Statistical Signal Processing of ESI-TOF-MS for Biomarker Discovery

by

Sai Buddi

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November, 2012 by the
Graduate Supervisory Committee:

Thomas Taylor, Co-Chair
Douglas Cochran, Co-Chair
Randall Nelson
Tolga Duman

ARIZONA STATE UNIVERSITY

December 2012

ABSTRACT

Signal processing techniques have been used extensively in many engineering problems and in recent years its application has extended to non-traditional research fields such as biological systems. Many of these applications require extraction of a signal or parameter of interest from degraded measurements. One such application is mass spectrometry immunoassay (MSIA) which has been one of the primary methods of biomarker discovery techniques. MSIA analyzes protein molecules as potential biomarkers using time of flight mass spectrometry (TOF-MS). Peak detection in TOF-MS is important for biomarker analysis and many other MS related application. Though many peak detection algorithms exist, most of them are based on heuristics models.

One of the ways of detecting signal peaks is by deploying stochastic models of the signal and noise observations. Likelihood ratio test (LRT) detector, based on the Neyman-Pearson (NP) lemma, is an uniformly most powerful test to decision making in the form of a hypothesis test. The primary goal of this dissertation is to develop signal and noise models for the electrospray ionization (ESI) TOF-MS data. A new method is proposed for developing the signal model by employing first principles calculations based on device physics and molecular properties. The noise model is developed by analyzing MS data from careful experiments in the ESI mass spectrometer. A non-flat baseline in MS data is common. The reasons behind the formation of this baseline has not been fully comprehended. A new signal model explaining the presence of baseline is proposed, though detailed experiments are needed to further substantiate the model assumptions. Signal detection schemes based on these signal and noise models are proposed. A maximum likelihood (ML) method is introduced for estimating the signal peak amplitudes.

The performance of the detection methods and ML estimation are evaluated with Monte Carlo simulation which shows promising results. An application of

these methods is proposed for fractional abundance calculation for biomarker analysis, which is mathematically robust and fundamentally different than the current algorithms. Biomarker panels for type 2 diabetes and cardiovascular disease are analyzed using existing MS analysis algorithms. Finally, a support vector machine based multi-classification algorithm is developed for evaluating the biomarkers' effectiveness in discriminating type 2 diabetes and cardiovascular diseases and is shown to perform better than a linear discriminant analysis based classifier.

In memory of my nephew, *Vicky*.

ACKNOWLEDGEMENTS

My dissertation has been a culmination of learning experience both at the technical and emotional level. I owe my gratitude to many people who inspired me and helped me through these years at Arizona State University.

I am greatly indebted to my advisor, Dr. Tom Taylor, for his support and guidance. I have always admired his passion and enthusiasm towards solving new problems and sparking useful insights. He is a very kind person and has always been available for research discussions. I am very fortunate to have him as my advisor.

I am grateful to my co-chair Dr. Douglas Cochran, who has provided helpful research inputs at various stages of this work. I learned the fundamentals of statistical signal processing from his insightful lectures which were vital to my research. I would also like to thank Dr. Tolga Duman for serving in my committee.

I am extremely thankful to Dr. Randall Nelson for welcoming me into the Molecular Biomarkers lab and for his professional, technical, and financial support. I am thankful to Dr. Chad Borges who was always there to answer my questions and help me comprehend the MS technology. I would also like to thank my colleagues in the lab Paul, Doug, Jason, Olga, and Nisha for their support and special thanks to Matt, who carried out experiments and provided useful data for my research.

In a home away from home, there have been many friends who helped me through thick and thin, and made sure that I had a life beyond graduate school. I would like to thank Indro, Eugene, Jon, Ceci, Mike, Varun, Matt, Cierra, Chiu, members of the Outing club and the Couchsurfing community for the adventurous and joyous moments through these years.

Finally, I would like to thank my parents, my siblings and their families for everything that they are to me. Without them, I would not be here.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The use of statistical signal processing techniques has been well established in a wide variety of applications in the past years. Traditionally, signal processing methods have been used with signals generated from classic physical phenomena such as electromagnetism. Quite often these signals are treated as stochastic processes to exploit their statistical properties. Such methods have been used in defense, communications, speech, and image systems among many others. Over the past few decades, signal processing in non-traditional signals has been gaining popularity within the research community. Such signals include, but are not limited to, measurements from biological, psychological, and seismological processes. Application of signal processing methods in these new research frontiers are based on the assumption that the underlying mathematical principles are compatible.

Many of the above mentioned applications require extraction of a signal or parameter of interest from degraded measurements. This is often accomplished by deploying fine grained statistical models, multidimensional signal representations or acquiring more spatial and temporal information through sensors. These approaches can be used to develop highly sensitive signal detection and estimation algorithms which can exploit statistical differences between signal and noise. Noise is a random disturbance generated in systems due to one of many reasons. Though various techniques exist to reduce the noise energy, it is never eliminated completely. The goal of signal processing algorithms is to be able to detect when events occur or decide which event has occurred by discerning the subtle differences between an information bearing and a random noise pattern. One of the ways of accomplishing this is by deploying stochastic models of the signal and noise observations based on some fundamental assumptions and how the noise interacts with

the information bearing signal when measurements are taken. The assumptions made for these statistical models should ideally mimic the underlying fundamental principles of the system. Sometimes the stochastic models have one or more parameters that describe the underlying properties of the system. Often these parameters have to be estimated from the degraded data as they are essential to completely describe the stochastic model of the system.

The fundamental theory behind detection and estimation has been developed in mathematical statistics and decision theory and the early application of these theories were by RADAR researchers. These were extended to more engineering systems such as SONAR, communication, speech, image etc., and recently to many non-traditional fields such as biological systems. This dissertation explores the use of statistical signal processing ideas to understand the properties of signal generated from protein mass spectrometry and introduce the detection and estimation problems in the measured data. There has been a few attempts in applying such techniques in this particular area of research but most of them are based on heuristics rather than mathematical rigor. The first step in developing any detection or estimation scheme is to adequately model the observed data. Here signal and noise models are developed by studying the fundamental principles of the measuring device and then finding a solution to the above mentioned detection and estimation questions under certain assumptions.

1.1 Likelihood Ratio Test

The signal detection problem in signal processing is generally formulated as a decision making problem, which is tackled by a statistical hypothesis test on the measured data. The most basic form of such a test is the binary hypotheses of whether an event has occurred or not, also known as the alternate and the null hypothesis

respectively. The Neyman-Pearson lemma [6] lays out the guidelines to reject one hypothesis in favor of the other using the likelihood ratio test (LRT), which is an uniformly most powerful test among competitor test models. Using probability density function, the likelihood describe a function of parameters given the measured data. The likelihood ratio essentially characterizes how much likely the measured data is under one model compared to the other. LRT compares the ratio of two statistical models, describing the presence and absence of an event, to a threshold. The optimum threshold is generally determined from the probability of false alarm or by plotting the receiver operating characteristic (ROC) plot. ROC is a plot of probability of detection (P_d) against the probability of false alarm (P_f) as a function of the threshold. The application of LRT spans numerous research areas [7].

One of the most basic LRT models is a signal embedded in additive white Gaussian noise with known parameters. The result of the LRT under such assumptions is the matched filter [8] which has been popular in numerous applications due to its simple and elegant form. In many cases, the matched filter provides satisfactory results irrespective of whether the model assumptions are valid. In certain applications, within the additive noise paradigm, noise statistics are not well modeled by a Gaussian distribution and is accompanied by unknown or nuisance parameters that lead to an incomplete model description under one or both hypotheses. Under such situations, an LRT is not possible and often a maximum likelihood estimate (MLE [9]) of the unknown parameters is used for the test and is known as the general likelihood ratio test (GLRT). Though GLRT does not usually lead to the optimal solution, it has been popular due to the ease of implementation and less restrictive assumptions compared to alternatives such as a Bayesian approach which requires the probability density function of the unknown parameter. GLRT has been used in traditional systems and recently in some non-traditional systems.

In particular, a Gamma distributed noise model will be an important theme in this dissertation.

For example, in recent years, voice activity detection (VAD) has become an emerging research problem due to the need of efficient use of limited bandwidth. VAD algorithms based on LRT has been proposed and shown to have good performances [10, 11]. In the conventional VAD algorithms, where the statistical models operate in the discrete Fourier transform (DFT) domain, the distributions of noisy speech spectra and noise spectra are assumed to be complex Gaussians [10]. Chang et. al. [12] used a Laplacian probability density function (PDF) to model the distributions of noisy speech spectra and noise spectra which was shown to be a better model for the distribution of clean speech [13, 14]. In [12], the variance of the Laplace distribution is unknown under the alternate hypothesis. The ML estimate of the variance is obtained by using a power subtraction method.

Recently, Shin et. al. [15] reported that the generalized gamma distribution provides a better model of the distribution of clean speech spectra than the Gaussian, Laplacian or Gamma PDFs. VAD algorithms based on GLRT, using the generalized Gamma distribution, has been demonstrated in [16, 17]. The ML estimate of the unknown parameters of the generalized gamma distribution is obtained by using a gradient descent algorithm, since an analytical solution could not be obtained [15].

In [18], the authors try target detection and parameter estimation using GLRT and propose an iterative GLRT for multi-input multi-output radar systems. The noise is assumed to be independent and identically distributed (IID) Gaussian with zero mean and unknown covariance matrix. For multiple targets, multiple hypotheses are used, assuming the total number of targets is known apriori. This is computationally intensive since it needs to search the multi-dimensional parameter

space and an iterative GLRT requiring an one-dimensional search, is proposed. The detection of radar targets against a background of unwanted clutter due to echoes from multiple unwanted surfaces has been a fundamental problem in radar signal processing [19]. Solution to the problem generally requires an understanding of the noise statistics before a detection scheme can be proposed. In [20] the problem of radar detection is handled by considering a log-normal clutter with white Gaussian noise. To understand the clutter model, attempts have been made to fit empirical models to data collected with radar systems. This is a reasonable approach for determining PDF of the clutter and such models have been proposed [21, 22].

Functional information from magnetic resonance imaging can be obtained using statistical tests based on the magnitude of image reconstructions. In [23], the authors propose a detector for fMRI using GLRT. It exploits the common phase property between the fMRI response signal and the baseline component. The noise is considered to be white Gaussian with an unknown variance and the signal model has unknown parameters as well. ML method is used to estimate these unknown parameters. Then a Monte Carlo simulation is performed to plot the ROC and find the exact threshold to use for the test. The GLRT method is compared with the more common magnitude correlation and complex correlation tests and is shown to perform better.

Seismic stations are a part of the International Monitoring System for the Comprehensive Test Ban Treaty. In [24], a GLRT based outlier detection method is proposed for the identification of seismic activities that might be nuclear explosions, using a large set of measured earthquake data. The test will look for outliers from the earthquake data. The noise is assumed white Gaussian with unknown mean and variance.

Considering these signal processing applications in a variety of research

fields, this dissertation delves into the realm of mass spectrometry. Despite the existence of the mass spectrometry technology for quite some time, very little effort has been made to understand and model the statistical characteristics of the underlying processes, for both signal and noise, with the same mathematical vigor.

1.2 Signal Detection in Mass Spectrometry

Mass spectrometry (MS) technique is used extensively in the field of biochemistry to study the presence of biomolecules such as proteins and peptides in a sample. Of the variety of MS methods, time-of-flight mass spectrometry (TOF-MS), involves first ionizing the molecules of interest and then separating them according to their mass-to-charge ratio by accelerating them using an electromagnetic field. The time taken by these ions to travel from the source to the detector is a measure of their mass-to-charge ratio. The ions form signal peaks at the detector with intensities proportional to their abundance in the sample. These signal peaks are usually embedded in noise from various sources. Hence, peak detection is an important step in MS based data analysis for protein/peptide identification. Many detection algorithms have been proposed according to the signal, noise sources, and ionization method.

For low-resolution peaks, one easy way to find peaks is to smooth the spectrum and then take the local maxima exceeding a certain threshold value [25]. A simple peak finding (SPF) thresholding algorithm based on the first derivative to find peak flanks is used in [26]. Wallace et al. [27] presented a different technique. The algorithm starts by finding the point in the raw spectrum that is farthest from the baseline formed by connecting the first and last points. Once a point with greatest orthogonal distance from the line has been identified, it joins the collection of strategic points and, in turn, becomes an end point for two new line segments from

a point with greatest orthogonal distance. This numerical scheme is performed until the greatest orthogonal distance to any end-point connecting line segment drops beneath a prescribed threshold value. Jarman et al. [28] used a metric called the intensity weighted variance (IWV) to check whether the histogram within a sliding window (ion counts vs. time or m/z bins), with varying width, resembles a uniform distribution or has a peaked shape.

Filters are used to enhance the resolution of a mass spectrum and remove background noise [29]. Wavelet transforms have been used to separate overlapping signals and detect peaks [30, 31]. Gras et al. [32] used a matched filter approach, using a Gaussian peak shape template, to locate potential peaks and then performed a non-linear regression to adjust the peak width and height. Overlapping peaks are found by subtracting the fitted models from the raw data and repeating the algorithm. Andreev et al. [33] used an algorithm called matched filtration with experimental noise determination (MEND) for peak picking. The matched filter is used in the frequency domain with a Guassian shape for the chromatographic peak.

Peak detection is vital for many MS applications such as peptide/protein identification, biomarker analysis etc., which require a robust detection scheme. This dissertation focusses on the biomarker analysis application. Biomarkers are a variety of biomolecules that are potentially capable of discriminating different states of a disease. TOF-MS is one of many techniques used to identify these biomarkers.

1.3 Mass Spectra Classification for Biomarker Analysis

Biomarkers are generally a set of protein molecules which can discriminate between different states of a given disease. These biomarkers are identified by analyzing the mass spectra of biological samples, such as serum, which have been labeled as one

of several disease states. A model is built using the concentrations (or abundances) of this labeled data set measured from the mass spectra. This model can then be used to diagnose new mass spectra cases by assigning them to one of the predefined classes. As a result, the biomarker discovery can be cast as a mass spectra classification problem. A variety of classification algorithms are used to model and evaluate the performance of biomarkers in discriminating disease states.

Fisher discriminant analysis (FDA [34, 35]) is one of the most popular approaches for biomarker classification. Dimensionality reduction methods, such as principal component analysis (PCA [36]), are used prior to FDA to keep the number of variables less than the population size. This is the case in [37] for studying prostate cancer biomarkers. Wu et al. [38] selected a variable set of size 15 by using a t-test score and applied a number of classification algorithms including linear discriminant analysis (LDA) with high accuracy. Qu et al. [39] used Mahalanobis distance to select a set of 11 wavelet coefficients before applying LDA for prostate cancer. The resulting classifier attained 96.7% sensitivity and 100% specificity on an independent test set. Similarly, Baggerly et al. [40] applied LDA to study lung cancer biomarkers.

K-nearest neighbor (KNN [41]) classification algorithm has been used in [42, 43]. A comparison study in [44] showed that Naive Bayes and KNN had roughly equivalent performance in ovarian cancer diagnosis. On a different ovarian cancer dataset, KNN outperformed LDA. Decision trees and rules (DT [45]) are sequential classification methods and use an embedded feature selection process. DTs have been used for biomarker analysis in various cancer studies [46–48]. However, DTs are shown to perform worse than other classification methods [44, 49].

Support vector machines (SVM [50]) have been applied extensively for classification and dimensionality reduction. In [51], features were selected from a large

set of variables by using SVM accuracy as a fitness function. SVMs have been used for studying biomarkers for diabetes and cardiovascular diseases in [52, 53]. SVMs have been shown to be one of the top performing biomarker classification algorithms when the number of features is very small and as dimensionality increases, their advantage over other methods becomes more pronounced [44, 54].

1.4 Research Motivation and Objective

The main objective of this dissertation is to understand the statistical characteristics of signal and noise in TOF-MS employing electrospray (ESI) method for ionizing molecules. Mathematical modeling of the observed data is the first step towards building robust detection and estimation algorithms. Though ESI-TOF-MS has been adopted widely in recent years as a technique for immunoassay biomarker analysis, this popularity has not translated to a better understanding of the statistical properties of the MS data. This is because the fundamental operations of mass spectrometers are considered too complex and involve many unknown variables for any reasonable modeling. Since the mass spectrometers are commercial instruments, their design details and data processing algorithms, used prior to the data being available to the user, are not readily accessible. Also, a highly interdisciplinary nature of the problem makes it difficult for model development. As a result, the statistical techniques for mass spectrometric data analysis are based on heuristics and some qualitative understanding of the process. In a recent paper [55], the authors have developed a statistical model using some of the basic characteristics of ESI-TOF-MS process where the rate of arrival of ions in the instrument is preserved by the ion detection scheme. However, when this assumption is not true for alternate ion detection methods, this model is not appropriate. As this MS technique is widely used for biomarker analysis, it is important to understand the stochastic

properties of the data and develop a model for peak detection.

The signal model is developed using the device's operating process. Time resolution issues based on a TOF instrument's physics have been studied in [56]. Similar principles are used in ESI-TOF-MS for developing a signal shape model. The noise model is developed by studying experimental data obtained under different conditions and with varying parameters of the instrument. Based on these noise and signal models, signal processing based peak detection and amplitude estimation algorithms are developed. Then an application of these algorithms for calculating fractional abundances for biomarker analysis is suggested. Finally, an SVM based multi-classification algorithm is introduced to study the effectiveness of potential biomarkers in predicting cardiovascular risks in type 2 diabetes patients.

1.5 ORGANIZATION

The rest of the dissertation is organized as follows. The concept of mass spectrometric immunoassay (MSIA) is introduced in Chapter 2. MSIA is a biomarker analysis technique that employs ESI-TOF-MS and thus serves the motivation for studying the statistical properties of TOF-MS data. The basic working process of a TOF-MS device, specifically, the data acquisition process of the Bruker's microTOFQ ESI mass spectrometer is explained in detail. Understanding the device and the underlying processes are important in constructing mathematical assumptions and models.

Chapter 3 is the core of this dissertation where the basic signal and noise models, with detailed analysis of the device physics and MS data, are developed. The signal model utilizes the basic physics of the mass spectrometer in terms of spatial and energy distributions to calculate the peak width. An algorithm for isotopic distribution estimation is introduced. Detector limitation is also considered

towards advancing a complete signal model. Experimental data is used for the noise model. Careful experiments are performed under different parameter settings and chemical compositions. Histograms and goodness of fit tests are performed on these observed data to define the noise probability density function. A time varying non-flat baseline is a typical, yet not well understood, phenomenon in TOF-MS. Along with discussing the traditional method of dealing with the baseline, a new manner of considering it as part of the desired signal is proposed. Once the adequate signal and noise models are built, robust peak detection schemes are explored. Specifically, an approximation to the optimal NP detector and a detector based on the general likelihood ratio test (GLRT), both under an additive noise model, are developed. Finally, a maximum likelihood (ML) method for estimating unknown signal amplitude is developed.

Fractional abundance estimation of potential biomarker molecules is an important step for biomarker discovery. In Chapter 4, the concept of fractional abundance calculation from ESI MS data is explained. A deconvolution algorithm, similar to the one used by the ESI instrument manufacturer, is described in detail. Additionally, an automated abundance calculation routine is introduced. Various limitations of these type of algorithms are discussed. Then the performance of the detection and estimation methods developed in the previous chapter are evaluated using Monte Carlo simulations. A new technique, that is principally different from the current deconvolution based algorithms and based on the above mentioned detection and estimation methods, is proposed for abundance calculations.

In Chapter 5, a support vector machine algorithm for multi-classification is proposed. Assessing the effectiveness of biomarkers in characterizing diseases is another important step in biomarker discovery and classification algorithms are necessary for such assessments. A set of biomarkers, from various cohorts, are

identified using MSIA and their abundances are calculated. The SVM algorithm is then used to evaluate the effectiveness of the biomarkers for predicting cardiovascular risk in type 2 diabetes patients. Few variations of the SVM algorithm along with an LDA algorithm are compared and the results are discussed.

Chapter 6 discusses the conclusions derived from this statistical signal processing of ESI-MS-TOF and multi-classification algorithm. Also, possible future opportunities for signal processing in mass spectrometry are discussed.

CHAPTER 2

IMMUNOASSAY

2.1 INTRODUCTION

Immunoassay is a technique used in biochemistry to detect and quantify a particular analyte in a solution, which frequently contains a complex mixture of molecules. Analytes can be ligands, proteins, antibodies etc. present in a clinical sample such as blood serum. Immunoassays are based on the fact that antibodies have a high affinity and specificity to bind to one or a limited group of molecules, called antigens. Immunoassays can be performed to quantify either an antigen or an antibody. In addition to the requirement of high binding specificity and affinity between antibody-antigen pairs, immunoassays require a means to produce a measurable signal in response to the bindings. Different immunoassays accomplish this by different methods. In radioimmunoassay (RIA [57]), antigens are made radioactive and mixed with an appropriate antibody. The radioactivity of the mixture in conjunction with a standard curve gives a measure of the amount of antigen in the sample. In enzyme linked immunosorbent assay (ELISA [58]), an enzyme linked secondary antibody is used to detect the antibody-antigen complex. The enzyme emits a fluorescent signal when certain chemical is added to the mixture. Other techniques include memory lymphocyte immunostimulation assay (MELISA [59]), magnetic immunoassay (MIA [60]) etc. Mass spectrometric immunoassay (MSIA [61], [1]) uses TOF-MS for identification and quantification of target antigens and their variants.

The MSIA concept is shown in Fig 2.1. As in any immunoassay method, the antibodies are incubated with the antigen sample. This is followed by repetitive washes of the antibody-antigen complex to get rid of any non-specific binding. A matrix solution is then used to extract (by means of absorption) the antigen onto a

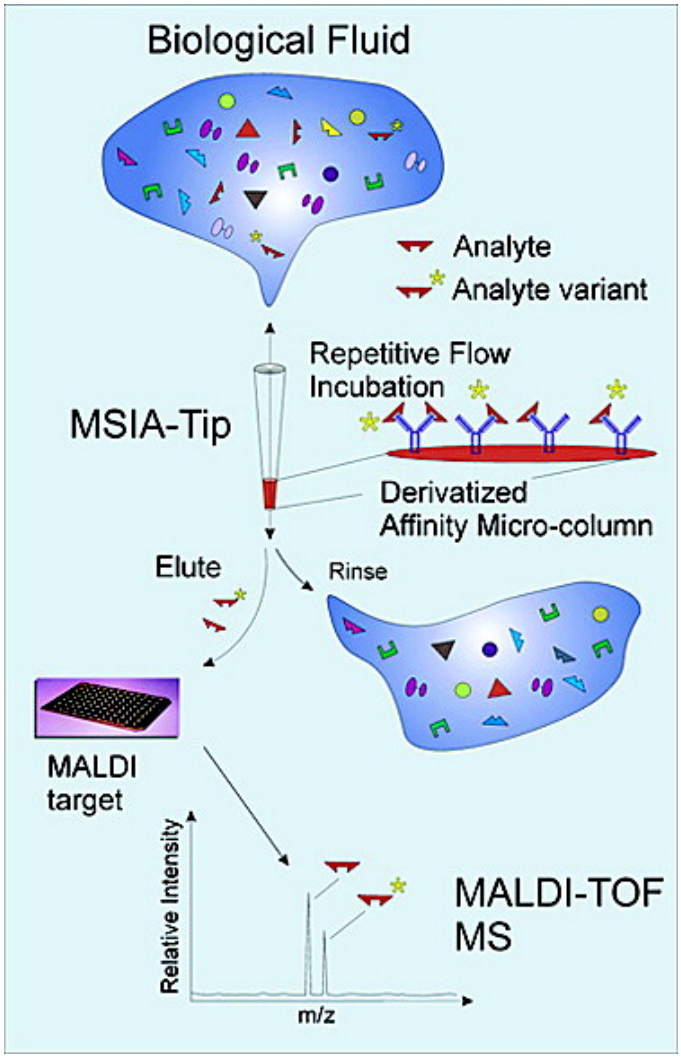


Figure 2.1: Mass Spectrometric Immunoassay concept [1]

mass spectrometer probe. The dried matrix solution is then put in a time of flight mass spectrometer to identify all the molecular variants present in the sample. TOF-MS is a very accurate technique to identify and estimate the abundance of molecular species.

2.2 Time of Flight Mass Spectrometry (TOF-MS)

Mass spectrometry is used for detecting molecular species in a sample by estimating their masses (via mass-to-charge ratio) accurately. MS begins with the ionization of

molecules at the source. These ions are then accelerated and separated according to their mass-to-charge ratio by electromagnetic fields. In TOF-MS [2], the instrument measures the time taken by an ion to travel from the source to detector, which is a function of the ion's mass (m) and charge (z). This follows from the kinetic energy equation,

$$\frac{1}{2}mv^2 = zV \tag{2.1}$$

where v is the velocity of the ion and V is the voltage across the source. The ions travel through a field free drift region (at a constant velocity v) before hitting the detector. The TOF for the ion traveling a drift distance D is,

$$t \approx \left(\frac{m}{2zV}\right)^{\frac{1}{2}} D \tag{2.2}$$

Some basic configurations of the TOF-MS analyzer are shown in Fig 2.2. The ionization technique used at the source can be any of the following: electron ionization (EI [62]), chemical ionization (CI [63]), plasma desorption mass spectrometer (PDMS [64]), laser desorption (LD [65]), matrix-assisted laser desorption/ionization (MALDI [66]), and electrospray ionization (ESI [67]- [68]).

MALDI is used to analyze biopolymers such as proteins and peptides. These molecules tend to fragment when ionized by traditional methods. MALDI provides a soft ionization by using a laser beam to vaporize and ionize the molecules. A matrix solution is mixed with the analyte (protein sample) and the solution is spotted on to a MALDI plate. The matrix provides protection from the destroying laser beams. When the laser is fired at the MALDI spots, the matrix absorbs the energy and becomes charged first. This charge is then transferred to the analyte molecule. Usually a proton is added to each molecules forming singly charged ions. Sometimes multiple charged ions can also be created.

The ESI technique is useful for large molecules as the tendency of these molecules to fragment during ionization is largely avoided. In this method, the analyte is dispersed into fine aerosol. A drying gas is used to evaporate the charged solvent molecules. The ions are then passed on to the TOF unit through various focussing stages. ESI almost always results in multiple charged ions resulting in a low mass-to-charge (m/z) ratio. The detailed construction and working of a typical ESI equipment is discussed in the next section.

2.3 ESI-TOF-MS

The basic block diagram of the Bruker's ESI-TOF-MS system is shown in Fig 2.3. The spectrometer is divided into four stages, each with multiple substages.

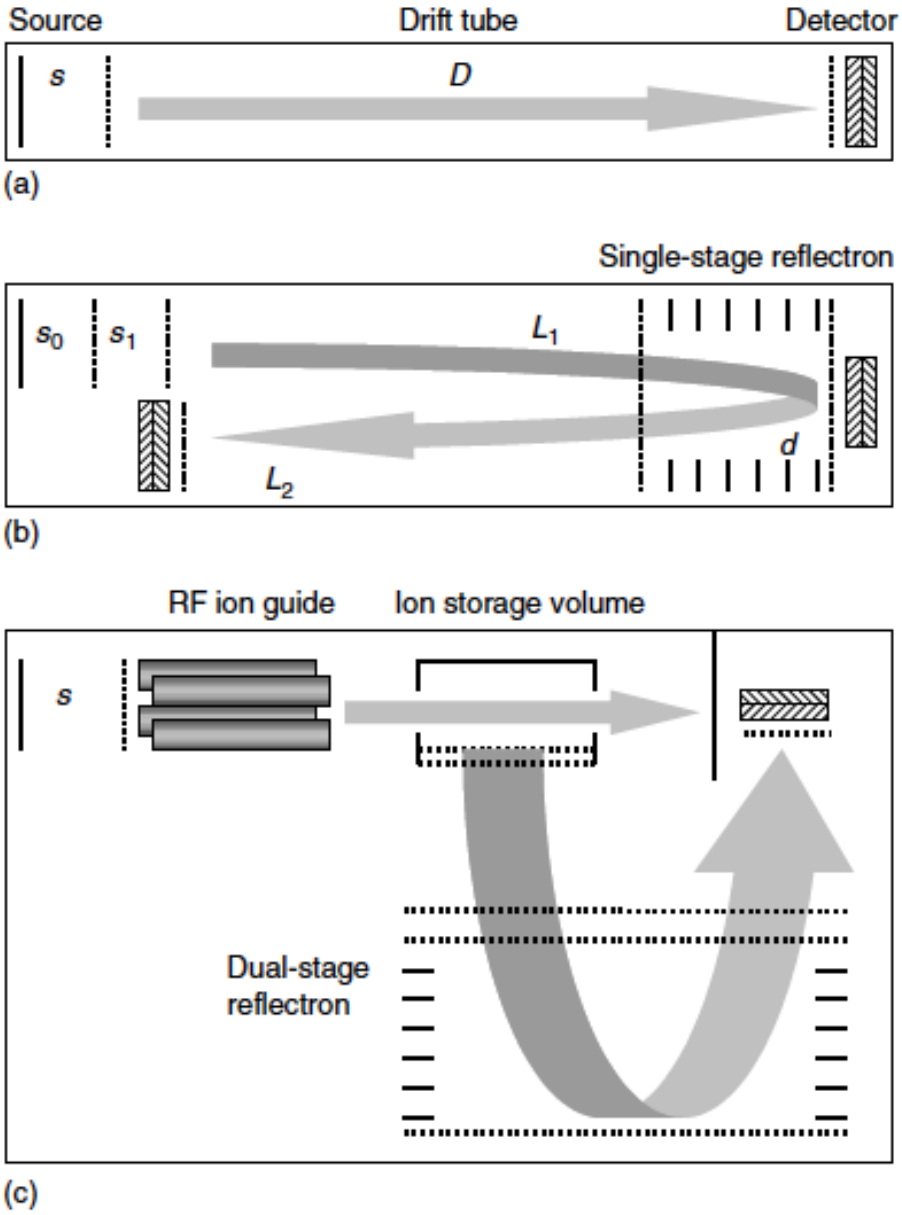


Figure 2.2: Basic configurations of time-of-flight mass spectrometers: (a) a simple linear TOF mass analyzer with a single-stage ionization source, (b) a reflectron TOF mass analyzer with a dual-stage ionization source, and (c) an orthogonal acceleration mass analyzer with a quadrupole ion guide and a dual-stage reflectron [2]

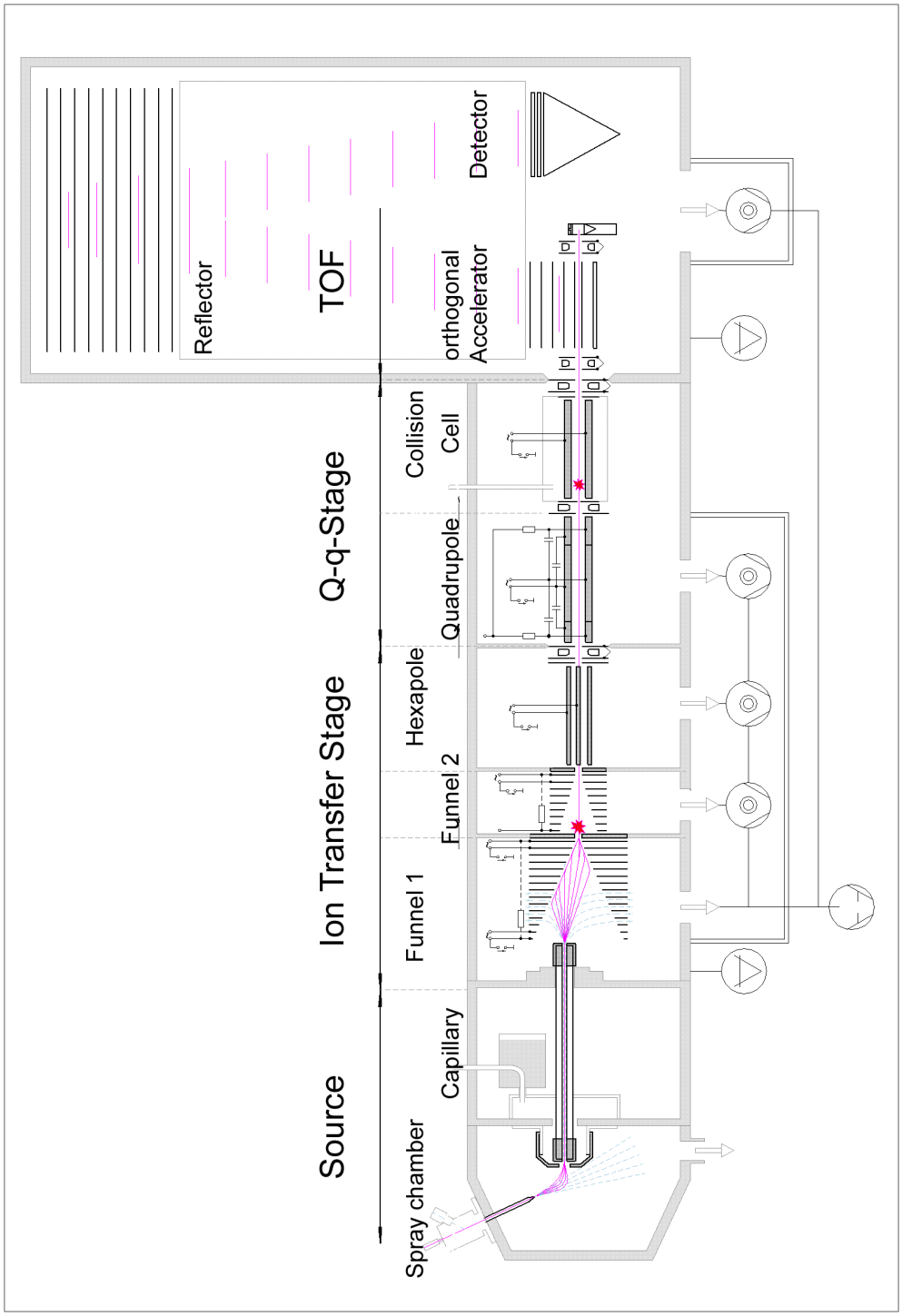


Figure 2.3: Block diagram of ESI-TOF-MS [3]

2.3.1 Source Stage

The function of the spray chamber is to ionize the sample molecules in solution after converting them into fine droplets, also called nebulizing. The nebulizer receives the sample and solvent from a liquid chromatograph (LC). This solution is sprayed into the chamber through a needle along with a pressurized gas, such as nitrogen, converting the solution into tiny droplets. The droplets are charged due to the high nebulizer voltage while the needle itself is at ground potential. The spray shield is kept at a negative voltage, thus channeling the positively charged droplets into the capillary. Heated nitrogen gas is used to evaporate the solvent in the charged droplets before they enter the glass capillary. This process requires high energy so that desolvation occurs without thermally decomposing the sample molecules and keeping a low droplet temperature. The ions are emitted out of the droplet when the electric field generated by the surface charges exceeds the surface tension of the droplet. The capillary transfers these ions, along with the heated gas and some solvent, from a high voltage source stage at atmospheric pressure to a low voltage vacuum system in the ion transfer stage.

2.3.2 Ion Transfer Stage

The ion transfer stage has three of the five vacuum substages. The pressure is systematically reduced in each passing substage. The two funnel substages are used to remove the dry gas and solvent with minimal ion loss. This is achieved by creating a DC voltage gradient, along the length of the funnel, to guide the charged particles while the uncharged particles are pumped away through the gaps in the funnel. The hexapole substage focusses the ions along the axis. The hexapole stage ends with a gate and focusing lens. The focusing lens forms a suitable beam shape for transferring the ions into the quadrupole of the Q-q stage.

2.3.3 Q-q Stage

The quadrupole is used as an additional ion guide and may also be used for isolating a defined mass range. In the collision cell, the isolated ions can be fragmented by using a neutral collision gas. The ions cool down due to the low pressure and are focussed very close to the axis. The collision cell ends with a gate and a transfer lens. The gate voltage is controlled such that the accumulated ions of certain mass range are transferred to the TOF-stage. This transfer time defines the beginning of the pre-pulse storage time of the next TOF voltage pulse and limits the transferred mass range. Together with the entrance lens of the following orthogonal accelerator the transfer lens provides a suitable parallel beam shape inside the acceleration stage. For an ion of mass m with a charge z, m/z defines the mass to charge ratio. The same molecular species of mass m can have a random number of multiple charges thus differing in m/z ratio.

2.3.4 TOF Stage

The layout of the TOF assembly is shown in Fig 2.4. The orthogonal accelerator stage consists of electrodes mounted on top of each other. These electrodes, except for the one at the base, are shaped like slot diaphragms. When ions from the collision cell move into the pulsing region of this stage, a high voltage pulse is applied to accelerate them towards the reflector through the slits of the electrodes. The pulses are timed such that the ions have enough time to reach the detector before the next batch of ions from the collision cell arrive. If the electrodes are at ground potential, or if the voltage pulse is not applied at the right instant, the incoming ions flow towards the secondary electron multiplier (SEM) and are lost for TOF analysis. The SEM is used for monitoring and troubleshooting the ESI system. The accelerated ions pass through the field free region to the reflector which helps in normalizing

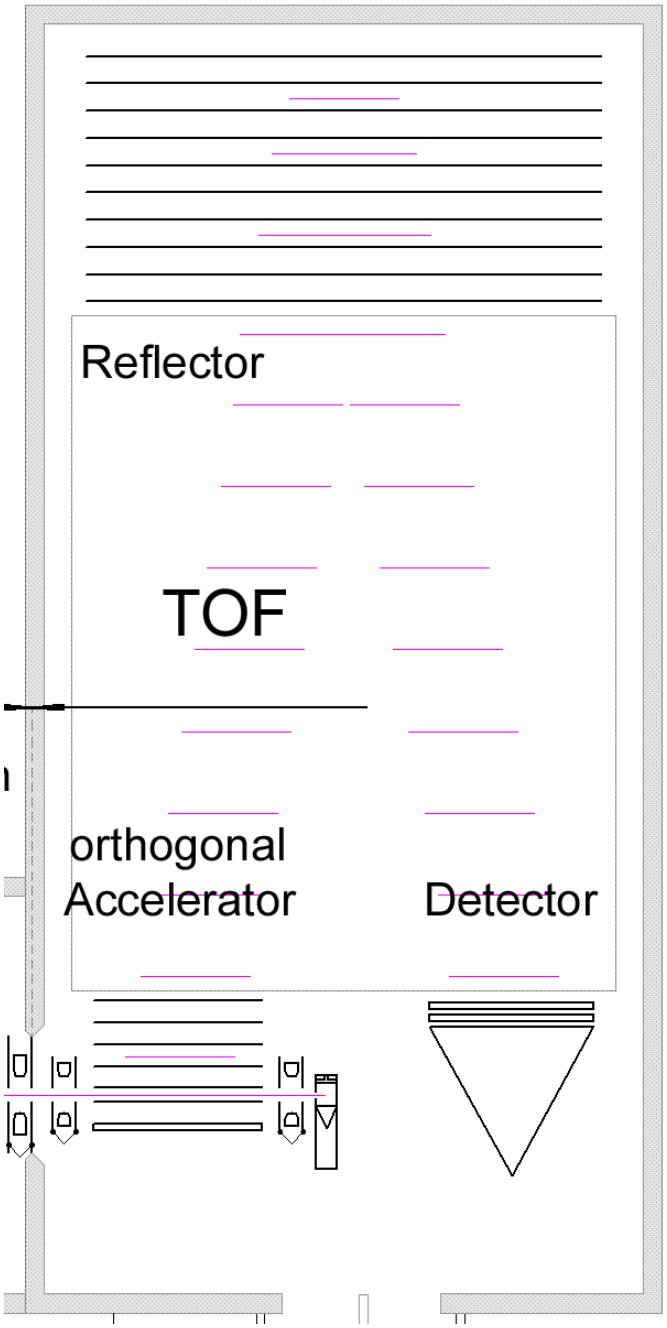


Figure 2.4: TOF assembly [3]

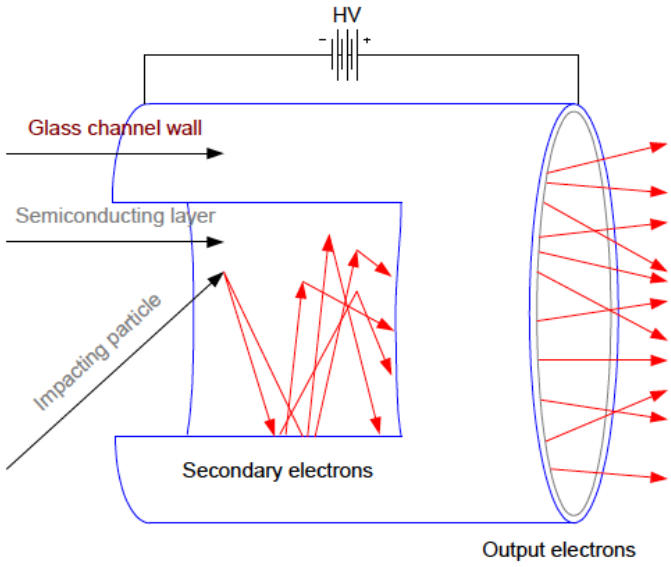


Figure 2.5: Electron multiplication in MCP detector [3]

the energy difference of ions, thus improving resolution. The detector, usually a micro channel plate (MCP [69]), converts an ion signal into an electrical signal. An MCP is a solid core assembly with millions of small pores, called microchannels, which are coated with a semi-conductive layer. Each channel works as an independent electron multiplier when an ion hits that channel, as shown in Fig 2.5. A high voltage is applied across the MCP resulting in a current flow through the channels. A high frequency analog to digital converter (ADC) measures the digital output from the detector.

2.4 Mass Spectra Data Generation

TOF mass spectrometers are used to accurately determine the masses of individual species in a sample. The mass spectrum is calculated from the time of flight spectrum of ions traveling from the accelerator to the detector, i.e., the time taken by individual ions to travel this distance is recorded. The recording process discussed here is specific to Bruker‘s ESI mass spectrometer [70].

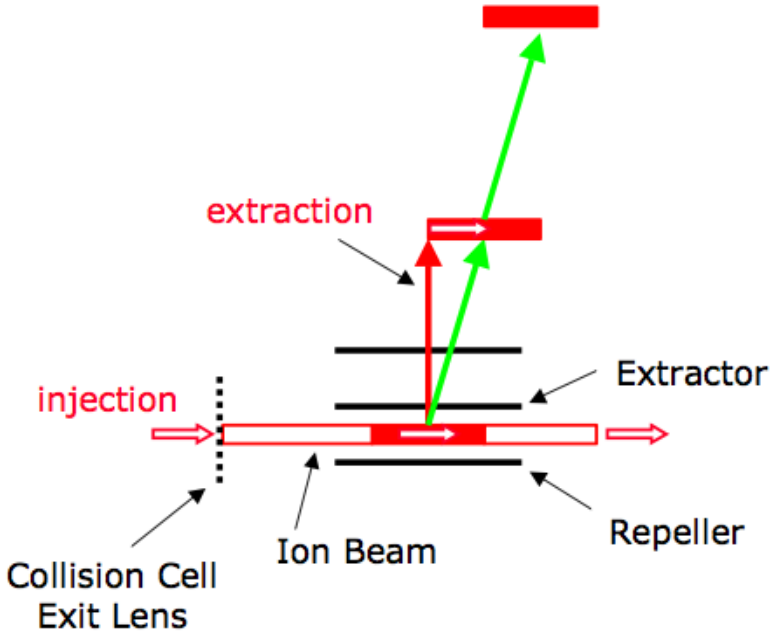


Figure 2.6: Orthogonal accelerator [3]

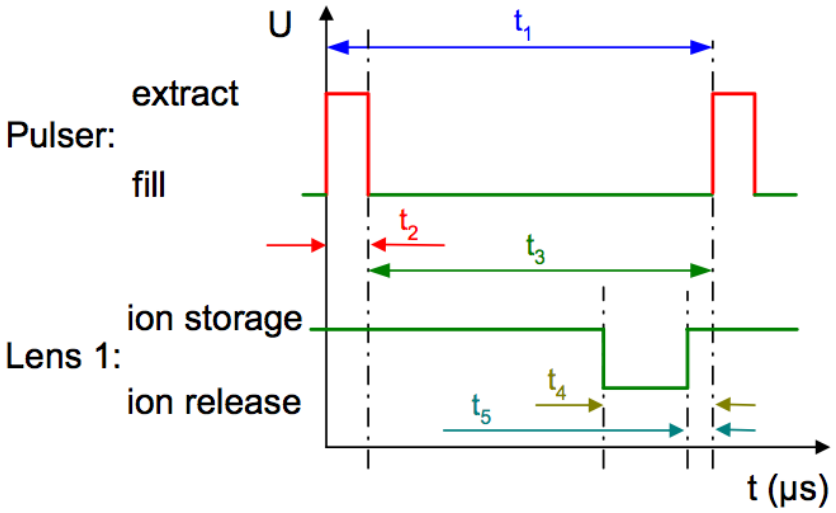


Figure 2.7: Accelerator pulse timing [4]. t_1 is the pulser time period with t_2 the on time and t_3 the off time. t_4 is the transfer time from the collision cell to the accelerator region. t_5 is the pre-pulse accelerator storage time.

Fig 2.6 shows the accelerator region of the TOF stage. The ions are moved into the accelerator region by applying an appropriate voltage at the collision cell lens. As shown in Fig 2.7, the lens voltage is turned on for t_4 microsecond before a short accelerating voltage pulse is applied between the repeller and extractor electrodes. The pulse on time, t_2, is on the order of 5 microseconds. The pulse off time, t_3, can be on the order of few hundreds of microseconds, depending on the mass of the ions being analyzed. Ions with larger m/z ratio take longer to reach the detector compared to those with smaller m/z ions (Eqn 2.2). This means that the pulse frequency is on the order of tens of kilohertz, i.e., thousands of TOF spectra are generated each second.

The ions hitting the detector, which may consist of more than one multichannel plate, create an avalanche of electrons. An ADC, present in the transient recorder and operating at a digitization rate of a few gigahertz, measures this ion current. The transient recorder is equipped with algorithms to check for the presence of ion peaks and to calculate time of flight and intensity for each individual TOF spectrum. The time and intensity pairs are stored in memory cells. For a 2 gigahertz ADC, the time resolution (bin) is 500 picoseconds. The high frequency ADC ensures high resolution and accurate time of flight calculation. If an individual ion hit generates an ADC output of 5, an 8-bit ADC will saturate if more than 50 ions hit the detector simultaneously. For an accelerator pulse with time period 100 microseconds ($t_1 = 100\mu s$), 200,000 memory cells are needed to store each TOF spectrum. The computation and memory requirement costs add up quickly as 10,000 spectra are generated each second and the usual recording time lasts a few minutes. Also, ions do not arrive at the detector in every 500 picosecond time bin, resulting in a lot of empty memory cells. Due to these reasons, thousands of spectra are usually added together, over a certain time period, and the resulting sum

spectrum (one frame) is stored in memory. There can be hundreds of frames in one MS data set, depending on the duration of the MS analysis. The TOF spectrum is then converted into a mass spectrum using a quadratic calibration equation. The calibration curve is generated by fitting time of flight data for a variety of known molecular samples.

It is worth noting that, unlike the ADC used in this device detector, counting detectors use a time-to-digital converter (TDC) to convert the electronic signal from the electron multiplier into a digital TOF signal. TDCs record the time of arrival of ions and cannot distinguish between multiple ions but they are robust to the variable gain of the electron multiplier such as MCP, unlike the ADC.

2.5 CONCLUSION

This chapter provided a brief introduction to immunoassay methods. MSIA is the motivation for studying the ESI-TOF-MS process and understanding the device physics, especially in the TOF stage. The data generation process at the detector of the TOF stage is complex. It involves amplification due to the electron multiplier and a peak detection algorithm before the data is digitized by ADC and stored. In the next chapter, a signal peak shape model is developed by using the physics of the TOF stage along with isotopic distribution and detector limitations. The effect of solvent molecules and detector thresholding is investigated to suggest a distribution for the chemical noise. This leads to the development of LRT based signal detector.

CHAPTER 3

STATISTICAL ANALYSIS OF MS DATA

3.1 INTRODUCTION

This chapter deals with developing a statistical model for the MS data from ESI MS TOF. The MS data (M_s) has typically been written as a sum of signal (s), noise (w), and baseline b.

$$M_s = s + w + b \tag{3.1}$$

While there is some value that maybe obtained from such a model, a new and more detailed signal model that takes the baseline into account has been proposed. Signal shape has been studied for various mass spectrometry applications. A Gaussian peak shape is considered for building a matched filter in [71], whereas in [72] an asymmetrical shape with a Gaussian leading edge and a Lorentzian trailing edge is proposed. A modified Gaussian model is used in [73]. In [74, 75], the peak shape is modeled as a sum of two Gaussian functions which are shifted relative to each other on the mass scale for a least square fit. Peak shape in the form of convolution of a Gaussian with an exponential function for the falling edge is used in [76]. Wavelet based peak shape is used in [31] for peak detection algorithms.

The current methods fit a theoretical peak shape to the signal by optimizing some parameters to minimize an error criterion. In section 3.2, a new method of estimating the peak shape and width by employing first principles calculations based on device physics and molecular properties is developed. Known issues that degrade signal resolution such as isotopic, spatial, and energy distributions are considered and a mathematical model for each is developed to account for the peak shape and width. Such a method is mathematically tractable and provides a sound basis for any peak shape assumptions.

Noise is ubiquitous in MS data which may result in the masking of signals

of interest. Studies about noise in MALDI [77, 78] describe three possible sources of small scale variability. Chemical noise due to matrix ions, shot noise due to the discrete nature of ions, and Johnson noise due to the electrical system. Chemical noise is one of the most significant sources of background in ESI mass spectra [79]. However, modeling of the MS-TOF noise has not been the primary target of interest in mass spectrometry studies. In most cases, noise is assumed to be Gaussian. Deisotoping, a method of recognizing peptides from MS peaks due to isotopes, requires modeling of noise. Methods described in [80–82] implicitly assume a Gaussian noise model. A multinomial noise model is used in [83]. However, the noise models are not usually verified with experiments to affirm the model assumptions. In section 3.3, a new chemical noise model is developed by investigating the experimental data from the mass spectrometer unlike the previous models suggested in the literature. Careful experiments are carried out by controlling various parameters of the device and goodness-of-fit tests are used to estimate the probability density with a fair degree of certainty.

Apart from the signal and additive noise, a non-flat baseline is commonly seen in the presence of sample ion peaks. Usually the signal spectrum sits on top of a time varying baseline. The exact cause of the baseline shape is not understood and is generally attributed to chemical noise and detector saturation leading to a slowly decaying charge. In other words, the baseline is considered as an artifact and various algorithms [84–91] have been proposed to estimate and remove it before further analysis is done. A more careful analysis of the baseline is proposed in section 3.4 and the statistical properties of chemical noise in the presence of such a baseline is investigated. Again, statistical goodness-of-fit tests are used to estimate the probability density of the baseline with noise from experimental data. This new approach eliminates the need for baseline correction algorithms since it is included

in the noise and signal model.

Detection procedures based on the likelihood ratio test are developed in section 3.5. The model assumes an additive noise. Though matched filters have been used [32, 71] for signal detection, they are heuristic methods and not based on a sound analysis of the signal and noise models. There have been a few likelihood ratio based detection schemes [55,83,92] for counting ion detectors. However, such tests have not been demonstrated for ADC based ESI-MS-TOF devices before. The statistical signal processing concepts applied to mass spectrometry has the potential to fundamentally change the way MS analysis is carried out currently, as it does not require for any pre- and post-processing steps and introduces a mathematically robust detection scheme.

3.2 Signal Peak Shape

Knowledge of the peak shape is necessary in many mass spectrometry applications. Known MS equipment physics can be used to come up with a theoretical peak shape. Resolution is the sharpness of a signal peak corresponding to any molecular species in the MS. Mathematically, as a standard, resolution (R) has been defined as the ratio of the mass value (m) and linear width (Δm) at half the signal height. Signal resolving power becomes an issue when the goal is to distinguish and quantify large protein molecules and their variants present in a sample. In MALDI, the instrumental resolving power begins to wane above the m/z range of about 30,000 [93]. MALDI typically produces singly charged ions, thus the m/z values are equivalent to the molecular weights. ESI, on the other hand, produces multiple charged ions, depending on the size of the molecules. This makes it possible to measure and distinguish large molecular weight proteins at high resolution. A higher resolution usually results in a better mass accuracy. Typically, modern ESI spectrometers

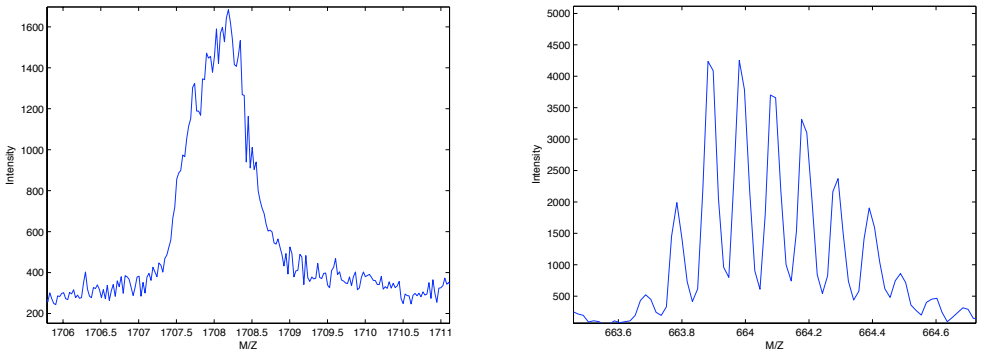
achieve resolutions in the order of 10^5. Though the ionization process differs between MALDI and ESI, the TOF sections work on the same principles. The time resolution issues of TOF-MS are discussed in [56, 94–96]. If all ions started in a plane at the center of the accelerating electrodes and with zero initial velocity, the flight time would be the same for all ions which had the same m/z. In such a case, the resolution is limited only by the detector. However, the resolving power of the TOF-MS depends on the time spread caused by the initial spatial and initial kinetic energy distributions as well as the ability to design the device to reduce that spread. To summarize, the TOF-MS usually suffers from the following effects that degrade signal resolution.

- Isotopic mass distribution - f_{isot}
- Space distribution - f_T
- Energy distribution - $f_{\mathscr{E}}$
- Detector limitation - f_{det}

The MS signal for a given molecular species can be modeled by understanding the physics of the instrumentation and mathematically quantifying these limitations. As the effects act independently, the signal shape f_S is the convolution of the distributions,

$$f_S = f_{isot} * f_T * f_{\mathscr{E}} * f_{det} \tag{3.2}$$

This is a new method of estimating the peak shape and width by employing first principles calculations based on device physics and molecular properties. Isotopic, spatial, and energy distributions degrade the signal resolution and a mathematical model for each is developed to provide a sound basis for any peak shape assumptions.



(a) ESI MS for DBP

(b) ESI MS for ApoC1

Figure 3.1: ESI MS peak shape for a particular charge state due to isotopic mass spread

3.2.1 Isotopic Distribution

Presence of isotopes is not a limitation of TOF-MS in itself, but they usually result in degraded resolution. Isotopes are variants of atoms with different number of neutrons in their nucleus. For example, there are two stable isotopes of carbon that appear in nature: C^{12} and C^{13}. They both have the same number of protons (atomic number = 6) but C^{13} has an extra neutron and hence a different atomic mass. The natural abundances of the elemental isotopes are known. Molecules contain elemental isotopes according to their natural abundances. The probability of occurrence of these isotopic variants (i.e., the isotopic distribution of a molecule) can be calculated from the atomic composition and the known elemental isotope abundances.

Proteins are usually made of thousands of atoms of H, C, N, O, S etc. When a protein sample is analyzed in a TOF mass spectrometer, instead of a single peak corresponding to the average molecular mass of the protein, multiple peaks corresponding to the isotopic distribution can exist. If the spectrometer has enough

resolution, the isotopic peaks can be resolved, especially for low mass proteins. For the high molecular weight proteins, usually investigated in ESI-TOF-MS, these peaks partially coalesce together resulting in an effective broadened peak. Fig 3.1 shows the MS peaks for vitamin D binding protein (DBP) and Apolipoprotein C-1 (ApoC1) for a particular charge state. For the low mass ApoC1, the isotopic peaks are well resolved but only partially resolved for DBP.

Calculation of isotopic distribution has been explored extensively in the field of mass spectrometry [97]. Earlier methods [98–101] used various stepwise and multinomial expansion techniques. However, due to computational efficiency and low memory requirements, convolution based methods [102–105] are currently popular. The method of [103] is based on using the nucleon count instead of the exact mass of molecules. This approximation is then adjusted to the exact mass after calculating the distribution using convolution.

To illustrate the use of convolution, consider the chlorine atom (Cl) and the molecule (Cl_2). Chlorine has two naturally occurring isotopes. Cl^{35} has a mass of 34.969 with a relative abundance of 0.7553. Cl^{37} has a mass of 36.966 with a relative abundance of 0.2447 [106]. Ideal isotopic distribution of chlorine atom can thus be represented as:

$$y_{Cl}(m) = 0.7553\,\delta(m-34.969) + 0.2447\,\delta(m-36.966)$$

where $\delta(\cdot)$ is the Dirac delta symbol. The isotopic distribution of Cl_2 can be calculated by considering it as the sum of two Cl atoms, i.e., the convolution of the two distributions:

$$\begin{aligned} y_{Cl_2}(m) &= [y_{Cl} * y_{Cl}](m) \\ &= 0.5705\,\delta(m-69.938) + 0.3696\,\delta(m-71.935) \\ &+ 0.0598\,\delta(m-73.932) \end{aligned}$$

This method of calculating the isotopic distribution can be easily implemented in the Fourier transform domain as the convolution is replaced by a product operation.

$$\begin{aligned} Y_{Cl}(\omega) &= 0.7553\, e^{i34.969\omega} + 0.2447\, e^{i36.966\omega} \\ Y_{Cl_2}(\omega) &= [Y_{Cl}(\omega)]^2 = 0.5705\, e^{i69.938\omega} + 0.3696\, e^{i71.935\omega} + 0.0598\, e^{i73.932\omega} \end{aligned}$$

The above method can be extended to all other molecules. For example, consider a hydrocarbon with n_1 carbon atoms and n_2 hydrogen atoms ($C_{n_1}H_{n_2}$). The isotopic distribution of the molecule can be written as the product of isotopic distributions of the carbon and hydrogen atoms in the Fourier domain ($Y_C(\omega), Y_H(\omega)$) and then converted to the mass domain by the inverse Fourier transform (IFT).

$$\begin{aligned} Y_{C_{n_1}H_{n_2}}(\omega) &= [Y_C(\omega)]^{n_1}[Y_H(\omega)]^{n_2} \\ y_{C_{n_1}H_{n_2}}(m) &= IFT[Y_{C_{n_1}H_{n_2}}(\omega)] = \sum_l Y_{C_{n_1}H_{n_2}}(\omega)\, e^{ilm\omega} \end{aligned}$$

The normalization factors are omitted.

To generalize this method for any molecule, consider the isotopic abundances as a mathematical product:

$$\prod_j [E(j)]^{n_j}$$

where $E(j)$ is the isotopic abundance, in Fourier domain, of the j^{th} element with n_j total atoms. Let m_{jk} be the mass of the k^{th} isotope of the j^{th} element with a natural abundance p_{jk}. The corresponding nucleon number, $m'_{jk} = [m_{jk}]$, is the nearest integer. The isotopic distribution is the convolution of individual elemental compositions. But it can be represented as a product in the Fourier domain as follows:

$$Y(\omega) = \prod_j \left[\sum_k p_{jk} e^{-i\omega m'_{jk}} \right]^{n_j} \tag{3.3}$$

To adjust the mass scales, $Y(\omega)$ is moved so that the center is at origin and then transformed to the mass domain using inverse Fourier transform as follows:

$$y(m^*) = IFT\left[Y(\omega)e^{i\omega M''_{av}}\right] \tag{3.4}$$

where $M''_{av} = [M'_{av}]$. M'_{av} is the average molecular mass of the molecule under consideration, using m'_{jk}. $y(m^*)$ is the isotopic distribution centered at the origin with a peak spacing of exactly 1 dalton. A correction to the mass scale is obtained by the following transform:

$$m = \frac{\sigma}{\sigma'}m^* + \frac{\sigma}{\sigma'}(M''_{av} - M'_{av}) + M_{av} \tag{3.5}$$

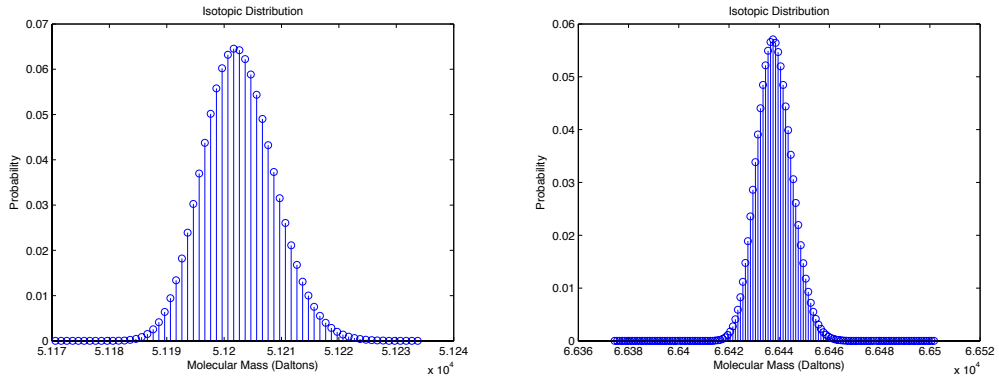
where σ and σ' are the standard deviations of the isotopic distribution using m_{jk} and m'_{jk} respectively. Fast Fourier transform (FFT) can be used for faster computation and efficient memory utilization. The total number of FFT points is chosen as the first power of 2 greater than or equal to $K(1+\sigma'^2)$. K is an integer between 1 and 10. The calculation is further simplified by using polar coordinates for the Fourier signal instead of the cartesian real and imaginary parts.

In TOF-MS, positive ions can be considered as protons (H^+) being added to the molecule. So, for the isotopic distribution of the ion in m/z axis can be achieved by using the following transform on Eqn 3.5:

$$m_z = \frac{m + z \times m_{H^+}}{z} = \frac{m}{z} + m_{H^+} \tag{3.6}$$

where z is the total charge of the ion and m_{H^+} is the mass of a proton. Fig 3.2 shows the isotopic distribution for DBP ($C_{2240}H_{3525}N_{582}O_{717}S_{35}$) and human serum albumin (HSA - $C_{2936}H_{4591}N_{786}O_{889}S_{41}$) molecules without any ionization and $K = 10$. The isotopic distribution ($f_{isot}(m_z)$) is transformed from the m/z axis to the time axis by the quadratic calibration equation which is of the form,

$$\frac{m}{z} = C_0 + C_1 t^2 \tag{3.7}$$



(a) DBP (b) HSA

Figure 3.2: Isotopic distribution

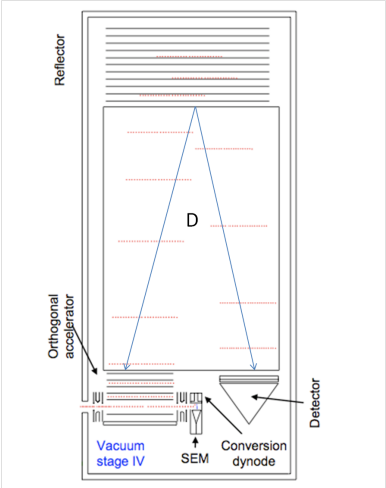
The calibration equation is used by the mass spectrometer for transforming the MS signal from time domain to the m/z domain. The equation is derived by using calibrate molecules with known masses and measuring the time taken for them to traverse through the TOF stage.

3.2.2 Spatial Distribution

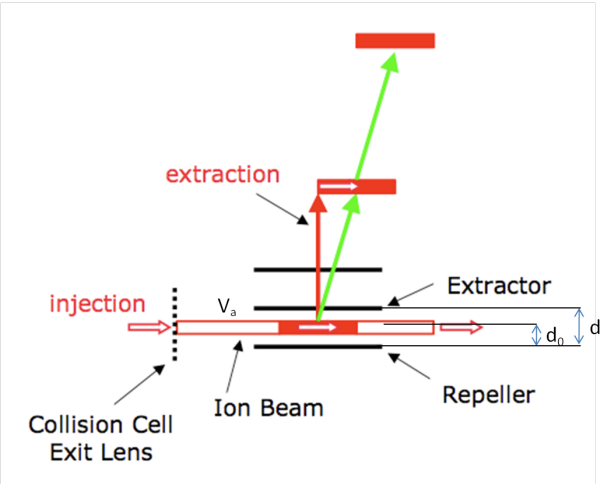
As seen in Fig 3.3(b), ions are focussed along the axis when entering the orthogonal accelerator region. The spatial distribution of ions between the electrodes will result in a velocity distribution and thus a TOF distribution; the ions closer to the repeller plate remain under acceleration longer and hence acquire a larger velocity. The larger the variance in ion spread, the larger the variance in TOF. The TOF distribution can be modeled by assuming a spatial distribution for the ions.

As shown in Fig 3.3(b), let d be the width of the accelerator region and d_0 be the initial position (from the repeller plate) of an ion with mass m and charge z when the pulse accelerating voltage V_a is applied between the electrodes. The initial position of the ion is a random variable. The potential at position d_0 is

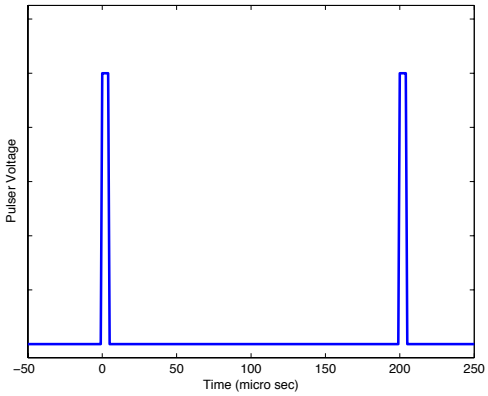
$$V_0 = V_a \frac{(d-d_0)}{d} = E_a(d-d_0) \tag{3.8}$$



(a) TOF Assembly



(b) Ion Acceleration



(c) V_a with $T = 200\mu S$, $t_{on} = 5\mu S$

Figure 3.3: TOF ESI-MS

where $E_a = \dfrac{V_a}{d}$ and is assumed to be uniform. If v_d is the drift velocity of the ion coming out of the accelerator,

$$\frac{1}{2}mv_d^2 = zV_0 \implies v_d = \sqrt{\frac{2zV_0}{m}} = \sqrt{\frac{2zE_a}{m}(d-d_0)} \tag{3.9}$$

As expected, ions starting near the bottom of the accelerator region (small d_0) will move with a higher velocity compared to ions starting near the extractor plate.

The time spent by the ion, moving with an acceleration $a = \dfrac{zE_a}{m}$, inside the

accelerator region is:

$$t_1 = \frac{v_d}{a} = \sqrt{\frac{2m}{zE_a}(d - d_0)} \tag{3.10}$$

The voltage pulse has an usual ON time, $t_{ON} = 5\mu S$. $t_1 < t_{ON}$ under the normal operating conditions of the mass spectrometer.

The time taken by the ion to travel through the field free drift distance, D (Fig 3.3(a)),

$$t_2 = \frac{D}{v_d} = \sqrt{\frac{mD^2}{2zE_a(d - d_0)}} \tag{3.11}$$

The reflector is used to normalize the spatial distribution since ions starting near the repeller plate will travel faster and thus penetrate deeper into the reflector. However, these ions will be ejected with the same velocity from the reflector. The time spent in the reflector, assuming an uniform decelerating electric field E_r (deceleration = $\frac{zE_r}{m}$) is:

$$t_3 = \sqrt{\frac{8mE_a}{zE_r^2}(d - d_0)} \tag{3.12}$$

The total time of flight, from Eqns 3.10, 3.11, and 3.12:

$$t = t_1 + t_2 + t_3 = \frac{\kappa_1}{\sqrt{d - d_0}} + \kappa_2\sqrt{d - d_0} \tag{3.13}$$

where, $\kappa_1 = \sqrt{\frac{mD^2}{2zE_a}}$ and $\kappa_2 = \sqrt{\frac{8mE_a}{zE_r^2}} + \sqrt{\frac{2m}{zE_a}}$.

The TOF, t, is a function of the initial position d_0. By substituting $x = \sqrt{d - d_0}$, then Eqn 3.13 becomes:

$$t(x) = \frac{\kappa_1}{x} + \kappa_2 x \tag{3.14}$$

$t(x)$ is plotted in Fig 3.4, with $t_{min} = 2\sqrt{\kappa_1\kappa_2}$ at $x = x_1 = \sqrt{\frac{\kappa_1}{\kappa_2}}$. It can be argued that an ion starting close to the repeller plate (smaller d_0 or larger x) will move with a faster velocity, hence will have a smaller t, compared to a similar ion starting near

the extractor plate. Hence, t should decrease with increasing x or $x \in [0, x_1]$ from Fig 3.4. Therefore, using Eqn 3.14, x can be written as:

$$x = \frac{t - \sqrt{t^2 - 4\kappa_1\kappa_2}}{2\kappa_2} \tag{3.15}$$

The initial position of the ion is a random variable with probability density function (PDF) $f_D(d_0)$. If x has a PDF $f_X(x)$, then (since $x = \sqrt{d - d_0}$) using the fundamental theorem of probability,

$$f_X(x) = \frac{f_D(d_0)}{|x'|} = 2x f_D(d_0) \qquad \text{for} \quad x \geq 0$$

If d_0 is assumed to have a Gaussian distribution, $d_0 \sim \mathcal{N}\left(\frac{d}{2}, \sigma^2\right) = f_D(d_0)$ then $x^2 = (d - d_0) \sim \mathcal{N}\left(\frac{d}{2}, \sigma^2\right) = f_D(d_0)$. Thus, the above equation can be written as,

$$f_X(x) = 2x f_D(x^2) \qquad \text{for} \quad x \geq 0 \tag{3.16}$$

From Eqn 3.14, the time distribution can be written in terms of the spatial distribution as follows:

$$f_T(t) = \frac{f_X(x)}{|t'(x)|} = \frac{x^2 f_X(x)}{|\kappa_2 x^2 - \kappa_1|} \qquad t \geq t_{min} > 0 \tag{3.17}$$

with x and $f_X(x)$ as defined in Eqns 3.15 and 3.16, respectively.

3.2.3 Energy Distribution

Similar to spatial distribution, ions may also enter the accelerator region with different velocities and thus kinetic energies. The energy variance of ions is minimized by controlling the temperature and lens voltage and typically has insignificant effect on signal resolution.

Let T be the temperature of a molecule of mass m when entering the accelerator region. The velocity of the molecule, u, can be approximated as Maxwell-Boltzman distribution assuming gaseous phase ions as follows:

$$f_U(u) = \sqrt{\frac{m}{2\pi kT}} \exp\left(-\frac{mu^2}{2kT}\right) \tag{3.18}$$

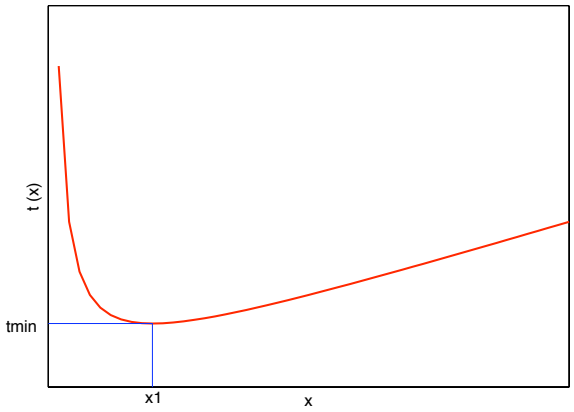


Figure 3.4: TOF as a function of ion position x

The standard deviation of this velocity is:

$$\Delta u = \sqrt{\frac{kT}{m}} \tag{3.19}$$

Similarly, the standard deviation in drift velocity (from Eqn 3.9) due to the initial position of the molecules, and hence the potential is:

$$\Delta v_d = \sqrt{\frac{z}{2mV_0}}\,\Delta V_0 \tag{3.20}$$

where z is the charge state of the molecule. V_0 depends on the initial position d_0 of the molecule and from Eqn 3.8,

$$\Delta V_0 = -V_a \frac{\Delta d_0}{d} \tag{3.21}$$

The total time of flight for a given molecule and charge state is, from Eqns 3.10, 3.11, and 3.12:

$$t = \frac{D}{v} + \frac{3v}{a} \tag{3.22}$$

It is assumed that the accelerator and reflector fields are equal ($E_a = E_r$) such that $t_3 = 2t_1$ and $v = u + v_d$. The pulse width in time of flight is:

$$\Delta t = \left(-\frac{D}{v^2} + \frac{3}{a}\right)\Delta v = \left(\frac{3}{a} - \frac{D}{v^2}\right)(\Delta u + \Delta v_d) \tag{3.23}$$

Substituting the values of Δu and Δv_d from Eqns 3.19, 3.20, $a = \frac{zE_a}{m}$ and assuming $u \ll v_d$ such that $v^2 \approx v_d^2$:

$$\Delta t = \left(\frac{3m}{zE_a} - \frac{mD}{2zV_0}\right)\left(\sqrt{\frac{kT}{m}} + \sqrt{\frac{z}{2mV_0}}\Delta V_0\right) \tag{3.24}$$

Re-arranging the terms, the above equation can be written as:

$$\frac{\Delta t}{m_z} = \alpha_1 + \frac{\alpha_2}{\sqrt{m_z}} \tag{3.25}$$

where $m_z = \frac{m}{z}$, and

$$\alpha_1 = \sqrt{\frac{kT}{m}}\left(\frac{3}{E_a} - \frac{D}{2V_0}\right) \tag{3.26}$$

$$\alpha_2 = \left(\frac{3}{E_a} - \frac{D}{2V_0}\right)\frac{\Delta V_0}{\sqrt{2V_0}} \tag{3.27}$$

Eqn 3.25 is a line with an intercept α_1 and slope α_2. By plotting this line from available data, the exact values of these constants can be estimated. The time distribution due to the energy spread is:

$$f_{\mathscr{E}}(t) \sim \mathscr{N}(0, \Delta t^2) \tag{3.28}$$

3.2.4 Limitation of Detector

The multi channel place (MCP) at the detector, works by secondary electron multiplication effect in the channel. This by itself causes the current signal to spread. Also, depending on the penetration depth of the ion in the channel, the trigger time for the electron avalanche will be different. These effects along with the finite scan rate of ADC limits the time resolution in MS-TOF. The pulse shape is considered to be Gaussian with a standard deviation of 1.5 nanosecond [70].

$$f_{det}(t) \sim \mathscr{N}(0, 2.25) \tag{3.29}$$

3.3 NOISE IN ESI-TOF-MS

Noise is introduced into the MS signal due to various reasons. The noise can be classified into chemical noise, shot noise, and Johnson noise. Chemical noise is one of the most significant source of background distortion in ESI. It is caused by the presence of solvent molecules. Even though the solvent is more volatile than the analyte, it is difficult to get rid of all solvent molecules by use of a heating gas or other such methods. They form clusters with each other and with analyte molecules [107]. Sometimes there are chemical impurities in the sample or in the MS equipment itself. These clusters and impurities get charged and traverse the flight path to reach the detector, just like the analyte molecules of interest. Impurities usually give rise to fixed pattern noise at specific m/z values but the solvent clusters cause interference in virtually all m/z values. Fragmentation of any of these various molecules at any step of the TOF-MS process adds to the complex nature of chemical noise.

Limited studies have been carried out to study the statistical nature of MS noise, especially for ion counting detectors. The noise distribution in these kind of detectors can be described by the Poisson distribution [55, 108]. In [92], a combination of multinomial and Poisson noise models is used for deisotoping with improved results. Counting detectors use a time-to-digital converter (TDC) to convert the electronic signal from the electron multiplier into a digital TOF signal. TDCs record the time of arrival of ions and cannot distinguish between multiple ions. They are robust to the variable gain of the electron multiplier such as MCP, unlike the ADC which measures the total signal output after the ion impact. Digital thresholding [70,109] to suppress unwanted interference further confounds the noise analysis and therefore a Poisson distribution is not the appropriate noise model for these

detectors. Furthermore, the proposed noise models are not usually verified with experimental data. Keeping these issues in mind, a new chemical noise model is developed next. Careful experiments are carried out by controlling threshold parameters of the device and goodness-of-fit tests are used to estimate the probability density with a fair degree of certainty unlike the previous suggested models.

3.3.1 Noise Statistics of ESI-TOF-MS

One source of noise in the detector arises from the dark current of the MCP. Thermal emission of electrons from in the channels of the MCP gives rise to an avalanche effect, knocking off more electrons along the channel. Under normal operating conditions, the dark current is low but as each frame of MS data is a sum of thousands of spectra, this current adds up. A frame is a snapshot of the sum spectra, generally over 1 second. Hence, if the MS analysis is carried out for 10 minutes, there will be $\sim$600 frames in the data set. A chromatogram is the plot of the total intensity (or mass) in the frames over time. The threshold is a negative noise suppression voltage (V_{TN}) applied at the anode of the detector. Dark current can be analyzed by MS analysis without any solvent or sample. The chromatogram of this analysis is shown in Fig 3.5. The mean, median, and variance are calculated per frame, using all intensities values in that frame. This gives a sense of the mean noise intensity and variance, which may not be inferred from the chromatogram plot alone.

Each point in the chromatogram plot is the sum of all noise intensities in that frame. The mean, median, and standard deviation plots (calculated from the intensities in each frame) show that the noise intensities are consistent throughout the frames. Fig 3.6 shows a typical MS frame under the positive V_{TN} and the histogram of the intensities through out the frame. The second histogram plot is for the m/z bin 1500-1510 taken from all 550 frames in the dataset. All the histograms

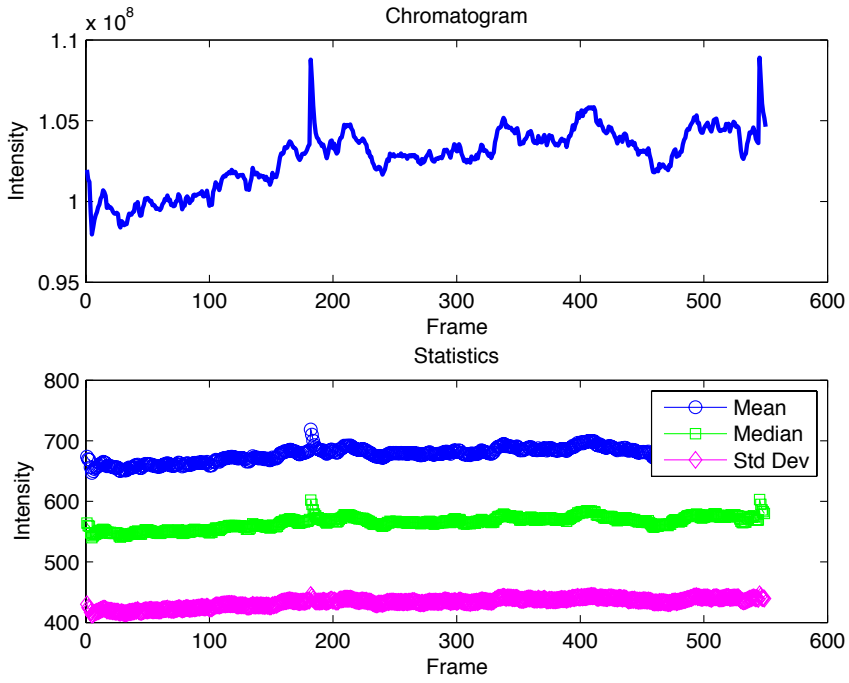


Figure 3.5: MS analysis of vacuum at $V_{TN} = 0.0001V$

are plotted with a bin size of 1 as the ADC gives an integer output as intensities.

The spectrometer is operated at a negative V_{TN} (-0.0023 V) for noise suppression. The chromatogram of the MS data obtained at this threshold is shown in Fig 3.7. As seen, almost all the dark current noise is eliminated at this threshold level.

Solvent induced clusters are major contributors to the chemical noise in ESI. The solvent is a mixture of Acetonitrile (CH_3CN) and water. The concentration of CH_3CN is increased from about 10% to 80% through a single run of a sample. This trend is easily noticed by looking at the chromatogram in Fig 3.8. A typical frame and its histogram is shown in Fig 3.9. Histograms are plotted in Figs 3.10 and 3.11 for data with m/z bin size of 10 taken from 10 frames at a time.

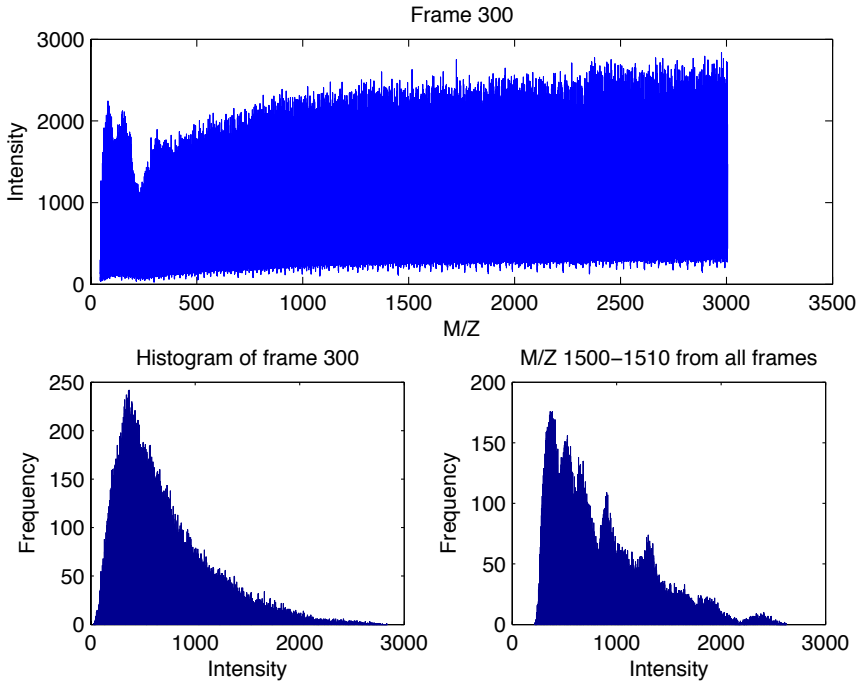


Figure 3.6: Typical frame and histogram at $V_{TN} = 0.0001V$

3.3.2 Goodness of Fit Test

A probability distribution model is desired to statistically describe the chemical noise in ESI MS-TOF and to enable development of suitable statistical signal processing algorithm. The test of goodness of fit (GOF) is used to evaluate the agreement between the distribution of these sets of observations and a theoretical distribution. Several such tests are described in statistical literature. Kolmogorov Smirnov test (KS test [110]) is a well known and popular GOF assessment.

The data is thought to have a theoretical cumulative distribution function (CDF) $F_0(x)$, $F_n(x)$ denotes the observed cumulative step function of n observations. The KS test statistic is the distance between the two in the supremum norm,

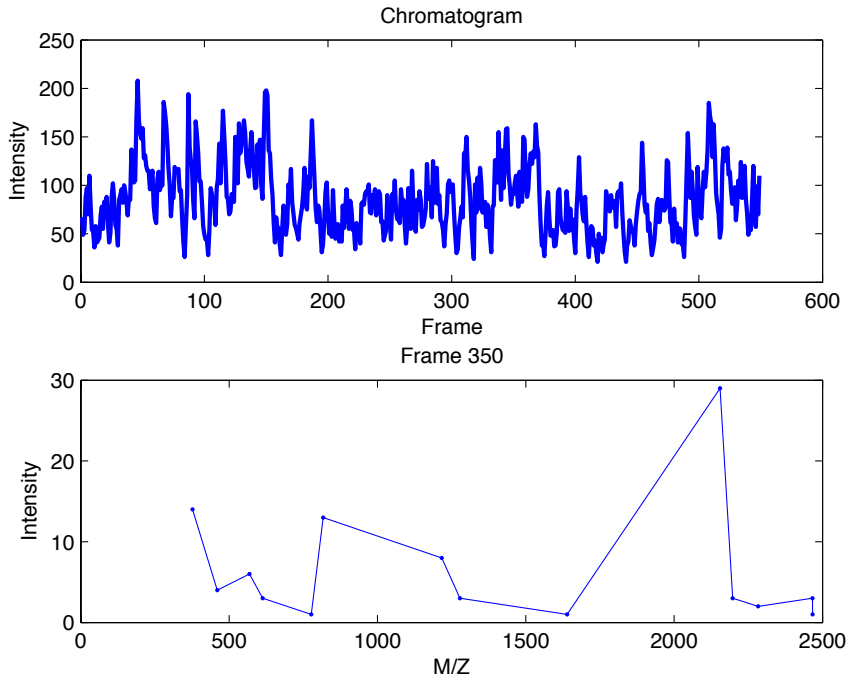


Figure 3.7: MS analysis of vacuum at $V_{TN} = -0.0023V$

i.e.,

$$d_n = \sup_x |F_0(x) - F_n(x)| \tag{3.30}$$

Under the null hypothesis that the dataset comes from the hypothesized distribution $F_0(x)$, d_n converges to Kolmogorov distribution for large n [111]. The null hypothesis is rejected at a significance level α if $\sqrt{n}d_n > k_\alpha$, where k_α is found from the Kolmogorov distribution table such that $P(K \leq k_\alpha) = 1 - \alpha$. In other words, the null hypothesis is rejected when the p-value is less than α. The p-value is the probability of observing a value as large as the calculated d_n, if the null hypothesis is true.

A simple modification of the KS test is the two-sample KS test which can be used to test if two sets of observations have the same underlying probability distribution functions. If $F_{n_1}(x)$ and $F_{n_2}(x)$ are the cumulative step functions of

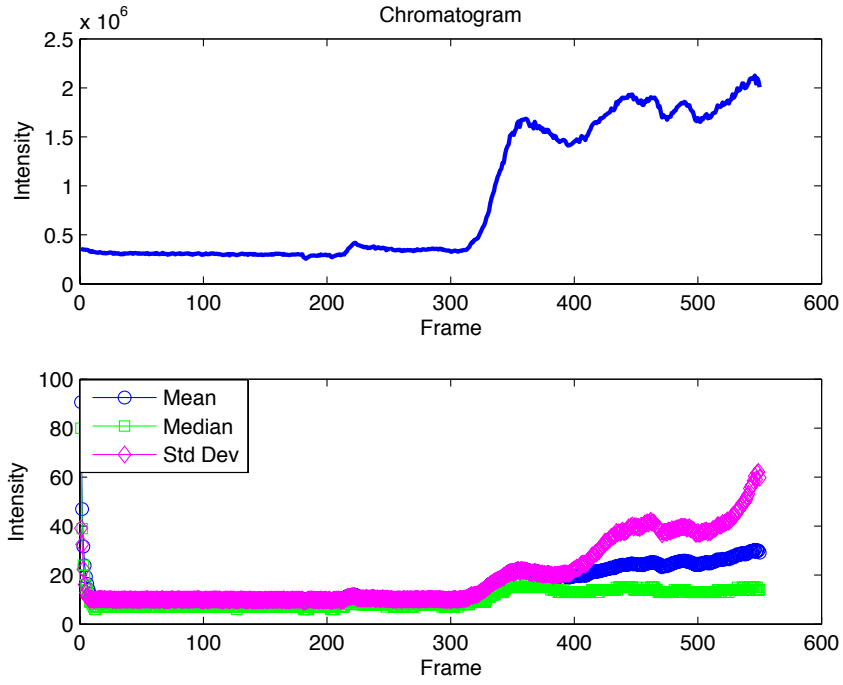


Figure 3.8: Chromatogram of solution at $V_{NT} = -0.0023V$

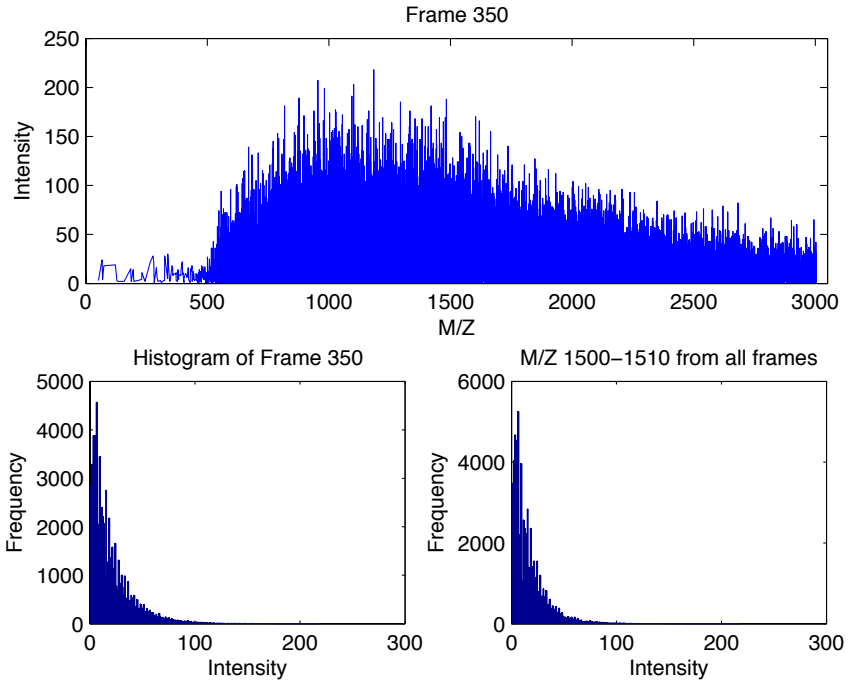


Figure 3.9: Typical frame and histogram at $V_{TN} = -0.0023V$

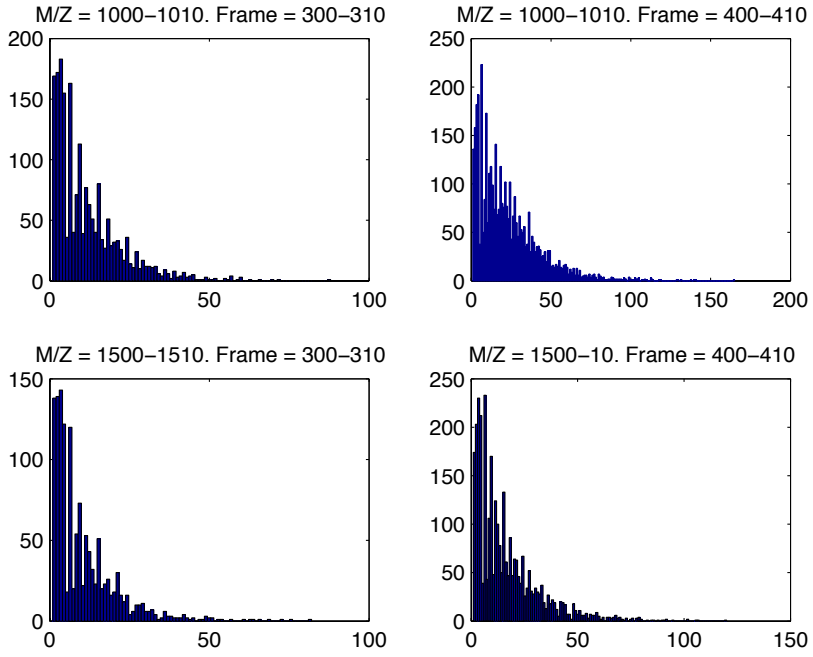


Figure 3.10: Histogram of solution at $V_{NT} = -0.0023V$

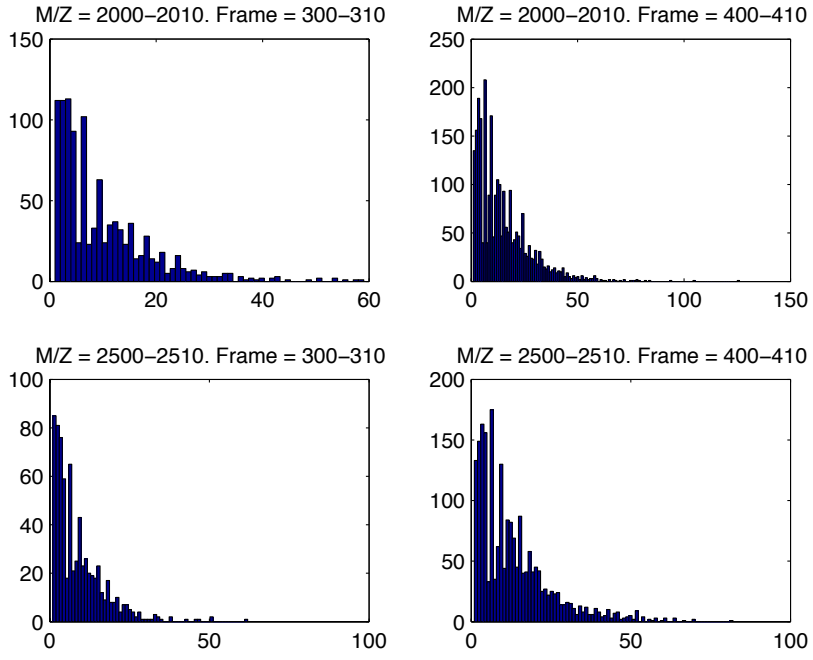


Figure 3.11: Histogram of solution at $V_{TN} = -0.0023V$

the two data sets with n_1 and n_2 observations respectively then the test statistic is defined as:

$$d_{n_1 n_2} = \sup_x |F_{n_1}(x) - F_{n_2}(x)| \tag{3.31}$$

The null hypothesis of both having the same underlying distribution is rejected when $\sqrt{\frac{n_1 n_2}{n_1 + n_2}} d_{n_1 n_2} > k_\alpha$. If the parameters of $F_0(x)$ are to be estimated from the observed data (n_1) then the two-sample KS test is used. The second set of samples are realized using the estimated parameters for the hypothesized distribution.

The solvent MS data has 550 frames with a m/z range of 500 - 3000 in each frame. Since the solvent concentration changes gradually and since the noise statistics may depend on the m/z value in a frame, the data is binned into an m/z bins of length 10 taken from 10 frames at a time. The m/z range is restricted to 1000 - 2500 and frame numbers 250 - 500 as almost all the sample molecules arrive at the detector in those ranges. So, there are 3,962 (151×26) data sets to be tested for a probability distribution model. By visually inspecting the noise histograms, as in Figs 3.10-3.11, it is very unlikely that the noise has a Gaussian, exponential or Poisson distribution. The null hypothesis was rejected for all the data sets when a two-sample KS test is performed for a log-normal distribution with $p = 0.05$. For almost 10% of data sets (365 out of 3,962) the null hypothesis was not rejected when a similar test is performed for Gamma distribution. The Gamma distribution has a PDF of the form,

$$f_0(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}} \tag{3.32}$$

where θ is the scaling parameter, k is the shape parameter, and $\Gamma(\cdot)$ is the gamma function defined by,

$$\Gamma(k) = \int_0^\infty e^t t^{k-1} dt$$

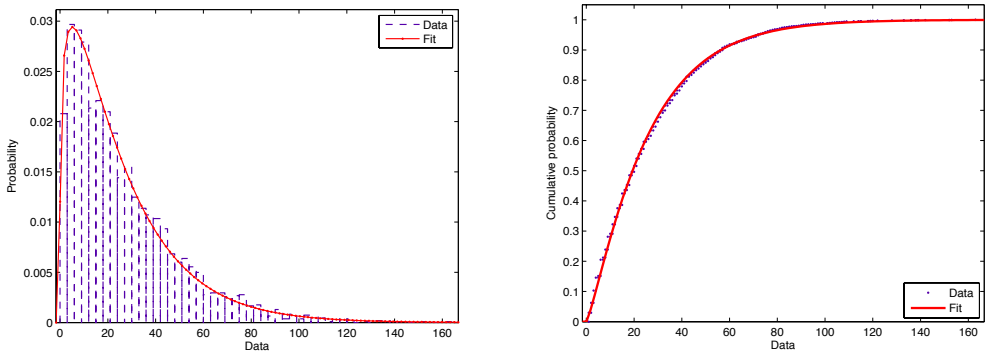
Fig 3.12 shows the PDF and CDF for a dataset (m/z 1500-1510, frame 350-360) where the KS test did not reject the null hypothesis. Fig 3.13 shows the PDF and CDF for a dataset (m/z 1500-1510, frame 300-310) where the KS test rejected the null hypothesis. Comparing the probability plots for the two cases, in Fig 3.14, it appears that the test fails at the tail of the distribution. Similar results are seen for most other data sets where the KS test rejected the null hypothesis.

Another GOF model is the Cramér-von Mises criterion (CM test [112]). The test statistic is defined as:

$$\begin{aligned} \omega^2 &= \int_{-\infty}^{+\infty} [F_n(x) - F_0(x)]^2 \, dF_0(x) \\ &= \int_{-\infty}^{+\infty} [F_n(x) - F_0(x)]^2 \, f_0(x) dx \end{aligned} \tag{3.33}$$

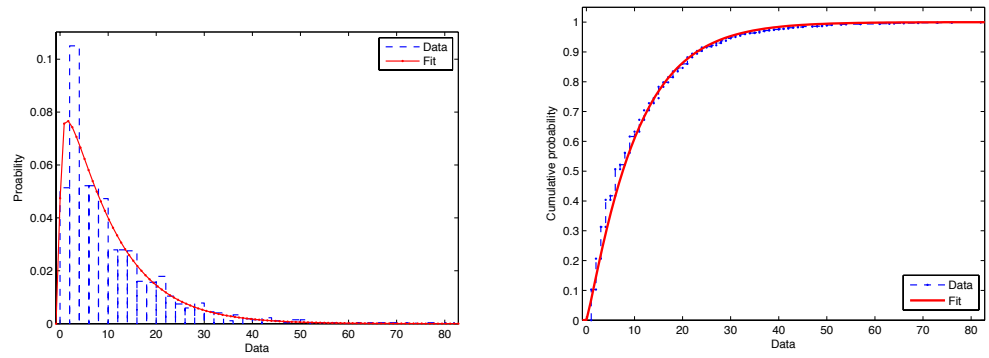
Since $dF_0(x) = f_0(x)\,dx$, the statistic $[F_n(x) - F_0(x)]^2$ is weighted according to $f_0(x)$. Small $f_0(x)$ at the tails will mean a smaller weight for the test statistic. As a result, the tail of the distribution is less emphasized for the GOF. The two-sample CM test is similar to the two-sample KS test where the second data set is generated using estimated parameters.

The two-sample CM test was performed on the same 3,962 noise vectors. Null hypothesis was not rejected for more than half of the vectors (2,121 out of 3,962) using the two-sample CM test for $\alpha = 0.05$. Ten more solvent runs were carried out to get more MS data. Two-sample CM test performed on the 3,962 vectors, with the additional data from the 10 runs, resulted in more than 88% of datasets (3,460 out of 3,962) for which the null hypothesis was not rejected. Based on these results, a Gamma distribution was adopted as a model for chemical noise for ESI-TOF-MS.



(a) PDF comparison (b) CDF comparison

Figure 3.12: H_0 not rejected for dataset m/z 1500-1509, frame 350-359



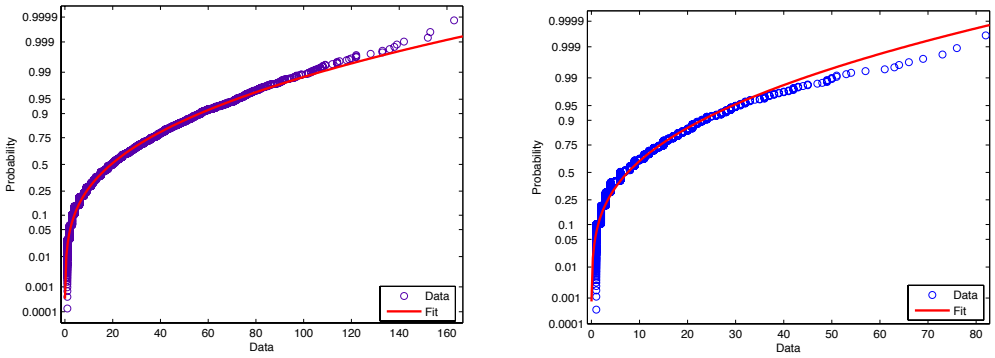
(a) PDF comparison (b) CDF comparison

Figure 3.13: H_0 rejected for dataset m/z 1500-1509, frame 300-309

3.3.3 Gamma Parameter Estimation

In view of the noise model developed in the previous section, signal processing algorithms suited to this model are desired. Here, a maximum likelihood method is used to estimate the parameters of the Gamma distribution from noise data. These parameters are then used for the two-sample GOF tests. From Eqn 3.32, the likelihood function of N independently distributed Gamma variates is:

$$L(k,\theta) = \prod_{n=0}^{N-1} \frac{1}{\Gamma(k)\theta^k} x[n]^{k-1} e^{-\frac{x[n]}{\theta}} \tag{3.34}$$



(a) H_0 not rejected (b) H_0 rejected

Figure 3.14: Probability plots

Hence, the log-likelihood function is:

$$\ln L(k,\theta) = (k-1)\sum_{n=0}^{N-1} \ln x[n] - \frac{x[n]}{\theta} - \ln\Gamma(k) - k\ln\theta \tag{3.35}$$

Taking the respective derivatives of the log-likelihood function with respect to the parameters θ and k yields:

$$\frac{\partial \ln L}{\partial \theta} = u(k,\theta) = \theta k N - \sum_{n=0}^{N-1} x[n] \tag{3.36}$$

$$\frac{\partial \ln L}{\partial k} = v(k,\theta) = \sum_{n=0}^{N-1} \ln x[n] - \ln\theta - \Psi(k) \tag{3.37}$$

where $\Psi(k) = \dfrac{\partial \ln\Gamma(k)}{\partial k}$. The ML estimates of the parameters are the solutions to the following simultaneous equations:

$$u(\hat{k},\hat{\theta}) = 0$$

$$v(\hat{k},\hat{\theta}) = 0$$

There is no closed form solution for the parameters and a numerical solution is provided in [113]. The estimated parameters from the MS noise samples are used for the GOF test and also for detection schemes developed later in the chapter.

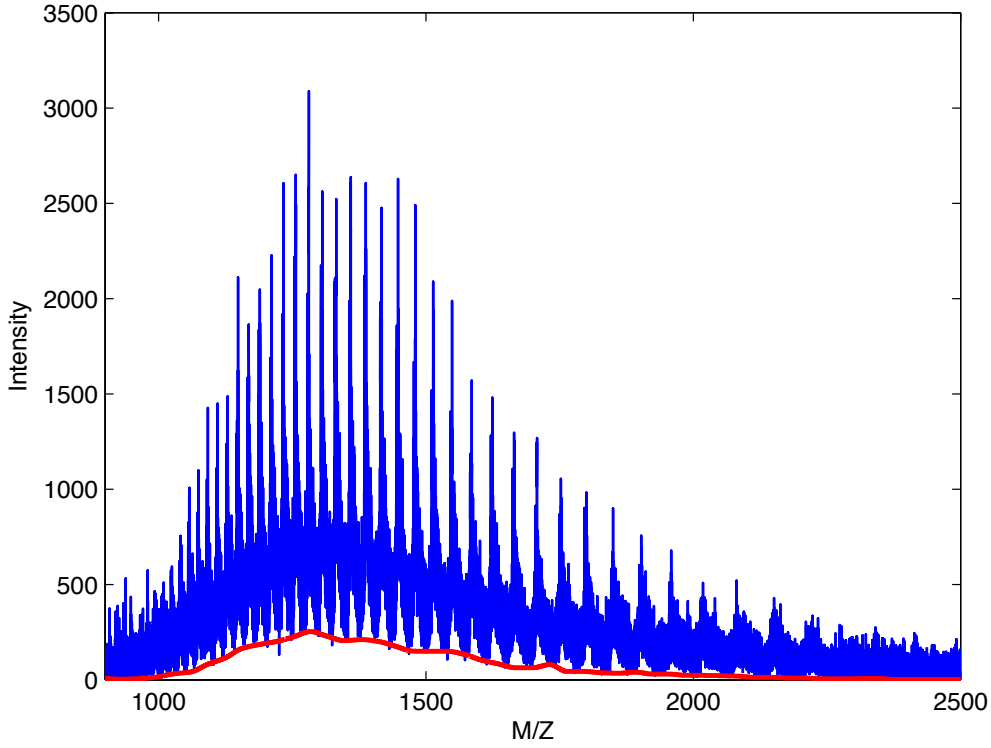


Figure 3.15: MS signal with a baseline

3.4 Baseline in Presence of Signal

Fig 3.15 shows an MS frame for HSA, which is a sum of thousands of spectra over a given period of time. The spectrum consists of signal peaks, due to the analyte, plus additive noise. However, there is a baseline component which is not flat, i.e., the spectrum seems to be sitting on top of a time varying baseline. The baseline is generally attributed to chemical noise and detector saturation leading to a slowly decaying charge [114]. A shifting window algorithm, similar to one described in [85], can be used to estimate the baseline. The mass spectrum is divided into overlapping rectangular windows and the minimum or an n^{th} quantile value of the intensity is determined for each window. Then a moving average filter is used to smooth the

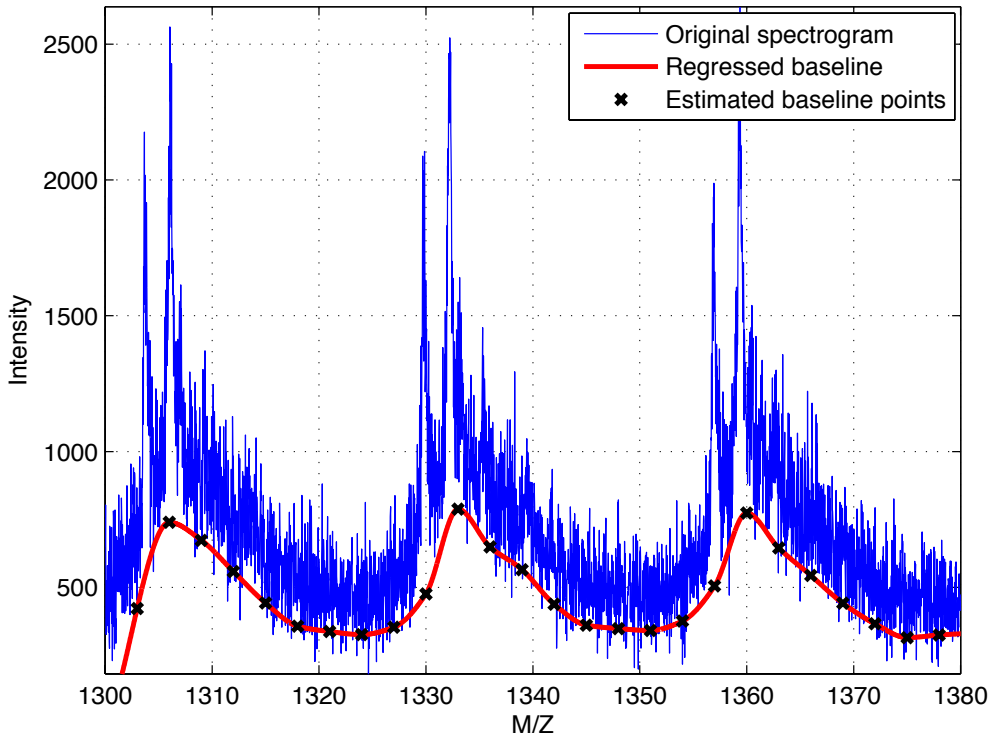


Figure 3.16: Baseline estimation

baseline points thus reducing or eliminating the effect of outliers. Then, an interpolation of the m/z values on to the smoothed baseline points is performed using a cubic spline algorithm [115]. The result is shown in Fig 3.16. Traditionally, this baseline is considered to be a distortion and is subtracted from the MS signal before further analysis is carried out.

3.4.1 Signal-Baseline Model

The reason for the formation of a typical baseline is not fully understood and is eliminated for MS analysis. Since a non-flat baseline is prominent when an analyte is present, it is safe to assume that the analyte molecules somehow contribute to this phenomenon. Looking at Fig 3.16, it can be argued that the decay of the sig-

nal around the peak is due to solvated and fragmented analyte molecules. In other words, the intact analyte molecules form the signal peak whereas the fragmented and solvated molecules are on the left and right respectively. Solvated molecules are analytes attached with some solvent molecules thus gaining mass and falling to the right of the main peak. As the number of attached solvent molecules increases, the mass of the solvated analyte increases. It is reasonable to assume that the probability of solvent molecules attaching to an analyte decreases as the number increases thus forming the decaying tail. Fragmentation on the other hand causes the loss of certain part of the analyte, thus decreasing its mass and falling to the left of the main peak. The faster decay on the left is possibly due to the fact that the analyte molecules are less prone to fragmentation in ESI; whereas solvation is much more likely resulting in a slower decay rate on the right of signal peak. It is to be noted that the area under both the tails is much larger than the area under the main peak. This means that the majority of the ions have some solvent molecules attached to them and thus falling on the right side decay region and only a small fraction of the ions form the main peak. The baseline is the result of the sum of all the decaying signals from all the neighboring charge states. This is illustrated in Fig 3.17.

The signal peak is shaped as a Gaussian pulse, as described in section 3.2. The lower part of the signal peak seems to have a Cauchy rate of decay, albeit with different rates for the two sides. The Cauchy decay model is found to be a better fit compared to an exponential decay model, though more experiments and model fitness tests have to be performed to substantiate the assumption. Assuming a Cauchy decay model for the tails, the signal model at an arbitrary position y_0 can

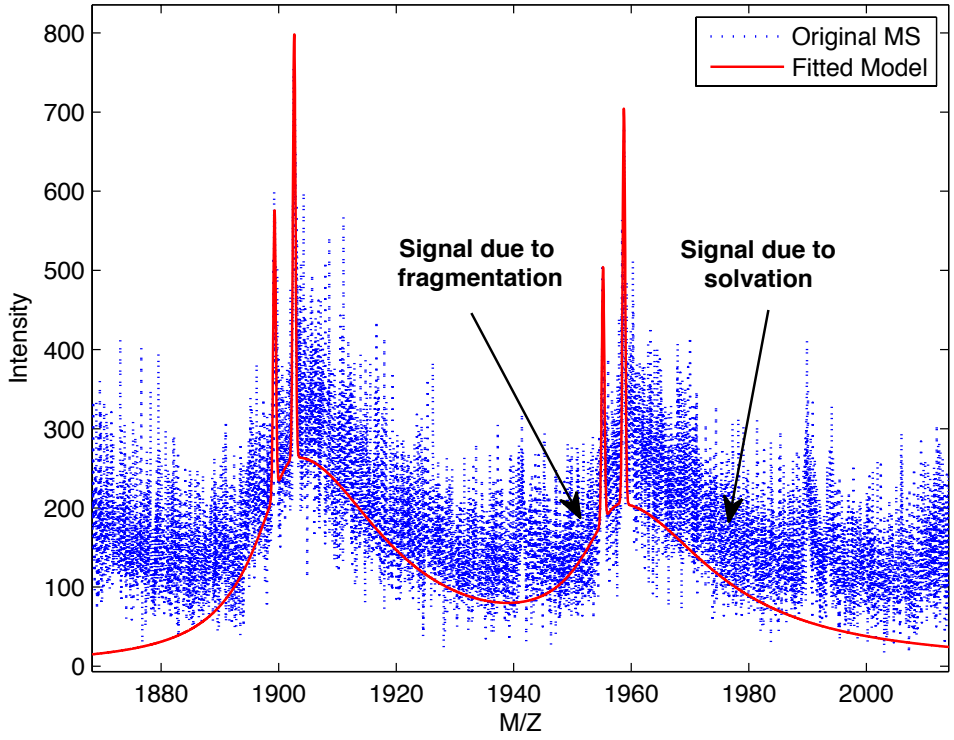


Figure 3.17: Baseline as a part of signal model

be described as a mixture model:

$$g(y) = \begin{cases} \dfrac{a_1}{1+\lambda_1(y-y_0)^2} + a_2\, e^{-\dfrac{(y-y_0)^2}{2\lambda_2^2}} & \text{if } y < y_0 \\ \dfrac{a_3}{1+\lambda_3(y-y_0)^2} + a_2\, e^{-\dfrac{(y-y_0)^2}{2\lambda_2^2}} & \text{if } y \geq y_0 \end{cases} \tag{3.38}$$

where a_i's are normalizing constants and λ_i's are the decay rates, for $i = 1,2,3$. A histogram of the samples drawn from $g(y)$ using Monte Carlo method is shown in Fig 3.18. The probability of a molecule ending up in a bin of size ε on the time spectrum is:

$$p_\varepsilon = \int_t^{t+\varepsilon} g(y)dy \tag{3.39}$$

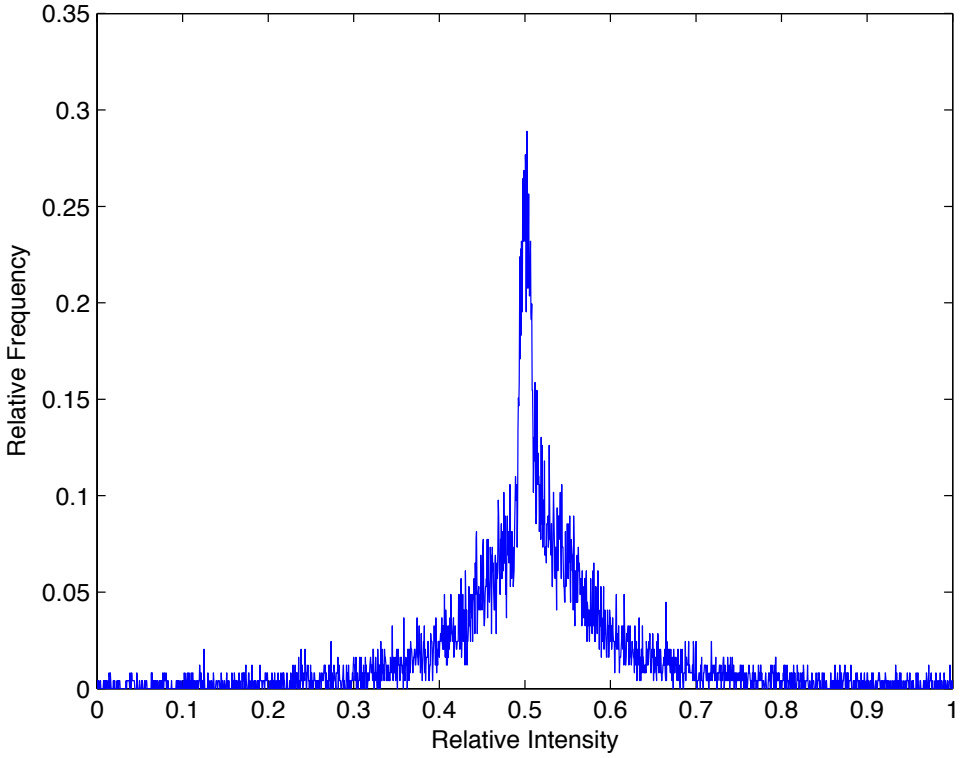


Figure 3.18: Histogram of samples drawn from $g(y)$

If there are a total of N_a analyte molecules in the sample that reach the detector, the probability of n_a of them reaching the detector in the time bin $[t, t+\varepsilon]$ is:

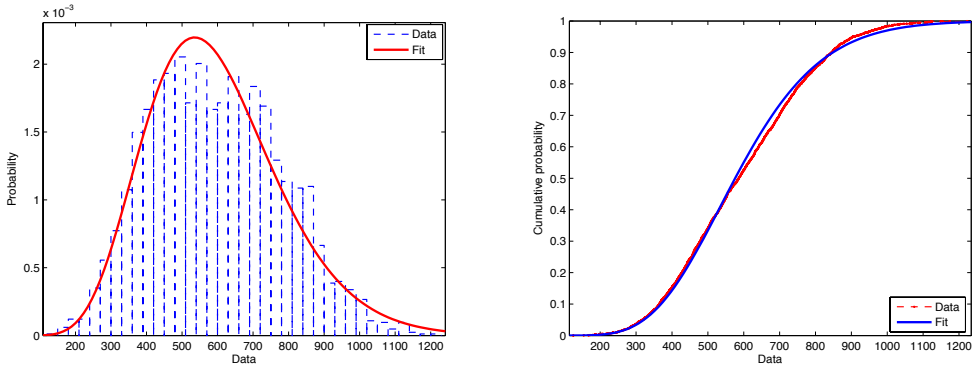
$$\begin{aligned} p_{n_a,\varepsilon} &= \binom{N_a}{n_a} p_\varepsilon^{n_a} (1-p_\varepsilon)^{N_a-n_a} \\ &= p_\varepsilon^{n_a} (1-p_\varepsilon)^{N_a-n_a} \frac{N_a!}{n_a!(N_a-n_a)!} \end{aligned}$$

This idea of considering the baseline as part of the signal and thus using it as a signal model is in its preliminary stage and goes against the traditional approach. Further analysis is needed before including it in the signal detection scheme presented in the next section. It is left to be explored in the future to see if any improvements in MS signal analysis can be achieved by doing so.

3.4.2 Noise-Baseline Model

All traditional MS analysis methods consider the baseline to be a distortion and devise of methods to eliminate it without causing any loss of signal information. Following this traditional assumption, the baseline is considered to be a part of the noise. The noise statistic, that includes the baseline, is evaluated for the MS data set of HSA. Acknowledging the time varying property of the baseline, a bin size of length 1 in m/z is considered. A total of 150 bins of size 1 in m/z from the range 1200 to 1600 (avoiding the peak signal part) were chosen for frames 340 to 400. The histograms and the GOF tests results show that the noise with baseline can still be adequately modeled with a Gamma distribution. The null hypothesis was rejected for 41 of the 150 datasets when a two-sample KS test was performed and it was rejected for 30 of the 150 datasets for a two-sample CM test. Both tests were at a significance level $\alpha = 0.05$. Analysis of more data from 5 sample runs of HSA show that the GOF test results are consistent for Gamma distribution. Fig 3.19 shows the empirical PDF and CDF comparisons for one of the datasets for which the null hypothesis was not rejected. Comparing the histograms in Figs 3.12 and 3.19 it is evident that the mean and variance of this noise with baseline is much larger compared to the mean and variance of only chemical noise.

The signal detection schemes being developed in this chapter use the traditional approach of considering the baseline as part of noise and include it in the noise model. However, from initial analysis and simulations, the signal-baseline model seems to be a better choice and can be considered for future signal analysis.



(a) PDF comparison (b) Probability Plot

Figure 3.19: H_0 not rejected

3.5 MS Signal Detection Using Likelihood Ratio Test

Consider the problem of a known signal embedded in an additive noise. The null and alternate hypotheses for detection problem are the following:

$$\begin{aligned} H_0 &: \underset{\sim}{\mathbf{x}} = \underset{\sim}{\mathbf{w}} \\ H_1 &: \underset{\sim}{\mathbf{x}} = A\mathbf{s} + \underset{\sim}{\mathbf{w}} \end{aligned} \tag{3.40}$$

where $\mathbf{s} = [s_0, \cdots, s_{N-1}]^T \in \mathbb{R}^N$ is a known deterministic signal vector with an unknown amplitude $A > 0$ and $\underset{\sim}{\mathbf{w}} = \left[\underset{\sim}{w}_0, \cdots, \underset{\sim}{w}_{N-1}\right]^T$ is a random N-vector whose elements $\underset{\sim}{w}_n$ are independent random variables with known distribution $f_{\underset{\sim}{w}_n}(w_n)$. The independent assumption implies:

$$f_{\underset{\sim}{\mathbf{w}}}(\mathbf{w}) = \prod_{n=0}^{N-1} f_{\underset{\sim}{w}_n}(w_n) \quad w_n \geq 0 \text{ and } n = 0, \cdots, N-1$$

The PDFs under the null and alternate hypotheses can be written as:

$$f_{\underset{\sim}{\mathbf{x}}}(\mathbf{x}; H_0) = \prod_{n=0}^{N-1} f_{\underset{\sim}{x}_n}(x_n) \tag{3.41}$$

$$f_{\underset{\sim}{\mathbf{x}}}(\mathbf{x}; H_1) = \prod_{n=0}^{N-1} f_{\underset{\sim}{x}_n}(x_n - As_n) \tag{3.42}$$

Thus, the likelihood ratio is:

$$L(\mathbf{x}) = \frac{\prod_{n=0}^{N-1} f_{\underset{\sim}{x}_n}(x_n; H_1)}{\prod_{n=0}^{N-1} f_{\underset{\sim}{x}_n}(x_n; H_0)} = \frac{\prod_{n=0}^{N-1} f_{\underset{\sim}{x}_n}(x_n - As_n)}{\prod_{n=0}^{N-1} f_{\underset{\sim}{x}_n}(x_n)} \tag{3.43}$$

The Neyman-Pearson (NP) detector, which maximizes probability of detection P_d for any chosen value of false alarm probability P_f, decides in favor of H_1 if $L(\mathbf{x})$ exceeds a threshold γ, i.e.,

$$\rightarrow H_1 = \left\{\mathbf{x} \middle| L(\mathbf{x}) > \gamma\right\}$$

The threshold γ is typically chosen to correspond to a desired value of P_f according to:

$$P_f = \int_{\rightarrow H_1} f_{\underset{\sim}{\mathbf{x}}}(\mathbf{x}; H_0) d\mathbf{x}$$

Often times the log-likelihood ratio is convenient for the threshold test and is defined as:

$$\ln L(\mathbf{x}) = l(\mathbf{x}) = \sum_{n=0}^{N-1} \ln \frac{f_{\underset{\sim}{x}_n}(x_n - As_n)}{f_{\underset{\sim}{x}_n}(x_n)} \tag{3.44}$$

3.5.1 Detection Equation in Gamma Distributed Noise

In this case, $\underset{\sim}{w}_n$ are independent Gamma distributed random variables, i.e.,

$$f_{\underset{\sim}{w}_n}(w_n) = \frac{1}{\Gamma(k_n)\theta_n^{k_n}} x_n^{k_n - 1} e^{-\frac{w_n}{\theta_n}} \quad \text{with } w_n > 0 \text{ and } n = 0, \cdots, N-1 \tag{3.45}$$

The PDF is parameterized by a shape parameter $k_n > 0$ and a scale parameter $\theta_n > 0$. So, as per Eqns 3.41, 3.42 and 3.43:

$$f_{\underset{\sim}{\mathbf{x}}}(\mathbf{x}; H_0) = \prod_{n=0}^{N-1} \frac{1}{\Gamma(k_n)\theta_n^{k_n}} x_n^{k_n - 1} e^{-\frac{x_n}{\theta_n}} \qquad x_n \geq 0$$

$$f_{\underset{\sim}{\mathbf{x}}}(\mathbf{x}; H_1) = \prod_{n=0}^{N-1} \frac{1}{\Gamma(k_n)\theta_n^{k_n}} (x_n - As_n)^{k_n - 1} e^{-\frac{(x_n - As_n)}{\theta_n}} \qquad x_n \geq As_n$$

$$L(\mathbf{x}) = \frac{f_{\underset{\sim}{\mathbf{x}}}(\mathbf{x}; H_1)}{f_{\underset{\sim}{\mathbf{x}}}(\mathbf{x}; H_0)} = \prod_{n=0}^{N-1} \left(\frac{x_n - As_n}{x_n}\right)^{k_n - 1} e^{\frac{As_n}{\theta_n}} \tag{3.46}$$

It is to be noted that, $L(\mathbf{x})$ is zero if $x_n < As_n$ and is infinite if $x_n < 0$. The NP detector decides in favor of the alternate hypothesis (H_1) when $L(\mathbf{x}) > \gamma$.

$$\rightarrow H_1 = \left\{ \prod_{n=0}^{N-1} \left(\frac{x_n - As_n}{x_n} \right)^{k_n - 1} e^{\frac{As_n}{\theta_n}} > \gamma \right\} \tag{3.47}$$

In the NP detector, A is deterministic but unknown. Taking the log of the likelihood ratio in Eqn 3.46:

$$l_A(\mathbf{x}) = \sum_{n=0}^{N-1} (k_n - 1) \ln \left(\frac{x_n - As_n}{x_n} \right) + A \frac{s_n}{\theta_n} \tag{3.48}$$

Assuming a small A, the Taylor series approximation of $l_A(\mathbf{x})$ can be written as follows:

$$\begin{aligned} l_A(\mathbf{x}) &\approx l_A(\mathbf{x})\Big|_{A=0} + \frac{\partial l_A(\mathbf{x})}{\partial A}\Big|_{A=0} A \\ &= 0 + \sum_{n=0}^{N-1} \left[\frac{-(k_n - 1)s_n}{x_n - As_n} + \frac{s_n}{\theta_n} \right] \Big|_{A=0} A \\ &= A \sum_{n=0}^{N-1} \left[\frac{-(k_n - 1)s_n}{x_n} + \frac{s_n}{\theta_n} \right] \end{aligned} \tag{3.49}$$

The NP detector, in terms of the log-likelihood function, is:

$$\rightarrow H_1 = \left\{ A \sum_{n=0}^{N-1} \left[\frac{-(k_n - 1)s_n}{x_n} + \frac{s_n}{\theta_n} \right] > \gamma' \right\}$$

Since $A, \theta_n > 0$ and s_n is known, the 2nd term in the sum can be absorbed into the threshold as follows:

$$\begin{aligned} \rightarrow H_1 &= \left\{ -\sum_{n=0}^{N-1} (k_n - 1) \frac{s_n}{x_n} > \gamma'' \right\} \\ \rightarrow H_1 &= \left\{ T(\mathbf{x}) > \gamma'' \right\} \end{aligned} \tag{3.50}$$

where, $\gamma'' = \dfrac{\gamma'}{A} - \displaystyle\sum_{n=0}^{N-1} \frac{s_n}{\theta_n}$ and $T(\mathbf{x})$ is the sufficient statistic for the test. Note that the unknown amplitude A and the scale parameter θ_n do not appear in $T(\mathbf{x})$, though

they will matter in determining the value of γ'' corresponding to a desired P_f. Also, the detector is invariant to unknown channel gains, i.e., multiplying both signal and noise by a positive constant does not alter the value of $T(\mathbf{x})$. In the event that the noise components are identically distributed, $k_n = k$ for all n, the detection statistic simplifies to:

$$T_i(\mathbf{x}) = -\sum_{n=0}^{N-1} \frac{s_n}{x_n} \tag{3.51}$$

The decision rules according to Eqns 3.50 and 3.51 are approximately optimal when the signal amplitude is unknown, provided it is small. For signals of large amplitude (i.e., high SNR regime), optimality of the detector is generally less important because such signals are less difficult to detect even with suboptimal detectors.

Alternately, an estimate of A can be used in Eqn 3.47 and a general likelihood ratio test (GLRT) can be performed. Considering $l_A(\mathbf{x})$ for the GLRT:

$$\begin{aligned} \rightarrow H_1 &= \left\{ \sum_{n=0}^{N-1} (k_n - 1) \ln\left(\frac{x_n - \hat{A}s_n}{x_n}\right) + \hat{A}\frac{s_n}{\theta_n} > \ln(\gamma) \right\} \\ &= \left\{ \sum_{n=0}^{N-1} (k_n - 1) \ln\left(\frac{x_n - \hat{A}s_n}{x_n}\right) > \ln(\gamma) - \hat{A}\sum_{n=0}^{N-1}\frac{s_n}{\theta_n} \right\} \\ &= \left\{ T_{GLRT}(\mathbf{x}) > \gamma''' \right\} \end{aligned} \tag{3.52}$$

The sufficient statistic for the GLRT, considering $k_n = k$, is given by:

$$T_{GLRT_i}(\mathbf{x}) = \sum_{n=0}^{N-1} \ln\left(\frac{x_n - \hat{A}s_n}{x_n}\right) \tag{3.53}$$

3.5.2 Maximum Likelihood Estimation of Amplitude

The maximum likelihood (ML) estimator is one of the most popular approaches of obtaining practical estimates of unknown parameters. The MLE is not optimal in general but in special cases it is optimal for large enough data records. The MLE of a parameter is defined as the value of the parameter that maximizes likelihood function for a fixed observation [116]. This maximization is performed over the

range of the parameter by differentiating the likelihood function. In this sense MLE is an extremum estimator. It is also common to resort to numerical techniques of maximization for finding the MLE.

In this Gamma noise case, the estimate of A in obtained by differentiating $l_A(\mathbf{x})$. From Eqn 3.48:

$$\frac{\partial l_A(\mathbf{x})}{\partial A} = \sum_{n=0}^{N-1} \left[\frac{-(k_n - 1)s_n}{x_n - As_n} + \frac{s_n}{\theta_n} \right] \tag{3.54}$$

Setting the partial derivative of the log-likelihood function to zero yields:

$$\begin{aligned} &\sum_{n=0}^{N-1} \left[\frac{-(k_n - 1)s_n}{x_n - \hat{A}s_n} + \frac{s_n}{\theta_n} \right] = 0 \\ \implies &\sum_{n=0}^{N-1} \frac{(k_n - 1)s_n}{x_n - \hat{A}s_n} - N\frac{\overline{s}}{\theta} = 0 \end{aligned} \tag{3.55}$$

where $\dfrac{\overline{s}}{\theta} = \dfrac{1}{N}\displaystyle\sum_{n=0}^{N-1} \frac{s_n}{\theta_n}$. There is no closed form solution for $\hat{A}$. The first term in Eqn 3.56 is a sum of monotonic functions of $\hat{A}$ and hence the equation has an unique solution. The solution $\hat{A}$ is such that $0 \leq \hat{A}s_n < x_n$ for all n. The solution is obtained by minimizing the square of Eqn 3.56 for the above bound using a golden section search [117] method.

$$\hat{A} : \min_{\hat{A}} \left(\sum_{n=0}^{N-1} \frac{(k_n - 1)s_n}{x_n - As_n} - N\frac{\overline{s}}{\theta} \right)^2 \qquad 0 \leq \hat{A} < \frac{x_n}{s_n} \text{ for } n = 0, \cdots, N-1 \tag{3.56}$$

The performance of the detectors and the estimator, in the form of Monte Carlo simulations, is evaluated in the next chapter.

3.6 Conclusion

Traditionally, MS data is analyzed as the combination of three components: signal, noise and baseline. In this chapter, each of these three components have been studied and a model is proposed. The signal peak shape and width is a function of the isotopic distribution of the molecule under investigation, combined with the

spatial and energy distribution and the physical limitations of the detector. Assuming a Gaussian spread as the ions spatial distribution, the signal shape turns out to be Gaussian as well. The width of this peak is a result of the isotopic distribution coalescing together, as limited resolution of the spectrometer can not resolve the individual isotopic peaks. In essence, the signal peak shape and width is estimated from first principles by mathematically modeling the isotopic, spatial and energy distributions. Such a method is mathematically tractable and provides a sound basis for any assumptions made regarding the shape and width of the signal peak.

A new chemical noise model was developed by investigating the experimental data from the mass spectrometer. This method is unlike most of the existing noise model assumptions that are not validated with experimental data. Dark current noise from the MCP is shown to be eliminated by the thresholding algorithm of the instrument and generally does not corrupt the MS signal at operating threshold levels. Chemical noise in MS arises due to the solvent molecules and clusters. The noise samples are observed by running only solvent through the instrument at operating threshold levels. Histograms and GOF tests are performed on this data. It is concluded that the chemical noise is adequately modeled by a Gamma distribution.

The cause of a baseline in the MS is not understood completely and it has traditionally been considered to be a distortion caused by noise saturation. It is shown that the noise along with the baseline distortions is also modeled by a Gamma distribution. A new model, with the baseline as part of the signal, is proposed. However, this is left for future research as further studies need to be carried out to investigate the assumptions and efficacy of such a signal model.

Finally, detection schemes, using the signal and additive noise models, were developed using the signal and noise models. An approximation to the optimal NP detector is developed for small signal amplitudes. A GLRT based detector is also

proposed. MLE method is used to estimate unknown amplitude in the model, to be used in the GLRT. The performance of this detection and estimation method, and an application to biomarker analysis is explored in the next chapter.

CHAPTER 4

DECONVOLUTION AND ABUNDANCE CALCULATION

4.1 INTRODUCTION

ESI MS TOF is less prone to fragmentation of macro molecules, such as proteins, and thus is useful for biomarker analysis. ESI almost always produces multiple charged ions resulting in a low mass-to-charge (m/z) ratio and thus a higher resolving power. However, it is not easy to interpret because the charge states are not precisely controlled which results in the same molecules transiting the mass spectrometer with a spectrum of flight times. This multiplexed information must be detangled to obtain species abundance in a sample and this detangling is referred to as "deconvolution" in much of the chemical literature. "Convoluted" can mean "complicated" to some and this might be the origin of using deconvolution to classify an algorithm that makes the mass spectrum less complicated [81], even though there is no time invariant process involved. It should be noted that the use of the term "deconvolution" is different than what is common in signal processing literature.

As seen in Fig 4.1, the ESI MS can have multiple charge states for the same protein molecules in a sample. These multiple charge states complicate the mass spectral interpretation as the charge state corresponding to each peak must be assigned to determine the mass of the species. Also, larger molecules tend to have more charges and peaks compared to smaller molecules.

In ESI, ions are formed by the addition of protons (with mass m_{H^+}) when operated in the positive ion mode. For an ion with a charge z, the ion peaks occur at a mass-to-charge ratio $(m + zm_{H^+})/z$ in the MS, where m is the molecular mass of the species. For easy interpretation, a routine is required to generate a zero-charge mass spectrum by transforming the peaks on a mass-to-charge ratio scale to a

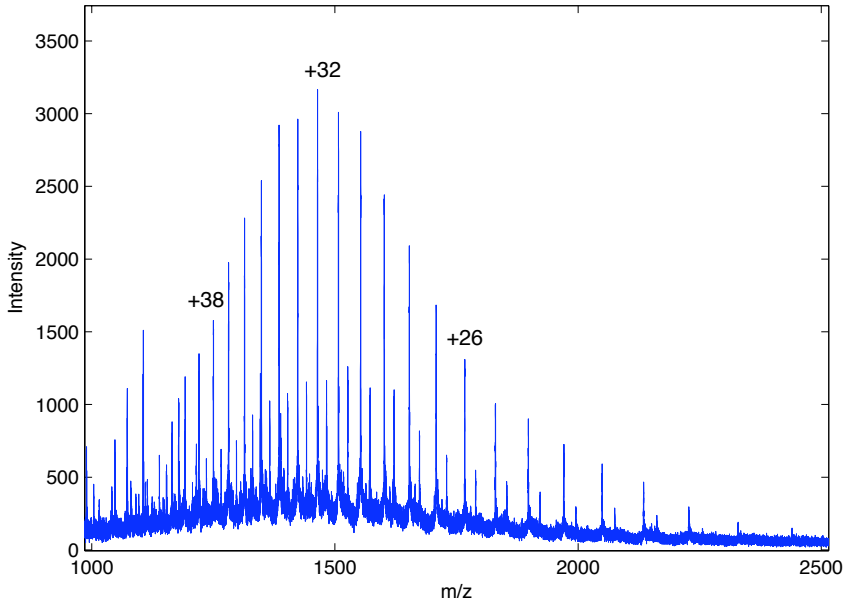
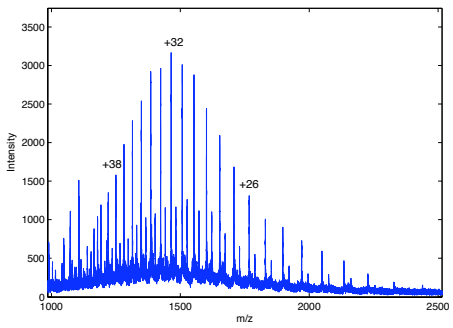
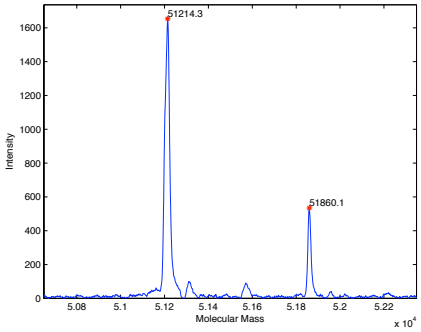


Figure 4.1: ESI MS for human Vitamin D Binding Protein (DBP)

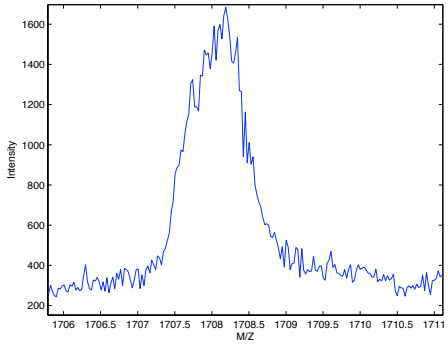
single peak on a molecular mass scale. This transformation is called deconvolution. When there are multiple species with multiple charge states, it becomes difficult to assign a particular peak to a species as required for deconvolution. As seen in Fig 4.2(a), there are two peak series corresponding to the protein (DBP) and its glycated variant. After deconvolution, shown in Fig 4.2(b), it is easy to interpret the MS by looking at the molecular masses of the two species. In addition to the multiple peaks due to charge states, there are also peaks due to the presence of isotopes. For low mass species, individual isotopic peaks might be resolved in the MS. For larger molecules however, the isotopes result in a peak spread. Fig 4.2(c) & (d) shows the peak spread for DBP and ApoC1 proteins. The isotopic peaks of the low molecular mass ApoC1 are somewhat resolved where as the isotopic peaks of DBP coalesce together forming a wider signal. It is to be noted that the ESI MS, Fig 4.1, is obtained by averaging multiple frames of MS data where most of the



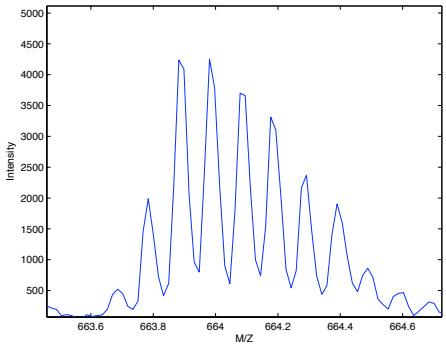
(a) Average Mass Spectrum for DBP



(b) Deconvoluted ESI for DBP



(c) ESI MS for DBP (For a single charge state)



(d) ESI MS for ApoC1 (For a single charge state)

Figure 4.2: Deconvolution and peak spread in ESI MS

peak signal is accumulated. These frames are usually chosen by visually inspecting the chromatogram plot. Chromatogram is the sum of all the masses hitting the ion detector over time, as shown in Fig 4.3. The x-axis is time, which corresponds to frame numbers, and y-axis is the total intensity. The average ESI MS is generally obtained over the time range in which most of the sample data reaches the detector, e.g. between 300 to 500 s. This is usually the case for most proteins, as the sample molecules take about 5 minutes to reach the TOF stage, also called the elution time.

There are various deconvolution algorithms [118–120] proposed in the literature. The method in [118] tends to produce artifacts and baseline distortions. A

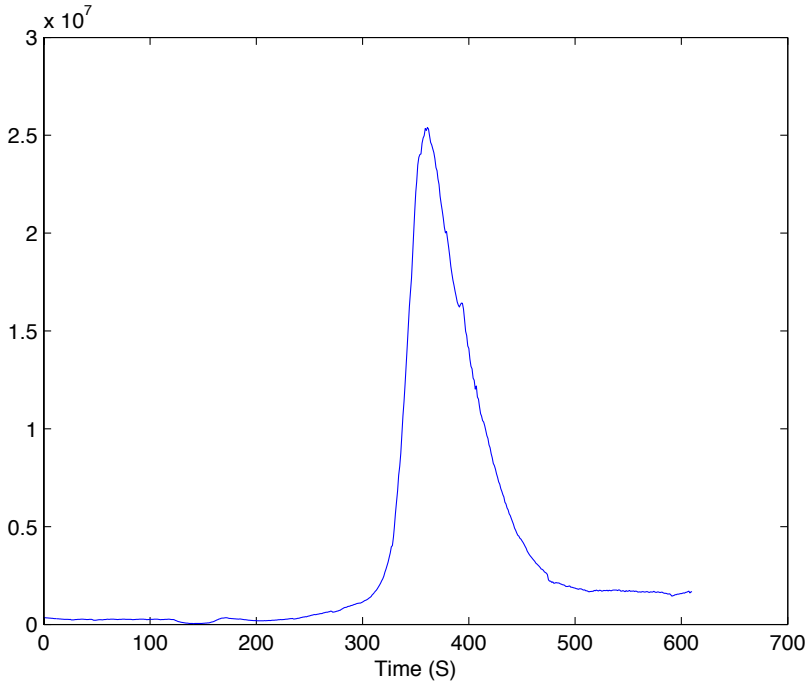


Figure 4.3: Chomatogram of HSA for one sample run

maximum-entropy based method, proposed in [119], tends to be time consuming. A fully automated score based deconvolution algorithm is proposed in [120]. There are several commercial and public software packages [121–125] available for the pre-processing and analysis of TOF-MS data.

For identifying biomarkers, it is important to estimate the relative abundance of all the protein variants present in a sample, once spectrum deconvolution is achieved. The signal peak intensity is a measure of the number of molecules hitting the detector. However, due to limited equipment resolution and isotopic distribution issues the signal peaks are wider. The area under a peak is considered to be a fair measure of the abundance of a molecule in the sample being tested. The abundance of a species is proportional to the sum of areas of each charge state belonging to the species or the area under the single peak after deconvolution. Commercially

available software (Bruker [126]) allow fractional abundance calculation after deconvolution, but require manual identification of boundaries of all the peaks. This process may become time consuming when analyzing a large number of samples, as is required in the biomarker identification application. The accuracy of such area estimation is questionable when there are overlapping peaks from closely related molecular species or because of isotopic spread.

In section 4.2, one of the current methods based on Bruker's software is explained in detail, along with the addition of an automated abundance calculation routine. Some of the results and limitations of this type of algorithm are discussed. In section 4.3, a new method of fractional abundance calculation, based on the detection and estimation schemes developed in Chapter 3, is introduced.

4.2 Current Method

Bruker's MS analysis software [126], which is provided with the instrument, has a few disadvantages as far as the area calculation is concerned. The user has to select the boundaries for each peak of interest in a given deconvoluted MS. In this section, an algorithm that works on the same principle as the software is explained with the addition of an automatic fractional abundance calculation routine.

4.2.1 Input Parameters

It is assumed that the user has some knowledge of the molecular species (protein) under investigation. Accordingly, the algorithm requires the following parameters (similar to that of [126]).

1. m/z Range $[(M/Z)_{min}, (M/Z)_{max}]$ - The raw MS may contain impurities or unwanted species. This parameter gives the range of interest for deconvolution.

2. Range of Molecular mass $[M_{min}, M_{max}]$ - This specifies the final deconvolution range of the MS. It is assumed that all the species of interest fall in this range.

3. Range of Charge States $[Z_{min}, Z_{max}]$ - In general molecules in ESI MS have multiple charge states. An approximate range for the charge states is required for deconvolution.

4. Number of Peaks $[N_{min}, N_{max}]$ - The algorithm searches for all the charge states of a particular protein species in the MS. As each charge states is a peak in the MS, the number of peaks range specifies the number of charge states to look for.

5. Peak Width (for peak separation, W_{sep}) - This is the minimum separation (in m/z) between two consecutive peaks in the raw spectrum. The user should be able to get the approximate peak width from the raw MS. This parameter should stay the same for all sample runs of the same molecule.

6. Peak Width (for area calculation, W_{area}) - This width parameter is needed to calculate the peak boundaries in the deconvoluted MS, for area calculation. W_{area} can be the same as W_{sep}.

7. Detection Threshold (γ_{det}) - As the name suggests, this parameter is used as a threshold to detect a peak in the raw MS. It's a number between $(0, 1)$, and is a fraction of the largest peak. The peaks are compared to this threshold before they can be considered for deconvolution.

8. Acceptance Threshold (γ_{acc}) - When looking for different charge states of a particular protein species, the theoretical peak location may differ from the

actual peak location in the MS. The acceptance threshold parameter defines the allowed separation (jitter) between the theoretical and the actual peak locations in the MS.

4.2.2 Pre-Processing

The raw MS tend to be noisy with baseline distortions. Preprocessing helps in removing or toning down these artifacts before the deconvolution algorithm can be applied. The MS is usually non-homogenous in m/z i.e. the sampling is non-uniform. This is usually due to the quadratic transform from the time domain to the m/z domain. Also, zero valued samples are not stored in the spectrometer, to save on memory usage, thus creating gaps. A resampling may be done to make it homogenous and have a manageable set of data points without losing any information from the spectrum. This step is not necessary and not used here. The pre-processing is applied only to the desired m/z range, which is an input parameter.

Baseline distortion is a common occurrence in mass spectrometry. As discussed in section 3.4, there are various algorithm available for baseline correction. A windowing based approach is used to estimate the time varying baseline and subtracted from the MS signal. The window length is chosen according to the peak width parameter.

There are methods to denoise the MS [127], usually by a moving average filter, without visibly distorting the peak shape or sharpness. A Savitzky-Golay smoothing filter [128] is used here for smoothing. The filter window length is fixed. This step is used for peak detection only. The non-smoothed version of the data is used for further analysis after peak detection.

4.2.3 Deconvolution

The deconvolution process starts by locating the highest peak (P_1) in the spectrum. Let the location of the peak be $m/z = M_{z_1}$. The first task is to estimate the charge state of this peak and the peak location for all other possible charge states. The algorithm is explained as follows:

1. The charge state range $[Z_{1_{min}}, Z_{1_{max}}]$ for P_{k1} is estimated using the input parameter for molecular mass range $[M_{min}, M_{max}]$ and M_{z_1}.

2. For every $z_{1_i} \in [Z_{1_{min}}, Z_{1_{max}}]$, construct a signal $y_i(\cdot)$:

$$y_i(M_z) = \begin{cases} 1 & \text{for } M_z = \frac{(M_{z_1} \times z_{1_i}) - z_{1_i}}{z_n} \\ 0 & \text{otherwise} \end{cases}$$

   where $z_n \in [Z_{min}, Z_{max}]$, the input parameter.

3. Calculate the cross-correlation coefficients:

$$r_i(l) = \sum_{m=-\infty}^{\infty} x(m) y_i(m+l) \tag{4.1}$$

   where $x(m)$ is the MS signal and the number of lags, l, is restricted by the parameter γ_{acc}. i corresponds to the different charge states.

4. The largest cross-correlation coefficient, $\max\{r_i\}$, gives the charge state (z_1) and hence the molecular mass (M_1) of the species at P_{k1}.

5. The charge states of all other peaks can be easily estimated from M_1. The deconvoluted MS is the average of all charge states taken over the range $[M_{min}, M_{max}]$. This also combines the multiple charge states of other molecular species present in the sample, as long as they are in the range $[M_{min}, M_{max}]$.

The deconvolution algorithm requires the knowledge of peak locations [25,30,129–131] in the processed mass spectrum.

4.2.4 Post Processing

The post-processing of the deconvoluted MS is very similar to pre-processing. Smoothing and baseline correction steps are repeated after deconvolution.

4.2.5 Area Under Deconvoluted Peaks

The goal is to calculate the relative abundances of a protein and its post-translational modifications (PTMs), if present. The deconvoluted MS may contain multiple peaks corresponding to a protein and its PTMs. Estimating the area under each of these peaks is needed for calculating the relative abundances. Bruker's software does not have an automated area calculation routine and manual identification of the boundaries of each peak is needed. This can be time consuming if there are multiple signal peaks in each spectrum and a lot of protein samples. An "automated area calculation" routine is highly desirable. The idea is to fit an n^{th} order polynomial, $P(x)$, to each of the detected peaks.

$$P(x) = \sum_{i=0}^{n} p_i x^i$$

The inflection points of $P(x)$ are the roots of the equation, $\frac{d}{dx}P(x) = 0$. Boundary points are the inflection points on either side of the peak. Since there are at least 3 inflection points in a peak, the polynomial order is chosen as $n \geq 4$. The curve fit itself need not be highly accurate as it is used for locating the boundaries (inflection points) of only the peak signal. The area is calculated using trapezoidal numerical integration method [132] on the original deconvoluted MS peak.

4.2.6 Results

The algorithm is implemented in the MATLAB version 7.7 software package. The GUI for the input parameters is shown in Fig 4.4. The default parameters values are for the protein DBP.

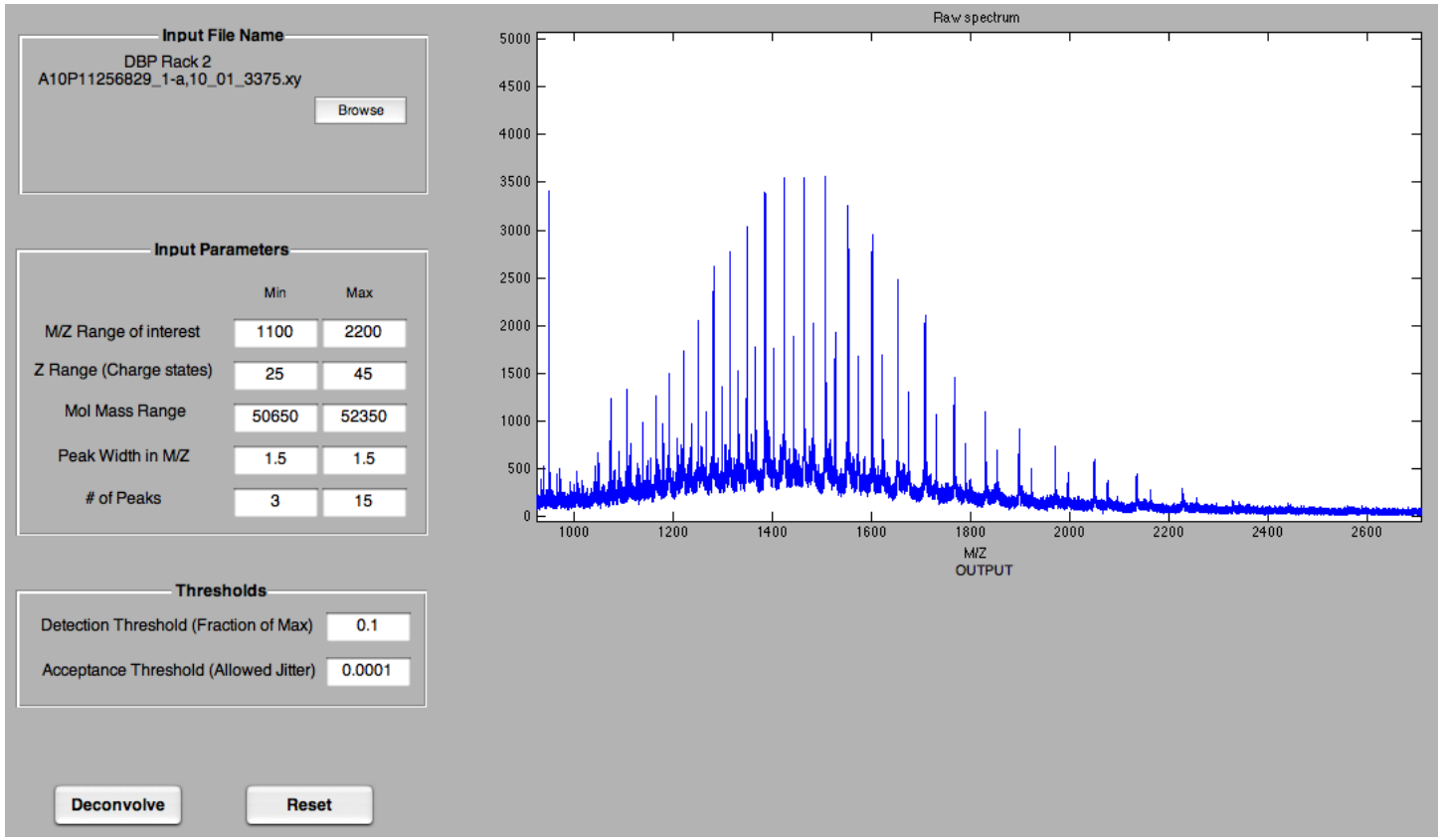
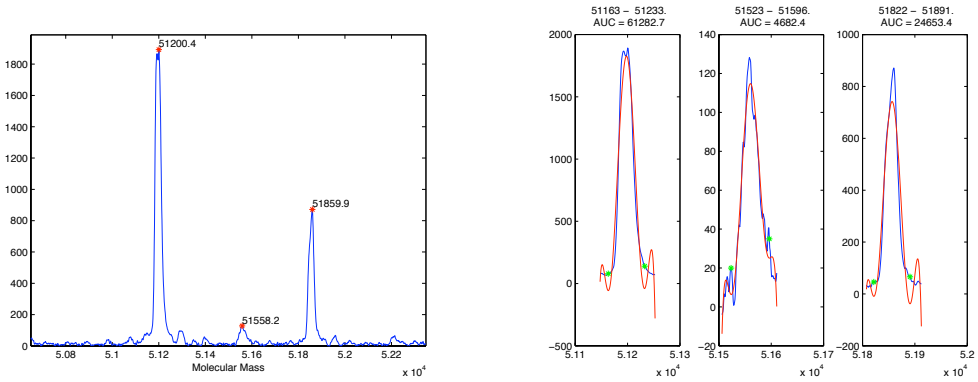


Figure 4.4: GUI with input parameters (for DBP)

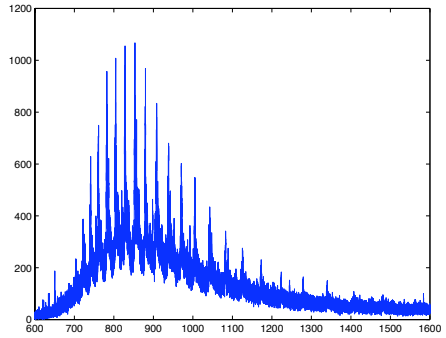


(a) Deconvoluted MS for DBP (b) Curve fitting for Area calculation

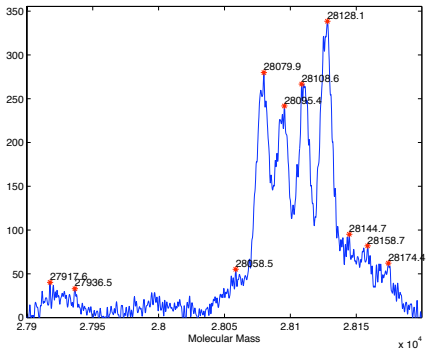
Figure 4.5: Deconvolution of a DBP sample

When an MS file is selected, it is displayed within the GUI. The deconvoluted MS is shown in Fig 4.5(a). The peak locations are marked in the deconvoluted spectrum. Detection threshold, γ_{det}, can be adjusted to include a peak of interest or conversely to exclude a peak. The curve fitting and area under curve (AUC) for each of the peaks is shown in Fig 4.5(b). Fig 4.6 shows another example for protein ApoA1. Changing the range of molecular mass and γ_{det} (from 0.15 to 0.30) helps in getting rid of unwanted peaks in the deconvoluted spectrum. As seen, the peaks have good boundary estimates for area calculation.

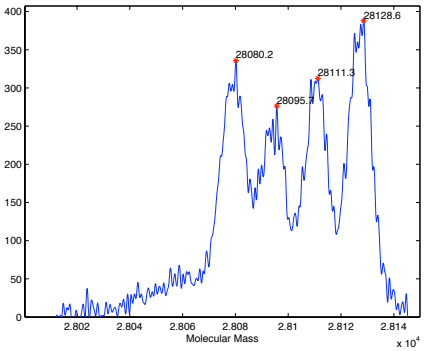
This algorithm can process multiple MS files without requiring any user intervention as long as the parameters remain the same. This is very helpful in processing MS data for the same protein taken from multiple samples. The deconvoluted spectrum and area under the peaks are the output. The proposed algorithm is much faster than Bruker's software, in the sense that the area calculation is fully automated. The deconvolution process itself is comparable to the existing method.



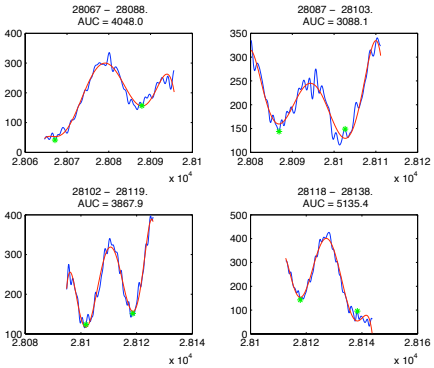
(a) Raw MS for ApoA1



(b) Deconvoluted MS for ApoA1



(c) Deconvoluted MS with changed parameters



(d) Curve fitting for Area calculation

Figure 4.6: Deconvolution of an ApoA1 sample

4.3 New Method of Peak Detection and Area Calculation

Peak detection and amplitude estimation algorithms were developed in the previous chapter. It was demonstrated that the noise statistics, along with baseline, is adequately modeled by a Gamma distribution. Signal detection in the presence of Gamma distributed noise is to be carried out according to the test statistics in Eqns 3.50 and 3.52, henceforth called the approximate detector and GLRT detector respectively. In the next section, the detector and estimator performance is evaluated.

4.3.1 Detector Performance

Monte Carlo simulation method is used to evaluate the performance of the two detectors by getting ROC plots. The Gamma parameters are fixed ($k = 1.25, \theta = 40$), so that the noise signal is generated $w_{mc} \sim \Gamma(1.25, 40)$ with variance $\sigma_w^2 = k\theta^2 = 2000$. The signal peak is a Gaussian shaped pulse generated as follows:

$$s_{mc}[n] = \exp\left(\frac{-h[n]^2}{2\sigma^2}\right)/(2\sigma^2) \quad n = 0, \cdots, N-1$$

where $h \in [-5, 5]$ is a vector with $N = 100$ and σ is the pulse width fixed at 0.25. This peak width is similar to the m/z widths seen in the protein peaks. s_{mc} amplitude is then normalized to 1. The variance of s_{mc} is the sample variance (σ_s^2). The signal-to-noise ratio (SNR) is defined as $10\log_{10}\frac{\sigma_s^2}{\sigma_w^2}$. The amplitude A_{mc} is then chosen such that the SNR in the range of -20 dB to -10 dB, and $x_{mc} = w_{mc} + A_{mc}s_{mc}$.

A total of 25,000 iterations is performed for each amplitude level. A false alarm is counted when the detector output (sufficient statistic in Eqns 3.50 and 3.52) exceeds the threshold (γ_{mc}) when $A_{mc} = 0$. A detection is counted if the detector output exceeds γ_{mc} when $A_{mc} \neq 0$. The false alarm and detection counts are then normalized by the total number of iterations to get P_f and P_d. The receiver operating characteristic (ROC) is a plot of P_d Vs. P_f obtained by varying γ_{mc} from $-\infty$ to $+\infty$. ROC curves at the above mentioned SNR levels, corresponding to different A_{mc} is plotted. The ROC plots for the two detectors is shown in Figs 4.8 and 4.7.

As expected, the detection performance improves with SNR for both the detectors. At all SNR levels, the approximate detector, for low amplitude levels, has a better detection performance (in terms of P_f) compared to the GLRT detector. For the GLRT detector, an estimate of the amplitude ($\hat{A}_{mc}$) is needed, as given in Eqn 3.56. The mean and variance of all the $\hat{A}_{mc}$s from the Monte Carlo simulation

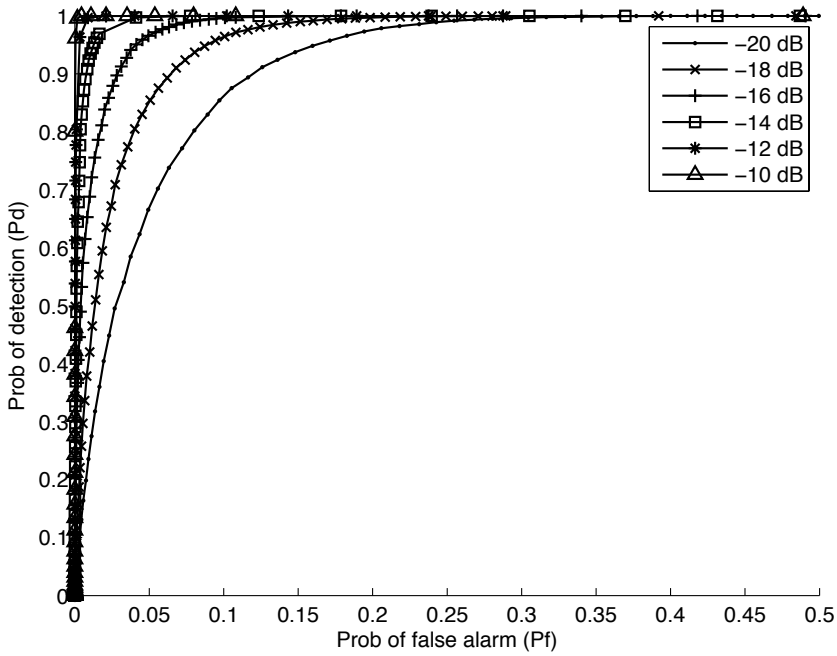


Figure 4.7: ROC for detector in Eqn 3.50

is shown in Table 4.1, normalized to $\frac{\hat{A}_{mc}}{A_{mc}}$. There is a positive bias ($E(\hat{A}_{mc}) \geq A_{mc}$) in the estimates and the bias decreases with increasing SNR. The variance of the estimates (var($\hat{A}_{mc}$)) also decreases with increasing SNR.

4.3.2 Fractional Abundance Calculation

This detection and estimation method, in conjunction with trapezoidal integration [132] is used to estimate the area under the relevant peaks in ESI MS. Instead of calculating an average MS from multiple frames chosen from the chromatogram (Fig 4.3), raw data from all the relevant frames are used for peak detection and estimation. Again, care is taken to choose the frames where most of the sample molecules are reaching the detector. The individual MS frame data are not usually accessible to the end users as they are stored in a proprietary data format. For the particular ESI spectrometer used for the experiments in this research, CompassX-

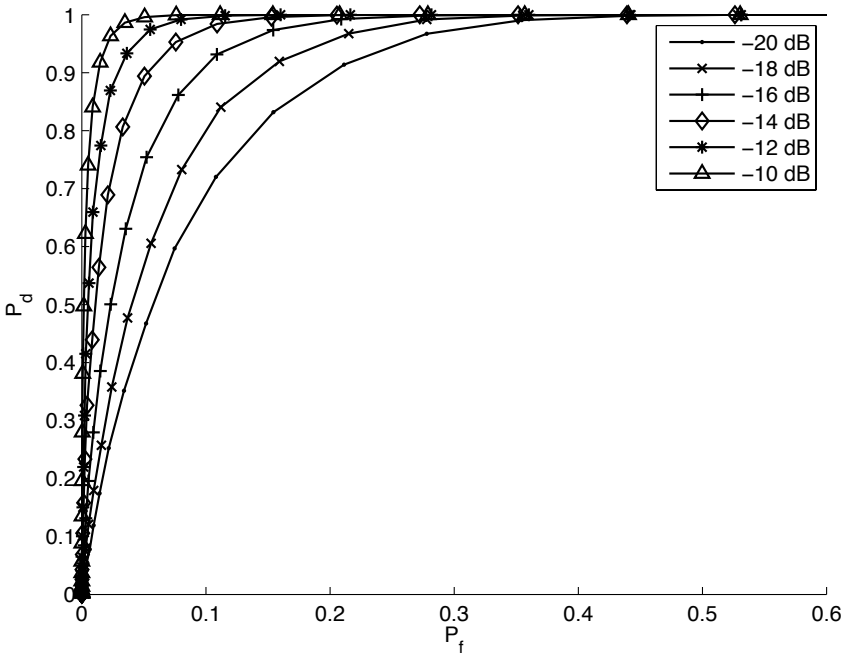


Figure 4.8: ROC for detector in Eqn 3.52

Table 4.1: $\hat{A}_{mc}$

SNR (dB)	Mean	Variance ($\times 10^{-3}$)
-20	1.57	147.5
-18	1.45	99.7
-16	1.35	65.9
-14	1.26	40.5
-12	1.20	26.0
-10	1.15	16.9
-5	1.03	1.1
0	1.01	0.3

port [133] software is utilized to convert the raw MS from the proprietary Bruker data format (.BAF) to an universal mzXML [134] data format.

Atomic composition and a range of possible charge states of the molecular species are the only input required for the abundance calculation routine. There are no pre-processing steps involved and no need for a deconvolution routine. Using the isotopic distribution of the molecule, the peak width is estimated for each charge state. The isotopic distribution provides the theoretical peak location in the MS. Often the MS frames are mis-aligned and the actual peak location may shift along the m/z axis. Hence, the peak detection algorithm is used to search over a window around the theoretical location. The maximum detector output, when greater than the threshold, is the location of the peak in the MS. The estimated amplitude at that location is then used to calculate the AUC of that charge state for that species. Unlike the previous algorithm, the AUC is calculated from the theoretical peak instead of the MS peak. This way the noise is not accounted for the abundance calculations. Fig 4.9 shows the result of such a detection and estimation method applied to a single frame and charge state of HSA ($C_{2936}H_{4591}N_{786}O_{889}S_{41}$) sample. The top-left figure shows the raw MS data with the theoretical location of the modeled peak for the particular charge state. The detection algorithm is used to locate the actual location of the peak with in a window of the theoretical location, as seen in the bottom right figure. The top-right figure shows the estimated amplitude and the location of the peak. As the noise model includes the baseline, the estimated amplitude is significantly smaller than the peak, so a baseline correction routine is unnecessary. The bottom-right figure shows the effect of adding the baseline to the estimated peak. These plots show that the baseline line is automatically eliminated before abundance calculation is carried out. This step is repeated for all the charge states and selected frames for a given molecular species.

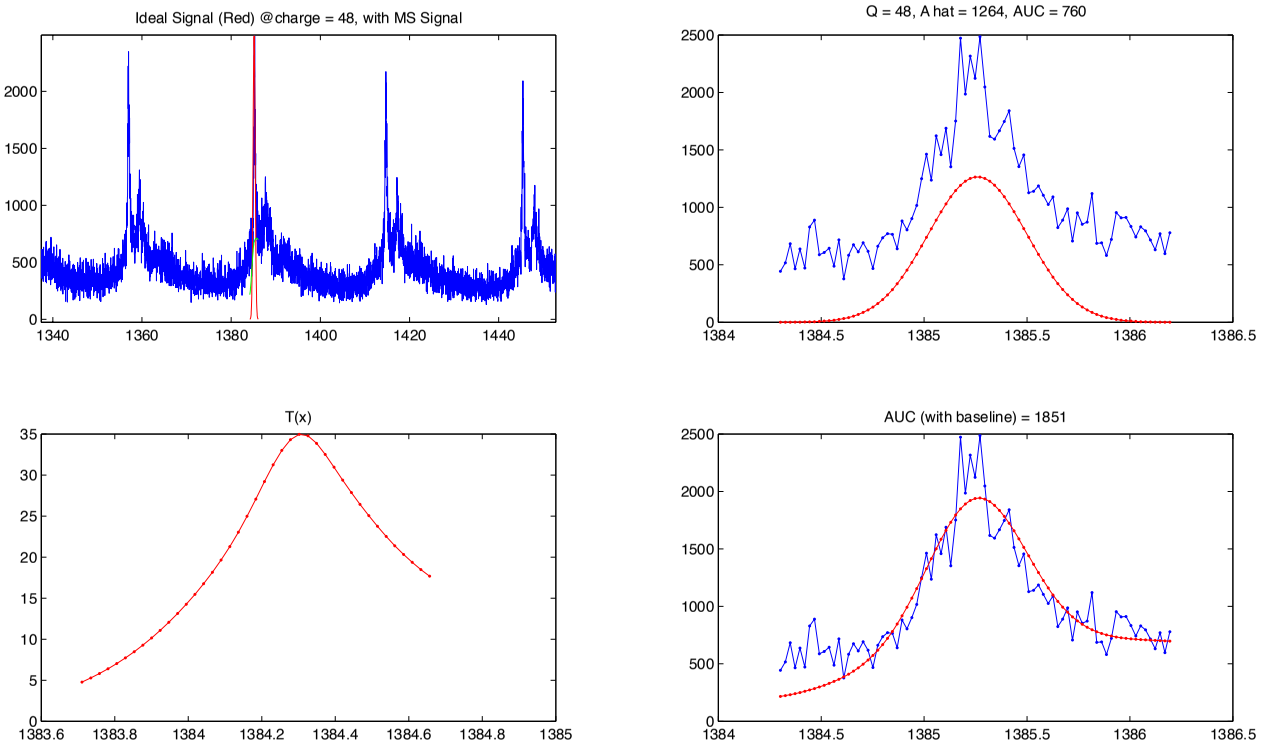


Figure 4.9: Abundance estimation for one charge state of HSA

Table 4.2: Fractional Abundances of HSA

No	Current Method	Proposed Method	Avg SNR (dB)
1	0.87	0.67	-0.3
2	0.83	0.65	-3.4
3	0.86	0.73	-6.7
4	0.74	0.60	-0.6
5	0.81	0.71	-3.0

The abundance of a molecular species is measured in terms of its AUC. Using the new method, the abundance of a molecular species is the sum of the AUCs of all charge state peaks in all the frames. The relative abundance is the ratio of the area of targeted species over the total area of all different molecular species present in the sample. In the example presented here, HSA is the primary molecular species and Cysteinylated HSA (Cys-HSA) is one of its molecular variant. Cys-HSA has a larger molecular weight due to the addition of the Cysteine residue attached via a disulfide bond. The net molecular mass change is a $\sim$120 Da increase compared to the intact HSA molecule. The relative abundances of Cys-HSA and HSA are calculated for 5 samples runs. The results from both the current and the proposed algorithms is presented in Table 4.2.

There is no ground truth for comparing the fractional abundances, as the molecular species (of HSA) are from samples drawn from human serum. The abundance estimates in the proposed method are consistently smaller compared to the current method. One probable reason may be due to the summing of frames in the current method. As the user selects a window of frames from the chromatogram, there is the possibility of noisy frames (low SNR) being included in the final MS, which can drown out the smaller peaks. For the new proposed method, the abundances calculation is performed by taking the AUC of the theoretical peak, once the location and amplitudes are estimated, instead of the actual MS peak. A noisy

MS peak will result in a higher abundance estimate compared to the clean theoretical peak, as the noise variance remains similar across the different species for a given charge state, a smaller peak will have a relatively higher AUC in the current algorithm compared to the proposed algorithm.

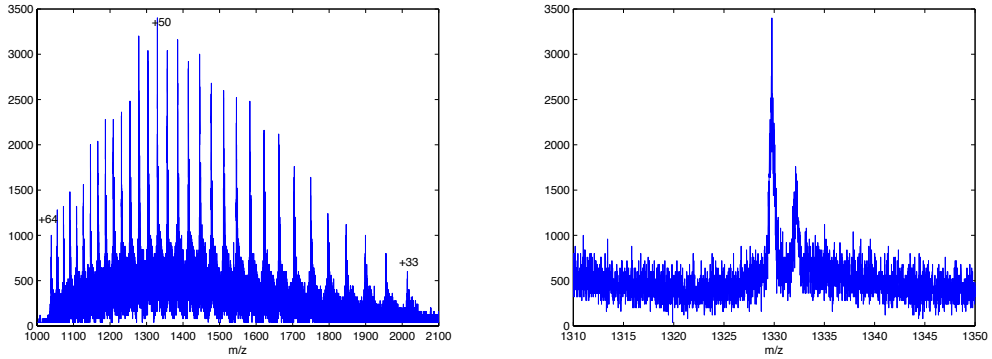
4.3.3 Monte Carlo Simulation

The fractional abundance estimates can be compared to a true abundances through a Monte Carlo simulation performed according to the signal-baseline model (Eqn 3.38):

$$g(y) = \begin{cases} \dfrac{a_1}{1+\lambda_1(y-y_0)^2} + a_2\, e^{-\dfrac{(y-y_0)^2}{2\lambda_2^2}} & \text{if } y < y_0 \\ \dfrac{a_3}{1+\lambda_3(y-y_0)^2} + a_2\, e^{-\dfrac{(y-y_0)^2}{2\lambda_2^2}} & \text{if } y \geq y_0 \end{cases}$$

For the HSA sample with or without the presence of Cys-HSA, a_1, a_2, and a_3 are chosen to be 0.15, 0.10, and 0.75 respectively. These values are selected to mimic the true MS data. Only 10% of the HSA molecules are present in the signal peak and the rest end up in the tails. The tail decay rates are determined for each possible charge state of HSA using the true MS data as a guide. The position of the peaks, y_0, is easily calculated from the average molecular mass of HSA and the charge state, which is chosen to be a number between 33 and 64. The total number of molecules of the two HSA forms is fixed and is distributed among the different charge states. The true relative abundance can be varied by varying the number of molecules used in the simulation. Fig 4.10 shows the result of a simulation with 500,000 HSA and 200,000 Cys-HSA molecules, i.e., a relative abundance of 0.2857.

The true relative abundance is varied by varying the total number of Cys-HSA molecules while keeping the HSA abundance constant at 500,000. A total of 10 relative abundances (of Cys-HSA) between 0 to 0.5 is used for the Monte Carlo simulation while iterating each 10 times. Both the current and proposed



(a) All charge states (b) Charge = 50

Figure 4.10: Simulated MS of HSA and Cys-HSA (Relative abundance = 0.28)

Table 4.3: Simulation results (Average of 10 iterations)

No	True Abundance	Current Method	Proposed Method
1	0.0196	0	0.0384
2	0.0530	0	0.0678
3	0.0842	0.0967	0.0990
4	0.1135	0.1236	0.1259
5	0.1490	0.1462	0.1547
6	0.1667	0.1710	0.1813
7	0.2857	0.2899	0.2968
8	0.3750	0.3809	0.3861
9	0.4444	0.4486	0.4580
10	0.5000	0.5051	0.5166

algorithms are used to estimate the relative abundances. The abundance estimates for the proposed algorithm is adjusted for the bias of amplitude estimates as given in Table 4.1. Only the signal peaks are considered for the abundance estimation. The average of the abundance estimates from both the algorithms is shown in Table 4.3.

4.3.4 Discussion

The true fractional abundances of HSA and Cys-HSA drawn from real biological samples are not available and hence a Monte Carlo method is used to simulate MS

data using the signal-baseline mixture model. Table 4.3 shows the result of the abundance estimates from both algorithms. It is clear that the proposed algorithms is more useful for low abundances (< 10%). The current algorithm doesn't yield good estimates in this range because lowering the threshold results in the identification of a lot of spurious peaks in the deconvolution. The estimates for higher abundances in the proposed algorithm are not as accurate as the results from the current algorithm. The error is less than 10% for abundances greater than 0.15 and decreases with increasing relative abundance. These results suggest that other algorithms of estimating peak amplitudes needs to be explored.

The main advantage of the proposed method is the lack of any pre- or post-processing of the raw MS data. As the noise and baseline is built into the model, the algorithm estimate the fractional abundance due to the signal part only. There is also no need for any MS alignment methods to align all the frames with respect to a reference peak, as the detection algorithm searches for the signal within a window of the theoretical location which is estimated from the isotopic distribution of the molecular species. Fewer number of input parameters are needed in this proposed algorithm compared to the current algorithm. The parameters required for the proposed algorithm is the molecular formula of the primary molecule of interest and a range of charge states. By adjusting the threshold (for $T(x)$ or $\hat{A}$ or SNR), the user can choose between signal peaks that go into the fractional abundance estimation.

For the proposed algorithm, Gamma noise parameters are estimated from simulated MS data without the presence of MS signal peaks. These parameters are used for all the other simulated data. As these parameters are estimated from the baseline in the MS, an inaccurate parameter estimate will result in inaccurate abundances. This algorithm also tends to be slower than the current algorithm, as

individual frames are processed instead of the sum of a range of frames.

This proposed algorithm can be improvised to be used for detecting overlapping peaks. This can be achieved by calculating the isotopic distribution and hence the peak shape of multiple species together. The proposed algorithm can be useful in other MS applications such as protein/peptide identification. It can also be modified for different peak shapes, eg., the signal-baseline model.

4.4 CONCLUSION

This chapter introduced the concept of deconvolution of ESI MS and fractional abundance estimation in the context of biomarker analysis. One of the many current software used for these applications is discussed in detail. The algorithm requires many input parameters and has a plethora of pre-processing routines before deconvolution and fractional abundance calculations take place. Moreover, the Bruker software requires manual identification of peaks for abundance calculation. A new, automatic, fractional abundance estimation routine is added to the algorithm and shown to provide satisfactory abundance results while speeding up the process multi-folds.

The performance of the detector and estimator models developed in Chapter 3 are evaluated using Monte Carlo simulation. ROC plots of the two detectors show that the approximate detection scheme performs better than the GLRT detector at the low SNRs. The amplitude estimator has a bias at the low SNRs and improves with increasing SNR from -20 dB to -10 dB. A fundamentally novel abundance estimation technique is proposed in the last part, which is based on these detection and ML estimation schemes. This has the potential to provide more accurate estimates, because of the use of raw MS data and no pre-processing routines. Further tests and analysis has to be carried out to extend the method to more complex spec-

tra and make the process faster. A signal model that includes the baseline instead of considering it a distortion can be considered as well for future research in this area.

CHAPTER 5

BIOMARKER DISCOVERY

5.1 INTRODUCTION

Diabetes is a chronic disease that has reached epidemic proportions in the US. It is caused by high level of blood sugar, otherwise called hyperglycemia. Most people with diabetes have either type 1 or type 2 diabetes. Type 1 diabetes results from lack of insulin, a hormone that regulates the level of glucose in blood. Type 2 diabetes (T2D) results from insulin resistance of the cell thus making insulin less effective in regulating glucose. Traditionally, T2D is diagnosed by measuring the absolute concentration of glucose. The level of glucose is also reflected by hemoglobin, an oxygen carrying protein in the blood. Hemoglobin undergoes glycation, defined as the bonding of a protein and a sugar molecule, when it is exposed to glucose. This is also one form of post translational modification (PTM) of the protein. Increased levels of glycated hemoglobin (HbA1c) in the blood is an indicator of hyperglycemia. This has led to the acceptance of HbA1c as a marker for diabetes [135] recently. Fig 5.1 shows the relative abundance of HbA1c (using mass spectrometry) in an individual diagnosed with T2D in comparison to a healthy person. As seen, HbA1c has a higher relative abundance for the T2D sample than the healthy sample. The glucose molecule adds 162 Daltons to the molecular mass of hemoglobin.

The lack of insulin in type 1 diabetes means that insulin therapy is the only effective treatment. Type 2 diabetes has a wider range of therapies available. However, diabetes has been associated with an elevated risk of cardiovascular disease (CVD), the leading cause of mortality among the patients. Hence the T2D therapies have to be evaluated for their effect on cardiovascular risks. Recently, much T2D research has centered around the connection between poor glucose control

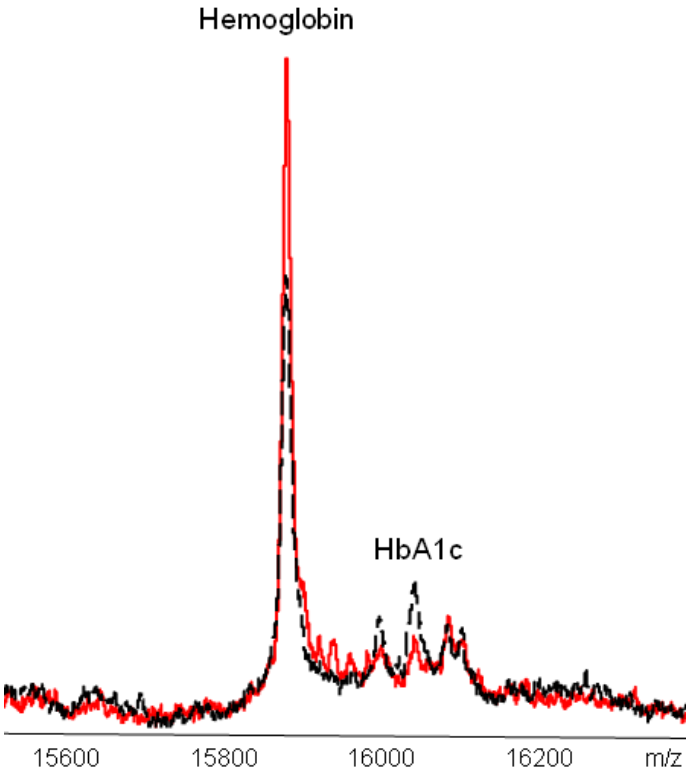


Figure 5.1: The relative abundance of Hb1Ac in a healthy (Solid) and T2D (Dash) sample. (Hemoglobin is made of 2 chains A & B which split apart during the mass spectrometry process. Clinically, HbA1c from the B chain is measured for T2D. And as seen, the diabetic sample has a higher abundance of HbA1c)

and CVD. In many cases, even a tight control on HbA1c has not resulted in cardiovascular benefits. As a result, FDA issued a Guidance of Industry statement [136] suggesting that developers of new anti-diabetic drugs demonstrate that the therapies will not result in an increase in cardiovascular risk. This led to an extended definition of diabetes, challenging the drug development industry to monitor markers for both T2D and CVD during drug trials which in turn led researchers to work towards identifying markers for monitoring T2D and related cardiovascular complications.

It is hypothesized that proteins, like HbA1c, in addition to carrying time-

cumulative marks of hyperglycemia, may also carry similar post-translational information with respect to systemic oxidative stress and aberrant enzymatic signaling that would be indicative of cardiovascular risk [52]. Various proteins that undergo PTM (glycation, oxidation, truncation) are identified. Immunoassay technique is used to measure the relative abundances of the proteins and their modified versions in different patient groups. This information is then evaluated with respect to the effectiveness in distinguishing groups with a history of T2D, cardiac heart failure (CHF) and myocardial Infraction (MI) by using multidimensional approaches such as support vector machines.

Support vector machines (SVMs [50]) have been popular classification tool, more so in the field of biomarker discovery. SVMs have been shown to be one of the top performing biomarker classification algorithms when compared to other methods such as LDA, KNN, etc [44,54]. In this chapter, a new multi-classification SVM algorithm is developed and the performance is compared to LDA.

5.2 Multi-classification using SVM

Support vector machines (SVMs) [5, 50], are a supervised learning method used for binary classification. The idea behind SVMs is to find a hyperplane such that it separates a multi-dimensional data set into two classes. This approach makes it a non-probabilistic classifier. As in any supervised learning methods, the SVM algorithm needs training data, belonging to either of two classes, to build a model that can be used to classify new/test data.

5.2.1 Introduction

Consider a training dataset of n, d-dimensional points:

$$\mathbf{D} = \left\{ (\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\} \right\}_{i=1}^{n}$$

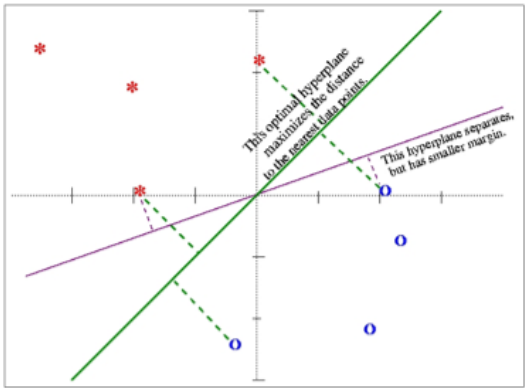


Figure 5.2: Hyperplane maximizing margin [5]

where y_i denotes the class, either 1 or -1, indicating the class to which the point $\mathbf{x}_i$ belongs. The goal is to find a hyperplane that divides the points according to their class. Any hyperplane can be written as the set of points $\mathbf{x}$ satisfying,

$$\mathbf{w} \cdot \mathbf{x} - b = 0$$

where, $\mathbf{w}$ is a vector perpendicular to the hyperplane and $\frac{b}{\|\mathbf{w}\|}$ is the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$. If the data $\mathbf{D}$ are linearly separable, the canonical hyperplane can be defined as the one which separates the data from the hyperplane by a functional distance of at least 1.

$$\mathbf{w} \cdot \mathbf{x}_i - b \geq +1 \qquad \text{when} \qquad y_i = +1$$

$$\mathbf{w} \cdot \mathbf{x}_i - b \leq -1 \qquad \text{when} \qquad y_i = -1$$

This can be rewritten as:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1, \qquad \forall i \tag{5.1}$$

For a given hyperplane $(\mathbf{w}, b)$, all pairs $\{\lambda\mathbf{w}, \lambda b\}$ are also the exact same hyperplane, but each has a different functional distance to a given data point, as

shown in Fig 5.2. The geometric distance from the hyperplane to a data point is obtained by normalize the distance by the magnitude of $\mathbf{w}$:

$$d\left((\mathbf{w},b),\mathbf{x}_i\right) = \frac{y_i(\mathbf{x}_i \cdot \mathbf{w} + b)}{\|\mathbf{w}\|} \geq \frac{1}{\|\mathbf{w}\|} \tag{5.2}$$

i.e. $(\mathbf{w},b)$ is a hyperplane that maximizes the geometric distance to the closest data points. This can be accomplished by minimizing $\|\mathbf{w}\|$, subject to the distance constraints. The solution will be the same if $\|\mathbf{w}\|$ is substituted with $\frac{1}{2}\|\mathbf{w}\|^2$. This quadratic programming (QP) optimization problem, with non-negative Lagrange multipliers α_i can be expressed as:

$$\min_{\mathbf{w},b} \max_{\alpha} \{\frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x_i} - b) - 1]\} \tag{5.3}$$

The solution can be expressed in terms of linear combination of the training vectors as:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i} \tag{5.4}$$

Only for a few cases, $\alpha_i > 0$ and the corresponding $\mathbf{x_i}$ are the support vectors, which lie on the margin and satisfy $y_i(\mathbf{w} \cdot \mathbf{x_i} - b) = 1$. Hence, the offset parameter is:

$$b = \mathbf{w} \cdot \mathbf{x_i} - y_i \tag{5.5}$$

By substituting $\mathbf{w}$ from Eqn 5.4 in Eqn 5.3, the dual problem can be written as:

$$\text{maximize:} \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \tag{5.6}$$

$$\text{subject to:} \quad \alpha_i \geq 0, \qquad \forall i \tag{5.7}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \qquad \forall i \tag{5.8}$$

In case the data is not linearly separable but instead can be done using a polynomial curve, then a different kernel $k(\mathbf{x}_i, \mathbf{x}_j)$ can be used instead of the linear kernel $\mathbf{x}_i \cdot \mathbf{x}_j$ in Eqn 5.6. Examples are polynomial, gaussian, hyperbolic tangent etc.

5.2.2 SVM for Multiclass Classification

SVMs are binary classifiers by default, but some datasets have more than two groups to be classified. There are some proposed methods in the literature [137, 138], that extend the binary SVM to a multiclass case (MSVM). The two general strategies are either to solve the multiclass case by solving a series of binary problems, or to consider all the classes at once. There are two common and simple methods that use the first strategy, wherein each classifier distinguishes between (a) one of the classes to the rest (one-versus-all) or (b) between every pair of classes (one-versus-one). In one-versus-all case, classification of a test data is done by a winner-takes-all strategy, in which the classifier with the highest output function assigns the class. For the one-versus-one approach, classification is done by a max-wins voting strategy, where the class with most votes (from each binary classifier) determines the result. The authors in [139] compare the different strategies and algorithms. The one-vs-rest approach is expanded to an any-vs-rest approach where the binary groups are formed by taking any combination of classes in one group. This is a super set of one-vs-rest.

Consider a multiclass dataset (k groups):

$$\mathbf{D} = \left\{ (\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{1, 2, \cdots, k\} \right\}_{i=1}^{n}$$

The k groups can be divided into $\kappa = 2^{(k-1)} - 1$ any-vs-rest binary classes (each binary class can have one or more of the original groups). The training data, $\mathbf{x_i}$, will now have a binary class, $(I_i \in \{+1, -1\},\ \text{for } i = 1, 2, \cdots, \kappa)$, as a result of the κ SVM classifiers. The data set with the new class assignment can be expressed as:

$$\mathscr{D} = \left\{ (\mathbf{x}_i, y_i, \mathbf{I}_i) \mid \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{1, 2, \cdots, k\}, \mathbf{I}_i \in \{-1, 1\}^{\kappa} \right\}_{i=1}^{n}$$

For a new test data, $\mathbf{x}_{test}$, a MAP decision rule can be used to estimate $\hat{y}_{test}$ as

follows:

$$\begin{aligned}\text{Classify}(\hat{y}_{test} = i) &= \arg\max_i P(y_{test} = i|\mathbf{I_{test}}) \\ &= \arg\max_i \frac{P(\mathbf{I_{test}}|y_{test} = i) \cdot P(y_{test} = i)}{P(\mathbf{I}_{test})}\end{aligned} \tag{5.9}$$

5.2.3 Parameter Estimation

There might not be enough data to estimate the likelihood probability for all possible values of $\mathbf{I}$ and y, especially when k is large. In such instances, Monte Carlo methods can be used to simulate data from the available statistics. Even with enough data, sometimes particular sequences of $\mathbf{I}$ may be more abundant and others less so. This results in the need for estimating the likelihood probability of an object (in this case, $\mathbf{I}$) that has never been seen before. Good-Turing methods [140, 141] are useful in estimating these probabilities.

The simple Good-Turing (SGT) method [141], uses a straight line to smooth the regions of inaccurate probability estimates. If there are a total of N samples with N_r distinct species represented exactly r times, $\sum_r rN_r = N$. N_r is called the "frequency of frequencies" because r is the frequency of occurrence. The original data sample may have $N_r = 0$ for certain values of r. For the maximum likelihood estimate, $p_r = \frac{r}{N}$ and $p_0 = 0$. So, estimates of p_0 are to be obtained. In SGT this is accomplished as:

$$p_r = \frac{r^*}{N} \tag{5.10}$$

where

$$r^* = (r+1)\frac{E(N_{r+1})}{E(N_r)} \tag{5.11}$$

There is a need to account for the N_r's which are zero. One way of achieving this is by averaging with each non-zero N_r, the zero N_r's that surround it. In other words, order the non-zero N_r by r and let q, r, t be successive indices of non-zero values.

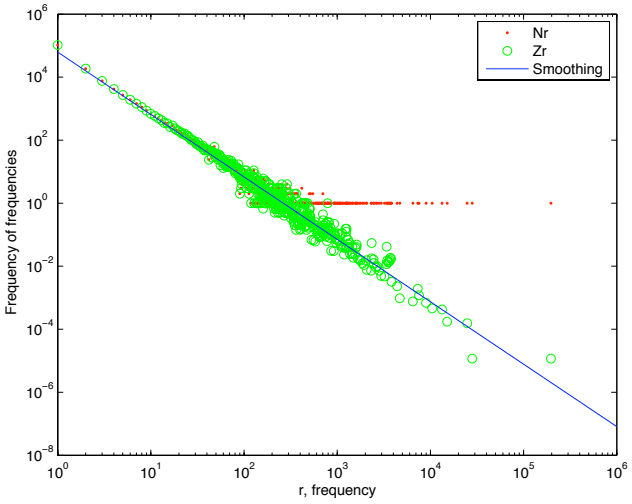


Figure 5.3: Smoothing in SGT (y-axis is in 'log' scale)

N_r is replaced by:

$$Z_r = \frac{N_r}{0.5(t-q)} \tag{5.12}$$

For small r there is no difference since $t - q = 2$, but for large r, there can be significant changes. The result can be seen in Fig 5.3, for the T2D data (includes simulated data). Smoothing, in SGT, is done by fitting a line of the form,

$$log(Z_r) = a + b\ log(r) \tag{5.13}$$

Estimates of a, b can be substitute $E(N_r) = Ar^b$ in Eqn 5.11, and eventually estimate p_r in Eqn 5.10.

5.3 Experimental Setup

Binary classification is performed using linear and quadratic kernel SVMs. Both One Vs Rest and Any Vs Rest multi-classification methods are evaluated. A linear discriminant analysis (LDA) is also performed for comparison. Half of the available data, chosen at random, is used for training all the classifiers and it is iterated 10 times. The data is normalized to zero mean and unit variance.

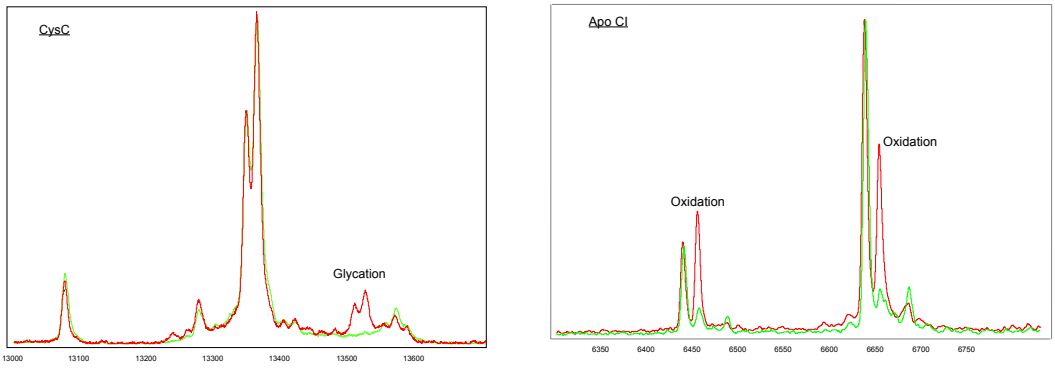
Table 5.1: Patient Classes

Patient Class (y_i)	Abbreviation	# Samples
Healthy	Hea	66
Type 2 Diabetics	T2D	50
Congestive heart failure	CHF	29
CHF, T2D	C/T	25
CHF, Myocardial Infarction	C/M	25
CHF, T2D, MI	C/T/M	17

5.3.1 Data

Standardized mass spectrometric immunoassay techniques were used to analyze proteins and their variants in plasma samples from different classes of patients. The patients are clinically diagnosed to have a history of type 2 diabetes and/or cardiovascular disease. The 6 classes of patients and the number of patients in each group, totaling 212 ($= n$) individuals, is shown in Table 5.1. The 6 classes are divided into 31 ($2^{6-1}-1$) binary groups. A binary SVM classifier is designed for each of these 31 binary groups.

The proteins under investigation included: albumin, apolipoprotein A-1, C-1, and C-2 (ApoA1, ApoC1, and ApoC2, respectively), vitamin D binding protein (DBP), transthyretin (TTR), β-2 microglobulin (B2M), cystatin c (CysC), serum amyloid P (SAP), c-reactive protein (CRP), and the chemokine RANTES. In total, $\sim 2,300$ assays were performed (212 individuals x 11 proteins), during which 41 ($= d$) different molecular species (proteins and their variants) were identified, producing $8,692$ data points, stored as a 212 x 41 matrix. The mass spectra of PTM forms of some of these protein, comparing healthy and T2D groups is shown in Figs 5.4.



(a) CysC and Glycation (b) ApoC1 and Oxidation

Figure 5.4: PTM of proteins comparing Healthy (green) and T2D (red) sample

5.3.2 Forming any-vs-rest groups

The 6 classes are divided into 31 ($2^{6-1}-1$) any-vs-rest binary groups. The list of all 31 binary classes is shown in Table 5.2. As seen in the table, No. 1-6 are the one-vs-rest binary classifiers. A binary SVM classifier is designed for each of the

Table 5.2: Binary grouping for the 6 classes (Any Vs Rest)

No	Super Class	No	Super Class
1	Hea	17	C/T/M+C/T
2	T2D	18	C/T/M+C/M
3	C/T/M	19	CHF+C/T
4	CHF	20	CHF+C/M
5	C/T	21	C/T+C/M
6	C/M	22	Hea+T2D+C/T/M
7	Hea+T2D	23	Hea+T2D+CHF
8	Hea+C/T/M	24	Hea+T2D+C/T
9	Hea+CHF	25	Hea+T2D+C/M
10	Hea+C/T	26	Hea+C/T/M+CHF
11	Hea+C/M	27	Hea+C/T/M+C/T
12	T2D+C/T/M	28	Hea+C/T/M+C/M
13	T2D+CHF	29	Hea+CHF+C/T
14	T2D+C/T	30	Hea+CHF+C/M
15	T2D+C/M	31	Hea+CHF+C/T
16	C/T/M+CHF		

31 binary data groups using 50% of the data as training, chosen randomly from each group. The same process is repeated using the entire data set as training. Both linear and quadratic kernels are used in each case.

5.4 Results And Discussion

The classifiers are implemented in the MATLAB Ver. 7 software. The 31 any-vs-rest classifiers are trained using both linear and quadratic kernel with 50% of available data chosen at random. The (6) one-vs-rest binary classifiers are a subset of these 31. A linear discriminant analysis is also evaluated for comparison. All results are averaged over 10 iterations. For multiclass classification using the MAP decision rule of Eqn 5.9, SGT method is used to make a table of likelihoods. Both the linear and quadratic kernel SVMs were used. The result from these two any-vs-rest multiclass classifiers (train with 50% and test with rest) is shown in Table 5.3. It is in the form of a confusion matrix, where the columns represent the true classes, and rows represent the classifier output. The top half of the table is for linear kernel and bottom half is for quadratic kernel. Each column, for a given kernel, adds up to one. This is repeated for one-vs-rest multiclass SVM classifier and linear discriminant analysis. The results are shown in Tables 5.4 and 5.5 respectively.

The linear kernel results, in Tables 5.3 and 5.4, show that the biomarkers are very effective in separating the healthy and T2D patients from the rest with a small error rate. The classification result for the group of T2D patients with a history of CVD (C/T/M) has the smallest P_{cr} (= 0.56). This may be attributed to lack of training data as the C/T/M group has only 17 subjects. The C/T group has comparatively significant misclassification error with CHF ($P = 0.23$) and this is true across all the different classifiers/kernels. So it is difficult to distinguish C/T (CHF associated with T2D) from CHF. In other words, C/T lies between the T2D-

Table 5.3: Any vs rest Confusion Matrix (Avg. of 10 iterations)

Linear Kernel						
	Hea	T2D	C/T/M	CHF	C/T	C/M
Hea	0.90	0.02	0.03	0.08	0.05	0.01
T2D	0	0.92	0.06	0	0	0.13
C/T/M	0	0.02	**0.56**	0	0	0.10
CHF	0.06	0	0.08	0.80	**0.23**	0.04
C/T	0.02	0	0.09	0.09	**0.65**	0.02
C/M	0.02	0.04	**0.18**	0.03	0.07	0.70
Quadratic Kernel						
Hea	0.90	0	0.01	0.05	0.06	0
T2D	0	0.89	0.02	0	0	0.13
C/T/M	0.01	0.04	**0.51**	0.02	0.05	0.12
CHF	0.04	0	0.08	**0.64**	**0.31**	0.02
C/T	0.03	0	0.14	**0.26**	**0.50**	0.03
C/M	0.02	0.07	**0.24**	0.03	0.08	0.70
	Hea	T2D	C/T/M	CHF	C/T	C/M

Table 5.4: One vs rest Confusion Matrix (Avg. of 10 iterations)

Linear Kernel						
	Hea	T2D	C/T/M	CHF	C/T	C/M
Hea	0.89	0	0.07	0.08	0.06	0.02
T2D	0.01	0.94	0.09	0	0	0.15
C/T/M	0.01	0.01	**0.26**	0.05	0.09	0.15
CHF	0.07	0.03	**0.26**	0.74	**0.34**	0.12
C/T	0.02	0	0.08	0.13	**0.48**	0.02
C/M	0	0.02	**0.24**	0	0.03	**0.54**
Quadratic Kernel						
Hea	0.87	0	0.05	0.03	0.06	0
T2D	0	0.90	0	0	0	0.14
C/T/M	0.07	0.07	**0.51**	**0.21**	**0.21**	**0.21**
CHF	0.03	0.01	0.06	**0.52**	**0.27**	0.03
C/T	0.02	0	0.10	**0.20**	**0.40**	0.07
C/M	0.01	0.02	**0.28**	0.04	0.06	**0.55**
	Hea	T2D	C/T/M	CHF	C/T	C/M

Table 5.5: LDA Confusion Matrix (Avg. of 10 iterations)

	Hea	T2D	C/T/M	CHF	C/T	C/M
Hea	0.89	0	0.02	0.10	0.05	0
T2D	0	0.93	0.15	0	0.02	0.08
C/T/M	0	0.01	**0.34**	0.04	0.06	**0.39**
CHF	0.06	0	0	**0.53**	**0.22**	0
C/T	0.03	0	0	**0.31**	**0.63**	0
C/M	0.02	0.06	**0.49**	0.02	0.03	**0.53**

Table 5.6: Comparison of Correct Rate of Classification (Avg. of 10 iterations)

Classifier	**Kernel**	$\mathbf{P_{cr}}$
Any Vs Rest	Linear	0.82
Any Vs Rest	Quadratic	0.77
One Vs Rest	Linear	0.75
One Vs Rest	Quadratic	0.72
Discr Analysis	Linear	0.74

CHF continuum in this biomarker feature space with some overlap.

The overall correct rate of classification is shown in Table 5.6. Both the multi-classification SVMs, with linear kernel, have better P_{cr} than LDA. Any-vs-rest classification method performs better than one-vs-rest for both linear and quadratic kernels. The linear kernel has a better P_{cr} than quadratic kernel in both classification approaches. This results suggests that the quadratic model, based on training samples is not a good fit for the test samples. Since the aim is to understand how the potential biomarkers contribute towards the identification of diseased states, the linear kernel any-vs-rest SVM classifier is of most interest in this experiment. Other kernels types are not considered as they will not provide a simple cause and effect relation between the biomarkers and the diseased states, which is important for diagnosis and treatment, as in the case of HbA1c and T2D.

5.5 CONCLUSION

The MSIA analysis of 11 proteins among 212 individuals, belonging to various disease groups, resulted in the identification of 41 molecular species as potential markers. SVM based multiclass classification and LDA algorithms are investigated to measure the effectiveness of using these species as detectors of the diseased states. This SVM multiclass classification problem is tackled by using a series of any-vs-rest binary classifiers and using a MAP decision rule. It is found that the any-vs-rest multiclass SVM classifier, with a linear kernel, has a better classification result than the one-vs-rest and LDA methods. Overall, the biomarkers are able to distinguish, with high accuracy, between the groups of T2D with or with out a history of CVD; although the sample of subjects tested in each group may be of inadequate size to explore the full diagnostic strength of the combined biomarker panel. The data adapted discretization, using SVM, for the MAP decision rule is very useful in measuring the effectiveness of the biomarker panel.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

The main objective of the dissertation was to understand the signal and noise characteristics in ESI-TOF-MS and develop statistical models using signal processing techniques. Though ESI MS has been popular in recent years for immunoassay biomarker analysis, the statistical properties of the MS data is not fully understood. This is because the fundamental operations of mass spectrometers are considered too complex with many unknown parameters for any reasonable modeling. This coupled with the fact that the spectrometers are commercial instruments and their design details, data processing algorithms are not readily accessible, makes the modeling problem difficult. Recently, researchers have made efforts to mathematically characterize MS data but most of these models are based on heuristics and some qualitative understanding of the process. Statistical modeling of ESI MS data incase of counting ion or TDC detectors have been developed recently in [55, 92]. TDC detector preserves the rate of arrival of ions in the instrument. This assumption is not valid for ADC type detectors, rendering these models ineffective.

Chapter 2 provided a brief introduction to the working of ESI-TOF-MS in the context of MSIA. The data acquisition or the TOF stage is the most important part of the instrument, in the context of statistical modeling. The data acquisition process using the ADC detector was explained in detail. The ADC samples data in the order of gigahertzs. Almost a thousand spectra are generated per second and the sum of all these spectra, called a frame, is stored in the machine. A typical data acquisition process runs for ten minutes, creating more than 500 frames in each sample run. Though the ADC improves the peak resolution, it introduces new challenges for statistical analysis in the form of electronic noise and variable gain.

Combined with the chemical noise inherent in ESI, the signal detection problem becomes quite involved.

In Chapter 3, the signal and noise models were developed. Unlike the current methods, a novel method of estimating the peak shape and width by employing first principles calculations based on device physics and molecular properties is developed. Particularly the isotopic, spatial and energy distributions were modeled to account for the peak width. An FFT based technique was used to estimate the isotopic distribution of a molecule. Peak width due to spatial and energy distributions of ions in the spectrometer was developed using the kinetics principles of the TOF stage. The peak shape and width calculations are mathematically tractable and provide a sound basis for any assumptions. Chemical noise is a major source of background distortion in ESI MS. Most of the existing noise models are not usually verified with experiments. A new chemical noise model was developed by investigating the experimental data from the mass spectrometer. Careful experiments were carried out by controlling various parameters of the device and GOF tests were used to conclude that the noise samples are distributed according to a Gamma PDF.

Apart from the signal and additive noise, a non-flat baseline is common in MS data. The exact cause of the baseline shape is not understood and it is considered to be an artifact. A more careful analysis of the baseline was proposed in section 3.4 and the statistical properties of chemical noise in the presence of such a baseline was investigated. Again, the GOF tests concluded that the noise is adequately modeled by a Gamma distribution. This new approach eliminates the need for baseline correction algorithms since it is included in the noise and signal models.

Detection methods based on the NP detector and GLRT were developed in section 3.5. The detection scheme was built by considering signal with an un-

known amplitude buried in an additive and independently distributed Gamma noise. An approximation of the optimal NP detector was developed for small signal amplitudes. MLE method was used to estimate the amplitude of the signal peaks for use in the GLRT. These statistical signal processing concepts applied to mass spectrometry has the potential to fundamentally change the way MS analysis is carried out currently, as it eliminates any requirement of pre- and post-processing steps and introduces a mathematically robust detection scheme.

In Chapter 4, the performance of the detection and estimation models was evaluated using Monte Carlo simulation. ROC curves were plotted for various SNR levels, which demonstrated that the detection performance is satisfactory even at low SNRs. The ML estimate of the amplitude does not have a closed form solution and a numerical method was used. The estimates are positively biased at low SNRs and the bias decreases with increasing SNR.

A current algorithm of abundance estimation was explained in detail. Fractional abundance estimation is an important process for biomarker discovery. A new method using the detection and estimation schemes was proposed for abundance calculations. This approach is very different from the current methods as frames are processed individually instead of the sum of the frames. By processing individual frames, the chances of small peaks being buried by noise from other frames is reduced. Though peak identification schemes are not included in this work, it can be considered for future research. The most important advantage of this method is that the usual pre-processing steps, such as, spectral alignment, baseline correction, smoothing, denoising, peak picking etc. are completely eliminated. There is no need for a manual peak boundary identification routine as the area is estimated from the modeled peak with adjusted amplitude.

Classification algorithms are necessary for assessing the effectiveness of biomarkers in discriminating disease groups. In Chapter 5, a new support vector machine (SVM) algorithm for multi-classification was proposed by dividing the multiple groups into binary supergroups. Both linear and quadratic kernels were tested for the any-vs-rest and one-vs-rest SVM multi-classification methods. The biomarker fractional abundance data was derived from the MSIA analysis of 11 proteins among 212 individuals, belonging to T2D, CVD and their combinations. The algorithms were trained using 50% of the available data, chosen at random, and then tested on the remaining. This process was iterated 10 times. The results show that the any-vs-rest classification method performed better than one-vs-rest for both linear and quadratic kernels and all SVM methods performed better than LDA. Overall, the biomarkers were able to make distinctions, with a high accuracy, between the disease groups.

6.2 Future Work

In section 3.4 a careful analysis of the baseline is made to argue that it can be considered as a part of the signal. Decaying signal around the peak is probably due to solvated and fragmented analyte molecules. The baseline is formed by the sum of the decaying signals from all the neighboring charge states. A Monte Carlo simulation showed that the histogram of the samples realized from a hypothetical mixture distribution has visible characteristics of a typical signal peak in MS. This proposed paradigm needs to be investigated further to verify the assumptions as it goes against all the current standards in MS analysis. Based on the results of further analysis, new signal detection and estimation schemes can be built based on the new signal model.

The signal peaks are modeled for one molecular species at a time and then detection of those species are carried out in the MS using the proposed detection algorithm. This strategy can be extended to modeling multiple species at once, to detect overlapping peaks in the MS. Overlapping peaks occur when two protein variants are too close to be resolved by the MS. Sometimes, isotopic distributions is also resolved for lighter molecules. New amplitude estimation algorithms can be explored to improve the fractional abundance calculations.

A software package for fractional abundance estimation can be built based on the detection/estimation scheme to process multiple MS data files and analyze the performance further. As it is difficult to establish a ground truth for the real abundances, the estimates can only be compared with other algorithms to see if consistent results are obtained.

Other applications, such as, protein/peptide identification, deisotoping etc. that require MS signal detection are potential problems that can be explored using the statistical signal processing algorithms developed in this dissertation.

REFERENCES

[1] R. W. Nelson, D. Nedelkov, K. A. Tubbs, and U. A. Kiernan, "Quantitative mass spectrometric immunoassay of insulin like growth factor 1," *Journal of Proteome Research*, vol. 3, no. 4, pp. 851–855, 2004.

[2] R. J. Cotter, *Time-of-flight mass spectrometry*. John Wiley and Sons, Ltd, 2004.

[3] *MicrOTOF-Q User Manual, Version 1.0*, Bruker Daltonics, Bremen, Germany, 2006.

[4] *MicrOTOF-Q User Manual, Version 2.1*, Bruker Daltonics, Bremen, Germany, 2006.

[5] D. Boswell, "Introduction to support vector machines," http://www.work.caltech.edu/~boswell/IntroToSVM.pdf, Accessed: Sep. 23rd, 2012.

[6] J. Neyman and E. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933.

[7] T. Kailath and H. V. Poor, "Detection of stochastic processes," *IEEE Transaction on Information Theory*, vol. 44, no. 6, 1998.

[8] G. Turin, "An introduction to matched filters," *IRE Transactions on Information Theory*, vol. 6, no. 3, pp. 311 –329, Jun. 1960.

[9] J. Aldrich, "R. A. Fisher and the making of maximum likelihood 1912-1922," *Statistical Science*, vol. 12, no. 3, pp. 162–176, 1997.

[10] N. S. K. J. Sohn and W. Sung, "A statistical model based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[11] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 1998, pp. 365–368.

[12] J.-H. Chang and N. S. Kim, "Voice activity detection based on complex laplacian model," *Electronics Letters*, vol. 39, no. 7, pp. 632–634, Apr. 2003.

[13] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204 –207, Jul. 2003.

[14] R. Martin, "Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2002*, vol. 1, May 2002, pp. I–253 –I–256.

[15] J. W. Shin, J.-H. Chang, and N. S. Kim, "Statistical modeling of speech signals based on generalized gamma distribution," *IEEE Signal Processing Letters*, vol. 12, no. 3, pp. 258–261, Mar. 2005.

[16] J. W. Shin, J.-H. Chang, S. Barbara, H. S. Yun, and N. S. Kim, "Voice activity detection based on generalized gamma distribution," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1, 2005, pp. 781 – 784.

[17] J.-H. Chang, N. S. Kim, and S. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.

[18] L. Xu and J. Li, "Iterative generalized-likelihood ratio test for mimo radar," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2375 –2385, Jun. 2007.

[19] K. Sangston and K. Gerlach, "Coherent detection of radar targets in a non-gaussian background," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 30, no. 2, pp. 330 –340, Apr. 1994.

[20] A. Farina, A. Russo, and F. Studer, "Coherent radar detection in log-normal clutter," *Communications, Radar and Signal Processing, IEE Proceedings F*, vol. 133, no. 1, pp. 39 –53, Feb. 1986.

[21] A. Farina, A. Russo, F. Scannapieco, and S. Barbarossa, "Theory of radar detection in coherent weibull clutter," *Communications, Radar and Signal Processing, IEE Proceedings F*, vol. 134, no. 2, pp. 174 –190, Apr. 1987.

[22] F. Aluffi Pentini, A. Farina, and F. Zirilli, "Radar detection of targets located in a coherent k distributed clutter background," *Radar and Signal Processing, IEE Proceedings F*, vol. 139, no. 3, pp. 239 –245, Jun. 1992.

[23] F. Nan and R. Nowak, "Generalized likelihood ratio detection for fMRI using complex data," *IEEE Transactions on Medical Imaging*, vol. 18, no. 4, pp. 320 –329, Apr. 1999.

[24] S. Taylor and H. Hartse, "An evaluation of generalized likelihood ratio outlier detection to identification of seismic events in western China," *Bulletin of the Seismological Society of America*, vol. 87, no. 4, 1997.

[25] Y. Yasui, M. Pepe, M. L. Thompson, B. Adam, G. L. Wright, Y. Qu, J. D. Potter, M. Winget, M. Thornquist, and Z. Feng, "A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection," *Biostatistics*, vol. 4, no. 3, pp. 449–463, 2003.

[26] V. P. Andreev, T. Rejtar, H. S. Chen, E. V. Moskovets, A. R. Ivanov, and B. L. Karger, "Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization," *Clinical Chemistry.*, vol. 49, no. 10, pp. 1615–1623, Oct. 2003.

[27] W. E. Wallace, A. J. Kearsley, and C. M. Guttman, "An operator-independent approach to mass spectral peak identification and integration," *Analytical Chemistry*, vol. 76, no. 9, pp. 2446–2452, May 2004.

[28] K. H. Jarman, D. S. Daly, K. K. Anderson, and K. L. Wahl, "A new approach to automated peak detection," *Chemometrics and Intelligent Laboratory Systems*, vol. 69, no. 1-2, pp. 61–76, 2003.

[29] J. A. Carroll and R. C. Beavis, "Using matrix convolution filters to extract information from time-of-flight mass spectra," *Rapid Communications in Mass Spectrometry*, vol. 10, no. 13, pp. 1683–1687, 1996.

[30] K. R. Coombes, S. Tsavachidis, J. S. Morris, K. A. Baggerly, M.-C. Hung, and H. M. Kuerer, "Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform," *Proteomics*, vol. 5, no. 16, pp. 4107–4117, 2005.

[31] W. A. K. Pan Du and S. M. Lin, "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching," *Bioinformatics*, vol. 22, pp. 2059–2065, 2006.

[32] R. Gras, M. Muller, E. Gasteiger, S. Gay, P. A. Binz, W. Bienvenut, C. Hoogland, J. C. Sanchez, A. Bairoch, D. F. Hochstrasser, and R. D. Appel, "Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection," *Electrophoresis*, vol. 20, no. 18, pp. 3535–3550, Dec. 1999.

[33] V. P. Andreev, T. Rejtar, H. S. Chen, E. V. Moskovets, A. R. Ivanov, and B. L. Karger, "A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain," *Analytical Chemistry*, vol. 75, no. 22, pp. 6314–6326, Nov. 2003.

[34] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[35] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Mullers, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.*, Aug. 1999, pp. 41 –48.

[36] J. Shlens, "A tutorial on principal component analysis," http://www.snl.salk.edu/~shlens/pca.pdf, Accessed: Sep. 23rd, 2012.

[37] R. H. Lilien, H. Farid, and B. R. Donald, "Probabilistic Disease Classification of Expression-Dependent Proteomic Data from Mass Spectrometry of Human Serum," *Journal of Computational Biology*, vol. 10, pp. 925–946, 2003.

[38] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data," *Bioinformatics*, vol. 19, no. 13, pp. 1636–1643, 2003.

[39] Y. Qu, B.-l. Adam, M. Thornquist, J. D. Potter, M. L. Thompson, Y. Yasui, J. Davis, P. F. Schellhammer, L. Cazares, M. Clements, J. Wright, George L., and Z. Feng, "Data reduction using a discrete wavelet transform

in discriminant analysis of very high dimensionality data," *Biometrics*, vol. 59, no. 1, pp. 143–151, 2003.

[40] K. A. Baggerly, J. S. Morris, J. Wang, D. Gold, L.-C. Xiao, and K. R. Coombes, "A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples," *Proteomics*, vol. 3, no. 9, pp. 1667–1672, 2003.

[41] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21 –27, 1967.

[42] L. Li, D. M. Umbach, P. Terry, and J. A. Taylor, "Application of the ga/knn method to seldi proteomics data," *Bioinformatics*, vol. 20, no. 10, pp. 1638–1640, 2004.

[43] W. Zhu, X. Wang, Y. Ma, M. Rao, J. Glimm, and J. S. Kovach, "Detection of cancer-specific markers amid massive mass spectral data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 25, pp. 14 666–14 671, 2003.

[44] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," *Genome Inform*, vol. 13, pp. 51–60, 2002.

[45] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[46] A. J. Rai, Z. Zhang, J. Rosenzweig, I. e. M. Shih, T. Pham, E. T. Fung, L. J. Sokoll, and D. W. Chan, "Proteomic approaches to tumor marker discovery," *Arch. Pathol. Lab. Med.*, vol. 126, no. 12, pp. 1518–1526, Dec. 2002.

[47] W. Clarke, B. C. Silverman, Z. Zhang, D. W. Chan, A. S. Klein, and E. P. Molmenti, "Characterization of renal allograft rejection by urinary proteomic analysis," *Annals of Surgery*, vol. 237, no. 5, pp. 660–664, May 2003.

[48] L. L. Banez, P. Prasanna, L. Sun, A. Ali, Z. Zou, B. L. Adam, D. G. McLeod, J. W. Moul, and S. Srivastava, "Diagnostic potential of serum proteomic patterns in prostate cancer," *Journal of Urology*, vol. 170, no. 2 Pt 1, pp. 442–446, Aug. 2003.

[49] P. Neville, P. Y. Tan, G. Mann, and R. Wolfinger, "Generalizable mass spectrometry mining used to identify disease state biomarkers from blood serum," *Proteomics*, vol. 3, no. 9, pp. 1710–1715, Sep. 2003.

[50] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. Springer, 1999.

[51] K. Jong, E. Marchiori, M. Sebag, and A. van der Vaart, "Feature selection in proteomic pattern data with support vector machines," in *Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB '04. Proceedings of the 2004 IEEE Symposium on*, Oct. 2004, pp. 41 – 48.

[52] C. R. Borges, P. E. Oran, S. Buddi, J. W. Jarvis, M. R. Schaab, D. S. Rehder, S. P. Rogers, T. Taylor, and R. W. Nelson, "Building multidimensional biomarker views of type 2 diabetes on the basis of protein microheterogeneity," *Clinical Chemistry*, vol. 57, no. 5, pp. 719–728, May 2011.

[53] S. Buddi, T. Taylor, C. Borges, and R. Nelson, "SVM Multi-classification of T2D/CVD patients using biomarker features," in *10th International Conference on Machine Learning and Applications and Workshops (ICMLA), 2011*, vol. 2, Dec. 2011, pp. 338 –341.

[54] M. Wagner, D. Naik, and A. Pothen, "Protocols for disease classification from mass spectrometry data," *Proteomics*, vol. 3, no. 9, pp. 1692–1698, Sep. 2003.

[55] A. Ipsen and T. Ebbels, "Prospects for a statistical theory of lc/tofms data," *Journal of The American Society for Mass Spectrometry*, vol. 23, pp. 779–791, 2012.

[56] M. L. Vestal, "Modern MALDI time-of-flight mass spectrometry," *Journal of Mass Spectrometry*, vol. 44, no. 3, pp. 303–317, 2009.

[57] R. S. Yalow and S. A. Berson, "Immunoassay of endogenous plasma insulin in man," *Journal of Clinical Investigation*, vol. 39, pp. 1157–1175, Jul. 1960.

[58] E. Engvall, P. Perlman, and A. Lindvall, "Enzyme-linked immunosorbent assay (ELISA). Quantitative assay of immunoglobulin G," *Immunochemistry*, vol. 8, no. 9, pp. 871–874, 1971.

[59] V. Stejskal, K. Cederbrant, A. Lindvall, and M. Forsbeck, "MELISA - an in vitro tool for the study of metal allergy," *Toxicology in Vitro*, vol. 8, no. 5, pp. 991–1000, 1994.

[60] P. Nikitin, P. Vetoshko, and T. Ksenevich, "Magnetic immunoassays," *Sensor Letters*, vol. 5, no. 1, pp. 296–299, 2007.

[61] R. W. Nelson, J. R. Krone, A. L. Bieber, and P. Williams, "Mass spectrometric immunoassay," *Analytical Chemistry*, vol. 67, no. 7, pp. 1153–1158, Apr. 1995.

[62] W. C. Wiley and I. H. McLaren, "Time-of-flight mass spectrometer with improved resolution," *The Review of Scientific Instruments*, vol. 26, pp. 1150–1157, 1955.

[63] J. Futrell, T. Tiernan, F. Abramson, and C. Miller, "Modification of a time-of-flight mass spectrometer for investigation of ion-molecule reactions at elevated pressures," *The Review of Scientific Instruments*, vol. 39, pp. 340–345, 1969.

[64] D. F. Torgerson, R. P. Skowronski, and R. D. Macfarlane, "New approach to the mass spectroscopy of non-volatile compounds," *Biochemical and Biophysical Research Communications*, vol. 60, no. 2, pp. 616–621, Sep. 1974.

[65] R. B. V. Breemen, M. Snow, and R. J. Cotter, "Time-resolved laser desorption mass spectrometry. I. Desorption of preformed ions," *International Journal of Mass Spectrometry and Ion Physics*, vol. 49, no. 1, pp. 35–50, 1983.

[66] M. Karas and F. Hillenkamp, "Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons," *Analytical Chemistry*, vol. 60, no. 20, pp. 2299–2301, Oct. 1988.

[67] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse, "Electrospray ionization for mass spectrometry of large biomolecules," *Science*, vol. 246, no. 4926, pp. 64–71, Oct. 1989.

[68] A. N. Verentchikov, W. Ens, and K. G. Standing, "Reflecting time-of-flight mass spectrometer with an electrospray ion source and orthogonal extraction," *Analytical Chemistry*, vol. 66, no. 1, pp. 126–133, Jan. 1994.

[69] J. L. Wiza, "Microchannel plate detectors," *Nuclear Instruments and Methods*, vol. 162, pp. 587–601, 1979.

[70] O. Rather, "High resolution detection for time-of-flight mass spectrometers," U.S. Patent US 6 870 156, Mar. 22, 2005.

[71] V. Raznikov and M. Raznikova, "Deconvolution of overlapping mass spectral peaks following ion-counting data acquisition," *International Journal of Mass Spectrometry and Ion Processes*, vol. 63, no. 2-3, pp. 157–186, 1985.

[72] D. I. Malyarenko, W. E. Cooke, E. R. Tracy, M. W. Trosset, O. J. Semmes, M. Sasinowski, and D. M. Manos, "Deconvolution filters to enhance resolution of dense time-of-flight survey spectra in the time-lag optimization range," *Rapid Communications in Mass Spectrometry*, vol. 20, no. 11, pp. 1661–1669, 2006.

[73] O. N. Peregudov and O. M. Buhay, "The peak shape model for magnetic sector and time-of-flight mass spectrometers," *International Journal of Mass Spectrometry*, vol. 295, no. 1-2, pp. 1–6, 2010.

[74] E. F. Strittmatter, N. Rodriguez, and R. D. Smith, "High mass measurement accuracy determination for proteomics using multivariate regression fitting: Application to electrospray ionization time-of-flight mass spectrometry," *Analytical Chemistry*, vol. 75, no. 3, pp. 460–468, 2003.

[75] M. Kempka, J. Sjödahl, A. Björk, and J. Roeraade, "Improved method for peak picking in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 18, no. 11, pp. 1208–1212, 2004.

[76] R. A. Zubarev, P. Håkansson, and B. Sundqvist, "Accurate monoisotopic mass measurements of peptides: Possibilities and limitations of high resolution time-of-flight particle desorption mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 10, no. 11, pp. 1386–1392, 1996.

[77] J. M. K. Hyunjin Shin, Miray Mutlu and M. K. Markey, "Parametric power spectral density analysis of noise from instrumentation in maldi tof mass spectrometry," *Cancer Informatics*, vol. 3, pp. 219–230, 2007.

[78] A. N. Krutchinsky and B. T. Chait, "On the nature of the chemical noise in MALDI mass spectra," *Journal of the American Society for Mass Spectrometry*, vol. 13, no. 2, pp. 129–134, Feb. 2002.

[79] N. B. Cech and C. G. Enke, "Practical implications of some recent studies in electrospray ionization fundamentals," *Mass Spectrometry Reviews*, vol. 20, no. 6, pp. 362–387, 2001.

[80] W. Wang, H. Zhou, H. Lin, S. Roy, T. A. Shaler, L. R. Hill, S. Norton, P. Kumar, M. Anderle, and C. H. Becker, "Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards," *Analytical Chemistry*, vol. 75, no. 18, pp. 4818–4826, 2003.

[81] D. M. Horn, R. A. Zubarev, and F. W. McLafferty, "Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules," *Journal of the American Society for Mass Spectrometry*, vol. 11, no. 4, pp. 320–332, 2000.

[82] C. Hundertmark, R. Fischer, T. Reinl, S. May, F. Klawonn, and L. J*ä*nsch, "MS-specific noise model reveals the potential of iTRAQ in quantitative proteomics," *Bioinformatics*, vol. 25, no. 8, pp. 1004–1011, 2009.

[83] P. Kaur and P. B. O'Connor, "Use of statistical methods for estimation of total number of charges in a mass spectrometry experiment," *Analytical Chemistry*, vol. 76, no. 10, pp. 2756–2762, 2004.

[84] G. A. Pearson, "A general baseline-recognition and baseline-flattening algorithm," *Journal of Magnetic Resonance (1969)*, vol. 27, no. 2, pp. 265–272, 1977.

[85] B. Williams, S. Cornett, B. Dawant, A. Crecelius, B. Bodenheimer, and R. Caprioli, "An algorithm for baseline correction of maldi mass spectra," in *Proceedings of the 43rd annual Southeast regional conference - Volume 1*, ser. ACM-SE 43. New York, NY, USA: ACM, 2005, pp. 137–142.

[86] D. Chang, C. D. Banack, and S. L. Shah, "Robust baseline correction algorithm for signal dense nmr spectra," *Journal of Magnetic Resonance*, vol. 187, no. 2, pp. 288–292, 2007.

[87] L. Andrade and E. S. Manolakos, "Signal background estimation and baseline correction algorithms for accurate DNA sequencing," *The Journal of VLSI Signal Processing*, vol. 35, pp. 229–243, 2003.

[88] D. I. Malyarenko, W. E. Cooke, B.-L. Adam, G. Malik, H. Chen, E. R. Tracy, M. W. Trosset, M. Sasinowski, O. J. Semmes, and D. M. Manos, "Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques," *Clinical Chemistry*, vol. 51, pp. 65–74, 2005.

[89] C. S. Tan, A. Ploner, A. Quandt, J. Lehti*ö*, and Y. Pawitan, "Finding regions of significance in SELDI measurements for identifying protein biomarkers," *Bioinformatics*, vol. 22, pp. 1515–1523, 2006.

[90] A. Antoniadis, J. Bigot, S. Lambert-Lacroix, and F. Letue, "Nonparametric pre-processing methods and inference tools for analyzing time-of-flight mass spectrometry data," *Current Analytical Chemistry*, vol. 3, pp. 127–147, 2007.

[91] X.-G. Shao, A. K.-M. Leung, and F.-T. Chau, "Wavelet: A new trend in chemistry," *Accounts of Chemical Research*, vol. 36, no. 4, pp. 276–283, 2003.

[92] P. Du, G. Stolovitzky, P. Horvatovich, R. Bischoff, J. Lim, and F. Suits, "A noise model for mass spectrometry based proteomics," *Bioinformatics*, vol. 24, no. 8, pp. 1070–1077, 2008.

[93] C. R. Borges, J. W. Jarvis, P. E. Oran, S. P. Rogers, and R. W. Nelson, "Population studies of intact vitamin D binding protein by affinity capture ESI-TOF-MS," *Journal of Biomolecular Techniques*, vol. 19, no. 3, pp. 167–176, Jul. 2008.

[94] W. C. Wiley and I. H. McLaren, "Time-of-flight mass spectrometer with improved resolution," *Review of Scientific Instruments*, vol. 26, no. 9, pp. 1150–1157, 1955.

[95] M. Vestal and P. Juhasz, "Resolution and mass accuracy in matrix-assisted laser desorption ionization - time-of-flight," *Journal of the American Society for Mass Spectrometry*, vol. 9, no. 9, pp. 892–911, 1998.

[96] M. Vestal and K. Hayden, "High performance MALDI-TOF mass spectrometry for proteomics," *International Journal of Mass Spectrometry*, vol. 268, pp. 83–92, 2007.

[97] D. Valkenborg, I. Mertens, F. Lemiére, E. Witters, and T. Burzykowski, "The isotopic distribution conundrum," *Mass Spectrometry Reviews*, vol. 31, no. 1, pp. 96–109, 2012.

[98] R. Robinson, C. Warner, and R. Gohlke, "Calculation of relative abundance of isotope clusters in mass spectrometry," *Journal of Chemical Education*, vol. 47, pp. 467–468, 1970.

[99] H. Kubinyi, "Calculation of isotope distributions in mass spectrometry. a trivial solution for a non-trivial problem," *Analytica Chimica Acta*, vol. 247, no. 1, pp. 107–119, 1991.

[100] M. L. Brownawell and J. San Filippo, "Simulation of chemical instrumentation. II: A program for the synthesis of mass spectral isotopic abundances," *Journal of Chemical Education*, vol. 59, no. 8, p. 663, 1982.

[101] J. A. Yergey, "A general approach to calculating isotopic distributions for mass spectrometry," *International Journal of Mass Spectrometry and Ion Physics*, vol. 52, no. 23, pp. 337–349, 1983.

[102] A. L. Rockwood, S. L. Van Orden, and R. D. Smith, "Rapid Calculation of Isotope Distributions," *Analytical Chemistry*, vol. 67, no. 15, pp. 2699–2704, 1995.

[103] A. L. Rockwood and S. L. Van Orden, "Ultrahigh-Speed Calculation of Isotope Distributions," *Analytical Chemistry*, vol. 68, no. 13, pp. 2027–2030, 1996.

[104] A. L. Rockwood, M. M. Kushnir, and G. J. Nelson, "Dissociation of individual isotopic peaks: predicting isotopic distributions of product ions

in MS," *Journal of the American Society for Mass Spectrometry*, vol. 14, no. 4, pp. 311–322, 2003.

[105] A. L. Rockwood and P. Haimi, "Efficient calculation of accurate masses of isotopic peaks," *Journal of the American Society for Mass Spectrometry*, vol. 17, no. 3, pp. 415–419, 2006.

[106] D. Lide, Ed., *CRC Handbook of Chemistry and Physics*, 71st ed. Boca Raton, Fl: CRC Press, 1990.

[107] M. Mann, J. Fenn, and S. Wong, *Where do all the charges go in electrospray ionization?* John Wiley and Sons, Ltd, 1990, pp. 139–144.

[108] V. Raznikov, A. Dodonov, and E. Lanin, "Data acquisition and processing in high-resolution mass spectrometry using ion counting," *International Journal of Mass Spectrometry and Ion Physics*, vol. 25, no. 3, pp. 295–313, 1977.

[109] S. A. Hofstadler, J. J. Drader, and A. Schink, "Selective ion filtering by digital thresholding: A method to unwind complex esi-mass spectra and eliminate signals from low molecular weight chemical noise," *Analytical Chemistry*, vol. 78, no. 2, pp. 372–378, 2006.

[110] J. Massey, Frank J., "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[111] G. Marsaglia, W. Tsang, and J. Wang, "Evaluating kolmogorov's distribution," *Journal of Statistical Software*, vol. 8, no. 18, pp. 1–4, 2003.

[112] T. W. Anderson, "On the distribution of the two-sample cramer-von mises criterion," *Annals of Mathematics Statistics*, vol. 33, no. 3, pp. 1148–1159, 1962.

[113] S. C. Choi and R. Wette, "Maximum likelihood estimation of the parameters of the gamma distribution and their bias," *Technometrics*, vol. 11, no. 4, pp. pp. 683–690, 1969.

[114] M. Hilario, A. Kalousis, C. Pellegrini, and M. Müller, "Processing and classification of protein mass spectra." *Mass spectrometry reviews*, vol. 25, no. 3, pp. 409–449, 2006.

[115] F. Fritsch and R. Carlson, "Monotone piecewise cubic interpolation," *SIAM Journal on Numerical Analysis*, vol. 17, no. 2, pp. 238–246, 1980.

[116] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, New Jersey: Prentice Hall, 1998, vol. 1.

[117] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Golden Section Search in One Dimension*, 3rd ed. New York: Cambridge University Press, 2007, ch. 10.

[118] M. Mann, C. K. Meng, and J. B. Fenn, "Interpreting mass spectra of multiply charged ions," *Analytical Chemistry*, vol. 61, no. 15, pp. 1702–1708, 1989.

[119] A. G. Ferrige, M. J. Seddon, B. N. Green, S. A. Jarvis, J. Skilling, and J. Staunton, "Disentangling electrospray spectra with maximum entropy," *Rapid Communications in Mass Spectrometry*, vol. 6, no. 11, pp. 707–711, 1992.

[120] Z. Zhang and A. G. Marshall, "A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra," *Journal of the American Society for Mass Spectrometry*, vol. 9, no. 3, pp. 225–233, 1998.

[121] J. M. Mikko Katajamaa and M. Oresič, "Mzmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data," *Bioinformatics*, vol. 22, no. 5, pp. 634–636, 2006.

[122] K. C. Leptos, D. A. Sarracino, J. D. Jaffe, B. Krastins, and G. M. Church, "Mapquant: Open-source software for large-scale protein quantification," *Proteomics*, vol. 6, no. 6, pp. 1770–1782, 2006.

[123] O. Kohlbacher, K. Reinert, C. Gröpl, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, and M. Sturm, "Topp-the OpenMS proteomics pipeline," *Bioinformatics*, vol. 23, no. 2, pp. e191–e197, 2007.

[124] A. Lommen, "Metalign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing," *Analytical Chemistry*, vol. 81, no. 8, pp. 3079–3086, 2009.

[125] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak, "XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification," *Analytical Chemistry*, vol. 78, no. 3, pp. 779–787, 2006.

[126] "ESI Deconvolution, MicrOTOF-Q," Bruker Daltonics, Bremen, Germany, 1993, version 4.0.

[127] C. J. Rowlands and S. R. Elliott, "Denoising of spectra with no user input: a spline-smoothing algorithm," *Journal of Raman Spectroscopy*, vol. 42, no. 3, pp. 370–376, 2011.

[128] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures." *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[129] J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly, and R. Kobayashi, "Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum," *Bioinformatics*, vol. 21, no. 9, pp. 1764–1775, May 2005.

[130] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, pp. 1200–1224, 1995.

[131] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, 2nd ed. Wellesley College, 1996.

[132] J. H. Mathews, "Module for the trapezoidal rule for numerical integration," http://math.fullerton.edu/mathews/n2003/TrapezoidalRuleMod.html, Accessed: Oct. 23rd, 2012.

[133] "Compassxport," Bruker Daltonics, Bremen, Germany, 2010, version 3.0.5.

[134] P. G. A. Pedrioli, J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, J. Randall K Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu, and R. Aebersold, "A common open representation of mass spectrometry data and its application to proteomics research," *Nature Biotechnology*, vol. 22, pp. 1459–1466, 2004.

[135] *Diagnosis and Classification of Diabetes Mellitus*, American Diabetes Association, Jan. 2010.

[136] *FDA Guidance for Industry: Diabetes Mellitus - Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes*, U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), 2008.

[137] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data," *Journal of the American Statistical Association*, vol. 99, pp. 67–81, 2004.

[138] Y. Liu and Y. Zheng, "One-against-all multi-class svm classification using reliability measures," in *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on*, vol. 2, 2005, pp. 849 – 854.

[139] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415 –425, 2002.

[140] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, pp. 237–264, 1953.

[141] W. A. Gale and G. Sampson, "Good turing frequency estimation without tears*," *Journal of Quantitative Linguistics*, vol. 2, no. 3, pp. 217–237, 1995.