, classifiers can only generalize well to newly presented data if the real-world variability is adequately represented in the training set. As shown above, the biological and/or technical variability in our study is very high. Thus, our training set, which only contains samples from 12 respectively 14 sets representing the two tumor classes, might not be large enough. Indeed, figure 5.7 shows that the learning curve of the classifier is still positive if $(3+3)$ sets are left out for testing (see experiments)[11]. Thus, the classifier might indeed benefit from more training data from additional patients.

The Three-class Problem is More Difficult. Whereas Rauser et al. restricted their analysis to a binary classification problem, we worked with a three class problem where–in addition to the two types of breast cancer tissue with positive and negative HER2 status–we also considered connective tissue. Obviously, this task is more demanding of the classifier, especially if the connective tissue samples might already have undergone molecular changes such that the differences of healthy and diseased tissue are less pronounced.

[11] Note that whereas we always left out 6 datasets for testing, Rauser et al. used a ten folds cross validation where each fold contained 3 sets for testing.

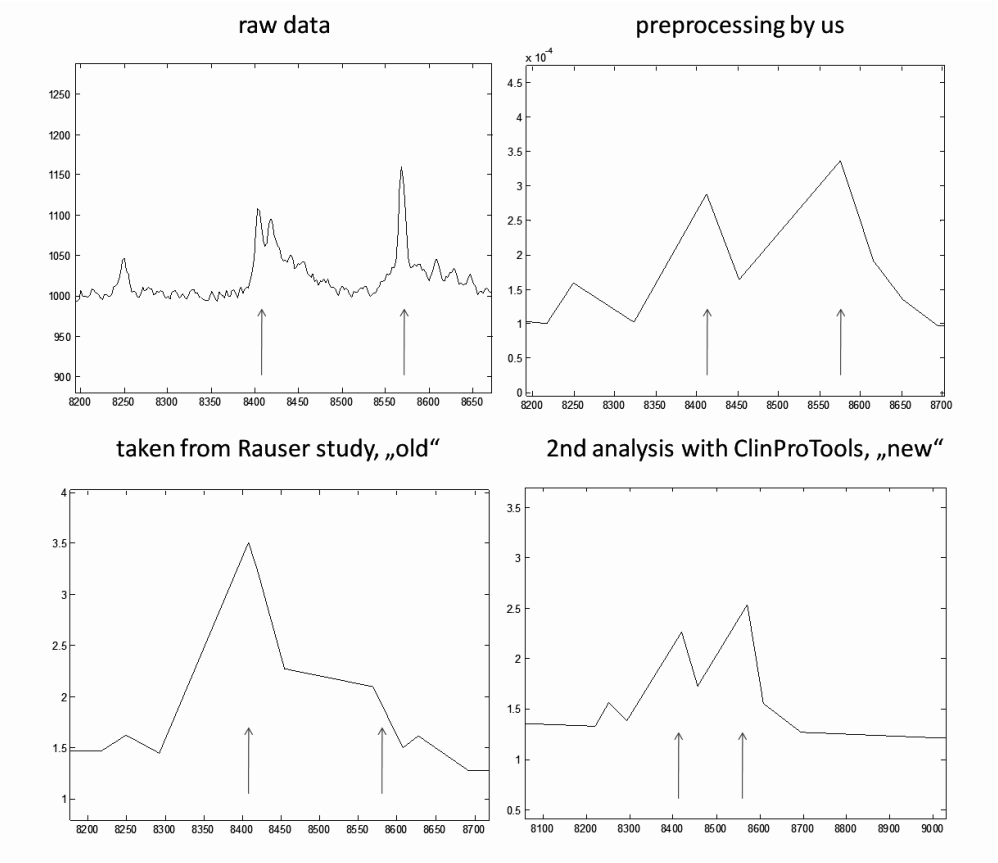

Figure 5.6.: The mean raw cancer spectrum of dataset 5 (left) shows two peaks at around 8,400 and 8,570 Da where the second one is higher than the first one. Whereas we observe a similar pattern in the peak picked mean spectrum obtained with our preprocessing methods (middle left), the mean spectrum taken from the Rauser study (which used a randomly sampled subset of all labeled samples to estimate the mean spectrum) shows an inverted pattern (middle right). A second analysis with ClinProTools using all labeled spectra in the set yielded a mean spectrum that was very similar to the one obtained by our preprocessing algorithms (right).

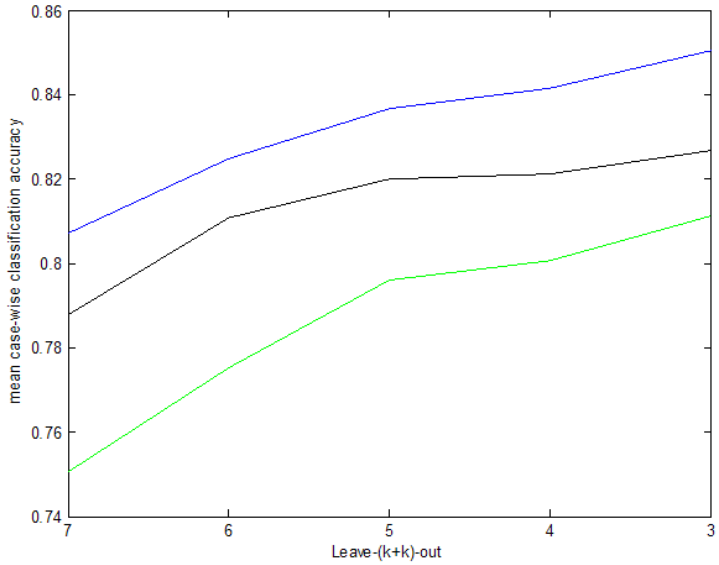

Figure 5.7.: In this experiment, we used the best-performing pipeline (see table 5.2) where we left out different numbers of positive and negative sets in the training of the classifier (7+7 up to 3+3). All experiments were repeated 500 times. We observe that the mean recall (blue), precision (green) and dataset-wise classification accuracies (black) increase if more datasets are used in the training (and thus, more variability is included in the training set). This suggests that the classifier might benefit from additional datasets.

Altered Data Pool Complicates Comparisons. A further difference between the two studies is that Rauser et al. used four additional datasets featuring tumor samples (sets 9, 12, 27, 29). The corresponding labels were removed for our study since the assessment of the HER2 status was less reliable than for the other 26 cases. Note that this removal of datasets can have a positive or negative effect on the classification accuracies. On the one hand, we discard less reliable labels, but on the other hand, we remove variability from the training set.

5.5. Conclusion

Our study suggests that the selection of methods used for preprocessing MALDI MSI spectra strongly influences the obtainable classification performance. Differences of up to 30% in precision and recall were observed when different (optimal and suboptimal) preprocessing pipelines were used. Nevertheless, many studies only briefly (if at all)

touch this topic. We thus hope to have raised awareness to this problem.

On the MALDI MSI data from our HER2 study, the best pipelines yielded pixel-wise classification rates of up to 80–85%. Although these results are promising and in line with the findings in the previous chapter, we think that higher rates would be possible by getting a grip on the high variability that was even observed between samples from the same tumor.

In case that this variability has technical reasons, more reliable instruments or sample preparation protocols might be needed to increase classification accuracies. As a temporary solution, the spatial resolution of the data may indeed have to be decreased to increase its signal to noise ratio [171] and to allow for the training of more robust classifiers. As an alternative, the analysis of the individual datasets should be repeated, since in mass spectrometry "significant variations in the data are usually evident, even for the same sample run on the same instrument on the same day" [65]. This, however, would prolong data acquisition times.

In case that the variability is of biological nature, the following aspects should be considered: Classifiers can only generalize well to newly acquired data if the real-world variability is adequately represented in the training set. If the within-class variability is high, large training sets (in our case $\gg 26$) might be needed to obtain robust classifiers. The learning curve is a good indicator if additional data might be beneficial. Secondly, it may be worth to reinspect the labels in the training set. Obviously, labeling errors should be avoided. They can, e.g., occur if the training is performed on individual pixels but the labels were assigned to large and potentially heterogeneous regions. It might also be worth to refine the label classes, e.g. by introducing additional tissue subclasses, to reduce the within-class variability and obtain higher performance levels. However, in this case, an even larger number of training sets might be required.

Chapter 6

Active Learning for Efficient Labeling and Classification of Mass Spectrometry Images

In the previous chapters, we have demonstrated that the random forest classifier is well suited for automated annotation of MSI data. However, the genetic variabilities between patients can be significant [148], and we have seen that at least some of these variabilities are also reflected in the MSI spectra. As a consequence, training of universal classifiers that generalize well to newly acquired MSI data is difficult, especially if only few training sets are available. In such scenarios, more robust and reliable results may be obtained by training the classifier anew for each newly acquired MSI set. This would at the same time account for fluctuations in sample preparation and/or instrument settings. However, labeling training data is both costly and time-consuming. It is thus desirable to reduce the number of required labels (i.e., labeling time for the expert) without compromising on classification accuracy. This motivates the application of semi-supervised learning techniques (SSL) [238, 40], active learning (AL) strategies (see Settles [196] for a comprehensive review) or hybrid approaches [168].

In the following chapter we build on work by Röder et al. [178, 177, 176] and propose a novel multi-class active learning strategy for the random forest classifier [25][1]. We show on real world MSI data, that our approach results in high classification accuracies after only a few learning steps and is thus suitable for efficient annotation of MSI datasets. We further demonstrate that, given the same number of labels, our querying strategy outperforms traditional passive learning by a large margin.

[1] Here we use the standard random forest classifier [25] but note that an online version [182, 76] or even a different classifier could be used instead [177].

6.1. Introduction

Brush tools as known from graphics editing programs like Adobe Photoshop or GiMP allow for efficient and fast annotation of large MSI training datasets. However, their application is inept if the analyzed tissue samples are highly heterogeneous as it is the case in many current MSI studies [32, 223]. Pixel-wise labeling, in contrast, is more accurate but typically requires significantly more labeling time.

In active learning, the algorithm iteratively queries the user to label the pixel positions where additional knowledge may be most beneficial for improving the classifier's performance. By labeling the samples in a smart order, a high performance level can often be obtained with only few training points. Although active learning methods have shown excellent performance in many fields of research such as speech recognition [175], image classification [112] or remote sensing [215], only very few researchers have applied them to mass spectrometry data. Zomer et al. [239] proposed an active learning algorithm for kernel-free support vector machines (SVM) for binary classification problems. To our knowledge, only Schleif et al. [186, 187] considered multi-class problems, which frequently occur in practical applications. Their multi-class method is adapted from margin based active learning strategies for generalized relevance learning vector quantization (GRLVQ). However, they use the term active learning with a different connotation than we do. While we aim at minimizing the number of labels that are requested from the expert, their active strategies rather optimize computation time, and usually all available points are used at least once when training the classifier (personal communication). We are not aware of any study that considers MSI datasets.

6.2. Materials and Methods

6.2.1. Active Learning

Active learning is motivated by the observation that often a classifier can benefit more from few but informative training examples than from large amounts of possibly non-informative examples. Typically, approaches are turn-based (i.e., iterative) and "guide" the labeler in the sense that the algorithm chooses the data from which it learns [196]. In each learning round, the algorithm requests a label for the sample for which a defined criterion is optimal. A powerful active learning strategy should select samples for labeling according to two principles:

- Principle 1: Labels should be provided for points that "matter". That is, we don't want to label unlikely samples unless their misclassification would cause significant harm. Sample importance should therefore depend on the sample's likelihood (as measured by the probability density in feature space) and the associated misclassification loss.

- Principle 2: New labels should have a reasonable chance to change the classifier's mind. That is, we don't want to label points in regions where the classifier is already very certain. Sample importance should therefore be controlled by an estimate of the likelihood that the output of the classifier might change.

Considering a sample's likelihood and/or its misclassification loss (cf. principle 1) is the core of many sampling strategies like uncertainty sampling or random sampling, but its impact on training the classifier is less frequently modeled (cf. principle 2). For instance, random sampling, the standard sampling strategy, does not meet all of these requirements because, implicitly, it only depends on the probability density of unlabeled samples.

In the following, let $S = \{(x^1, y^1), \ldots, (x^N, y^N)\}$ be the set of available M-dimensional training samples, that is mass spectra $x^k \in \mathbb{R}^M$ with M channels and corresponding class labels $y^k \in \{1, \ldots, D\}$ (cf. section 4.2.1).

6.2.2. Novel Active Learning Strategy

Outline. Recently, Röder et al. proposed a novel active learning strategy [177] for binary classification problems. The rest of this section follows the presentation in [177] and generalizes their method for the classification of data with an arbitrary number of classes (i.e., for the multi-class case).

In short, in each learning step, the next point to be labeled is determined by calculating the *training utility value (TUV)* of each pixel. This measure quantifies the benefit for the classifier of requesting an additional label for spectrum/pixel x. Such a criterion is at the heart of almost all active learning strategies, however, many different variants are possible. In our case, the TUV is defined such that it conforms to principles 1 and 2 and fulfills the following properties: 1) points in areas of high density in feature space are preferred, that is points with spectra that are highly similar to many others; 2) points with posterior class probability estimates that are close to the decision boundary are preferred, that is points where the classifier is unsure; and 3) points with spectra that are similar to other spectra which have already been labeled are selected with less probability, since the information gain for those are expected to be low. In each round, the algorithm requests a label for the sample x that has the maximum TUV value among all unlabeled points U and is thus supposed to contribute most to improving the classifier's performance. After label assignment, the classifier is retrained with the augmented label set, all unlabeled points are reclassified, and the algorithm continues with the next learning step. This iterative scheme is continued until either the human expert is satisfied with the classification result or a predefined stopping criterion is met. An overview of the method is given in algorithm 1.

Formally, the TUV is the difference between two different estimates given sample x: the estimated loss associated with the point estimate (termed $\hat{R}_x$) and the estimated

Algorithm 1 : Overview of our active learning procedure.

query label at point x with the largest density in feature space
for $k = 1$ to maxIterations **do**
1. uniformly sample R times from the bounding box enclosing all observations in feature space and label the obtained samples as "0"
2. combine user-labeled points (D classes) and "0"-samples to train a random forest classifier with $D + 1$ classes
3. classify all unlabeled points
4. drop random forest votes for class "0" to obtain D-dimensional vectors α_x for all unlabeled points x with $\alpha_x(i) = 1 + v_x(i), i = 1, \ldots, D$ where $v_x(i)$ is the number of trees that given x vote for class i
5. query label at point x with the highest *training utility value (TUV)* among all so far unlabeled points (i.e., $\max_{x \in U} TUV_x(\alpha_x)$; see equation (6.6))
end for

loss based on the distribution estimate ($\hat{R}_x^B$). In the following paragraphs, these two estimates are discussed in more detail.

Minimizing the Overall Expected Loss. Assume, that for all unlabeled points $x \in U$ a density estimate $\hat{d}(x)$ of the unknown real density $d(x)$ is available. In the multi-class case, the estimated local loss of the classifier that we want to minimize is given by [177]

$$\mathbb{E}[L] = d(x) \sum_{j=1}^{D} \sum_{i \neq j} L_{ij} p(y = i|x) \mathbf{1} \left\{ \text{classification}(x) = j \right\}. \tag{6.1}$$

L_{ij} is the loss associated with misclassifying a point of class i as j, and the indicator function $\mathbf{1}\{\text{classification}(x) = j\}$ is equal to 1 if the classifier assigns class j and 0 else (for a given sample x). Further assume that the classifier not only provides us with a point estimate for the posterior class probabilities but also with a distribution estimate $\hat{F} := \hat{F}(p(y = 1|x), \ldots, p(y = D|x))$ for the posterior class probabilities $p(y = i|x), i = 1, \ldots, D$. This distribution estimate will turn out to be an important ingredient of our active learning approach.

Estimate Loss for the Point Estimate: $\hat{R}_x$. We freeze the active learning iterations after a specific number of learning steps to obtain a crisp classification. For each sample x, the random forest classifier yields a vector $p(y|x) = (p(y = 1|x); \ldots; p(y = D|x))$ of posterior class probability estimates $p(y = i|x), i = 1, \ldots, D$. When interpreted as a point, it lies on a D-dimensional simplex Ψ (cf. figure 6.1). The larger the maximum value of $p(y|x)$, the closer it lies to a corner of the simplex and thus, the more unambiguous the classification is. Given that all losses L_{ij}, $i \neq j$ are equal (e.g., 0-1 loss that assigns a loss of 0 for correct classifications and 1 for *all* misclassifications), we assign

class i if and only if $\hat{p}(y|x)$ is closest to the i-th corner of the simplex. Thus, we introduce a threshold point T^{0-1} that lies in the center of the simplex and partitions Ψ into D parts $\Psi_i, i = 1, \ldots, D$, and we assign class i if and only if $\hat{p}(y|x) \in \Psi_i$. Working with differing L_{ij} corresponds to shifting T on the simplex. The i-th component of T is then calculated from $T(i) = \sum_{l \neq i} L_{li} / \sum_{i,l} L_{il}$. Given that we always assign the class with the highest posterior probability, we obtain a plug-in estimate for the local loss at x [177]:

$$\hat{R}_x(\alpha_x(1), \ldots, \alpha_x(D)) = d(x) \min_j \left[\sum_{i \neq j} p(y = i|x) L_{ij} \right]. \tag{6.2}$$

Note that the loss at x is weighted by the density $d(x)$ and the losses L_{ij} that are associated with the misclassification of x. Thus, a low loss is attributed to outlying samples unless misclassifying these samples significantly increases the overall loss (cf. the definition of the local loss in equation (6.1) and principle 1). The density $d(x)$ can be estimated by a variety of different density estimation schemes (here we use random forest density estimation [178]), and we estimate the unknown class probability $p(y = i|x), i = 1, \ldots, D$ by the i-th component of the expected value of the distribution estimate $\hat{F}$.

Obtaining Distribution Estimates. Following Röder et al. [177], in each iteration we enhance the set of currently labeled points by a certain number of samples from an artificial reference class, where the ratio between the two is defined by the user-defined resampling parameter R. These additional samples, labeled as class "0", are uniformly drawn from the hypercube enclosing all observations in feature space (that is the M-dimensional bounding box). Thus, in each learning step the random forest is actually trained with $D + 1$ classes and then used to classify all unlabeled $x \in U$. In regions of the feature space with no or only few user-labeled samples, the individual trees of the forest tend to vote for the reference class "0", for which one or more examples are likely to be close. A high fraction of such votes for the reference class indicates a low number of labeled points from the original training set in the neighborhood of x and thus high uncertainty of the classifier with respect to the classes that it is actually supposed to learn.

More formally, let $v_x(i)$ be the number of trees that given sample x vote for class i. The distribution over class votes can be modeled as a multinomial distribution (cf. Appendix 6A). Under the mild assumption of a uniform prior (i.e., all classes are equally likely), it follows that in the two-class case $\hat{F}(p(y = 2|x))$ (and analogously $\hat{F}(p(y = 1|x))$) is described by a Beta distribution $B(a, b)$ with parameters $a = 1 + v_x(2)$ and $b = 1 + v_x(1)$ [177]. In case that many trees vote for the reference class "0" (i.e., $v_x(0)$ is high), the parameters of this Beta distribution will be small, and, as a consequence, the distribution will be broader, which reflects higher uncertainty. Under similar assumptions, in the multi-class case $\hat{F}$ follows the multivariate generalization of the Beta distribution, which

is the Dirichlet distribution [36] with parameters $\alpha_x(i) = 1 + v_x(i), i = 1, \ldots, D$ (see Appendix 6A for details). $\tilde{F} \sim Dir$ if and only if

$$\tilde{F}(p(y=1|x), ..., p(y=D|x)|\alpha_x(1), ..., \alpha_x(D)) = \frac{\Gamma\left(\sum_{i=1}^{D} \alpha_x(i)\right)}{\prod_{i=1}^{D} \Gamma\left(\alpha_x(i)\right)} \prod_{i=1}^{D} p(y=i|x)^{\alpha_x(i)-1} \tag{6.3}$$

where Γ is the Gamma function. A closed-form solution for the mean M of this distribution estimate exists [36], and its components can be calculated from $M(i) = \alpha_x(i) / \sum_{k=1}^{D} \alpha_x(k), i = 1, \ldots, D$. By using the mean as an estimate for $p(y = i|x), i = 1, \ldots, D$ in equation (6.2) we obtain a natural estimate $\hat{R}_x$ for the expected loss at x (in the current active learning step), where

$$\hat{R}_x(\alpha_x(1), \ldots, \alpha_x(D)) = \hat{d}(x) \min_j \left[\sum_{i \neq j} \frac{\alpha_x(i) L_{ij}}{\sum_{k=1}^{D} \alpha_x(k)} \right]. \tag{6.4}$$

Estimate Loss for the Distribution Estimate: $\hat{R}_x^B$. We conclude from the last paragraphs that samples from a reference class can be used to get a distribution estimate $\tilde{F}$, and that we can derive an estimate $\hat{R}_x$ for the loss associated with the point estimate. While so far we have only used crisp class assignments (see equations (6.2) and (6.4)), we now assume that we do not stop the active learning iterations, but continue learning. By doing so, we account for the fact that indeed distributions and not only point estimates for the posterior class probabilities are available and calculate an estimate for the loss associated with the distribution estimate. The difference between the loss of the point estimate and the latter will provide us with an estimate for the possible benefit for requesting an additional label for sample x, and will allow us to conform to principle 2 that requests that a suitable active sampling strategy prefers samples that have a reasonable chance to change the classifier.

Let the simplex Ψ be partitioned into D parts Ψ_j as described above, and define $\tilde{\alpha}_{x,i}$ to be equivalent to α_x except that $\tilde{\alpha}_{x,i}(i) = \alpha_x(i) + 1$. Let further $I_{\Psi_j}(\tilde{\alpha}_{x,i})$ be the multivariate equivalent to an incomplete Beta function [1], which integrates the distribution estimate $\tilde{F}$ (given by a Dirichlet distribution with parameters $\tilde{\alpha}_{x,i}$) over part Ψ_j of the simplex (see figure 6.1 and Appendix 6B for details). With the decision rule from above an estimate $\hat{R}_x^B$ for the loss of the distribution estimate [20] at position x is obtained:

$$\hat{R}_x^B(\alpha_x(1), \ldots, \alpha_x(D)) = \hat{d}(x) \left[\sum_{j=1}^{D} \sum_{i \neq j} \underbrace{\frac{\alpha_x(i)}{\sum_{k=1}^{D} \alpha_x(k)}}_{term\ 1} \underbrace{L_{ij}}_{term\ 2} \underbrace{I_{\Psi_j}(\tilde{\alpha}_{x,i})}_{term\ 3} \right]. \tag{6.5}$$

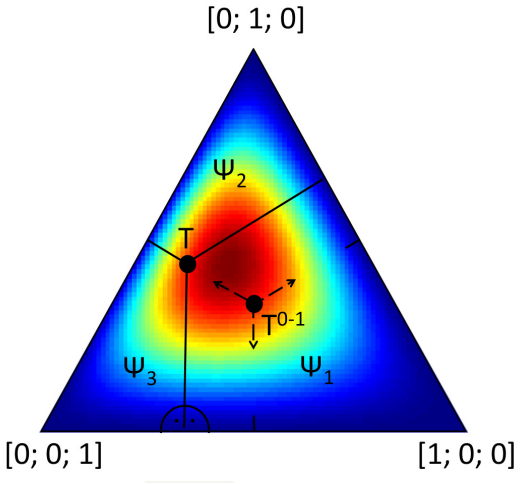

Figure 6.1.: The figure shows a Dirichlet distribution on a 3-dimensional simplex Ψ (blue indicates low, red indicates high probability). The threshold point T partitions Ψ into three parts Ψ_j that define the integration boundaries for I_{Ψ_j} (see equation (6.5) and text for details). Its position is determined by the loss function. In case that all losses $L_{ij}, i \neq j$ are equal (e.g., 0-1 loss), T resides in the center of the simplex.

Essentially, this formula is a generalization of equation (6.4) for distributional estimates where each point in the simplex is weighted according to the distribution estimate[2]. An estimate for term 1 can directly be calculated, term 2 is given, and term 3 can, e.g., be approximated by Monte Carlo integration [93]. Our MATLAB implementation is based on Minka's Fastfit toolbox [150] for sampling from a Dirichlet distribution. The original code was optimized for sampling from a series of Dirichlet distributions with similar parameterizations (for details refer to Appendix 6D) to reduce computation time.

Training Utility Value. Combining equations (6.4) and (6.5) yields the *training utility value (TUV)* criterion introduced above for selecting the next instance for querying a label [177]:

$$TUV_x(\alpha_x) = \hat{R}_x(\alpha_x(1), \ldots, \alpha_x(D)) - \hat{R}_x^B(\alpha_x(1), \ldots, \alpha_x(D)). \tag{6.6}$$

Remember that the first summand $\hat{R}_x$ estimates the contribution of x to the risk that is obtained from the current best estimate (the mean of $\hat{F}$) of the true posterior, and the second summand $\hat{R}_x^B$ estimates the contribution of x to the average risk that can be achieved when considering the entire distribution of possible posterior probabilities. Querying for additional labels is most interesting where it does seem possible to obtain a lower minimal risk than the one obtained with the current best estimate of the posterior,

[2]In equation (6.4), exactly one point has mass 1 and the rest mass 0. For further intuition see [176].

which is why we use the difference of the two estimates as criterion. Thus, when selecting the next query point we always choose the instance for which the expected average decrease of local loss is largest, that is $\max_{x \in U} TUV_x(\alpha_x)$.

Note that in accordance with principle 2 we prefer points that have the potential to actually change the classifier's mind and at the same time focus on points that "matter" (principle 1). Analogously to [177], given equal density $d(x)$ our TUV criterion favors to request labels at points where the posterior class probability estimates $\hat{p}(y = i|x)$ are close to $1/D$, i.e., the parameters $\alpha_x(j)$ are small and equal and we can expect a large decrease in expected loss (high prediction uncertainty). Such points are preferred over points with high but equal parameters, which in turn are preferred over points with low prediction uncertainty, that is for which the parameters are unequal (see figure 6.2). The TUV integrates three sources of information [177]–the distance of an instance x to the current decision boundary, its density $d(x)$ of the marginal distribution of features, and the number of labeled points in the neighborhood of x. The latter is implicitly included since a a low number of neighboring training points results in many votes for the reference "0" class and thus a Dirichlet distribution with low parameter values.

6.3. Experiments

6.3.1. Research Questions

We evaluated if active selection of training samples with our method indeed improves the classification performance. Therefore, we compared our approach (AL-RF) to passive learning, i.e., a strategy that in each round randomly selects one of the remaining unlabeled points for which it queries a label (referred to as random sampling (RS) in the following). Performance was compared after training both approaches with the same number of labels.

6.3.2. Data

To compare our active learning method to random sampling, we used the secondary ion mass spectrometry (SIMS) data acquired from orthotopic human breast cancer xenografts (MCF-7) that was also used in the preceding chapters (see sections 3.3.2 and 4.3.1). Three out of the six available slices were selected–one from the bottom (entitled S4), middle (S7) and top (S11) of the stack of available parallel slices of the tumor. The spectra in the three datasets were baseline corrected by channelwise subtraction of the minimum, normalized by their total ion count, and features were extracted with a peak picker based on local maximum detection (cf. section 5.2.4). The dimensionality of the resulting spectra varied from 64 to 69 for the three sets. Labels corresponding to the five classes of interest where assigned using chemical staining of parallel slices (see figure

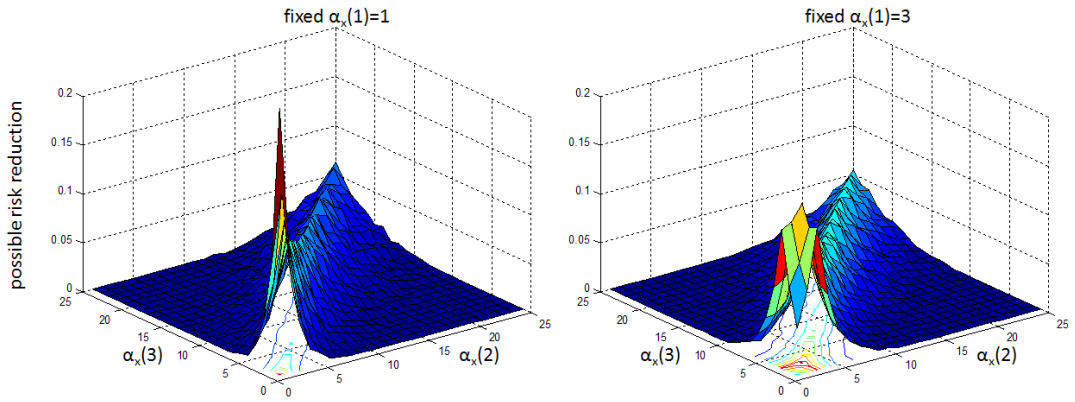

Figure 6.2.: The expected risk reduction $TUV_x(\alpha_x) = \hat{R}_x(\alpha_x) - \hat{R}_x^B(\alpha_x)$ for a sample x in the 3D case. The density $d(x)$ was fixed to 1 in both examples. We further fixed $\alpha_x(1)$ to 1 and 3 respectively (from left to right). Note that the TUV is symmetric in the parameters $\alpha_x(i)$ such that fixing $\alpha_x(2)$ or $\alpha_x(3)$ instead leads to the same results. The highest TUV scores are obtained for $\alpha_x(2)$ and $\alpha_x(3)$ equal to 1 respectively 3, i.e., for uniform parameters (see contour plot in bottom plane). Since the point $\alpha_x(i) = 3\ \forall i$ corresponds to a lower level of uncertainty than the point $\alpha_x(i) = 1\ \forall i$, the maximum on the right is lower than the one on the left. Also note that the TUV is symmetric with respect to the two varying parameters (some distortions due to the random nature of the Monte Carlo integrations occur). Equal values for $\alpha_x(2)$ and $\alpha_x(3)$ lead to higher $TUVs$ than differing values.

6.3 as well as sections 3.3.2 and 4.3.1 for a detailed description of the data). All points for which label information is available were used in the evaluation of our method.

6.3.3. Evaluation Criteria

Prediction accuracy was measured by sensitivity (SE) and positive predictive value (PPV) (see section 4.3.3 for definitions). The obtained SE and PPV rates were averaged over all four respectively five classes.

6.4. Results

Due to the probabilistic nature of the random sampling strategy and the Dirichlet sampling required for Monte Carlo integration, we repeated our active learning method and the random sampling approach 100 times and averaged the obtained results in each learning step. To obtain reliable quality estimates, in addition, we repeated the random

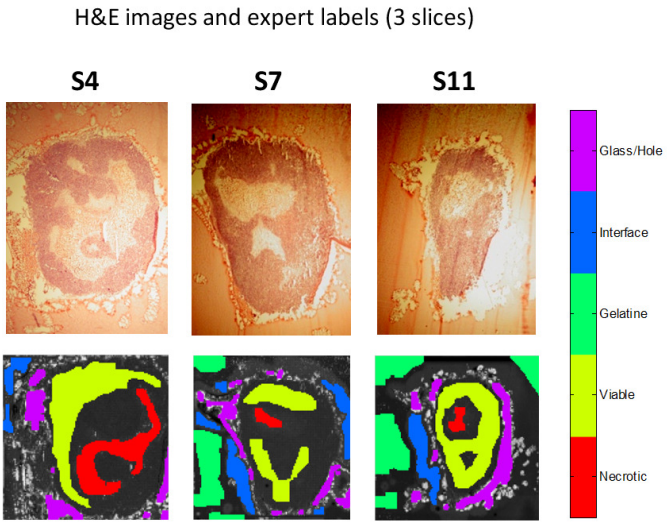

Figure 6.3.: The gold standard labels for the three MSI datasets (bottom row) are obtained from hematoxylin-eosin staining of parallel slices (top row). The labeling is only partial: labels for the five tissue classes of interest are color-coded whereas black and white indicates that no label information is available.

forest training and classification in each learning step five times. We used 10 trees per forest which constituted a good compromise between accuracy and computation time, drew 300 samples to perform the Monte Carlo integrations, and employed stratified sampling to balance the labels in the training set. In both approaches, the learning was started with an empty set of labeled points (in practical applications a number of initial labels might already be given), exactly one label was queried in each active learning step where the ground truth label map served as oracle, and a 0-1 loss function was assumed (i.e., T resided in the center of the simplex). With these settings, identifying the next query point required approximately one second. Classification results on the SIMS data are reported in figures 6.4, 6.5, 6.6, and 6.7 as well as in table 6.1.

6.5. Discussion

6.5.1. Performance on Slices S4, S7, and S11

Slice S4. Figure 6.4 reveals that our active learning scheme (AL-RF) performed competitive to random sampling (RS) in the first few learning steps and significantly outperformed RS as soon as more than $\approx$ 20 learning steps were executed. After 100 iterations, AL-RF improved on RS by about 10% in sensitivity. With an increasing

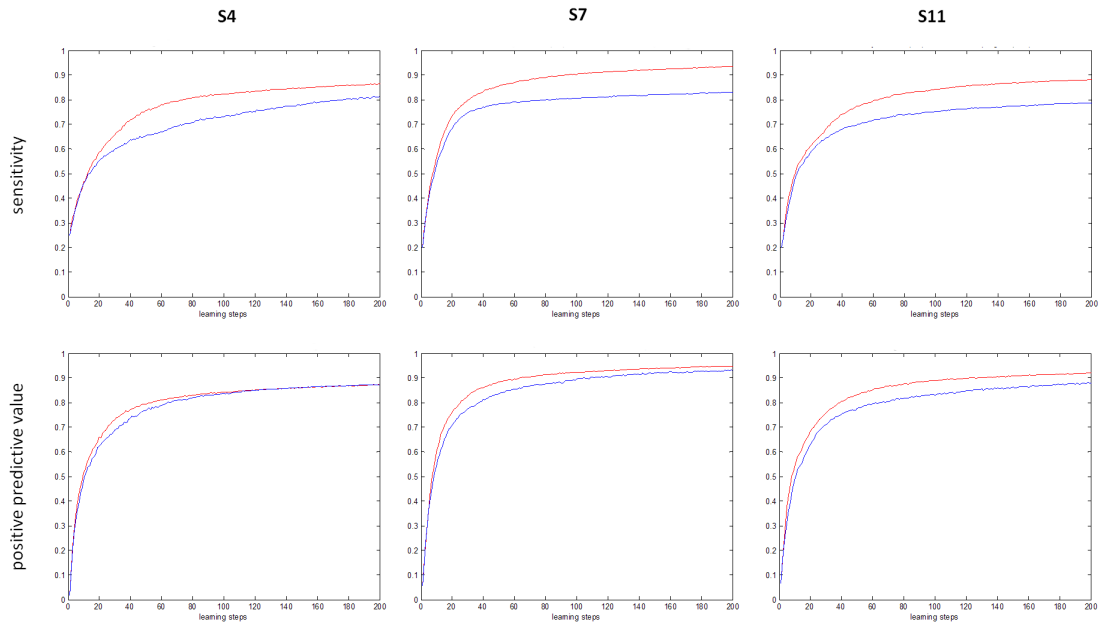

Figure 6.4.: Comparison of the sensitivities and positive predictive values obtained with random sampling (blue) and our active learning approach (red) (averaged over 100 repeats). On all sets, our method outperforms random sampling with respect to sensitivity and positive predictive value.

number of learning steps, RS slowly converged toward the sensitivity rates obtained with AL-RF, however, the margin was still more than 5% (cf. table 6.1) after 200 iterations. Regarding positive predictive value, AL-RF slightly outperformed RS in the first ≈ 70 learning steps and performed competitive afterwards.

However, the results after 200 steps were still below the reference values given in table 6.1, which were obtained by training a random forest classifier with 90% of all available samples per class and predicting the remaining samples. One possible explanation is that the labels in the ground truth map were assigned by looking at regions in the H&E stains rather than checking individual pixels and thus, there might be some errors in the ground truth that require additional learning steps. For instance, the necrotic class is likely to contain some holes, that is pixels that should belong to the glass class. Yet, this does not occur in the label maps. Furthermore, training with 90% of all available pixels might actually lead to overfitting.

Slice S7. Over the whole range of the first 200 iterations and especially for low numbers of learning steps, our approach outperformed RS with respect to PPV. At the same time, it significantly outperformed RS regarding sensitivity, leading to a gain of more than 10% after 100 and also after 200 learning steps. The sensitivity of the RS

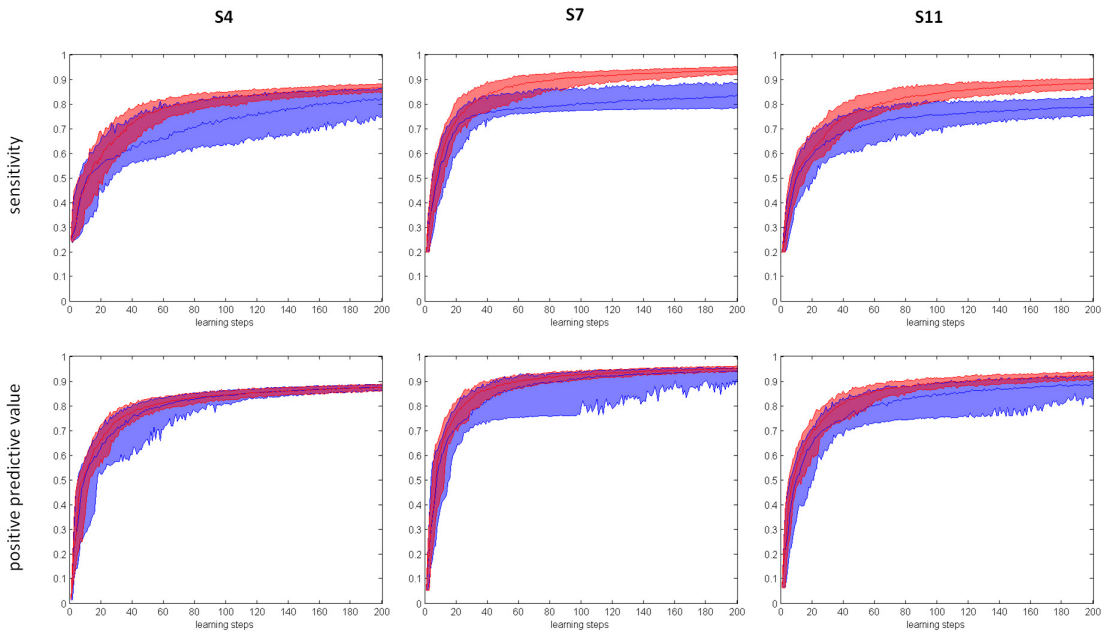

Figure 6.5.: The plots are complementary to the ones in figure 6.4 and show the band between the 95% quantile and the 5% quantile as well as the median for the 100 repeats of each of the methods. In contrast to random sampling (blue), our active learning approach (red)exhibits significantly lower variance between the different learning runs and the band around the median gets thinner and thinner over the course of iterations.

algorithm increased very slowly such that after 500 iterations the sensitivity was still at a comparably low level of 86%. After 200 iterations, the AL-RF sensitivity and PPV estimates were close to the reference values (cf. table 6.1).

Figure 6.6 reveals that RS failed to correctly discriminate the necrotic class (indicated in red) from the viable class (light green). This is because the necrotic area has only small spatial extent and thus, random sampling only selected few corresponding training points. In contrast, AL-RF usually selected more than twice as many necrotic samples leading to a significantly better classification result (cf. figure 6.6). Discriminating viable and necrotic tumor is probably the most challenging part of the classification task since those classes are relatively close in feature space. In contrast, the gelatin and glass spectra show less overlap in feature space, which simplifies their classification. Indeed, AL-RF requested less samples from these two classes than RS, and the corresponding areas in figure 6.6 are less densely sampled. Figure 6.7 displays intermediate results for one repeat of our active learning algorithm.

Slice S11. Regarding sensitivity as well as positive predictive value, the results ob-

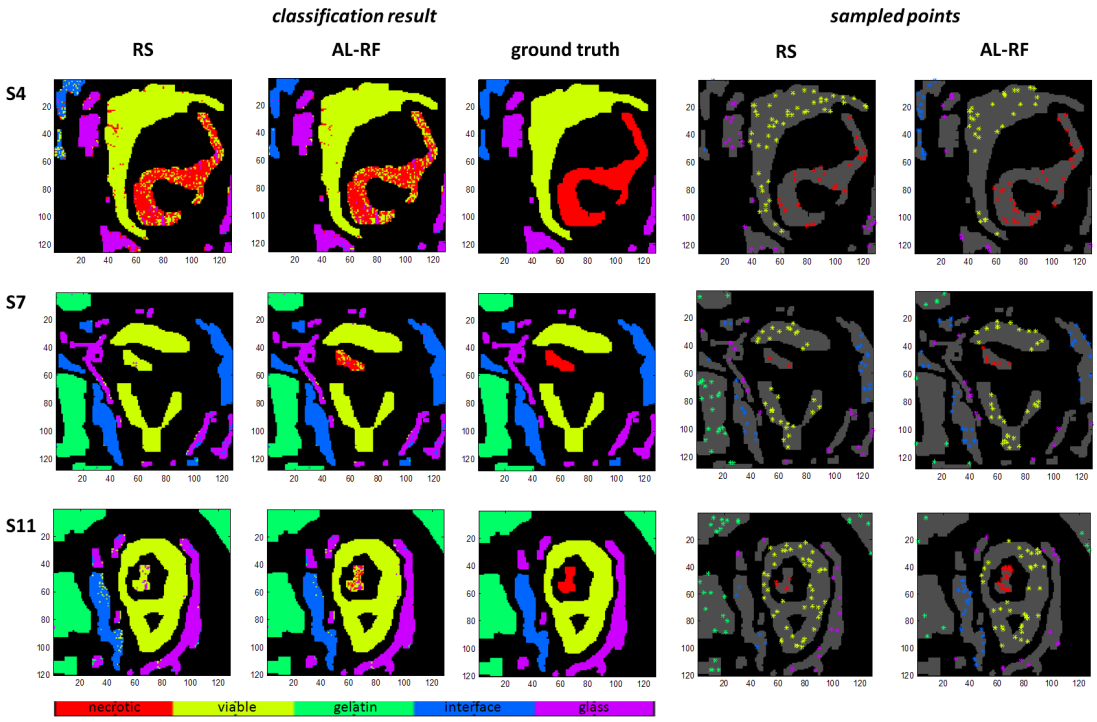

Figure 6.6.: The classification results after 100 learning steps with our active learning method (AL-RF) and random sampling (RS). To obtain the crisp classification, we first averaged the probability maps gathered in the 100 repeats and then took the maximum likelihood estimate in each pixel. On the right, the selected training points for representative learning runs are plotted (we refrain from plotting the training points for all 100 repeats to keep the images uncluttered). Since the area of the necrotic class is comparatively small in slices S7 and S11, random sampling only selects very few training points for that class, leading to a bad classification result. In contrast, AL-RF requests more training samples for that class yielding a superior classification. At the same time, it samples less points from the gelatin and glass classes, which have less overlap with the other classes in feature space than, e.g., necrotic and viable tissue and are thus easier to learn.

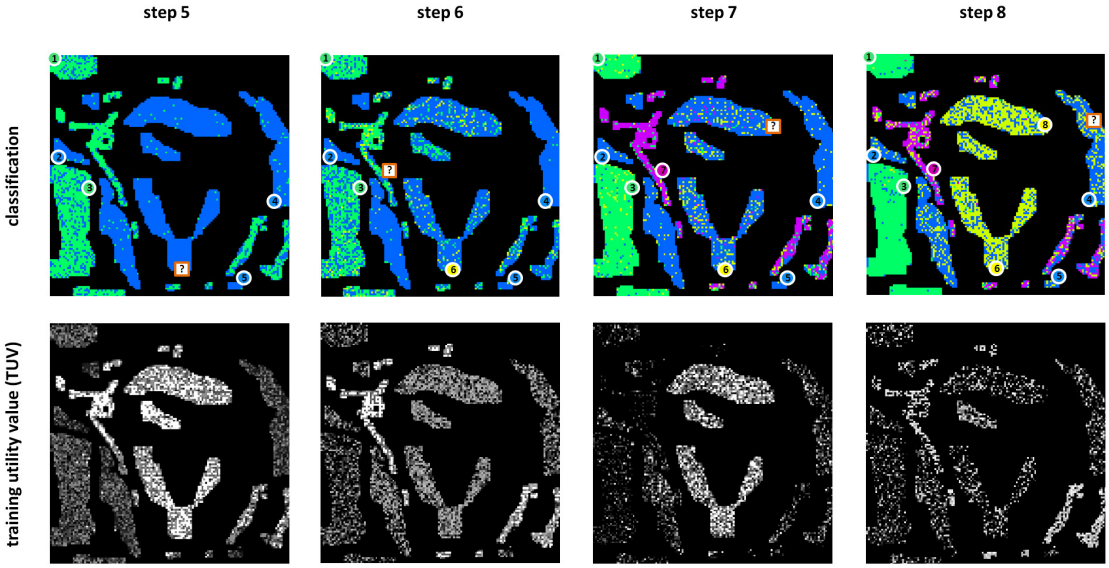

Figure 6.7.: Here, we present intermediate steps for one run of our active learning algorithm. The top row displays the classification results obtained after learning steps 5–8 (color coding as in figure 6.6). The bottom row shows the corresponding training utility value (TUV) maps where light areas correspond to high TUVs, i.e., points with high possible risk reduction. For instance, in step 5 the TUV is high for the pixels corresponding to the necrotic, viable and glass classes (see also figure 6.3) and a sample from the viable area is selected (indicated by the question mark). Consequently, in the next step the classification result for that class *slightly* improves and the respective TUV values decrease. Whereas the TUV values for the necrotic class also decrease since the viable and necrotic class are close in feature space, the TUV values for the glass area stay high, such that in the next step, a glass sample is picked, leading to significant improvement in the classification accuracy of that class.

set	criterion	method	50	100	150	200	reference value
S4	SE	RS	65.3	73.0	77.7	81.4	
		AL	**75.6**	**82.4**	**84.8**	**86.4**	
							91.9
	PPV	RS	76.8	83.7	**86.1**	**87.3**	
		AL	**79.4**	**84.2**	**86.1**	87.2	
							92.7
S7	SE	RS	78.3	80.6	82.0	83.1	
		AL	**85.5**	**90.3**	**92.2**	**93.6**	
							94.0
	PPV	RS	83.5	89.5	92.0	93.1	
		AL	**88.4**	**92.2**	**93.7**	**94.8**	
							97.7
S11	SE	RS	70.1	75.2	77.3	78.6	
		AL	**77.3**	**84.3**	**87.0**	**88.1**	
							91.3
	PPV	RS	77.7	83.1	86.3	87.9	
		AL	**83.4**	**89.0**	**90.8**	**92.1**	
							96.2

Table 6.1.: Average sensitivities and positive predictive values for the three datasets after 50, 100, 150 and 200 learning steps. Our active learning approach significantly outperforms random sampling on all sets (see also figure 6.4). The reference values are obtained from training a random forest classifier with 90% of all available samples per class, i.e., using up to 6,000 samples in training, and classifying the remaining (test) samples where we again averaged over 100 repeats. We note that, although our strategy performs well, the obtained accuracies after 200 iterations are still exceeded by the reference values. Possible explanations are given in the text. However, even in cases where many labels are required to learn a problem, active learning is still an efficient strategy since it requests the labels in a smart order.

tained for slice S11 proofed to be highly similar to the results for slice S7. AL-RF again outperformed RS with respect to both sensitivity and positive predictive value. After 100 and 200 learning steps it resulted in SE and PPV rates which were approximately 9% respectively 4–6% higher than the results yielded with RS. Figure 6.6 shows that RS again failed to achieve good classification performance for the necrotic class. AL-RF performed significantly better but still confused several necrotic samples with viable cancer and some with glass. Apparently, more learning steps (or better features) are necessary to learn to reliably discriminate necrotic and viable tumor in this dataset. Again, AL-RF selected more training examples from the difficult classes (necrotic, interface) and less

samples from the easier classes (gelatin, glass) than RS.

Overall, we observed that AL-RF resulted in positive predictive values which were slightly higher or competitive to the ones obtained with RS. Simultaneously, AL-RF significantly outperformed RS in sensitivity as soon as more than 15–20 learning steps were executed to train the classifier.

6.5.2. Computation Time

On a standard desktop PC with a dual core 2 GHz processor and 2 GBytes of RAM, training of the random forest and subsequent classification took less than 1 second, and Monte Carlo integration took around 1.5–2 seconds for the parameterizations and datasets used in this study. While a speed-up is of course desirable (see Outlook), this is still below the time that an expert typically needs for labeling the current query point.

6.5.3. Different Labeling Strategies

One might argue that an expert usually does not label in a completely uninformed, random way and that random sampling is thus not a realistic strategy to compare to. However, we think that it constitutes a reasonable baseline strategy, especially since it is unclear how such an alternative strategy should be defined[3]. Independent from the exact way in which an expert labels, our learning strategy has an intrinsic advantage: Although the expert might be able to provide a meaningful, initial training set, his information is limited since the stained picture only contains partial information. For instance, it is not obvious from the stain which classes are easy and/or difficult to discriminate *given the MSI data*. Whereas it may be easy to distinguish a subset of classes given a stained image, the corresponding spectra might be close in feature space (complicating the classification) or vice versa. Since this information is not available to the expert, it is unclear how many labeled examples he/she should provide for the individual classes. Another risk is that less prominent regions in the stained image might be overlooked. In that sense, the expert's decisions are uninformed and, in a way, also random. We believe that in such cases, feedback of the classifier is very helpful.

Finally, we note that our strategy does not necessarily start with an empty label set but can be initialized with a user-defined set. After assigning obvious labels, the expert can switch to our learning strategy to find additional candidates.

6.6. Outlook

Currently, we see two main directions of future research:

[3]The labeling strategy is probably different for each expert and dataset.

6.6.1. Speed-up

Computation times may be reduced by using an online version of the random forest classifier [182, 76]. Alternatively, more than one label may be queried in each active learning step before retraining the classifier. However, simply selecting the instances x with the highest TUVs (equation (6.6)) results in picking instances which are likely to be very close in feature space. One way of dealing with this problem is to introduce a heuristic that after picking the first instance decreases the TUV of other instances which are close [38] and thus reduces their chance of being selected next.

6.6.2. Fractional Labels

Furthermore, it would be interesting to incorporate support for fractional labels into our algorithm since fractional labels sometimes better reflect the true nature of a heterogeneous sample than crisp labels. Since the resolution of the stained images is usually higher than the resolution obtained with SIMS and especially MALDI imaging, estimates for the true class mixture behind individual MSI pixels could be obtained from the stained image (given that the registration between MSI data and stained slice is sufficiently accurate). Assigning such fractional labels is even more tedious than assigning crisp labels, which further emphasizes the need for active learning strategies like ours.

6.7. Conclusion

Robust training of supervised classifiers requires a set of expert labels that well reflects the true variabilities between different patients and instrument settings. Since these variabilities tend to be very high in MSI and often, data from only few patients is available, it might be beneficial to request several labels for each newly acquired dataset from an expert. However, labeling is both time-consuming and expensive. Consequently, novel algorithms are needed that yield the highest possible classification accuracies and at the same time require as few user-interaction as possible. We have demonstrated in a proof-of-concept study, how active learning can be used for the efficient annotation and classification of MSI data, especially in scenarios where the tissue samples are heterogeneous. To this end, a novel active learning approach for multi-class problems was introduced. Experiments demonstrate that it outperforms passive learning by a large margin if the same number of learning steps is used. We thus regard our study as a stepping stone toward clinical application of mass spectrometry imaging.

Appendix

6a. Derivation of the Distribution Estimate

Multinomial and Uniform Prior yields Dirichlet

Let $p(y = i|x), i = 0, 1, \ldots, D$ be the true probability that sample x is of class i, and let $v_x(i)$ be the number of trees voting for class i where $i = 0$ is the reference class. v_x can be modeled as a realization of a multinomially distributed random variable [177] with density

$$Mult(v_x(0), \ldots, v_x(D) \mid p(y = 0|x), \ldots, p(y = D|x); \nu) \tag{6.7}$$

$$= \binom{\nu}{v_x(0), \ldots, v_x(D)} \prod_{i=0}^{D} \hat{p}(y = i|x)^{v_x(i)} \tag{6.8}$$

where $\nu = \sum v_x(i)$ is the total number of trees. The Dirichlet distribution, given by

$$Dir(p(y = 0|x), \ldots, p(y = D|x) \mid \alpha_x(0), \ldots, \alpha_x(D)) \tag{6.9}$$

$$= \frac{\Gamma\left(\sum_{i=0}^{D} \alpha_x(i)\right)}{\prod_{i=0}^{D} \Gamma\left(\alpha_x(i)\right)} \prod_{i=0}^{D} \hat{p}(y = i|x)^{\alpha_x(i)-1} \tag{6.10}$$

is conjugate to the multinomial distribution [36], and uniform on the simplex for $\alpha_x(i) = 1, i = 0, \ldots, D$. Thus, applying Bayesian inference and multiplying the multinomial with the uniform prior yields the posterior distribution estimate [20]

$$\hat{F}(p(y = 0|x), \ldots, p(y = D|x)) \mid v_x(0), \ldots, v_x(D)) \propto \tag{6.11}$$

$$Mult(v_x(0), \ldots, v_x(D) \mid p(y = 0|x), \ldots, p(y = D|x); \nu) \tag{6.12}$$

$$\cdot\, Dir(p(y = 0|x), \ldots, p(y = D|x) \mid 1, ..., 1) \tag{6.13}$$

with

$$\hat{F}(p(y = 0|x), \ldots, p(y = D|x) \mid v_x(0), \ldots, v_x(D)) = Dir(1 + v_x(0), \ldots, 1 + v_x(D)). \tag{6.14}$$

In this work we assume that the probability output of the random forest classifier is a realization of the true posterior which is unknown. We note, however, that in reality, it might be corrupted, e.g., by the limited number of training samples. Furthermore, it is yet unclear if the random forest is consistent [18].

Dropping the Votes for the Reference Class

We are now interested in the distribution estimate for the scenario where the reference class (class 0) is dropped. It is known [36, 1] that for stochastically independent and Gamma-distributed $Y_k \sim Gamma(\alpha(k), 1), k = 0, \ldots, D$ it holds that

$$\left(\frac{Y_0}{\sum_{k=0}^{D} Y_k}, \frac{Y_1}{\sum_{k=0}^{D} Y_k}, \ldots, \frac{Y_D}{\sum_{k=0}^{D} Y_k}\right) \sim Dir(\alpha(0), \ldots, \alpha(D)). \tag{6.15}$$

Thus, defining

$$W_i := \frac{Y_i}{\sum_{k=0}^{D} Y_k}, i = 0, \ldots, D \tag{6.16}$$

yields that

$$(W_0, \ldots, W_D) \sim Dir(\alpha(0), \ldots, \alpha(D)). \tag{6.17}$$

Proposition: Analogously to the proof of the binary case (with $D = 2$) [177], it follows that for $\tilde{W}_i := \frac{W_i}{\sum_{l=1}^{D} W_l}$

$$(\tilde{W}_1, \ldots, \tilde{W}_D) \sim Dir(\alpha(1), \ldots, \alpha(D)). \tag{6.18}$$

Proof: Using equation (6.16) we obtain:

$$\left(\frac{W_1}{\sum_{l=1}^{D} W_l}, \ldots, \frac{W_D}{\sum_{l=1}^{D} W_l}\right) = \left(\frac{\frac{Y_1}{\sum_{k=0}^{D} Y_k}}{\sum_{l=1}^{D} \frac{Y_l}{\sum_{k=0}^{D} Y_k}}, \ldots, \frac{\frac{Y_D}{\sum_{k=0}^{D} Y_k}}{\sum_{l=1}^{D} \frac{Y_l}{\sum_{k=0}^{D} Y_k}}\right) \tag{6.19}$$

$$= \left(\frac{Y_1}{\sum_{l=1}^{D} Y_l}, \ldots, \frac{Y_D}{\sum_{l=1}^{D} Y_l}\right) \tag{6.20}$$

which is again Dirichlet-distributed with parameters $\alpha(1), \ldots, \alpha(D)$ [36].

□

One drawback of modeling the distribution estimates by Dirichlet distributions that are parameterized by the tree votes $v_x(i)$ is that by increasing the number of trees the parameters specifying the Dirichlet distributions are also increased, which results in narrower distributions. Thus, the uncertainty estimate is dependent on the number of trees and converges to zero as the number of trees goes to infinity. However, in real life applications, a fixed finite number of trees is used. Furthermore, only a relative order of uncertainty for the pixels is needed. This is guaranteed, since the same number of trees is used for each pixel within the sample. Indeed, good results were obtained in our experiments. We refer the interested reader to [176] in which an alternative modeling approach is discussed.

6b. Derivation of the Formula for Estimating the Loss for the Distribution Estimate

Above, we derived equation (6.5) for estimating the loss associated with the distribution estimate at position x. The derivation of the multi-class formula is similar to the derivation of its binary version [177]:

Let $B(\alpha) = \prod_{l=1}^{D} \Gamma(\alpha(l)) / \Gamma(\sum_{l=1}^{D} \alpha(l))$ be the multinomial Beta function and $\mathbf{1}\{\}$ the indicator function, which is 1 if the condition in brackets is fulfilled and 0 else. Further, define

$$\tilde{\alpha}_{x,i} = \begin{pmatrix} \tilde{\alpha}_{x,i}(1) \\ \vdots \\ \tilde{\alpha}_{x,i}(i) \\ \vdots \\ \tilde{\alpha}_{x,i}(D) \end{pmatrix} = \begin{pmatrix} \alpha_x(1) \\ \vdots \\ \alpha_x(i) + 1 \\ \vdots \\ \alpha_x(D) \end{pmatrix} \tag{6.21}$$

and let F be the Dirichlet density on the simplex Ψ, which is partitioned into D parts Ψ_j as described above. Finally, let I_{Ψ_j} be the multivariate equivalent of the incomplete Beta function [177, 1] where the integration over the Dirichlet distribution (defined by the parameters $\tilde{\alpha}_{x,i}$) is restricted to part Ψ_j of the simplex (see figure 6.1).

An estimate $\hat{R}_x^B$ can be obtained from

$$\hat{R}_x^B(\alpha_x(1), \ldots, \alpha_x(D)) \tag{6.22}$$

$$= \hat{d}(x) \sum_{j=1}^{D} \int_{\Psi} \sum_{i \neq j} \mathbf{1}\{\hat{p}(y|x) \in \Psi_j\} \hat{p}(y = i|x) L_{ij} d\hat{F} \tag{6.23}$$

$$= \hat{d}(x) \sum_{j=1}^{D} \sum_{i \neq j} \int_{\Psi} \mathbf{1}\{\hat{p}(y|x) \in \Psi_j\} \hat{p}(y = i|x) L_{ij} d\hat{F}. \tag{6.24}$$

By restricting the integration to the area of the simplex that corresponds to the condition in the indicator function, we obtain

$$= \hat{d}(x) \sum_{j=1}^{D} \sum_{i \neq j} \int_{\Psi_j} \hat{p}(y = i|x) L_{ij} d\hat{F}. \tag{6.25}$$

We are given a Lebesgue density for the distribution $\hat{F}$. Integration over the (Dirichlet distributed) posterior class probability estimates $\hat{p}(y|x)$ yields

$$= \hat{d}(x) \sum_{j=1}^{D} \sum_{i \neq j} \int_{\Psi_j} \psi(i) L_{ij} \left[\frac{1}{B(\alpha_x)} \prod_{l=1}^{D} \psi(l)^{\alpha_x(l)-1} \right] d\psi. \tag{6.26}$$

By defining $\tilde{\alpha}_{x,i}$ as above (cf. equation (6.21)) this can be rewritten as

$$= \hat{d}(x) \sum_{j=1}^{D} \sum_{i \neq j} \frac{B(\tilde{\alpha}_{x,i})}{B(\alpha_x)} L_{ij} \underbrace{\int_{\Psi_j} \frac{1}{B(\tilde{\alpha}_{x,i})} \prod_{l=1}^{D} \psi(l)^{\tilde{\alpha}_{x,i}(l)-1} d\psi}_{=: I_{\Psi_j}(\tilde{\alpha}_{x,i})} \tag{6.27}$$

which–using theorem 1 (cf. Appendix 6C)–is equivalent to

$$= \hat{d}(x) \left[\sum_{j=1}^{D} \sum_{i \neq j} \underbrace{\frac{\alpha_x(i)}{\sum_{k=1}^{D} \alpha_x(k)}}_{term\ 1} \underbrace{L_{ij}}_{term\ 2} \underbrace{I_{\Psi_j}(\tilde{\alpha}_{x,i})}_{term\ 3} \right]. \tag{6.28}$$

6c. Theorem 1

Proposition: Let $B(\alpha) = \frac{\prod_{l=1}^{D} \Gamma(\alpha(l))}{\Gamma(\sum_{l=1}^{D} \alpha(l))}$ be the multinomial Beta function and let $\tilde{\alpha}_{x,i}$ be defined as in equation (6.21). Then it holds that

$$\frac{B(\tilde{\alpha}_{x,i})}{B(\alpha_x)} = \frac{\alpha_x(i)}{\sum_{k=1}^{D} \alpha_x(k)}. \tag{6.29}$$

Proof:

$$\frac{B(\tilde{\alpha}_{x,i})}{B(\alpha_x)} = \frac{\Gamma(\alpha_x(1)) \cdot \ldots \cdot \Gamma(\alpha_x(i)+1) \cdot \ldots \cdot \Gamma(\alpha_x(D))}{\Gamma(\alpha_x(1) + \ldots + (\alpha_x(i)+1) + \ldots + \alpha_x(D))} \tag{6.30}$$

$$\cdot \frac{\Gamma(\alpha_x(1) + \ldots + \alpha_x(i) + \ldots + \alpha_x(D))}{\Gamma(\alpha_x(1)) \cdot \ldots \cdot \Gamma(\alpha_x(i)) \cdot \ldots \cdot \Gamma(\alpha_x(D))} \tag{6.31}$$

$$= \frac{\Gamma(\alpha_x(i)+1)}{\Gamma(\alpha_x(i))} \cdot \frac{\Gamma(\sum_{k=1}^{D} \alpha_x(k))}{\sum_k \alpha_x(k) \Gamma(\sum_{k=1}^{D} \alpha_x(k))} = \frac{\alpha_x(i)\Gamma(\alpha_x(i))}{\Gamma(\alpha_x(i))} \cdot \frac{1}{\sum_{k=1}^{D} \alpha_x(k)} \tag{6.32}$$

$$= \frac{\alpha_x(i)}{\sum_{k=1}^{D} \alpha_x(k)} \tag{6.33}$$

where we use the iterative definition of the Gamma function [20], that is $\Gamma(n+1) = n\Gamma(n)$.

□

6d. Integration over a Part of the Simplex

For each unlabeled pixel x we have to evaluate the following equation (cf. equation 6.5):

$$R_x^B(\alpha_x(1),\ldots,\alpha_x(D)) = \hat{d}(x)\left[\sum_{j=1}^{D}\sum_{i\neq j}\underbrace{\frac{\alpha_x(i)}{\sum_{k=1}^{D}\alpha_x(k)}}_{term\ 1}\underbrace{L_{ij}}_{term\ 2}\underbrace{I_{\Psi_j}(\tilde{\alpha}_{x,i})}_{term\ 3}\right] \tag{6.34}$$

$$= \hat{d}(x)\sum_{j=1}^{D}\sum_{i\neq j}\frac{B(\tilde{\alpha}_{x,i})}{B(\alpha_x)}L_{ij}\underbrace{\int_{\Psi_j}\frac{1}{B(\tilde{\alpha}_{x,i})}\prod_{l=1}^{D}\psi(l)^{\tilde{\alpha}_{x,i}(l)-1}d\psi}_{term\ 3}. \tag{6.35}$$

Terms 1 and 2 can directly be calculated (see section 6.2.2). We use Monte Carlo integration [93] to approximate term 3, which integrates a Dirichlet over a part Ψ_j of the simplex. For each x and each $i = 1,\ldots,D$, Q samples are drawn from the corresponding Dirichlet distribution.

Sampling from a Dirichlet distribution can efficiently be performed with Minka's Fastfit toolbox [150]. Inspection of the code revealed that, internally, sampling from a D-dimensional Dirichlet distribution parameterized by vector α_x boils down to drawing one sample from each of the D Gamma distributions $Gamma(\alpha_x(i),1)$ with subsequent normalization by division with the sum. Since Minka's code is fast, it can in theory be used to sequentially draw samples from Dirichlet distributions with different parameterizations α_x. However, performing these calculations independently for all pixels x is still very time-consuming. We speed up the procedure by exploiting the fact, that in our scenario the parameterizations of the individual Dirichlet distributions are highly similar. Let $v_x(i)$ be the number of trees voting for class i and let $\nu = \sum v_x(i)$ be the number of all trees. Since $\alpha_x(i) = 1 + v_x(i), i = 1,\ldots,D$, it follows that $\alpha_x(i) \in \{1,2,\ldots,\nu+1\}$. Consequently, the range of parameters is limited.

We can thus "reuse" the Gamma samples as follows: Assume, that for each pixel x we want to draw Q D-dimensional samples $u_l, l = 1,\ldots,Q$ from a Dirichlet distribution parameterized by vector α_x to perform the Monte Carlo integration. Therefore, we first draw Q samples from $\Gamma(l)\forall l = 1,\ldots,(\nu+1)$ and store the results in a $(\nu+1)\times Q$ matrix V, which serves as a look-up table. Then, the Q requested samples are "constructed" from V by first selecting the D rows that correspond to the parameters $\alpha(i), i = 1,\ldots,D$ and storing them in a $D \times Q$ matrix U. Next, we randomly permute each row of U to avoid bias in case of non-unique $\alpha_x(i)$ and do a column-wise normalization of U such that each column contains one Dirichlet sample $u_l, l = 1,\ldots,Q$[4].

[4]In our experiments, the procedure described above led to a significant speed-up factor of ≈ 100. We

Given a threshold point T, we then determine for each of the Q samples $u_k, k = 1, \ldots, Q$ to which part Ψ_j of the simplex Ψ it belongs (see figure 6.1). The set Z_j of samples $\tilde{u}$ that fall in part Ψ_j can be expressed by (cf. section 7.2.3 for details)

$$Z_j = \left\{ \tilde{u} \in \{u_1, \ldots, u_Q\} \left| 1 - \frac{\tilde{u}(j)}{\tilde{u}(j) + \tilde{u}(i)} < 1 - \frac{T(j)}{T(j) + T(i)} \forall i \in \{1, ..., D\} \setminus \{j\} \right. \right\}. \tag{6.36}$$

Note that in case of 0-1 loss the threshold point resides in the center of the simplex such that the formula simplifies and we only have to determine the column-wise maxima of U in order to calculate the assignments. Finally, the estimate for $I_{\Psi_j}(\tilde{\alpha}_{x,i})$ is calculated as $card(Z_j)/Q$ where $card(.)$ is the cardinality operator.

propose to construct V (and thus U) anew in each learning step of the active learning procedure.

Chapter 7

Multivariate Watershed Segmentation of Compositional Data

Recent improvements in instrumentation allow mass spectrometry imaging at cell scale [9, 41]. An important application of microscopic imaging is cell screening [23]. Typical research questions include the localization and quantitation of different kinds of cells in a sample. Especially in high-throughput settings, automated analysis is indispensable. Whereas the localization problem can be addressed with classification algorithms, the latter usually requires an additional segmentation step that ensures that cells that are close and have been assigned to the same class are indeed recognized as individuals. However, segmentation of multivariate data is challenging, especially if some of the channels are uncorrelated as in MS images. This motivates the development of novel segmentation techniques, e.g., based on the classic watershed transform.

Both probabilistic latent semantic analysis (pLSA, cf. chapter 3) and random forests (cf. chapters 4,5,6) can be used to assign class probabilities to the pixels of a mass spectrometry image. Such data, where the probability vector in each pixel sums to one, is termed *compositional* [5]. We propose three novel methods to obtain scalar boundary indicator maps from compositional data. We furthermore introduce the multivariate watershed which is a multi-class generalization of the classic watershed transform. A quantitative comparison to three established algorithms–Noyel's [157] sum/supremum of channel-wise morphological gradients and Malpica's [137] tensor approach–demonstrates good performance on both simulated and real-world MSI data.

7.1. Introduction

Segmentation is an important task in image processing (see, e.g., [58, 123, 159, 221]). The idea is to partition an input image into disjoint regions where the regions themselves show (some degree of) homogeneity with respect to a criterion like gray value or texture. Segmentation methods can roughly be grouped into two categories: energy-based segmentation and watershed-based segmentation [156]. In the following we will concentrate on the latter. The watershed transform [61] is a region-based segmentation algorithm for gray-scale images and is a popular method in image segmentation [179]. Figuratively speaking, the gray-valued boundary indicator image is considered as a height map which is flooded with water. Whenever two water basins that originate from different local minima meet, a watershed is constructed. Various definitions for the continuous and discrete case have been given, most of which operate on scalar-valued input. Typically the gradient is used as a boundary indicator, featuring high values at border locations and low values in homogeneous areas.

High-dimensional data is typically transformed into a scalar boundary map, such that the conventional watershed can be applied. A direct generalization for color images is the color gradient. Alternatively, the watershed transform is individually performed on each color band, and the obtained segmentation results are integrated into a single segmentation map [129]. Some authors have suggested to use color models that feature an intensity channel [39, 129]. However, most of these approaches are based on the assumption that color channels are highly correlated. This is typically not the case for more complex data , e.g., from remote sensing [105], medical data analysis [159], and mass spectrometry imaging experiments [142]. In fluorescence microscopy, experimentalists often deliberately select dyes that highlight different, uncorrelated structures [151].

To analyze such kind of data, authors have used (weighted) channel-wise gradients [185, 130, 157] or the metric-based gradient [157]. Noyel [157] proposed the sum or supremum of channel-wise morphological gradients that are computed as difference between channel dilation and erosion. Malpica [137] and Karvelis [115] construct the Jacobian matrix J of partial derivatives (in direction of the two spatial axes) in each pixel and calculate the eigenvalues of $J'J$. The difference between the two eigenvalues is used as a boundary indicator. Zhang [236] uses the spectral angle between a pixel's underlying spectrum and a reference spectrum to obtain a scalar boundary indicator. Angulo proposed a stochastic watershed algorithm [12] which was later extended to multispectral images [158]. Soille [202] combines a histogram-based clustering with shape priors and pixel-wise gradients which are used as input for the classic watershed transform. Authors have also suggested dimensionality reduction [157, 158, 159] prior to segmentation to reduce noise artifacts.

7.2. Material and Methods

7.2.1. Watershed Segmentation of Compositional Data

The discrete and sequential watershed methods on scalar-valued input can be grouped into watershed by immersion [222] and watershed by topographic distance [147]. The following review is based on the presentation in [179]. Typically, it is assumed that the scalar-valued input image is lower-complete, that is each pixel that is not a minimum has a neighbor of lower gray value[1]. Let $G = (C, E, f)$ be a digital gray value image in graph representation with nodes C and edges E, and let $f : C \to \mathbb{N}$ be a function defined on C that assigns an integer value to each pixel $u \in C$. In most applications, a square grid in conjunction with a 4- or 8-connectivity neighborhood system is used.

Watershed by Immersion. Let h_{min} and h_{max} denote the minimum and maximum value of f, B_h the union of the set of basins $B_{h,i}, i = 1, ..., k$ at level h, MIN_h the union of regional minima at level h and $Q_h = \{u \in C | f(u) \leq h\}$ the threshold set of f at level h. A recursion is defined with gray level h increasing from h_{min} to h_{max}. Each connected component in Q_{h+1} can either be a new minimum or an extension of a basin in B_h. In the latter case, the assignments of points to basins is based on calculating geodesic influence zones $iz_{Q_{h+1}}$ of B_h within Q_{h+1}:

$$iz_{Q_{h+1}}(B_{h,i}) = \bigcup_{i=1}^{k} \left\{ u \in Q_{h+1} | \forall j \in \{1, ..., k\} \setminus \{i\} : d_{Q_{h+1}}(u, B_{h,i}) < d_{Q_{h+1}}(u, B_{h,j}) \right\} \tag{7.1}$$

where $d_{Q_{h+1}}(u, B_{h,i})$ is the distance within Q_{h+1} between pixel u and the point in $B_{h,i}$ that is closest to u. $IZ_{Q_{h+1}}(B_h)$ is the union of the geodesic influence zones of the basin set B_h in Q_{h+1}:

$$IZ_{Q_{h+1}}(B_h) = \bigcup_{i=1}^{k} iz_{Q_{h+1}}(B_{h,i}). \tag{7.2}$$

Essentially, for each point the shortest path within Q_{h+1} to a basin in B_h is calculated. If multiple shortest paths to different basins exist, the pixel is considered a watershed. This can be expressed with the following recursion formula [179]:

$$\begin{cases} B_{h_{min}} = \{u \in C | f(u) = h_{min}\} = Q_{h_{min}} \\ B_{h+1} = MIN_{h+1} \cup IZ_{Q_{h+1}}(B_h) \text{ for } h \in [h_{min}, h_{max}) \end{cases} \tag{7.3}$$

where $f(u)$ is the intensity at pixel u. Equation (7.3) results in the watershed image $Wshed(f) = C \setminus B_{h_{max}}$.

The flooding of the (scalar-valued) height map f starts from the pixels with the lowest intensity ($Q_{h_{min}}$, "seeds"). The basins are initialized and successively extended. In each

[1] Otherwise, an algorithm for obtaining lower completeness has to be applied first[179].

recursion step, first all pixels with intensities less or equal to the new height level $h+1$ are identified (Q_{h+1}). Whenever a pixel lies within the geodesic influence zone of an existing basin, it is merged with the respective basin. Pixels that lie in the influence zone of two or more basins are preliminarily labeled watersheds, but this assignment can change if higher levels are analyzed. Connected components that are not adjacent to any existing basin are used as seeds for new basins.

More generally spoken, the algorithm uses a priority queue that determines the sequence in which the individual pixels are processed. The method runs until a basin or watershed label has been assigned to all pixels. In some variants of the algorithm, the watersheds are not identified with the pixels but defined to lie between pixels.

Watershed by Topographical Distance. Watershed by topographical distance [147] is based on gradient-descent. Figuratively spoken, rain falls on the height map f. After hitting the surface in a pixel u a raindrop proceed along the path of steepest descent until a local minimum is reached. This can be formalized in the following way: Let $LS(u)$ ("lower slope") be the maximum slope of f linking a pixel u to any of its neighbors $neigh_C(u)$ of lower gray value

$$LS(u) = \underset{v \in neigh_C(u) \cup \{u\}}{\text{MAX}} \frac{f(u) - f(v)}{d(u,v)} \tag{7.4}$$

where $d(u,v)$ is a distance between pixels u and v which depends on the pixel grid and the associated neighborhood system (e.g., 4-connectivity). In the following, we assume $d(u,v) = 1 \; \forall v \in neigh_C(u)$. Equation (7.4) is zero if $f(u)$ is a local minimum. The cost for walking from u to a neighboring pixel v is defined by

$$cost(u,v) = \begin{cases} LS(u) \cdot d(u,v) & \text{if } f(u) > f(v) \\ LS(v) \cdot d(u,v) & \text{if } f(u) < f(v) \\ \frac{1}{2}(LS(u) + LS(v)) \cdot d(u,v) & \text{if } f(u) = f(v) \end{cases} \tag{7.5}$$

The topographic distance between $u^0 = u$ and $u^s = v$ along a path $\pi = (u^0, \ldots, u^s)$ is calculated from [19]

$$\tau_f^\pi(u,v) = \sum_{i=0}^{s-1} cost(u^i, u^{i+1}) \tag{7.6}$$

and the topographical distance between two points u and v is the length of the shortest path between the two points. A path π is of steepest descent if and only if $\tau_f^\pi(u,v) = f(u) - f(v)$ for $f(u) > f(v)$ [179].

An efficient algorithm proceeds as follows: for each pixel we store a reference to the neighboring pixel with the maximum slope. Minima pixels point towards themselves, and a distinct label is assigned to each local minimum. Starting from the minima, the constructed paths are traversed in reverse order, and all pixels are labeled with the corresponding label. In this implementation, the resulting watersheds lie *between* pixels (inter-pixel boundaries), and the algorithm has linear complexity in the number of pixels.

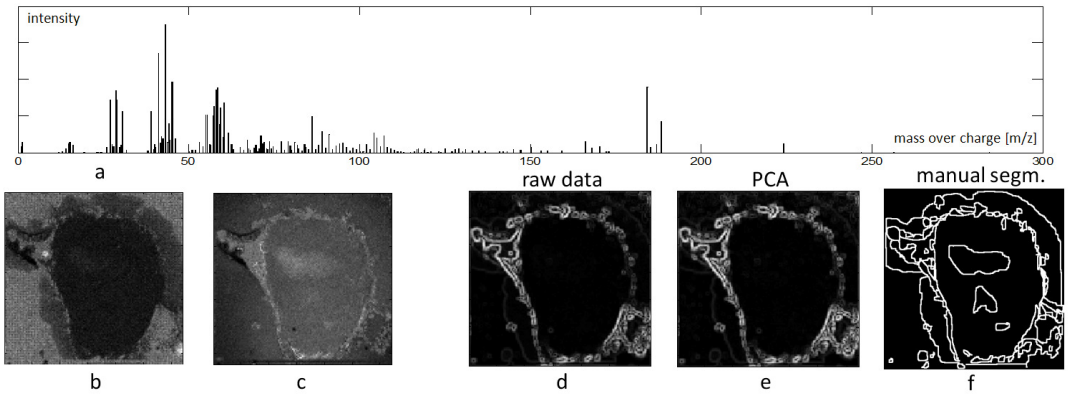

Figure 7.1.: An excerpt from a typical mass spectrum (a) and two channels of the MSI dataset (b,c). Only parts of the images are correlated. A boundary map is obtained with the sum of channel-wise morphological gradients on the raw data (d) or the first few principal components (e), revealing only few boundaries of the manual segmentation (f).

Compositional Data. In the following, we focus on watershed by topographic distance and work with spectral data in two spatial dimensions. Note, however, that our results carry over to the watershed by immersion setting and that the methods presented below are also applicable to other high-dimensional data. Let $S = \{(x^1, y^1), \ldots, (x^N, y^N)\}$ be a set of available M-dimensional training samples, i.e., N spectra x^i with M channels and corresponding class labels $y^i \in L_1, \ldots, L_D$–e.g., "cancerous", "healthy tissue", and "blood vessels" in medical applications or "faulty" or "intact" in quality control. We train a random forest classifier [25] (see section 4.2.1) on S and classify the whole dataset comprising K data points to obtain the posterior probability for each of the D classes in each pixel. The resulting dataset has D dimensions in each pixel and is compositional since each probability vector sums to one.

We use the slice S7 from the SIMS dataset described in sections 3.3.2 and 4.3.1 to illustrate the performance of the different methods. Figure 7.1a shows an example spectrum of this set. The dataset features five different tissue classes of interest and thus, random forest classification yields a compositional dataset with five dimensions in each pixel. Noyel's sum of morphological gradients [157] (see below) was used to calculate boundary indicator maps directly from the raw input and the first principal components of the raw input (see figure 7.1d+e) as well as from the probability maps (see figure 7.2). Visual inspection suggests that the latter contains more information. Therefore, we used probability maps as input for all following boundary map computations.

7.2.2. Multivariate Gradients

We next present three methods from the literature as well as three novel methods that create scalar boundary maps from such compositional input data. We then introduce the multivariate watershed, which is a generalization of the classic watershed definition.

Methods by Noyel and Malpica. Noyel's sum of morphological gradients [157] is calculated from

$$SMG(f_u) = \sum_{j=1}^{D} grad_{morph}(f_u(j)) \tag{7.7}$$

where $f_u(j)$ is the j-th value in vector f_u at pixel u and $grad_{morph}$ is the morphological gradient. The morphological gradient is a marginal gradient and is defined as the difference between channel dilation and channel erosion [157, 111][2]. The supremum version can be obtained analogously. Note however that each morphological gradient must be normalized between $[0, 1]$ before the supremum is taken. In Malpica's tensor approach [137], first the Jacobian matrix J of partial derivatives

$$J = \begin{bmatrix} \frac{\delta f_u(1)}{\delta x} & \frac{\delta f_u(1)}{\delta y} \\ \frac{\delta f_u(2)}{\delta x} & \frac{\delta f_u(2)}{\delta y} \\ \dots & \dots \\ \frac{\delta f_u(D)}{\delta x} & \frac{\delta f_u(D)}{\delta y} \end{bmatrix} \tag{7.8}$$

is constructed in each pixel u. Next, the eigenvalues of $J'J$ are calculated, and the difference of the two eigenvalues $\lambda_1 - \lambda_2$ is used as a boundary indicator for pixel u.

These latter three methods will be used to evaluate our novel boundary indicators which will be presented next.

Gini Impurity. The underlying idea for the Gini impurity watershed is that the class impurity in the classification results can be used to identify borders between different regions. Essentially, Gini impurity [25] is a measure of vector sparseness. For probability distribution f_u at pixel location u it is defined by

$$GI(f_u) = \sum_{j=1}^{D} f_u(j)(1 - f_u(j)) = 1 - \sum_{j=1}^{D} f_u(j)^2 \tag{7.9}$$

where again $f_u(j)$ is the j-th value in f_u. The minimum degree of impurity is obtained if one of the $f_u(j)$ equals one. We calculate the Gini impurity index for each pixel and perform the conventional watershed segmentation on the obtained scalar boundary map. Slight smoothing of the probability maps with a channel-wise Gaussian filter (zero mean, unit variance) prior to calculation of the impurity indices preserves the

[2]Note that f_u is not required to be compositional in this approach.

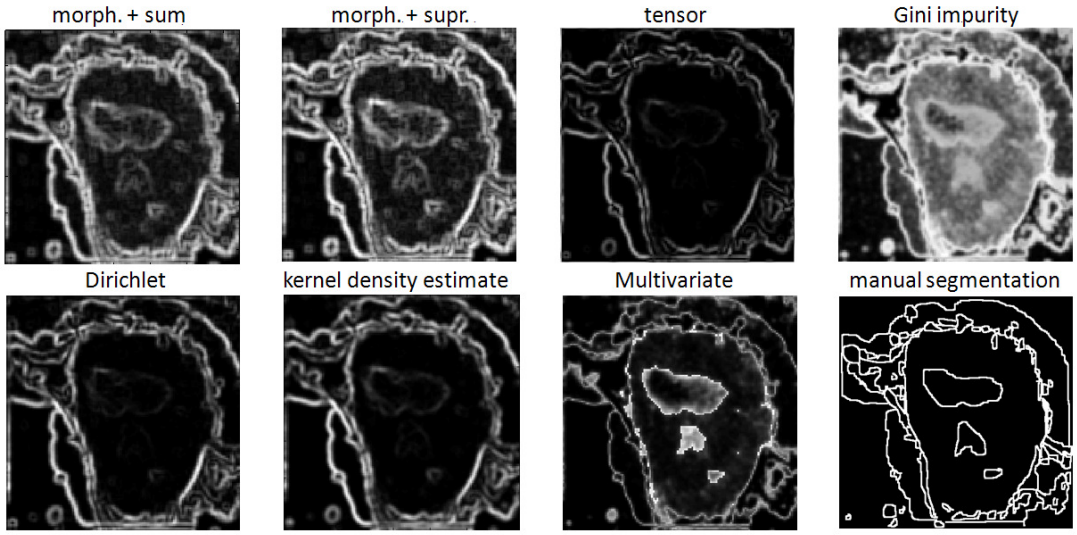

Figure 7.2.: The probability map based boundary indicators obtained with Noyel's sum/supremum of channel-wise morphological gradients, Malpica's tensor and the methods introduced here. The result of an approximate manual labeling is also shown.

sum-constraint and ensures that the border between two adjacent points with (different) pure components has indeed higher impurity.

Dirichlet Boundary Indicator. The observed probability vectors at pixel u and its neighbors $neigh(u)$ can be interpreted as realizations of a (single) Dirichlet distribution [36]. Its D-dimensional realizations sum to one as do the class probabilities for each pixel. The Dirichlet distribution is parameterized by a vector $\alpha = (\alpha(1), \ldots, \alpha(D))$ with $\alpha(j) > 0 \; \forall j = 1, \ldots, D$, i.e., $f \sim Dir$ with

$$f(x(1), ..., x(D)|\alpha(1), ..., \alpha(D)) = \frac{\Gamma\left(\sum_{j=1}^{D} \alpha(j)\right)}{\prod_{j=1}^{D} \Gamma\left(\alpha(j)\right)} \prod_{j=1}^{D} x(j)^{\alpha(j)-1} \tag{7.10}$$

for all $x(j) > 0$ with $\sum_j^D x(j) = 1$ where Γ is the Gamma function. For given observations at u and its neighbors $neigh(u)$, its optimal parameters $\hat{\alpha}_u$ can be estimated by maximizing the log-likelihood of the data [36]. This maximum likelihood estimation is performed in a neighborhood of each pixel. The obtained parameters $\hat{\alpha}_u(j)$ determine the shape of the distribution. At pixel locations within homogeneous regions of the spectral image their sum is high, and we obtain highly peaked distributions with low variances. In contrast, in the vicinity of borders the sum of the $\hat{\alpha}_u(j)$ is low and the distributions are broad. Thus, we propose to use the inverted precision–defined as 1 divided by the sum of all $\hat{\alpha}_u(j)$–as boundary indicator at pixel location u. The resulting

boundary indicator map is used as input for the classic watershed.

Kernel Density Estimate. This method performs a kernel density estimation [89] in each pixel u. Here we use Gaussian kernels k, but other choices are possible. The density KDE at a pixel u is calculated from

$$KDE(u) = \frac{c}{K\sigma_{spat}^2\sigma_{prob}^D} \sum_{v=1}^{K} \underbrace{k_{spat}\left(\left\|\frac{g_u - g_v}{\sigma_{spat}}\right\|^2\right)}_{\text{term 1}} \underbrace{k_{prob}\left(\left\|\frac{f_u - f_v}{\sigma_{prob}}\right\|^2\right)}_{\text{term 2}} \tag{7.11}$$

where v iterates over the K pixels of the MSI dataset, g_v is the two-dimensional vector of spatial coordinates corresponding to pixel v, and c is a normalization constant. Contributions are weighted by the distance in space (term 1) as well as by distance in composition (term 2). σ_{spat} and σ_{prob} are the corresponding (Gaussian) kernel bandwidths. The inverse density is used as a boundary indicator map for the classic watershed. The density estimation formula in equation (7.11) is well known from the literature. It constitutes a link of the watershed algorithm with the mean shift procedure [49] and bilateral filtering [213].

7.2.3. Multivariate Watershed

In the classic watershed algorithm only one type of basin is known. For compositional data, we generalize this by introducing one basin type per class, that is basin types $B_1, \ldots, B_D$. Each pixel is assigned to the class (basin type) whose posterior probability dominates. In absence of prior knowledge, dominance is established by a simple "winner takes all" rule (corresponding to a fair arbitrator); but certain classes *can* be favored if desired by introducing bias (corresponding to a bribed arbitrator). We first consider the two-class case (cf. figure 7.3). For pixel u the probabilities for classes 1 and 2 are given by $f_u(1)$ and $f_u(2) = 1 - f_u(1)$. Maximum likelihood estimation (MLE) for the assignment of pixels to basin types corresponds to introducing a threshold $T = 0.5$ where u is assigned to basin type 1, formally stated as $w_T(u) = B_1$, if $f_u(1) \geq T$ and to B_2 otherwise. By moving T, different risks can be assigned to the two classes. The basin assignment is then obtained in a risk-weighted maximum a-posteriori decision.

When compositional data of higher dimension is analyzed, the points live on a simplex. In the $3D$ case, the MLE-based assignment of points to basins is defined by the threshold point $T = [1/3; 1/3; 1/3]$ and the perpendiculars to the lines connecting the corners of the simplex (see figure 7.9). By moving T on the simplex, the emphasis of the different classes can be controlled. If the threshold point is close to one corner of the simplex, the corresponding class will be less influential in the segmentation. In the D-dimensional case, the set of pixels that are assigned to basin B_k is given by $Z_k := \{u \mid w_T(u) = B_k\}$,

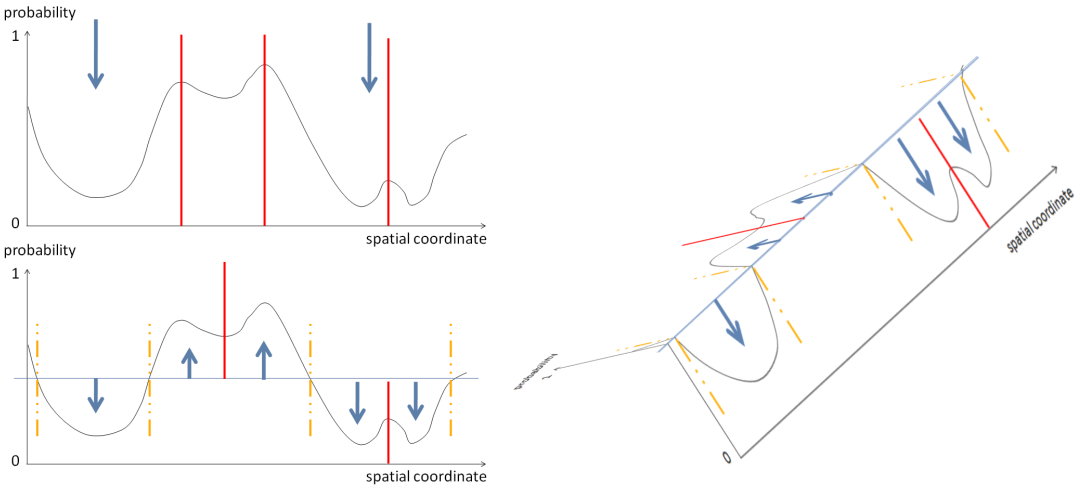

Figure 7.3.: Visualization of the multivariate watershed for $D = 2$. The classic approach (top left) only has one basin type. Figuratively, rain falls from above and fills the basins. Whenever two basins meet, a watershed is constructed (vertical lines). In the multivariate case (bottom left), a threshold is introduced that controls the assignment of each point to one of the two basin types. Figuratively, the rain now falls from a higher dimension onto a surface that has been folded at the threshold value (right). New watersheds emerge for the parts above the threshold and between basins of different type. Classic watershed emerges as a special case in which the threshold has been set to the maximum intensity.

that is

$$Z_k = \left\{ u \left| \underbrace{1 - \frac{f_u(k)}{f_u(k) + f_u(j)}}_{=:\theta_{f_u,k\to j}} < \underbrace{1 - \frac{T(k)}{T(k) + T(j)}}_{=:\theta_{T,k\to j}} \forall j \in \{1, ..., D\} \setminus \{k\} \right. \right\}. \tag{7.12}$$

Consider the following example in three dimensions: Let $T = [1/4; 3/8; 3/8]$ be the threshold point. Assume that we want to identify the separation lines that determine the assignment of points to the three basin types represented by $B_1 = [1; 0; 0]$, $B_2 = [0; 1; 0]$ and $B_3 = [0; 0; 1]$. Using equation (7.12), we calculate the pairwise class-ratios of the

threshold point: $\theta_{T,1\to2} = 1 - (1/4)/(1/4 + 3/8) = 1 - 2/5 = 3/5$, $\theta_{T,2\to1} = 2/5$, $\theta_{T,1\to3} = 3/5$, $\theta_{T,3\to1} = 2/5$, $\theta_{T,2\to3} = 1 - 1/2 = 1/2$ and $\theta_{T,3\to2} = 1/2$. A point $f_{\tilde{u}}$ is assigned to basin B_1, that is $f_{\tilde{u}} \in Z_1$, if its ratios are component-wise smaller or equal than the corresponding threshold point's ratios. This is, e.g., the case for $f_{\tilde{u}} = [1/3; 1/3; 1/3]$ for which $\theta_{f_{\tilde{u}},1\to2} = \theta_{f_{\tilde{u}},1\to3} = 1/2 \leq \theta_{T,1\to2} = \theta_{T,1\to3} = 3/5$. By default, we use $T = [1/D; ...; 1/D] \in \mathbb{R}^D$.

Note that the assignment of a point $f_{\tilde{u}}$ to a basin B_k can efficiently be found in $D-1$ pairwise comparisons. In each step, we compare the ratios in (7.12) with respect to two dimensions, e.g. $k = 1$ and $j = 2$. The dimension for which the inequality holds (here 1 or 2) is declared the "winner" which in the next step is compared to the next dimension ($k = winner(\{1, 2\}), j = 3$) and so on.

For each set of points $Z_k, k = 1, \ldots, D$ the distance between each point in the set and the respective basin is calculated. On the resulting scalar map, a watershed-like segmentation is performed. It differs from the conventional watershed transform in the following way: The basin assignments define the areas of influence for each basin type. Figuratively, water is never allowed to cross borders between different influence zones. This can be incorporated into the conventional watershed algorithm by changing its distance function $d(u, v)$ such that points that have been assigned to different basin types have an infinite distance: $d(u, v) = \infty \; \forall v \in neigh(u) : w_T(v) \neq w_T(u)$.

7.3. Experiments

We used the topographic distance version of the watershed algorithm and a neighborhood system with 8-connectivity. For the real-world data studied at the end of this section, no exact ground truth is available (label assignment is especially difficult in the border areas, see also figure 4.3). For a quantitative evaluation, we hence resort to simulated data, which is described next.

7.3.1. Data

Simulated Data. First, three spectra from a real-world mass spectrometry image were taken and defined as "pure" spectra (see figure 7.1a for an example). Then, different mixture maps were generated that contain pure areas as well as impure ones (see figure 7.4). The "observed" data was created by mixing the pure spectra according to the mixture maps. Again, a Poisson noise model was used to simulate instabilities in the data acquisition process (cf. section 3.3.1). Samples from the noisy dataset were used to train a random forest with 250 trees and 100 samples per class. After training, the whole dataset was classified and a set of probability maps was obtained. A total of three experiments was performed: In the first experiment, the mixture map contained pure *and* impure regions, but training was performed with samples from the pure regions only (yielding a 3 class problem). For experiment 2, the same setting as in experiment

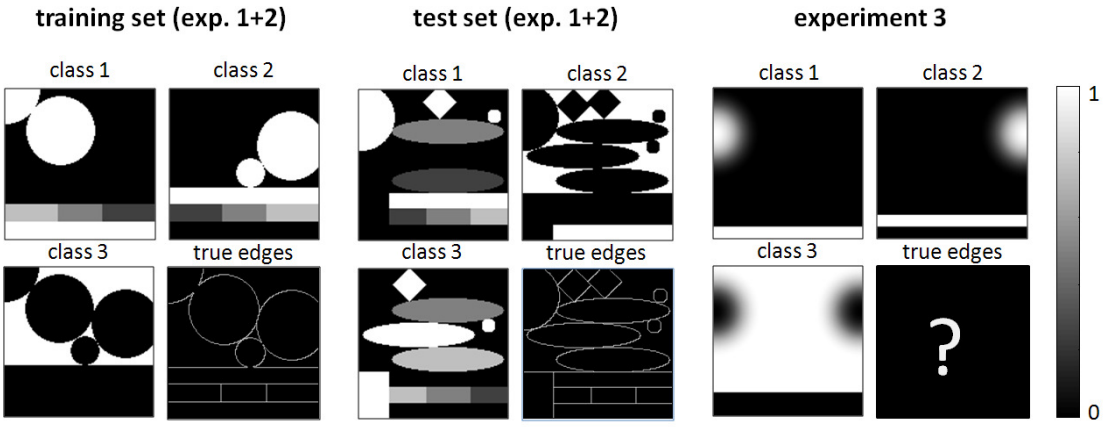

Figure 7.4.: The figure shows the ground truth mixture maps that were used for simulating data. In all experiments, we used three ground truth spectra and mixed them according to the mixture maps shown above. White areas correspond to a pure concentration, black indicates that this class is absent at the respective location. The correct boundaries–if unambiguous–are also given.

1 was used, but each mixture area was considered an individual class, and the classifier was trained with samples from each of them (yielding a 6 class problem). The last experiment demonstrates the influence of the threshold of the multivariate watershed on the obtained segmentation.

Real-World Data. We also applied the methods to the SIMS MSI data of slice S7 (cf. section 4.3.1). Two of its over 4,000 mass over charge channels are shown in figure 7.1. A random forest classifier with 250 trees and 100 samples per class was used.

7.3.2. Postprocessing

The obtained segmentation maps typically contain oversegmentation. We use watershed dynamics [91, 30] to amend this problem. For each edge, that is for each pair of adjacent basins, a dynamics value is calculated from the distances between the basins' minima and the minimum height of the dividing edge. Edges with a dynamics value that is below a given threshold are removed, and the respective basins are merged.

7.3.3. Evaluation Criteria

The true edges as well as the watersheds obtained with the topographic distance watershed algorithm lie between the pixels of the grid. We use inter-pixel edges represented on a half-integer grid to quantitatively compare the estimated edge map E_{est} with the ground truth edge map E_{gt}. The Baddeley distance [13] over all pixels v is employed as

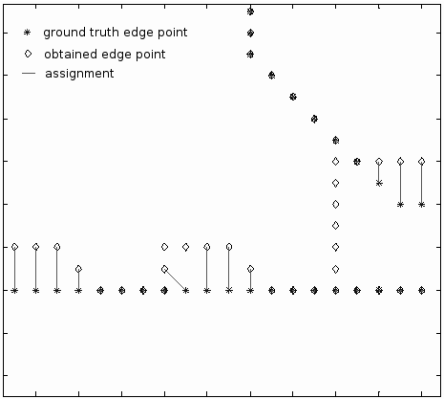

Figure 7.5.: The Gale-Shapley algorithm for solving the stable marriage algorithm is used to assign points in the obtained edge map E_{est} to points in the ground truth edge map E_{est_i}. Note that not all points have a partner.

a dissimilarity measure:

$$dist_{Bad}(E_{gt}, E_{est}) = \left[\frac{1}{K} \sum_{v=1}^{K} |DT(v, E_{gt}) - DT(v, E_{est})|^q \right]^{1/q}. \tag{7.13}$$

$DT(v, E)$ is the closest distance between pixel v and any of the edges in the edge map E [181]. In our study, q was set to 2, and we only considered distances up to 5 pixels, i.e., pixels that were more than 5 pixels away from any edge in both of the edge maps were ignored. The Baddeley distance penalizes both over- and undersegmentation. Since the optimal dynamics threshold value is different for the various watershed algorithms under consideration, for each method the best dynamics threshold with respect to the Baddeley distance was chosen.

Besides using the Baddeley distance, we quantify the segmentation quality by means of sensitivity and specificity. The former measures which percentage of the true edges are identified by a method, the latter how many background points are wrongly classified as edges. Since the edges in the obtained segmentation maps E_{est_i} can be slightly displaced from their true positions in E_{gt}, we first match them to the ground truth edges. To this aim, we employ the Gale-Shapley algorithm [77] that is best known for solving the stable marriage problem. This method uniquely assigns each pixel in E_{est_i} to a close pixel in E_{gt} as long as the maximum distance is below a given threshold, here 2 pixels (see figure 7.5). Edge pixels in E_{gt} without a "partner" in E_{est_i} are considered false negatives (FN, indicating undersegmentation), edge pixels in E_{est_i} without a partner in E_{gt} are false positives (FP, oversegmentation), pairs are considered true positives (TP), and the rest are true negatives (TN). We then calculate $sensitivity = \frac{TP}{TP+FN}$ and

$specificity = \frac{TN}{FP+TN}$. An estimate for the test error is obtained by testing the watershed methods on an independent test image set using the same parameter settings. Training and test set differ in geometry and class mixtures (cf. figure 7.4).

7.4. Results

The kernel density estimation watershed features two kernel parameters σ_{spat} and σ_{prob} that need to be specified. To allow for a fair comparison, we calculated the segmentations for a variety of different parameter settings from a given range, and the best settings for the training set were used. In experiment 1, the best choices were $\sigma_{spat} = 2.0$ and $\sigma_{prob} = 1.0$, and in experiment 2 the best settings proved to be $\sigma_{spat} = \sigma_{prob} = 1.0$. Similarly, different structuring elements can be used for the calculation of the morphological gradient. We experimented with discs of varying sizes and found a radius of 1 to be most adequate. Results for the three experiments are given in figures 7.6,7.7,7.8,7.97.10 as well as in table 7.1.

7.5. Discussion

We next discuss the outcome of the experiments described in section 7.3.

7.5.1. Experiment 1: 3-Class Problem

Regarding the training set, the results of Malpica's tensor and especially the Dirichlet approach are very close to the ground truth (cf. figure 7.6 and table 7.1). The Dirichlet boundary indicator reliably finds edges between different pure mixtures and impure mixtures and results in straight boundary lines (cf. figure 7.7a). The optimal dynamics threshold for the kernel density-based method leads to oversegmentation in the lower right part of the image, but the remaining part of the image is well segmented. The Gini impurity watershed and the multivariate watershed accurately identify boundaries between pure mixture areas but have some problems to detect boundaries that separate impure regions (cf. figure 7.7b). This task is indeed difficult for most of the methods since the classifier output shows a relatively low gradient in these areas. The Gini impurity watershed itself cannot detect edges between pure and highly impure regions since highly impure mixtures have a high Gini boundary indicator and are therefore interpreted as boundaries instead of regions. However, in the postprocessing step, some of these boundaries are removed by merging basins, and the real edges can be recovered. The multivariate watershed assigns both of the two regions in the lower left of figure 7.7c to the same basin type (class). Consequently, the boundary pixels are reduced to jumps in the distance function, which are used to construct the boundary indicator maps. The oversegmentation of the 50% to 50% mixture area results from the fact that

depending on the Poisson noise, the pixels are randomly assigned to classes 1 and 2. Here, some smoothing prior to the watershed segmentation could improve results. In contrast to all other methods (including the summation of morphological gradients), the multivariate watershed is able to reconstruct the contours in areas of narrow bends (cf. figure 7.7d+e).

On the test set, the morphological gradients, the Dirichlet boundary indicator and the Gini impurity perform best. The latter leads to the best distance value, partly because it is least oversegmented. However, some of the boundaries are less accurate but displaced by a few pixels.

7.5.2. Experiment 2: 6-Class Problem

One can argue that the 3 mixture areas (with 25%, 50% and 75% contributions) show clear spatial extent and constitute classes in their own right. Thus, in experiment 2 we trained the classifier with samples from 6 classes. Table 7.1 and figure 7.8 show that in this scenario both the Gini impurity and the multivariate watershed achieve good to very good results. Especially the edges between pure and impure areas are now much better identified by the Gini impurity watershed (cf. figure 7.7f). The multivariate watershed even results in a perfect reconstruction of all boundaries and thus in the best sensitivity and specificity values. The kernel density estimate and Dirichlet boundary indicators again compete well with the existing methods.

7.5.3. Experiment 3: Influence of the Threshold Parameter

Figure 7.9 shows how the threshold parameter of the multivariate watershed can be used to influence the segmentation result. The threshold is varied such that the proportions of classes 2 and 3 remain constant. In contrast, the first class is emphasized compared to the other two. It can be seen from the two segmentation results that the border between classes 2 and 3 remains unchanged whereas the border between classes 1 and 3 is shifted in favor of class 1. By setting the threshold, the user can control the class weights. Thus, the multivariate watershed provides the user with more control over the segmentation result than other methods like the watershed based on morphological gradients.

7.5.4. Real-World Data

Figure 7.10 shows the results of the different watershed methods applied to the MSI dataset (postprocessed with watershed dynamics). The optimal parameters have been tuned manually. Judging from visual inspection and considering the approximate manual segmentation (cf. figure 7.2), all approaches lead to reasonable results. However, some oversegmentation remains that cannot be removed with the concept of watershed dynamics.

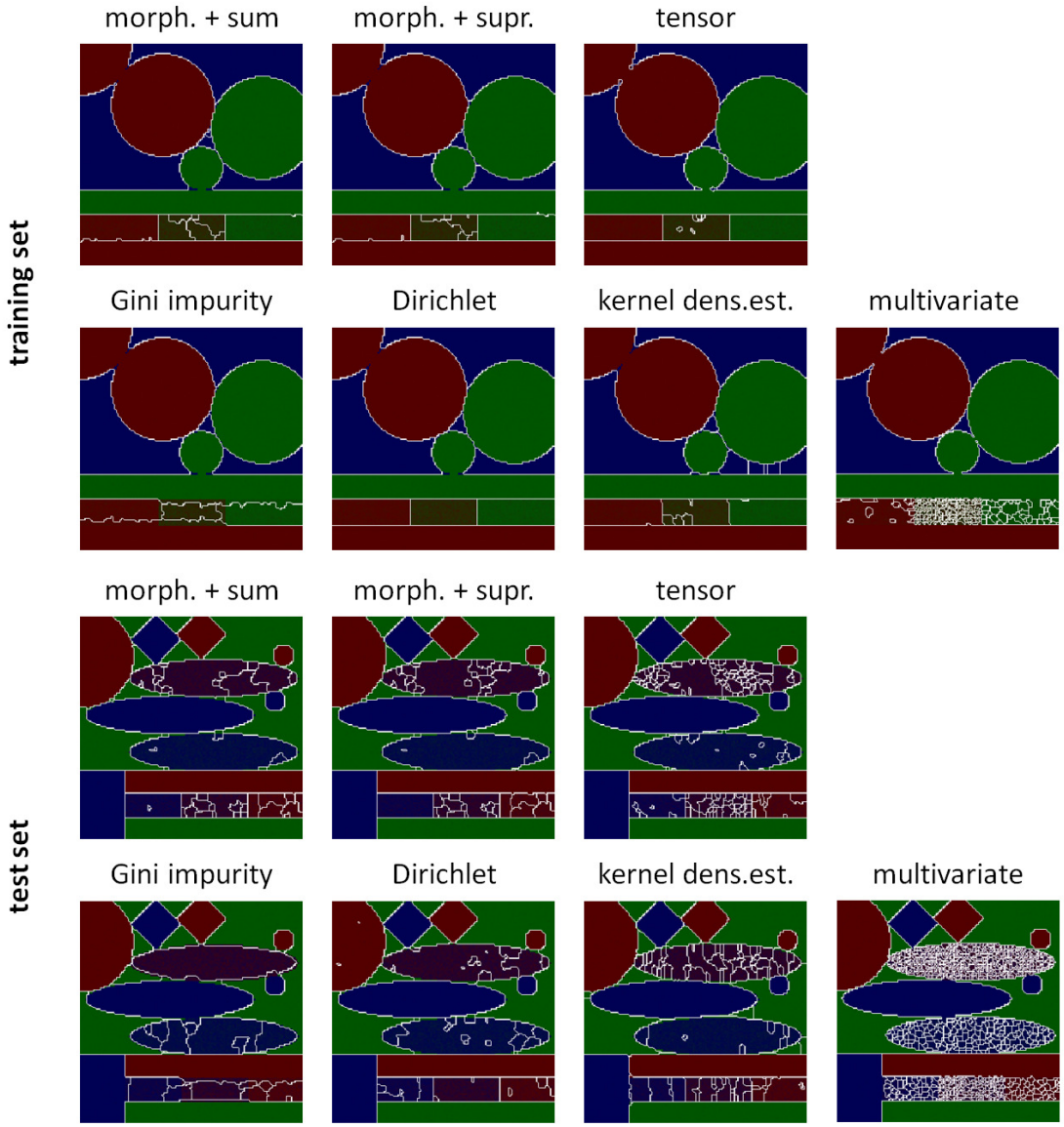

Figure 7.6.: Experiment 1: Segmentation results on training (top) and test set (bottom) after training of the classifier with samples from pure mixture regions only.

7.6. Conclusion

We have introduced four watershed-based methods for the segmentation of multivariate compositional data: the Gini impurity watershed, the Dirichlet boundary indicator

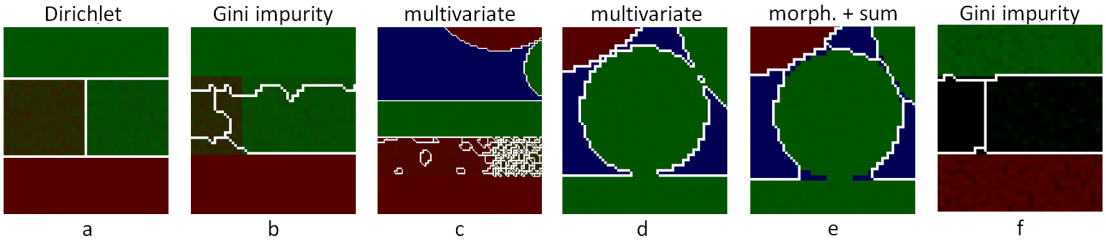

Figure 7.7.: Experiment 1: Zoomed-in areas of the segmentation results. See text for interpretations.

		Experiment 1		
method	threshold	distance	sensitivity	specificity
morphological+sum	0.0502	2.4958	0.9396	0.9864
morphological+supremum	0.0752	2.4835	0.9423	0.9855
tensor	0.0019	3.1687	0.9710	0.9702
Gini impurity	0.0576	**2.3467**	0.8734	**0.9906**
Dirichlet	0.0033	2.5908	0.9311	0.9882
kernel density estimation	0.0013	3.2484	0.9206	0.9757
multivariate	0.0118	4.7705	**0.9965**	0.8917
		Experiment 2		
method	threshold	distance	sensitivity	specificity
morphological+sum	0.0502	0.2193	0.9408	0.9995
morphological+supremum	0.0752	0.3150	0.9253	0.9997
tensor	0.0019	0.2000	0.9648	0.9996
Gini impurity	0.0576	0.3755	0.9199	0.9999
Dirichlet	0.0033	0.2955	0.9210	0.9999
kernel density estimation	0.0013	0.4054	0.9001	0.9998
multivariate	0.0118	**0.0**	**1.0**	**1.0**

Table 7.1.: The results obtained on the test set show that there is no clear winner. However, the methods introduced in this chapter compete well with existing ones.

watershed, the kernel density estimate-based watershed and the multivariate watershed. The former three approaches use novel techniques to obtain a scalar boundary indicator map that is used as an input for the classic watershed transform. The latter generalizes the definition of the classic watershed for multispectral compositional data. In our experiments on simulated and real-world MSI data, no overall best performing method could be identified. However, the methods introduced in this chapter have been shown to compete well with existing methods and are superior in some scenarios.

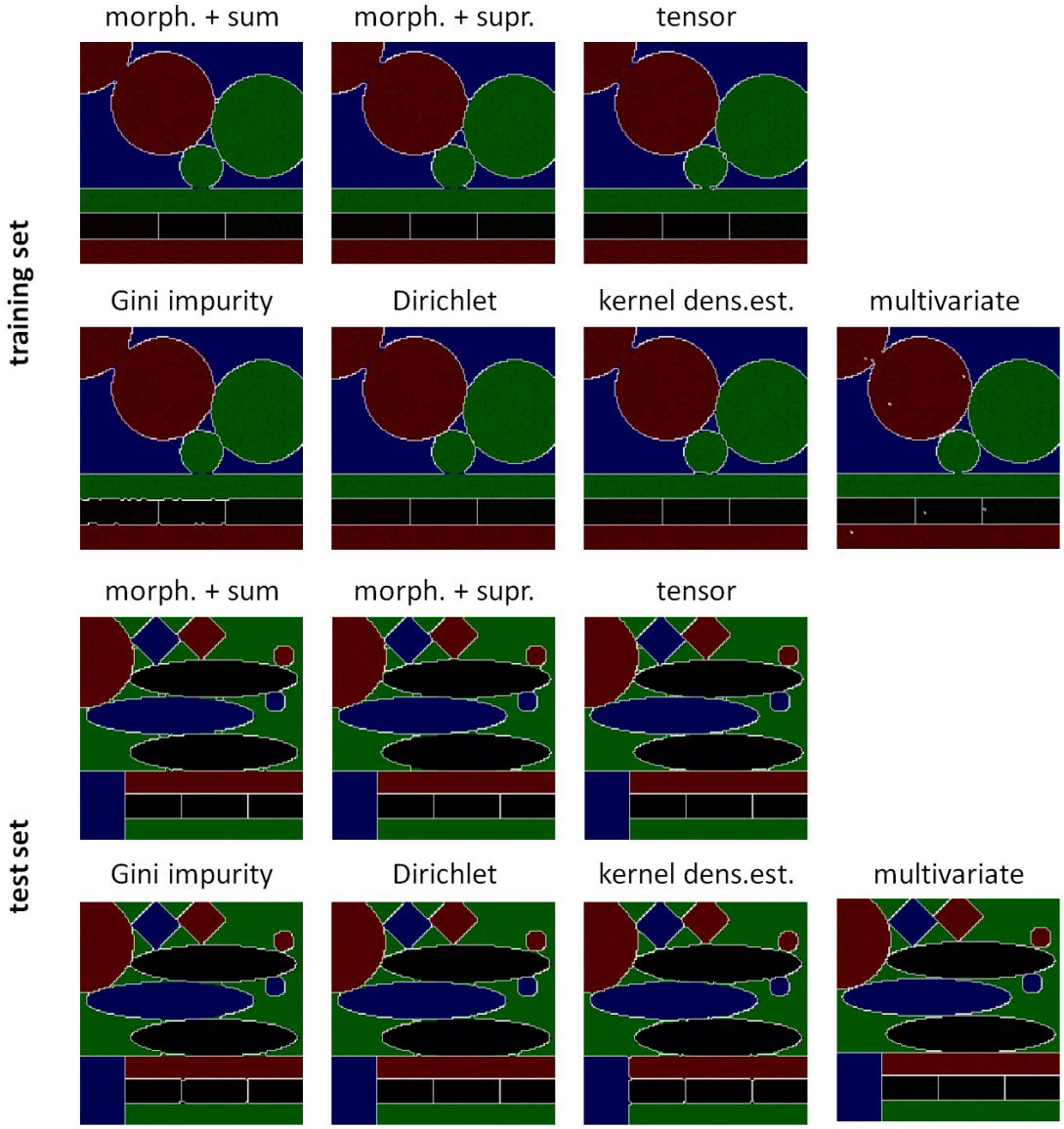

Figure 7.8.: Experiment 2: Segmentation results on training (top) and test set (bottom) after training of the classifier with samples from six classes.

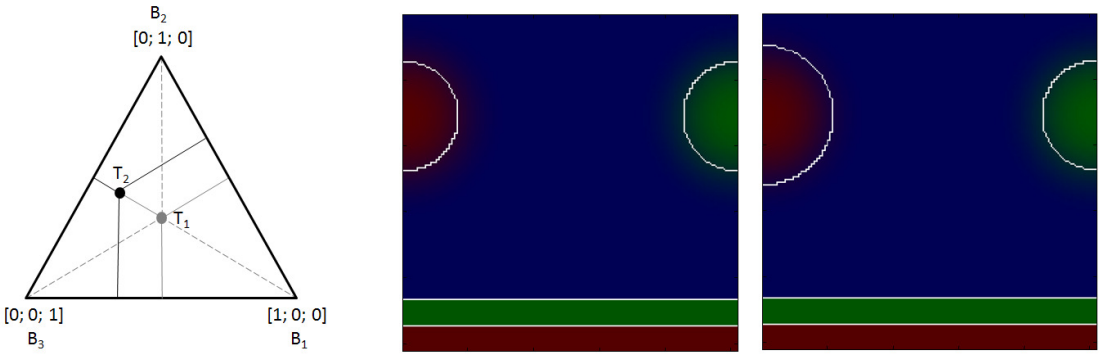

Figure 7.9.: Experiment 3: The user can influence the outcome of the segmentation by selecting the threshold parameter of the multivariate watershed. Changing the threshold from $T_1 = [1/3; 1/3; 1/3]$ to $T_2 = [1/7; 3/7; 3/7]$ corresponds to shifting the threshold on the simplex (left) and leads to an accentuation of class 1 (red). The boundaries between classes 2 and 3 remain unchanged.

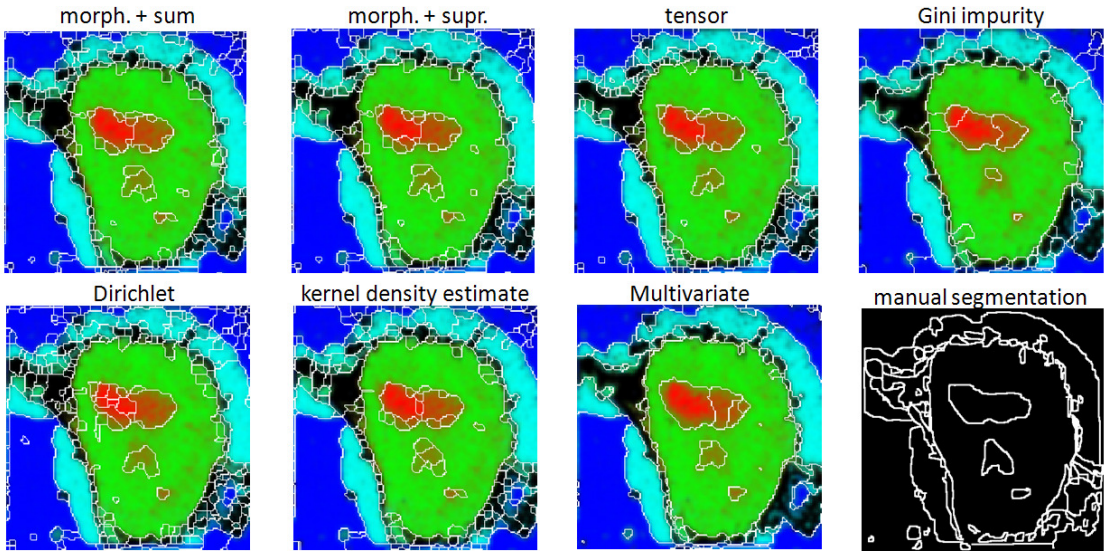

Figure 7.10.: Segmentation results on the real-world data with manually selected parameters.

Chapter 8

Conclusions and Perspectives

The potential of mass spectrometry imaging (MSI) has been demonstrated in many case studies (e.g., [92, 127, 190, 191, 233, 32]), and the technology has been attributed as "very promising" [161] or even "pioneering" [126]. However, it is doomed to fail without the availability of efficient, reliable, and robust algorithms that can handle the enormous amount of data produced by state-of-the-art instruments [66].

In this thesis, we have developed and deployed novel methods for the analysis of mass spectrometry images. First, we have proposed the application of probabilistic latent semantic analysis (pLSA) for concise representation of mass spectrometry images, and have shown that this method constitutes a valuable tool for unsupervised decomposition of MSI data in exploratory settings. We have then successfully used the random forest classifier for automated annotation of MSI data ("digital staining") as well as analyzed the influence of the preprocessing methods on the obtainable classification results. To reduce labeling costs and time, we have established a novel multi-class active learning algorithm for efficient annotation and classification of large MSI datasets. Our method is especially useful if the variability between the analyzed datasets is high or the classifier has to be trained anew on each set. A natural step after classification is to segment the data into coherent regions. For this purpose, a multivariate generalization of the popular classic watershed transform as well as three novel strategies to create boundary indicator maps were introduced.

Detailed discussions of these methods as well as concluding remarks can be found at the end of the respective chapters and are omitted here. Instead, we focus on interesting directions of future MSI research.

8.1. Standardization and Improved Reproducibility

MS and MSI are continuously moving closer to clinical application [73, 161, 193, 192, 223]. However, we observe that reproducibility is still a major challenge in MSI ex-

periments (cf. chapter 5). Depending on the instrument and the spatial and spectral resolution at which the tissue is analyzed, the data may be characterized by rather low signal to noise ratios and high variability between measurements. Recent efforts thus seek to standardize the output of proteomics experiments to improve data quality and comparability [210, 200, 65]. Similar considerations exist for mass spectrometry imaging [24, 74, 90, 135]. This is an important stepping stone on the way to clinical application of MSI.

8.2. Advances in Biomarker Discovery

The determination of biomarkers for diseases like cancer or Alzheimer plays a pivotal role in clinical research. In this thesis we have employed the mean decrease accuracy criterion to identify biomarker candidates for different kinds of human breast cancer tissue based on high-dimensional MSI data. However, although good results were obtained, this feature importance score does not explicitly consider correlation effects which are intrinsic to MSI data. It would thus be interesting to explore recently introduced alternative approaches [207, 62, 143, 241] that have been reported to deliver superior performance under potentially strong feature correlation.

8.3. Survival Analysis and Personalized Medicine

Another main direction of current MS and MSI research is personalized medicine. For instance, mass spectra can be correlated with survival information or information on therapy success [191, 233]. The hope is to obtain reliable estimates for the survival time of a patient and to determine which therapy is the most promising one. Here, the development and application of novel methods like random survival forests [108] may be promising.

8.4. Improved Feature Extraction, Identification, and Quantitation

The continuous increase in spatial and spectral resolution makes high demands on data analysis procedures but at the same time takes MSI analysis to a new level. For instance, improvements in the spectral resolution of MS images will finally allow for feature extraction with advanced peak pickers like Nitpick [173] or THRASH [102] that require isotope resolution. In this setting, spatially regularizing the peak picking, e.g., by using an adaptive lasso [240]-based approach, may make feature extraction more robust to noise.

In recent years, the combination of MSI with identification by means of MS^2 has become increasingly popular [135]. By acquiring MS^2 information for the most prominent peaks in the MS^1 spectra of an MS image, identification of compounds becomes possible. However, acquiring MS^2 information for all major peaks, that is in all spectra, is extremely time-consuming, and the acquisition times can turn into a serious bottleneck. Here, the application of active learning algorithms that exploit the redundancy in MSI datasets may be instrumental. With a smart selection of spatial positions and precursor ions, the acquisition of MS^2 spectra might be guided such that acquisition times are significantly reduced while all relevant information is kept.

With recent advances in technology, the quantitative side of mass spectrometric analysis is starting to take center stage (e.g., see [237, 226, 149]). Local quantitation is not yet possible with MSI [98] but is likely to draw major attention as soon as the technology makes this possible.

Frequently used Abbreviations

AIC	Akaike Information Criterion
AL	Active Learning
BIC	Bayesian Information Criterion
Da	Dalton
FN	False Negatives
FP	False Positives
FT-ICR MS	Fourier Transform Ion Cyclotron Resonance Mass Spectrometry
HE	Hematoxylin and Eosin
HER2	Human Epidermal growth factor Receptor 2
ICA	Independent Component Analysis
KL divergence	Kullback-Leibler divergence
LC/MS	Liquid Chromatography Mass Spectrometry
MALDI	Matrix Assisted Laser Desorption/Ionization
MLE	Maximum Likelihood Estimation
MRF	Markov Random Field
MS	Mass Spectrometry
MS^2	synonym for tandem MS or MS/MS
MSI	Mass Spectrometry Imaging (also: imaging mass spectrometry)
NN-PARAFAC	Non-Negative PARAllel FACtor analysis

Table 8.1.: Frequently used abbreviations (1).

PCA	Principal Component Analysis
pLSA	probabilistic Latent Semantic Analysis
ppm	parts per million
PPV	Positive Predictive Value
RS	Random Sampling
SE	SEnsitivity
SIMS	Secondary Ion Mass Spectrometry
SSL	Semi-Supervised Learning
SVM	Support Vector Machine
TIC	Total Ion Count (also: Total Ion Current)
TN	True Negatives
TOF	Time Of Flight
TP	True Positives
VVM	Vector-Valued Median

Table 8.2.: Frequently used abbreviations (2).

List of Tables

List of Figures

List of Publications

In Preparation

- M. Hanselmann, J. Röder, U. Köthe, B.Y. Renard, R.M.A. Heeren, F.A. Hamprecht. Active Learning for Efficient Annotation and Classification of Imaging Mass Spectrometry Data.

- M. Hanselmann, B. Balluff, A. Walch, F.A. Hamprecht (preliminary list of authors). Differential Diagnostics of Breast Cancer using MALDI MS Imaging: The Role of Preprocessing and Technical and Biological Variability.

- J. Röder, B. Nadler, M. Hanselmann, F.A. Hamprecht. Bayesian Distributional Uncertainty Estimates for Local Averaging Classifiers.

- M. Hanselmann*, B. Voss*, B.Y. Renard, M.S. Lindner, U. Köthe, M. Kirchner, F.A. Hamprecht. SIMA: Simultaneous Multiple Alignment of LC/MS Peak Lists, submitted to Bioinformatics.

Journals

- M. Hanselmann, U. Köthe, M. Kirchner, B.Y. Renard, E.R. Amstalden, K. Glunde, R.M.A. Heeren, F.A. Hamprecht (2009). Toward Digital Staining using Imaging Mass Spectrometry and Random Forests, Journal of Proteome Research 8(7):3558–3567.

- M. Hanselmann, M. Kirchner*, B.Y. Renard*, E.R. Amstalden, K. Glunde, R.M.A. Heeren, F.A. Hamprecht (2008). Concise Representation of Mass Spectrometry Images by Probabilistic Latent Semantic Analysis, Analytical Chemistry 80(24):9649–9658.

Book Chapters

- M. Hanselmann, U. Köthe, B.Y. Renard, M. Kirchner, R.M.A. Heeren, F.A. Hamprecht (2009). Multivariate Watershed Segmentation of Compositional Data. Proceedings of the 15th Int. Conf. on Discrete Geometry for Computer Imagery (DGCI), Montreal, Canada; Lecture Notes in Computer Science 5810:180–192, Springer. (proceedings and talk)

Extended Abstracts

- M. Hanselmann and F.A. Hamprecht (2011). Automated Analysis of Mass Spectrometric Images. 2nd Int. Workshop on Protein Analysis of Tissues, Munich, Germany.

- M. Hanselmann, J. Röder, U. Köthe, B.Y. Renard, A. Kreshuk, R.M.A. Heeren, F.A. Hamprecht (2010). Active Learning for Efficient Labeling and Classification of Imaging Mass Spectrometry Data. Proc. of the 58th ASMS Conf. on Mass Spectrometry and Allied Topics, Salt Lake City, Utah, USA.

- A. Kreshuk, M. Kirchner, B.Y. Renard, D. Winter, B.X. Kausler, X. Lou, M. Hanselmann, J.A.J. Steen, H. Steen, W.D. Lehmann, F.A. Hamprecht (2010). Automatic Relative Quantification for High-Resolution LC/MS 16/18O Labeling Experiments. Proc. of the 58th ASMS Conf. on Mass Spectrometry and Allied Topics, Salt Lake City, Utah, USA.

- D.F. Smith, M.C. Duursma, M. Hanselmann, F.A. Hamprecht, N.A. Giese, R.M.A. Heeren (2009). Imaging Mass Spectrometry for the Characterization of Human Pancreatic Disease. 18th Int. Mass Spectrometry Conf. (IMSC), Bremen, Germany.

- D.F. Smith, M. Hanselmann, F.A. Hamprecht, N.A. Giese, R.M.A. Heeren (2009). FT-ICR Imaging Mass Spectrometry for the Characterization of Pancreatic Disease. 7th North American FT MS Conf., Key West, Florida, USA.

- D.F. Smith, M. Hanselmann, F.A. Hamprecht, N.A. Giese, R.M.A. Heeren (2009). MALDI FT-ICR MS and SIMS-TOF for Chemical and Spatial Characterization of Pancreatic Disease. Symp. of the Dutch Mass Spectrometry Society (NVMS) and the Belgium Society for Mass Spectrometry (BSMS), Kerkrade, The Netherlands.

- M. Hanselmann, U. Köthe, M. Kirchner, B.Y. Renard, E.R. Amstalden, R.M.A. Heeren, F.A. Hamprecht (2009). Automated Classification and Grading of Tumors in Mass Spectrometric Images using postprocessed Random Forests. Proc. of

the 57th ASMS Conf. on Mass Spectrometry and Allied Topics, Philadelphia, Pennsylvania, USA.

- B.M. Voss, B.Y. Renard, A. Kreshuk, M. Hanselmann, U. Köthe, H. Steen, J.A.J. Steen, M. Kirchner, F.A. Hamprecht. (2009). Simultaneous Multiple Alignment for LC/MS Peak Lists. Proc. of the 57th ASMS Conf. on Mass Spectrometry and Allied Topics, Philadelphia, Pennsylvania, USA.

- D.F. Smith, M.C. Duursma, M. Hanselmann, F.A. Hamprecht, N.A. Giese, R.M.A. Heeren (2009). FT-ICR and SIMS-TOF Imaging Mass Spectrometry for the Characterization of Human Pancreatic Disease. Proc. of the 57th ASMS Conf. on Mass Spectrometry and Allied Topics, Philadelphia, Pennsylvania, USA.

- M. Hanselmann, M. Kirchner*, B.Y. Renard*, A. Kharchenko, L.A. Klerk, U. Köthe, R.M.A. Heeren, F.A. Hamprecht (2008). Concise Representation Of MS Images By Probabilistic Latent Semantic Analysis. Proc. of the 56th ASMS Conf. on Mass Spectrometry and Allied Topics, Denver, Colorado, USA.

- M. Hanselmann, C. Winter, T. Wittenberg, T. Zerfaß (2007). Auflösungssteigerung für die Mikroskopie am Beispiel der Hämatologie. 41. DGBMT Jahrestagung, Aachen, Germany. (extended abstract and talk)

- F. Müller, M. Hanselmann, T. Liebig, O. Noppens (2006). A Tableaux-based Mobile DL Reasoner - An Experience Report. Proc. of the Int. Workshop on Description Logics, Lake District, United Kingdom.

Review

- CVPR 2010: IEEE Conf. on Computer Vision and Pattern Recognition.

* contributed equally

Bibliography

[1] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions.* Dover, 1965.

[2] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair. A comparison of machine learning techniques for phishing detection. *ACM Proc. of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit*, 8:60–69, 2007.

[3] R.P. Adams. *Identification of Essential Oil Components by Gas Chromatography/Mass Spectrometry.* Allured Publication Corporation, 4 edition, 2007.

[4] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422:198–207, 2003.

[5] J. Aitchison. *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability.* Chapman and Hall, 1986.

[6] H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, 19:716–723, 1974.

[7] A.F.M. Altelaar. *Biomolecular Imaging Mass spectrometry: mapping molecular distributions in cells and tissue sections.* PhD thesis, University of Amsterdam, The Netherlands, 2007.

[8] A.F.M. Altelaar, S.L. Luxembourg, L.A. McDonnell, S.R. Piersma, and R.M.A. Heeren. Imaging mass spectrometry at cellular length scale spatial resolution. *Nature Protocols*, 2:1185–1196, 2007.

[9] A.F.M. Altelaar, I.M. Taban, L.A. McDonnell, P.D.E.M. Verhaert, R.P.J. de Lange, R.A.H. Adan, W.J. Mooi, R.M.A. Heeren, and S.R. Piersma. High-resolution MALDI imaging mass spectrometry allows localization of peptide distributions at cellular length scales in pituitary tissue sections. *International Journal of Mass Spectrometry*, 260:203–211, 2007.

[10] E.R. Amstalden van Hove, D.F. Smith, and R.M.A. Heeren. A concise review of mass spectrometry imaging. *Journal of Chromatography A*, 1217(25):3946–3954, 2010.

[11] M. Andersson, M.R. Groseclose, A.Y. Deutch, and R.M. Caprioli. Imaging mass spectrometry of proteins and peptides: 3D volume reconstruction. *Nature Methods*, 5(1):101–108, 2008.

[12] J. Angulo and D. Jeulin. Stochastic watershed segmentation. *Int. Symp. on Mathematical Morphology*, 8:265–276, 2007.

[13] A.J. Baddeley. An error metric for binary images. *Robust Computer Vision*, pages 59–78, 1992.

[14] J.H. Barrett and D.A. Cairns. Application of the random forest classification method to peaks detected from mass spectrometric proteomic profiles of cancer patients and controls. *Statistical Applications in Genetics and Molecular Biology*, 7(2), 2008. article 4.

[15] A.M. Belu, M.C. Davies, J. M. Newton, and N. Patel. TOF-SIMS characterization and imaging of controlled-release drug delivery systems. *Analytical Chemistry*, 72(22):5625–5638, 2000.

[16] F. Benvenuto, A. La Camera, C. Theys, A. Ferrari, H. Lanteri, and M. Bertero. The study of an iterative method for the reconstruction of images corrupted by Poisson and Gaussian noise. *Inverse Problems*, 24:1–20, 2008.

[17] J.M. Berg, J.L. Tymoczko, and L. Stryer. *Biochemistry.* Palgrave Macmillian, 5 edition, 2002.

[18] G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9:2015–2033, 2008.

[19] A. Bieniek and A. Moga. A connected component approach to the watershed segmentation. *Proc. of the 4th Int. Symp. on Mathematical Morphology and its Applications to Image and Signal Processing*, pages 215–222, 1998.

[20] C.M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2007.

[21] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[22] S. Boppel, B.Y. Renard, M. Kirchner, H. Steen, U. Köthe, and F.A. Hamprecht. Sparse profile reconstruction for LC/MS feature extraction. *56th ASMS Conf. on Mass Spectrometry and Allied Topics*, 2008.

[23] K. Börner, J. Hermle, C. Sommer, N.P. Brown, B. Knapp, B. Glass, J. Kunkel, G. Torralba, J. Reymann, N. Beil, J. Beneke, R. Pepperkok, R. Schneider, T. Ludwig, M. Hausmann, F.A. Hamprecht, H. Erfle, L. Kaderali, H.-G. Kräusslich, and M. Lehmann. From experimental setup to bioinformatics: An RNAi screening platform to identify host factors involved in HIV-1 replication. *Biotechnological Journal*, 5(1):39–49, 2010.

[24] H. Brandt, T. Ehmann, and M. Otto. Toward prediction: Using chemometrics for the optimization of sample preparation in MALDI-TOF MS of synthetic polymers. *Analytical Chemistry*, 2010. DOI: 10.1021/ac101526w.

[25] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[26] L. Breiman. Consistency of a simple model of random forests. Technical report, University of California, Berkeley, 2004.

[27] R. Bro. *Multi-way Analysis in the Food Industry - Models, Algorithms, and Applications*. PhD thesis, University of Amsterdam, The Netherlands, 1998.

[28] R. Bro and C.A. Andersson. N-way toolbox. http://www.models.kvl.dk/source/nwaytoolbox/index.asp, 2007.

[29] A. Broersen, R. van Liere, and R.M.A. Heeren. Comparing three PCA-based methods for the 3D visualization of imaging spectroscopy data. *Proc. of the 5th IASTED Int. Conf. on Visualization, Imaging, and Image Processing*, pages 540–545, 2005.

[30] L. Brun, M. Mokhtari, and F. Meyer. Hierarchical watersheds within the combinatorial pyramid framework. *Proc. of the 12th Int. Conf. on Discrete Geometry for Computer Imagery (DGCI), Lecture Notes in Computer Science (LNCS)*, 2005.

[31] K.P. Burnham and D.R. Anderson. *Model Selection and Multimodel Inference: A Practical-Theoretic Approach*. Springer, 2nd edition, 2002.

[32] R.L. Caldwell and R.M. Caprioli. Tissue profiling by mass spectrometry - a review of methodology and applications. *Molecular and Cellular Proteomics*, 4:394–401, 2005.

[33] M. Cannataro, P.H. Guzzi, G. Tradigo, and P. Veltri. On the preprocessing of mass spectrometry proteomics data. *Neural Nets: 16th Italian Workshop on Neural Nets, WIRN/NAIS 2005, Lecture Notes in Computer Science (LNCS)*, 3931:127–131, 2005.

[34] R.M. Caprioli, T.B. Farmer, and J. Gile. Molecular imaging of biological samples: Localization of peptides and proteins using MALDI-TOF MS. *Analytical Chemistry*, 69:4751–4760, 1997.

[35] J.D. Carroll and J.J. Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. *Psychometrika*, 35:283–819, 1970.

[36] G. Casella and R.L. Berger. *Statistical Inference*. Duxbury Advanced Series, 2002.

[37] R. Castaing and G.J. Slodzian. Optique corpusculaire - premiers essais de microanalyse par emission ionique secondaire. *Microscopie*, 1:395–399, 1962.

[38] N. Cebron and M.R. Berthold. Active learning for object classification: from exploration to exploitation. *Data Mining and Knowledge Discovery*, 18(2):283–299, 2009.

[39] C.-Y. Chang, W.-S. Shie, and J.-H. Wang. Color image segmentation via fuzzy feature tuning and feature adjustment. *IEEE Conf. on Systems, Man and Cybernetics*, 4:6, 2002.

[40] O. Chapelle, A. Zien, and B. Schölkopf. *Semi-Supervised Learning*. MIT Press, 2006.

[41] P. Chaurand, K.E. Schriver, and R.M. Caprioli. Instrument design and characterization for high resolution MALDI-MS imaging of tissue sections. *Journal of Mass Spectrometry*, 42:476–489, 2007.

[42] P. Chaurand, S.A. Schwartz, and R.M. Caprioli. Imaging mass spectrometry: a new tool to investigate the spatial organization of peptides and proteins in mammalian tissue sections. *Current Opinion in Chemical Biology*, 6:676–681, 2002.

[43] P. Chaurand, M. Stoeckli, and R.M. Caprioli. Direct profiling of proteins in biological tissue sections by MALDI mass spectrometry. *Analytical Chemistry*, 71:5263–5270, 1999.

[44] C. Chen, A. Liaw, and L. Breiman. Using random forest to learn imbalanced data. Technical report, University of California, Berkeley, 2004.

[45] S. Cho, S.G. Park, D.O.H. Lee, and B.C. Park. Protein-protein interaction networks: from interactions to networks. *Journal of Biochemistry and Molecular Biology*, 37(1):45–52, 2004.

[46] K. Chughtai and R.M.A. Heeren. Mass spectrometric imaging for biomedical tissue analysis. *Chemical Reviews*, 110:3237–3277, 2010.

[47] A. Cichocki and R. Zdune. Regularized alternating least squares algorithms for non-negative matrix/tensor factorization. *Proc. of the 4th Int. Symp. on Neural Networks: Advances in Neural Networks III, Lecture Notes in Computer Science (LNCS)*, 2007:793–802, 2007.

[48] M.P. Colombini. *Organic Mass Spectrometry in Art and Archaeology.* John Wiley and Sons, 2009.

[49] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Trans. on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[50] European Commission. Fundamental genomics website. http://ec.europa.eu/research/health/genomics/tabs/gene_exp_prot_en.htm, 2009. accessed September 2009.

[51] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[52] K.R. Coombes, S. Tsavachidis, J.S. Morris, K.A. Baggerly, M.C. Hung, and H.M. Kuerer. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, 5(16):4107–4117, 2005.

[53] D.S. Cornett, S.L. Frappier, and R.M. Caprioli. MALDI-FTICR imaging mass spectrometry of drugs and metabolites in tissue. *Analytical Chemistry*, 80(14):5648–5653, 2008.

[54] D.S. Cornett, M.L. Reyzer, P. Chaurand, and R.M. Caprioli. MALDI imaging mass spectrometry: molecular snapshots of biochemical systems. *Nature Methods*, 4:828–833, 2007.

[55] A.C. Crecelius, D.S. Cornett, R.M. Caprioli, B. Williams, B.M. Dawant, and B. Bodenheimer. Three-dimensional visualization of protein expression in mouse brain structures using imaging mass spectrometry. *Journal of the American Socitey for Mass Spectrometry (JASMS)*, 16(7):1093–1099, 2005.

[56] S. Datta and L.M. DePadilla. Feature selection and machine learning with mass spectrometry data for distinguishing cancer and non-cancer samples. *Statistical Methodology*, 3(1):79–92, 2006.

[57] S.-O. Deininger, M.P. Ebert, A. Fütterer, M. Gerhard, and C. Röcken. MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *Journal of Proteome Research*, 7(12):5230–5236, 2008.

[58] A. Delong and Y. Boykov. Globally optimal segmentation of multi-region objects. *Int. Conf. on Computer Vision (ICCV)*, pages 285–292, 2009.

[59] J. D'Errico. Shape language modeling (SLM) toolbox. http://www.mathworks.com/matlabcentral/fileexchange/24443-slm-shape-language-modeling, 2009.

[60] R. Diaz-Uriate and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3):1–13, 2006.

[61] H. Digabel and C. Lantuéjoul. Iterative algorithms. *Actes du Second Symp. Europèen d'Analyse Quantitative des Microstructures en Sciences des Matériaux, Biologie et Médecine*, pages 85–99, 1978.

[62] D. Donoho and J. Jin. Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences (PNAS)*, 105(39):14790–95, 2008.

[63] K. Downard. *Mass Spectrometry of Protein Interactions*. John Wiley and Sons, 2007.

[64] F. Dubois, R. Knochenmuss, R. Zenobi, A. Brunelle, C. Deprun, and Y.L. Beyec. A comparison between ion-to-photon and microchannel plate detectors. *Rapid Communications in Mass Spectrometry*, 13(9):786–791, 1999.

[65] M.W. Duncan, R. Aebersold, and R.M. Caprioli. The pros and cons of peptide-centric proteomics. *Nature Biotechnology*, 28(7):659–664, 2010.

[66] I. Eidhammer, K. Flikka, L. Martens, and S.-O. Mikalsen. *Computational Methods for Mass Spectrometry Proteomics*. John Wiley and Sons, 2007.

[67] G.B. Eijkel, B. Kükrer-Kaletas, I.M. van der Wiel, J.M. Kros, T.M. Luider, and R.M.A. Heeren. Correlating MALDI and SIMS imaging mass spectrometric datasets of biological tissue surfaces. *Surface and Interface Analysis*, 41:675–685, 2009.

[68] P.H.C. Eilers and H.F.M. Boelens. Baseline correction with asymmetric least squares smoothing. Technical report, University of Amsterdam, 2005.

[69] J.B. Fenn, M. Mann, C.K. Meng, S.F. Wong, and C.M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246:64–71, 1989.

[70] D. Figeys, L.D. McBroom, and M.F. Moran. Mass spectrometry for the study of protein-protein interactions. *Methods*, 24(3):230–239, 2002.

[71] J.S. Fletcher. Molecular SIMS imaging; spatial resolution and molecular sensitivity: Have we reached the end of the road? Is there light at the end of the tunnel? *Surface and Interface Analysis*, 2010. DOI: 10.1002/sia.3488.

[72] L. Fornai, I. Klinkert, A. Angelini, F. Giskes, L.A. Klerk, M. Fedrigo, G. Thiene, and R.M.A. Heeren. 3d imaging mass spectrometry of the heart. *58th ASMS Conf. on Mass Spectrometry and Allied Topics*, 2010.

[73] I. Fournier, M. Wisztorski, and M. Salzet. Tissue imaging using MALDI-MS: a new frontier of histopathology proteomics. *Expert Reviews in Proteomics*, 5(3):413–424, 2008.

[74] J. Franck, K. Arafah, M. Elayed, D. Bonnel, D. Vergara, A. Jacquet, D. Vinatier, M. Wisztorski, R. Day, I. Fournier, and M. Salzet. MALDI imaging mass spectrometry - state of the art technology in clinical proteomics. *Molecular and Cellular Proteomics*, 8:2023–2033, 2009.

[75] A.M. Frank, M.M. Savitski, M.N. Nielsen, R.A. Zubarev, and P.A. Pevzner. De novo peptide sequencing and identification with precision mass spectrometry. *Journal of Proteome Research*, 6(1):114–123, 2007.

[76] T.J. Fuchs and J.M. Buhmann. Inter-active learning of randomized tree ensembles for object detection. *3rd IEEE ICCV Workshop on On-line Computer Vision*, 2009.

[77] D. Gale and L.S. Shapley. College admissions and the stability of marriage. *American Mathematical Monthly*, 69:9–14, 1962.

[78] D. Gao, Y.-X. Zhang, and Y.-H. Zhao. Random forest algorithm for classification of multiwavelength data. *Research in Astronomy and Astrophysics*, 9:220–226, 2009.

[79] E. Gaussier and C. Goutte. Relation between PLSA and NMF and implications. *Proc. of the 28th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 601–602, 2005.

[80] H. Gävert, J. Hurri, J. Särelä, and A. Hyvärinen. FastICA toolbox. http://www.cis.hut.fi/projects/ica/fastica/, 2007.

[81] G. Ge and G.W. Wong. Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. *BMC Bioinformatics*, 9:275–287, 2008.

[82] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[83] M. Gerhard, S.-O. Deininger, and F.-M. Schleif. Statistical classification and visualization of MALDI-imaging data. *Symp. on Computer-Based Medical Systems*, 20-22:403–405, 2007.

[84] P. Geurts, M. Fillet, D. de Seny, M.-A. Meuwis, M. Malaise, M.-P. Merville, and L. Wehenkel. Proteomic mass spectra classification using decision tree based ensemble methods. *Bioinformatics*, 21(14):3138–3145, 2005.

[85] P. Ghosh and W.A. Brand. Stable isotope ratio mass spectrometry in global climate change research. *International Journal of Mass Spectrometry*, 228:1–33, 2003.

[86] P.O. Gislason, J.A. Benediktsson, and J.R. Sveinsson. Random forests for land cover classification. *Pattern Recognition Letters*, 27:294–300, 2006.

[87] K. Glunde, E. Ackerstaff, K. Natarajan, D. Artemov, and Z.M. Bhujwalla. Real-time changes in 1H and 31P NMR spectra of malignant human mammary epithelial cells during treatment with the anti-inflammatory agent indomethacin. *Magnetic Resonance in Medicine*, 48:819–825, 2002.

[88] K. Glunde, C. Jie, and Z.M. Bhujwalla. Mechanisms of indomethacin-induced alterations in the choline phospholipid metabolism of breast cancer cells. *Neoplasia*, 8(9):758–771, 2006.

[89] F. Godtliebsen, J.S. Marron, and P. Chaudhuri. Significance in scale space for bivariate density estimation. *Journal of Computational and Graphical Statistics*, 11:1–21, 2002.

[90] F.M. Green, I.S. Gilmore, J.L.S. Lee, S.J. Spencer, and M.P. Seah. Static SIMS-VAMAS interlaboratory study for intensity repeatability, mass scale accuracy and relative quantitation. *Surface and Interface Analysis*, 42(3):129–138, 2010.

[91] M. Grimaud. New measure of contrast: the dynamics. *Image Algebra and Morphological Image Processing III*, pages 292–305, 1992.

[92] M.R. Groseclose, P.P. Massion, P. Chaurand, and R.M. Caprioli. High-throughput proteomic analysis of formalin-fixed paraffin-embedded tissue microarrays using MALDI imaging mass spectrometry. *Proteomics*, 8:3715–3724, 2008.

[93] J.M. Hammersley. Monte carlo methods for solving multivariable problems. *The Annals of the New York Academy of Science*, 86:844–874, 1960.

[94] D.J. Hand. Breast cancer diagnosis from proteomic mass spectrometry data: A comparative evaluation. *Statistical Applications in Genetics and Molecular Biology*, 7(2), 2008. article 15.

[95] R.A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, 1970.

[96] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[97] A.M. Hawkridge and D.C. Muddiman. Mass spectrometry-based biomarker discovery: Toward a global proteome index of individuality. *Annual Review of Analytical Chemistry*, 2:265–77, 2009.

[98] R.M.A. Heeren, B. Kükrer-Kaletas, I.M. Taban, L. MacAleese, and L.A. McDonnell. Quality of surface: The influence of sample preparation on MS-based biomolecular tissue imaging with MALDI-MS and (ME-)SIMS. *Applied Surface Science*, 255(4):1289–1297, 2008.

[99] F. Hillenkamp, M. Karas, R.C. Beavis, and B.T. Chait. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Analytical Chemistry*, 63(24):1193A–1203A, 1991.

[100] T. Hofmann. Probabilistic latent semantic analysis. *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence*, 1999.

[101] T. Hofmann. Probabilistic latent semantic indexing. *Proc. of the 22nd Annual Intern. SIGIR Conf. on Research and Development in Information Retrieval*, 1999.

[102] D.M. Horn, R.A. Zubarev, and F.W. McLafferty. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry*, 4:320–32, 2000.

[103] P. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[104] Q. Hu, R.J. Noll, H. Li, A. Makarov, M. Hardman, and R. Graham Cooks. The orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry*, 40(4):430–443, 2005.

[105] A.B. Huguet, M.C. de Andrade, R.L. Carceroni, and A.A. Araujo. Color-based watershed segmentation of low-altitude aerial images. *17th Brazilian Symp. on Computer Graphics and Image Processing*, 17-20:138–145, 2004.

[106] T.W. Hutchens and T.T. Yip. New desorption strategies for the mass spectrometric analysis of macromolecules. *Rapid Communications in Mass Spectrometry*, 7:576–580, 1993.

[107] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[108] H. Ishwaran, E.H. Blackstone, C.A. Hansen, and T.W. Rice. A novel approach to cancer staging: application to esophageal cancer. *Biostatistics*, 10(4):603–620, 2009.

[109] G. Ivosev, L. Burton, and R. Bonner. Dimensionality reduction and visualization in principal component analysis. *Analytical Chemistry*, 80(13):4933–4944, 2008.

[110] G. Izmirlian. Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. *Annals of the New York Academy of Sciences*, 1020:154–174, 2005.

[111] B. Jähne. *Digital image processing (3rd ed.): concepts, algorithms, and scientific applications*. Springer, 1995.

[112] A.J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[113] A.B. Kanu, P. Dwivedi, M. Tam, L. Matz, and H.H. Hill Jr. Ion mobility-mass spectrometry. *Journal of Mass Spectrometry*, 43(1):1–22, 2008.

[114] M. Karas and F. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, 60(20):2299–2301, 1988.

[115] P.S. Karvelis, D.I. Fotiadis, A. Tzallas, and I. Georgiou. Region based segmentation and classification of multispectral chromosome images. *IEEE Trans. on Medical Imaging*, 27:697–708, 2008.

[116] B.X. Kausler, M. Kirchner, A. Kreshuk, B.Y. Renard, H. Hahne, B. Küster, J.A.J. Steen, H. Steen, and F.A. Hamprecht. Resolution as a function of m/z for TOF, FT-ICR, and Orbitrap: predicted and confirmed. *58th ASMS Conf. on Mass Spectrometry and Allied Topics*, 2010.

[117] B.F. Keele, J.H. Jones, K.A. Terio, J.D. Estes, R.S. Rudicell, M.L. Wilson, Y. Li, G.H. Learn, T.M. Beasley, J. Schumacher-Stankey, E. Wroblewski, A. Mosser, J. Raphael, S. Kamenya, E.V. Lonsdorf, D.A. Travis, T. Mlengeya, M.J. Kinsel, J.G. Else, G. Silvestri, J. Goodall, P.M. Sharp, G.M. Shaw, and A.E. Pusey. Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz. *Nature*, 460:515–519, 2009.

[118] H.A. Kiers. Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, 56:97–212, 1991.

[119] M. Kirchner, B.Y. Renard, U. Köthe, D.J. Pappin, F.A. Hamprecht, H. Steen, and J.A.J. Steen. Computational protein profile similarity screening for quantitative mass spectrometry experiments. *Bioinformatics*, 26(1):77–83, 2010.

[120] L.A. Klerk, A.F.M. Altelaar, M. Froesch, L.A. McDonnell, and R.M.A. Heeren. Fast and automated large-area imaging MALDI mass spectrometry in microprobe and microscope mode. *International Journal of Mass Spectrometry*, 285(1-2):19–25, 2009.

[121] L.A. Klerk, A.T. Jackson, I.W. Fletcher, and R.M.A. Heeren. Imaging techniques for the analysis of polymers and polymer additives. *Sanibel Conf. on Imaging Mass Spectrometry, Sanibel Island, Florida, USA*, 2007.

[122] I. Klinkert, L.A. McDonnell, S.L. Luxembourg, Altelaar A.F.M., E.R. Amstalden, S.R. Piersma, and R.M.A. Heeren. Tools and strategies for visualization of large image data sets in high-resolution imaging mass spectrometry. *Review of Scientific Instruments*, 78(5):053716, 2007.

[123] E.D. Kolaczyk, J. Ju, and S. Gopal. Multiscale, multigranular statistical image segmentation. *Journal of the American Statistical Association*, 100:1358–1369, 2005.

[124] N. Komodakis and G. Tziritas. Image completion using efficient belief propagation via priority scheduling and dynamic pruning. *IEEE Trans. on Image Processing*, 16(11):2649–2661, 2007.

[125] A. Kreshuk, M. Stankiewicz, X. Lou, M. Kirchner, F.A. Hamprecht, and M.P. Mayer. Automated detection and analysis of bimodal isotope peak distributions in H/D exchange mass spectrometry using HeXicon. *International Journal of Mass Spectrometry*, 2010. DOI: 10.1016/j.ijms.2010.08.025.

[126] C.S. Lane. Mass spectrometry-based proteomics in the life sciences. *Cellular and Molecular Life Sciences*, 62(7–8):848–869, 2005.

[127] R. Lemaire, S.A. Menguellet, J. Stauber, V. Marchaudon, J.-P. Lucot, P. Collinet, M.-O. Farine, D. Vinatier, R. Day, P. Ducoroy, M. Salzet, and I. Fournier. Specific MALDI imaging and profiling for biomarker hunting and validation: Fragment of the 11S proteasome activator complex, reg alpha fragment, is a new potential ovary cancer biomarker. *Journal of Proteome Research*, 6:4127–4234, 2007.

[128] K. Lerch. Discontinuity preserving filtering of spectral images. Master's thesis, University of Heidelberg, Germany, 2006.

[129] O. Lezoray. Supervised automatic histogram clustering and watershed segmentation. *Image Analysis in Stereology*, 22:113–120, 2003.

[130] P. Li and X. Xiao. Evaluation of multiscale morphological segmentation of multispectral imagery for land cover classification. *Proc. of the IEEE Geoscience and Remote Sensing Symp.*, 4:2676–2679, 2004.

[131] H.J. Liebl. Ion microprobe mass analyzer. *Journal of Applied Physics*, 38:5277–5280, 1967.

[132] Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Society*, 101:578–590, 2006.

[133] X. Lou, M. Kirchner, B.Y. Renard, U. Köthe, C. Graf, C. Lee, J.A.J. Steen, H. Steen, M.P. Mayer, and F.A. Hamprecht. Deuteration distribution estimation with improved sequence coverage for HDX/MS experiments. *Bioinformatics*, 26(12):1535–1541, 2010.

[134] S.L. Luxembourg, T.H. Mize, L.A. McDonnell, and R.M.A. Heeren. High-spatial resolution mass spectrometric imaging of peptide and protein distributions on surface. *Analytical Chemistry*, 76:5339–5344, 2004.

[135] L. MacAleese, J. Stauber, and R.M.A. Heeren. Perspectives for imaging mass spectrometry in the proteomics landscape. *Proteomics*, 9:819–834, 2009.

[136] A. Makarov. Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry*, 72(6):1156–1162, 2000.

[137] N. Malpica, J.E. Ortuno, and A. Santos. A multichannel watershed-based algorithm for supervised texture segmentation. *Pattern Recognition Letters*, 24:1545–1554, 2003.

[138] N.E. Manicke, A.L. Dill, D.R. Ifa, and R.G. Cooks. High-resolution tissue imaging on an orbitrap mass spectrometer by desorption electrospray ionization mass spectrometry. *Journal Of Mass Spectrometry*, 45(2):223–6, 2010.

[139] D. Mantini, F. Petrucci, P. del Boccio, D. Pieragostino, M. di Nicola, A. Lugaresi, G. Federici, P. Sacchetta, C. di Ilio, and A. Urbani. Independent component analysis for the extraction of reliable protein signal profiles from MALDI-TOF mass spectra. *Bioinformatics*, 24(1):63–70, 2008.

[140] D. Mantini, F. Petrucci, D. Pieragostino, P. del Boccio, M. di Nicola, C. di Ilio, G. Federici, P. Sacchetta, S. Comani, and A. Urbani. LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise. *BMC Bioinformatics*, 8:101, 2007.

[141] A.G. Marshall, C. Hendrickson, and G.S. Jackson. Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrometry Reviews*, 17:1–35, 1998.

[142] L.A. McDonnell and R.M.A. Heeren. Imaging mass spectrometry. *Mass Spectrometry Reviews*, 26:606–643, 2006.

[143] N. Meinshausen. Hierarchical testing for variable importance. *Biometrika*, 95(2):265–278, 2008.

[144] G. Menschaert, T.T.M. Vandekerckhove, G. Baggerman, B. Landuyt, J.V. Sweedler, L. Schoofs, W. Luyten, and W. Van Criekinge. A hybrid, de novo based, genome-wide database search approach applied to the sea urchin neuropeptidome. *Journal of Proteome Research*, 9(2):990–996, 2010.

[145] B.H. Menze, B.M. Kelm, M.A. Weber, P. Bachert, and F.A. Hamprecht. Mimicking the human expert: pattern recognition for an automated assessment of data quality in MR spectroscopic images. *Magnetic Resonance in Medicine*, 59(6):1457–66, 2008.

[146] W. Meuleman, J.Y.M.N. Engwegen, M.-C.W. Gast, J.H. Beijnen, M.J.T. Reinders, and L.F.A. Wessels. Comparison of normalisation methods for surface-enhanced laser desorption and ionisation (SELDI) time-of-flight (TOF) mass spectrometry data. *BMC Bioinformatics*, 9:88, 2009.

[147] F. Meyer. Topographic distance and watersheds lines. *Signal Processing*, 38:113–125, 1994.

[148] H.E. Meyer and K. Stühler. High-performance proteomics as a tool in biomarker discovery. *Proteomics*, 7 Suppl 1:18–26, 2007.

[149] A.M.A. Mingels, J.L.J. van Dongena, and M. Merkx. Mapping preferred sites for fluorescent labeling by combining fluorescence and MS analysis of tryptic CNA35 protein digests. *Journal of Chromatography B*, 863(2):293–297, 2008.

[150] T.P. Minka. Fastfit toolbox for MATLAB, version 1.2. http://research.microsoft.com/en-us/um/people/minka/software/fastfit/, 2004. accessed March 2009.

[151] J. Moffat, D.A. Grueneberg, X. Yang, S.Y. Kim, A.M. Kloepfer, G. Hinkle, B. Piqani, T.M. Eisenhaure, B. Luo, J.K. Grenier, A.E. Carpenter, S.Y. Foo, S.A. Stewart, B.R. Stockwell, N. Hacohen, W.C. Hahn, E.S. Lander, D.M. Sabatini, and D.E. Root. A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell*, 124(6):1283–1298, 2006.

[152] J.S. Morris, K.R. Coombes, J. Koomen, K.A. Baggerly, and R. Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9):1764–1775, 2005.

[153] E.A. Muchmore. Chimpanzee models for human disease and immunobiology. *Immunological Reviews*, 183(1):83–93, 2001.

[154] K. Murphy. Bayes net toolbox for MATLAB. http://www.cs.ubc.ca/ murphyk/Software/BNT/bnt.html, 1997-2002. accessed October 2008.

[155] D. Nanavati, M. Gucek, J.L.S. Milne, S. Subramaniam, and S.P. Markey. Stoichiometry and absolute quantification of proteins with mass spectrometry using fluorescent and isotope-labeled concatenated peptide standards. *Molecular and Cellular Proteomics*, 7:442–447, 2008.

[156] H.T. Nguyen, M. Worring, and R. van den Boomgaard. Watersnakes: Energy-driven watershed segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(3):330–342, 2003.

[157] G. Noyel, J. Angulo, and D. Jeulin. Morphological segmentation of hyperspectral images. *Image Analysis in Stereology*, 26:101–109, 2007.

[158] G. Noyel, J. Angulo, and D. Jeulin. Random germs and stochastic watershed for unsupervised multispectral image segmentation. *KES, Lecture Notes in Computer Science (LNCS)*, 4694/2008:17–24, 2008.

[159] G. Noyel, J. Angulo, D. Jeulin, D. Balvay, and C.-A. Cuenod. Filtering, segmentation and region classification by hyperspectral mathematical morphology of DCE-MRI series for angiogenesis imaging. *Int. Symp. on Biomedical Imaging: From Nano to Macro*, pages 1517–1520, 2008.

[160] M. Pal. Random forest classifier for remote sensing classification. *International Journal for Remote Sensing*, 26(1):217–222, 2005.

[161] M. Palmblad, A. Tiss, and R. Cramer. Mass spectrometry in clinical proteomics - from the present to the future. *Proteomics - Clinical Applications*, 3:6–17, 2009.

[162] M. Pardo and G. Sberveglieri. Random forests and nearest shrunken centroids for the classification of sensor array data. *Sensor and Actuators*, 131:93–99, 2008.

[163] D.M. Parkin, F. Bray, J. Ferlay, and P. Pisani. Global cancer statistics, 2002. *CA: A Cancer Journal for Clinicians*, 55:74–108, 2005.

[164] W. Paul. Electromagnetic traps for charged and neutral particles. *Proc. of the Int. School of Physics Enrico Fermi Course CXVIII on Laser Manipulation of Atoms and Ions*, pages 497–517, 1992.

[165] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[166] S. Petrie and D.K. Bohme. Ions in space. *Mass Spectrometry Reviews*, 26(2):258–280, 2006.

[167] T.M. Preuss, M. Caceres, M.C. Oldham, and D.H. Geschwind. Human brain evolution: Insights from microarrays. *Nature Reviews Genetics*, 5:850–860, 2004.

[168] S. Rajan, J. Ghosh, and M.M. Crawford. An active learning approach to hyperspectral data classification. *IEEE Trans. on Geoscience and Remote Sensing*, 46(4):1231–1242, 2008.

[169] F.A. Rajgara, D. Mathur, T. Nishide, T. Kitamura, H. Shiromaru, Y. Achiba, and N. Kobayashi. Multi-hit, position-sensitive, time-of-flight spectroscopy using a modified backgammon-weighted capacitor anode. *International Journal of Mass Spectrometry*, 215:151–162, 2001.

[170] S. Rauser, H. Höfler, and A. Walch. In-situ-Proteomanalyse von Geweben mittels bildgebender Massenspektrometrie (MALDI imaging). *Der Pathologe*, 30(2):140–145, 2009.

[171] S. Rauser, C. Marquardt, B. Balluff, S.-O. Deininger, C. Albers, E. Belau, R. Hartmer, D. Suckau, K. Specht, M.P. Ebert, M. Schmitt, M. Aubele, H. Höfler, and A. Walch. Classification of HER2 receptor status in breast cancer tissues by MALDI imaging mass spectrometry. *Journal of Proteome Research*, 9(4):1854–1863, 2010.

[172] B.Y. Renard, M. Kirchner, F. Monigatti, A.R. Ivanov, J. Rappsilber, D. Winter, J.A.J. Steen, F.A. Hamprecht, and H. Steen. When less can yield more - computational preprocessing of MS/MS spectra for peptide identification. *Proteomics*, 9(21):4978–84, 2009.

[173] B.Y. Renard, M. Kirchner, H. Steen, J.A.J. Steen, and F.A. Hamprecht. NITPICK: peak identifcation for mass spectrometry data. *BMC Bioinformatics*, 9:355, 2008.

[174] Thomson Reuters. Web of science and ISI web of knowledge. http://scientific.thomson.com/products/wos/ and http://apps.isiknowledge.com/.

[175] G. Riccardi and D. Hakkani-Tür. Active learning: Theory and applications to automatic speech recognition. *IEEE Trans. on Speech and Audio Processing*, 13(4):1–8, 2006.

[176] J. Röder, B. Nadler, M. Hanselmann, and F.A. Hamprecht. Confidence random forest. Unpublished work, 2010.

[177] J. Röder, B. Nadler, U. Köthe, and F.A. Hamprecht. An improved active learning strategy for label query in linear time. Unpublished work, 2009.

[178] J. Röder, B. Nadler, U. Köthe, M. Hanselmann, R.M.A. Heeren, and F.A. Hamprecht. Fast outlier detection and uncertainty estimation for random forest. Unpublished work, 2009.

[179] J.B.T.M. Roerdink and A. Meijster. The watershed transform: Definitions, algorithms and parallelization strategies. *Fundamenta Informaticae*, 41:187–228, 2001.

[180] T.C. Rohner, D. Staab, and M. Stoeckli. MALDI mass spectrometric imaging of biological tissue sections. *Mechanisms of Ageing and Development*, 126(1):177–185, 2005.

[181] A. Rosenfeld and J. Pfaltz. Sequential operations in digital picture processing. *Robust Computer Vision*, 13(4):471–494, 1966.

[182] A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof. On-line random forests. *3rd IEEE ICCV Workshop on On-line Computer Vision*, 2009.

[183] M.E. Sanders, E.C. Dias, B.J. Xu, J.A. Mobley, D. Billheimer, H. Roder, J. Grigorieva, M. Dowsett, C.L. Arteaga, and R.M. Caprioli. Differentiating proteomic biomarkers in breast cancer by laser capture microdissection and MALDI MS. *Journal of Proteome Research*, 7(4):1500–1507, 2008.

[184] A. Savitzky and M.J.E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36:1627–1639, 1964.

[185] P. Scheunders. Multivalued image segmentation based on first fundamental form. *Proc. of the Int. Conf. on Image Analysis and Processing*, pages 185–190, 2001.

[186] F.-M. Schleif, B. Hammer, and T. Villmann. Margin based active learning for LVQ networks. *Proc. of the European Symp. on Artificial Neural Networks*, pages 539–544, 2006.

[187] F.-M. Schleif, B. Hammer, and T. Villmann. Margin based active learning for LVQ networks. *Neurocomputing*, 70:1215–1224, 2007.

[188] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[189] O. Schulz-Trieglaff, E. Machtejevas, K. Reinert, H. Schlüter, J. Thiemann, and K. Unger. Statistical quality assessment and outlier detection for liquid chromatography-mass spectrometry experiments. *BioData Mining*, 2(1):4, 2009.

[190] K. Schwamborn, R.C. Krieg, M. Reska, G. Jakse, R. Knuechel, and A. Wellmann. Identifying prostate carcinoma by MALDI-imaging. *International Journal of Molecular Medicine*, 20:155–159, 2007.

[191] S.A. Schwartz, R.J. Weil, R.C. Thompson, Y. Shyr, J.H. Moore, S.A. Toms, M.D. Johnson, and R.M. Caprioli. Proteomic-based prognosis of brain tumor patients using direct-tissue matrix-assisted laser desorption ionization mass spectrometry. *Cancer Research*, 65(17):7674, 2005.

[192] E. Seeley and R.M. Caprioli. Imaging mass spectrometry: Towards clinical diagnostics. *Proteomics - Clinical Applications*, 2:1435–1443, 2008.

[193] E.H. Seeley and R.M. Caprioli. Molecular imaging of proteins in tissues by mass spectrometry. *Proceedings of the National Academy of Sciences (PNAS)*, 105(47):18126–18131, 2008.

[194] The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437:69–87, 2005.

[195] J. Serra. *Image Analysis and Mathematical Morphology*. New York Academic Press, 1982.

[196] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[197] F. Simpkins, J.A. Czechowicz, L. Liotta, and E.C. Kohn. SELDI-TOF mass spectrometry for cancer biomarker discovery and serum proteomic diagnostics. *Pharmacogenomics*, 6(6):647–653, 2005.

[198] A. Sinz. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. *Mass Spectrometry Reviews*, 25(4):663–682, 2006.

[199] K. Sköld, M. Svensson, A. Nilsson, X. Zhang, K. Nydahl, R.M. Caprioli, P. Svenningsson, and P.E. Andren. Decreased striatal levels of PEP-19 following MPTP lesion in the mouse. *Journal of Proteome Research*, 5(2):262–269, 2006.

[200] A. Slany, V.J. Haudek, N.C. Gundacker, J. Griss, T. Mohr, H. Wimmer, M. Eisenbauer, L. Elbling, and C. Gerner. Introducing a new parameter for quality control of proteome profiles: Consideration of commonly expressed proteins. *Electrophoresis*, 30(8):1306–28, 2009.

[201] V.S. Smentkowski, S.G. Ostrowski, F. Kollmer, A. Schnieders, M.R. Keenan, J.A. Ohlhausen, and P.G. Kotula. Multivariate statistical analysis of non-mass-selected TOF-SIMS data. *Surface and Interface Analysis*, 40(8):1176–1182, 2008.

[202] P. Soille. *Morphological Image Analysis*. Springer, 1999.

[203] J. Stauber, L. MacAleese, J. Franck, E. Claude, M. Snel, B. Kükrer-Kaletas, I.M. van der Wiel, M. Wisztorski, I. Fournier, and R.M.A. Heeren. On-tissue protein identification and imaging by MALDI-ion mobility mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 21(3):338–47, 2009.

[204] W.E. Stephens. A pulsed mass spectrometer with time dispersion. *Physical Review*, 69:691, 1946.

[205] R.A. Stine. Model selection using information theory and the MDL principle. *Sociological Methods and Research*, 33:230–260, 2004.

[206] M. Stoeckli, P. Chaurand, D.E. Hallahan, and R.M. Caprioli. Imaging mass spectrometry: A new technology for the analysis of protein expression in mammalian tissues. *Nature Medicine*, 7(4):493–496, 2001.

[207] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307, 2008.

[208] E.B. Sudderth, A.T. Ihler, W.T. Freeman, and A.S. Willsky. Nonparametric belief propagation. *Proc. of the 2003 IEEE Conf. on Computer Vision and Pattern Recognition*, 1:605–612, 2003.

[209] F. Taguchi, B. Solomon, V. Gregorc, H. Roder, R. Gray, K. Kasahara, M. Nishio, J. Brahmer, A. Spreafico, V. Ludovini, P.P. Massion, R. Dziadziuszko, J. Schiller, J. Grigorieva, M. Tsypin, S.W. Hunsucker, R.M. Caprioli, M.W. Duncan, F.R. Hirsch, P.A. Bunn, and D.P. Carbone. Mass spectrometry to classify non-small-cell lung cancer patients for clinical outcome after treatment with epidermal growth factor receptor tyrosine kinase inhibitors: A multicohort cross-institutional study. *Journal of the National Cancer Institute*, 99(11):838–846, 2007.

[210] C.F. Taylor, N.W. Paton, K.S. Lilley, P.A. Binz, R.K. Jr Julian, A.R. Jones, W. Zhu, Apweiler R., R. Aebersold, E.W. Deutsch, M.J. Dunn, A.J. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T.A. Neubert, S.D. Patterson, P. Ping, S.L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T.M. Vondriska, J.P. Whitelegge, M.R. Wilkins, I. Xenarios, J.R. Yates, and H. Hermjakob. The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology*, 25(8):887–893, 2007.

[211] M. Thevis and W. Schänzer. Mass spectrometry in sports drug testing: Structure characterization and analytical assays. *Mass Spectrometry Reviews*, 26:79–107, 2007.

[212] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, and Q.-T. Le. Sample classification from protein mass spectrometry, by peak probability contrasts. *Bioinformatics*, 20(17):3034–3044, 2004.

[213] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. *Proc. of the IEEE Conf. on Computer Vision*, pages 836–846, 1998.

[214] P.J. Trim, S.J. Atkinson, A.P. Princivalle, P.S. Marshall, A. West, and M.R. Clench. Matrix-assisted laser desorption/ionisation mass spectrometry imaging of lipids in rat brain tissue with integrated unsupervised and supervised multivariant statistical analysis. *Rapid Communications in Mass Spectrometry*, 22:1503–1509, 2008.

[215] D. Tuia, F. Ratle, F. Pacifici, M.F. Kanevski, and W.J. Emery. Active learning methods for remote sensing image classification. *IEEE Trans. on Geoscience and Remote Sensing*, 47(7):2218–2232, 2009.

[216] M. Uhlen, E. Björling, C. Agaton, C.A. Szigyarto, B. Amini, E. Andersen, A.C. Andersson, P. Angelidou, A. Asplund, C. Asplund, L. Berglund, K. Bergström, H. Brumer, D. Cerjan, M. Ekström, A. Elobeid, C. Eriksson, L. Fagerberg, R. Falk, J. Fall, M. Forsberg, M.G. Björklund, K. Gumbel, A. Halimi, I. Hallin, C. Hamsten, M. Hansson, M. Hedhammar, G. Hercules, C. Kampf, K. Larsson, M. Lindskog, W. Lodewyckx, J. Lund, J. Lundeberg, K. Magnusson, E. Malm, P. Nilsson, J. Odling, P. Oksvold, I. Olsson, E. Oster, J. Ottosson, L. Paavilainen, A. Persson, R. Rimini, J. Rockberg, M. Runeson, A. Sivertsson, A. Sköllermo, J. Steen, M. Stenvall, F. Sterky, S. Strömberg, M. Sundberg, H. Tegel, S. Tourle, E. Wahlund, A. Walden, J. Wan, H. Wernerus, J. Westberg, K. Wester, U. Wrethagen, L.L. Xu, S. Hober, and F. Pontén. A human protein atlas for normal and cancer tissues. *Molecular and Cellular Proteomics*, 4(12):1920–32, 2005.

[217] P.J. Ulintz, J. Zhu, Z.S. Qin, and P.C. Andrews. Improved classification of mass spectrometry database search results using newer machine learning approaches. *Molecular and Cellular Proteomics*, 5:497–509, 2006.

[218] S.E. van Bramer. An introduction to mass spectrometry. Technical report, Widener University, Chester PA, 1998.

[219] R. van de Plas, F. Ojeda, M. Dewil, L. van den Bosch, B. de Moor, and E. Waelkens. Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis. *Proc. of the Pacific Symp. of Biocomputing*, 12:458–469, 2007.

[220] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, J.D. Gocayne, P. Amanatides, R.M. Ballew, D.H. Huson, J.R. Wortman, Q. Zhang, C.D. Kodira, X.H. Zheng, L. Chen, M. Skupski, G. Subramanian, P.D. Thomas, J. Zhang, G.L. Gabor Miklos, C. Nelson, S. Broder, A.G. Clark, J. Nadeau, V.A. McKusick, N. Zinder, A.J. Levine,

R.J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A.E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T.J. Heiman, M.E. Higgins, R.R. Ji, Z. Ke, K.A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G.V. Merkulov, N. Milshina, H.M. Moore, A.K. Naik, V.A. Narayan, B. Neelam, D. Nusskern, D.B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N.N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M.J. Campbell, K.V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[221] S. Vicente, V. Kolmogorov, and C. Rother. Graph cut based image segmentation with connectivity priors. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[222] L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm

based on immersion simulations. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 13:583–598, 1991.

[223] A. Walch, S. Rauser, S.-O. Deininger, and H. Höfler. MALDI imaging mass spectrometry for direct tissue analysis: a new frontier for molecular histology. *Histochemistry and Cell Biology*, 130(3):421–434, 2008.

[224] C. Weickhardt, F. Moritz, and J. Grotemeyer. Time-of-flight mass spectrometry: State-of-the-art in chemical analysis and molecular science. *Mass Spectrometry Reviews*, 15:139–162, 1996.

[225] M. Welk, C. Feddern, B. Burgeth, and J. Weickert. Median filtering of tensor-valued images. *Proc. of DAGM Symp. 2003, Lecture Notes in Computer Science (LNCS)*, pages 17–24, 2003.

[226] S. Wiese, K.A. Reidegeld, H.E. Meyer, and B. Warscheid. Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics*, 7(3):340–350, 2006.

[227] B. Williams, S. Cornett, B. Dawant, A. Crecelius, B. Bodenheimer, and R.M. Caprioli. An algorithm for baseline correction of MALDI mass spectra. *Proc. of the 43rd Annual Southeast Regional Conf. on Algorithms and Theory*, 1:137–142, 2005.

[228] G. Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction*. Springer, 2003.

[229] J.M. Wiseman, D.R. Ifa, Q. Song, and R.G. Cooks. Tissue imaging at atmospheric pressure using desorption electrospray ionization (DESI) mass spectrometry. *Angewandte Chemie (International Edition)*, 45(43):7188–7193, 2006.

[230] A.C. Wolff, M.E. Hammond, J.N. Schwartz, K.L. Hagerty, D.C. Allred, R.J. Cote, M. Dowsett, P.L. Fitzgibbons, W.M. Hanna, A. Langer, L.M. McShane, S. Paik, M.D. Pegram, E.A. Perez, M.F. Press, A. Rhodes, C. Sturgeon, S.E. Taube, R. Tubbs, G.H. Vance, M. van de Vijver, T.M. Wheeler, and D.F. Hayes. American society of clinical oncology/college of american pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Journal of Clinical Oncology*, 25(1):118–45, 2007.

[231] J.K. Wu and R.W. Odom. Matrix-enhanced secondary ion mass spectrometry: a method for molecular analysis of solid surfaces. *Analytical Chemistry*, 68(5):873–882, 1996.

[232] L. Wu, X. Lu, K.S. Kulp, M.G. Knize, E.S.F. Berman, E.J. Nelson, J.S. Felton, and K.J.J. Wu. Imaging and differentiation of mouse embryo tissues by TOF-SIMS. *International Journal of Mass Spectrometry*, 2-3(1):137–145, 2007.

[233] K. Yanagisawa, Y. Shyr, B. Xu, P. Massion, P. Larsen, B. White, J. Roberts, M. Edgerton, A. Gonzalez, S. Nadaf, J.H. Moore, R.M. Caprioli, and D.P. Carbone. Proteomic patterns of tumour subsets in non-small-cell lung cancer. *The Lancet*, 362(9382):433–439, 2003.

[234] I. Yao, Y. Sugiura, M. Matsumoto, and M. Setou. In situ proteomics with imaging mass spectrometry and principal component analysis in the scrapper-knockout mouse brain. *Proteomics*, 8(18):3692–3701, 2008.

[235] D. Zelterman. *Advanced Log-Linear Models Using SAS*. SAS Publishing, 2002.

[236] Y. Zhang, X. Feng, and X. Le. Segmentation on multispectral remote sensing image using watershed transformation. *Proc. of the 08 Congress on Image and Signal Processing*, 4:773–777, 2008.

[237] W. Zhu, J.W. Smith, and C.-M. Huang. Mass spectrometry-based label-free quantitative proteomics. *Journal of Biomedicine and Biotechnology*, 2010:1–6, 2009.

[238] X. Zhu. Semi-supervised learning literature survey. Computer Sciences Technical Report 1530, University of Wisconsin–Madison, 2005.

[239] S. Zomer, M. del Nogal Sánchez, R.G. Brereton, and J.L. Pérez Pavón. Active learning support vector machines for optimal sample selection in classification. *Journal of Chemometrics*, 18:294–305, 2004.

[240] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[241] V. Zuber and K. Strimmer. Gene ranking and biomarker discovery under correlation. *Bioinformatics*, 20:2700–7, 2009.