DOCTORAL THESIS

Large-scale multiple testing under arbitrary covariance dependency and topological data analysis for mass spectrometry imaging applications

---

by

Vladimir Vutov

A thesis submitted in partial fulfillment of the requirements for the degree of Dr. rer. nat.

Institute for Statistics, University of Bremen
2023

Supervisor and First Reviewer:	Prof. Dr. Thorsten Dickhaus
Second Reviewer:	Prof. Dr. Werner Brannath
Submission Date:	August 2023
Date of Defense:	26.01.2024

"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less." – Marie Curie

Dedication

"Life is beautiful. It's about giving. It's about family." - Walt Disney

First and foremost, this work is dedicated to the memory of my beloved late mother, Nadia S. Vutova. Her words have had a profound impact on my life and have served as a guiding light for me: "In life, only very few good things come easy. For everything else, you have to put in dedicated effort and hard work." I am grateful to my father, Krasimir V. Vutov, for educating me on the values of determination and commitment.

To my beloved aunt, Diana, I wish to convey my deep gratitude for her unwavering support. Thank you, Aunt Didi, for being an incredible source of inspiration and believing in me.

Acknowledgements

First and foremost, I would like to extend my sincere gratitude to Prof. Dr. Thorsten Dickhaus, my primary supervisor, for giving me a chance to conduct this Ph.D. project, offering generous help throughout it, and organizing additional opportunities for all members of the (statistical) group to share ideas and develop, such as the Reading Club and the ZeSob colloquiums.

I am immensely thankful for the financial support provided by the German Research Foundation within the framework of Research Training Group (RTG) 2224, entitled: "π^3 Parameter Identification–Analysis, Algorithms, Applications". Also, I am thankful to the University of Bremen for providing me with the necessary equipment during my stay and research.

I am grateful to have been part of the 2nd cohort of the RTG. The program has been instrumental in supporting me throughout my research and contributing to my growth as a researcher. Also, I want to thank Prof. Dr. Werner Brannath for offering valuable feedback and insightful ideas following each of my presentations. A big thanks to our previous and recent RTG advisors, Dr. Daniel Baguer and Dr. Pascal Fernsel.

Here, I would like to take the opportunity to thank my non-statistical co-authors, Gideon Klaila and Prof. Dr. Anastasios Stefanou, for their collaboration and time.

I would also like to thank Dörte Mindermann, Judith Barthel, and Martina Titze for their administrative assistance. Furthermore, I really appreciated the technical assistance of Victor Wilden, Ingo Jauer, and Jannis Wilken.

I want to extend a warm shout-out to all RTG and statistical colleagues and friends, especially [A-Z] Anna, Derick, Daniel Odipo, Eva, Friederike, Jean, Johannes, Justus, Max, Maryam, Louisa, Torben, Pascal, and others.

To Sofiya, I am filled with immense gratitude for the all-around support and motivation you provided me throughout my Ph.D. project. Words can hardly describe my thanks and appreciation to you.

Throughout my higher education journey, I have been incredibly fortunate to have had numerous exceptional mentors. I would like to take this opportunity to express my gratitude to a few of them. I want to sincerely thank Denitsa Grigorova, Ralitza Gueorguieva, Dimitar Kodjabachev, Vladimir Mitankin, and Vessela K. Stoimenova for their profound contributions. A special note of thanks goes to my Master's thesis advisor, Prof. Dr. Neyko M. Neykov, for his invaluable advice and continuous support.

I would like to express my appreciation to all my friends outside the academia for their tremendous support, particularly Delyan, Lubomir, Mincho, Stanislav, Simeon, Ivailo, Andreas, Yvan Pacheco, Hanna, and many others. A big shout-out to Dr. Iliyana Angelova for the (long) wine-based evenings in Bremen.

Contents

Chapter 1

Synopsis

1.1 Purpose of the study

1.1.1 Challenges in biomedical data analysis

With the recent surge of high-throughput biomedical devices, contemporary biomedical datasets have reached a large scale, i.e., containing numerous (measured) variables ([92, 3]). Additionally, those datasets often exhibit noise due to the extreme sensitivity and complexity of the data acquisition processes. As a result, analyzing modern biomedical data becomes increasingly challenging, specifically in high-dimensional and noisy data regimes. This surge poses new challenges for researchers while (also) creating new avenues of opportunity. Frequently, those biomedical datasets contain well-studied and biologically meaningful data variables (e.g., biomarkers; [94]) or biological patterns (such as an isotope pattern; [100, 66]). These variables and patterns can be utilized to verify the effectiveness of any (novel) statistical model. In essence, the biologically meaningful characteristics underlying the data can serve as a means to validate the (suggested) data analysis approach.

This work considers an advanced biological device called Matrix-assisted laser desorption/ionization mass spectrometry imaging (MALDI) MSI, commonly known as MALDI Imaging. MALDI is a label-free instrument that provides molecular information of a given biological sample (e.g., tissue) in a spatial fashion (e.g., [16, 59, 60]). Specifically, the MALDI tool measures mass spectra at various spatial positions and generates an image for each specific spot within tissues. Each spot is called a mass spectrum, and it represents the ”relative abundances” of molecules with numerous mass-to-charge ratio (m/z) values – encompassing a range from several hundred to a few tens of thousands of m/z values (corresponding to molecular masses) (cf. [3]), for more details, see, e.g., in [3, 16, 59, 60].

1.1.2 Aims and objectives

This thesis aims to address two challenging tasks in analyzing MALDI-related data. The first task is feature selection or, viz., discovering features that are associated with the outcome variable in high-dimensional settings. The second task entails tumor classification or, simply put, classifying observational units (mass spectra) into tumor (sub-)types (cf. [9]). These tasks have garnered great attention and have led to the development of numerous methodologies (for example, among many others, [5, 9, 16, 63, 93, 117]). Throughout this work, the (discrete) outcome variables describe cancer (sub-)types while the independent variables are m/z values in the magnitude of thousands ([3]). Subsequently, a brief explanation of the analysis mechanism for each of the aforementioned tasks is provided below.

To begin with the first challenging task, we make a contribution by proposing novel methodologies that evaluate the strength of associations between either binary or multi-class (nominal, i.e., categorical classes are not ordered) outcome variables and a great number of explanatory variables. Chapters 2 - 3 introduce these methodologies by means of multiple testing under arbitrary covariance dependency among test statistics. Generally speaking, these methods approximate the (realized) false discovery proportion (FDP) of a thresholding procedure for the marginal p-values. The false discovery proportion presents the proportion of false discoveries among all rejections. In simple terms, this quantity provides insights to researchers regarding how many features are spuriously claimed as statistically associative. Furthermore, the proposed (inferential) methods can work under scenarios whereby the number of explanatory variables (m/z channels) is either quite large or even (much) larger than the number of observational units (mass spectra).

In Chapter 4, we present our contribution to the second (challenging) task. The entire framework consists of two integral steps: Retrieving the informative parts from spectral data by means of topological data analysis (TDA) and then performing either linear or nonlinear classifiers on the resulting (topological) vectors to classify the observational units into (binary) cancer subtypes. Our contribution is related to the extraction of peak-related information from spectral data (by means of TDA), where the importance of the peak growths with its relative height. In the context of TDA, this is called "topological persistence". Broadly speaking, high persistent peaks are anticipated to be signals, while low persistent peaks are presumably to be noise. Our proposed topological framework offers superior denoising and compression properties within the context of MALDI applications. To further clarify these properties, data compression techniques, such as the one discussed in [5], reduce the volume/size of data while retaining the same number of variables. In contrast, data reduction techniques such as principal component analysis (PCA), independent component analysis (ICA), and non-negative factor approximation (NMF) (ideally) aim to reduce the number of variables (data columns), i.e., representing the original data into a lower-dimensional subspace. Regarding the denoising property, the objective is to capture informative parts of a (mass) spectrum signal while removing noise-associated fluctuations. We investigate the aforementioned property

of our topological framework based on artificially contaminated mass spectrometry images given varying levels (i.e., signal-to-noise ratios) and types of noise. Finally, our supervised framework can be utilized in multi-class tumor classification studies in conjunction with random forest.

To sum up, this thesis deals with the aforementioned challenges in high-dimensional data regimes, specifically focusing on MALDI data. In the context of MALDI, we model each spectrum individually as it is measured from small tissue core regions – this is common practice in modeling MALDI data (e.g., [16, 9, 63, 93]).

1.2 Analysis methodologies

1.2.1 Large-scale multiple testing

Large-scale multiple testing refers to the concept whereby a large number of statistical tests are conducted simultaneously on a (single) dataset. For example, in genomic studies, thousands or tens of thousands of (genetic) markers are analyzed for association with a particular phenotype, e.g., disease. In the context of MALDI, we aim at discovering a group of variables (i.e., molecular masses) that are the most (statistically) distinguishing for cancer associations.

In multiple testing, the common (type I) rates that researchers aim at controlling are family-wise error rate (FWER) and false discovery rate (FDR). By definition, the FWER is the probability of making at least one false rejection in a family of hypothesis-testing problems, i.e., $FWER = \mathbb{P}(V \geq 1)$, where V denotes the (random) number of type I errors. The FDR is defined as the expected proportion of falsely rejected null hypotheses among all the rejections (i.e., rejected null hypotheses). Classical methodologies that control the FWER are inclined to have substantially less statistical power than procedures that control the FDR (cf. [11]). In this connection, the FWER is a conservative criterion, and many interesting discoveries can be disregarded, given a significant number of null hypotheses. As mentioned above (Subsection 1.1.1), modern biomedical datasets contain a large number of (measured) variables, frequently in a magnitude of a couple of thousands or even more ([92, 3]).

To highlight, an extensive comparison of various multiple testing procedures has been reported in [29], where the authors compared different FWER procedures versus FDR procedures ([29]). In Section 4.2 of their study ([29]), the authors investigated the performance of selected procedures by means of simulations in the case of a moderate testing regime involving 500 hypotheses. As a result, they have concluded that the considered FDR procedures "provided substantial increases in power compared to the more conservative FWER procedures" ([29]).

Since (at least) its premiere in the seminal paper by Benjamini and Hochberg ([11]), the FDR control has been widely utilized in high-dimensional studies. However, employing the widely recognized Benjamini-Hochberg procedure ([11]) or Storey's procedure ([105, 106]), initially designed for stochastically independent or weakly dependent p-values, can result in an FDR inflation under certain forms of (strong) dependencies (cf., among others

[33, 32, 38, 37]). In [33], the author highlighted the importance of taking into consideration correlation effects when deciding which null hypotheses are (statistically) significant; let us call this decision process. In particular, the (strong) dependencies among test statistics can make the theoretical null distribution (this refers to the standard normal distribution on $\mathbb{R}$) effectively wider or narrower ([33]). In the former case, the aforementioned procedures are extremely liberal, i.e., rejecting a great number of (null) hypotheses. In the second case, the theoretical null is "too conservative", and these procedures are too conservative. This supports the goal of incorporating the correlation effects in the decision process.

High-throughput (biological) instruments generate data that display strong temporal or spatial dependencies due to the inherent biological or technological mechanisms (cf. [102]). Likewise, the MALDI instrument generates its datasets on multiple spatial spots within a patient's tissue. As a result, mass spectra (data rows) are highly related due to the heterogeneous (biological) structures among these (discrete) spots. Multiple data analysis approaches have been designed to extract the important information hidden in the context of MALDI data by "exploring statistical correlations" ([16], cf. also [25, 51, 55, 63]). Similarly, our objective is explicitly utilizing correlation effects in our inferential procedures. In Chapter 2, the utilized dataset has been generated solely on cancerous (annotated) regions (called regions of interest) ([9]) – based on the two most lethal lung cancer entities (for more details, see Section 2.4) – mass spectra are highly related.

This thesis employs a multi-factor model in order to (model and) integrate dependencies among test statistics into the decision process. This way of modeling aims to describe the underlying dependency structure through latent factors. In particular, we employ a generic framework to approximate the false discovery proportion (FDP) proposed in [38]. In short, the framework relies on the assumption that the test statistics (approximately) follow a multivariate normal distribution with a known covariance matrix of the test statistics. The idea of the approach is to conduct a spectral decomposition of the aforementioned covariance matrix and then to deduct the principal factors that induce the strong correlation across the z-values before evaluating the FDP. Furthermore, the (aforementioned) covariance matrix can have an arbitrary dependence structure. This estimator is called principal factor approximation (PFA), and it relies on the sparsity assumption in the sense that the number of false null hypotheses (i.e., rejected null hypotheses) is very small compared to the total number of hypotheses. FDR and FDP control are (closely) related concepts. The FDR is defined as the expected value of the FDP.

In this thesis, we utilize marginal regressions to individually model (marginal) associations for each covariate. The purpose for considering marginal modeling stems from our objective in analysis scenarios where the number of covariates is extremely large or (possibly) exceeds the number of observations. Hence, it is not possible to (reliably) fit a single model encompassing all covariates. The utilization of this way of modeling has been considered in several studies under different regression models (e.g., see in [38, 8, 7]), where the focus of those studies is on performing a simultaneous statistical inference for each predictor.

The indispensable step of our frameworks is the approximation of the joint null distributions among all marginal

fits and the estimation of the (limiting) covariance matrix across the regression coefficients. We achieve this by employing the framework of multiple marginal models (MMM) (cf. [83]). The conventional scenario of the latter is provided by modeling data with multiple outcome variables ([80, 79, 83]). As a result, those studies employ procedures to control the FWER regarding type I errors. However, in this thesis, we approach it differently, viz., each "endpoint" (comparison) corresponds to an explanatory variable, and the number of independent variables under examination in our studies amounts to thousands. As aforementioned, the procedures that control FWER are conservative and tend to result in lower statistical power vis-à-vis procedures that control the proportion of false discoveries.

1.2.2 Topological data analysis

Topological data analysis (TDA) arose from numerous works in applied (algebraic) topology and computational geometry, making it a relatively novel field ([21]). TDA strives to provide well-founded mathematical, statistical, and algorithmic methods to infer, analyze, and exploit the complex topological and geometric structures underlying data ([30]).

In the context of MALDI, as highlighted in multiple studies (e.g., in [16, 9, 119]), a straightforward approach to extracting meaningful variables is on the basis of the concept of detecting significant signal peaks, commonly referred to as peak detection, in the literature. To this end, significant peaks are assumed to carry valuable information for distinguishing mass spectra originating from different tumor classes (see [9]). Furthermore, concentrating on the meaningful peaks, one can disregard those highly associated with noise, as acknowledged in [3]. To clarify, peak picking or the selection of meaningful peaks is a pre-processing method used in MALDI data analysis. It involves choosing m/z values that correspond to high and relevant peaks (for more details, see [3]). A more recent and novel approach (see in [66]) suggests integrating the isotope pattern ([100]) around the chosen peaks. Loosely speaking, the isotope pattern corresponds to the situation where a biologically meaningful peak (i.e., biomarker) is surrounded by other peaks with a relatively close height.

Inspired by these studies, we aim at utilizing the peak-related information (i.e., the geometry structure) by means of TDA and employing this information for classification. Our topological approach depends upon the claim that the importance of a peak increases with its relative height, also known as topological "persistence" ([117]). As mentioned earlier, persistent low peaks are considered noise, which is closely related to the pre-processing step of picking high/significant peaks. Furthermore, our goal is not only to pick (or detect) such peaks but also to extract valuable information from them. Briefly, given a real-valued function f on a compact set M, $f : M \to \mathbb{R}$, the concept of persistence implies pairing the critical points of f, i.e., minima and maxima (i.e., peaks) jointly in our study. The "relative height" refers to the concept of the difference between these critical values.

A commonly adopted method for determining the persistence of a peak is by means of upper-level set filtration,

where each peak represents a *topological feature*. These (topological) features are tracked from their emergence (so-called "birth") until they merge with larger (topological) features (so-called "death"). However, there is a significant drawback to utilizing the upper-level set filtration in our context. Namely, while tracking the persistence of each peak, we lose information about their positions (viz., the column indexes of the peaks), which is crucial for MALDI classification applications. Statistically speaking, the upper-level set filtration does not store any information on the explanatory variables' permutation indexes (i.e., the positions) and their respective persistence values. To overcome this limitation, we utilize an alternative analysis method called persistence transformation (cf. [117]). This approach simultaneously tracks the information regarding the positions along with the persistence of the peaks, allowing its application to MALDI data.

In Chapter 4, following applying the persistence transformation to the MALDI data, the subsequent step in our framework involves the classification of the persistence values into tumor classes. One of the classifiers we employ is the random forest (RF) algorithm, a machine-learning technique. An essential aspect of using this classifier is the selection of hyperparameters, also referred to as tuning parameters. In this thesis, we consider default tuning values for the random forest classifier, as implemented in *scikit-learn* ([81]) (*version 1.2.1*) for the function *'RandomForestClassifier()'*, save in the number of (random) trees. We consider 1000 trees for each experiment, while the default value is 100 trees. However, as concluded in [86], the number of trees is a parameter that is not "tunable in the classical sense", and higher values are generally preferred over smaller values in terms of performance (for more details, see [85]), especially for variable importance. Furthermore, it has concluded, by assessing the "tunability" of (different) algorithms and their associated hyperparameters, that random forest is much less "tunable" compared to other algorithms, such as support vector machines (for more details, see in [86] and the reference therein). All in all, LR is an algorithm that performs well with the default settings.

1.3 Organization of the thesis

This thesis follows a cumulative structure comprising a compilation of scientific publications. Accordingly, each chapter can be comprehended as an independent piece of work. Hereafter, we provide a concise overview of every chapter and the corresponding data application for that particular chapter.

1.3.1 Summary and applications

In Chapter 2, we introduce our inferential framework for two-sample comparisons in high-dimensional data regimes when the test statistics have an arbitrary covariance structure. Here, the outcome variable is binary (describing two lung cancer subtypes), and the procedure screens out all explanatory variables so as to discover the most associative ones with the target variable. In order to model the marginal associations, we consider logistic

regression with the logit (canonical) function. The criterion for controlling type I error is to maintain the approximated value for the FDP at a specific significance level α, let's say 10%. We utilize our procedure on a MALDI imaging dataset containing a large number of covariates (m/z values). Several researchers have previously analyzed this particular dataset. The outcomes obtained from our data analysis are in line with biologically meaningful findings documented in other studies. Specifically, these findings encompass five biomarkers, certain monoisotopic peaks, and isotopic patterns surrounding the biomarkers.

Chapter 3 is a multivariate extension of our previous work (i.e., Chapter 2), allowing simultaneous inference with more than two nominal classes. Here, we consider multinomial regression to model marginal associations, a specific case of multivariate generalized linear models (GLMs), wherein the outcome variables are expressed as multi-dimensional vectors. A single response category is selected as a baseline, and the multinomial logistic model is instituted by pairing each, except one, nominal response category with the baseline. Our framework approximates the FDP for each baseline-category logit pair. From an application point of view, we apply our procedure to a real MALDI dataset where the nominal target variable consists of three cancer (sub-)types, specifically two lung cancer statuses and pancreatic status. We assign pancreatic adenocarcinoma as the baseline, viz., the one human organ versus two cancer lung subtypes. Hence, our analysis aims to concurrently execute two (sub-)tasks: the pancreatic vs. lung adenocarcinoma and the pancreatic vs. lung squamous cell carcinoma comparison. To this end, our analysis aims to discover m/z values that can be distinctive for either cancer (sub-)type for each baseline-category pair.

Chapter 4 introduces our contribution to the second challenging task in terms of MALDI. This study introduces our algebraic topological framework that captures peak-related information from MALDI data and converts it into topological persistence representations. Subsequently, we utilize logistic regression and random forest classifiers on the derived topological persistence vectors to automate the process of tumor typing or subtyping. In this study, we applied our topological framework to a real MALDI dataset for binary classification. In order to showcase the competitiveness of our proposed framework, we conduct experiments on this dataset through two different cross-validation schemes. Additionally, we illustrate the efficacy of the single denoising parameter by evaluating its performance on (multiple) synthetic MALDI images. In brief, we first simulate multiple synthetic mass spectrometry (MS) images, which serve as ground truth images. Each pixel corresponds to the average value of a unique mass spectrum in these images. Following this, we artificially contaminate the spectral data responsible for generating the ground truth images. This (image) contamination is established by (artificially) adding different types and levels of noise to the ground truth images. Lastly, we apply our topological framework to the contaminated images, demonstrating pictorially how the persistence transformation effectively denoises these compromised images.

Chapter 5 serves as the concluding discussion of the thesis. This chapter is dedicated to summarizing the conclusions drawn from each preceding chapter, addressing computational challenges in the context of MALDI

and how (different) pre-processing steps affect our frameworks, outlining the contributions made by this thesis, discussing relevant works, and presenting an outlook with potential directions for future research.

As a side remark, we have received a comment from an anonymous reviewer regarding our last publication (i.e., Chapter 4) that the abbreviation imaging mass spectrometry (IMS) can be mistakenly understood as Ion-Mobility Spectrometry and the abbreviation mass spectrometry imaging (MSI) is preferred. To this end, the latter abbreviation has been used in Chapter 4, Synopsis, and Chapter 5 of the thesis, but not in the first two publications (i.e., Chapters 2 - 3). Throughout this thesis, it is important to note that the abbreviations "MSI" and "IMS" are synonyms (i.e., interchangeable) and refer to the same concept.

Chapter 2

Multiple two-sample testing under arbitrary covariance dependency with an application in imaging mass spectrometry

Information: This chapter is a slightly modified version of Vutov and Dickhaus (2023a) ([115]) published in *Biometrical Journal*.

Authors

Vladimir Vutov, Institute for Statistics, University of Bremen, Bremen, Germany

Prof. Dr. Thorsten Dickhaus, Institute for Statistics, University of Bremen, Bremen, Germany.

Declaration of individual contributions: Co-author and supervisor Prof. Dr. Thorsten Dickhaus has conceptualized the research project. I have performed data modeling and analysis, as well as R programming. Both authors have written the manuscript together.

Data Availability Statement: R code, with which all results presented in the present manuscript can be reproduced, is provided in the data file, which is available as supplementary material for this article. Biometrical Journal 65.2 (Feb. 2023), vol. 65, issue 2, 2100328 https://doi.org/10.1002/bimj.202100328.

Abstract: Large-scale hypothesis testing has become a ubiquitous problem in high-dimensional statistical inference, with broad applications in various scientific disciplines. One relevant application is constituted by imaging mass spectrometry (IMS) association studies, where a large number of tests are performed simultaneously in order to identify molecular masses that are associated with a particular phenotype, e.g., a cancer subtype. Mass spectra obtained from Matrix-assisted laser desorption/ionization (MALDI) experiments are dependent when considered

as statistical quantities. False discovery proportion (FDP) estimation and control under arbitrary dependency structure among test statistics is an active topic in modern multiple testing research. In this context, we are concerned with the evaluation of associations between the binary outcome variable (describing the phenotype) and multiple predictors derived from MALDI measurements. We propose an inference procedure in which the correlation matrix of the test statistics is utilized. The approach is based on multiple marginal models. Specifically, we fit a marginal logistic regression model for each predictor individually. Asymptotic joint normality of the stacked vector of the marginal regression coefficients is established under standard regularity assumptions, and their (limiting) correlation matrix is estimated. The proposed method extracts common factors from the resulting empirical correlation matrix. Finally, we estimate the realized FDP of a thresholding procedure for the marginal p-values. We demonstrate a practical application of the proposed workflow to MALDI IMS data in an oncological context.

2.1 Introduction

Imaging mass spectrometry (IMS) is a technique that acquires spatially resolved mass spectral information of small to large molecules. Provided a thin tissue section, mass spectra are collected in a spatially orientated pattern within the tissue. This produces an image where each discrete spot represents a mass spectrum. Mass spectra associate molecular masses to their relative molecular abundances. Hence, this provides insights into the chemical decomposition of a unique and specific region in the tissue. A promising technology that has evolved over recent years is Matrix-assisted laser desorption/ionization (MALDI) imaging mass spectrometry, also known as MALDI imaging. This technology allows for analyzing a wide range of analytes (e.g., proteins, peptides, lipids, etc.) from many types of biological samples. MALDI imaging is a versatile tool and has the advantage of combining spatial and molecular information from biological samples. This makes the technology interesting for biomedical and cancer research (for more applications in pathology, see [2, 59], among others). The latter is possible by virtue of its applicability to analyzing formalin-fixed paraffin-embedded (FFPE) tissue samples. One of the key benefits of utilizing MALDI IMS on fixed samples is that multiple FFPE core biopsies can be arranged in a single tissue microarray (TMA) block (see, [84, 16]). In other words, multiple tumor cores from different patients can be examined simultaneously (for more details, see [9]). As pointed out in [16], modern MALDI-IMS instruments manage to acquire molecular information with a high signal-to-noise ratio at short time measurements.

A challenging task for advanced bioinformatics tools, as acknowledged in [3], is stable feature extraction or, in other words, extracting biologically meaningful evidence out of a huge amount of spectra. Statistically, we model each spectrum individually as it is measured from small tissue core regions with slight fluctuating structures within a single core. This is a standard practice in modeling MALDI data (cf. [63, 16, 9]).

Large-scale multiple testing is a widely used methodology in the analysis of high-dimensional data and has a variety of applications in scientific fields like, e.g., genomics, proteomics, brain-computer interfacing, etc. (for

more life science applications, cf. Chapters 9-12 in [26]). Starting with the highly influential work by Benjamini and Hochberg ([11]), control of the expected proportion of false positive findings, called false discovery rate (FDR), has become a standard type I error criterion in large-scale multiple testing. Another well-known technique to control the FDR has been proposed in [105], and is often referred to as Storey's procedure. Its main idea is to fix a rejection threshold value t for the marginal p-values, then to estimate the FDR of the resulting thresholding procedure, and finally to choose t such that the estimated FDR is lower than or equal to the pre-defined FDR level α. Early FDR research has mainly established FDR control of the aforementioned procedures in the case of independent test statistics. However, high-dimensional studies seldom involve the analysis of independent variables. In contrast, most studies involve many related variables simultaneously (cf., among many others, [102, 44]). Similarly, MALDI-IMS data consist of a couple of thousands of variables, and many of them are related. Explicitly taking into account these dependencies can increase the power of the multiple test, cf. [27] for an overview of so-called multivariate multiple tests.

There are multivariate multiple tests, which are based on block structures in the data. For instance, in [102], it has proposed to control the family-wise error rate (FWER) in blocks of adjacent genetic markers; see also Section 5 in [103]. Likewise, in [104], the authors have reported an extensive study to compare different controlling methods based on the assumption of block-correlation positively dependent tests. However, there is no evidence that MALDI-IMS data can be grouped straightforwardly into adjacent blocks. Other methods utilize a multi-factor model in order to describe the dependencies among the test statistics, meaning that the latter dependency structure may be explained by latent factors.

In addition to modeling the dependencies, a further task is to integrate the correlation effects in the decision process; see, for example, [33, 32, 62]. In [38] has been introduced a general setting for approximating the false discovery proportion (FDP). They have assumed that the test statistics are (approximately) following a multivariate normal distribution with an arbitrary and known covariance matrix. The idea of their approach is to carry out a spectral decomposition of the covariance matrix of the test statistics and then to subtract the principal factors that cause the strong dependency across the z-values before evaluating the FDP. This method is called principal factor approximation (PFA). The authors ([37]) have established a fully data-driven process to estimate the FDP, where the authors adopted a POET estimator (see [39]) to estimate an unknown covariance matrix and subsequently compute the realized FDP. Recently, in [40] has addressed the problem when the assumption of normality is violated, for instance, in the context of multiple testing under arbitrary dependency and heavy-tailed data. The method utilizes a robust covariance estimator and constructs factor-adjusted test statistics.

Since the FDR is the expected value of the FDP, FDP control and FDR control are related concepts. In the present work, we explore the problem of two-sample multiple hypotheses testing under arbitrary correlation dependency under the scope of multiple marginal logistic regression models by making use of PFA. We propose to estimate the realized value of the FDP of a thresholding procedure in this context and to choose the rejection thresh-

old such that this estimated FDP value equals a given constant $\alpha \in (0, 1)$. Furthermore, we apply our proposed method to MALDI imaging data.

One common approach for extracting meaningful variables in the MALDI context is based on discovering significant signal peaks, which is also known as peak detection (cf. [119]). These peaks are anticipated to help distinguish spectra from different cancerous classes. For example, [118] introduced a peak detection method that carries out dual-tree complex wavelet transformation. Moreover, a novel spatial approach has been proposed in [66]. Namely, the latter study proposed to incorporate the neighboring information (an isotope pattern) around the selected peaks, which can (potentially) enhance the peak picking process (a so-called "spatially-aware approach"). In the isotope context, e. g. in [100], a methodology has been proposed, which discovers this type of pattern in MALDI data. For the aforementioned studies, the (column) indices of the peaks (i. e., the indices of explanatory variables) are important for their estimation procedures. In methods of that type, based on selected peaks, further analysis is performed in order to find a subset of those peaks that are statistically related to the response variable. This has been pointed out by, for instance, in [16].

On the other hand, there are also other methods for MALDI data analysis which use "more of the intrinsic information hidden in the data by exploring statistical correlations" ([16]; cf. also [55], [25], and [51]). Since the spectral data are non-negative, a popular approach is the usage of non-negative matrix factorization. Methods of that type aim at representing the original data in a lower-dimensional space by combining "characteristic spectral patterns" (cf. [63, 16]) given the non-negativity constraint. The latter studies proposed so-called "automated tumor typing": After mapping the original data into a lower-dimensional space, supervised classification methods are carried out on the resulting feature vectors in order to classify observational units into tumor subtypes. These methods are invariant with respect to the permutation of the indices of the explanatory variables, but rather they exploit similarities among these variables. Our proposed methodology is more in the spirit of these latter methods, which are invariant with respect to the indices of the explanatory variables and instead use statistical similarities (in our case, quantified by correlations) to reduce the (effective) dimensionality of the feature space.

The remainder of the work is structured as follows. In Section 2.2, we describe the proposed two-sample multiple testing framework. Section 2.3 is devoted to computer simulations. We demonstrate the practical application to a MALDI-IMS study in Section 2.4, and we conclude with a discussion in Section 2.5.

2.2 Proposed methodology

In MALDI Imaging-related studies, data are commonly stored in an $n \times p$ matrix $X = (x_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le p}}$, where spectra are stored as rows and columns correspond to mass-to-charge (m/z) values (in the context of MALDI interpreted as molecular masses, cf. [3]). Usually, both n and p are in thousands. We address the biological question of the association between m/z values and a cancerous status by testing multiple hypotheses. More specifically, we test

for each j the null hypothesis H_{0j} which states there is no association between a particular m/z value X_j and the cancer subtype.

2.2.1 Marginal Modelling

The first step of the proposed framework is to model the marginal associations for each j separately. Therefore, let j be arbitrary, but fixed throughout this section. The motivation for this marginal modeling is that the number p of potential regressors is very large in our context, such that the available sample size n does not allow for fitting a multivariate model which incorporates all regressors in one model. Screening for potentially relevant regressors by means of marginal logistic regression models is common practice, as remarked, for instance, in the section entitled "Variables inclusion and selection" of [101]. Marginal logistic regression models have also been considered and analyzed in [82]. Let Y denote the (random) binary outcome, and let X_j denote the random variable describing the j-th m/z value. Thus, the tuple (X_j, Y) takes its values in $\mathbb{R}\times\{0,1\}$. We are interested in the conditional distribution $\mathbb{P}(Y|X_j)$. To this end, we assume a (marginal) binary regression model with the canonical (logit) link function. This model has two parameters, namely, the intercept α_j and the regression coefficient β_j. We denote the observational units for the j-th marginal regression problem by $(X_j^{(i)}, Y^{(i)})_{1\le i\le n}$, and we assume that they are independent copies of (X_j, Y). Letting, for a given $i \in \{1,\dots,n\}$, $\pi_j^{(i)} = \mathbb{P}(Y^{(i)} = 1|X_j^{(i)})$, the model equation for the j-th binary logistic regression models is given by

$$g(\pi_j^{(i)}) := \log\left(\frac{\pi_j^{(i)}}{1-\pi_j^{(i)}}\right) = \alpha_j + X_j^{(i)}\beta_j, \tag{2.1}$$

where $g = \text{logit}$ is the canonical link function mentioned before. The unknown parameters (α_j,β_j) are estimated by the principle of the maximum (log-) likelihood. The log-likelihood function pertaining to the model in (2.1) is given by

$$l(\alpha_j,\beta_j) = \sum_{i=1}^{n} Y^{(i)}\left[\log\pi_j^{(i)} - \log(1-\pi_j^{(i)})\right] + \log(1-\pi_j^{(i)}). \tag{2.2}$$

By substituting

$$\pi_j^{(i)} = \frac{\exp(\alpha_j + X_j^{(i)}\beta_j)}{1+\exp(\alpha_j + X_j^{(i)}\beta_j)} \quad \text{as well as} \quad 1-\pi_j^{(i)} = \frac{1}{1+\exp(\alpha_j + X_j^{(i)}\beta_j)}$$

in (2.2), we obtain that

$$l(\hat\alpha_j,\hat\beta_j) = \max_{(\alpha_j,\beta_j)} \sum_{i=1}^{n} Y^{(i)}\left(\alpha_j + X_j^{(i)}\beta_j\right) - \log\left(1+\exp(\alpha_j + X_j^{(i)}\beta_j)\right), \tag{2.3}$$

where the estimation is performed conditionally to the actually observed values $X_j^{(i)} = x_j^{(i)}$ for $1 \le i \le n$.

In this study, we are concerned with simultaneous testing of the pairs of hypotheses

$$H_{0j} : \beta_j = 0 \text{ versus } H_{1j} : \beta_j \neq 0, \quad j = 1, \ldots, p. \tag{2.4}$$

This means, that we test a family $\mathcal{H}_p = \{H_{01}, \ldots, H_{0p}\}$ of p null hypotheses, and a binary decision (rejection or non-rejection) is made for each of these p null hypotheses on the basis of the study data at hand. The result of the data analysis, therefore, is a binary decision vector $\mathbf{d} \in \{0, 1\}^p$. The j-th entry d_j of this decision vector encodes the decision referring to the (marginal) test problem H_{0j} versus H_{1j} for $1 \leq j \leq p$, where $d_j = 1$ means (by convention) that H_{0j} is rejected in favor of H_{1j}, while $d_j = 0$ means that H_{0j} is not rejected. Biologically speaking, we aim at discovering the most distinctive m/z values for a cancer association.

2.2.2 Multiple Marginal Models

The second step of the proposed procedure is to combine all p marginal models and to approximate the joint null distribution of all estimators. To this end, we follow the framework described in [83] for jointly estimating multiple marginal association parameters, and apply this framework to the marginal models described in the previous section. Notice that we assume that regression coefficients are unique to one model j and not shared between any two models $j_1 \neq j_2$. Furthermore, the intercepts $(\alpha_j)_{1 \leq j \leq p}$ are nuisance parameters in the sense that the hypotheses in (2.4) only refer to the β_js.

The main goal of this section is to establish a central limit theorem for the vector $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\top$, which is achieved by stacking the score contributions of the $\hat{\beta}_j$'s across all p marginal models. Following [83], we consider the asymptotic ($n \to \infty$) expansion

$$(\hat{\beta}_j - \beta_j)\sqrt{n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_{ij} + o_{\mathbb{P}}(1), \tag{2.5}$$

where

$$\Psi_{ij} = F^{(i)}(\beta_j)^{-1} \widetilde{\Psi}_{ij}(\beta_j). \tag{2.6}$$

In (2.6), $F^{(i)}(\beta_j)^{-1}$ is the row corresponding to β_j of the inverse Fisher information matrix $F^{(i)}(\alpha_j, \beta_j)^{-1}$ for the i-th observational unit, $\widetilde{\Psi}_{ij}(\beta_j)$ is the score function (the first derivative of the log-likelihood function) pertaining to β_j for the i-th observational unit, and $o_{\mathbb{P}}(1)$ indicates a sequence of random variables converging to zero in probability. For our (marginal) logistic model, $\widetilde{\Psi}_{ij}$ is the score function pertaining to α_j and β_j, see the derivation above Equation 2.9 (i.e., $\hat{\Psi}_{ij}$).

The score function and the Fisher information matrix for the i-th observational unit are under our model given

by

$$\begin{aligned}\widetilde{\Psi}_{ij}(\alpha_j,\beta_j) &= \begin{pmatrix} Y^{(i)} - \pi_j^{(i)} \\ X_j^{(i)}(Y^{(i)} - \pi_j^{(i)}) \end{pmatrix}, \\ F^{(i)}(\alpha_j,\beta_j) &= \begin{pmatrix} \pi_j^{(i)}(1-\pi_j^{(i)}) & X_j^{(i)}\pi_j^{(i)}(1-\pi_j^{(i)}) \\ X_j^{(i)}\pi_j^{(i)}(1-\pi_j) & X_j^{(i)}X_j^{(i)}\pi_j^{(i)}(1-\pi_j^{(i)}) \end{pmatrix}.\end{aligned}$$

Furthermore, due to the stochastic independence of the observational units, the score function and the Fisher information matrix for the entire sample (in coordinate j) are given by summing up the respective contributions of the n observational units. The maximum likelihood estimators $\hat{\alpha}_j$ and $\hat{\beta}_j$ appearing in (2.3) can be found in practice by (numerically) solving the score equations, meaning that $(\hat{\alpha}_j, \hat{\beta}_j)$ are (numerically) found such that $\sum_{i=1}^n \widetilde{\Psi}_{ij}(\hat{\alpha}_j, \hat{\beta}_j) = (0,0)^\top$ is fulfilled.

Now, we define the vectors $\beta := (\beta_1, \ldots, \beta_p)^\top$, $\hat{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\top$, and $\Psi_i := (\Psi_{i1}, \ldots, \Psi_{ip})^\top$, and consider the asymptotic expansion

$$(\hat{\beta} - \beta)\sqrt{n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \Psi_i + o_{\mathbb{P}}(1), \tag{2.7}$$

which follows from (2.5) under standard regularity assumptions like, for instance, finiteness of the Fisher information and non-vanishing (limiting) proportion of data points corresponding to $Y = 1$ and $Y = 0$, respectively. The left-hand side of (2.7) converges in distribution, by the multivariate central limit theorem, to a p-variate normal distribution, i. e.,

$$(\hat{\beta} - \beta)\sqrt{n} \xrightarrow{d} N_p(0, \Sigma).$$

The limiting variance-covariance matrix Σ can be estimated in a consistent manner, namely by

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \hat{\Psi}_i^\top \hat{\Psi}_i, \tag{2.8}$$

whence $\hat{\Psi}_i$ are yielded by plugging the parameter estimates from all marginal models into Ψ_i.

In our study, we are genuinely interested in the effect of β_j for $j \in \{1, \ldots, p\}$. However, the intercepts $(\alpha_j)_{1 \le j \le p}$ contribute to the estimation and standardisation of the β_j's. Specifically, for the logit model described in the previous section, $\hat{\Psi}_{ij}$ is given by the second coordinate of the bivariate vector

$$\left\{\hat{\pi}_j^{(i)}(1-\hat{\pi}_j^{(i)})(1, X_j^{(i)})^\top (1, X_j^{(i)})\right\}^{-1} (1, X_j^{(i)})^\top (Y^{(i)} - \hat{\pi}_j^{(i)}),$$

where $\hat{\pi}_j^{(i)} = \frac{\exp(\hat{\alpha}_j + X_j^{(i)}\hat{\beta}_j))}{1+\exp(\hat{\alpha}_j + X_j^{(i)}\hat{\beta}_j)}$ and $\hat{\alpha}_j, \hat{\beta}_j$ are as in (2.3).

Next, we denote by $Z_1, \ldots, Z_p$ the Studentized versions of $\hat{\beta}_1, \ldots \hat{\beta}_p$, meaning that

$$Z_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}}, \quad j = 1, \ldots, p, \tag{2.9}$$

where $\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}$ is the square root of the j-th diagonal element of $\hat{\Sigma}$, divided by $\sqrt{n}$. Then, we have that

$$(Z_1, Z_2, \ldots, Z_p)^\top \underset{\text{approx.}}{\sim} N_p((\mu_1, \mu_2, \ldots, \mu_p)^\top, \hat{\Sigma}^*), \tag{2.10}$$

where $\mu_j = \beta_j / \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}$ for $1 \leq j \leq p$, $\hat{\Sigma}^* = \mathrm{diag}[\hat{\Sigma}]^{-1/2} \cdot \hat{\Sigma} \cdot \mathrm{diag}[\hat{\Sigma}]^{-1/2}$ is the correlation matrix pertaining to $\hat{\Sigma}$, and the notation $\underset{\text{approx.}}{\sim}$ indicates the approximate distribution for large n. Assuming that $\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}$ is positive for all $j \in \{1, \ldots, p\}$, $\beta_j = 0$ if and only if $\mu_j = 0$. Thus, the family of hypotheses from (2.4) can then equivalently be expressed as

$$H_{0j} : \mu_j = 0 \ \text{ versus } \ H_{1j} : \mu_j \neq 0, \ \ j = 1, \ldots, p, \tag{2.11}$$

although μ_j depends on the data via $\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}$ and is therefore not a statistical parameter.

2.2.3 Approximation of the false discovery proportion

Throughout this manuscript, we consider the multiple test problem which is given by the p pairs of null and alternative hypotheses specified in (2.11). Let $p_0 = \#\{j : \mu_j = 0\}$ denote the number of true null hypotheses and $p_1 = \#\{j : \mu_j \neq 0\}$ the number of false null hypotheses, such that $p = p_0 + p_1$. Throughout the remainder, we make the following sparsity assumption.

Assumption 1 *The number p_1 of false null hypotheses is very small compared to the number p of all null hypotheses.*

In an asymptotic setting where $p = p(n)$ tends to infinity as n tends to infinity, Assumption 1 can be formalized as $p_0(n)/p(n) \to 1$. In the present work, however, we rely only on asymptotics in the sample size n and regard p as fixed. For the calibration of a multiple test with respect to type I error control, we proceed similarly as in Storey's method (see [105]). Namely, for a (data-dependent) threshold t, we will reject the null hypotheses, which correspond to those p-values that are not exceeding t. This approach has been broadly used in practice (e.g., see, [38, 37, 33, 32, 105]). The aim of the proposed method is to estimate the realized FDP for any fixed t in the multiple testing setting given by (2.11), based on the Z-statistics (2.10) under an arbitrary structure of Σ.

To this end, we consider empirical processes given by

$$V(t) = \#\{\text{true null } P_j : P_j \leq t\},$$
$$S(t) = \#\{\text{false null } P_j : P_j \leq t\},$$
$$R(t) = \#\{P_j : P_j \leq t\},$$

where t ranges in $[0, 1]$. For a given value of t, the null hypothesis $H_{0,j}$ is rejected if and only if its corresponding p-value p_j does not exceed t. This decision rule leads to the decision pattern, which is displayed in Table 2.1. The random variables $V(t)$, $S(t)$, and $R(t)$ are the number of false discoveries (i. e., false rejections), the number of true discoveries, and the total number of discoveries, respectively. Clearly, $R(t) = V(t) + S(t)$. The latter random variables depend on the test statistics $Z_1, Z_2, \ldots, Z_p$, because every p-value P_j is a transformation of the corresponding Z-statistic Z_j, $1 \leq j \leq p$, as we will describe below. Furthermore, $V(t)$ and $S(t)$ are both unobservable, whereas $R(t)$ is observable. We recall here the definition of the FDP, namely, $\text{FDP}(t) = V(t)/\max\{R(t), 1\}$. Table 2.1 provides a summary of the relevant quantities.

Table 2.1: Decision pattern of the multiple test which thresholds the marginal p-values at a given value $t \in [0, 1]$.

Number	Number accepted	Number rejected	Overall
True nulls	$U(t)$	$V(t)$	p_0
False nulls	$T(t)$	$S(t)$	p_1
All nulls	$p - R(t)$	$R(t)$	p

2.2.4 Principal Factor Approximation

The next step of the analysis is to model and to utilize the dependency structure of the test statistics in an approximation of $\text{FDP}(t)$ for a given t. The proposed technique relies on an approximation of a normally distributed random vector with a factor model involving weakly dependent, normally distributed random errors. In our case, we use the factor model as a tool to approximate the correlation matrix $\hat{\Sigma}^*$ with a reduced number of parameters, without actually assuming that latent factors are involved in the data-generating process. To this end, we first employ a spectral decomposition of the correlation matrix $\hat{\Sigma}^*$ (cf. [38]). Namely, $\hat{\Sigma}^*$ is represented in terms of its eigenvalue-eigenvector pairs $(\lambda_j, \gamma_j)_{1 \leq j \leq p}$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. The representation can be written as

$$\hat{\Sigma}^* = \lambda_1 \gamma_1 \gamma_1^\top + \lambda_2 \gamma_2 \gamma_2^\top + \ldots + \lambda_p \gamma_p \gamma_p^\top. \tag{2.12}$$

For a fixed integer $k \geq 1$, we let $A_k = \sum_{j=k+1}^{p} \lambda_j \gamma_j \gamma_j^\top$, and we note that

$$\|A_k\|^2 = \lambda_{k+1}^2 + \ldots \lambda_p^2,$$

where $\|.\|$ is the Frobenius norm. We further let $L_k = (\sqrt{\lambda_1}\gamma_1, \sqrt{\lambda_2}\gamma_2, \ldots, \sqrt{\lambda_k}\gamma_k)$, which presents a $p \times k$ matrix. Thus, $\hat{\Sigma}^*$ can be written as,

$$\hat{\Sigma}^* = L_k L_k^\top + A_k.$$

Respectively, $Z_1, \ldots, Z_p$ can be approximated by

$$\mu_j + b_j^\top W + K_j = \mu_j + \eta_j + K_j, \quad j = 1, \ldots, p, \tag{2.13}$$

where $b_j = (b_{j1}, \ldots, b_{jk})^\top$ and $(b_{1j}, \ldots, b_{pj})^\top = \sqrt{\lambda_j}\gamma_j$. The vector $W = (W_1, \ldots, W_k)^\top \sim N_k(0, I_k)$ is called the vector of common factors, and these factors are stochastically independent of each other. The random vector $(K_1, \ldots, K_p)^\top \sim N_p(0, A_k)$ is called the vector of random errors, and it is assumed that factors and random errors are stochastically independent. We can think of (2.13) as an approximation of the data-generating process for $Z_1, \ldots, Z_p$. In this interpretation, $\mu_j = 0$ corresponds to the true null hypotheses, and $\mu_j \neq 0$ corresponds to the false null hypotheses.

It is essential to choose the number k of factors carefully. On the one hand, it is important to choose k large enough to capture most of the dependencies among $Z_1, \ldots, Z_p$. On the other hand, a small k stabilizes the computations, both from a numerical and from statistical point of view. In [38], the authors have discussed one way to determine a suitable value of k. Concretely, the authors proposed to choose the smallest k such that

$$\frac{\sqrt{\lambda_{k+1}^2 + \ldots + \lambda_p^2}}{\lambda_1 + \ldots + \lambda_p} < \epsilon, \tag{2.14}$$

where ϵ is some small number, for example, 0.01. It has been pointed out in [37] that an overestimation of k does not invalidate the approximation of the FDP, as long as the factor approximation of $\hat{\Sigma}^*$ can still be estimated with reasonable accuracy.

Based on the aforementioned derivations, we consider the "principal factor" FDP estimator from Proposition 2 in [38], which is given by

$$\widehat{\mathrm{FDP}}(t) = \min\left\{\sum_{j=1}^{p}\left[\Phi(a_j(z_{t/2} + \widehat{\eta}_j) + \Phi(a_j(z_{t/2} - \widehat{\eta}_j))\right], R(t)\right\}/R(t) \tag{2.15}$$

whenever $R(t) \neq 0$, and $\widehat{\mathrm{FDP}}(t) = 0$ in the case of $R(t) = 0$. In (2.15), $a_j = (1 - \sum_{h=1}^{k} b_{jh}^2)^{-1/2}$ and $R(t) = \{j : 2\Phi(-|Z_j|) \leq t\}$ is the (total) number of rejections for a given t, where Φ and $z_{t/2} = \Phi^{-1}(t/2)$ are the cumulative distribution function and the lower $t/2$-quantile of the standard normal distribution on $\mathbb{R}$, respectively. The (unadjusted) two-sided (random) p-value corresponding to Z_j is given by $P_j = 2\Phi(-|Z_j|) = 2(1 - \Phi(|Z_j|))$, and this p-value is thresholded at t for every $j \in \{1, \ldots, p\}$ when computing $R(t)$. Furthermore, $\hat{\eta}_j = \sum_{h=1}^{k} b_{jh}\hat{w}_j$ is an

estimator for $\eta_j = b_j^\top w$, where $w = (w_1, \ldots, w_k)^\top$ denotes the value of $W = (W_1, \ldots, W_k)^\top$.

The estimator in (2.15) relies on the intuition that large $|\mu_j|$'s tend to generate large $|z_j|$'s, meaning that false null hypotheses tend to produce large Z-statistics (in absolute value). Furthermore, the estimator in (2.15) relies on Assumption 1, namely, that the number p_0 of true null hypotheses is close to p. This assumption justifies the summation over all j from one to p in (2.15). Different FDP estimators have been compared in [95], and under sparsity in the aforementioned sense, the author has proposed to use the estimator from (2.15). There are several reasons why the assumption of sparsity is plausible in our study. Firstly, due to high sensitivity during sample preparation and acquisition, there is evidence of a small signal-to-noise ratio. Secondly, a reasonable assumption is that solely a tiny fraction of molecular masses are distinctive for a cancer association. In fact, we have applied the proposed method to real MALDI data, where there have been characterized five biomarkers (i. e., biologically meaningful covariates) out of a couple of thousands of measured covariates.

In order to evaluate (2.15) in practice, it remains to specify the estimator of w. In [38], the authors have proposed to construct this estimator by means of L_2-regression or by means of L_1-regression, respectively. For the former, the authors proposed to include only the 95% smallest $|z_j|$'s in the regression fit. Specifically, the estimator based on L_2-regression is given by

$$\hat{w} = \arg\min_{v \in \mathbb{R}^k} \sum_{j=1}^{\lfloor 0.95p \rfloor} (Z_j - b_j^\top v)^2, \tag{2.16}$$

where we assume that the Z_j's in (2.16) are ordered from small to large according to their absolute values. This estimator has been used in our simulation study. The estimator based on L_1-regression is given by

$$\hat{w} = \arg\min_{v \in \mathbb{R}^k} \sum_{j=1}^{p} |Z_j - b_j^\top v|. \tag{2.17}$$

We adopted L_1-regression rather than L_2-regression, because it is more robust to outliers. The consistency of $\hat{w}$ has been discussed in [38] under model assumptions that are similar to ours.

Finally, the dependency-adjusted (random) p-values corresponding to the Z_j's are given by

$$\tilde{P}_j = 2\Phi(-|a_j(Z_j - b_j^\top \hat{w})|). \tag{2.18}$$

The null hypothesis H_{0j} from (2.11) gets rejected based on the observed data, iff $\tilde{p}_j \leq t$, $1 \leq j \leq p$. In this, the data-dependent rejection threshold is chosen as the largest value $t = t_\alpha \in [0, 1]$ such that $\widehat{\mathrm{FDP}}(t_\alpha)$ is not exceeding a pre-defined level α. In practice, a (grid) search algorithm can be employed to find the value t_α for a given level α.

2.2.5 Schematic description of the entire data analysis workflow

Algorithm 1 provides a step-by-step description of the proposed data analysis workflow.

Algorithm 1: The Logit-PFA method

1: Fit the marginal logistic regression model with the logit link function for each $j \in \{1, \ldots, p\}$ separately on the basis of $(X_j^{(i)} : 1 \leq i \leq n)$ and $(Y^{(i)} : 1 \leq i \leq n)$.
(1.1) Find the maximum-likelihood estimates for $\hat{\beta}_j$ and $\hat{\alpha}_j$.
(1.2) Calculate the standardized score contributions $\hat{\Psi}_{ij}$ based on $\hat{\beta}_j$ for $i \in \{1, \ldots, n\}$ and stack them on top of each other to build a vector $\hat{\Psi}_i$.
(1.3) Calculate the estimated covariance matrix $\hat{\Sigma}$, given in (2.8), based on $(\hat{\Psi}_i : 1 \leq i \leq n)$, and obtain the correlation matrix pertaining to $\hat{\Sigma}$.
(1.4) Calculate the Z-statistics given in (2.9).
2: Choose a grid $\mathcal{G} \subset [0, 1]$ of candidate values for the rejection threshold t.
3: Based on the Z-statistics, evaluate $R(t)$ for each $t \in \mathcal{G}$.
4: Apply singular value decomposition to the correlation matrix pertaining to $\hat{\Sigma}$, and determine an appropriate number of factors k. Then, extract the corresponding factor loading coefficients $\{b_{jh} : j = 1, \ldots, p; h = 1, \ldots, k\}$.
5: Obtain the estimate $\hat{w}$ of the values of the common factors by means of regression; cf. (2.16) and (2.17), respectively. Plug these factor estimates into (2.15), and obtain the estimate $\widehat{\mathrm{FDP}}(t)$, for each $t \in \mathcal{G}$.
6: For a given value of $\alpha \in (0, 1)$, find the largest value $t_\alpha \in \mathcal{G}$ fulfilling $\widehat{\mathrm{FDP}}(t_\alpha) \leq \alpha$.
7: Obtain adjusted p-values according to (2.18).
8: Threshold the adjusted p-values at t_α.

2.3 Computer simulations

In this section, we illustrate the performance of the proposed approach based on simulated data under different data-generating processes. Specifically, we consider the sample size $n = 400$, the number of false nulls hypotheses $p_1 = 10$, and the total number of hypotheses $p \in \{500, 1000\}$. For each combination of these parameters, 1,000 simulation runs have been performed. For a given value of p_1, we assume without loss of generality that $\beta_j \neq 0$ for $j \in \{1, \ldots, p_1\}$, while the p_0 true nulls with $\beta_j = 0$ correspond to the coordinates $j \in \{p_1 + 1, \ldots, p\}$. We employed the least-squares estimator, defined in (2.16), for the estimation of the values of the common factors. For each observational unit $i \in \{1, \ldots, n\}$, the simulation data have been generated according to the model

$$\mathbb{P}_\beta(Y^{(i)} = 1 | X^{(i)}) = \frac{\exp\left(\sum_{j=1}^{p_1} \beta_j X_j^{(i)}\right)}{1 + \exp\left(\sum_{j=1}^{p_1} \beta_j X_j^{(i)}\right)} \tag{2.19}$$

for the response variable $Y^{(i)}$ given the covariate vector $X^{(i)} = \left(X_1^{(i)}, \ldots, X_p^{(i)}\right)$. Moreover, we have set $\beta_j = 1$ for all $j \in \{1, \ldots, p_1\}$. As remarked, for instance, in Section 3.2 of [116], the marginal logistic regression model given by (2.1) is, in general, incorrect if the true model is a multiple logistic regression model as in (2.19) with continuous covariates. However, the regression coefficient pertaining to covariate X_j is zero in both models if the distribution of Y does not depend on X_j and X_j is stochastically independent of all those covariates which have an influence on Y. In this sense, screening for potentially relevant regressors by marginal models is still meaningful under

the multiple logistic regression model given in (2.19). The considered data-generating distributions for the vector $X = (X_1, \ldots, X_p)^\top$ are provided in Model 1.

Model 1

Scenario 1: $X_1, \ldots, X_p$ *are stochastically independent and identically* $N(0,1)$*-distributed random variables.*
Scenario 2: $X_1, \ldots, X_p$ *are jointly normally distributed on* $\mathbb{R}^p$*. The parameters of their p-variate joint normal distribution have been chosen such that each* X_j *is marginally* $N(0,1)$*-distributed,* $j = 1, \ldots, p$*. Furthermore, the correlation coefficient* $Corr(X_{j_1}, X_{j_2})$ *equals* ρ *for all* $1 \leq j_1 < j_2 \leq p_1$ *as well as for all* $p_1 + 1 \leq j_1 < j_2 \leq p$ *(Gaussian equi-correlation model). The subvector* $(X_j : 1 \leq j \leq p_1)$ *is stochastically independent of the subvector* $(X_j : p_1 + 1 \leq j \leq p)$*, to avoid spurious effects of covariates* X_j *with* $p_1 + 1 \leq j \leq p$ *on the response variable which arise from confounding of covariates* X_j *with* $1 \leq j \leq p_1$*.*
Scenario 3: *As Scenario 2, but now* $(X_j : 1 \leq j \leq p_1)$ *are stochastically independent and identically* $N(0,1)$*-distributed random variables.*

Table 2.2: Simulation results under Scenario 1 (I). The total number of hypotheses equals $p = 500$ in the first row and $p = 1{,}000$ in the second row; the number of factors equals $k = 10$; the rejection threshold equals $t = 10^{-4}$, except for the last column.

Median of $\widehat{\mathrm{FDP}}(t)$	Standard Error of $\widehat{\mathrm{FDP}}(t)$	Average of $R(t)$	Standard Error of $R(t)$	Average of $S(t)$	Standard Error of $S(t)$	Median of $t_{0.05}$
0.004144	0.000918	6.930	1.247	6.892	1.236	1.24e-03
0.009116	0.002175	6.965	1.270	6.898	1.228	6.6e-04

Table 2.3: Simulation results under Scenario 1 (II). The total number of hypotheses equals $p = 500$ in the first row and $p = 1{,}000$ in the second row; the number of factors equals $k = 10$; the rejection threshold equals $t = 0.005$.

Median of $\widehat{\mathrm{FDP}}(t)$	Standard Error of $\widehat{\mathrm{FDP}}(t)$	Average of $R(t)$	Standard Error of $R(t)$	Average of $S(t)$	Standard Error of $S(t)$
0.158180	0.022796	11.774	1.664	9.519	0.632
0.277878	0.045845	14.062	2.194	9.517	0.650

Table 2.4: Simulation results under Scenario 2 (I). The total number of hypotheses equals $p = 500$; the number of factors equals $k = 1$; the rejection threshold equals $t = 10^{-4}$, except for the last column.

ρ	Median of $\widehat{\mathrm{FDP}}(t)$	Standard Error of $\widehat{\mathrm{FDP}}(t)$	Average of $R(t)$	Standard Error of $R(t)$	Median of $t_{0.05}$
0.2	0.001395	0.009750	10.030	0.182	2.41e-03
0.5	0.000131	0.024349	10.025	0.250	7.41e-03
0.8	0.000099	0.018323	10.021	0.572	3.49e-02

Table 2.5: Simulation results under Scenario 2 (II). The total number of hypotheses equals $p = 1000$; the number of factors equals $k = 1$; the rejection threshold equals $t = 10^{-4}$, except for the last column.

ρ	Median of $\widehat{\text{FDP}}(t)$	Standard Error of $\widehat{\text{FDP}}(t)$	Average of $R(t)$	Standard Error of $R(t)$	Median of $t_{0.05}$
0.2	0.002842	0.014190	10.082	0.306	1.28e-03
0.5	0.000167	0.028907	10.059	0.400	4.69e-03
0.8	0.000099	0.024967	10.028	0.447	2.74e-02

Table 2.6: Simulation results under Scenario 3 (I). The total number of hypotheses equals $p = 500$; the number of factors equals $k = 1$; the rejection threshold equals $t = 10^{-4}$, except for the last column.

ρ	Median of $\widehat{\text{FDP}}(t)$	Standard Error of $\widehat{\text{FDP}}(t)$	Average of $R(t)$	Standard Error of $R(t)$	Average of $S(t)$	Standard Error of $S(t)$	Median of $t_{0.05}$
0.2	0.002292	0.012646	6.915	1.256	6.902	1.240	2.41e-03
0.5	0.000221	0.026264	6.920	1.313	6.898	1.240	7.41e-03
0.8	0.000142	0.021452	6.907	1.295	6.895	1.237	3.49e-02

Table 2.7: Simulation results under Scenario 3 (II). The total number of hypotheses equals $p = 1000$; the number of factors equals $k = 1$; the rejection threshold equals $t = 10^{-4}$, except for the last column.

ρ	Median of $\widehat{\text{FDP}}(t)$	Standard Error of $\widehat{\text{FDP}}(t)$	Average of $R(t)$	Standard Error of $R(t)$	Average of $S(t)$	Standard Error of $S(t)$	Median of $t_{0.05}$
0.2	0.004194	0.021576	6.938	1.318	6.872	1.272	1.28e-03
0.5	0.000250	0.039067	6.916	1.328	6.870	1.273	4.69e-03
0.8	0.000143	0.024121	6.871	1.288	6.877	1.273	2.74e-02

For the simulation of correlated independent variables, we have used the function `rmvnorm` from the `R` (see [87]) package `mvtnorm`. Furthermore, we have used the function `pfa.test()` from [36] for implementing the principal factor approximation method. In Tables 2.2 - 2.7, we report summaries (over the 1,000 simulation runs) of $\widehat{\text{FDP}}(t)$, $R(t)$, and $S(t)$ for fixed values of t, and we report the median value of t_α for the common choice of $\alpha = 0.05$.

Tables 2.2 - 2.3 summarize our simulation results under Scenario 1. Here, due to joint independence of the test statistics, $t_{0.05}$ is rather small because the "effective number of tests" (in the sense of Section 3.4 in [27] and the references therein) equals p under joint independence of the test statistics, meaning that a rather strong multiplicity correction is required. On the other hand, the standard error of $\widehat{\text{FDP}}(t)$ is rather small under Scenario 1, too, because the FDP concentrates well around its expectation (the FDR) under joint independence of all p test statistics. The

results given in Tables 2.4 - 2.5 refer to Scenario 2 of Model 1, and Tables 2.6 - 2.7 refer to Scenario 3. Under Scenarios 2 and 3, the effective number of tests is smaller than p whenever $\rho > 0$, and it decreases with increasing ρ. Thus, $t_{0.05}$ increases with ρ, too. Under our Scenario 2, the considered multiple test always rejected all ten false null hypotheses. For this reason, we do not report summaries of $S(t)$ in Tables 2.4 and 2.5. The reason for the high power of the multiple test under Scenario 2 is, that the correlation among $(X_j : 1 \leq j \leq p_1)$ amplifies the signal strength for each $j \in \{1, \ldots, p_1\}$. Under Scenario 3, where the relevant regressors are stochastically independent, the power of the multiple test is smaller than under Scenario 2, such that on average, only approximately seven of the ten false null hypotheses can be rejected by the multiple test considered in Tables 2.6 and 2.7.

2.4 Real data analysis

2.4.1 Description of the dataset

We applied the proposed multiple testing approach to a MALDI IMS data frame introduced in [60]. The authors ([60]) have characterized five biomarkers. Broadly speaking, biomarkers are biologically meaningful molecules indicative of a distinct biological state or condition (cf. [94]). Statistically speaking, biomarkers are well-identified predictors that can be used to accurately predict relevant clinical outcomes, and also, they are an apt starting point for an evaluation of any statistical model. The aforementioned data frame has been re-analyzed by several researchers; cf. [16], [63], and [9]. Therefore, we refer to the aforementioned references for an extensive description of the data frame. Here, we only give a brief overview of sample acquisition, data preparation, measurement, and data processing.

FFPE lung tumor tissue samples for this study were provided by the bank of the National Center for Tumour Diseases (NCT, Heidelberg, Germany). Cylindrical tissue cores of non-small cell lung cancer were taken from 304 patients, where 168 patients were associated with primary lung adenocarcinoma (ADC), and 136 patients were associated with primary squamous cell carcinoma (SqCC). Cylindrical tissue cores of all tissue samples were collected in eight TMA blocks in total. Lung cancer is the leading reason for cancer-related deaths worldwide, with around 1.59 million reported deaths in 2012 (for more concrete numbers, see, e. g., [60] or [88]). Two primary lung cancer categories are determined, namely small cell lung cancer and non-small cell lung cancer (NSCLC), whence the latter constituted around 85% of all cases. The two most fatal histological NSCLC entities are ADC and SqCC, compromising approx. 50% and approx. 40% of all lung cancers, respectively. Differentiation of these two subtypes is critical for the choice of chemotherapy regimens and further test strategies.

Tissue sections were cut from all TMA blocks and treated in accordance with a previously published protocol for tryptic peptide imaging; cf. [20]. MALDI data were obtained through an Autoflex speed MALDI-TOF instrument (Bruker Daltonik) in positive ion reflector mode. Spectra were measured in the mass range 500-5000 m/z

at 150 μm spatial resolution using 1600 laser shots. Tumour status and typing for all cores were confirmed by standard histopathological examination; cf. [16]. Afterwards, the raw spectral data was loaded into SCiLS Lab (version 2016b, Bruker Daltonik), the standard baseline correction was performed (convolution method of 20), and total-ion-count (TIC) normalization was employed. The normalizing step is crucial in order to reduce the laboratory variation resulting from day-to-day instrument fluctuations or biological artifacts coming from sample preparation. Finally, spectral smoothing was performed to intervals of 0.4 Da (dalton) width (cf. [97]), and the remaining 4,669 spectra were pruned to the mass range of 800 – 2,500 (outside of this interval, m/z values were not considered), resulting in 1,699 m/z channels (columns).

In summary, we worked on a MALDI data set where all data-processing steps are based on standard protocols. It is out of the scope of this paper to compare different data-processing steps, like normalization, smoothing, etc.

2.4.2 Results of the data analysis

Figure 2.1 displays two mass spectra from both cancer subtypes, where the m/z values are illustrated on the horizontal axes, while the vertical axes refer to the relative abundances (intensities values) of ionizable molecules. These two graphs represent unique and specific spots within a patient's tissue and correspond to two mass spectra. Therefore, we have modeled each pixel marginally to identify which m/z values (based on 0.4 DA) are distinctive for a particular cancer subtype. We refer to [9] (see their Figure 1) for a more detailed illustration of the pipeline from a tissue to a single spectrum.

As discussed in [33, 32], the density of the empirical distribution of all Z-values does, in general, not coincide with the density of the standard normal distribution on $\mathbb{R}$, even if almost all p null hypotheses are true. The reason for this phenomenon, which can also be observed in our data (see Figure 2.2), is the presence of dependencies among the Z-statistics, as well as the presence of some extreme outliers, which presumably correspond to strong effect sizes. In particular, these effects lead to an inflation of the variance of the null distribution of the Z-statistics. However, we nevertheless have that the Z-statistics of the previously identified biomarkers lie in the tails of the distribution. Namely, their Z-statistics are large in absolute value and might be declared as statistically significant. Note that we consider an absolute value for the Z-values since we wish to find distinctive m/z values for either cancer subtype.

To disentangle the two sources for variance inflation (correlations among the Z-statistics and extreme outliers), we employed the method described in Section 5 (specifically, around Equations (53)-(55)) in [33]. This method suggested for our data a spread of approximately 2.63 in the central part (which presumably corresponds to true null hypotheses) of the empirical Z-score distribution. Thus, following the recommendation in [33] to "empirically correct" the null distribution, we divided all Z-scores by 2.63 prior to the following steps of data analysis. Notice that this re-scaling of the Z-scores is not an (essential) part of our proposed methodology but has been applied to

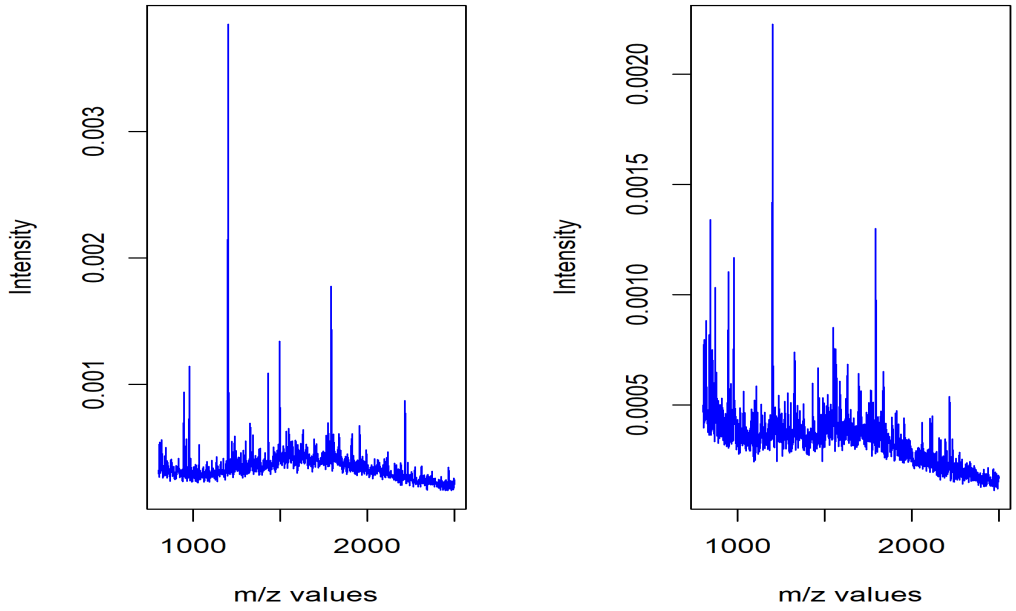


Figure 2.1: Two exemplary MALDI spectra. Two unique and specific spots within a tissue. Each of these spots represents a mass spectrum.

our real data for the sake of better comparability with the simulation results presented in Section 2.3. If we would omit the re-scaling, the p-values reported in Table 2.9 as well as the plausible rejection thresholds listed in Table 2.8 would both be given on a smaller scale. The test decisions, however, would stay the same for all j, because the re-scaling does not alter the ordering of the Z-scores.

The next step of the data analysis has been to determine an appropriate number k of common factors. To this end, we performed the proposed data analysis workflow described in Algorithm 1 over a range of different candidate values for k and compared the results. As documented in Figure 2.3, the estimated number of false discoveries as well as the estimated FDP are rather stable for $k \in \{6, \ldots, 9\}$. Based on this and for stability reasons, we chose $k = 6$ for our actual data analysis.

The main results of our real data analysis are illustrated in Figure 2.4. It is evident that $R(t)$, $\widehat{V}(t)$ and $\widehat{\mathrm{FDP}}(t)$ are increasing in the rejection threshold t. For $t \in [6 \times 10^{-4}, 8 \times 10^{-3}]$, the estimated FDP lies between approximately 6 % and approximately 23%. This indicates that most of the smallest p-values correspond to false nulls (leading to true discoveries). Table 2.8 lists the total number of rejections as well as the estimated FDP for several plausible choices of t.

The Logit-PFA method indicates, as highly significant, m/z values that are closely related to the five biomarkers identified by [60], for all considered thresholds t. These findings were confirmed by the dependency-adjustment method and also by the original Z-statistics with a fixed threshold value. Table 2.9 lists the 20 top-ranked (i. e., most

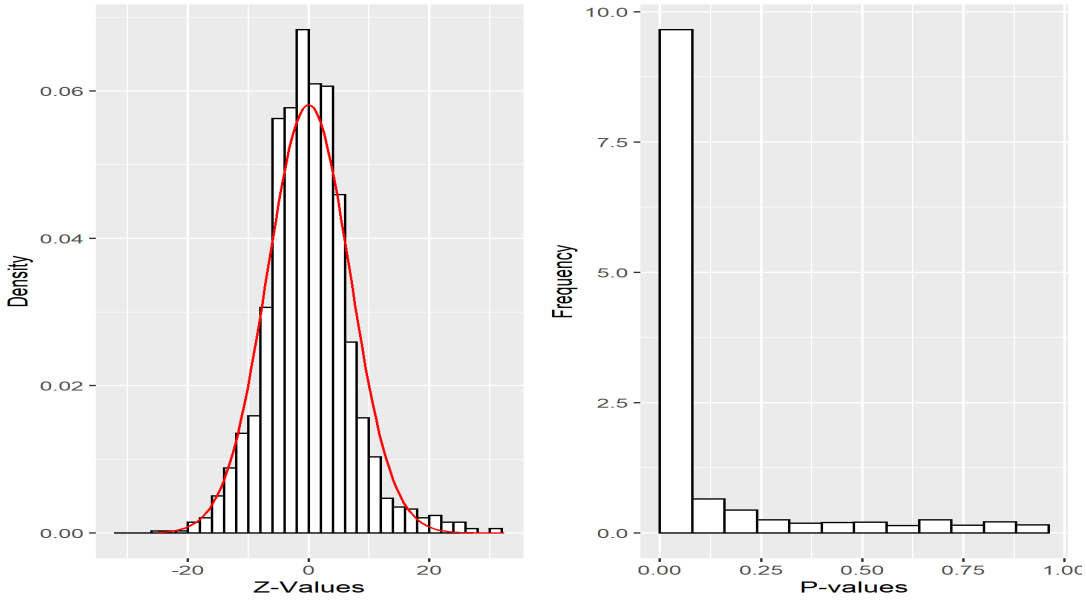


Figure 2.2: The empirical distribution and fitted normal density curve of the Z-values for the MALDI data. Due to dependencies among the Z-values, they are not following the theoretical $N(0,1)$ distribution. Instead, a closer look at the empirical distribution reveals that it can best be approximated by $N(-0.03455, 6.86^2)$. Consequently, the non-adjusted p-values have a lot of mass around zero.

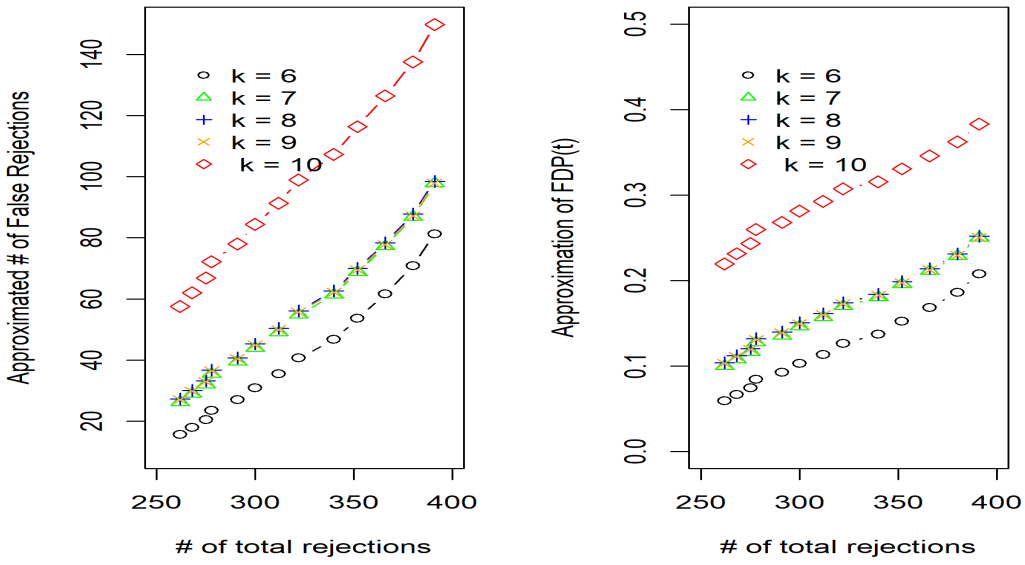


Figure 2.3: The approximated number of false discoveries as well as the approximated FDP as functions of the total number of rejections. Each curve corresponds to a different choice of the number k of common factors, where $k \in \{6, 7, 8, 9, 10\}$ has been considered.

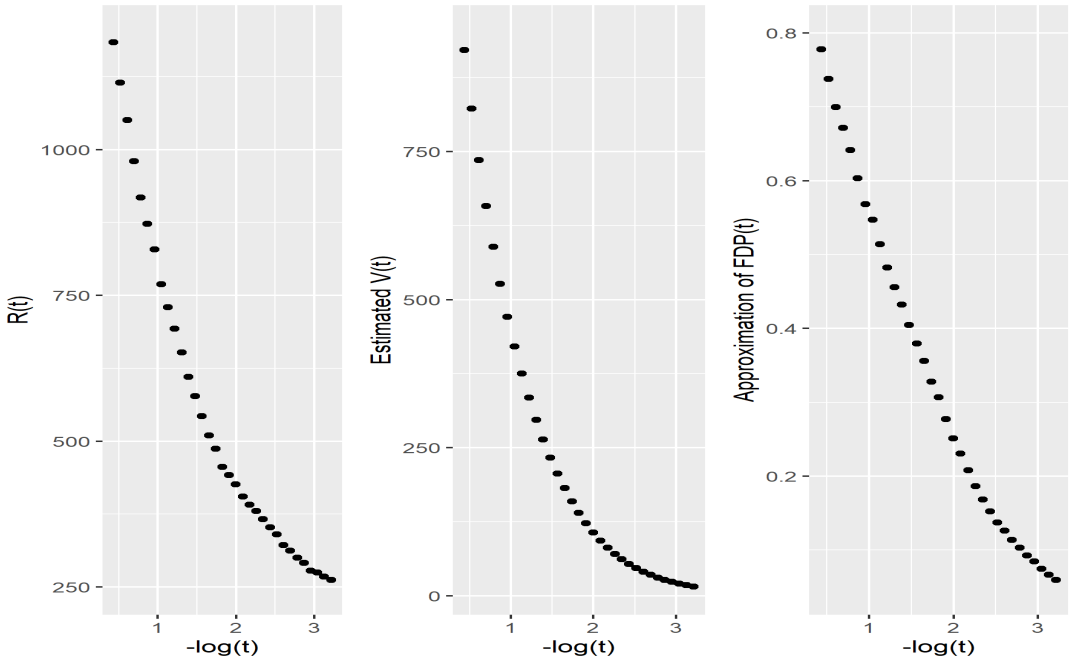


Figure 2.4: Main results: Total number of rejections, estimated number of false rejections, and estimated FDP, as functions of the threshold t. For the sake of presentation, the values of the t-axis are provided on a negative logarithmic (base 10) scale.

significant) null hypotheses (m/z values) for both cancer subtypes and thereby illustrates the overall significance of the five previously identified biomarkers which are indicated by stars in Table 2.9. For a comparison to the findings published previously in [60, 16, 63], we attribute the values m/z = 1410.7 Da and m/z = 1411.7 Da to the peak of a peptide of the CK5 protein and its second isotopic peak. In addition, the values m/z = 1877.9 Da and m/z = 1905.9 Da appearing in Table 2.9 are likely to be attributable to peptides of the proteins CK15 and HSP27. The m/z value = 1406.7 Da, indicating a negative direction, is likely associated to a peptide of the CK7 protein, which indicates to be efficient for the identification of ADC in the lung. The biomarker at m/z = 1821.9 can be attributed to a peptide CK15 protein distinctive for SqCC.

2.5 Discussion

From the statistical perspective, we have proposed an inferential framework for two-sample comparisons in high-dimensional settings when the test statistics have an arbitrary correlation structure. The major assumptions underlying the proposed methodology are (i) sparsity in the sense of Assumption 1, (ii) asymptotic normality of the vector of test statistics, and (iii) that the dependency structure among the test statistics can be described accurately by a factor model. To account for the high multiplicity of the considered applications, our criterion for type I error control is to keep the estimated value of the false discovery proportion at a given value of α. In the presence of strong dependencies among test statistics, the FDP is typically not well concentrated around its mean (the FDR),

and hence many authors have considered FDP control as the more appropriate criterion than FDR control under strong dependencies; see, e. g., ([13]) and the references therein. Under slightly different model assumptions than ours, in [38], the authors have provided conditions under which the approximated FDP value is close to the true value of the FDP with high probability if PFA is applied. Whenever such conditions are fulfilled, (approximate) FDP control can be achieved with the proposed methodology.

From the application perspective, we have applied the proposed method to a MALDI imaging data frame with a large number of covariates (m/z values). The results derived from the proposed method are consistent with already reported insights about this data frame. Reliable statistical modeling of MALDI data is a challenging task; cf., e. g., ([93]). Our approach based on MMM does not rely on heavy assumptions. Essentially, it is assumed that the (binary) phenotype of interest is associated with certain m/z-values and that this association can be described by a (marginal) logistic regression model for each m/z-value separately. These assumptions are well established in the statistical theory of modeling binary data; see, e. g., ([1]).

There are several possible directions for future research: First, it may be interesting to consider other supervised statistical learning models (for instance, neural networks with more than one layer) instead of the logistic regression model proposed in this work. Second, it is of interest to quantify the uncertainty about the realized FDP for different threshold values, with the goal of providing a confidence region for this realized FDP, in addition to a mere point estimate. Third, it is of interest to analyze the statistical properties of MALDI data in a more detailed manner, which may allow for joint modeling (after potential dimension reduction) instead of the MMM-based approach presented here. Finally, it will be worthwhile to consider categorical response variables with more than two categories.

Table 2.8: Number of rejections and estimated FDP for several plausible rejection thresholds.

Threshold t	$R(t)$	$\widehat{\text{FDP}}(t)$
8.23e-03	405	0.2304
3.70e-03	352	0.1524
2.48e-03	322	0.1263
1.36e-03	291	0.0926
9.12e-04	275	0.0746
6.11e-04	262	0.0596

Table 2.9: The top 20 ranked m/z values based on their re-scaled Z-scores for both cancer subtypes. Those m/z value that are presumably related to the five previously identified biomakers are indicated by the symbol * in both subtables.

(a) The most sign. m/z-values for ADC

m/z values	Z-scores	P-values
1407.7*	-9.52	$< 10^{-6}$
1406.7	-9.05	$< 10^{-6}$
1234.6	-7.76	$< 10^{-6}$
975.48	-7.58	$< 10^{-6}$
1813.9	-7.56	$< 10^{-6}$
1293.6	-7.12	$< 10^{-6}$
1476.7	-7.04	$< 10^{-6}$
1516.8	-7.03	$< 10^{-6}$
1277.6	-6.80	$< 10^{-6}$
2247.1	-6.80	$< 10^{-6}$
2246.1	-6.59	$< 10^{-6}$
1292.6	-6.46	$< 10^{-6}$
2248.1	-6.45	$< 10^{-6}$
1812.9	-6.35	$< 10^{-6}$
1838.9	-6.30	$< 10^{-6}$
1641.8	-6.06	$< 10^{-6}$
1814.9	-5.95	$< 10^{-6}$
1738.9	-5.93	$< 10^{-6}$
1705.8	-5.85	$< 10^{-6}$
1512.7	-5.79	$< 10^{-6}$

(b) The most sign. m/z-values for SqCC

m/z values	Z-scores	P-values
1410.7*	12.31	$< 10^{-6}$
1411.7	11.9	$< 10^{-6}$
865.43	11.44	$< 10^{-6}$
810.4	10.5	$< 10^{-6}$
878.43	10.15	$< 10^{-6}$
1877.9*	9.88	$< 10^{-6}$
1821.9*	9.73	$< 10^{-6}$
1878.9	9.58	$< 10^{-6}$
1439.7	9.37	$< 10^{-6}$
1822.9	9.24	$< 10^{-6}$
1437.7	8.98	$< 10^{-6}$
1412.7	8.96	$< 10^{-6}$
1879.9	8.92	$< 10^{-6}$
1438.7	8.42	$< 10^{-6}$
1425.7	8.42	$< 10^{-6}$
866.43	8.27	$< 10^{-6}$
1905.9*	8.19	$< 10^{-6}$
1823.9	8.13	$< 10^{-6}$
1906.9	8.07	$< 10^{-6}$
879.44	8.01	$< 10^{-6}$

Chapter 3

Multiple-multi sample testing under arbitrary covariance dependency

Information: This chapter is a slightly modified version of Vutov and Dickhaus (2023b) ([114]) published in *Statistics in Medicine*.

Authors

Vladimir Vutov, Institute for Statistics, University of Bremen, Bremen, Germany

Prof. Dr. Thorsten Dickhaus, Institute for Statistics, University of Bremen, Bremen, Germany

Declaration of individual contributions: Co-author and supervisor Prof. Dr. Thorsten Dickhaus has conceptualized the research project and proposed the usage of MMM. I have performed data modeling and analysis, as well as R programming. Both authors have written the manuscript together.

Data Availability Statement: R code, with which all results presented in the present manuscript can be reproduced, is provided in the data file, which is available as supplementary material for this article. Statistics in Medicine 42.17 (July 2023), vol. 42, issue 17, pp. 2944-2961. DOI: https://doi.org/10.1002/sim.9761.

Abstract: Modern high-throughput biomedical devices routinely produce data on a large scale, and the analysis of high-dimensional datasets has become commonplace in biomedical studies. However, given thousands or tens of thousands of measured variables in these datasets, extracting meaningful features poses a challenge. In this article, we propose a procedure to evaluate the strength of the associations between a nominal (categorical) response variable and multiple features simultaneously. Specifically, we propose a framework of large-scale multiple testing under arbitrary correlation dependency among test statistics. First, marginal multinomial regressions are performed for each feature individually. Second, we use an approach of multiple marginal models for each baseline-category pair to establish asymptotic joint normality of the stacked vector of the marginal multinomial regression coeffi-

cients. Third, we estimate the (limiting) covariance matrix between the estimated coefficients from all marginal models. Finally, our approach approximates the realized false discovery proportion of a thresholding procedure for the marginal p-values for each baseline-category logit pair. The proposed approach offers a sensible trade-off between the expected numbers of true and false findings. Furthermore, we demonstrate a practical application of the method on hyperspectral imaging data. This dataset is obtained by a matrix-assisted laser desorption/ionization (MALDI) instrument. MALDI demonstrates tremendous potential for clinical diagnosis, particularly for cancer research. In our application, the nominal response categories represent cancer (sub-)types.

3.1 Introduction

Many datasets in various scientific disciplines, such as neuroimaging, genomics, brain-computer interfacing, and others, are nowadays high-dimensional. Such datasets are frequently analyzed by means of large-scale multiple testing simultaneously for some phenotype – for example, cancer (sub-)types – on each feature among thousands of features. This approach has multiple applications in the life sciences (see, [26]). At least since the seminal publication by Benjamini and Hochberg ([11]), control of the false discovery rate (FDR) is a commonly used type I error criterion in high-dimensional multiple test problems. However, applying the well-known Benjamini-Hochberg (henceforth, BH) procedure (cf. [11]) or Storey's procedure (see [105, 106]), which have originally been designed for stochastically independent or weakly dependent p-values, respectively, can lead to an FDR inflation under certain forms of (strong) dependencies (see, among others, in [33, 32, 38, 62]). Under the assumption of positive regression dependency on subsets (PRDS), the BH procedure still controls the FDR (cf. [12]), but checking the validity of the PRDS assumption on the basis of a sample is far from trivial and only established for a limited number of special cases like, for example, elliptically distributed vectors of test statistics ([14]).

For these reasons and with the goal of optimizing the statistical power of the multiple test under the constraint of (at least asymptotic) FDR control at the desired level, multivariate (FDR-controlling) multiple tests have drawn a lot of attention over recent years (for more details, see [27]). A multivariate multiple test incorporates the dependence structure among test statistics or p-values, respectively, or an estimate thereof, explicitly in its decision rule. Large-scale multiple testing under general and strong dependency remains challenging and an active research topic in modern statistics ([38, 62, 108, 44]). Among other approaches like, for instance, considering block dependencies in genetics (cf. [102, 103, 50]), several multi-factor models have been proposed to model dependencies among test statistics (see in [38, 62, 44, 33, 32]). In particular, a general framework to approximate the false discovery proportion (FDP) has been introduced in [38] (cf. also [37]). It employs a so-called principal factor approximation (PFA) of the (estimated limiting) covariance matrix of test statistics, assuming that these test statistics are (at least asymptotically for large sample sizes) jointly normally distributed. The general strategy of the approach is to apply a spectral decomposition of the aforementioned covariance matrix and then to deduct its principal factors

that mostly induce the strong correlation dependency. A more recent procedure has been proposed in the context of FDR control under arbitrary (covariance) dependency (cf. [28]). This approach does not heavily rely on the assumption of the (asymptotic) normality of the test statistics and can be utilized even in the case of an unknown covariance matrix among the test statistics.

Biological datasets generated by advanced high-throughput devices typically contain thousands of measured variables, many of which are related. In this work, we consider a dataset obtained by a matrix-assisted laser desorption/ionization (MALDI) imaging mass spectrometry (IMS) tool, also known as MALDI imaging. MALDI has advanced considerably and demonstrates immense potential in numerous pathological applications (see in [59, 16]). MALDI-generated datasets contain thousands of measured variables (corresponding to molecular masses), many of which are (highly) related (cf. [63]). Since MALDI data do not generally exhibit dependencies in blocks, it has recently been advocated to model dependencies among MALDI features by means of factor models and to employ PFA in the context of multiple two-sample comparisons based on MALDI data (for more details, see [115]). In the present work, we extend our previous work by proposing a (multivariate) generalization to multi-sample comparisons, meaning that simultaneous inference with more than two nominal classes can be performed.

In this work, we deal with multiple multi-sample hypotheses testing under arbitrary correlation dependency, in particular in the case of more than two samples or categories, respectively. We do so under the scope of multiple marginal multinomial regression models and by applying the PFA methodology. In general, multinomial regression is a specific case of multivariate generalized linear models (GLMs), where (categorical) outcome variables are expressed as multi-dimensional vectors ([34]). As a result, our methodology performs multi-sample group testing simultaneously across these categorical groups. An essential step in our framework is the approximation of the joint null distributions among all marginal fits. In the present work, we exploit the framework of multiple marginal models (MMM) (cf. [83]). The classical setting of the latter is provided by modeling data with multiple outcome variables ([79]). Often, those studies consider procedures to control the family-wise error rate (FWER) with respect to type I errors (cf. [80, 89]). In our study, however, each "endpoint" (comparison) refers to a single independent (i.e., explanatory) variable, where the number p of independent variables under consideration is in the thousands. In the context of high-dimensional studies, it is well-known that FWER-controlling procedures are conservative and typically lead to substantially less statistical power than procedures that control the proportion of false discoveries (e.g., [29, 11]). A computationally efficient alternative to the MMM approach, based on score test statistics, has been proposed recently (see in [50, 91, 96]). The latter methodology needs only one fit of a GLM model. However, this methodology requires a data regime in which $p \ll n$, where n denotes the sample size, in order to approximate the (limiting) covariance matrix with sufficient accuracy. We call our method multinomial-principal factor approximation ("Multi-PFA").

The remainder of this article is structured as follows. In Section 3.2, we present the proposed multi-sample multiple testing approach. Section 3.3 is dedicated to a simulation study given different data-generating scenarios.

We demonstrate the practical application of the proposed approach to hyperspectral data in Section 3.4, and we conclude with a general discussion and outlook in Section 3.5.

3.2 Proposed Methodology

3.2.1 Data Structures

MALDI imaging data are typically stored in an $n \times p$ matrix $X = (x_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le p}}$, where the entry x_{ij} corresponds to an intensity value of the i-th mass spectrum at the j-th mass-to-charge ratio (m/z) value. Since MALDI data represent molecular masses of ionizable molecules, all entries of the data matrix X are non-negative (cf. [4]). However, our proposed framework makes no explicit assumption that data points have to be non-negative. Furthermore, our framework can also work with discrete covariates. We approach the biological challenge of analyzing the associations between (individual) m/z values (predictors $(X_j)_{1 \le j \le p}$) and several cancer (sub-)types (categorical response Y) by carrying out multiple statistical hypothesis tests simultaneously. Under the j-th null hypothesis H_{0j}, there is no association between X_j and Y, and we aim at testing the p null hypotheses $H_{01}, \ldots, H_{0p}$ (against their two-sided alternatives $H_{11}, \ldots, H_{1p}$) simultaneously based on one and the same dataset. Throughout the remainder, we assume that Y is a (random) nominal outcome variable, meaning that the response categories are unordered. Hence, for each $j \in \{1, \ldots, p\}$, the tuple (X_j, Y) takes its values in $\mathbb{R}_{\ge 0} \times \{1, 2, \ldots, q\}$, where we denote by q the number of (response) categories, and $\mathbb{R}_{\ge 0}$ denotes the set of non-negative real numbers.

3.2.2 Marginal Modelling for Categorical Responses

First, we model marginal associations for each predictor X_j individually. The reason for considering marginal modeling relies on the fact that we consider an inferential framework in which the number of covariates is either quite large or even (much) larger than the number of observational units ($n \ll p$). Thus, a model that includes all predictors in a single model cannot be fitted (reliably). However, this condition is not obligatory, and the proposed approach is also applicable to other data regimes. In the context of large-scale multiple (association) testing, marginal linear regression models have been adopted to test an individual hypothesis for each predictor (see, [7, 38]). Marginal GLMs, for positive outcome variables, have been considered in situations where the focus is on an inferential process around the mean of the response given a group of predictors ([8]). Also, marginal logistic regressions have been utilized to analyze child obesity in [82] with the usage of generalized estimating equations for jointly estimating (univariate) associations (for more details, see [64]).

For this section, let $j \in \{1, \ldots, p\}$ be arbitrary, but fixed. To model the conditional distribution $\mathbb{P}(Y|X_j)$, we assume a marginal multinomial regression with a canonical (logit) link function. In general, these response categories cannot be treated as a one-dimensional response, and we have to establish a dummy variable for each

category. We, therefore, have a multivariate outcome variable ([1]). One (response) category is used as a baseline, and the multinomial logistic model is established by pairing each (except one) nominal response category with the baseline. Throughout this article, the response category q is used as the baseline. Furthermore, we mainly consider the case of $q = 3$ throughout the remainder for concreteness and because of its relevance to the application that we will present in Section 3.4. For each (marginal) fit, we consider intercept terms α_{jc} and slope parameters β_{jc}, for $c \in \{1, \ldots, q-1\}$. We denote the observables (regressor and response variable) for the j-th marginal regression fit by $(X_j^{(i)}, Y^{(i)})_{1 \le i \le n}$, and we assume that these tuples are stochastically independent and identically distributed bivariate random vectors, each having the same (joint) distribution as (X_j, Y). For observational unit $i \in \{1, \ldots, n\}$, our model equation for the c-th baseline-category logit is given by

$$\log\left(\mathbb{P}(Y_c^{(i)} = 1|X_j^{(i)})/\mathbb{P}(Y_q^{(i)} = 1|X_j^{(i)})\right) = \log\left(\pi_{jc}^{(i)}/\pi_{jq}^{(i)}\right) = \alpha_{jc} + \beta_{jc} X_j^{(i)}, \tag{3.1}$$

where $Y_c^{(i)}$ denotes the dummy indicator for category $c \in \{1, \ldots, q-1\}$ pertaining to observational unit $i \in \{1, \ldots, n\}$, and $\pi_{jc}^{(i)} = \mathbb{P}(Y_c^{(i)} = 1|X_j^{(i)})$ by definition. The unknown model parameters $(\alpha_{jc}, \beta_{jc})_{1 \le c \le q-1}$ can be estimated by the maximum (log-) likelihood (ML) principle. To this end, we follow the derivations in Section 7.1.4 of the book by [1]. Given that $\pi_{jq}^{(i)} = 1 - \pi_{j1}^{(i)} - \pi_{j2}^{(i)} - \ldots - \pi_{j,q-1}^{(i)}$ as well as $Y_q^{(i)} = 1 - Y_1^{(i)} - Y_2^{(i)} - \ldots - Y_{q-1}^{(i)}$, the log-likelihood contribution for the i-th observational unit is under the model specified in (3.1) given by

$$\begin{aligned}\log\left(\prod_{c=1}^{q}\left\{\pi_{jc}^{(i)}\right\}^{Y_c^{(i)}}\right) &= \sum_{c=1}^{q-1} Y_c^{(i)} \log \pi_{jc}^{(i)} + \left(1 - \sum_{c=1}^{q-1} Y_c^{(i)}\right) \log\left[1 - \sum_{c=1}^{q-1} \pi_{jc}^{(i)}\right] \\ &= \sum_{c=1}^{q-1} Y_c^{(i)} \log \frac{\pi_{jc}^{(i)}}{1 - \sum_{c=1}^{q-1} \pi_{jc}^{(i)}} + \log\left[1 - \sum_{c=1}^{q-1} \pi_{jc}^{(i)}\right].\end{aligned} \tag{3.2}$$

Substituting $\log(\pi_{jc}^{(i)}/\pi_{jq}^{(i)}) = \alpha_{jc} + X_{jc}^{(i)}\beta_{jc}$ as well as $\pi_{jq}^{(i)} = 1/\{1 + \sum_{c=1}^{q-1} \exp(\alpha_{jc} + X_{jc}^{(i)}\beta_{jc})\}$ in (3.2), we find that

$$\ell\left(\hat{\alpha}_{jc}, \hat{\beta}_{jc}|Y^{(i)}, X_j^{(i)}\right) = \max_{\alpha_{jc}, \beta_{jc}} \sum_{i=1}^{n} \left\{ \sum_{c=1}^{q-1} Y_c^{(i)}(\alpha_{jc} + X_j^{(i)}\beta_{jc}) - \log\left[1 + \sum_{c=1}^{q-1} \exp(\alpha_{jc} + X_j^{(i)}\beta_{jc})\right]\right\}, \tag{3.3}$$

where ℓ stands for the log-likelihood function, and $\hat{\alpha}_{jc}$ and $\hat{\beta}_{jc}$ denote the ML estimators. In practice, the estimation is carried out conditionally to the actual observed values of the response indicators and the regressors, leading to the numerical values of the ML estimates.

In this work, we are interested in testing simultaneously two families of hypothesis-alternative pairs, which are given by

$$H_{0jc} : \beta_{jc} = 0 \text{ versus } H_{1jc} : \beta_{jc} \neq 0, \quad j = 1, \ldots, p, \ c \in \{1, 2\}, \tag{3.4}$$

for $q = 3$. By means of testing the hypotheses in (3.4), we make binary decisions (rejection or non-rejection)

for both baseline-category sets, consisting of p null hypotheses each, based on the data at hand. In the MALDI context, this means that we aim at identifying the most distinctive m/z values for the cancer association, where three different cancer types are considered in our application.

3.2.3 Multiple Marginal Models

The following (second) step of the proposed framework is to combine all marginal fits and to approximate the joint null distribution of all ML estimators for each baseline-category pair. To do so, we exploit the approach for jointly estimating the parameters of multiple marginal models developed in [83] and apply this approach to the marginal models explained in the previous section. Notice that our modeling approach implies that the regression parameters are unique to one model j and are not shared by any two models $j_1 \neq j_2$. In addition to that, the intercepts $(\alpha_{jc})_{1\leq j\leq p,\, c\in\{1,2\}}$ are nuisance parameters in the sense that the hypotheses in (3.4) solely concern the β_{jc}'s. However, the intercepts $(\alpha_{jc})_{1\leq j\leq p,\, c\in\{1,2\}}$ contribute to the estimation and the standardization of β_{jc}'s.

The methodology in [83] yields a central limit theorem for the two p-dimensional random vectors $\hat{\beta}_c = (\hat{\beta}_{1c},\ldots,\hat{\beta}_{pc})^\top$, $c \in \{1,2\}$. This is accomplished by stacking the standardized score contributions of the $\hat{\beta}_{jc}$'s across all p marginal models, for each baseline-category pair. To this end, it is used that, under standard regularity assumptions (like finite variances and non-vanishing limiting relative category frequencies), each ML estimator $\hat{\beta}_{jc}^{(i)}$ admits the asymptotic (i. e., $n \to \infty$) representation (see [83])

$$(\hat{\beta}_{jc} - \beta_{jc})\,\sqrt{n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_{ijc} + o_{\mathbb{P}}(1), \tag{3.5}$$

where $\Psi_{ijc} = \left(F_j^{(i)}\right)^{-1} \widetilde{\Psi}_{ijc}$, $\left(F_j^{(i)}\right)^{-1}$ is the relevant row of the inverse Fisher information matrix of the model specified in (3.1) that corresponds to coordinate j for the i-th observational unit, $\widetilde{\Psi}_{ijc}$ is the score function pertaining to coordinate j and category c for the i-th observational unit, and $o_{\mathbb{P}}(1)$ indicates a sequence of random variables converging to zero in probability.

Elementary calculations yield that $(\Psi_{ij1}, \Psi_{ij2})^\top$ is given by the second and the fourth coordinate of the four-variate vector

$$\left(\left(\left(\pi_{jc}^{(i)}\left[\mathbf{I}(c=c') - \pi_{jc'}^{(i)}\right](1, X_j^{(i)})^\top (1, X_j^{(i)})\right)_{c,c'\in\{1,2\}}\right)^{-1} \left(Y_1^{(i)} - \pi_{j1}^{(i)}, Y_2^{(i)} - \pi_{j2}^{(i)}\right)^\top\right) \otimes (1, X_j^{(i)})^\top, \tag{3.6}$$

where

$$\mathbf{I}(c=c') = \begin{cases} 1, & \text{if } c = c', \\ 0, & \text{if } c \neq c', \end{cases}$$

$$\pi_{jc}^{(i)} = \frac{\exp\left(\alpha_{jc} + X_j^{(i)}\beta_{jc}\right)}{1 + \sum_{h=1}^{2} \exp\left(\alpha_{jh} + X_j^{(i)}\beta_{jh}\right)},$$

and $\otimes$ denotes the Kronecker product. For more (computational) details regarding multinomial regression, we refer to previous literature on the topic (see [15, 52]).

Now, let $\beta_c := (\beta_{1c}, \ldots, \beta_{pc})^\top, \hat{\beta}_c := (\hat{\beta}_{1c}, \ldots, \hat{\beta}_{pc})^\top$, and $\Psi_{ic} := (\Psi_{i1c}, \ldots, \Psi_{ipc})^\top$ for $c \in \{1, 2\}$ and $i \in \{1, \ldots, n\}$. Then, we can conclude from (3.5) the asymptotic expansion $(\hat{\beta}_c - \beta_c)\sqrt{n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_{ic} + o_{\mathbb{P}}(1)$, where now $o_{\mathbb{P}}(1)$ indicates a sequence of p-dimensional random vectors converging to the zero vector in probability. Since the observational units are assumed to be stochastically independent and identically distributed, the multivariate central limit theorem yields convergence in distribution (indicated by the symbol $\xrightarrow{d}$) to a centered p-variate normal distribution, i. e., that $(\hat{\beta}_c - \beta_c)\sqrt{n} \xrightarrow{d} N_p(0, \Sigma_c)$ for $c \in \{1, 2\}$. For each $c \in \{1, 2\}$, the limiting variance-covariance matrix Σ_c can be estimated consistently by $\widehat{\Sigma}_c = \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_{ic}^\top \hat{\Psi}_{ic}$ where each $\hat{\Psi}_{ic}$ is obtained by inserting the parameter estimates $(\hat{\alpha}_{jc}, \hat{\beta}_{jc})_{1 \le j \le p,\, c \in \{1,2\}}$ from the p marginal fits (cf. (3.3)).

For each $c \in \{1, 2\}$, let $Z_{1c}, \ldots, Z_{pc}$ be the standardised versions of $\hat{\beta}_{1c}, \ldots \hat{\beta}_{pc}$, which are given by

$$Z_{jc} = \frac{\hat{\beta}_{jc}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{jc})}}, \quad j = 1, \ldots, p, \quad c \in \{1, 2\}, \tag{3.7}$$

where $\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{jc})}$ is the square root of the j-th diagonal element of $\hat{\Sigma}_c$, divided by $\sqrt{n}$. We have that

$$(Z_{1c}, Z_{2c}, \ldots, Z_{pc})^\top \underset{\text{approx.}}{\sim} N_p((\mu_{1c}, \mu_{2c}, \ldots, \mu_{pc})^\top, \hat{\Sigma}_c^*), \tag{3.8}$$

where $\mu_{jc} = \beta_{jc} / \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{jc})}$ for $1 \le j \le p$ and $c \in \{1, 2\}$, $\hat{\Sigma}_c^* = \mathrm{diag}[\hat{\Sigma}_c]^{-1/2} \hat{\Sigma}_c \, \mathrm{diag}[\hat{\Sigma}_c]^{-1/2}$ denotes the correlation matrix pertaining to $\hat{\Sigma}_c$, and the notation $\underset{\text{approx.}}{\sim}$ indicates the approximate distribution for large n. Therefore, since $\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{jc})} > 0$ holds true with probability one under our regularity assumptions, the families of hypotheses from (3.4) can be equivalently written as

$$H_{0jc} : \mu_{jc} = 0 \;\text{ versus }\; H_{1jc} : \mu_{jc} \neq 0, \quad j = 1, \ldots, p, \; c \in \{1, 2\}. \tag{3.9}$$

For the sake of presentation, we use henceforth without loss of generality $\hat{\Sigma}_c^*$ and $Z_{1c}, \ldots, Z_{pc}$ only for the first baseline-category (i. e., for $c = 1$) and omit the index c, for convenience of notation. However, we consider the same inference procedure simultaneously for both baseline-category pairs. In order to address the fact that two (or, in general, $q - 1$) families of hypotheses are considered together, one may adjust the significance level α occurring in Section 3.2.5 below accordingly.

3.2.4 False Discovery Proportion

Considering the multiple test problem given by (3.9) (for $c = 1$), let $p_0 = \#\{j : \mu_j = 0\}$ denote the number of true null hypotheses, and $p_1 = p - p_0 = \#\{j : \mu_j \neq 0\}$ the number of false null hypotheses. Throughout the remainder, we make the following sparsity assumption.

Assumption 2 *The number p_1 of false null hypotheses is very small in comparison to the number p of all null hypotheses.*

In an asymptotic context, where $p = p(n)$ tends to infinity as n tends to infinity, Assumption 2 can be formalized as $p_0(n)/p(n) \to 1$. For controlling type I errors, we employ empirical process techniques in the spirit of Storey's method (cf. [105]). In particular, we consider thresholding rules which reject a null hypothesis H_{0j} if and only if a corresponding p-value p_j is smaller than or equal to a (data-dependent) threshold t. Such types of multiple tests have been considered widely in previous literature (e.g., [38, 37, 105]) Conceptually, the goal of the proposed method is to approximate the proportion of false discoveries among all rejections (commonly referred to as the false discovery proportion, FDP for short) for a fixed threshold t under an arbitrary correlation matrix Σ^*. To this end, we consider the three empirical processes

$$V(t) = \#\{\text{true null } P_j : P_j \leq t\},$$
$$S(t) = \#\{\text{false null } P_j : P_j \leq t\},$$
$$R(t) = \#\{P_j : P_j \leq t\},$$

where t ranges in $[0, 1]$ and the notation P_j is used to indicate that we consider the j-th p-value as a random variable here. For fixed $t \in [0, 1]$, the quantities $V(t)$, $S(t)$, and $R(t)$ represent the (random) number of false discoveries (i. e., false rejections or, synonymously, type I errors), the (random) number of true discoveries, and the (random) total number of discoveries, respectively, hence $R(t) = V(t) + S(t)$. The FDP is for a fixed t given by $\mathrm{FDP}(t) = V(t)/\max\{R(t), 1\}$, where the maximum is taken to avoid an expression of the form $0/0$. All of the latter random variables depend on the Z-statistics $Z_1, Z_2, \ldots, Z_p$, because each (random) p-value P_j will be obtained by a transformation of Z_j, for $1 \leq j \leq p$, see below. Notice also that $V(t)$ and $S(t)$ are both unobservable, whereas $R(t)$ is observable.

3.2.5 Principal Factor Approximation

The next step of our proposed framework is to incorporate the correlation effects of the Z-statistics in an approximation of $\mathrm{FDP}(t)$ for a fixed threshold t. The procedure relies on the identification of a low-dimensional linear space of random vectors, which captures most of the dependence structure of the Z-statistics.

To this end, we carry out the spectral decomposition of the correlation matrix Σ^*. We let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ denote the eigenvalues of Σ^*, and $\gamma_1, \ldots, \gamma_p$ the corresponding eigenvectors. Hence, $\hat{\Sigma}^*$ is represented by its eigenvalue-eigenvector pairs $(\lambda_j, \gamma_j)_{1 \leq j \leq p}$. For a fixed number k of common factors, where $1 \leq k \ll p$, the factor approximation of Σ^* is given by

$$\hat{\Sigma}^* = L_k L_k^\top + A_k, \tag{3.10}$$

where $A_k = \sum_{j=k+1}^{p} \lambda_j \gamma_j \gamma_j^\top$ and $L_k = (\sqrt{\lambda_1}\gamma_1, \sqrt{\lambda_2}\gamma_2, \ldots, \sqrt{\lambda_k}\gamma_k)$. If k is chosen appropriately, the $p \times k$- matrix L_k, corresponding to k latent variables, can describe most of the dependence structure among the Z-statistics. Accordingly, $Z_1, \ldots, Z_p$ can be decomposed as

$$Z_j = \mu_j + b_j^\top W + K_j = \mu_j + \eta_j + K_j, \quad j = 1, \ldots, p, \tag{3.11}$$

where $b_j = (b_{j1}, \ldots, b_{jk})^\top$ and $(b_{1j}, \ldots, b_{pj})^\top = \sqrt{\lambda_j}\gamma_j$. The vector $W = (W_1, \ldots, W_k)^\top \sim N_k(0, I_k)$ is the vector of (latent) common factors. Due to orthogonality projection, these common factors are jointly stochastically independent. The random vector $(K_1, \cdots, K_p)^\top \sim N_p(0, A_k)$ is called the vector of random errors, and it is assumed that factors and random errors are stochastically independent. In our previous work ([115]), we elaborated more on the importance and the choice of k.

On the basis of the above considerations, we consider the FDP estimator from Proposition 2 in [38], which is for a given $t \in [0, 1]$ defined as

$$\widehat{\mathrm{FDP}}(t) = \min\left\{ \sum_{j=1}^{p} \left[\Phi(a_j(z_{t/2} + \widehat{\eta}_j) + \Phi(a_j(z_{t/2} - \widehat{\eta}_j)) \right], R(t) \right\} / R(t) \tag{3.12}$$

if $R(t) > 0$, and $\widehat{\mathrm{FDP}}(t) = 0$ if $R(t) = 0$. In (3.12), $a_j = (1 - \sum_{h=1}^{k} b_{jh}^2)^{-1/2}$ and $R(t) = \{j : P_j \leq t\}$, where Φ and $z_{t/2} = \Phi^{-1}(t/2)$ are the cumulative distribution function (cdf) and the lower $t/2$-quantile of the standard normal distribution on $\mathbb{R}$, respectively. In this, the dependency-unadjusted, two-sided and random p-value corresponding to Z_j is given by $P_j = 2\Phi(-|Z_j|) = 2(1 - \Phi(|Z_j|))$, and $\hat{\eta}_j = \sum_{h=1}^{k} b_{jh}\widehat{W}_j$ is a linear estimator for $\eta_j = b_j^\top W$.

The FDP estimator given in (3.11) is based on a sparsity assumption (for more details, see [38, 37]). Specifically, it is assumed that the numbers p and p_0 are large, while the number p_1 of false nulls is relatively low. This assumption allows for the summation of all hypotheses (not only over "true nulls") in the numerator of (3.11). In the context of MALDI modeling, this sparsity assumption is justified for three reasons: First, several researchers who have analyzed MALDI datasets have reported a low number of m/z values that are highly associative for cancer (sub-)type (cf. [9, 16]). Second, to the best of our knowledge, the reported number of biologically meaningful molecules, so-called biomarkers, in MALDI-related studies is low – from three to five per study (see, [60, 63]). Third, there is evidence for considerable noise in the data ([63, 16]).

To evaluate (3.12) in a practical application, one needs to determine the (linear) estimator $\widehat{W} = (\widehat{W}_1, \ldots, \widehat{W}_k)^\top$ of the common factors. The authors have proposed to estimate $\widehat{W}$ via a linear regression (L_2-regression) or by using a quantile regression (L_1-regression) (cf. [38]). Regarding the L_2-estimator, it has been proposed to carry out the estimation of $(\widehat{W}_1, \ldots, \widehat{W}_k)^\top$ only on the basis of the subset of length $0.95p$ of the smallest (in absolute values) Z-statistics. This leads to the L_2-estimator

$$\hat{w} = (\hat{w}_1, \ldots, \hat{w}_k)^\top = \arg\min_{W \in \mathbb{R}^k} \sum_{j=1}^{\lfloor 0.95p \rfloor} (Z_j - b_j^\top W)^2, \tag{3.13}$$

where the Z-values in (3.13) are sorted in ascending order based on their absolute values. We have utilized this estimator in our simulation study (see Section 3.3). The estimator based on the L_1-regression is given by

$$\hat{w} = (\hat{w}_1, \ldots, \hat{w}_k)^\top = \arg\min_{W \in \mathbb{R}^k} \sum_{j=1}^{p} |Z_j - b_j^\top W|. \tag{3.14}$$

For the analysis of the MALDI data (Section 3.4), we have used the L_1-regression rather than the L_2-regression because it is more robust to highly untypical observations (i. e., outliers).

Finally, the dependency-adjusted p-values related to $Z_1, \ldots, Z_p$ are given by

$$\tilde{P}_j = 2\Phi(-|a_j(Z_j - b_j^\top \widehat{W})|). \tag{3.15}$$

For a given threshold $t \in [0, 1]$, the null hypothesis H_{0j} from (3.4) gets rejected based on the data at hand, if and only if $\tilde{p}_j \leq t$, $1 \leq j \leq p$. For the purpose of FDP control, t can be selected as the largest value $t = t_\alpha \in [0, 1]$ fulfilling that $\widehat{\mathrm{FDP}}(t_\alpha)$ is not exceeding a pre-determined level $\alpha \in (0, 1)$, for example, $\alpha = 10\%$. In practice, one can carry out a grid search over a set of candidate values of t in order to find the value t_α for a given α (for more details, see in [115]).

3.3 Simulation Study

3.3.1 Simulation Setup

We carried out a simulation study to assess the performance of our proposed methodology under different data-generating schemes. We considered the parameter settings $n = 500$, $p \in \{500, 1000\}$, and $p_1 = 10$ (for each baseline-category pair). For each combination of the aforementioned parameter values, we performed $1,000$ Monte Carlo repetitions. Without loss of generality, we set $\beta_{j1} \neq 0$ as well as $\beta_{j2} \neq 0$ for $j \in \{1, \ldots, p_1\}$, and we call these first p_1 coordinates "active". The remaining p_0 coordinates have been set "inactive", meaning that $\beta_{jc} = 0$ for all in $j \in \{p_1 + 1, \ldots, p\}$ and $c \in \{1, 2\}$. Furthermore, we set all intercept terms $(\alpha_{jc})_{1 \leq j \leq p,\, c \in \{1,2\}}$ to zero in all simulations.

We estimated the common factors $\{W_h : h = 1, \ldots, k\}$ by applying the least-squares estimator as given in (3.13). The utilized number k of common factors is reported in each caption of Tables 3.1 - 3.4 below.

For each observational unit $i \in \{1, \ldots, n\}$, the simulation data have been generated according to the model

$$\mathbb{P}_\beta(Y_1^{(i)} = 1|X^{(i)}) = \frac{\exp\left(\sum_{j=1}^{p_1} \beta_{j1} X_j^{(i)}\right)}{1 + \exp\left(\sum_{j=1}^{p_1} \beta_{j1} X_j^{(i)}\right) + \exp\left(\sum_{j=1}^{p_1} \beta_{j2} X_j^{(i)}\right)},$$

$$\mathbb{P}_\beta(Y_2^{(i)} = 1|X^{(i)}) = \frac{\exp\left(\sum_{j=1}^{p_1} \beta_{j2} X_j^{(i)}\right)}{1 + \exp\left(\sum_{j=1}^{p_1} \beta_{j1} X_j^{(i)}\right) + \exp\left(\sum_{j=1}^{p_1} \beta_{j2} X_j^{(i)}\right)},$$

$$\mathbb{P}_\beta(Y_3^{(i)} = 1|X^{(i)}) = \frac{1}{1 + \exp\left(\sum_{j=1}^{p_1} \beta_{j1} X_j^{(i)}\right) + \exp\left(\sum_{j=1}^{p_1} \beta_{j2} X_j^{(i)}\right)}.$$

The considered dependency structures for $X_1, \ldots, X_p$ are given in Model 2.

Model 2 *Scenario 1:* $X_1, \ldots, X_p$ *are stochastically independent and identically standard normally distributed random variables.*
Scenario 2: $X_{p_1+1}, \ldots, X_p$ *are jointly normally distributed on* $\mathbb{R}^{p_0}$*, namely, according to* $N_{p_0}(0, \Sigma^*)$*, where* Σ^* *is an equi-correlation matrix with diagonal elements equal to one, and all off-diagonal elements equal to the same given value* ρ*. Furthermore, the subvector* $(X_{p_1+1}, \ldots, X_p)^\top$ *is stochastically independent of the subvector* $(X_1, \ldots, X_{p_1})^\top$*. This is in order to avoid spurious effects of covariates* X_j *with* $p_1 + 1 \leq j \leq p$ *on the response, which would arise from confounding if the two aforementioned subvectors would be dependent. The random variables* $X_1, \ldots, X_{p_1}$ *are stochastically independent and identically* $N(0, 1)$*-distributed. The equi-correlation model is a one-factor model with a dominating first eigenvalue; see, e.g., Example 2.1 in [43].*

This simulation study has been run in *R version 4.1.1* ([87]) employing the R function *rmvnorm()* in [46] to simulate correlation dependency among the independent variables. In addition, we have employed the function *vglm()* in [109] for the estimation of the marginal multinomial estimates. The function *pfa.test()* in [36] has been used for applying the PFA method.

3.3.2 Simulation Results

We summarize the results from our computer simulations in Tables 3.1 - 3.4 in terms of $\widehat{\text{FDP}}(t)$, $R(t)$, and $S(t)$ for a fixed threshold $t = 10^{-4}$. The estimates for $\widehat{\text{FDP}}(t)$ have been derived by applying Equation (3.12), and the first columns of each table correspond to the respective baseline-category pair. In addition, we report the median value of t_α for the conventional choice of $\alpha = 0.05$ as well as the average (over the 1,000 Monte Carlo repetitions) of $S(t_\alpha)$. Due to the choice of $p_1 = 10$, the largest possible value of $S(t)$ equals ten for any choice of the threshold t. The purpose of the summary tables is to illustrate that the proposed framework can produce meaningful results under different strengths of dependency among the test statistics. In this sense, the simulations serve as a proof of concept for the Multi-PFA approach.

Tables 3.1 and 3.2 contain our simulation results under Scenario 1 from Model 2. Because the test statistics

are jointly independent (i.e., joint independence among all X_j), $t_{0.05}$ is rather small because the "effective number of tests" (in the sense of Section 3.4 in [27]) equals p under joint independence of all test statistics or p-values, respectively. This means that a rather strong multiplicity correction is necessary. However, the standard error of $\widehat{\text{FDP}}(t)$ is also quite small under Scenario 1, since the FDP is well concentrated around its expectation (the FDR) under joint independence across among all test statistics or p-values, respectively. With regard to type II errors, the reported average values of $S(t_{0.05})$ demonstrate that, on average, $7-8$ out of 10 active coordinates could be identified by our proposed multiple test procedure.

Simulation results under Scenario 2 (Gaussian equi-correlation model) are summarised in Tables 3.3 and 3.4. Here, the effective number of tests is smaller than p whenever $\rho \geq 0$, and it decreases as ρ increases. Consequently, $t_{0.05}$ grows with ρ, too. In turn, this also leads to an improved power of the proposed multiple test, which is reflected by the numbers reported for $S(t)$, which are under Scenario 2 on average larger than the corresponding values under Scenario 1.

Table 3.1: **Simulation results under Scenario 1 (I)**

c	Median of $\widehat{\text{FDP}}(t)$	Std. Error of $\widehat{\text{FDP}}(t)$	Mean of $R(t)$	Std. Error of $R(t)$	Mean of $S(t)$	Std. Error of $S(t)$	Median of $t_{0.05}$	Mean $S(t_{0.05})$
1	0.00433	0.00147	5.80	1.39	5.75	1.36	1.60e-03	8.37
2	0.00442	0.00134	5.78	1.36	5.74	1.35	1.60e-03	8.31

The total number of hypotheses equals $p = 500$; the number of false null hypotheses equals $p_1 = 10$ per baseline-category pair (non-zero regression coefficients equal 1); the number of factors equals $k = 10$; the rejection threshold equals $t = 10^{-4}$, apart from the last two columns.

Table 3.2: **Simulation results under Scenario 1 (II)**

c	Median of $\widehat{\text{FDP}}(t)$	Std. Error of $\widehat{\text{FDP}}(t)$	Mean of $R(t)$	Std. Error of $R(t)$	Mean of $S(t)$	Std. Error of $S(t)$	Median of $t_{0.05}$	Mean $S(t_{0.05})$
1	0.00946	0.00325	5.95	1.35	5.86	1.33	5.52e-04	7.50
2	0.00943	0.00360	5.83	1.43	5.76	1.40	5.52e-04	7.46

The total number of hypotheses equals $p = 1000$; the number of false null hypotheses equals $p_1 = 10$ per baseline-category pair (non-zero regression coefficients equal 1); the number of factors equals $k = 10$; the rejection threshold equals $t = 10^{-4}$, apart from the last two columns.

Table 3.3: **Simulation results under Scenario 2 (I)**

c	ρ	Median of $\widehat{\text{FDP}}(t)$	Std. Error of $\widehat{\text{FDP}}(t)$	Mean of $R(t)$	Std. Error of $R(t)$	Mean of $S(t)$	Std. Error of $S(t)$	Median of $t_{0.05}$	Mean $S(t_{0.05})$
1	0.2	0.00287	0.0199	5.83	1.36	5.81	1.31	2.33e-03	8.63
2	0.2	0.00273	0.00995	5.90	1.33	5.89	1.31	2.33e-03	8.72
1	0.5	0.000274	0.0475	5.84	1.48	5.81	1.31	6.99e-03	9.28
2	0.5	0.000265	0.0168	5.88	1.33	5.89	1.31	6.99e-03	9.31
1	0.8	0.000171	0.0455	5.87	2.46	5.81	1.31	2e-02	9.78
2	0.8	0.000168	0.0141	5.86	1.31	5.90	1.32	2e-02	9.77

The total number of hypotheses equals $p = 500$; the number of false null hypotheses equals $p_1 = 10$ per baseline-category pair (non-zero regression coefficients equal 1); the number of factors equals $k = 1$; the rejection threshold equals $t = 10^{-4}$, apart from the last two columns.

Table 3.4: **Simulation results under Scenario 2 (II)**

c	ρ	Median of $\widehat{\text{FDP}}(t)$	Std. Error of $\widehat{\text{FDP}}(t)$	Mean of $R(t)$	Std. Error of $R(t)$	Mean of $S(t)$	Std. Error of $S(t)$	Median of $t_{0.05}$	Mean $S(t_{0.05})$
1	0.2	0.00497	0.0208	5.86	1.36	5.81	1.32	1.01e-03	8.00
2	0.2	0.00505	0.0265	5.84	1.38	5.79	1.32	1.01e-03	7.98
1	0.5	0.000299	0.0316	5.84	1.37	5.81	1.32	5.06e-03	9.08
2	0.5	0.000315	0.0496	5.84	1.63	5.79	1.32	5.06e-03	9.07
1	0.8	0.000168	0.0122	5.81	1.33	5.81	1.32	2.9e-02	9.73
2	0.8	0.000169	0.0402	5.84	2.69	5.79	1.32	2.9e-02	9.75

The total number of hypotheses equals $p = 1000$; the number of false null hypotheses equals $p_1 = 10$ per baseline-category pair (non-zero regression coefficients equal 1); the number of factors equals $k = 1$; the rejection threshold equals $t = 10^{-4}$, apart from the last two columns.

3.4 Real Data Application: MALDI Imaging Data

3.4.1 Description of the dataset

We applied our proposed inferential procedure to a hyperspectral dataset obtained from a MALDI imaging instrument. For a detailed description of this dataset in terms of its acquisition protocols, tissue sections, tissue blocks, etc. (see, [9, 16]). In the aforementioned studies, the researchers analyzed this dataset by merging two lung cancer

subtypes into a single category. By doing so, the authors worked with two cancer types based on two distinctive human organs (the so-called "LP task", i.e., lung vs. pancreas). In our study, however, we model this dataset with a three-class model, using pancreatic adenocarcinoma as a baseline. The resulting goal of the statistical analysis is to simultaneously perform two (sub-)tasks, namely the pancreatic vs. lung adenocarcinoma and the pancreatic vs. lung squamous cell carcinoma comparison. As mentioned in the introduction, the high-throughput device that motivated this study is MALDI IMS. In general, this technology provides molecular information about a given analyte (for example, tissue) in a spatial manner. More concretely, the MALDI IMS tool measures mass spectra at multiple discrete spatial positions and yields an image for each unique spot within a provided tissue. For an illustration, Figure 3.1 displays instances of outputs from a MALDI experiment from three spatial spots. In Figure 3.1, the m/z values are plotted on the horizontal axes, and the relative abundances (intensities values) of ionizable molecules are plotted on the vertical axes. This spatial molecular "biochemical information may then be used for the determination of the cancer subtypes or the identification of the origin of the primary tumor in patients" ([9]). Each spot is called a mass spectrum ([6, 4]) and it depicts relative abundances of ionizable molecules with numerous mass-to-charge (m/z) values – "ranging [..] from several hundred up to a few tens of thousands of m/z" ([4]) – whereas a single m/z value in the context of MALDI is interpreted as a molecular mass. The full pipeline of a MALDI experiment, from a tissue spot to data analysis, is thoroughly illustrated in Figure 1 in [16] and Figure C2 in [63]. In the dataset analyzed here, formalin-fixed paraffin-embedded (FFPE) (for more details regarding FFPE see in [76]) tumor samples from 445 patients were provided by the tissue bank of the National Center for Tumor Diseases (NCT) Heidelberg, Germany. Tissue cores of three cancer (sub-)types were used, comprising lung adenocarcinoma (ADC) (168 patients), lung squamous cell carcinoma (SqCC) (136 patients) and pancreatic adenocarcinoma (141 patients). The corresponding dataset is publicly available from (`https://gitlab.informatik.uni-bremen.de/digipath/Deep_Learning_for_Tumor_Classification_in_IMS` – last accessed 7 February 2022).

3.4.2 Data preprocessing

We performed the following (further) data preprocessing steps. First, due to the high dimensionality of the feature space (27, 286 m/z values), spectral filtering was carried out. Namely, we loaded the data into MATLAB R2018 and used the internal library MSClassifyLib (function MSAdaptiveResampleTrigger()) to apply the spectral filtering with the default value of 0.4 Da (cf. [63, 16]). Second, the mass range was pruned up to the mass range of 2100 m/z, and we used one mass spectrum from each patient. Finally, the well-established normalization step TIC (total ion count) was employed on the chosen mass spectra range ([6]). This data preprocessing resulted in a dataset with 445 mass spectra (observational units) and 1579 m/z channels (features).

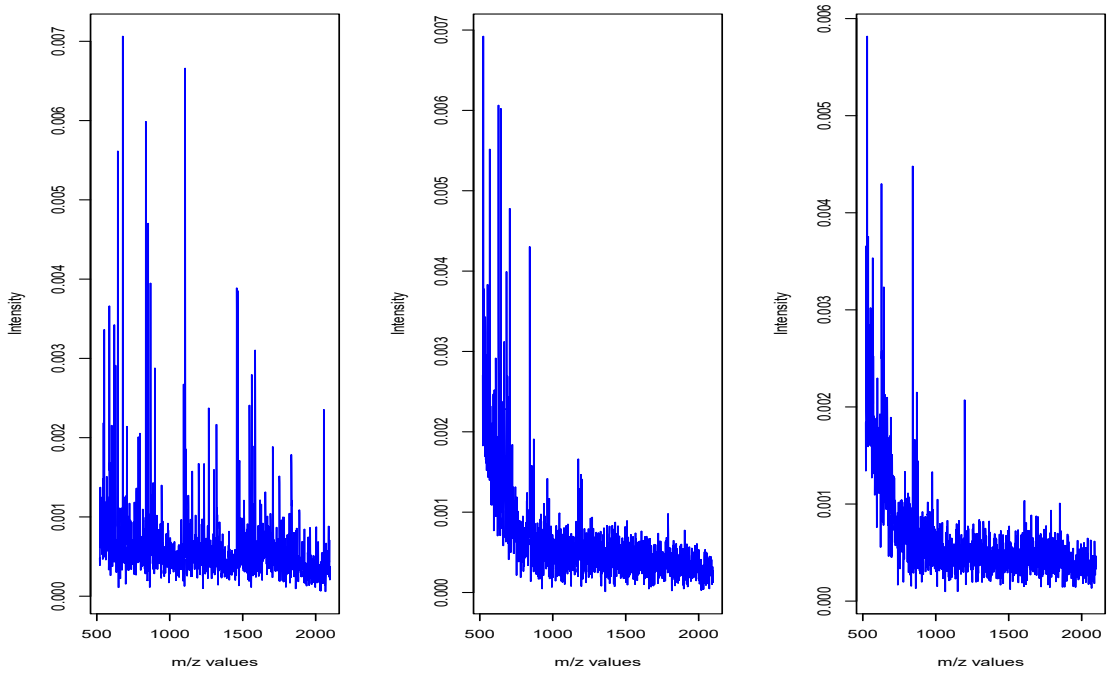


Figure 3.1: Examples of mass spectra for three cancer (sub-)types.

3.4.3 Analysis of the MALDI dataset

We have modeled each pixel (based on the averaging filter around 0.4 Da, as explained in the previous section) individually to determine those m/z values that are highly associative for certain cancer (sub-)type. Thus, the m/z values take the role of the X_j's, and the cancer (sub-)type takes the role of Y from our general setup, where $j \in \{1, \ldots, p = 1{,}579\}$ here. Statistically speaking, our procedure identifies significant differences in distributions across the (sub-)types and is not only identifying significant peaks in the spectra.

Figure 3.2 displays the empirical distributions of the Z-statistics, along with histograms of the non-adjusted p-values, for both tasks. Clearly, the Z-values (for both tasks) do not resemble the realization of a random sample from the theoretical null distribution, which is the standard normal distribution on $\mathbb{R}$. In particular, the empirical variances are much larger than one in both histograms displayed in Figure 3.2. There are two reasons for this overdispersion, namely, (i) the presence of correlation effects among the Z-values (see [33, 32]) and (ii) the presence of extremely large (in absolute value) Z-values, which is likely caused by significant effects, peaks and isotopic patterns. In our previous study, we have observed a similar phenomenon and have described a procedure (based on the "empirical null distribution") which can be used to account for the part of the overdispersion which is caused by correlations among the Z-statistics (cf. [115]).

The main results of our real data analysis are presented in Figure 3.3, where we have chosen $k = 3$ for both tasks. According to the number $p = 1{,}579$ and taking into account the magnitude of the observed correlation effects

(a) Task: Pancreatic vs. ADC

(b) Task: Pancreatic vs. SqCC

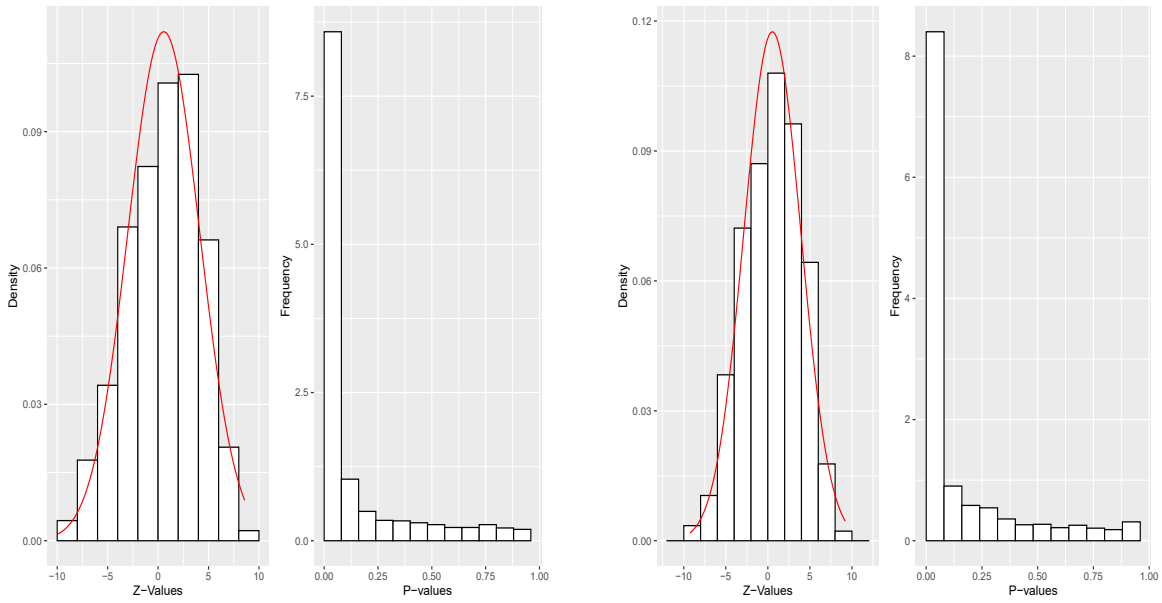


Figure 3.2: Empirical distributions and fitted normal density curves of the Z-values, as well as histograms of the unadjusted p-values, for both tasks. The fitted normal distributions are $N(0.565, 3.365^2)$ for the first task and $N(0.543, 3.394^2)$ for the second task. As a result of overdispersion and non-zero effect sizes, there are many non-adjusted p-values which are close to zero, in both graphs.

among the Z-statistics, a plausible range for the rejection threshold t has been selected for being displayed on the horizontal axes. The subplots in Figure 3.3 illustrate the total number of rejections ($R(t)$), the estimated number of false discoveries ($V(t)$), and the estimated FDP ($FDP(t)$) over this range. Apparently, all three aforementioned quantities decrease with decreasing t (notice the negative logarithmic scale of the horizontal axes in Figure 3.3). For both tasks, $\widehat{\text{FDP}}(t)$ ranges in $[5\%, 15\%]$ for $t \in [10^{-7}, 10^{-5}]$. Table 3.5 tabulates $R(t)$ and $\widehat{\text{FDP}}(t)$ for several threshold values t from the latter range.

Table 3.5: **Main results of the Multi-PFA analysis for both tasks.**

(a) Pancreatic vs. ADC.

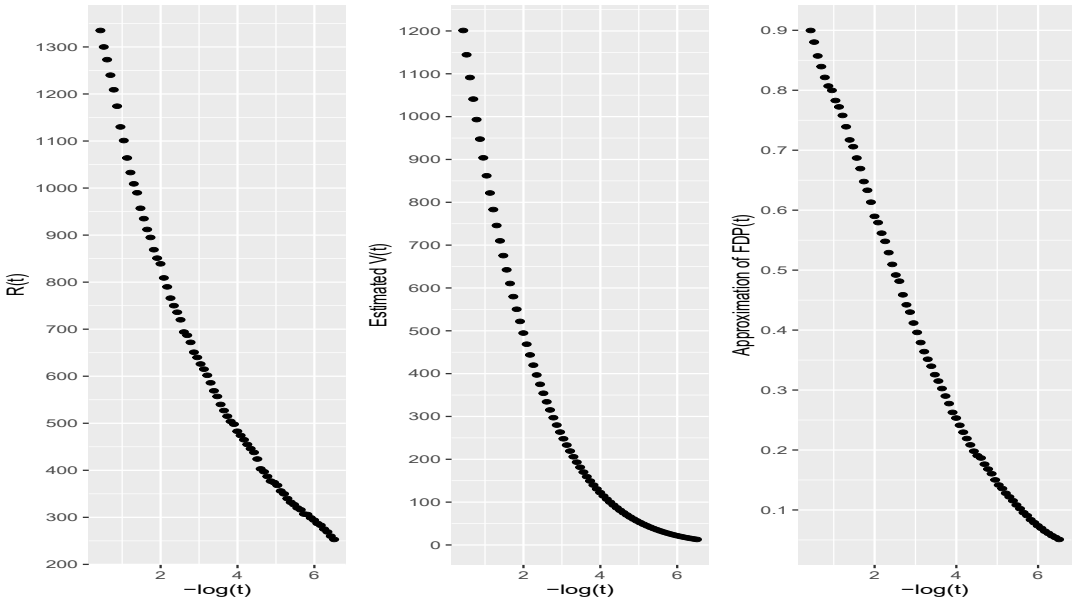
Threshold t	$R(t)$	$\widehat{\text{FDP}}(t)$
1.12e-05	374	0.1502
7.50e-06	356	0.1357
4.12e-06	332	0.1154
2.76e-06	320	0.1022
5.57e-07	275	0.0609
3.06e-07	253	0.0508

(b) Pancreatic vs. SqCC.

Threshold t	$R(t)$	$\widehat{\text{FDP}}(t)$
5.03e-06	295	0.1505
3.37e-06	289	0.1349
2.76e-06	280	0.1304
1.85e-06	262	0.1217
8.32e-07	237	0.1013
6.18e-07	168	0.0490

To the best of our knowledge, no biomarkers have been confirmed yet for this dataset (pancreas vs. lung task) (cf. [9]). However, according to the previous reports, m/z values at 836.5 Da, 852.4 Da and 868.5 Da might be potential biomarkers (as discussed in [9, 16]). Since we have used data based on a different resolution (0.4 Da) compared to the aforementioned reports, we identified as statistically significant m/z-values which are closely

(a) Task: Pancreatic vs. ADC.



(b) Task: Pancreatic vs. SqCC.

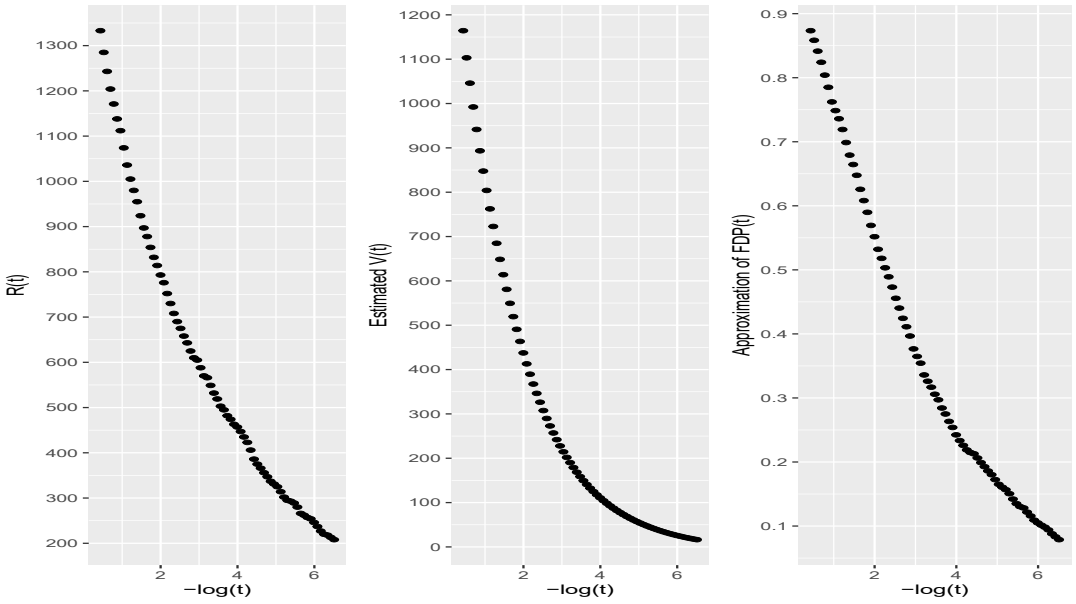


Figure 3.3: Main results: Overall number of rejections, estimated number of false rejections, and estimated FDP, as functions of the threshold t, for both tasks. The values for t on the horizontal axis are plotted on the minus log_{10} scale.

Table 3.6: **Top 10 ranked m/z values based on their original Z-values for the task Pancreatic vs. ADC.**

(a) The most sign. m/z-values for Panc.

m/z values	Z-values	P-values
886.44	-9.917	$< 10^{-6}$
840.42	-9.570	$< 10^{-6}$
679.34	-9.300	$< 10^{-6}$
899.45	-9.176	$< 10^{-6}$
854.42	-8.999	$< 10^{-6}$
771.38	-8.930	$< 10^{-6}$
874.43	-8.866	$< 10^{-6}$
841.42	-8.528	$< 10^{-6}$
898.44	-8.430	$< 10^{-6}$
785.39	-8.404	$< 10^{-6}$

(b) The most sign. m/z-values for ADC

m/z values	Z-values	P-values
1647.82	8.585	$< 10^{-6}$
1575.78	8.528	$< 10^{-6}$
535.26	8.300	$< 10^{-6}$
521.26	8.252	$< 10^{-6}$
1694.84	8.245	$< 10^{-6}$
1599.79	8.197	$< 10^{-6}$
531.26	8.174	$< 10^{-6}$
522.26	7.963	$< 10^{-6}$
532.26	7.844	$< 10^{-6}$
529.26	7.672	$< 10^{-6}$

Table 3.7: **Top 10 ranked m/z values based on their original Z-values for the task Pancreatic vs. Sqcc.**

(a) The most sign. m/z-values for Panc.

m/z values	Z-values	P-values
771.38	-9.285	$< 10^{-6}$
795.39	-8.839	$< 10^{-6}$
678.34	-8.727	$< 10^{-6}$
785.39	-8.390	$< 10^{-6}$
840.42	-8.334	$< 10^{-6}$
874.43	-8.276	$< 10^{-6}$
886.44	-8.274	$< 10^{-6}$
796.39	-8.266	$< 10^{-6}$
841.42	-8.202	$< 10^{-6}$
854.42	-8.122	$< 10^{-6}$

(b) The most sign. m/z-values for SqCC

m/z values	Z-values	P-values
535.26	9.196	$< 10^{-6}$
1575.78	8.860	$< 10^{-6}$
521.26	8.664	$< 10^{-6}$
530.26	8.406	$< 10^{-6}$
531.26	8.381	$< 10^{-6}$
522.26	8.339	$< 10^{-6}$
1694.84	8.125	$< 10^{-6}$
1611.80	7.946	$< 10^{-6}$
565.28	7.917	$< 10^{-6}$
1764.87	7.910	$< 10^{-6}$

related to the previously published insights. Namely, the adjacent molecules have appeared to be extremely statistically significant. In particular, for the m/z-value 836.5 Da, based on the spectral filtering used in our analysis, adjacent m/z values (836.42, 837.41 and 838.42) have been indicated as highly significant for both tasks. Similarly, for the m/z value 852.4 Da, adjacent m/z values have occurred as highly considerable (853.42 and 854.42). Similarly, we observed the m/z value at 868.50 and the adjacent m/z value at 869.43. Figure 3.4 illustrates the identified peaks related to the pancreatic status. The Multi-PFA method indicates m/z values around 886.42 Da and 852.42 Da as highly statistically significant. Tables 3.6 and 3.7 tabulate the 10 top-ranked m/z values for both tasks. Their corresponding Z-values lie in the very tails of the empirical distributions plotted in Figure 3.2. Negative signs in Tables 3.6 and 3.7 indicate those m/z values that are distinctive for the pancreatic cancer type.

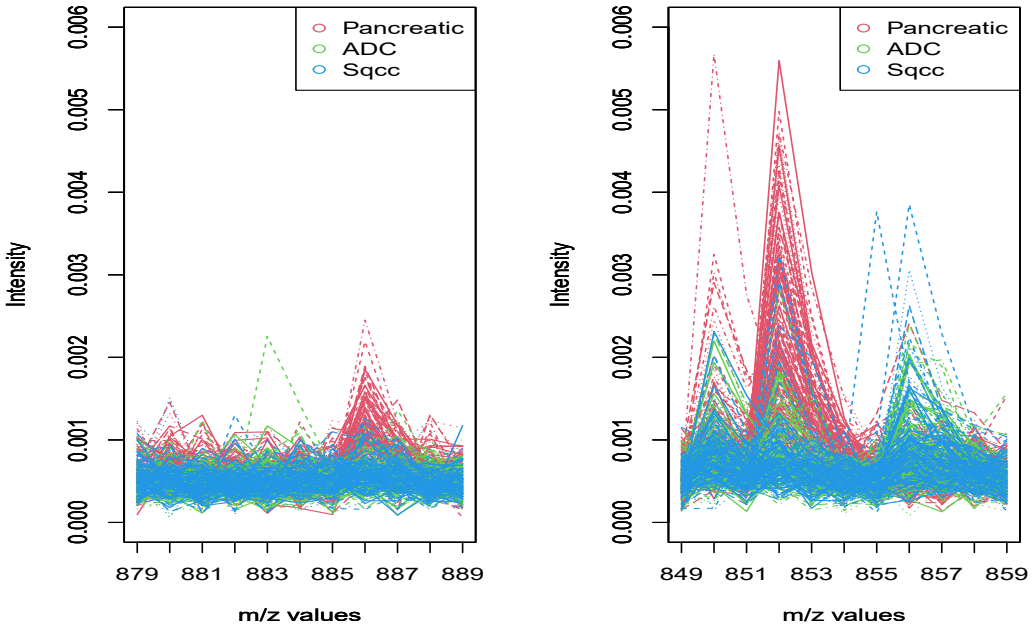


Figure 3.4: A comparison of the relative intensities at each m/z value across the (sub-)types for different zooms related to very significant m/z values for the pancreatic association.

Finally, we compare our methodology with two classical procedures for controlling the FDR, namely the BH ([11]) procedure and the Benjamini-Yekutieli (BY) ([12]) procedure, in Table 3.8. We used the R function p.adjust() to carry out the latter multiple tests. These comparisons are relevant since the FDR is the expected value of the FDP, implying that FDP control and FDR control are related concepts. Both the BH and the BY procedures operate on (marginal) p-values. To obtain these p-values, we consider for each $c \in \{1, 2\}$ the standardized versions $Z_{1c}, \ldots Z_{pc}$ of $\hat{\beta}_{1c}, \ldots \hat{\beta}_{pc}$ as given in (3.7). In particular, the empirical standard errors used for Studentization are taken as the square roots of the diagonal elements of $\hat{\Sigma}_c$, divided by $\sqrt{n}$. Next, each Z_{jc} (for $1 \leq j \leq p$) is transformed into the corresponding two-sided p-value $P_{jc} = 2\Phi(-|Z_{jc}|) = 2(1 - \Phi(|Z_{jc}|))$ (cf. Figure 3.2). The

resulting numbers of rejections are provided in the columns labeled "BH" and "BY" in Table 3.8. It is apparent that both BH and BY are (too) liberal in this context. For example, for a nominal level of $\alpha = 0.1$, BH and BY reject approximately 70% of all null hypotheses. The reason for this liberal behavior is that the scaling of the "empirical null distribution" of $Z_{1c}, \ldots Z_{pc}$ is different from the unit scale due to strong dependencies, even though each individual Z_{jc} is properly Studentized (for more details, see in [33, 115]). In order to further substantiate this point and to rule out the possibility that the observed liberal behavior may result from an inaccurate estimation of Σ_c, we present the columns labeled BH* and BY* in Table 3.8. For these columns, the p-values have been extracted directly from the marginal model fits (with standard errors based on the marginal Fisher information matrices), without carrying out the MMM approach. The results remain essentially the same. To diminish the variance inflation effects of the dependencies and to make the comparisons with Multi-PFA "fair" and meaningful, we applied Efron's "empirical correction" of the null distribution ([33]). Briefly put, the idea here is to divide the Z-values by an empirical value based on their central spread, where it is assumed that the central part of the empirical distribution of the Z-values presumably corresponds to true null hypotheses. In our case, these empirical values equal 2.48 (Part (a) of Table 3.8) and 2.50 (Part (b) of Table 3.8), respectively. We denote by BH^E and BH^{E*} the columns where this empirical correction has been performed. Now, the variance inflation is corrected, but the generic procedures turn out to be conservative in comparison with Multi-PFA. The reason for this is that the correlation effects among the m/z-values are not accounted for by the BH procedure. We may remark here that the aforementioned variance inflation is taken care of automatically by Multi-PFA. This can be seen by comparing the values in the columns labeled "Threshold t" in Table 3.5 with the value α/p.

Table 3.8: **The numbers of rejections for the Multi-PFA, the BH and the BY procedures are compared for three significance levels.**

(a) Pancreatic vs. ADC.

Sign. lev.	Multi-PFA	BH	BY	BH*	BY*	BH^E	BH^{E*}
$\alpha = 0.05$	253	946	714	957	739	0	20
$\alpha = 0.10$	320	1045	785	1068	812	18	35
$\alpha = 0.15$	374	1120	839	1136	853	115	142

(b) Pancreatic vs. SqCC.

Sign. lev.	Multi-PFA	BH	BY	BH*	BY*	BH^E	BH^{E*}
$\alpha = 0.05$	168	901	665	922	688	0	0
$\alpha = 0.10$	237	1018	735	1034	764	18	65
$\alpha = 0.15$	295	1097	783	1103	811	53	104

3.5 Discussion and Outlook

Motivated by MALDI association studies, this study's objective has been to evaluate the strength of associations between a nominal variable of interest – describing cancer (sub-)types in our application – and a large number

of measured variables (in our case given by m/z values). We have proposed an approach to screen all features and identify the most associative ones with the multi-class outcome. Our approach decomposes the outcome variable into multiple baseline-category pairs and approximates the false discovery proportion under arbitrary correlation dependency within each pair. As demonstrated on simulated data, this approach leads to a sensible balance between false and true findings. Furthermore, both our simulation study and the presented application to real data demonstrate that the proposed method can be applied even in cases where the sample size is considerably smaller than the number of features. This makes the procedure attractive for medical screening applications. Two central assumptions underlie the proposed methodology: First, we assume sparsity in the sense that the number of active features is small. Second, we assume that the dependency structure amongst the considered Z-statistics can accurately be approximated by a multi-factor model. The approach of MMM, however, is flexible in the sense that it does not depend on heavy assumptions. For its applicability, we only have to assume that the multi-class outcome is associated with features (m/z values), and that this association can accurately be described by a (marginal) multinomial regression model for each m/z-value separately. A clear limitation of MMM is that the approach relies on an asymptotic setting, such that small data regimes (small sample sizes) cannot be handled with the proposed methodology. Furthermore, the computational effort of the proposed methodology is quite high because p marginal models have to be fitted.

From the application perspective, we have applied the Multi-PFA method to a MALDI imaging dataset consisting of a large number of m/z values and one spectrum from each patient. To the best of our knowledge, we are the first to address the three-class problem for this dataset. In this way, we have provided a more detailed analysis for both lung cancer subtypes than in previous studies.

There are several potential directions for further research. First, it might be captivating to consider different supervised learning methods (for instance, neural networks with more than one layer) instead of the multinomial regression model proposed in this paper. Second, it may be of interest to evaluate the uncertainty regarding the realized FDP in order to provide a confidence region for it, in addition to mere point estimation of the FDP. Namely, we aim at establishing an "exceedance control" of the FDP, where the latter is approximated by the PFA estimator ([45, 54]). Based on this, one can construct upper confidence bounds for the FDP, ideally simultaneously over a grid of rejection thresholds. Third, disentangling correlation effects and regression effects on the empirical Z-statistics distribution in a more detailed manner is an interesting follow-up research topic in our context.

Chapter 4

Supervised topological data analysis for MALDI mass spectrometry imaging applications

Information: This chapter is a slightly modified version of Klaila, Vutov, and Stefanou (2023) ([57]) published in *BMC Bioinformatics*.

Authors:

Gideon Klaila, Institute for Algebra, Geometry, Topology and their Applications (ALTA), University of Bremen, Bremen, Germany.

Vladimir Vutov, Institute for Statistics, University of Bremen, Bremen, Germany.

Prof. Dr. Anastasios Stefanou, Institute for Algebra, Geometry, Topology and their Applications (ALTA), University of Bremen, Bremen, Germany.

Declaration of individual contributions: This paper is based on the joined work of the three main authors as equal contributors. Gideon Klaila implemented the algorithm and proved the complexity. I took care of the statistical treatment. Specifically, the choice of classifiers and the design and implementation of the statistical experiments (including the idea of using the mass spectrometry images). Prof. Anastasios Stefanou helped clarify the notion of persistent transformation and its connection to the persistence diagram via the elder rule and proposed examples illustrating this connection. All authors have written the manuscript and approved the final version.

Availability of data and materials: Python code with the numerical results presented in the paper, including the MALDI data, is available at: https://github.com/klailag/SupervisedTDAMethodForMALDI

Background: Matrix-assisted laser desorption/ionization mass spectrometry imaging (MALDI MSI) displays significant potential for applications in cancer research, especially in tumor typing and subtyping. Lung cancer is the primary cause of tumor-related deaths, where the most lethal entities are adenocarcinoma (ADC) and squamous cell carcinoma (SqCC). Distinguishing between these two common subtypes is crucial for therapy decisions and successful patient management.

Results: We propose a new algebraic topological framework, which obtains intrinsic information from MALDI data and transforms it to reflect topological persistence. Our framework offers two main advantages. Firstly, topological persistence aids in distinguishing the signal from noise. Secondly, it compresses the MALDI data, saving storage space and optimizes computational time for subsequent classification tasks. We present an algorithm that efficiently implements our topological framework, relying on a single tuning parameter. Afterwards, logistic regression and random forest classifiers are employed on the extracted persistence features, thereby accomplishing an automated tumor (sub-)typing process. To demonstrate the competitiveness of our proposed framework, we conduct experiments on a real-world MALDI dataset using cross-validation. Furthermore, we showcase the effectiveness of the single denoising parameter by evaluating its performance on synthetic MALDI images with varying levels of noise.

Conclusion: Our empirical experiments demonstrate that the proposed algebraic topological framework successfully captures and leverages the intrinsic spectral information from MALDI data, leading to competitive results in classifying lung cancer subtypes. Moreover, the framework's ability to be fine-tuned for denoising highlights its versatility and potential for enhancing data analysis in MALDI applications.

4.1 Background

Matrix-assisted laser desorption/ionization mass spectrometry imaging (MALDI MSI), also known as MALDI Imaging, is a label-free tool for spatially furnishing molecular weight information of compounds like proteins, peptides, and many others (see, e.g., [70, 19]). Provided a thin biological sample (usually a tissue section [4, 16]), MALDI collects mass spectra at multiple discrete positions within the biological sample. As a result, an image is obtained where each spatial spot presents a mass spectrum ([3]). The latter depicts the relative abundances of ionizable molecules with a significant number of mass-to-charge ratio (m/z) values, ranging from a couple of hundreds to a few tens of thousands of m/z values (see [4, 3, 9]). MALDI has been demonstrated to be a valuable instrument for many pathological applications (for more details, see [59, 112]), which is possible because of its feasibility in examining formalin-fixed paraffin-embedded (FFPE) tissue samples. In other words, the MALDI tool allows analyses of multiple tumor cores across many patients by aggregating them in a single tissue microarray (TMA) (cf. [19]). As discussed in [63], "the pathological diagnosis of a tumor found in a tissue specimen, including the determination of the tumor origin and genetic subtype" ([63]) is essential for adequate treatment of patients.

This study reports our findings on a MALDI dataset based on two lung cancer (LC) subtypes.

As pointed out by several studies (e.g., in [16, 9, 115]), a plain approach to discovering meaningful m/z values relies upon the idea of identifying significant signal peaks, also known in the literature as peak detection. Focusing on the relevant peaks, one can neglect those highly associated with noise, as acknowledged in [3]. Significant peaks are assumed to provide information for discriminating mass spectra from different cancer subtypes (see, among others, in [118, 110]). Different peak-detection algorithms have been compared in [119]. Furthermore, in [66], a more recent and novel approach proposes to incorporate an isotope pattern ([100]) around the chosen peaks, which can boost the peak detection methodology.

Other methods aim to extract "characteristic spectral patterns (CSP)" from MALDI-MSI data (cf. [16]). Namely, these methods combine spectral information from different correlated spectral features into a lower dimensional subspace of the data. Afterward, classification models are performed on the extracted feature vectors to classify data units into class labels, i.e., the tumor types or subtypes. Some other frameworks tend to perform feature selection first. Then, based on the selected features, such frameworks execute supervised classification methods to classify observational units into response labels (e.g., [93]). In the context of variable selection, in [115, 114], the authors have proposed approaches by means of large-scale simultaneous testing so as to identify the most associative m/z values with the (cancerous) outcome variables.

The usual challenge modeling methods face is a significant amount of spectral data. As more and more data are being gathered, analyzing the data efficiently with short computational time becomes increasingly more challenging. Topological data analysis (TDA) is a contemporary scientific area that arose from diverse works in applied topology and computational geometry (see [77, 21]). TDA offers an algebraic way of reducing the dimensionality of datasets and extracting essential features in short computational times.

TDA is a relatively novel field of data analysis, and more theoretical work needs to be done to improve the methods (cf. [58]). Even so, TDA has been successfully employed in various fields of science, e.g., in physics, chemistry, and bio-medicine ([99]), as well as in oncology (see [18, 67]). Motivated by these approaches, this study's objective is to propose a novel framework based on the algebraic topology of MALDI imaging to get improved classification results in a shorter computational time.

In the context of MALDI, one can take advantage of TDA by filtering out the most relevant part of the data, namely the peak-related information. We hypothesize that the importance of a peak increases with its relative height, also known as topological "persistence". Accordingly, low persistent peaks are more likely to be noise. To this end, one can benefit from utilizing our topological framework due to its superior characteristics of denoising and compression.

A general approach to determining a peak's persistence is the upper-level set filtration, where each peak corresponds to a topological feature. These topological features are tracked from their appearance until they merge with a larger feature. This way of analyzing the data is fast in its computational time but has a significant drawback.

While tracking the persistence of each peak, one loses the information about their positions, which is paramount to carrying out data analysis applications. For example, in the context of MALDI-related studies, the locations of the biomarkers (cf. [16, 115]) are relevant information for the analysis. To circumvent this limitation, we introduce a different analysis method: the persistence transformation (cf. [117]). The proposed approach keeps track of each peak's position while determining its persistence. This enables the application of this methodology to spectral data.

4.2 Topological Data Analysis

4.2.1 MALDI data structure

The first step of our approach is to transform the input MALDI data in order to reflect the topological persistence. MALDI-MSI datasets are commonly stored in an $n \times q$ matrix, denoted by X, and X takes its values in $\mathbb{R}_{\geq 0}^{n \times q}$, where each data record corresponds to an intensity value. Usually, mass spectra are stored as rows. While every data column corresponds to an intensity plot for a certain m/z value (see in [63, 42]), n corresponds to the number of mass spectra, and q is the number of m/z values within each mass spectrum. Furthermore, the data records are non-negative since MALDI data presents information on molecular masses of ionizable molecules (cf. [3]). For the convenience of notation, we assume that each mass spectrum is represented as a set of tuples $M := \{(x_1, s_1), (x_2, s_2), \cdots (x_q, s_q)\}$, where x_j is the j-th m/z value and s_j is the j-th intensity value for $1 \leq j \leq q$. Note that this set M is equipped with a real-valued function f corresponding to the projection to the second coordinate. That means that the intensity values induce a map $f : M \to \mathbb{R}$ with $f(x_j) := s_j$ for $1 \leq j \leq q$ within each mass spectrum.

To sum up, our approach processes each mass spectrum (defined by the set M) individually in order to extract the topological properties of the data, i.e., the topological persistence.

4.2.2 Topological Persistence

Let $f : M \to \mathbb{R}$ be a real-valued function on a compact set, then for $a \in \mathbb{R}$ the upper-level set can be defined as $M_a := \{x \in M \,|\, f(x) \geq a\}$. Note that $M_a \subseteq M_{a'}$ for any $a > a'$. This yields the "upper-level set filtration" $M_{a_1} \subseteq \ldots \subseteq M_{a_i}$ for $a_1 > \ldots > a_i \in \mathbb{R}$ (for more details; see Chapter 18 in [58]). In this study, we aim at utilizing the upper-level set filtration instead of the common alternative, i.e., the sublevel set filtration (defined as $M_{\leq a} := \{x \in M \,|\, f(x) \leq a\}$). Since the latter, filtration detects local minima and tracks them until they merge with other minima. Conversely, the maxima are highly important in MALDI applications because they correspond to the underlying spectral peaks. As mentioned, peaks provide the necessary information to distinguish mass spectra from different cancerous subtypes ([3, 115]). To this end, the upper-level set filtration is of interest in this study since it tracks the local maxima.

Let $x, x' \in M$ with a path-connection, i.e. there exists a continuous function $\rho : [0,1] \to M$ such that $\rho(0) = x$ and $\rho(1) = x'$. We denote the image of the map ρ by $[x, x']_\rho$. Then we say that x and x' are path-connected in M_a, and we write $x \sim_a x'$, if there exists a path connection ρ of x, x', such that $\forall \hat{x} \in [x, x']_\rho : \hat{x} \in M_a$, i.e. $f(\hat{x}) \geq a$, for all $\hat{x} \in [x, x']_\rho$. To track the homology in the upper-level set filtration, we now identify all path-connected points to each other. The degree of the 0-the homology group h_0 is given by $b_0(M_a) = |M_a / \sim_a|$, i.e. the number of different path-connected components in M_a. These components are called "topological features". Henceforth, we refer to a (topological) feature in this pure topological sense, not a feature in a statistical sense, like a covariate or an explanatory variable. Remark that there are no features of dimension one or higher since each data unit (mass spectra; see Figure 4.5) is represented as a curve.

In the upper-level set filtration, a topological feature is detected for x at a^* if and only if $x \in M_{a^*}$ and $\forall x' \in M_{a^*} : x \not\sim_{a^*} x'$, i.e., x is not path-connected to any other element in M_{a^*}. The feature in x merges with another feature in M_{a^+}, if $a^+ = \max\{a \,|\, x \in M_a \wedge \exists x' \in M_a : x \sim_a x' \wedge f(x') > f(x)\}$, i.e., the largest upper-level set in which the peak gets path-connected to a larger peak. We call a^* the *birth* of the feature, a^+ the *death* of the feature, and $p := a^* - a^+$ the *persistence* of the feature. Since the largest peak (the global maximum) does not merge with any other feature, its death is defined as the global minimum. The induced merging order is according to the "elder rule" (see [30]).

To encode all information about features and their persistence in a meaningful and comparable way, the persistence diagram is utilized. Figure 4.1 displays an example of the encoding process for the upper-level set filtration. Here, the birth and death axis are swapped since the birth values are always greater or equal to the death values. In this way, all the feature points appear above the diagonal line $x = f(x)$ (cf. Section 4 in [41]). In the persistence diagram, each feature is represented by a point (a^*, a^+) with a multiplicity for similar features. The closer a point is to the diagonal of the diagram $\{(x, f(x) = x) \mid \forall x\}$, the lesser its persistence is and vice versa. For more details of the standard persistence diagram approach, we refer to [31] and [22].

Given two real-valued functions on a compact set $f, g : M \to \mathbb{R}$, the resulting persistence diagrams $\mathbf{dgm}(f)$ and $\mathbf{dgm}(g)$ can be compared with a suitable metric on the diagrams, e.g., the bottleneck distance (cf. [22, 113]). This metric defines the closeness of persistence diagrams and can also indicate the closeness of the corresponding functions in the sense that if the persistence diagrams are different, the functions are different. In the context of MALDI-MSI data, the persistence diagram can be used as a proxy for distinguishing cancer (sub-)types. The opposite case that similar persistence diagrams imply similar functions does not hold, as illustrated in Figure 4.4.

4.2.3 Persistence Transformation

Using persistence homology to track the features and applying the bottleneck distance on the resulting persistence diagrams is a viable way to distinguish functions. However, there is a disadvantage to this approach. While

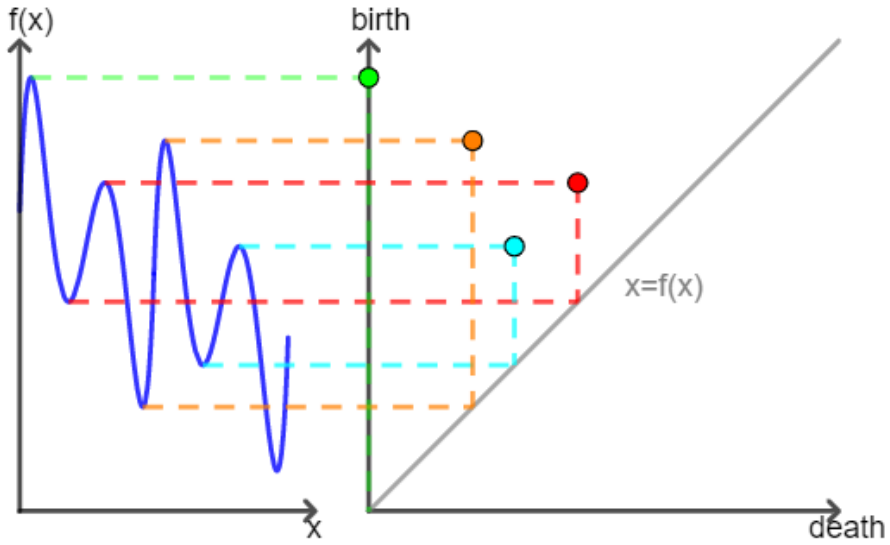


Figure 4.1: The persistence diagram. **Left hand**: A real-valued function $f : M \to \mathbb{R}$ is plotted with $x \in M$ on the x-axis and $f(x)$ on the y-axis. **Right hand**: The corresponding persistence diagram is illustrated. Here, the y-axis represents the birth values, while the x-axis represents the death values. Each point stands for a topological feature for the function f. The first feature, corresponding to the global maxima, never dies. To indicate this, we mark its death value with 0.

tracking the persistence of each feature in a comprehensible way, its position is not being tracked. Nevertheless, as mentioned above, in data-driven applications, including MALDI Imaging, the position of variables is a required property.

To process the information regarding the positions along with the persistence of peaks, we approach the topological manipulation of the spectral data differently. Instead of creating a point $(a^*, a^+) \in \mathbb{R}^2$ for each feature and displaying it in the persistence diagram, we introduce a new dimension to track the position. For each feature, this approach gives a point $(x, a^*, a^+) \in M \times \mathbb{R}^2$, where $x \in M$ is the position of each peak. We define a pairing function $\mu : M \to \mathbb{R}$ with $\mu(x) = a$. The value a is defined to be the highest value smaller or equal to $f(x)$, which upper-level set M_a contains a point x' having a greater function value (similarly as in [117]):

$$\mu(x) := \sup\{a \leq f(x) \,|\, \exists x' \in M_a : f(x') > f(x) \wedge x \sim_a x'\}. \tag{4.1}$$

For the global maximum $\hat{x}$ the pairing value $\mu(\hat{x})$ is not defined, so we define $\mu(\hat{x}) := \min\{a \in \mathbb{R} \,|\, \exists x : f(x) = a\}$ instead. Then the birth of a topological feature is given by $a^* = f(x)$, while the death is calculated by $a^+ = \mu(x)$. The persistence p of a point x can now be defined to be

$$p(x) := f(x) - \mu(x) = a^* - a^+.$$

For any point x not being a local maximum there is a local maximum $x' \in M_{f(x)}$ with $f(x') > f(x)$ and

$x \sim_{f(x)} x'$. Then the pairing of x is trivial, i.e. $\mu(x) = f(x)$ with the resulting persistence of $p(x) = f(x) - \mu(x) = 0$. Alternatively, for a point x being a local maximum, the pairing is non-trivial, i.e., there is $\mu(x) < f(x)$, which results in a non-zero persistence $p(x) = f(x) - \mu(x) > 0$. This pairing value always corresponds to $f(\tilde{x})$ for a unique local minimum $\tilde{x}$, i.e.

$$\mu(x) = f(\tilde{x}). \tag{4.2}$$

The persistence transformation $t : M \to M \times \mathbb{R}^2$ can then be defined for each $x \in M$ to be $t(x) = (x, f(x), \mu(x)) = (x, a^*, a^+) \in M \times \mathbb{R}^2$. For each $x \in M$, the feature triple $t(x) = (x, a^*, a^+)$ consists of the position, the birth value, and the death value. The storage can be reduced to $3 \cdot m$, with m being the number of peaks, by neglecting the trivial tuples, resulting in the "persistence transformation vector".

Similar to the persistence diagram of the upper-level set filtration, the merging of features in the persistence transformation occurs according to the elder rule (see [30]), i.e., the topological feature with the higher birth value persists when merged to another feature. The process can be illustrated in the corresponding merge tree (e.g., in Figure 4.2).

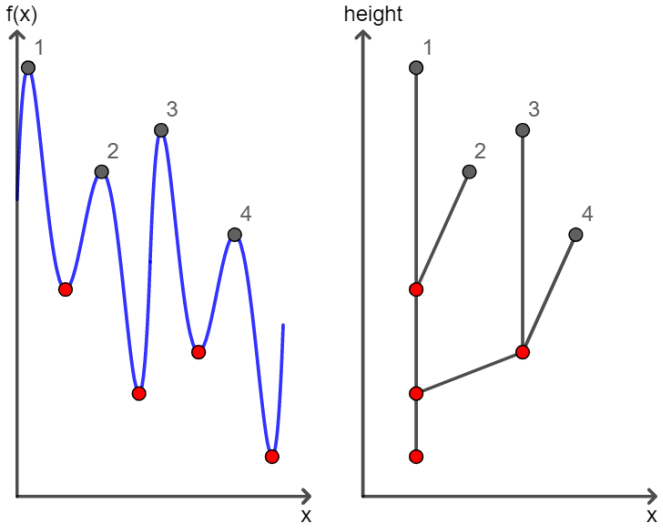


Figure 4.2: The merge tree. A merge tree of a real-valued function $f : M \to \mathbb{R}$ is depicted. On the x-axis are the $x \in M$ values. On the y-axis is $f(x) \in \mathbb{R}$ (left-hand side) and the height (right-hand side).

4.2.4 Application

The notion of the persistence transformation is, in theory, a great way to store more topological information about a graph in general. But many applications do not need the whole information provided by the persistence feature. In these cases, the greater dimensionality of the persistence transformation is rather disadvantageous in calculations. This could be evaded by introducing a "reduced persistence transformation", where instead of $t(x) = (x, a^*, a^+)$

only the position and the persistence are stored: $\tilde{t}(x) = (x, a^* - a^+) \in M \times \mathbb{R}$ for $x \in M$. This computational reduction has two benefits. First, it compresses the persistence vector. Second, it can track the persistence of each topological feature and its position.

Another computational improvement for applications can be made by associating low persistent features with noise. By omitting the possible noise by only considering the $k\%$ most persistent features, the accuracy of the analysis can be improved. In addition, reducing the number of stored features might also improve the run-time of subsequent approaches.

Finally, in many applications, there exists a total order on the set M, e.g., if $M \subseteq \mathbb{R}$. This order can be passed to the feature space, such that the elder rule for features with equal birth value can be applied deterministically by using the induced order.

4.2.5 Comparison

The persistence transformation is strictly a better invariant in distinguishing two functions $f, g : M \to \mathbb{R}$ than the persistence diagram of their 0-dimensional upper-level set filtration. By taking the projection, $p_i((x, a^*, a^+)) = (a^*, a^+)$ of the persistence transformation, the persistence diagram of the upper-level set filtration can be obtained. Hence, all functions that can be distinguished by the persistence diagram can also be distinguished by the persistence transformation. Furthermore, there are cases where the persistence transformation differentiates two functions successfully while the zero-dimensional persistence diagram of the upper-level set filtration fails to do so (see Figure 4.3).

The reduced persistence transformation, on the other hand, is not strictly better than the persistence diagram of the upper-level set filtration. However, there are cases where the reduced persistence transformation outperforms the persistence diagram (see Figure 4.4). For MALDI-related applications, one seeks to utilize exactly the advantages that the (reduced) transformation offers. For example, the total height of a peak is less interesting in these applications than the relative height, and the position of the peak can be used to backtrack molecules. To this end, in the MALDI-MSI applications, the reduced persistence transformation performs better in analyzing the spectral data than the persistence diagram of the upper-level set filtration.

The benefit of the persistence transformation can be utilized in different scientific areas where the position of the peaks is of interest (e.g., [49]). In the context of TDA, data analyses generally only work reliably if the applied method is stable, meaning that a slight change in the data (given a suitable metric) only leads to small changes in the numerical results. The persistence diagram has been proven stable (see [22]). Furthermore, there are stability theorems for other topological methods (see [23, 68]), but to the best of our knowledge, there has not yet been a proven stability theorem for the persistence transformation. This may be done in further work.

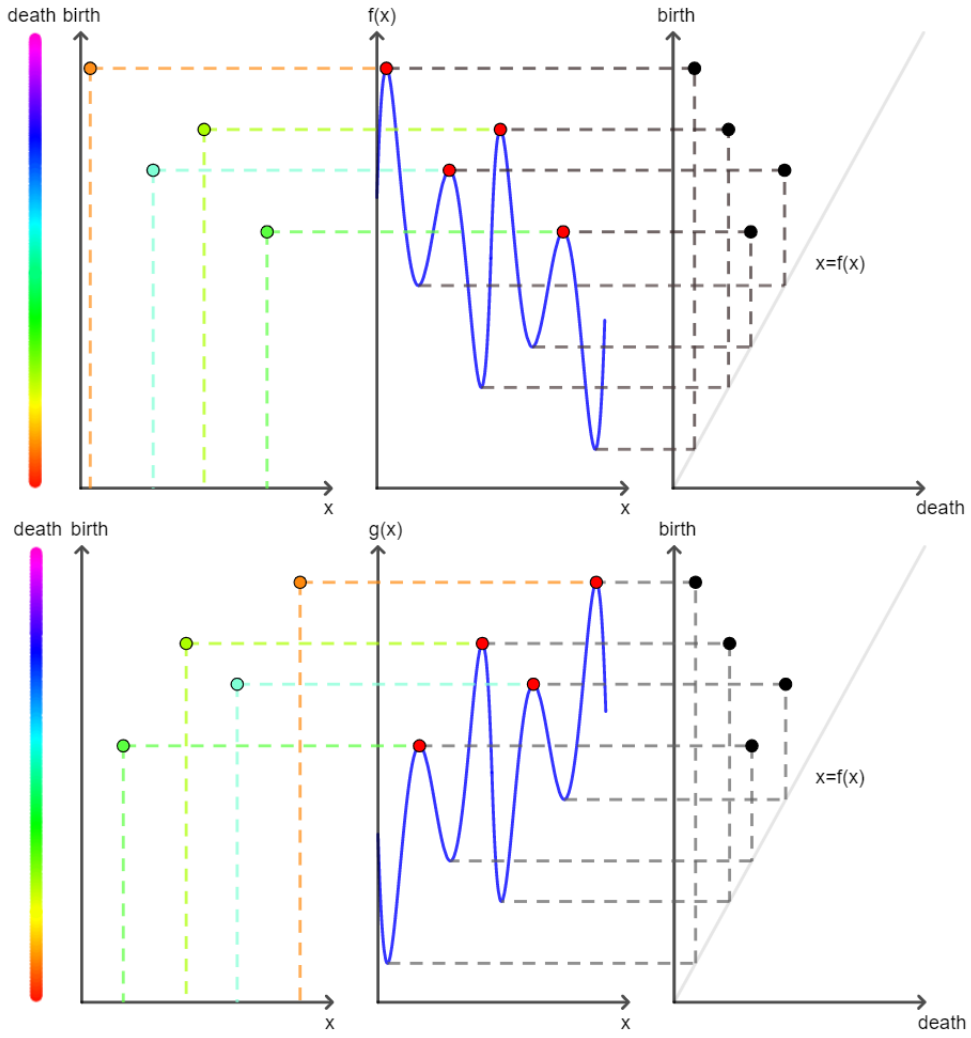


Figure 4.3: The persistence transformation in comparison to the persistence diagram. **Mid**: A real-valued function $f : M \to \mathbb{R}$ (top row) and its mirror equivalent $g : M \to \mathbb{R}$ (bottom row) are illustrated, with $x \in M$ on the x-axis and $y \in \mathbb{R}$ on the y-axis. **Left hand:** The persistence transformation of the functions f and g are shown with x values on the x-axis and birth values on the y-axis. The color coding of the points corresponds to the third dimension, i.e., the magnitude of death values. The features (i.e., the points) of the functions f and g can be distinguished clearly. **Right hand**: The persistence diagram of the upper-level set filtration of the functions f and g is shown with the birth values on the y-axis and the death values on the x-axis. In contrast to the persistence transformation, the persistence diagrams are identical.

4.2.6 Implementation and Analysis of the Algorithm

To analyze the MALDI dataset, we implemented a custom-made computer algorithm for the reduced persistence transformation. The recursive algorithm is based on pairing peaks with their unique local minimum (cf. Equation (4.2)) to determine their persistence. The pseudo code with a detailed analysis by run-time and storage usage is given in Appendix A.1, proving the following theorem:

Theorem 1 *The algorithm always terminates and has a complexity of $\sigma(q) + \sigma(m^2)$, where m is the number of peaks for each spectrum and returns all features with their persistence.*

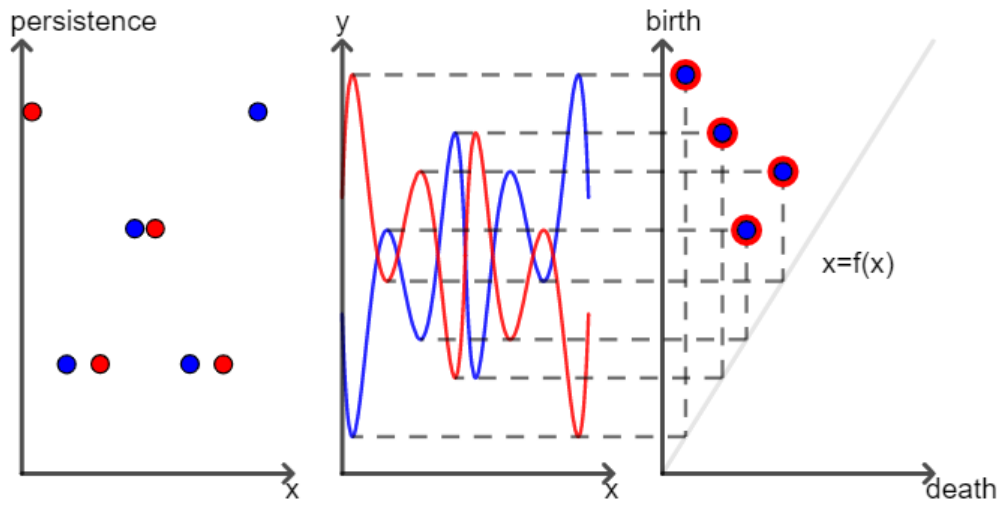


Figure 4.4: The reduced persistence transformation in comparison to the persistence diagram. **Mid**: There are two real-valued functions $f : M \to \mathbb{R}$ (blue) and $g : M \to \mathbb{R}$ (red) plotted with $x \in M$ on the x-axis and $y \in \mathbb{R}$ on the y-axis. **Left hand**: The persistence transformation of the functions f (blue) and g (red) are displayed with the x values on the x-axis and the persistence values on the y-axis. The features of g are distinct from the features of f (blue). **Right hand**: The persistence diagram of the functions f (blue) and g (red) are shown. The y-axis indicates the birth values, while the x-axis shows the death values. The features of g (red) cannot be distinguished from the features of f (blue).

Note that the implementation of the algorithm stores for each feature the tuple $(x, p(x))$, where $p(x) = a^* - a^+$ is the persistence of the feature. Without the further cost of calculation, the algorithm could calculate the persistence transformation instead of the reduced persistence transformation by storing the triple (x, a^*, a^+). Furthermore, the algorithm could be adjusted to determine the persistence diagram of the upper-level set filtration instead by storing the tuple (a^*, a^+).

By considering only the peak information while tracking the position of the peaks, the storage space can be compressed to $2 \cdot m$, i.e., the x value and the p value.

4.3 Supervised Methods

The second step of the proposed methodology is to carry out a classification method on the resulting persistences to classify observational units into class labels, i.e., LC subtypes. To do so, we consider two classifiers, logistic regression (LR) and random forest (RF). Note that our goal is to investigate the performance of the proposed topological framework in the context of MALDI modeling, not a benchmark study that compares RF versus LR. For benchmark studies, see, e.g., in [24, 56].

4.3.1 Logistic Regression

Throughout the remainder, we denote the topologically transformed matrix by $Z = (z_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le q}}$, where the entry z_{ij} corresponds to a persistence value for the i-th mass spectrum at the j-th m/z value. Let Y be a (random)

binary outcome variable, meaning that it takes its values in $\{0, 1\}$, in our application, describing two LC subtypes. Further, we denote by Z_j the j-th persistence vector corresponding to its j-th m/z value alternative. The aim of the logistic regression is to model and estimate the effects of the available covariates on the conditional probability, $\pi_i = P(Y_i = 1|Z_{i1}, \ldots, Z_{iq})$ for the outcome variable $(Y_i)_{1\leq i\leq n}$ and the numerical realizations of the covariates $(Z_{ij})_{\substack{1\leq i\leq n \\ 1\leq j\leq q}}$. In this setting, the observations $(Y_i, Z_{ij})_{1\leq j\leq q}$ are assumed to be independent and identically distributed for all $i \in \{1, \ldots, n\}$. LR models combine the probability π_i with the linear predictor η_i via a "structural" (functional) component given in the linear form $\pi_i = h(\eta_i) = h(\beta_0 + \beta_1 Z_{i1} + \cdots + \beta_q Z_{iq})$ (for more details, see Section 2 in [35]). In this study, we consider the logit (canonical) link function. Then, the logistic response function is given by

$$\pi_i = h(\eta_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}.$$

Correspondingly, the logit link function can be expressed as

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 Z_{i1} + \cdots + \beta_q Z_{iq} = Z_i^T \beta. \tag{4.3}$$

Here, $\beta := (\beta_0, \beta_1, \ldots, \beta_q)^T$ and $Z_i := (1, Z_{i1}, \ldots, Z_{iq})^T$, where the first coordinate corresponds to an intercept term and for all $i \in \{1, \ldots, n\}$. The unknown parameters are (usually) estimated by the maximum (log-) likelihood principle. The log-likelihood function related to model (4.3) is expressed by

$$l(\beta) = \sum_{i=1}^{n} \Big\{ Y_i \left[\log(\pi_i) - \log(1 - \pi_i)\right] + \log(1 - \pi_i) \Big\}. \tag{4.4}$$

For the aforementioned (logit) model, by plugging

$$\pi_i = \frac{\exp(Z_i^T \beta)}{1 + \exp(Z_i^T \beta)} \quad \text{along with} \quad 1 - \pi_i = \frac{1}{1 + \exp(Z_i^T \beta)}$$

in (4.4), it yields that

$$l(\beta) = \sum_{i=1}^{n} \Big\{ Y_i (Z_i^T \beta) - \log\left(1 + \exp(Z_i^T \beta)\right) \Big\}. \tag{4.5}$$

Finally, the probability that an unobserved (new) observation is assigned to class 1 is estimated by substituting $(\beta_0, \ldots, \beta_q)$ by their fitted counterparts and Z's by their numerical realizations for the considered new observation in the conditional $P(Y = 1|Z_0, \ldots, Z_q) = \frac{\exp(Z^T \beta)}{1+\exp(Z^T \beta)}$, where we have $q + 1$ covariates since the first coordinate corresponds to the intercept term. Respectively, the new observation is assigned to class $Y = 1$ if the conditional probability, $P(Y = 1) > c$, is greater than a pre-specified threshold c, and oppositely to class $Y = 0$. In this study, we set $c = 0.5$ – a commonly used threshold (cf. [24]). To obtain the numerical results, we adopted the "LogisticRegression()" function in *scikit-learn, v. 1.2.1* [81] (with no penalty) to obtain our numerical results.

4.3.2 Random Forest

The RF algorithm has become an established non-parametric procedure for regression and classification tasks. It has been broadly used in various scientific disciplines ([78, 75, 98]), including subtyping of lung cancer [72]. RF was originally introduced by Leo Breiman in [17] and it presents an "ensemble learning" approach constituting the aggregation of a collection of a great number of decision trees ([53]). RF takes advantage of numerous decision trees, which leads to a reduction of empirical variance in comparison to a single (decision) tree and significant enhancements in its prediction accuracy ([56]). RF utilizes decision trees in order to calculate the majority votes in the leaf nodes when deciding a class label for each observational unit ([56, 17]). In essence, RF consists of two steps. The first step is to build an RF tree. The following step is to classify the data on the basis of an RF tree that has been generated in the first step. For more details, e.g., see in [65, 75].

In this study, we employ the original variant of RF (see [17]), where each tree of the RF algorithm is constructed on a bootstrap sample drawn arbitrarily from the data by employing the classification and regression trees method and minimizes the Gini impurity (GI) regarding the splitting criterion. When constructing each tree (for each split), solely a pre-specified number of randomly selected (data) covariates are deemed as candidates for splitting.

An important step when using RF is selecting hyperparameters, also called tuning parameters. Their values have to be optimized attentively since the optimal quantities depend on the data at hand. An essential concept regarding tuning optimization is "overfitting". In other words, tuning parameters related to complex rules be inclined to "overfit" the training data. As a result, they produce prediction rules overly specific to the training data, performing well for that (training) data but potentially underperforming when applied to independent data. As discussed in [86], the choice of less-than-optimal parameter quantities can be (at least) partially prevented by utilizing a test set or cross-validation (CV) procedures for tuning. However, it is out of the scope of this study to identify the (most) optimal tuning parameters in the context of MALDI modeling. Instead, we are predominantly interested in evaluating the performance of the proposed TDA approach. To this end, we select the "typical default values" for the RF algorithm, as listed in *Table 1* in [86]. Specifically, we set the tuning parameters in our numerical experiments as follows: the number of trees equals 1000. The number of drawn candidate variables per split is equal to $\sqrt{q}$ (often referred to as "mtry", *max_features* in *scikit-learn*). The splitting criterion in the nodes is the Gini impurity. The minimum number of samples in a terminal node is equal to one (*min_samples_leaf* in *scikit-learn*). Regarding the sampling scheme, the number of observational units that are (randomly) drawn for training each tree is determined by the sample size parameter. The default value corresponds to n (i.e., to the overall number of data samples). Respectively, observational units are drawn with replacement when generating each tree. The seed for all experiments was set to 1234.

We carried out the RF approach on the basis of the resulting algebraic vectors $Z := (Z_{i1}, \cdots, Z_{iq})^T$ and the binary outcome Y_i for all $i \in \{1, \ldots, n\}$. Notice that a unit vector has not been considered, i.e., without an intercept

term, as in the case of LR. We adopted the function *'RandomForestClassifier()'* in [81] (*version 1.2.1*) to yield the numerical results.

4.4 Real Data Analysis

4.4.1 Description of the MALDI-MSI data

Here, we present the empirical results obtained by applying multiple classification schemes to MALDI-MSI data based on different levels of persistence extraction. Several studies have previously analyzed this dataset (cf. [16, 9, 63, 59, 115]). We refer to [16, 59, 63] for an in-depth description of the aforementioned dataset concerning its acquisition protocols, tissue sections, tissue blocks, etc. Here, we provide only a brief outline of the dataset.

Cylindrical tissue cores (CTCs) of non-small cell lung cancer were taken from 304 patients. Specifically, 168 patients were associated with primary lung adenocarcinoma (ADC), while 136 patients were associated with primary squamous cell carcinoma (SqCC). CTCs of all patients were gathered into eight tissue microarray (TMA) blocks (for descriptive statistics, see Table 4.1). As discussed in [63], the tumor status and subtyping for all CTCs were affirmed by standard "histopathological examination". Furthermore, this dataset has been generated only on annotated subregions called regions-of-interest (cf. [9]), i.e., subregions comprising only tumor cells.

For illustrative purposes, Figure 4.5 depicts an example of an output from a MALDI experiment taken from a single spatial location in the provided tissue. In the left panel of Figure 4.5, the m/z values are illustrated on the x-axis, while the intensity values of "ionizable molecules" are charted on the vertical axis. This spatial information can be used in two directions, namely, for the determination of the subtyping (i.e., the cancer subtypes) or the identification of the source of the tumor in tissue. The right panel of Figure 4.5 illustrates the data granularity following one of the data-processing steps, i.e., the spectral filtering. Namely, the latter means that m/z values were centered around their expected peptide masses (for more details, see [63] and the references therein). Other data-processing steps applied to this dataset were baseline correction and total ion count (TIC) normalization.

As pointed out in different studies (e.g., [107, 60]), LC is the primary cause of cancer-related fatalities globally; for example, there were 1.59 million reported deaths in 2012 (see [60]). Two major LC categories are recognized, i.e., small cell lung cancer (SCLC) and non-small cell lung cancer (NSLC). The latter constitutes approximately 85% of all LC cases as reported. The two prevailing NSLC entities are ADC and SqCC, compromising approx. 50% and 40% of all lung-related cancers, respectively ([107, 60]). As discussed in [63], the distinction between these two common subtypes is of great importance for the therapy choice of patients.

The used dataset can be found on Gitlab, as provided in [63]. Note we did not apply any further data-processing steps to this dataset. Specifically, this dataset contains $n = 4669$ (observational units, the number of mass spectra), and the number of m/z values is $q = 1699$.

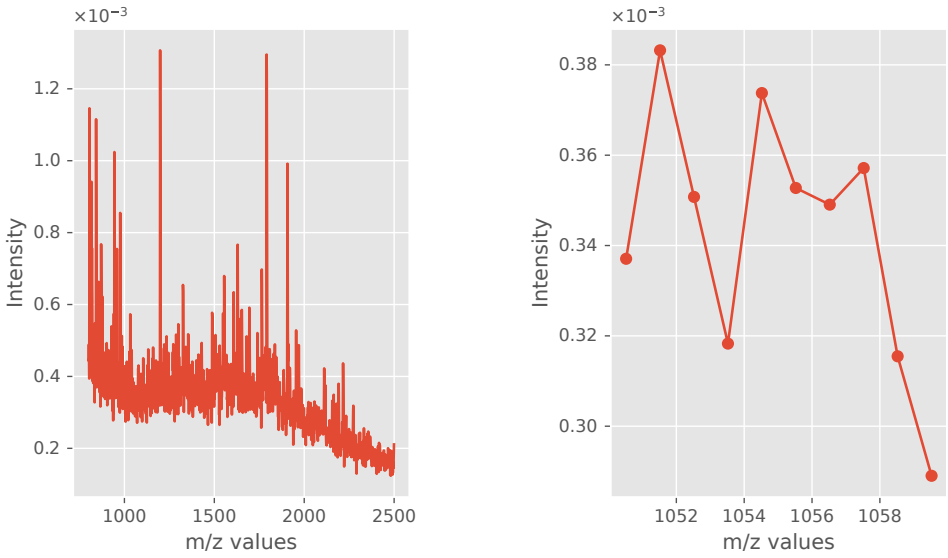


Figure 4.5: An example of a mass spectrum. **Left hand:** An example of a mass spectrum from a single cancerous spot within a patient tissue. **Right hand:** A closer look at spectral data following spectral filtering.

Table 4.1: Descriptive statistics for each TMA

TMA	Number of spectra	Ratio ADC: SqCC
TMA_1	680	59.3 %
TMA_2	437	60.4 %
TMA_3	563	41.2 %
TMA_4	601	49.3 %
TMA_5	512	63.5 %
TMA_6	650	76.8 %
TMA_7	536	50.4 %
TMA_8	690	54.3 %

4.4.2 Classification evaluation

To evaluate the performance of the classification schemes, we mimic the realistic scenario proposed in [63]. Namely, the data is split into training and test sets. The upcoming results were derived by performing k-fold cross-validation (CV) on a TMA level. We followed both scenarios as proposed in [63], specifically 8-fold CV and 2-fold CV. Regarding the 8-fold CV, eight distinct test subsets were created based on each TMA from the overall set of eight TMA blocks. Then, in each of the eight CV folds, each classification scheme was applied to seven TMAs and predicted on the remaining test set —not considered in the training process. Likewise, we carried out 2-fold CV, creating two subsets $A := \{TMA_1, \ldots, TMA_4\}$ and $B := \{TMA_5, \ldots, TMA_8\}$. We reported the obtained results on the basis of all test sets for the 2-fold and 8-fold CVs.

The classification accuracy was assessed by computing the balanced accuracy. The latter metric is computed as the average proportions of correctly classified spectra for each class separately. As a result, this metric is indepen-

dent with respect to imbalanced binary categories, i.e., when one of the target classes appears far more often than the other in the test set. To illustrate the numerical performances appertaining to the persistence transformation, we set a tuning parameter k based on different percentages of peaks extraction.

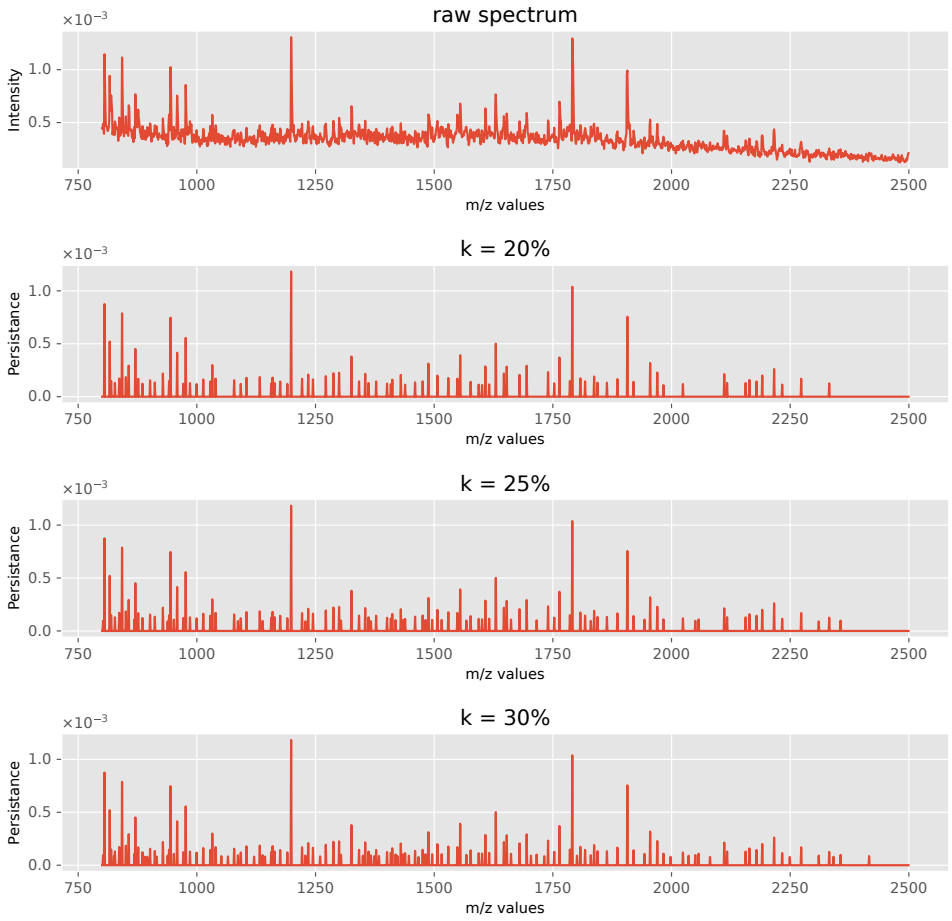


Figure 4.6: Application of the reduced persistence transformation. A visualizing inspection of the proposed topological framework. The top row subplot depicts the raw spectrum, while the following subplots illustrate simplified representations of this raw spectrum given different percentages of peak extraction (abbreviated with $k\%$). Namely, the raw spectrum is transformed to sparse tuples based on a pre-specified k level of peaks extraction.

4.4.3 Data-analysis results

Figure 4.6 depicts an example of the proposed topological feature extraction. Namely, the top row illustrates an example of a raw spectrum (cf. Figure 4.5), whereas the next subplots depict extracted persistence values. Apart from the first row, the m/z values are plotted on the horizontal axes, while the vertical axes show the derived persistence values, not the intensity values as in the first row of the figure.

Tables 4.2 - 4.3 summarize the classification results obtained by applying LR and RF classifiers based on several levels of the extracted persistence vectors. The percentage of extracted persistence employed for the evaluation, denoted by k, ranges in a grid of pre-specified levels —$k \in \{10\%, 20\%, 25\%, 30\%, 40\%, 50\%\}$. By compressing the

Statistic	k = 10%	k = 20%	k = 25%	k = 30%	k = 40%	k = 50%
mean	0.826	0.845	0.859	0.866	0.859	0.859
min	0.722	0.716	0.724	0.743	0.737	0.723
max	0.917	0.937	0.931	0.950	0.935	0.935
median	0.827	0.869	0.886	0.888	0.880	0.878
std	0.074	0.079	0.077	0.076	0.076	0.078

Table 4.2: Comparison table of 8-fold cross-validation. The table tabulates the obtained balanced accuracy results given different percentages of peaks extraction and non-extracted data based on the LR classifier.

Statistic	k = 10%	k = 20%	k = 25%	k = 30%	k = 40%	k = 50%
mean	0.831	0.863	0.871	0.878	0.877	0.868
min	0.702	0.734	0.760	0.774	0.776	0.746
max	0.936	0.945	0.940	0.959	0.930	0.949
median	0.849	0.873	0.880	0.892	0.900	0.890
std	0.083	0.076	0.066	0.067	0.061	0.079

Table 4.3: Comparison table of 8-fold cross-validation. The table tabulates the obtained balanced accuracy results given different percentages of peaks extraction and non-extracted data based on the RF classifier.

raw data with respect to different k, we observed a significant gain in the computational time for executing RF. For example, to execute the 8-fold CV task on 7 CPUs on a standard machine, it takes 35 seconds when using $k = 5\%$ and 70 seconds when using $k = 50\%$, in contrary to the 215 seconds it takes using the raw data, i.e., the original variant where each data entry corresponds to an intensity value.

To put our results in the context of other competitors for this dataset, we performed a comparison with a popular method for retrieving informative parts of the spectral data and executing automated cancer (sub-)typing. Briefly put, in [63], the authors proposed novel supervised non-negative matrix factorization methods (NMF): the classification tasks are executed in parallel to feature extractions (in the context of NMF extraction), which differs from the more classical NMF-related scenario (cf. [16]). The authors introduced 13 distinct classification schemes. From these, we selected the top 2 competitors; for the remaining ones, we refer interested readers to Figure 3 and Figure 4 in [63]. These top 2 competitors from [63] are $Flog\_int$ and $Flog\_log$, where the number of NMF "features" is 60, as suggested in the provided code. Our competitors are the persistence transformation where k is either 30% or 40% and the FR classifier. These schemes are abbreviated to $PT\_RF\_30\%$ and $PT\_RF\_40\%$, respectively.

Table 4.4 illustrates that $Flog\_int$ and $Flog\_log$ perform slightly better than our best performers for this dataset for the 8-fold CV task and 2-fold Train B. However, the topological-based competitors outperformed the other 11 NMF-based schemes for this dataset, even in some scenarios $PT\_RF\_40\%$ produces numerically similar results vis-à-vis all NMF-based competitors, cf. Bal. Acc. (2-fold) Train A. The authors [63] concluded that apart from the

Method	Avg. Bal. Acc. (8-Fold)	Bal. Acc. (2-Fold) Train A	Bal. Acc. (2-Fold) Train B
PT_RF_30%	87.8% ± 6.70	90.75%	84.40%
PT_RF_40%	87.7% ± 6.01	91.15%	83.53%
Flog_int	89.8% ± 4.35	90.8%	89.1%
Flog_log	88.8% ± 6.36	91.1%	87.1%

Table 4.4: Performance of the proposed classification algorithms vis-à-vis the best classification competitors from [63]. *PT_RF_k%* stands for Persistence Transformation, using the Random Forest classifier with k as a hyper-parameter. According to [63], Flog stands for the Frobenius norm utilizing the logistic regression classifier. The suffix "*_int*" denotes the Integrated approach, while "_log" stands for the Optimized approach.

top three classification schemes, most of the other methods achieved, on average, balanced accuracy values below 80%. This table stands for the proof of concept that our topological framework accompanying the RF classifier can produce competitive results. Moreover, the RF (non-linear) algorithm can operate in pure high-dimensional scenarios, i.e., when covariates exceed the observational units $n \ll q$. Therefore, the proposed classification scheme $PT + RF$ can also be applied in different data regimes.

4.5 Image Denoising with Persistence Transformation

4.5.1 Simulation setup

A (big) challenge in examining real-world applications is the presence of noise that can corrupt the data and lead to incorrect data-analysis results (see [66, 61]). To this end, researchers have to be careful when analyzing datasets with a possibility of noise and address it appropriately to improve the accuracy of the results (see [111, 115, 114]).

To assess the effectiveness of the proposed topological framework given the presence of noise, we proceeded as follows. First, we simulated multiple synthetic mass spectrometry (MS) images, where each pixel of these images corresponds to a unique mass spectrum. Specifically, each pixel presents an average value for its respective mass spectrum. Second, we artificially contaminated the spectral data that generated the MS images by adding different types and levels of noise. Finally, for each figure, we plotted the ground truth, the noise image, and two variants of denoised MS images in a row. As a result, one can pictorially identify the ability of our TDA approach to differentiate signal from noise in the images. Two of these results are displayed in Figure 4.7 and in Figure 4.9.

We utilized the *"Cardinal"* package ([10] *v. 3.0.1*) in *R* ([87] to simulate noiseless (ground truth) MS images. Accordingly, we employed the "SimulateImage()" function with the following parameters: the preset image is two (i.e., there are two figures, a circle in the top-left corner and a square in the bottom-right corner), the m/z range lies in $500 - 2000$ (resulting in 3466 m/z values), the number of peaks $k^* = 50$, and a noiseless image (i.e., "sdnoise" equals to zero). We aim to demonstrate the efficacy of our methodology with varying baseline levels so as to cover baseline value ranges $\in \{0, 5, 15\}$. Based on these parameters, we simulated multiple MS images with sizes $\{30 \times 30, 42 \times 42, 60 \times 60\}$. Following the simulation of the ground truth images, we contaminated the spectral data

by adding either Gaussian or Poisson noise. We chose increasing values for the standard deviation for the Gaussian noise [1] and λ for the Poisson noise, i.e., artificially contaminated synthetic data more and more. These distribution parameters are given in the caption of each subplot and can serve as a proxy for different signal-to-noise ratio variants. Due to its relevance to our real-world data application, we proceeded by picking a percentage of the most significant peaks – signals outside these fractions were considered noise.

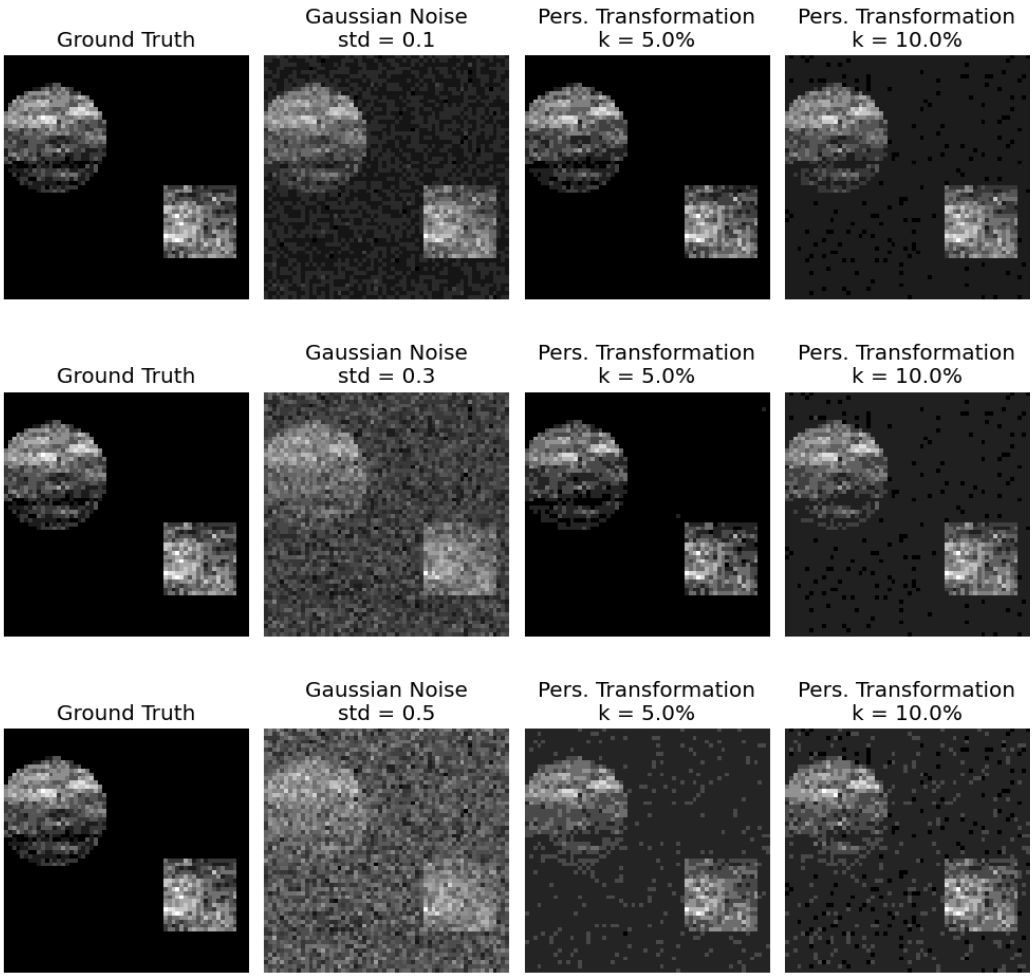


Figure 4.7: Denoising with the persistence transformation. On the leftmost column, the ground truths of synthetic MALDI-Images are displayed. In the second column, distinct levels of Gaussian noise are added to the spectral data. The processed images based on two choices of k are displayed in the third and fourth columns.

Adding artificial noise to the MALDI spectra has two significant effects. First, the height of the existing signal peaks could be altered. Second, new noise peaks can be established. For low levels of noise, the added noise peaks are less persistent than the signal peaks. By optimizing the tuning parameter k, our algorithm can differentiate the bulk of the signal peaks by neglecting the noisy ones in spectral data. A good example of low-level noise is the Gaussian noise, which is displayed in Figure 4.7 and Figure 4.8 as can be seen in Figure 4.8, most of the signal peaks can be distinguished from the noise peaks by their height. Hence, the original shapes from the ground

[1] with a fixed location parameter of 0.1

truth images can be reconstructed in the denoised images in Figure 4.7 when the tuning parameter k is chosen accordingly.

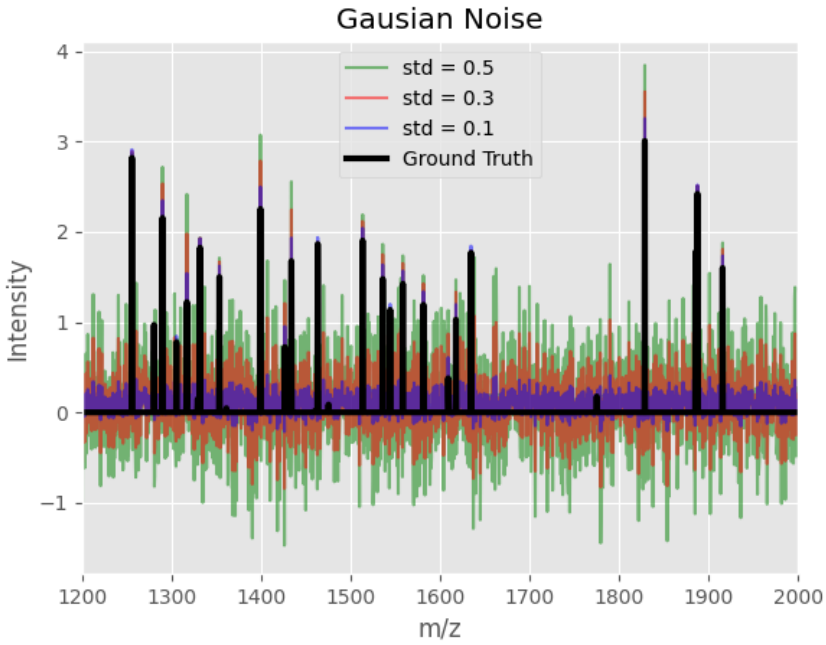


Figure 4.8: An example of a synthetic spectrum of the images in Figure 4.7. The ground truth spectrum is displayed in black, and its different noisy counterparts are displayed in color.

Higher levels of noise, on the other hand, can be challenging for the persistence transformation. By adding noise peaks larger than the signal peaks, the ground truth can be compromised so that the persistence transformation can not reconstruct the original shapes. An example of such kind of noise is the Poisson noise, and it is displayed in Figure 4.9 and Figure 4.10. In Figure 4.10, it is shown that most of the noise spectra exceed the signal peaks. Even so, with an appropriate k value, the shapes of the ground truth can be reconstructed to some degree in Figure 4.9.

More simulation results can be found in Appendix A.2. We want to highlight the time used for the (denoising) analysis. It can be seen that doubling the number of pixels results in doubled time for the algorithm, e.g., for Gaussian noise with a standard deviation of 0.1, the analysis of a 30×30 image takes approx. 29 seconds, for a 42×42 image it takes approx. 61 seconds, and for a 60×60 image it takes approx. 113 seconds on a standard computer. This illustrates the almost linearity of the implementation.

4.6 Discussion

4.6.1 Summary

Motivated by the MALDI classification studies, the objective of this study has been to propose a novel custom-made approach for modeling MALDI-MSI data. In general, the study's methodology consists of two steps. First,

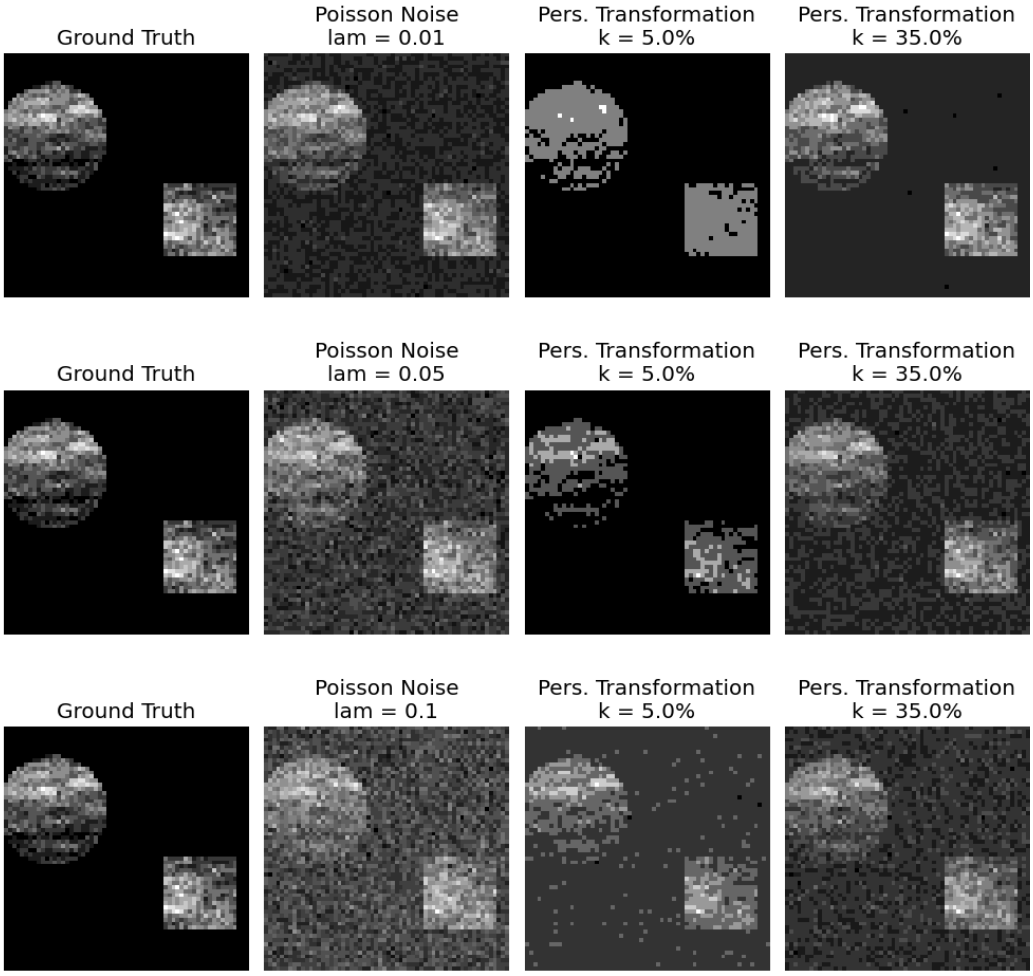


Figure 4.9: Denoising with the persistence transformation. On the leftmost column, the ground truths of synthetic MALDI-Images are displayed. In the second column, distinct levels of Poisson noise are added to the spectral data. The processed images based on two choices of k are displayed in the third and fourth columns.

we carry out the introduced topological framework to obtain the intrinsic information from each mass spectrum, given thousands of m/z values. Generally speaking, this step can be considered as a data-compression method for MALDI-MSI data. Second, we execute two supervised classification methods based on the resulting persistence vectors so as to classify the observational units into lung cancer subtypes.

The usefulness of the proposed topological framework consists of three perspectives. First, our numerical classification results illustrate that the topological framework extracts the necessary information, which can be used for further classification tasks. The obtained results are competitive with other data-analysis methods for this dataset (cf. [63]). Second, the proposed framework compresses MALDI-MSI data, resulting in a significant computational gain for the RF classifier. Third, we have demonstrated its effectiveness in retrieving the informative parts of spectral signals under different noisy scenarios. The proposed topological framework can be adopted in a computationally efficient algorithm depending on a single tuning parameter, i.e., the fraction of used peaks.

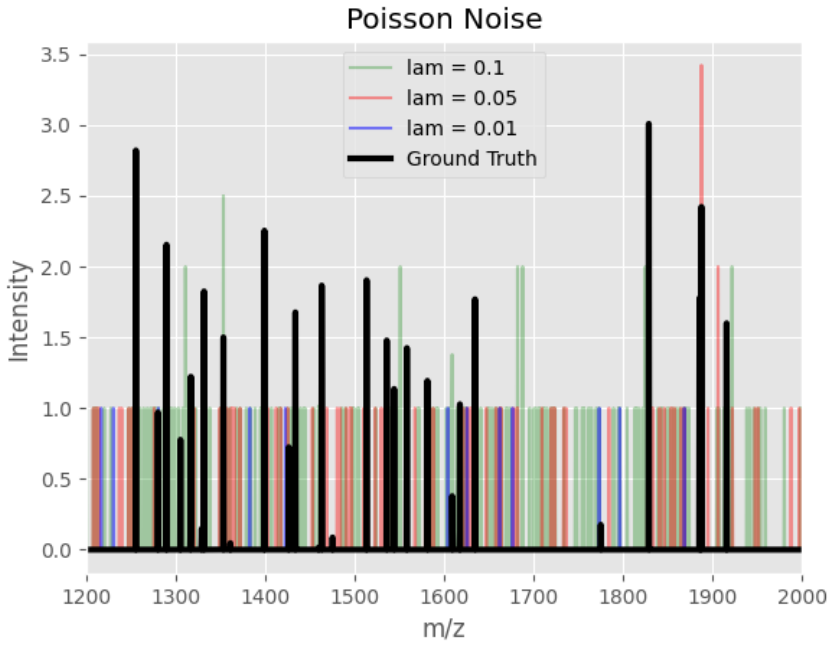


Figure 4.10: An example of a synthetic spectrum of the images in Figure 4.9. The ground truth spectrum is displayed in black, and its different noisy counterparts are displayed in color.

4.6.2 Outlook

The persistence transformation is a novel tool for topological data analysis, which can be employed in different real-world applications. Future work to extend the introduced framework might include the application of the heat equation to reduce the impact of noisy peaks (see [117]), which might increase the classification performance. Alternatively, one can use the persistence pairs as input points for a second persistence homology analysis. However, any computational improvement would lead to the extension of the computational time. Furthermore, the information on the position of the peaks might be used for backtracking to identify (biologically) relevant molecules given as peaks. Finally, the algorithm used in this paper shows similarity to Morse Theory ([69]) and especially to the matching theorems in [58]. This presents an interesting topic for follow-up research.

Chapter 5

Discussion

This thesis has proposed novel approaches to address two challenging tasks in analyzing high-dimensional datasets, particularly MALDI-related data: feature selection and tumor classification. This chapter is structured as follows. First, we draw conclusions for each chapter in the thesis. The following subsection outlines the contributions of this thesis. Next, we discuss computational challenges in the context of MALDI and how different pre-processing steps affect our frameworks. Lastly, we provide an outlook for future research to build upon the contributions of this thesis.

5.1 Conclusions

Regarding our inferential methodologies, which are presented in Chapters 2 - 3, we can conclude the importance of incorporating correlation effects in the decision process (i.e., when deciding whether to reject or retain a null hypothesis), especially for datasets generated under strong temporal or spatial conditions, such as MALDI-related datasets. We have illustrated this importance by conducting computer simulations under different data-generating scenarios and real data applications. Referring to the latter, as depicted in Figure 2.2 and Figure 3.2, strong dependencies stemming from the (spatial) data acquisition processes impact the theoretical null distribution, leading to classical FDR procedures being either too liberal or too conservative, in our cases being extremely liberal. In Chapter 2, the dataset utilized for this study exclusively comprises cancerous cells measured from multiple spots from each patient, resulting in highly related observational units. We found it convenient (but not essential) to "empirically correct" the null distribution (for further details, we refer to Subsection 2.4.2) due to the strong dependencies among the Z-statistics in order to diminish the variance inflation. However, even following this "empirical correction," the empirical Z-values do not resemble the density of the standard normal distribution (i.e., the theoretical null distribution) on $\mathbb{R}$; however, as illustrated, it is not a challenge for our (inferential) procedures.

With respect to our topological framework, we propose exploiting the peak-related information by means

of persistence transformation and utilizing this information for further classification tasks. We have illustrated the effectiveness of our approach on a (real) MALDI dataset and compared the obtained classification results with those of competitive frameworks. Additionally, through simulations on synthetic mass spectrometry images and artificially contaminating them, we have demonstrated the capacity of our topological framework to retrieve intrinsic information from spectral signals while disregarding noise-associated fluctuations. All in all, it leads us to conclude that utilizing the (peaks) geometry structure is viable for MALDI-related applications. It is important to mention that a central assumption of our topological methodology is the existence of peaks and their importance for considered applications. Therefore, our topological approach is limited to applications in which this property holds. An anonymous reviewer has commented that our topological framework could be applicable not only to MALDI-MSI applications but also to desorption electrospray ionization (DESI) applications and secondary ion mass spectrometry (SIMS) applications (for more details about DESI and SIMS, we refer to [90]). We have not (yet) examined such datasets to delve deeper into these potential applications. Nevertheless, the feasibility of applying our framework to DESI and SIMS datasets is worth considering for future investigations.

5.2 Contributions

Motivated by addressing the challenge of discovering associative m/z values in the context of MALDI, the starting point for our research was proposing inferential procedures capable of considering the following properties in the analysis of MALDI-related data: the strong spatial dependencies, a sparse assumption in the sense only a small number of m/z values (synonymously covariates or features) are expected to be associative, high-dimensional settings (i.e., containing a great number of covariates), and binary or nominal target variables (describing cancer classes).

Considering all the aforementioned characteristics, from the statistical perspective, our primary contribution in Chapter 2 is to propose a novel procedure for two-sample comparisons. Briefly put, by utilizing the framework of multiple marginal logistic regression models and making use of the principal factor approximation estimator (cf. [38]), our methodology incorporates the correlation matrix of the test statistics. We propose to employ the multiple marginal models (MMM) framework ([83]), which provides us the capability to approximate the (limiting) covariance matrix under different data regimes, including situations where the number of features exceeds the number of observations. Respectively, our inferential procedure can work in high-dimensional setups, viz., when the number of covariates exceeds the number of observations by magnitude.

The primary contribution of Chapter 3 consists of presenting a procedure for analyzing datasets with categorical outcome variables in high-dimensional scenarios. This methodology (effectively) extends our prior work in Chapter 2 to a multivariate context. From the statistical perspective, we employ multinomial regression and adapt the necessary adjustments to tailor the MMM approach in lieu of binary logistic regression. It would, in principle,

be possible to fit two times our procedure for a two-sample comparison, given some appropriate modifications (i.e., re-coding of the outcome variable and filtering out the observational units for the respective pairing), in the case of three nominal categories. Yet it is nevertheless advantageous to employ our multivariate version. Specifically, this approach allows for utilizing all (i.e., n) observational units in each marginal fit; otherwise, one would have to exclude a part of the observational units for each of the two separate runs of the methodology presented in Chapter 3. So, there is a computational benefit since one can analyze datasets with categorical outcome variables without a need to carry out the aforementioned data manipulations. Our implementation in R ([87]) also works in the case the (random) outcome variable is not only a vector, but it might be a matrix of random counts, and our generic function works with more than three classes. Lastly, the class of models that can be taken into consideration under the multivariate setup is richer compared to the class of the univariate models as in Chapter 2.

From the application point of view, we have carried out our methodologies on two (real MALDI) datasets given multiple mass spectra per patient and solely a single (mass) spectrum per patient, in this way illustrating the versatility of our inferential procedures. Our procedures' results are in accordance with the already published findings by other researchers for the used datasets.

Now, let us further outline our contributions in the context of MALDI applications. For obtaining their "baseline method", the authors ([9]) utilized the Mann–Whitney–Wilcoxon test to discover relevant features. Then they chose "the highest test statistic" in a range of 5 to 100 features, while our approaches contribute by proposing a statically grounded way to identify the number of relevant features under the constraint of FDP control. In [5], the authors employed different multiple testing procedures, including the Benjamini-Hochberg procedure and the Benjamini-Yekutieli (BY) procedure, to perform feature selection based on discrete wavelet coefficients (i.e., not directly on intensity values). By contrast, our procedures operate on intensity values and explicitly take into consideration the (arbitrary and strong correlation) dependencies in the decision process. We have illustrated the importance of employing dependencies in the decision process. In the context of MALDI, it is fair to say that signals appear as significant peaks, and multiple methods have been proposed to discover peaks (e.g., [66, 119]). Since peak detection is an unsupervised process, as highlighted in [16], additional analysis is conducted on detected peaks with the intention of identifying a (relevant) subset of peaks that exhibit statistical significance with respect to the response variables. Our screening frameworks reveal the magnitude of (genuine) Z-values of well-researched biomarkers (signal peaks), isotope peaks, and already reported monoisotopic peaks as highly distinctive for cancer associations (see Table 2.9 and Figure 3.4). As a result, the dependency-adjusted p-values for the aforementioned peaks are exceptionally low. Thus, under the constraint of FDP control, we could declare them as false nulls (i.e., statistically relevant covariates). Additionally, the signs of the Z-statistics provide with the information of being distinguishing for particular cancer (sub-)type, specifically ADC ($Z_j < 0$) and SqCC ($Z_j > 0$), as illustrated in Chapter 2, for instance. These insights provide empirical evidence that our proposed methodologies contribute to automatically identifying statistically (associative) significant peaks, e.g., for the task of biomarker discovery.

Regarding Chapter 4 ([57]), our work mainly contributes in the following direction. We propose an algorithm to calculate the reduced persistence transformation and prove its complexity. Of note, Weise et al. (2020) ([117]) also advocate for the utilization of the persistence transformation (PT) within the context of MALDI. However, as aforementioned, our contribution consists of proposing a novel algorithm to carry out the proposed topological framework, specifically to detect peaks and compute their persistence values. Now, we proceed to outline the contributions from the application point of view.

We have applied our framework to a real MALDI dataset for binary classification. Particularly, we demonstrated the classification performance of our supervised framework at varying levels of topological extraction using this dataset. Also, we compared the resulting classification results with those obtained from competitive frameworks (cf. [63]) based on two different cross-validation schemes. These comparisons serve as validations of our complete tumor (sub-)typing process. Furthermore, to the best of our knowledge, we are the first authors to illustrate the denoising and compression properties of the PT in the context of MALDI. Regarding the former, we have demonstrated this aspect through multiple artificially contaminated images subjected to different levels and types of contamination. We then displayed the resulting denoised images after applying the PT framework (for more pictorial details, see Appendix A.2). As regards the compression property, following applying the PT, we have presented evidence that the running time for the random forest classifier based on compressed data is considerably faster compared to the non-compressed data. Regarding related works, as aforementioned, the first step of our topological framework is to discover all peaks; therefore, this step is closely related to peak detection (e.g., in [66, 119]), which is an active research topic in the context of MALDI. Furthermore, our contribution is that we strive not solely to identify significant peaks but to utilize peak-related information further for tumor classification.

All in all, our topological framework can serve as a valuable instrument for researchers and data curators, as it can be solely applied to raw MALDI data aiming for size compression and denoising (not only for tumor classification) while preserving the necessary information in the context of MALDI.

5.3 Computational challenges

As discussed in [3], MALDI mass spectra (i.e., the observational units) exhibit considerable heterogeneity due to (some) technical reasons, such as noise, ions diffusion, and more. Consequently, an essential step in the analysis of MALDI-related datasets is conducting (some) data-processing steps. These processing steps involve a series of data manipulations aiming to prepare data for further analysis (for more details on mass spectra pre-processing, see [74]). Generally speaking, we can group these pre-processing steps into two categories. The first category aims to increase the comparability among mass spectra, while the second category aims to decrease the number of measured variables. The latter helps reduce the feature space (synonymously, the number of measured variables) and, therefore, the computational time required to analyze the data at hand without losing valuable information.

We refer to [3] for more details regarding computational challenges in the context of MALDI.

A crucial step of pre-processing MALDI-Imaging data involves spectra normalization, which refers to scaling each spectrum up to a specific factor for a more reasonable "intercomparison" of intensity values across different (mass) spectra, facilitating better comparability (for more details, see [3]). As pointed out in [3], a standard method for normalization is total ion count (TIC) normalization. In this method, the TIC, the sum of all intensity values within each spectrum, is computed, then all (spectrum) intensity values are divided by the TIC value, respectively the summation of intensity values within each mass spectrum is 1 (cf. [117]). The author in [3] (see also the references therein) discussed the necessity of normalization. Furthermore, this processing step has also been considered in advanced analysis methods, such as deep learning (cf. [9]). Another pre-processing step that aims to increase comparability is baseline correction. It involves removing or adjusting the baseline for each spectrum to align for fluctuations among intensity values.

Regarding the second group of pre-processing steps, researchers typically perform spectral filtering ([63, 16, 93]). Specifically, a non-processed dataset is resampled (or binned) to specific Dalton (Da) unit intervals, which are centered around expected peptide masses (for more details, see [63]). By applying this step, researchers aim to significantly reduce the number of measured variables and enable more manageable subsequent analyses. Generally speaking, m/z values are binned around a higher Da zoom, for example, 0.4 Da. The non-processed dataset can be quite large, often spanning a couple of gigabytes, and loading such (massive) datasets into RAM memory demands substantial hardware capacity. The authors in [16] examined different combinations of pre-processing steps and how those steps affect their analysis approach. Also, the authors in [74] introduced a data-processing workflow to enhance the extraction of valuable information from mass spectrometry. Another optional step is peak selection, which also aims to reduce the number of m/z values (MALDI-related covariates) to only a subset of relevant peaks. Spectral binning, baseline correction, and peak selection require specialized software, whereas the TIC normalization is straightforward and does not require dedicated software. The R-based package "MALDIquant" ([47]) offers different pre-processing steps and routines for processing and analyzing MALDI-related datasets.

In this thesis, we employed "standard" pre-processing workflows using in-house functions built in Matlab. The influence of different data-processing settings has not been studied throughout this thesis and would require further investigation. Let us elaborate more on the importance and necessitates of the aforementioned processing steps in the context of our frameworks.

Regarding our inferential frameworks, baseline correction and data normalization are essential steps to increase the comparability among mass spectra or, in other words, to diminish the external (extra) variability among the observational units and align their distributions. The spectral binning step is optional, and our frameworks can be applied to different Da granularity. However, analyzing datasets with a significantly small (Da) granularity, e.g., a zoom of 0.08 Da, could result in over 50,000 m/z values. To carry out our procedures, one needs to run a marginal

regression for each m/z value, meaning 50 thousand marginal fits. It is worth mentioning that marginal fits and the multiple marginal models (MMM) approach can be parallelized (since we run them independently) in order to optimize the execution time, and the computational burden can be lessened. However, the subsequent step in our procedures is to execute the eigenvalue decomposition on the covariance matrix of the test statistics, meaning one needs lots of RAM to load this 50 by 50 thousand covariance matrix. To this end, the pre-processing step of binning is good practice. Moreover, it has been proven to be successful in analyzing MALDI-related data (e.g., [63, 93, 16]). Lastly, the peak detection procedure is redundant since our (screening) procedures tend to identify m/z values that are associative with the target variable regardless of whether they are (only) monoisotopic peaks or biomarkers.

In the context of the pre-processing, we observed one benefit of utilizing our topological framework. Specifically, we simulated multiple spectral data with distinct baseline levels (not only around 0; see Section 4.5) to demonstrate its capability to eliminate noise from mass spectrometry images. As illustrated in Chapter 4, our framework successfully retrieved signals from noise, even under this scenario. Furthermore, the authors ([117]) employed the PT framework. The authors argue that the PT pipeline requires only specific intensity normalization (cf. Section 3.1 in ([117])). They illustrated the superiority of the PT framework over "complex conventional pre-processing" on a (real) MALDI dataset for binary classification. In this thesis, we applied our framework to an already processed dataset and compared the classification evaluation with other frameworks on the same pre-processed dataset. To this end, we did not investigate how distinct pre-processing steps impact the classification efficacy of our entire framework. As aforementioned, we introduce an algorithm (see, Appendix A.1) that implements our topological framework. The proposed algorithm's execution time is linearly dependent on the number of m/z values and the number of detected peaks within each spectrum. In other words, a reduced number of m/z values results in a faster running time for our algorithm. As discussed, a feasible way to reduce the number of m/z values is by applying the spectral filtering step; however, this processing step is optional. It is worth noting that our algorithm processes each spectrum individually, enabling the algorithm's parallelization. This means multiple (mass) spectra can be processed simultaneously on multiple workers, leading to faster execution and improved computational efficiency.

5.4 Outlook

Regarding our multiple testing frameworks, several follow-up questions might be interesting. First, we have examined MALDI datasets with relatively balanced categorical classes (binary and nominal). As discussed in [91] (cf. also in [50]), the imbalance in terms of binary outcomes can lead to an inflation of the type I error rates. In this thesis, we have not studied the approximation accuracy of our methodologies in the context of imbalanced categorical, either binary or nominal (target) classes. To this end, we need first to comprehend the performance

of inferential frameworks given the phenomenon of class imbalance. If the FDP control is not that sufficient, I assume, for example, a plausible treatment would be to balance datasets artificially, viz., to correct class imbalance (cf. [48]). I have to admit that a reviewer has stated this intriguing question, but from my perspective, it is excellent and deserves further investigation. Second, it might be captivating to evaluate the uncertainty regarding the realized FDP so as to provide a confidence region for it, in addition to mere point estimation of the FDP ([54]). The objective is to establish an "exceedance control" (cf. [45]) of the FDP estimator, as considered in the previous chapters. As a result, we can construct upper confidence bounds for the FDP, ideally simultaneously over a grid of rejection thresholds. Last but not least, we employ the maximum likelihood estimator (MLE) to estimate the unknown parameter in the proposed methodologies. However, the presence of outliers in data can (significantly) ruin the obtained estimates (cf. [71]). Some outliers in the independent variables can cause parameter estimates to be (arbitrarily) large or small. To this end, we could draw wrong conclusions because of not correctly yielded estimates. A plausible remedy might be by adapting robust alternatives of the MLE, for example, in the spirit of trimmed likelihood estimators (TLE) (for more details, see [71, 73]).

With regards to the algebraic framework, there are two more open questions (from my point of view) apart from those stated in the Outlook Subsection 4.6.2 of Chapter 4. Firstly, we extracted equal persistence information from each spectrum (observational unit) for our (supervised) MALDI data analysis, defined as a k percentage of all (detected) m peaks. Since we define a set M for each spectrum and extract its persistence information individually, it would be beneficial to extract this k percentage more dynamically from each spectrum. Namely, extracting more topological information in the case of class misclassification for the respective observational units. Secondly, we consider a trivial pairing when the topological feature is a low persistent peak (i.e., its relative height is low). Therefore, we assume that this topological feature is noise and encode its value as 0. However, an alternative solution might be to encode trivial pairings as missing values (i.e., NAs) and employ a more sophisticated machine-learning classifier that explicitly deals with missing values.

Finally, this thesis provides an outlook regarding incorporating multiple testing with the persistence transformation. Firstly, I am inspired by this study ([5]) in the context of biomarker discovery in MALDI. The authors employed another compression technique, namely, discrete wavelet transformation, and conducted different multiple testing procedures based on the obtained (discrete) wavelet coefficients at different levels to identify the most distinguishing m/z values for cancer association. A similar idea might be an excellent application within the scope of our topological framework. In simpler terms, one could conduct a multiple testing procedure (and by incorporating dependency effects) based on persistence values (rather than m/z values) at different peak extraction levels, resulting in topological feature selection.

Secondly, we propose to employ "reduced persistence transformation" (see Subsection 4.2.4), where in lieu of $t(x) = (x, a^*, a^+)$ solely the position and the persistence are stored: $\tilde{t}(x) = (x, a^* - a^+) \in M \times \mathbb{R}$ for $x \in M$. However, another interesting application might be simultaneous statistical inference based on pairwise comparisons of "birth"

and "death" values (a^*, a^+) (not their difference) for each m/z value (i.e., at every position for x) across all mass spectra (observational units). The latter application is plausible since our topological algorithm needs to pair all inflation points (i.e., minima and maxima). In other words, the algorithm always estimates $k = 100\%$, and then one can pick a lower percentage for k (i.e., $k < 100\%$) in order to compress the data at hand effectively. Note that the algorithm executes once the process and the post-selection choices of k do not require any further executions. Applying a multiple testing procedure based on the aforementioned (all) pairwise comparisons (for $k = 100\%$) would be beneficial to discover a subset of persistence vectors with statistically significant differences in their relative heights. As pointed out in [3], "[s]electing many peaks slows down the segmentation and can introduce additional variation". Incorporating statistical inference with our peak detection algorithm would allow us to control the selection of a peak-based subset among detected peaks, addressing this concern effectively. Note that such application differs from our inferential procedures (i.e., Chapters 2 - 3). Since the former application tends to answer the question of significant peaks, no associative peaks, meaning this is unsupervised peak detection.

Appendix A

Appendix

A.1 The persistence transformation algorithm

Here, we provide the pseudo-code of the algorithm introduced in Chapter 4. Furthermore, it contains the proof of the Theorem 1.

Algorithm 2: Recursion start

Input: $[f(x_0), \ldots, f(x_{q-1})]$
Return: $[(\hat{x}, p(\hat{x})), \ldots]$
maxima, minima, featurePairs $\leftarrow \varnothing$
for all $x_j \in [x_0, x_{q-1}]$ **do**
  if x_j is maximum **then**
    maxima $\leftarrow (x_j, f(x_j))$
  else if x_j is minimum **then**
    minima $\leftarrow (x_j, f(x_j))$
SORT(maxima, f(x), >)
SORT(minima, f(x), <)
$(\hat{x}, f(\hat{x})) \leftarrow$ maxima.pop(0)
featurePairs $\leftarrow (\hat{x}, f(\hat{x}) -$ minima[0][1])
RecursionStep(x_0, $\hat{x}$, maxima.copy(), minima.copy(), featurePairs)
RecursionStep(x_n, $\hat{x}$, maxima.copy(), minima.copy(), featurePairs)
return featurePairs

Proof 1 (Proof of Theorem 1:) *The algorithm to calculate the reduced persistence transformation is divided into two parts: the recursion start (Algorithm 2) and the recursion step (Algorithm 3).*

The recursion start (Algorithm 2) gets as input the list of all the intensity values for each m/z value, marked as $f(x_j)$. Let m be the number of maxima in this mass spectrum. In the beginning, empty lists are created for the maxima, the minima, and the results (called "featurePairs"). The latter list stores the x value, i.e., the position, as well as the persistence of each peak. Notice that each recursion step updates the list instead of returning results.

In the next step, all the minima and the maxima are stored in the corresponding lists in tuples of the form $(x, f(x))$. For this, the algorithm iterates through the list of the $f(x_j)$. If the value is larger than its neighbors, it

Algorithm 3: Recursion Step

Input: start, end, maxima, minima, featurePairs
for all $(x_j, f(x_j)) \in$ maxima **do**
  if $x_j \notin$ [start, end) **then**
    maxima $\leftarrow$ maxima$\backslash(x_j, f(x_j))$
if |maxima| = 0 **then**
  return
$(\hat{x}, f(\hat{x})) \leftarrow$ maxima.pop(0)
RecursionStep(start, $\hat{x}$, maxima.copy(), minima.copy(), featurePairs)
for all $(x_j, f(x_j)) \in$ minima **do**
  if $x_j \notin (\hat{x}, \text{end})$ **then**
    minima $\leftarrow$ minima$\backslash(x_j, f(x_j))$
$(x', f(x')) \leftarrow$ minima.pop(0)
featurePairs $\leftarrow (\hat{x}, f(\hat{x}) - f(x'))$
RecursionStep(x', $\hat{x}$, maxima.copy(), minima.copy(), featurePairs)
RecursionStep(x', end, maxima.copy(), minima.copy(), featurePairs)

is marked as maximum and stored in the corresponding list. Correspondingly, values that are smaller than their neighbors are marked as a minimum. This identification of extremal points can be made in linear run-time since the list is traversed just once, resulting in a complexity of $\sigma(q)$*. The two lists maxima and minima are then sorted by their corresponding value* $f(x_j)$ *(the minima list inverted) with a complexity of* $\sigma(m \cdot \log m)$*.*

For the largest feature, the global maximum, the persistence is defined to be the difference to the global minimum (see Equation 4.1). These values are the first elements of their corresponding lists. After calculating the persistence, the maximum is removed from the list, and the found feature $(x, p(x))$ *is stored in the list persistencePairs. The recursion step is called afterwards with the intervals* $[x_0, \hat{x})$ *and* $[x_{q-1}, \hat{x})$*. Notice that the second interval is reversed. As input, the recursion step gets a copy of the two lists minima and maxima, as well as the original list featurePairs. All these computations can be done in constant time, i.e.,* $\sigma(1)$*. After the last recursion step, all the features are detected and stored in the featurePairs list and can be returned.*

The input for the recursion step (Algorithm 3) consists of two indices, namely start and end (indicating the part of the data which is processed in the current recursion step, i.e., the positions of m/z values), a list of maxima and a list of minima, and the shared list of featurePairs. In the first step, the routine removes all the maxima not in the currently processed part of the data. There are at most m elements in the maximum list, so the complexity of this task is $\sigma(m)$*. If the list is empty after the removing step, the recursion step reaches the end and can return. If not, it removes the first element* $\hat{x}$ *from the list. This is the most persistent feature (in terms of topology) in the processed part of the data. The elder rule (cf. [30]) states that the feature can only merge with a feature with a larger persistence, which is per construction at the index end. The corresponding minimum (cf. Equation (4.2) to* $\hat{x}$ *can only be in the interval* $(\hat{x}, end)$ *so the minima list can be filtered in a similar fashion to the maxima list with the same complexity. Since there are more possible features in the interval* $(start, \hat{x})$*, the recursion step is called once more for this interval.*

The values $\hat{x}$ and the smallest minimum x' from the minima list generate a topological feature. This feature is updated in the original featurePairs list, and the recursion step can be repeated with the two intervals $(x', \hat{x})$ and (x', end).

The recursion step is processed at least once for each maximum with three extra calls after no more maxima are left, i.e., it runs at most $4 \times m$ times. Given the complexity of each step of $\sigma(m)$, the complexity of all the recursion steps together is $\sigma(m^2)$. This gives an overall complexity of the algorithm of

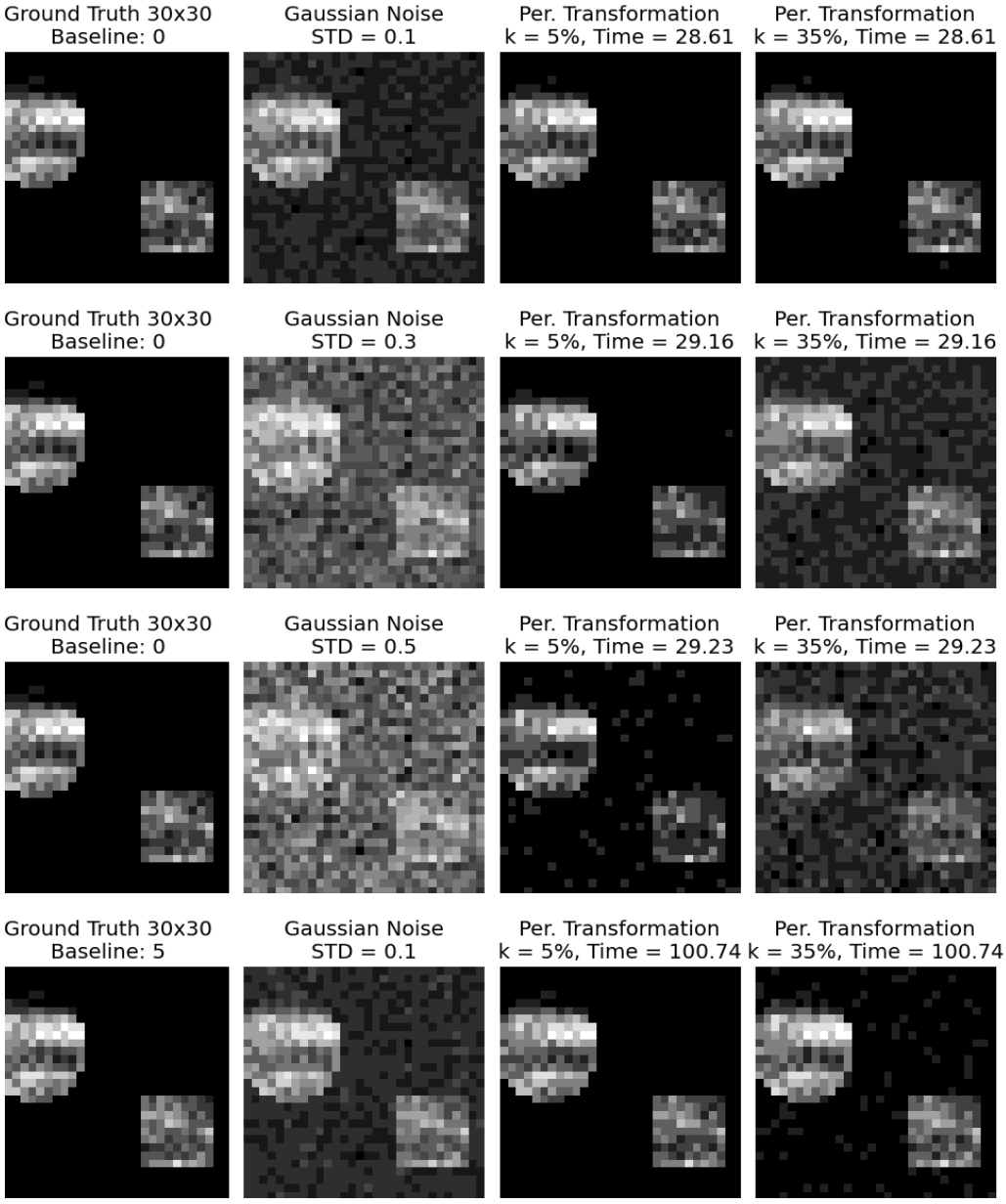
$$\sigma(q) + \sigma(m^2) + \sigma(m \cdot \log m) = \sigma(q) + \sigma(m^2).$$

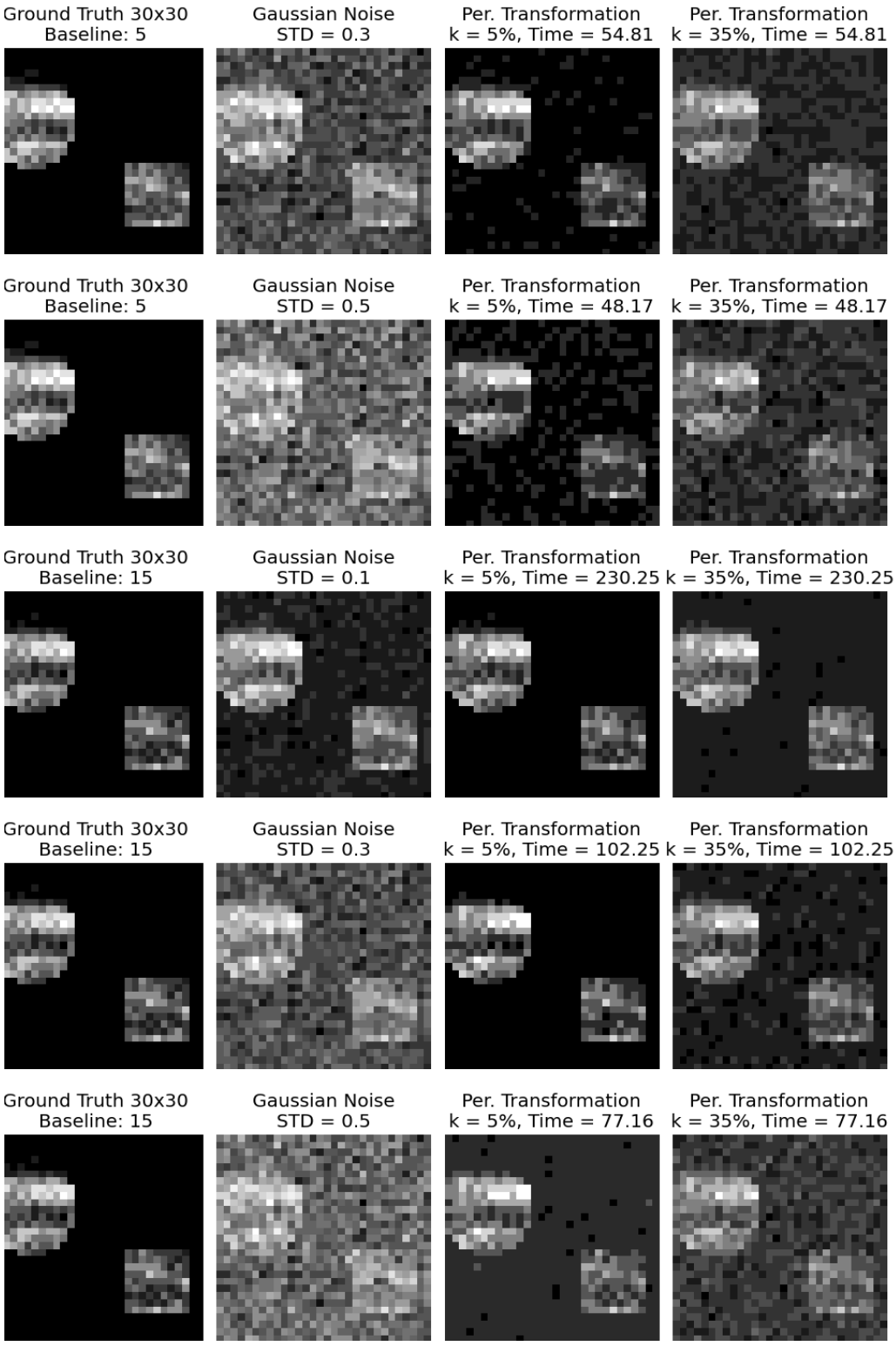
For each maximum, there is a tuple stored which contains the information of the position and the persistence, resulting in an overall storage use of $2m$ elements.

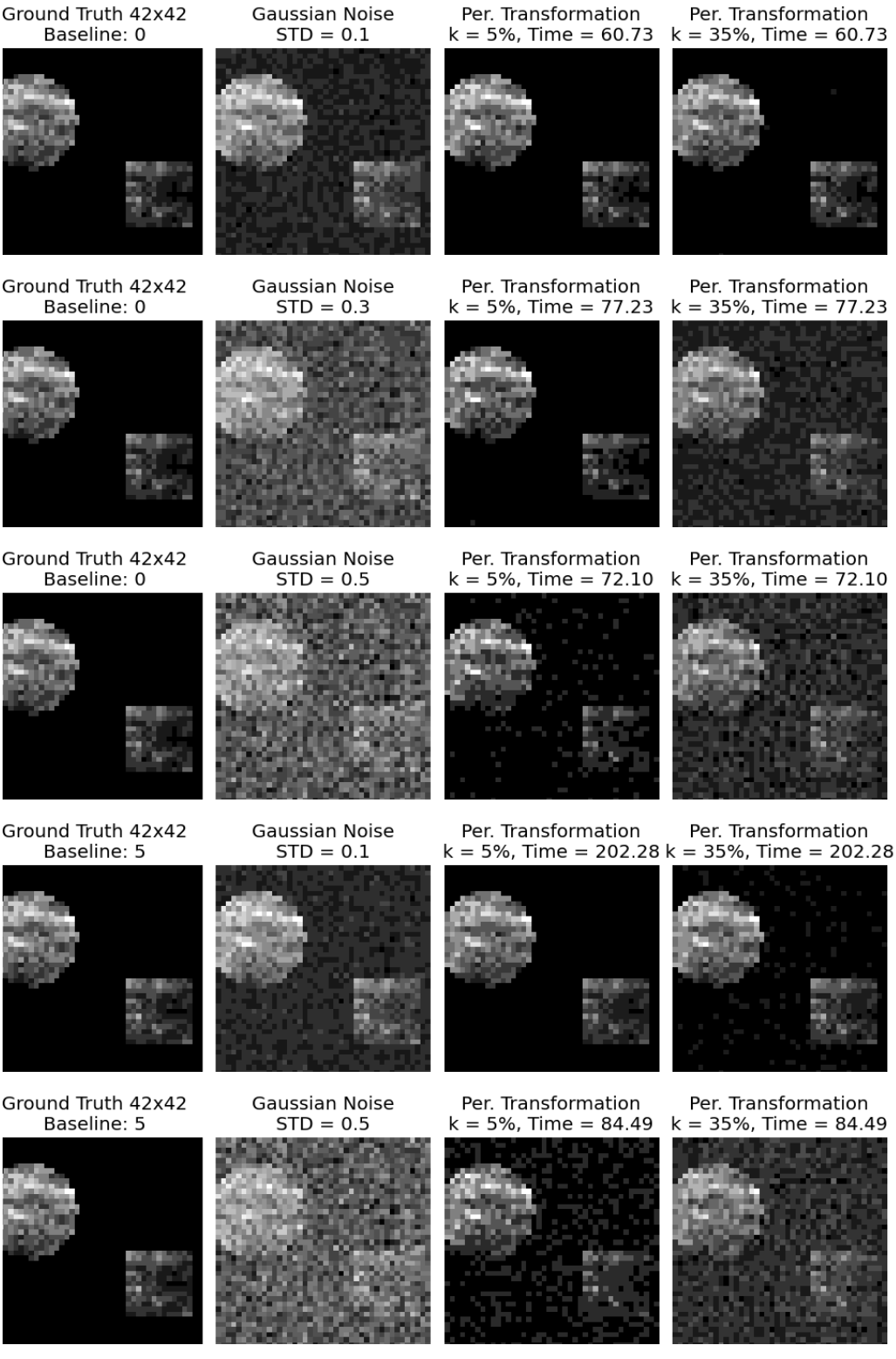
The algorithm always terminates since, at each recursion step, one maximum is removed from the list of maxima —if it is not already empty. Likewise, each part of the input list is being processed. Since there is only a finite number of elements in the maxima list (i.e., m), the algorithm terminates after all are processed. Even more, the algorithm returns all the features with their persistence. Each maximum creates a feature, and all maxima are processed. They are paired with the correct minimum between themselves and a feature with a higher persistence according to the elder rule (see [30]). Hence, the algorithm always terminates and returns the correct solution in $\sigma(q) + \sigma(m^2)$ run-time.

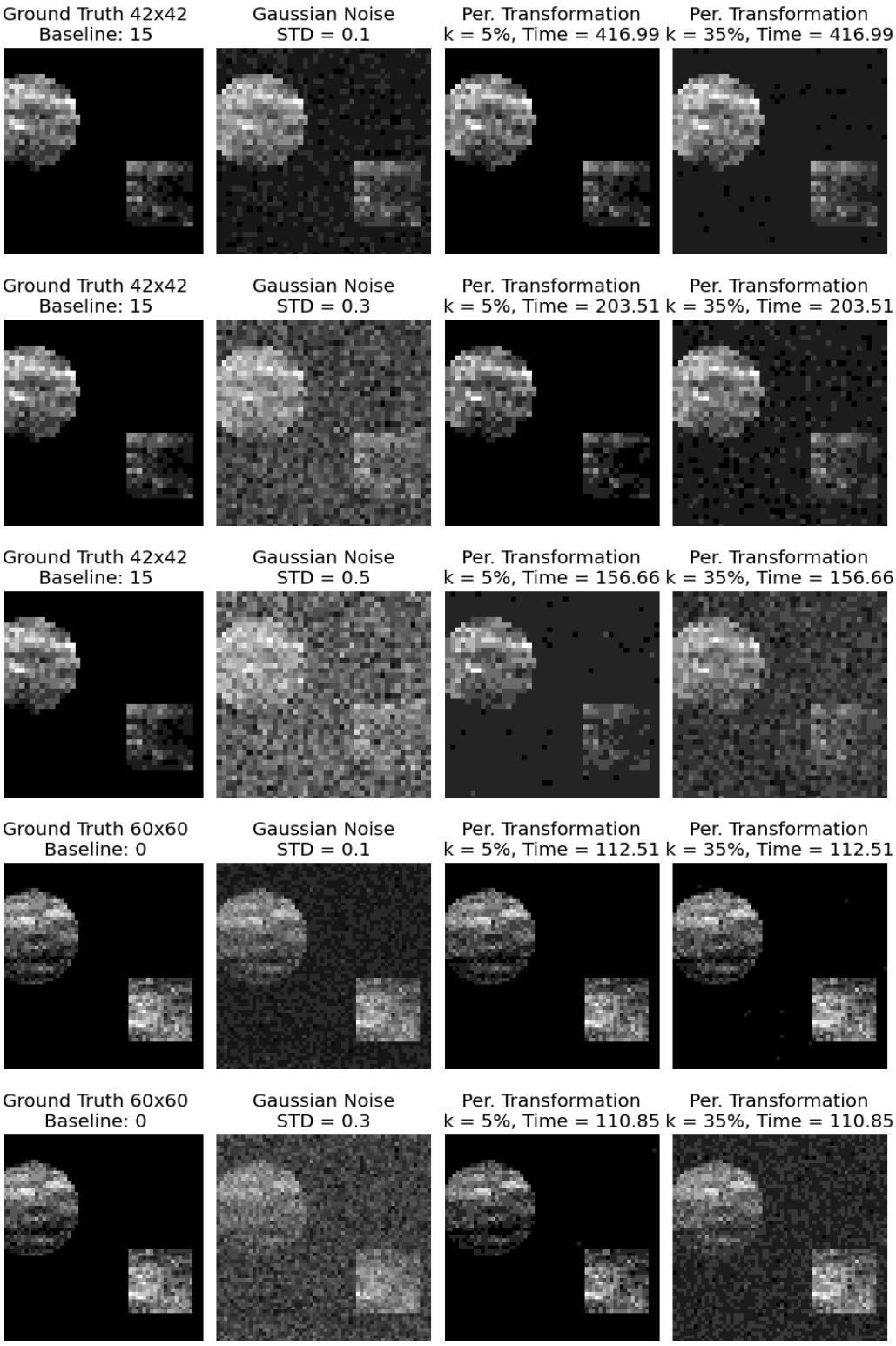
A.2 Additional simulations

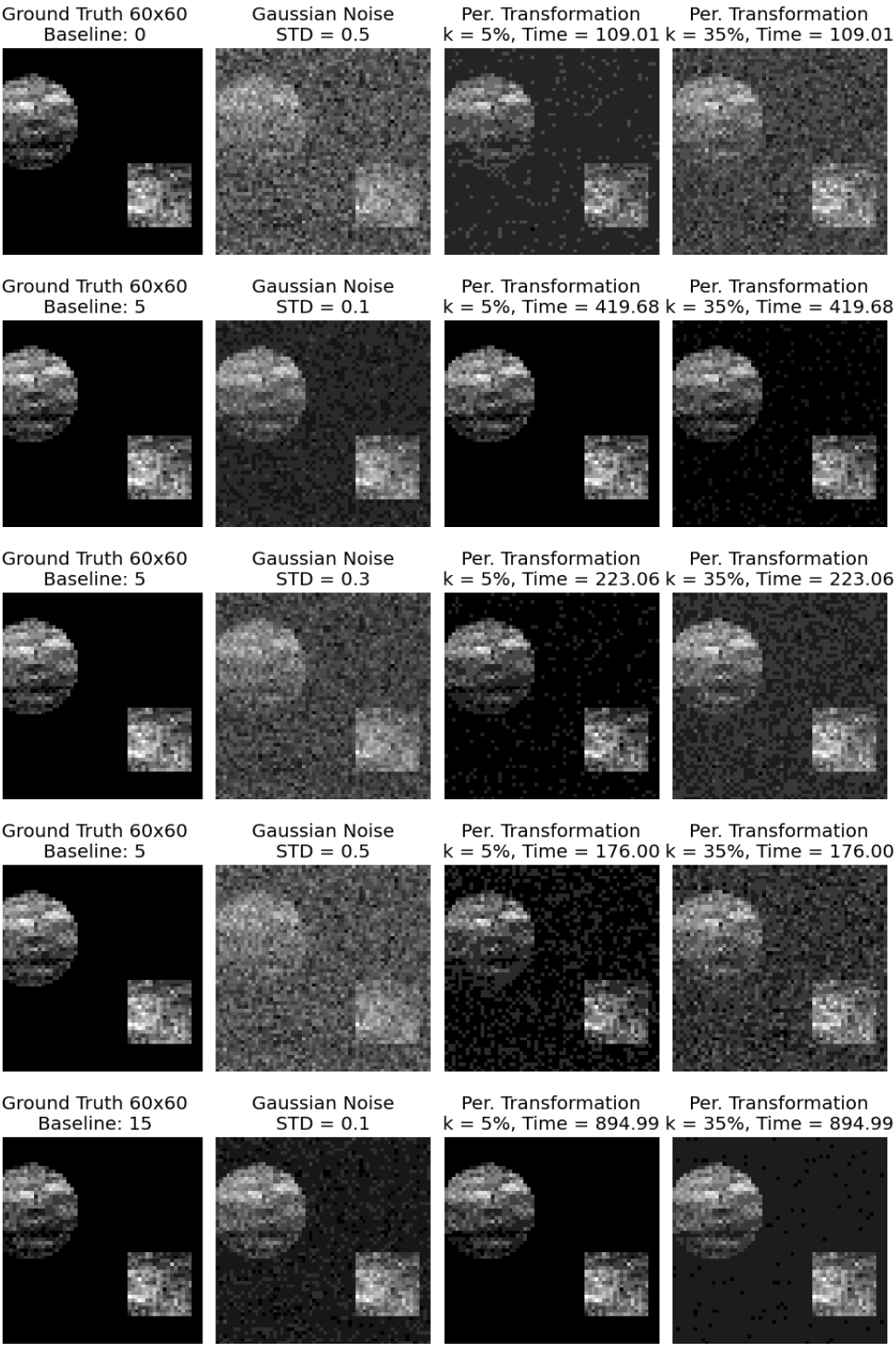
Distinct types and levels of noise are added to the ground truth and displayed. Finally, the results of the denoising with the persistence transformation are depicted.

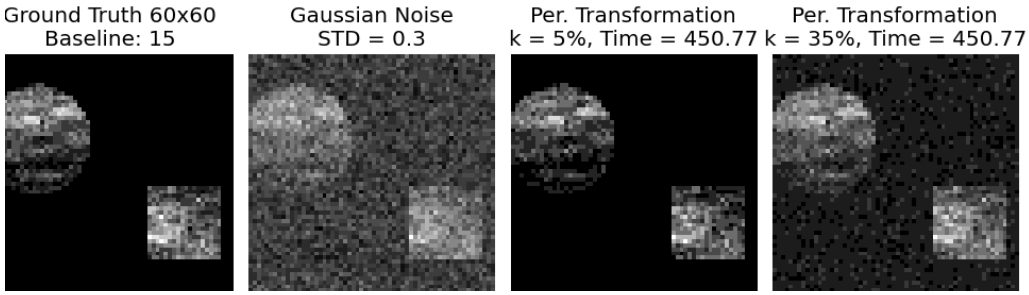

Ground Truth 30x30
Baseline: 0
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 28.61
Per. Transformation
k = 35%, Time = 28.61
Ground Truth 30x30
Baseline: 0
Gaussian Noise
STD = 0.3
Per. Transformation
k = 5%, Time = 29.16
Per. Transformation
k = 35%, Time = 29.16
Ground Truth 30x30
Baseline: 0
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 29.23
Per. Transformation
k = 35%, Time = 29.23
Ground Truth 30x30
Baseline: 5
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 100.74
Per. Transformation
k = 35%, Time = 100.74

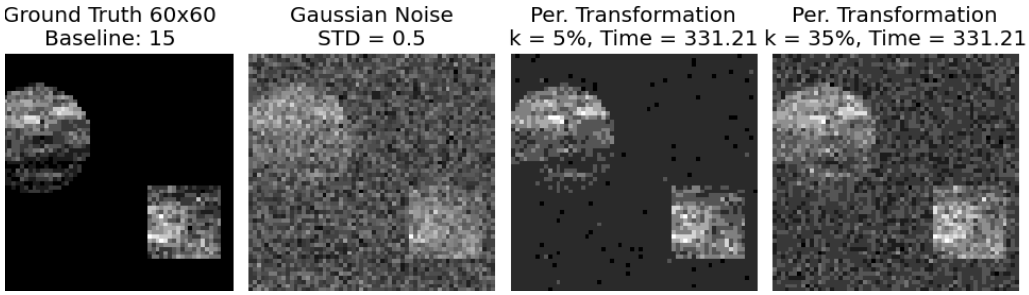

Ground Truth 30x30
Baseline: 5
Gaussian Noise
STD = 0.3
Per. Transformation
k = 5%, Time = 54.81
Per. Transformation
k = 35%, Time = 54.81
Ground Truth 30x30
Baseline: 5
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 48.17
Per. Transformation
k = 35%, Time = 48.17
Ground Truth 30x30
Baseline: 15
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 230.25
Per. Transformation
k = 35%, Time = 230.25
Ground Truth 30x30
Baseline: 15
Gaussian Noise
STD = 0.3
Per. Transformation
k = 5%, Time = 102.25
Per. Transformation
k = 35%, Time = 102.25
Ground Truth 30x30
Baseline: 15
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 77.16
Per. Transformation
k = 35%, Time = 77.16

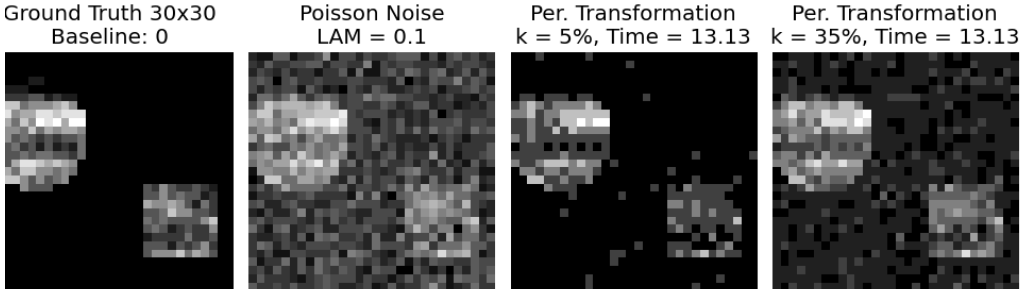

Ground Truth 42x42
Baseline: 0
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 60.73
Per. Transformation
k = 35%, Time = 60.73
Ground Truth 42x42
Baseline: 0
Gaussian Noise
STD = 0.3
Per. Transformation
k = 5%, Time = 77.23
Per. Transformation
k = 35%, Time = 77.23
Ground Truth 42x42
Baseline: 0
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 72.10
Per. Transformation
k = 35%, Time = 72.10
Ground Truth 42x42
Baseline: 5
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 202.28
Per. Transformation
k = 35%, Time = 202.28
Ground Truth 42x42
Baseline: 5
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 84.49
Per. Transformation
k = 35%, Time = 84.49

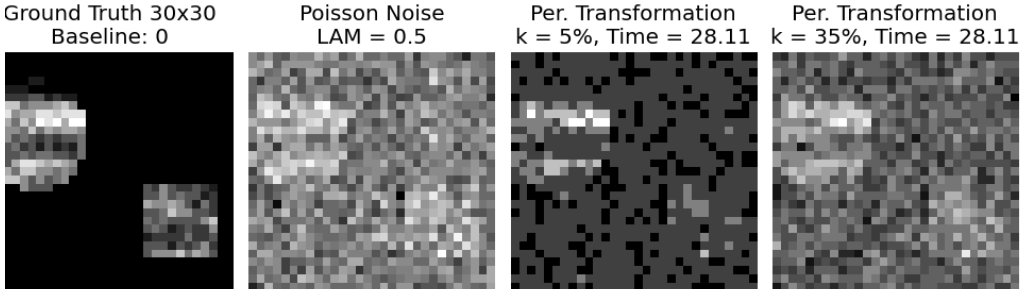

Ground Truth 42x42
Baseline: 15
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 416.99
Per. Transformation
k = 35%, Time = 416.99
Ground Truth 42x42
Baseline: 15
Gaussian Noise
STD = 0.3
Per. Transformation
k = 5%, Time = 203.51
Per. Transformation
k = 35%, Time = 203.51
Ground Truth 42x42
Baseline: 15
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 156.66
Per. Transformation
k = 35%, Time = 156.66
Ground Truth 60x60
Baseline: 0
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 112.51
Per. Transformation
k = 35%, Time = 112.51
Ground Truth 60x60
Baseline: 0
Gaussian Noise
STD = 0.3
Per. Transformation
k = 5%, Time = 110.85
Per. Transformation
k = 35%, Time = 110.85

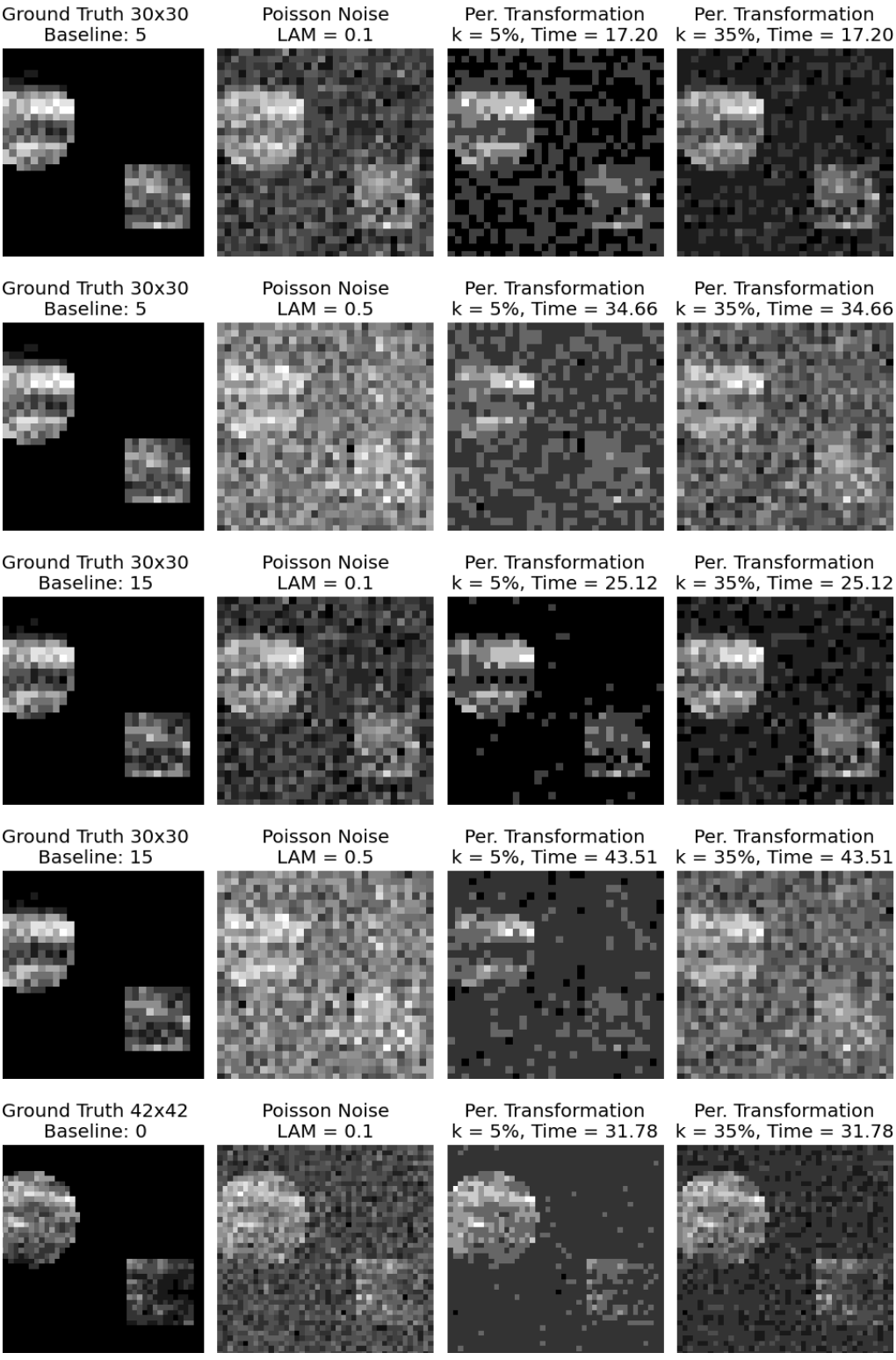

Ground Truth 60x60
Baseline: 0
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 109.01
Per. Transformation
k = 35%, Time = 109.01
Ground Truth 60x60
Baseline: 5
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 419.68
Per. Transformation
k = 35%, Time = 419.68
Ground Truth 60x60
Baseline: 5
Gaussian Noise
STD = 0.3
Per. Transformation
k = 5%, Time = 223.06
Per. Transformation
k = 35%, Time = 223.06
Ground Truth 60x60
Baseline: 5
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 176.00
Per. Transformation
k = 35%, Time = 176.00
Ground Truth 60x60
Baseline: 15
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 894.99
Per. Transformation
k = 35%, Time = 894.99

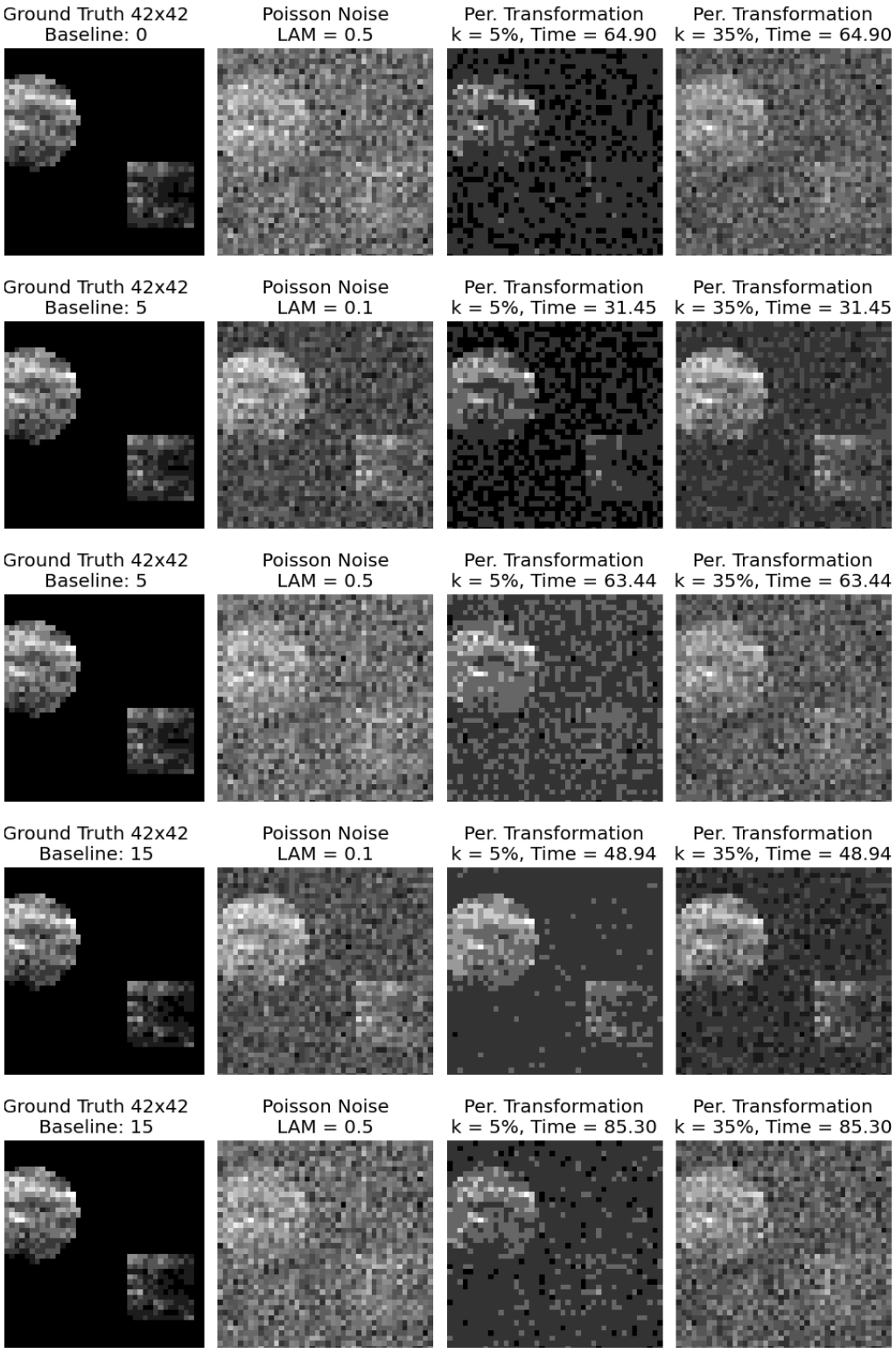

Ground Truth 60x60
Baseline: 15
Gaussian Noise
STD = 0.3
Per. Transformation
k = 5%, Time = 450.77
Per. Transformation
k = 35%, Time = 450.77

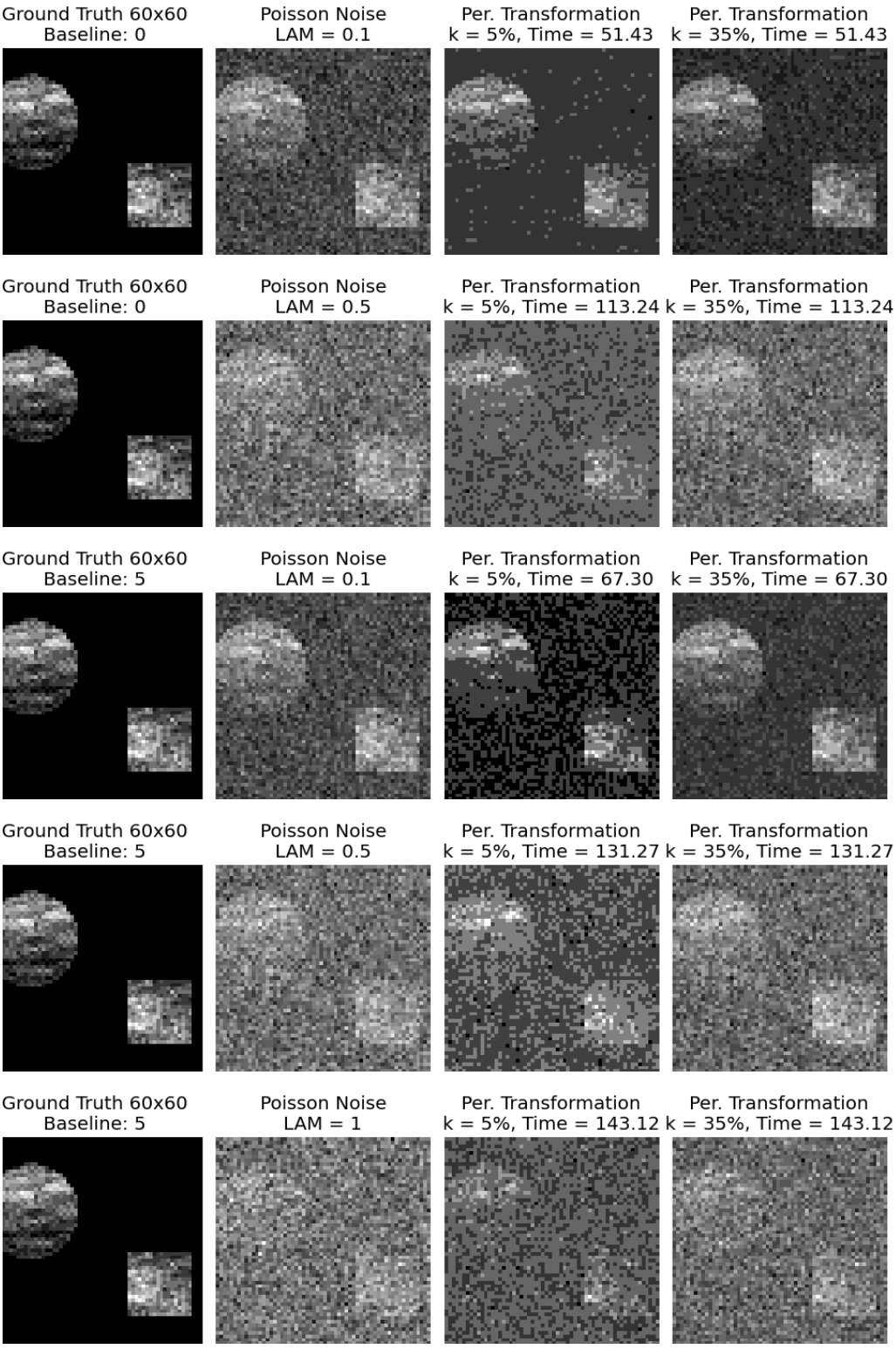

Ground Truth 60x60
Baseline: 15
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 331.21
Per. Transformation
k = 35%, Time = 331.21


Ground Truth 30x30
Baseline: 0
Poisson Noise
LAM = 0.1
Per. Transformation
k = 5%, Time = 13.13
Per. Transformation
k = 35%, Time = 13.13


Ground Truth 30x30
Baseline: 0
Poisson Noise
LAM = 0.5
Per. Transformation
k = 5%, Time = 28.11
Per. Transformation
k = 35%, Time = 28.11


Ground Truth 30x30
Baseline: 5
Poisson Noise
LAM = 0.1
Per. Transformation
k = 5%, Time = 17.20
Per. Transformation
k = 35%, Time = 17.20
Ground Truth 30x30
Baseline: 5
Poisson Noise
LAM = 0.5
Per. Transformation
k = 5%, Time = 34.66
Per. Transformation
k = 35%, Time = 34.66
Ground Truth 30x30
Baseline: 15
Poisson Noise
LAM = 0.1
Per. Transformation
k = 5%, Time = 25.12
Per. Transformation
k = 35%, Time = 25.12
Ground Truth 30x30
Baseline: 15
Poisson Noise
LAM = 0.5
Per. Transformation
k = 5%, Time = 43.51
Per. Transformation
k = 35%, Time = 43.51
Ground Truth 42x42
Baseline: 0
Poisson Noise
LAM = 0.1
Per. Transformation
k = 5%, Time = 31.78
Per. Transformation
k = 35%, Time = 31.78


Ground Truth 42x42
Baseline: 0
Poisson Noise
LAM = 0.5
Per. Transformation
k = 5%, Time = 64.90
Per. Transformation
k = 35%, Time = 64.90
Ground Truth 42x42
Baseline: 5
Poisson Noise
LAM = 0.1
Per. Transformation
k = 5%, Time = 31.45
Per. Transformation
k = 35%, Time = 31.45
Ground Truth 42x42
Baseline: 5
Poisson Noise
LAM = 0.5
Per. Transformation
k = 5%, Time = 63.44
Per. Transformation
k = 35%, Time = 63.44
Ground Truth 42x42
Baseline: 15
Poisson Noise
LAM = 0.1
Per. Transformation
k = 5%, Time = 48.94
Per. Transformation
k = 35%, Time = 48.94
Ground Truth 42x42
Baseline: 15
Poisson Noise
LAM = 0.5
Per. Transformation
k = 5%, Time = 85.30
Per. Transformation
k = 35%, Time = 85.30


Ground Truth 60x60
Baseline: 0
Poisson Noise
LAM = 0.1
Per. Transformation
k = 5%, Time = 51.43
Per. Transformation
k = 35%, Time = 51.43
Ground Truth 60x60
Baseline: 0
Poisson Noise
LAM = 0.5
Per. Transformation
k = 5%, Time = 113.24
Per. Transformation
k = 35%, Time = 113.24
Ground Truth 60x60
Baseline: 5
Poisson Noise
LAM = 0.1
Per. Transformation
k = 5%, Time = 67.30
Per. Transformation
k = 35%, Time = 67.30
Ground Truth 60x60
Baseline: 5
Poisson Noise
LAM = 0.5
Per. Transformation
k = 5%, Time = 131.27
Per. Transformation
k = 35%, Time = 131.27
Ground Truth 60x60
Baseline: 5
Poisson Noise
LAM = 1
Per. Transformation
k = 5%, Time = 143.12
Per. Transformation
k = 35%, Time = 143.12

Bibliography

[1] Alan Agresti. *Categorical data analysis*. Second. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], New York, 2002, pp. xvi+710. ISBN: 0-471-36093-7. DOI: 10.1002/0471249688. URL: https://doi.org/10.1002/0471249688.

[2] M. Aichler and A. Walch. "MALDI Imaging mass spectrometry: current frontiers and perspectives in pathology research and practice." In: *Lab Invest* 95.4 (2015), pp. 422–431. DOI: 10.1038/labinvest.2014.156. URL: https://doi.org/10.1038/labinvest.2014.156.

[3] T. Alexandrov. "MALDI imaging mass spectrometry: statistical data analysis and current computational challenges". In: *BMC Bioinformatics* 13 Suppl 16 (2012), S11. DOI: 10.1186/1471-2105-13-S16-S11.

[4] T. Alexandrov and A. Bartels. "Testing for presence of known and unknown molecules in imaging mass spectrometry". In: *Bioinform.* 29.18 (2013), pp. 2335–2342. DOI: 10.1093/bioinformatics/btt388. URL: https://doi.org/10.1093/bioinformatics/btt388.

[5] T. Alexandrov et al. "Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation". In: *Bioinformatics* 25.5 (Mar. 2009), pp. 643–649. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btn662. URL: https://doi.org/10.1093/bioinformatics/btn662.

[6] T. Alexandrov et al. "Super-resolution segmentation of imaging mass spectrometry data: Solving the issue of low lateral resolution". In: *J Proteomics.* 75.1 (2011). https://doi.org/10.1016/j.jprot.2011.08.002, pp. 237–45.

[7] David Azriel and Armin Schwartzman. "The empirical distribution of a large number of correlated normal variables". English. In: *J. Am. Stat. Assoc.* 110.511 (2015), pp. 1217–1228. ISSN: 0162-1459. DOI: 10.1080/01621459.2014.958156.

[8] Anirban Basu and Paul J. Rathouz. "Estimating marginal and incremental effects on health outcomes using flexible link and variance function models". In: *Biostatistics* 6.1 (Jan. 2005), pp. 93–109. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxh020.

[9] J. Behrmann et al. "Deep learning for tumor classification in imaging mass spectrometry". In: *Bioinformatics* 34.7 (Apr. 2018), pp. 1215–1223. DOI: 10.1093/bioinformatics/btx724.

[10] Kyle D. Bemis et al. "Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments". In: *Bioinformatics* 31.14 (July 2015), pp. 2418–2420. DOI: 10.1093/bioinformatics/btv146.

[11] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: A practical and powerful approach to multiple testing." In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 57.1 (1995), pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[12] Yoav Benjamini and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency." In: *Ann. Stat.* 29.4 (2001), pp. 1165–1188. DOI: 10.1214/aos/1013699998.

[13] Gilles Blanchard et al. "On least favorable configurations for step-up-down tests". In: *Statist. Sinica* 24.1 (2014), pp. 1–23. ISSN: 1017-0405. DOI: 10.5705/ss.2011.205.

[14] Taras Bodnar and Thorsten Dickhaus. "On the Simes inequality in elliptical models". In: *Ann. Inst. Statist. Math.* 69.1 (2017), pp. 215–230. ISSN: 0020-3157. DOI: 10.1007/s10463-015-0539-4. URL: https://doi.org/10.1007/s10463-015-0539-4.

[15] D. Böhning. "Multinomial logistic regression algorithm". In: *Annals Inst. Stat. Math.* 44.1 (1992). https://doi.org/10.1007/BF00048682, pp. 197–200.

[16] T. Boskamp et al. “A new classification method for MALDI imaging mass spectrometry data acquired on formalin-fixed paraffin-embedded tissue samples”. In: *Biochim Biophys Acta Proteins Proteom* 1865.7 (July 2017), pp. 916–926. DOI: 10.1016/j.bbapap.2016.11.003.

[17] Leo Breiman. “Random Forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[18] Anuraag Bukkuri, Noemi Andor, and Isabel K. Darcy. “Applications of Topological Data Analysis in Oncology”. In: *Frontiers Artif. Intell.* 4 (2021), p. 659037. DOI: 10.3389/frai.2021.659037. URL: https://doi.org/10.3389/frai.2021.659037.

[19] Richard M. Caprioli, Terry B. Farmer, and Joe Gile. “Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS.” In: *Analytical chemistry* 69.23 (Dec. 1997), pp. 4751–4760. DOI: doi:10.1021/ac970888i.

[20] R. Casadonte and R. Caprioli. “Proteomic analysis of formalin-fixed paraffin-embedded tissue by MALDI imaging mass spectrometry.” In: *Nat Protoc* 6 (2011), pp. 1695–1709. DOI: 10.1038/nprot.2011.388. URL: https://doi.org/10.1038/nprot.2011.388.

[21] Frédéric Chazal and Bertrand Michel. “An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists”. In: *Frontiers Artif. Intell.* 4 (2021), p. 667963. DOI: 10.3389/frai.2021.667963. URL: https://doi.org/10.3389/frai.2021.667963.

[22] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. “Stability of Persistence Diagrams”. In: *Discret. Comput. Geom.* 37.1 (2007), pp. 103–120. DOI: 10.1007/s00454-006-1276-5. URL: https://doi.org/10.1007/s00454-006-1276-5.

[23] Marco Contessoto et al. “Persistent cup-length”. In: *arXiv preprint arXiv:2107.01553* (2021).

[24] Raphaël Couronné, Philipp Probst, and Anne-Laure Boulesteix. “Random forest versus logistic regression: a large-scale benchmark experiment”. In: *BMC Bioinform.* 19.1 (2018), 270:1–270:14. DOI: 10.1186/s12859-018-2264-5. URL: https://doi.org/10.1186/s12859-018-2264-5.

[25] Sören-Oliver Deininger, Michael Becker, and Detlev Suckau. “Tutorial: Multivariate Statistical Treatment of Imaging Data for Clinical Biomarker Discovery”. In: *Mass Spectrometry Imaging: Principles and Protocols*. Ed. by Stanislav S. Rubakhin and Jonathan V. Sweedler. Totowa, NJ: Humana Press, 2010, pp. 385–403. DOI: 10.1007/978-1-60761-746-4_22. URL: https://doi.org/10.1007/978-1-60761-746-4_22.

[26] T Dickhaus. *Simultaneous Statistical Inference with Applications in the Life Sciences*. http://dx.doi.org/10.1007/978-3-642-45182-9. Berlin, Heidelberg: Springer, 2014, Chapters 9–12.

[27] T. Dickhaus, A. Neumann, and T. Bodnar. “Multivariate Multiple Test Procedures”. In: *Handbook of Multiple Comparisons*. Ed. by Xinping Cui et al. Boca Raton, FL: Chapman & Hall / CRC Press, 2021, Chapter 3.

[28] Lilun Du et al. “False discovery rate control under general dependence by symmetrized data aggregation”. In: *Journal of the American Statistical Association* 118.541 (2023), pp. 607–621. DOI: https://doi.org/10.1080/01621459.2021.1945459.

[29] Sandrine Dudoit, Juliet Popper Shaffer, and Jennifer C. Boldrick. “Multiple Hypothesis Testing in Microarray Experiments”. In: *Statistical Science* 18.1 (2003), pp. 71–103. DOI: 10.1214/ss/1056397487.

[30] Herbert Edelsbrunner and John L Harer. *Computational topology: An introduction*. Providence, USA: American Mathematical Society, 2010.

[31] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. “Topological Persistence and Simplification”. In: *Discret. Comput. Geom.* 28.4 (2002), pp. 511–533. DOI: 10.1007/s00454-002-2885-2. URL: https://doi.org/10.1007/s00454-002-2885-2.

[32] Bradley Efron. “Correlated z-values and the accuracy of large-scale statistical estimates”. In: *J. Amer. Statist. Assoc.* 105.491 (2010), pp. 1042–1055. ISSN: 0162-1459. DOI: 10.1198/jasa.2010.tm09129. URL: https://doi.org/10.1198/jasa.2010.tm09129.

[33] Bradley Efron. “Correlation and large-scale simultaneous significance testing”. In: *J. Amer. Statist. Assoc.* 102.477 (2007), pp. 93–103. ISSN: 0162-1459. DOI: 10.1198/016214506000001211. URL: https://doi.org/10.1198/016214506000001211.

[34] Ludwig Fahrmeir and Gerhard Tutz. "Models for multicategorical responses: Multivariate extensions of generalized linear models". In: *Multivariate statistical modelling based on generalized linear models*. Springer, 2001, pp. 69–137.

[35] Ludwig Fahrmeir et al. *Multivariate statistical modelling based on generalized linear models*. Vol. 425. Berlin, Germany: Springer, 1994.

[36] J. Fan et al. *pfa: Estimates False Discovery Proportion Under Arbitrary Covariance Dependence*. R package version 1.1, available from https://CRAN.R-project.org/package=pfa. 2016.

[37] Jianqing Fan and Xu Han. "Estimation of the false discovery proportion with unknown dependence". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 79.4 (2017), pp. 1143–1164. ISSN: 1369-7412. DOI: 10.1111/rssb.12204. URL: https://doi.org/10.1111/rssb.12204.

[38] Jianqing Fan, Xu Han, and Weijie Gu. "Estimating false discovery proportion under arbitrary covariance dependence". In: *J. Amer. Statist. Assoc.* 107.499 (2012), pp. 1019–1035. ISSN: 0162-1459. DOI: 10.1080/01621459.2012.720478. URL: https://doi.org/10.1080/01621459.2012.720478.

[39] Jianqing Fan, Yuan Liao, and Martina Mincheva. "Large covariance estimation by thresholding principal orthogonal complements". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 75.4 (2013). With 33 discussions by 57 authors and a reply by Fan, Liao and Mincheva, pp. 603–680. ISSN: 1369-7412. DOI: 10.1111/rssb.12016. URL: https://doi.org/10.1111/rssb.12016.

[40] Jianqing Fan et al. "FarmTest: factor-adjusted robust multiple testing with approximate false discovery control". In: *J. Amer. Statist. Assoc.* 114.528 (2019), pp. 1880–1893. ISSN: 0162-1459. DOI: 10.1080/01621459.2018.1527700. URL: https://doi.org/10.1080/01621459.2018.1527700.

[41] Brittany Fasy et al. "Statistical Inference For Persistent Homology: Confidence Sets For Persistence Diagrams". In: (Mar. 2013). DOI: 10.1214/14-AOS1252.

[42] Pascal Fernsel. "Spatially Coherent Clustering Based on Orthogonal Nonnegative Matrix Factorization". In: *Journal of Imaging* 7.10 (2021), p. 194. DOI: "doi.org/10.3390/jimaging7100194".

[43] H. Finner, T. Dickhaus, and M. Roters. "Dependency and false discovery rate: Asymptotics." In: *Ann. Stat.* 35.4 (2007), pp. 1432–1455. DOI: 10.1214/009053607000000046.

[44] Chloé Friguet, Maela Kloareg, and David Causeur. "A factor model approach to multiple testing under dependence". In: *J. Amer. Statist. Assoc.* 104.488 (2009), pp. 1406–1415. ISSN: 0162-1459. DOI: 10.1198/jasa.2009.tm08332. URL: https://doi.org/10.1198/jasa.2009.tm08332.

[45] Christopher R Genovese and Larry Wasserman. "Exceedance control of the false discovery proportion". In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1408–1417. DOI: 10.1198/016214506000000339.

[46] A. Genz et al. *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.1-2. https://CRAN.R-project.org/package=mvtnorm. Aug. 2021.

[47] Sebastian Gibb and Korbinian Strimmer. "MALDIquant: a versatile R package for the analysis of mass spectrometry data". In: 28.17 (2012), pp. 2270–2271. DOI: 10.1093/bioinformatics/bts447.

[48] Ruben van den Goorbergh et al. "The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression". In: *Journal of the American Medical Informatics Association* 29.9 (June 2022), pp. 1525–1534. ISSN: 1527-974X. DOI: 10.1093/jamia/ocac093. URL: https://doi.org/10.1093/jamia/ocac093.

[49] Florent Grélard et al. "Esmraldi: efficient methods for the fusion of mass spectrometry and magnetic resonance images". In: *BMC Bioinform.* 22.1 (2021), p. 56. DOI: 10.1186/s12859-020-03954-z. URL: https://doi.org/10.1186/s12859-020-03954-z.

[50] Kari Krizak Halle et al. "Computationally efficient familywise error rate control in genome-wide association studies using score tests for generalized linear models". In: *Scandinavian Journal of Statistics* 47.4 (2020), pp. 1090–1113. DOI: https://doi.org/10.1111/sjos.12451.

[51] M. Hanselmann et al. "Concise representation of mass spectrometry images by probabilistic latent semantic analysis". In: *Anal Chem* 80.24 (Dec. 2008), pp. 9649–9658. DOI: https://doi.org/10.1021/ac801303x.

[52] Asad Hasan, Zhiyu Wang, and Alireza S. Mahani. "Fast Estimation of Multinomial Logit Models: R Package mnlogit". In: *Journal of Statistical Software* 75.3 (2016), pp. 1–24. DOI: 10.18637/jss.v075.i03. URL: https://www.jstatsoft.org/index.php/jss/article/view/v075i03.

[53] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. New York, NY: Springer, 2009. ISBN: 9780387848570. DOI: 10.1007/978-0-387-84858-7. URL: https://doi.org/10.1007/978-0-387-84858-7.

[54] Jesse Hemerik and Jelle J Goeman. "False Discovery Proportion Estimation by Permutations: Confidence for Significance Analysis of Microarrays". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.1 (2018), pp. 137–155. DOI: https://doi.org/10.1111/rssb.12238.

[55] Emrys A. Jones et al. "Multiple Statistical Analysis Techniques Corroborate Intratumor Heterogeneity in Imaging Mass Spectrometry Datasets of Myxofibrosarcoma". In: *PLOS ONE* 6.9 (Sept. 2011), Article No. e24913. DOI: 10.1371/journal.pone.0024913. URL: https://doi.org/10.1371/journal.pone.0024913.

[56] Kaitlin Kirasich, Trace Smith, and Bivin Sadler. "Random forest vs logistic regression: binary classification for heterogeneous datasets". In: *SMU Data Science Review* 1.3 (2018), p. 9.

[57] Gideon Klaila, Vladimir Vutov, and Anastasios Stefanou. "Supervised topological data analysis for MALDI mass spectrometry imaging applications". In: *BMC Bioinform.* 24.279 (2023), pp. 1–22. DOI: 10.1186/s12859-023-05402-0. URL: https://doi.org/10.1186/s12859-023-05402-0.

[58] Dmitry N. Kozlov. "A combinatorial method to compute explicit homology cycles using Discrete Morse Theory". In: *J. Appl. Comput. Topol.* 4.1 (2020), pp. 79–100. DOI: 10.1007/s41468-019-00042-x. URL: https://doi.org/10.1007/s41468-019-00042-x.

[59] J. Kriegsmann, M. Kriegsmann, and R. Casadonte. "MALDI TOF imaging mass spectrometry in clinical pathology: A valuable tool for cancer diagnostics (review)". In: *Int J Oncol* 46.3 (2015), pp. 893–906. DOI: https://doi.org/10.3892/ijo.2014.2788.

[60] M. Kriegsmann et al. "Reliable Entity Subtyping in Non-small Cell Lung Cancer by Matrix-assisted Laser Desorption/Ionization Imaging Mass Spectrometry on Formalin-fixed Paraffin-embedded Tissue Specimens". In: *Mol Cell Proteomics* 15.10 (2016), pp. 3081–3089. DOI: 10.1074/mcp.m115.057513.

[61] Andrew N Krutchinsky and Brian T Chait. "On the mature of the chemical noise in MALDI mass spectra". In: *Journal of the American Society for Mass Spectrometry* 13.2 (2002), pp. 129–134.

[62] Jeffrey T. Leek and John D. Storey. "A general framework for multiple testing dependence". In: *Proc. Natl. Acad. Sci. USA* 105.48 (2008), pp. 18718–18723. ISSN: 0027-8424. DOI: 10.1073/pnas.0808709105.

[63] J. Leuschner et al. "Supervised non-negative matrix factorization methods for MALDI imaging applications". In: *Bioinformatics* 35.11 (2019), pp. 1940–1947. DOI: 10.1093/bioinformatics/bty909.

[64] KUNG-YEE Liang and SCOTT L. Zeger. "Longitudinal data analysis using generalized linear models". In: *Biometrika* 73.1 (Apr. 1986), pp. 13–22. ISSN: 0006-3444. DOI: 10.1093/biomet/73.1.13. eprint: https://academic.oup.com/biomet/article-pdf/73/1/13/679793/73-1-13.pdf. URL: https://doi.org/10.1093/biomet/73.1.13.

[65] Andy Liaw, Matthew Wiener, et al. "Classification and regression by randomForest". In: *R news* 2.3 (2002), pp. 18–22.

[66] F. Lieb, T. Boskamp, and H. G. Stark. "Peak detection for MALDI mass spectrometry imaging data using sparse frame multipliers". In: *Journal of Proteomics* 225 (2020), p. 103852. ISSN: 1874-3919. DOI: https://doi.org/10.1016/j.jprot.2020.103852.

[67] Ciara Frances Loughrey et al. "The topology of data: opportunities for cancer research". In: *Bioinform.* 37.19 (2021), pp. 3091–3098. DOI: 10.1093/bioinformatics/btab553. URL: https://doi.org/10.1093/bioinformatics/btab553.

[68] Facundo Mémoli, Anastasios Stefanou, and Ling Zhou. "Persistent Cup Product Structures and Related Invariants". In: *arXiv preprint arXiv:2211.16642* (2022).

[69] John Milnor. *Morse Theory. (AM-51), Volume 51*. Princeton: Princeton University Press, 1963. ISBN: 9781400881802. DOI: doi:10.1515/9781400881802. URL: https://doi.org/10.1515/9781400881802.

[70] Thomas Mortier et al. “Bacterial species identification using MALDI-TOF mass spectrometry and machine learning techniques: A large-scale benchmarking study”. In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 6157–6168. ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2021.11.004. URL: https://www.sciencedirect.com/science/article/pii/S2001037021004694.

[71] Christine H. Müller and Neyko Neykov. “Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models”. In: *Journal of Statistical Planning and Inference* 116.2 (2003), pp. 503–519. ISSN: 0378-3758. DOI: https://doi.org/10.1016/S0378-3758(02)00265-3. URL: https://www.sciencedirect.com/science/article/pii/S0378375802002653.

[72] Judith Martha Neumann et al. “Subtyping non-small cell lung cancer by histology-guided spatial metabolomics”. In: *Journal of cancer research and clinical oncology* 148.2 (2022), pp. 351–360. DOI: 10.1007/s00432-021-03834-w.

[73] N. Neykov et al. “Robust fitting of mixtures using the trimmed likelihood estimator”. In: *Computational Statistics & Data Analysis* 52.1 (2007), pp. 299–308. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2006.12.024. URL: https://www.sciencedirect.com/science/article/pii/S0167947306005019.

[74] Jeremy L Norris et al. “Processing MALDI Mass Spectra to Improve Mass Spectral Direct Tissue Analysis”. In: *International journal of mass spectrometry* 260.2-3 (2007), pp. 212–221.

[75] Zainib Noshad et al. “Fault Detection in Wireless Sensor Networks through the Random Forest Classifier”. In: *Sensors* 19.7 (2019), p. 1568. DOI: 10.3390/s19071568. URL: https://doi.org/10.3390/s19071568.

[76] J Oetjen et al. “An approach to optimize sample preparation for MALDI imaging MS of FFPE sections using fractional factorial design of experiments”. In: *Anal Bioanal Chem* 408 (2016). https://doi.org/10.1007/s00216-016-9793-4, pp. 6729–6740.

[77] Nina Otter et al. “A roadmap for the computation of persistent homology”. In: *EPJ Data Sci.* 6.1 (2017), p. 17. DOI: 10.1140/epjds/s13688-017-0109-5. URL: https://doi.org/10.1140/epjds/s13688-017-0109-5.

[78] Mahesh Pal. “Random forest classifier for remote sensing classification”. In: *International journal of remote sensing* 26.1 (2005), pp. 217–222.

[79] Philip Pallmann, Mias Pretorius, and Christian Ritz. “Simultaneous comparisons of treatments at multiple time points: Combined marginal models versus joint modeling”. In: *Statistical Methods in Medical Research* 26.6 (2017), pp. 2633–2648. DOI: https://doi.org/10.1177/0962280215603743.

[80] Philip Pallmann, Christian Ritz, and Ludwig A Hothorn. “Simultaneous small-sample comparisons in longitudinal or multi-endpoint trials using multiple marginal models”. In: *Statistics in Medicine* 37.9 (2018), pp. 1562–1576. DOI: https://doi.org/10.1002/sim.7610.

[81] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[82] M. S. Pepe, R. C. Whitaker, and K. Seidel. “Estimating and comparing univariate associations with application to the prediction of adult obesity”. In: *Stat. Med.* 18.2 (1999), pp. 163–173. DOI: 10.1002/(sici)1097-0258(19990130)18:2<163::aid-sim11>3.0.co;2-f.

[83] Christian Bressen Pipper, Christian Ritz, and Hans Bisgaard. “A versatile method for confirmatory evaluation of the effects of a covariate in multiple models”. In: *J. R. Stat. Soc. Ser. C. Appl. Stat.* 61.2 (2012), pp. 315–326. ISSN: 0035-9254. DOI: 10.1111/j.1467-9876.2011.01005.x. URL: https://doi.org/10.1111/j.1467-9876.2011.01005.x.

[84] N. Poté et al. “Imaging mass spectrometry reveals modified forms of histone H4 as new biomarkers of microvascular invasion in hepatocellular carcinomas”. In: *Hepatology* 58.3 (Sept. 2013), pp. 983–994. DOI: 10.1002/hep.26433.

[85] Philipp Probst and Anne-Laure Boulesteix. “To tune or not to tune the number of trees in random forest”. English. In: *J. Mach. Learn. Res.* 18 (2018). Id/No 181, p. 18. ISSN: 1532-4435. URL: jmlr.csail.mit.edu/papers/v18/17-269.html.

[86] Philipp Probst, Marvin N. Wright, and Anne-Laure Boulesteix. “Hyperparameters and tuning strategies for random forest”. In: *WIREs Data Mining Knowl. Discov.* 9.3 (2019). DOI: 10.1002/widm.1301. URL: https://doi.org/10.1002/widm.1301.

[87] R Development Core Team. *R: A Language and Environment for Statistical Computing*. Available from: http://www.R-project.org. 2021.

[88] M. Reck et al. "Management of non-small-cell lung cancer: recent developments". In: *Lancet* 382.9893 (Aug. 2013), pp. 709–719. DOI: 10.1016/S0140-6736(13)61502-0.

[89] Christian Ritz, Rikke Pilmann Laursen, and Camilla Trab Damsgaard. "Simultaneous inference for multilevel linear mixed models—with an application to a large-scale school meal study". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 66.2 (2017), pp. 295–311. DOI: https://doi.org/10.1111/rssc.12161.

[90] T. L. Salter et al. "A comparison of SIMS and DESI and their complementarities". In: *Surface and Interface Analysis* 43.1-2 (2011), pp. 294–297. DOI: https://doi.org/10.1002/sia.3412. eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/sia.3412. URL: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/sia.3412.

[91] Daniel J Schaid et al. "Score Tests for Association between Traits and Haplotypes when Linkage Phase Is Ambiguous". In: *The American Journal of Human Genetics* 70.2 (2002), pp. 425–434. DOI: https://doi.org/10.1086/338688.

[92] Konstantin Schildknecht, Karsten Tabelow, and Thorsten Dickhaus. "More specific signal detection in functional magnetic resonance imaging by false discovery rate control for hierarchically structured systems of hypotheses". In: *PloS one* 11.2 (2016), e0149016. DOI: https://doi.org/10.1371/journal.pone.0149016.

[93] Jonathan von Schroeder. *Stable Feature Selection with Applications to MALDI Imaging Mass Spectrometry Data*. Preprint, available via https://arxiv.org/abs/2006.15077. 2020.

[94] K. Schwamborn. "Imaging mass spectrometry in biomarker discovery and validation". In: *J Proteomics* 75.16 (Aug. 2012), 4990–4998. DOI: 10.1016/j.jprot.2012.06.015.

[95] Armin Schwartzman. "Comment: FDP vs FDR and the effect of conditioning". In: *J. Amer. Statist. Assoc.* 107.499 (2012), pp. 1039–1041. ISSN: 0162-1459. DOI: 10.1080/01621459.2012.712876. URL: https://doi.org/10.1080/01621459.2012.712876.

[96] Shaun R Seaman and Bertram Müller-Myhsok. "Rapid Simulation of P values for Product Methods and Multiple-Testing Adjustment in Association Studies". In: *The American Journal of Human Genetics* 76.3 (2005), pp. 399–408. DOI: https://10.1086/428140.

[97] M. W. Senko, S. C. Beu, and F. W. McLaffertycor. "Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions". In: *J Am Soc Mass Spectrom* 6.4 (Apr. 1995), pp. 229–233. DOI: 10.1016/1044-0305(95)00017-8.

[98] Binita Shrestha, Haroon Stephen, and Sajjad Ahmad. "Impervious Surfaces Mapping at City Scale by Fusion of Radar and Optical Data through a Random Forest Classifier". In: *Remote. Sens.* 13.15 (2021), p. 3040. DOI: 10.3390/rs13153040. URL: https://doi.org/10.3390/rs13153040.

[99] Yara Skaf and Reinhard C. Laubenbacher. "Topological data analysis in biomedicine: A review". In: *J. Biomed. Informatics* 130 (2022), p. 104082. DOI: 10.1016/j.jbi.2022.104082. URL: https://doi.org/10.1016/j.jbi.2022.104082.

[100] Martin Slawski et al. "Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching". In: *BMC Bioinform.* 13 (2012), p. 291. DOI: 10.1186/1471-2105-13-291. URL: https://doi.org/10.1186/1471-2105-13-291.

[101] S. Sperandei. "Understanding logistic regression analysis". In: *Biochem. Med.* 24.1 (2014), pp. 12–18. DOI: 10.11613/BM.2014.003.

[102] J. Stange et al. "Multiplicity- and dependency-adjusted p-values for control of the family-wise error rate". In: *Stat. Probab. Lett.* 111 (2016), pp. 32–40. DOI: 10.1016/j.spl.2016.01.005.

[103] Jens Stange, Nina Loginova, and Thorsten Dickhaus. "Computing and approximating multivariate chi-square probabilities". In: *J. Stat. Comput. Simul.* 86.6 (2016), pp. 1233–1247. DOI: 10.1080/00949655.2015.1058798.

[104] J. R. Stevens, A. Al Masud, and A. Suyundikov. “A comparison of multiple testing adjustment methods with block-correlation positively-dependent tests”. In: *PLoS One* 12.4 (2017), e0176124. DOI: 10.1371/journal.pone.0176124.

[105] John D. Storey. “A direct approach to false discovery rates”. English. In: *J. R. Stat. Soc., Ser. B, Stat. Methodol.* 64.3 (2002), pp. 479–498. ISSN: 1369-7412. DOI: 10.1111/1467-9868.00346.

[106] John D. Storey, Jonathan E. Taylor, and David Siegmund. “Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach”. English. In: *J. R. Stat. Soc., Ser. B, Stat. Methodol.* 66.1 (2004), pp. 187–205. ISSN: 1369-7412. DOI: 10.1111/j.1467-9868.2004.00439.x.

[107] Simon Sugár et al. “Proteomic Analysis of Lung Cancer Types - A Pilot Study”. In: *Cancers* 14.11 (2022). ISSN: 2072-6694. DOI: 10.3390/cancers14112629. URL: https://www.mdpi.com/2072-6694/14/11/2629.

[108] Wenguang Sun and T. Tony Cai. “Large-scale multiple testing under dependence”. In: *J. R. Stat. Soc., Ser. B, Stat. Methodol.* 71.2 (2009), pp. 393–424. ISSN: 1369-7412. DOI: 10.1111/j.1467-9868.2008.00694.x.

[109] Y Thomas. *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.1-5. https://CRAN.R-project.org/package=VGAM. 2021.

[110] Wiebke Timm et al. “Peak intensity prediction in MALDI-TOF mass spectrometry: A machine learning study to support quantitative proteomics”. In: *BMC Bioinform.* 9 (2008). DOI: 10.1186/1471-2105-9-443. URL: https://doi.org/10.1186/1471-2105-9-443.

[111] Dennis Trede et al. “On the importance of mathematical methods for analysis of MALDI-imaging mass spectrometry data”. In: *Journal of Integrative Bioinformatics (JIB)* 9.1 (2012), pp. 1–11.

[112] Kirill A. Veselkov et al. “Chemo-informatic strategy for imaging mass spectrometry-based hyperspectral profiling of lipid signatures in colorectal cancer”. In: *Proceedings of the National Academy of Sciences* 111.3 (2014), pp. 1216–1221. DOI: 10.1073/pnas.1310524111. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1310524111.

[113] Athanasios Vlontzos et al. “Topological Data Analysis of Database Representations for Information Retrieval”. In: *CoRR* abs/2104.01672 (2021). arXiv: 2104.01672. URL: https://arxiv.org/abs/2104.01672.

[114] Vladimir Vutov and Thorsten Dickhaus. “Multiple multi-sample testing under arbitrary covariance dependency”. In: *Statistics in Medicine* 42.17 (July 2023), pp. 2944–2961. DOI: https://doi.org/10.1002/sim.9761.

[115] Vladimir Vutov and Thorsten Dickhaus. “Multiple two-sample testing under arbitrary covariance dependency with an application in imaging mass spectrometry”. In: *Biometrical Journal* 65.2 (2023), p. 2100328. DOI: https://doi.org/10.1002/bimj.202100328.

[116] H. Wang et al. “Inconsistency Between Univariate and Multiple Logistic Regressions”. In: *Shanghai Arch Psychiatry* 29.2 (2017), pp. 124–128. DOI: 10.11919/j.issn.1002-0829.217031.

[117] Caroline Weis et al. “Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra”. In: *Bioinform.* 36.Supplement-1 (2020), pp. i30–i38. DOI: 10.1093/bioinformatics/btaa429. URL: https://doi.org/10.1093/bioinformatics/btaa429.

[118] C. D. Wijetunge et al. “A new peak detection algorithm for MALDI mass spectrometry data based on a modified Asymmetric Pseudo-Voigt model”. In: *BMC Genomics* 16 Suppl 12 (2015), Article No. S12. DOI: 10.1186/1471-2164-16-S12-S12.

[119] C. Yang, Z. He, and W. Yu. “Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis”. In: *BMC Bioinformatics* 10 (Jan. 2009), Article No. 4. DOI: 10.1186/1471-2105-10-4.