Statistical methods for mass spectrometry-based clinical proteomics

Kakourou, A.A.

Author: Kakourou, Alexia
Title: Statistical methods for mass spectrometry-based clinical proteomics
Date: 2018-03-08

1 Introduction

1.1 Introduction

Clinical proteomics is currently a growing and promising field of research which deals with the study and understanding of fundamental biological processes at the protein level. A key objective in clinical proteomic studies is the identification of biomarkers for early detection, diagnosis of disease, assessment of disease prognosis and monitoring of disease progression which could improve the long-term survival of patients. The technological progress in mass spectrometry (MS) and other related technologies over the last decade has elevated the use and potential of proteomic studies in clinical research. Mass spectrometry has become one of the key technologies for jointly measuring the expression of thousands of proteins in biological samples.

Following the acquisition of protein expression, several research questions/investigations could be of interest, among which the assessment of differential expression levels or the comparison of protein profiles across distinct groups such as those collected in a case-control manner. The latter may allow for the construction of discriminant rules to distinguish between individuals as to the presence or absence of a disease or the degree of the disease progression. Moreover, the statistical analysis of protein profiles collected from both cases and controls could contribute to the identification of a set of features potentially associated with the physiological and pathological condition of an individual and therefore could provide evidence to further exploit diagnostic and therapeutic potential.

The development of MS instrumentation gave rise to new statistical challenges in the processing and analysis of the acquired data. This is due to the complex nature of the spectral proteomic signal which is measured, as it consists of high-dimensional functions representing the within-patient proteome expression. This thesis considers novel methods to respond to these challenges. We present a series of data analyses for distinct case-control cancer studies. In particular, the main objective throughout this thesis is the construction of diagnostic rules for disease status allocation of future patients through the use of innovative statistical methodology specific for the type of data considered in each of the studies.

The remainder of the introduction is organized as follows. We first explain the process by which the MS proteomic data are generated and pre-processed to obtain an analysable/condensed data set. Next, we describe the specific characteristics of this particular type of data and the statistical challenges which may arise when analysing such data. Additionally, we introduce the basic steps involved in proteomic diagnosis which include the choice of model, the construction of the diagnostic rule and the evaluation of the resulting rule and we give a detailed description of each of these steps. We finish with an outline of the chapters presented in this thesis along with a brief description of the main methodological contributions in each study.

1.2 Data acquisition

In general, all types of mass spectrometers comprise of three basic components: an ionization source, a mass analyser and a detector. Here, the description of mass spectrometry (MS) is limited to biomolecular applications since the studies described in this thesis are focussed on the analysis of peptides and proteins. In the first part of a mass spectrometer, each biomolecule is ionised at either atmospheric pressure (ambient) or at decreased pressures. These ions are then transferred into the mass analyser in which the mass of the corresponding molecule is determined, or, more precisely, the mass-to-charge (m/z) ratio. Often the mass analyser is additionally used as a separation device, enabling the simultaneous mass analysis of a mixture of biomolecules. In the third part of a mass spectrometer, ions are detected and the resulting output is an array of intensity readings distributed over an m/z range generated from the detected ions. This intensity array is referred to as a mass spectrum. An example of a mass spectrum is shown in Figure 1.1.

Peptide and protein analysis by MS is generally performed by using electrospray ionization (ESI) or matrix-assisted laser desorption/ionisation (MALDI). The first one yields multiply charged species, whereas MALDI (predominantly) results in singly charged species. In this thesis we consider MALDI mass spectra that are obtained from either a time-of-flight mass analyser (MALDI-TOF) or a Fourier transform ion cyclotron resonance system (MALDI-FTICR). In a TOF mass analyser, ions are accelerated through an electric field followed by mass separation in a field-free drift tube. The flight time, that is, the time it takes for an ion to travel through the detector is determined and the corresponding mass is calculated after proper calibration (kinetic energy equals $1/2$ mass times velocity(sq)). The heavier ions travel longer than ions with lower mass and thus will

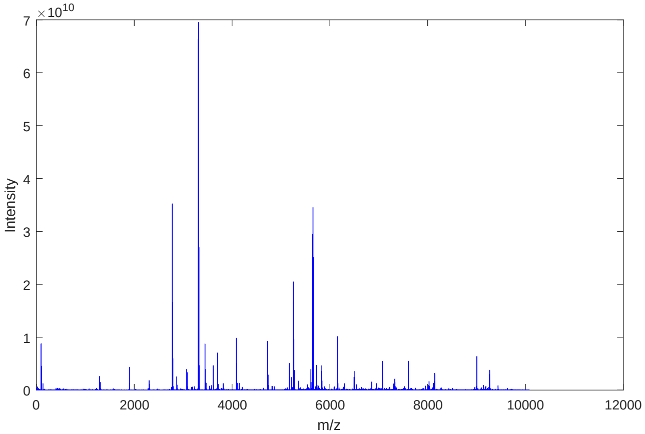

Figure 1.1: A mass spectrum for a single individual.

be detected sequentially. The separation power of modern TOF analysers is high and consequently TOF-based mass spectra are referred to as “high resolution” spectra (resolving powers up to 50,000). In an FTICR mass spectrometer, ions are trapped and mass analysed by measuring their cyclotron frequency in a high magnetic field. Since a frequency can be recorded more precisely than a flight time in TOF, this type of mass analysis is referred to as “ultrahigh resolution” (resolving powers higher than 100,000).

Both TOF and FT-based MS-technologies are widely used approaches for the analysis of complex mixtures and identification of biomarker signatures. Recently, FTICR MS-platforms have received a lot of attention due to their superiority, compared to TOF-MS, attributed to the ultrahigh mass resolving power that allows the analysis of large proteins and complex mixtures. Moreover, an improved mass accuracy and precision provide a more reliable identification of the detected species and allow a wide dynamic range for the detection of low abundant components. Particularly high mass accuracy and resolution are essential for protein and peptide identification. High resolving powers result in so-called isotopic profiles (or clusters) in mass spectra. These originate from different compositions of naturally occurring stable isotopes of carbon (C), nitrogen (N), oxygen (O) and hydrogen (H) atoms. The larger the mass of biomolecule, the greater the number of isotopes which are included, detected in a mass spectrum at higher m/z-values. In a MALDI-spectrum with mostly singly-charged ions, these isotopic peaks are approximately 1 Da apart, resulting in a series of locally correlated single peaks that reflect the isotopic distribution of the molecule of interest. Figure 1.2 plots an isotopic cluster at position m/z 2021,2.

The data analyses presented in Chapters 2 and 3 of this thesis are based on MALDI-TOF data while the last three chapters of the thesis are devoted to methods for preprocessing, summarizing and analysing MALDI-FTICR data while dealing with the ad-

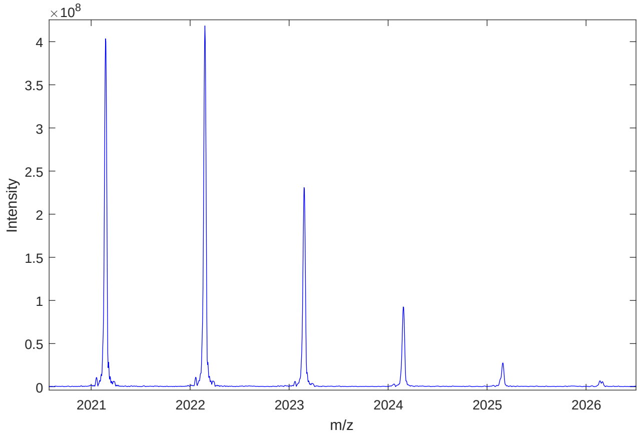

Figure 1.2: An isotopic cluster at position m/z 2021,2.

ditional statistical challenges which accompany this particular type of data.

1.3 Data pre-processing

Typically, in an MS experiment, the proteomic expression profiles of individuals are obtained in the form of m/z and intensity pairs. Upon acquisition of the proteomic profiles, data analysis can proceed in two ways. The first is to consider the complete high-dimensional proteomic signal which is measured without applying any data reduction prior to data analysis. The second is to reduce the high-dimensional individual profiles to a discrete set of peaks corresponding to potential proteins/peptides which, depending on the data resolution, might have an isotope clustering structure. In this thesis, both approaches are considered. In chapters 2 and 3, the MALDI-TOF proteomic profiles are kept intact, whereas in the last three chapters a preprocessing algorithm is applied to the MALDI-FTICR raw spectra in order to reduce the complete proteomic signal to cluster summaries of isotopic expression.

A common solution towards reducing the complete proteomic expression to a discrete set of peaks is to apply a peak detection approach to the raw spectra in order to locate the peptide-related isotopic peaks in the mass spectrum and determine their abundance. Various peak detection algorithms have been proposed over the last few years to preprocess mass spectral data. For high-resolution MS data, most of the reported peak detection algorithms make use of the fact that isotopic clusters are completely resolved - as opposed to low-resolution MS data where isotope clusters appear as single peaks in the mass spectrum - in order to obtain more accurate determination of the mass and the abundance of the protein/peptide corresponding to that mass. To this end, many different approaches have been proposed to interpret and process the acquired data, routinely by trying to match the observed isotope cluster pattern with a theoretical isotope distribution (Rockwood and Haimi, 2006; Horn et al., 2000). The advantage of completely resolved isotope clusters in high-resolution MS data is utilized in several methods through the exploitation of the property of 1 mass unit of distance between the successive peaks within the isotope distributions. This known property of isotope distributions is used in Horn et al. in order to identify isotope variants in isotopic clusters. However the proposed method is applicable to single spectra, hence it has not been evaluated in the context of clinical applications, and it is based on best fitting local models which can be computationally quite challenging. In Chapter 3 we propose a new approach which relies solely on the statistical property of successive isotopic peaks of a peptide molecule being separated by 1 Da in order to reduce the complete expression in the individual spectra to clusters of isotopic expression on which summary measures can be later defined. In contrast to the idea presented in Horn et al. (2000), the algorithm we propose uses information across all potential isotopic peaks across all patients in order to find m/z positions of peaks which belong to isotopic clusters. This is done in a completely non-parametric fashion and thus avoids any computationally intensive tasks such as model fitting to the observed spectra.

Peak detection can be trickier for low-resolution MS data where the isotope cluster patterns are not resolved but are instead displayed as single peaks in the mass spectrum. For lower resolution data, Morris et al. (2005) proposed a peak detection approach which is performed on the average spectrum rather than the individual spectra. The rational behind this idea is that a peak corresponding to a protein/peptide/molecule should stand out above the noise level as well as the baseline level, ideally in a sufficient amount of spectra, and these properties should be preserved in the mean spectrum which is simply the average across all spectra. In their paper they showed that using the mean spectrum to detect relevant peaks leads to greater sensitivity and specificity while it eliminates the difficult task of matching peaks detected on individual spectra (commonly referred to as alignment problem).

1.4 Limit of detection (LOD)

While detection and quantification of the proteomic expression in biological samples are essential steps in MS-based proteomics analysis, these steps can be complicated by measurements being subject to lower detection limits due to censoring mechanisms on low-abundance proteins/peptides. This issue is known as limit of detection (LOD) and occurs due to the limitation of MS equipment in measuring low-abundant proteins/peptides which may in fact be of greatest interest as potential biomarkers. Commonly, the lower the abundance of a protein/peptide, the more difficult it becomes for MS technologies to distinguish the peaks or isotopic clusters originating from that molecule from the background noise. This issue often gives rise to incomplete mass spectral data as low-abundance peptides below the LOD threshold which were undetectable and could not be quantified by the instrument are often reported as missing values.

Common strategies to deal with data subject to (lower or upper) detection limits are to delete the missing values and consider only the complete data for any further analysis or substitute the missing values with a fixed constant. However, both of these approaches can lead to severe loss of information and possibly bias in the results of any subsequent analysis as they do not take into account that the data have been left-censored. While the problem of handling data subject to LOD has seen many applications over the last years, these were reported mainly in the field of ecological and environmental research. Statistical methods to deal with proteomic data affected by the LOD have only recently been proposed in the field of MS clinical proteomics. Examples of such methods can be found in the papers by Dong et al. (2014) and Tekwe et al. (2012). Dong et al. addressed the problem of assessing bias in the estimation of distribution parameters of proteomic biomarkers whose measurements are affected by the LOD in distinguishing cancer patients from non-cancer patients. In their paper they showed that the estimates of receiver operating characteristic (ROC) curve parameters computed while adjusting for the LOD are much closer to the truth as compared to the estimates resulting when ignoring the LOD. On the other hand, Tekwe et al. treated the LOD issue in MS proteomic data as a problem of censored data analysis and proposed the use of survival methodology, in particular accelerated failure time (AFT) models, to investigate differential expression of proteins. They proved that AFT models have higher ability to detect differentially expressed proteins than standard testing procedures which ignore the left-censored nature of the proteomic data.

In Chapter 5 we present an approach to handle proteomic measurements subject to lower detection limits in the prediction framework. The approach is an adaptation of censored data methodology in which we use censored normal regression methods to estimate the expected expression within isotope clusters in order to obtain improved estimates of the overall isotope cluster abundance. The objective is to use these newly derived estimates as new input variables for the construction of more accurate prediction rules. We demonstrate that the proposed method can be used successfully to handle the LOD problem in determining the average expression in isotope clusters and calibrating diagnostic rules of comparable performance as if we had the complete information.

1.5 Proteomic prediction

Once the protein abundances have been obtained (and quantified) it is often of interest to evaluate the discriminatory information in mass spectral data collected in case-control studies for the detection of various types of cancer. An approach to address these questions is through the construction of prediction rules.

Building a prediction model can be a challenging task in proteomics research, due to the complex nature of the spectral proteomic signal measured for each individual. Especially when the number of spectral features that are measured is much larger than the number of samples collected in the proteomic case-control study, we may run into the risk of overfitting (over-interpreting) the data, a common phenomenon in proteomics and other omic-studies. This phenomenon gives rise to models which predict very well on the data on which the diagnostic rule was based but perform very poorly on new data. Therefore, there is a crucial need for obtaining a fully validated and unbiased assessment of the predictive performance of the resulting diagnostic rule as this may determine whether the overall research effort is worthwhile.

Various models have been developed for the construction of diagnostic rules and different estimation algorithms have been proposed to fit these models while avoiding overfitting. A few examples of such prediction methods as well as a description of the strategy we chose to use throughout this thesis in order to overcome the potential optimistic bias due to overfitting are given in the following sections. Additionally, we give an overview of evaluation methods and performance measures to assess the predictive potential of a diagnostic rule.

1.5.1 Types of models

In diagnostic (classification) problems, the objective is to construct a rule which best describes the relationship between the disease (class) outcome and the set of predictor variables. There is a great variety of established methods in the classification literature which attempt to model the relationship between a number of covariates and a binary outcome. While this relationship may often be non-linear in real life, it has been shown in applications that most classification potential can be achieved using relatively simple structures such as linear forms of linear or logistic discriminant analysis (Hand, 2006). Especially in high-dimensional data problems with relatively small sample sizes as in omics research, simple linear structures are considered to be more reliable and stable as compared to more complex and sophisticated structures such as neural networks or support vector machines. Several favourable results in the field of deep learning have led to a growing trend of applying the associated techniques to high-dimensional data structures in the hope of generating better and more reliable predictions. Deep learning refers to artificial neural networks with more complex architectures (neural networks which are composed of many layers). However, the impact of deep learning methods, especially in the field of clinical research, remains to be proven. Taking all the above into consideration, the approaches selected for the construction of diagnostic rules in the context of this thesis are restricted to linear prediction methods.

When the number of predictors is smaller than the number of observations, standard forms of linear classification methods such as standard linear discriminant analysis or standard logistic regression can be applied to the data in order to derive the classification rule. When the number of predictor variables exceeds the number of observations and/or when there are high correlations between them, as is the case with MS proteomic data, regularization schemes need to be employed in order to avoid overfitting which may result in unstable estimates and poor generalizations. Various regularization approaches have been proposed in the field of statistics which try to deal with the problem of overfitting and collinearity both in the case of linear discrimination and logistic regression.

Linear discriminant analysis (LDA) finds a direction in space that maximizes the between-group variations while minimizing the within-group variations. In high-dimensional data problems, linear discriminant analysis cannot be applied directly, as the pooled within-group sample dispersion matrix is singular and therefore can not be inverted to give the required estimates. Various methods have been proposed in the literature to deal with the problem of singularity while most popular approaches to tackle this problem are using shrinkage-based estimation for the pooled within-group covariance matrix. Shrinkage in this case is carried out through the use of principal component decomposition (PCA) (Krzanowski et al., 1995; Mertens, 2003; Mertens et al., 2006) or ridge regularization.

Regularized logistic regression (LR) shrinks the coefficients towards zero by imposing a penalty on their size. This is achieved by adding a penalty function to the likelihood. Commonly used penalty functions for regularized logistic regression are the quadratic and the absolute penalty functions. The quadratic penalty function leads to ridge logistic regression (RLR) and shrinkage of the parameter vector (Le Cessie and van Houwelingen, 1992). The absolute penalty function leads to lasso logistic regression (LLR), resulting in both shrinkage of the parameter vector and variable selection (Tibshirani, 1996). While ridge logistic regression models are known to outperform other classification models (including lasso logistic regression) with regards to predictive performance, the estimates of the regression effects are usually not interpretable. When the research objective is to obtain more interpretable prediction rules, lasso logistic regression is the preferred method. An alternative to ridge and lasso logistic regression is to use a convex combination of the quadratic and absolute penalties leading to the elastic net logistic regression (ENLR) (Zou and Hastie, 2005).

1.5.2 Construction of diagnostic rules

Application of the above mentioned classification methods requires tuning of the associated regularized parameters. When there is subsequent need for assessing the predictive performance of the rule, it is crucial that the choice of the optimal value of any tuning parameter is itself validated in order to avoid overfitting. To accomplish this, the optimization procedure for selecting the optimal tuning parameter can/should not be based neither on the left aside validation data which will be used to evaluate the resulting rule nor on the complete calibration data which will be used to calibrate the diagnostic rule. A common approach to make a validatory decision on the optimal penalty value, without the need of introducing a separate test (tuning) set in addition to the calibration and validation sets, is via cross-validation. Throughout this thesis, we use a leave-one-out cross validatory approach for predictive optimization of any tuning parameter. Leave-one-out cross-validation removes each individual i (for $i = 1, ..., n$) in turn from the calibration data, and uses the leftover data to construct the classification rule f_{θ}^{-i} for a grid of tuning parameters $\theta \in \{\theta_1, ..., \theta_m\}$. The resulting rule for each value of θ is then evaluated on the left-out-observation i by calculating the loss $l^i(\theta)$. This procedure is repeated across all individuals and for each observation separately and gives a cross-validated estimate of the optimal value of tuning parameter $\hat{\theta}$ which is chosen as the one that minimizes the cross-validated average loss across all individuals. The selected value of tuning parameter can be used thereafter to calibrate the final allocation rule which has yet to be validated as well, in order to obtain an unbiased estimate of the overall predictive performance. We discuss how to perform predictive validation of the allocation rule in an unbiased manner in the following subsection.

1.5.3 Assessment of diagnostic rules

When diagnosis is the main research objective, obtaining an unbiased evaluation of the predictive performance of the calibrated diagnostic rule is crucial. Evaluation of the calibrated classification rule can be achieved through the use of a set aside data set, not involved in the model fitting. When the sample size is not too small, this can be done by splitting the available data into calibration and validation sets. In this case, the calibration set is used for model fitting while the set aside validation set is used for model assessment by applying the fitted model to the validation data. The predictive performance of the resulting rule can then be summarized by calculating misclassification rates and other performance measures on the basis of the validated predictions.

Often, in case-control experiments, the study design is such as to define a calibration set and a separate validation set, which probably went through the same sampling handling and measurement process as the calibration set. The validation samples in such studies might even be collected in a later time period than the calibration samples. This offers the possibility to evaluate the diagnostic rule, derived based on one data set, on an independent data set, under the assumption that samples from both data sets stem from the same underlying population. This type of model evaluation is usually referred to as “external validation”. In cases of limited resources which may lead to small sample sizes, a way to mimic a situation where the complete available data set is composed of a calibration set and a separate validation set is through the use of cross-validation. We restrict attention to leave-one-out cross-validation for consistency with the previous subsection. Again here, as in the case of predictive optimization, each observation is left aside such that the model is fitted to the leftover data and evaluated on the single left-out datum which is the only validation sample. This is done repeatedly so that each observation serves exactly once as validation sample and gives for each observation a cross-validated prediction estimate of the class outcome. This type of evaluation is referred to as “internal” validation and it is a common approach in MS clinical applications which are often characterized by relatively small sample sizes. Irrespective of the type of model evaluation, it is recommended that the results of a study are replicated in another laboratory/hospital to ensure their validity and reliability.

There are various performance measures that can be used to assess the predictive performance of a diagnostic rule. These measures are indicative of either the calibration capacity, i.e. how close predictions are to the actual outcome, or the discriminative ability

(Steyerberg et al., 2010), i.e. to what extent patients who have the outcome have higher risk predictions than those who do not, of the prediction model. Examples of calibration measures, commonly used in classification literature for reporting performance of a prediction model are the Brier score

$$B = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - y_i)^2,$$

and the Deviance of the model

$$\begin{aligned} D &= -2\sum_{i=1}^{n} y_i \log \hat{p}_i + (1-y_i)(\log(1-\hat{p}_i) \\ &= -2\sum_{i=1}^{n} \log(1-|\hat{p}_i - y_i|) \end{aligned}$$

where $\hat{p}_i$ denotes the predicted class probability for individual i, y_i is the actual class outcome of that individual and n is the total sample size. Several measures have been proposed to evaluate the accuracy of a prediction model, that is, how well a model discriminates between individuals with and without the outcome. Among the most popular discrimination measures is the so called c-statistic which in the binary outcome case is identical to the area under the Receiver Operating Characteristic (ROC) curve, which plots the sensitivity (true positive rate) against 1 - (false positive rate) for consecutive thresholds of the class outcome probability. The c-statistic is a rank-order-based statistic of predictions against true outcomes and it is given by

$$c = \frac{1}{n_1 n_2}\sum_{i\in G_1}\sum_{j\in G_2}\left[I\left(\hat{p}_i > \hat{p}_j\right) + 0.5 \times I\left(\hat{p}_i = \hat{p}_j\right)\right]$$

where G_1 and G_2 are the index sets for $y = 1$ and $y = 0$ respectively and n_1 and n_2 are their respective sizes. The value of the c-statistic indicates the extent to which the calibrated class probabilities are higher for individuals in the groups defined by the true outcome y and it varies from 0 and 1. A value of 0.5 indicates that the calibrated class probabilities are distributed randomly among the two classes while a value equal to 1 indicates that all predicted probabilities for individuals with the outcome ($y = 1$) are higher than the those for individuals without the outcome ($y = 0$). As a rank-order-based statistic, the c-statistic is invariant under monotone transformations such as due to calibration errors.

1.6 Outline of the thesis

This thesis consists of a collection of four articles and one book chapter. In principle, the following chapters can be presented in any random order as the related methodologies have been either published in or submitted for publication to independent journals. Yet, we believe that the current structuring of the thesis in the following sequence of chapters respects a series of criteria linking one chapter with the other.

Chapters 2 and 3 are connected as they are both concerned with combination methods for enhancing calibration and prediction. Chapter 2 presents an approach for combining a collection of distinct classification methods whereas chapter 3 reviews various methods, among which an adaptation of the combination approach introduced in chapter 2, for combining a collection of distinct (omic) data sets. In both chapters, the methods are illustrated through the analysis of lower resolution MALDI-TOF proteomic data from a breast cancer case-control study.

The last three chapters deal with higher resolution MALDI-FTICR proteomic data from a pancreatic cancer case-control study. Chapter 4 introduces a method for preprocessing the raw spectra and investigates various methods of summarizing the acquired data using information on either the intensity or the shape of the isotope distributions for prediction purposes. Chapters 5 and 6 are progressions of the work presented in chapter 4. Chapter 5 investigates the problem of simultaneous isotope variable-selection and assessment of the added-value of shape on top of intensity which was left as an open question in chapter 4. Finally, chapter 6 presents methods to deal with the LOD which can be a problem when researchers are provided with a reduced list of peaks where any undetectable peak intensities are reported as missings. Although the limit of detection was not an issue for the methods presented in chapters 4 and 5, since we had access to the complete raw data, we consider the situation where this is not the case and return with a formal solution to the LOD in chapter 6.

In chapter 2, we analyse data collected in a proteomic mass spectrometry case-control study from breast cancer patients and healthy individuals. We present an approach for combining a collection of linear classifiers for the construction of a "combined" classification rule in order to improve calibration and predictive performance. In a first stage this is achieved by building individual classification models separately, each one using the entire proteomic data set to estimate the class-predicted probabilities which will be later used to calibrate the "combined" allocation rule. To estimate the class probabilities from each individual classifier we use a double leave-one-out cross-validatory approach. We evaluate the performance of the combination approach through the breast cancer proteomic data set as well as simulation studies. Results from both the real data analysis and the simulation studies showed that in many cases gains in classification performance and predictive accuracy can be achieved.

While chapter 2 presents methods for combining distinct classifiers, chapter 3 focuses on methods for combining distinct (omic) data sets. More specifically, in this chapter, we review various approaches for combining distinct omic sources in the context of case-control studies with binary outcomes. All the combination-based diagnostic rules discussed in this chapter are based on regularized regression models which are calibrated using a double cross-validatory scheme. The performance of each of the resulting "combined" rules is evaluated with respect to calibration and discrimination and it is compared with the performance achieved when using single-omic data sources. To illustrate the methods we present analyses based on real data from two different studies: 1) the breast cancer case-control study (also considered in chapter 2) in which we examine the combination of two fractions of proteomic mass spectrometry for the calibration of a diagnostic rule for the detection of early-stage breast cancer and 2) the Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) study, a population-based cohort from Finland, in which we consider transcriptomics and metabolomics as predictors of obesity. Results from both studies indicate that improved predictions can be obtained by combining the different omic predictor sources using one of the proposed approaches, as compared to methods based on single-omic data.

Chapters 4 until 6 deal with the additional statistical challenges introduced when analysing high-throughput MALDI-FTICR mass spectrometry data. All analyses presented in these chapters are based on data collected in the context of a pancreatic cancer case-control study. Chapter 4 presents methods to preprocess, summarize and analyse the MALDI-FTICR proteomic data while dealing with the challenges that accompany such data. To preprocess the raw spectra and translate the proteomic expression into a condensed data set we make use of the isotope clustering information, inherent in this specific type of data, and in particular the known statistical properties of the isotopic distribution of the peptide molecules. We present alternative ways to exploit the information on either the intensity level or the shape of the identified isotopic clusters in the individual mass spectra in order to derive summary measures on which diagnostic allocation rules can be based. Our experimental results indicate that both the shape of the isotope clusters and the overall intensity level carry information on the class outcome and can be used to predict the presence or absence of the disease.

The proposed methods and associated analyses presented in chapters 5 and 6 are based and dependent on the data generated by the pre-processing algorithm introduced in chapter 4. Chapter 5 explores the problem of isotope variable selection with paired intensity and shape measurements, derived based on the methods presented in chapter 4. Each pair in our data set corresponds to a distinct isotope cluster and each component within each pair represents a cluster summary of isotopic expression derived based on two different types of information: a) the overall intensity and b) the shape of the observed isotope cluster. Our objective in this work is a) to identify a collection of isotope clusters associated with the disease outcome and b) to optimally integrate the intensity and shape information while maintaining predictive performance. A Bayesian model formulation is used for this purpose which exploits the paired structure of the proteomic data and allows at the same time the evaluation of the added-value of the shape on top of the intensity information. We present a post-hoc analysis which allows researchers to focus on a restricted set of potentially interesting isotope clusters for further investigation and gives insight into the relative predictive capacity of shape when integrated with intensity. We finally present results from a simulation study to demonstrate how the method behaves under a controlled situation.

Chapter 6 considers the case where analysis is hindered due to spectral measurements being subject to the limit of detection. Often in clinical applications statisticians are provided with a reduced list of peaks where any low-concentration proteins below the detection limit, non-detectable by the mass spectrometer, are reported as missing values. Our objective is to investigate whether starting from a reduced set of incomplete peak intensities, we can develop methods which will allow us to construct diagnostic rules of comparable performance as if we had the complete information. We propose the use of censored data methodology to handle spectral measurements within the presence of limit of detection, under the assumption that these have been left-censored for low abundance proteins. The set of incomplete spectral measurements is replaced with estimates of the expected intensity which are then used as new predictor variables for the construction of a prediction model. We combine censored regression with borrowing of information through the addition of an individual-specific random effect formulation in order to account for lack of information and measurement uncertainty. Different variants of using the random censored regression model for the construction and assessment of prediction rules are considered and it is demonstrated how the random effect formulation may additionally allow for variable selection. To evaluate the proposed methods we compare their predictive performance with the one achieved using the complete information as well as competitive methods to deal with the LOD.