Dissertations, Theses, and Masters Projects   Theses, Dissertations, & Master Projects

2006

Automated peak identification for time -of -flight mass spectroscopy

Haijian Chen
College of William & Mary - Arts & Sciences

Follow this and additional works at: https://scholarworks.wm.edu/etd

Part of the Analytical Chemistry Commons, and the Physics Commons

Recommended Citation

AUTOMATED PEAK IDENTIFICATION

FOR TIME-OF-FLIGHT MASS SPECTROSCOPY

_____

A Dissertation

Presented to

The Faculty of the Department of Physics

The College of William and Mary in Virginia

In Partial Fulfillment

Of the Requirements for the Degree of

Doctor of Philosophy
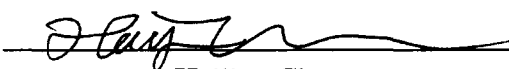
_____

by

Haijian Chen

2005

APPROVAL SHEET

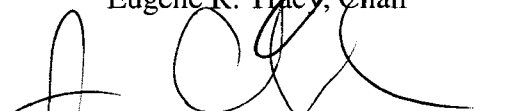This dissertation is submitted in partial fulfillment of

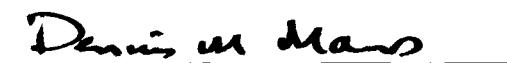The requirements for the degree of

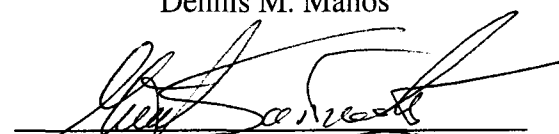Doctor of Philosophy

Haijian Chen
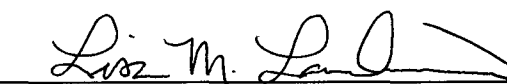
Approved by the Committee, December 2005

Eugene R. Tracy, Chair

Jan Chaloupka

William E. Cooke

Dennis M. Manos

Maciek Sasinowski

Lisa M. Landino, Chemistry

ii

To my wife, Mom and Dad

TABLE OF CONTENTS

iv

ACKNOWLEDGMENTS

I would first like to express my sincere gratitude to my advisor, Dr. Eugene R. Tracy, for his great effort he devoted to this project, for his invaluable guidance, endless support and for his patient. This work would not be possible if otherwise.

Special gratitude goes to Dr. William Cooke for his helpful insights, brilliant suggestions and excellent questions, which keep me moving forward. I would like to extend my gratitude to Dr. Maciek Sasinowski for bringing me into this project and for his continuous encouragements. I am also grateful to Dr. Dennis M. Manos and Dr. Michael W. Trosset for their helpful comments and scientific insights. I am also thankful to committee member Dr. Jan Chaloupka and Dr. Lisa M. Landino for their time and effort in reviewing this thesis.

I would also like to thank collaborators Dr. Dariya I. Malyarenko and research specialist Amy Wilkerson who taught me how to use TOF-SIMS and offer me their support.

My deepest possible gratitude goes to my parents. It is their hard working to create every opportunity for me and their encouragement since childhood that make me possible to go this far. No word would be able to express my thankfulness to them.

My loving thanks go to my wonderful wife, Mingyao Zhu, whom I am fortunate to share life with. Thank you for your care, your support and your patient.

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The high throughput capabilities of protein mass fingerprints measurements have made mass spectrometry one of the standard tools for proteomic research, such as biomarker discovery. However, the analysis of large raw data sets produced by the time-of-flight (TOF) spectrometers creates a bottleneck in the discovery process. One specific challenge is the preprocessing and identification of mass peaks corresponding to important biological molecules. The accuracy of mass assignment is another limitation when comparing mass fingerprints with databases.

We have developed an automated peak picking algorithm based on a maximum likelihood approach that effectively and efficiently detects peaks in a time-of-flight secondary ion mass spectrum. This approach produces maximum likelihood estimates of peak positions and amplitudes, and simultaneously develops estimates of the uncertainties in each of these quantities. We demonstrate that a Poisson process is involved for time-of-flight secondary ion mass spectrometry (TOF-SIMS) and the algorithm takes the character of the Poisson noise into account.

Though this peak picking algorithm was initially developed for TOF-SIMS spectra, it can be extended to other types of TOF spectra as soon as the correct noise characteristics are considered.

We have developed a peak alignment procedure that aligns peaks in different spectra. This is a crucial step for multivariate analysis. Multivariate analysis is often used to distill useful information from complex spectra.

We have designed a TOF-SIMS experiment that consists of various mixtures of three bio-molecules as a model for more complicated biomarker discovery. The peak picking algorithm is applied to the collected spectra. The algorithm detects peaks in the spectra repeatably and accurately. We also show that there are patterns in the spectra of pure bio-molecules samples. Furthermore, we show it is possible to infer the concentration ratios in the mixture samples by checking the strength of the patterns.

xi

Automated Peak Identification

for Time-Of-Flight Mass Spectroscopy

Chapter 1

Introduction

Though great success has been made in genome sequencing, it has been increasingly recognized that the genome, by itself, is not sufficient to understand the behavior and functions of cells, tissues, and biological systems. A current research focus in molecular biology is to test the hypothesis that proteins, instead of DNA, give more complete information related to cell function. Hence, proteins, the final product of genes, are receiving increased attention in biomedicine and a new field, proteomics, which focuses on protein characterization, protein identification and protein function, has emerged.

Although two-dimensional gel electrophoresis and amino acid sequencing retain their important roles in biochemical analysis, recent developments in Mass Spectrometry (MS) have now made it an additional analytical tool in proteome research [1, 2]. Mass spectrometry can give accurate mass "fingerprints" which, in conjunction with protein database searching, can rapidly provide information about protein identification, protein function and protein post-translational modification (*i.e.*, modifications after the polypeptide is synthesized).

2

A mass spectrometer is an apparatus that differentiates different molecular/atomic species in the sample under investigation according to their mass-to-charge ratio. It also can give information about the abundance of each species in the sample. We will discuss more about mass spectrometers in the next two chapters. In protein identification, matrix assisted laser desorption ionization mass spectrometry (MALDI-MS) and electrospray ionization mass spectrometry (ESI-MS) are often used because they can ionize large biological molecules 'softly' without breaking most of them into smaller pieces. To identify proteins, proteins are often digested by a protease into peptides and fingerprints of the resulting peptides are measured by MALDI or ESI. The mass fingerprints of peptides are then compared with a database (for example, Swiss-Prot) using programs such as Sequest, searching for sequences in the database that have the same masses as the experimental masses.

The high throughput, high sensitivity and quantitative analysis of mass spectrometry make it possible to analyze hundreds of analytes over a large mass range simultaneously even the sample volume is small. If a biological fluid, such as blood, is measured, a protein "profile" may be developed. This leads to the potential for finding biomarkers that are overexpressed or underexpressed or modified. Such biomarkers can then be used to differentiate pathological states (disease) from normal states or to assess and guide drug treatments. If desired, the discovered biomarker can be chemically extracted for further analysis. Progress has been made with this line of cancer research as summarized in [3, 4, 5, 6, 7, 8].

All information that mass spectrometry provides is encoded by peaks that occur at different masses with various amplitudes in the spectrum. The number of peaks present varies with the sample under investigation and the type of mass spectrometer used. If biological or organic samples are investigated, the peak number can easily rise to a few hundred. In blood serum, it is estimated that there may be up to 10,000 proteins with concentrations ranging over at least 9 orders of magnitude [9, 10]. However, the dynamic range of MS instruments is only 3~4 orders of magnitude [11], thus careful biochemical sample preparation is critical. Very often, as in the biomarker discovery for disease detection, it is not clear beforehand which peak is important. Thus all peaks potentially have to be taken into account. This is also true for protein identification.

As more effort has been devoted to improving the performance of MS instruments to provide more detailed information about the sample, to increase resolving power, and to lower the detection limit, the resulting mass spectra have inevitably become more complicated. When we use a Time-of-Flight Secondary Ion Mass Spectrometry (TOF-SIMS) apparatus to analyze biological samples, we produce spectra that contain over a million data points and contain a few hundred peaks. Figure 1.1(a) shows a low mass portion of a typical spectrum. One can see that peaks span the whole region at almost every mass unit. Even a small portion that looks negligible (inside the square), actually has peaks present (Fig.1.1 (b)). Compounding the problem of dense data sets, roboticized sample preparation and computerized data collection allow researchers to generate dozens, or hundreds, of such spectra in a few hours.

The analysis of such large raw data sets produced by survey mass spectrometers creates a bottleneck in the research process. To overcome this bottleneck, the first step is to simplify a spectrum that contains thousands, even millions, of data points down to only the essential information about peaks, positions and amplitudes. In this way, a spectrum can be reduced to only a few hundreds points that represent peak positions, amplitudes and uncertainties in the peak positions and amplitudes.

We should emphasize that not only mass spectrometry faces peak detection problem. In fact, peak detection is a quite general problem in many analytical instruments. A good automated peak detection procedure should run rapidly, and give repeatable and accurate results. It should find all significant peaks in a spectrum but not report false peaks. For some biological samples, the concentration is very small and consequently the spectrum has a low signal-to-noise ratio, hence finding peaks is difficult. Missing peaks in a spectrum, and reporting false peaks, can both potentially lead to discovery of false "biomarkers." This could lead to wasted further investment, which could potentially be costly and time consuming.

A good peak detector should give accurate peak position assignments and peak amplitude estimations. The importance of accuracy in peak position is obvious, especially when we want to compare the mass of a peak with a database such as Swiss-Prot. Accurate peak amplitude estimations are also important when quantitative analysis is required. For example, when looking for proteins that are associated with disease, it is very possible the proteins we are looking for are common in both healthy and sick people but are overexpressed or underexpressed. In this case, it is not a "yes

or no" problem, but rather a "more or less" problem, and the correct peak amplitude estimation is essential.

Another important factor is that the full peak detection procedure should be automated as much as possible for high-throughput data handling. An additional advantage of an automated peak picking algorithm is that it minimizes human interaction and thus eliminates potential bias introduced by investigators. It has been reported that in an inter-laboratory investigation conducted by the NIH, the same samples were analyzed by MALDI in different laboratories. It turned out that when comparing experimental results from different laboratories, the reduced data showed more differences than the raw data. The differences in the reduced data were traced back to detail decisions that investigators made when the data were reduced [12, 13]. This result highlights the need for adoption of common, well tested and well understood, automated methods to avoid *ad hoc* methods developed in each research group.

During the past few years, various algorithms have been developed for peak detection. Many of them detect peaks according to an assigned signal-to-noise ratio or other similar thresholds. This is similar to a one sided statistical test of whether the signal deviates from background noise. Others detect peaks by comparing the spectrum with a predefined peak lineshape. Sometimes, a spectrum is smoothed before peak detection. Below we give some typical peak detection procedures that have been reported.

Yasui *et al.* identify peaks by first finding points whose intensities are the highest among ±N points before and after them [14]. After trying different N value, they choose N to be 20. To avoid picking false peaks that are due to random noise, a requirement that the intensity must be higher than the average intensity over a broad neighborhood is applied.

Bryant *et al.* use cross-correlation to help peak detection [15]. First, a threshold is applied on intensity, the segments that are above the threshold form data blocks. Only blocks of length above a limit that is associated with peak width are qualified for further analysis. Shorter blocks are considered due to noise that occasionally runs above the intensity threshold. If the length of the qualified block is comparable to peak width, it is considered there is only one peak in the data block, and the total intensity is calculated and the centroid of the data is estimated. If the length of the qualified data block is larger, then there may be more than one peak in the data block, and cross-correlation is calculated using a pre-defined peak lineshape. If the cross-correlation satisfies certain conditions, peaks are "detected".

Gras *et al.* find peaks in a spectrum by comparing a segment of the spectrum with a template which describes the peak shape and isotopic pattern [16]. An error function, which is a modified Euclidian distance between the spectrum and the template, is defined to estimate the mismatch. Peaks are detected when the error function is smaller than a threshold error value and the peak heights are larger than a required value.

A similar approach is proposed by Sokolov *et al.* where peaks are detected based on "minimum average risk" criteria originally proposed by Kolmogorov [17]. Before discussing their method, let us first introduce some notations and we will use these notations throughout the thesis:

We are going to use $p(X)$ to denote the probability of some event X; use $p(X|I)$ to denote the conditional probability of X given relevant background information I at hand. We will, when talking about probability, always condition the statements on the background information, as the 'absolute probability' is not well posed. We will use $p(X,Y|I)$ to denote the joint probability of X and Y, conditioned on relevant background information I at hand.

To detect peaks using "minimum average risk" method, let $p(X_0|Y)$ be the probability that a peak is located at X_0 for a given data Y. If the investigator decides the peak is located at X, let $C(X_0,X)$ be the loss function, which represents the penalty for making an erroneous decision where the true peak position is X_0. Then the conventional risk function is defined as:

$$R(X|Y) = \int C(X_0,X)p(X_0|Y)dX_0 \tag{1.1}$$

The average risk function is defined as:

$$R(X) = \int R(X|Y)p(Y)dY \tag{1.2}$$

The loss function is often defined as:

$$C(X_0,X) = \begin{cases} 0 & X_0 = X \\ 1 & otherwise \end{cases} \tag{1.3}$$

then, equation (1.1) becomes:

$$R(X \mid Y) = 1 - p(X_0 \mid Y)\delta(X_0 - X) \qquad (1.4)$$

Now, note that according to Bayes' theorem, which we will introduce in the next chapter, we have:

$$p(X_0 \mid Y) \propto p(Y \mid X_0)p(X_0) \qquad (1.5)$$

In (1.5), $p(Y \mid X_0)$ is called the likelihood function, and $p(X_0)$ is the prior. As we are going to discuss likelihood function and prior in more details in the next chapter when we introduce Bayes' theorem, here we just state that $p(Y \mid X_0)$ and $p(X_0)$ are two probabilities. If $p(X_0)$ is uniformly distributed, then the maximum of $p(Y \mid X_0)$ would minimize (1.1), this then becomes a maximum likelihood method.

In order to make the above method work, an important assumption is made. It states that since there are many possible noise sources, then by the central limit theorem the resulting noise may be approximated by white Gaussian noise. The likelihood function can be written explicitly:

$$p(Y \mid X_0) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^M e^{\left(-\frac{\sum_{i=1}^{M}(Y(t_i)-S(t_i,X_0))^2}{2\sigma^2}\right)} \qquad (1.6)$$

where, $S(t_i, X_0)$ is the ideal signal without any noise. Equation (1.6) can be maximized by setting $\partial P(Y \mid X_0)/\partial X_0 = 0$.

While the above methods require knowledge about the peak lineshape, Wallace et $al.$ have developed an algorithm that makes no assumption about peak shape and does

not need to smooth the data before peak detection [12]. The essential idea of the algorithm is based on segmentation that is illustrated in Fig1.2. The segmentation begins with the first point (A) and the last point (B) of the spectrum. Then, the point (C) in the spectrum which has the largest orthogonal distance to line AB is found. This breaks one segment into two. Next, in the portion between A and C, find the point (D) that has the largest orthogonal distance to AC, which is the beginning point of the peak, and similarly point E, the ending point of the peak. Then, the heights of C, D and E are adjusted to best fit the data.

Fig1.2 is the simplest case. In general there will be more than one peak and the above process needs to iterate many times until the largest orthogonal distance is less than a predefined threshold. Thus, at the end of the procedure, each peak will be represented by three points, beginning point, center point and ending point. To avoid false detection, peaks less than threshold height will be discarded.

Another peak detection algorithm that does not depend on the peak shape is due to Jarman et al.[18]. In their approach, a spectrum is viewed as a histogram. In regions where there is no peak, the spectrum is relatively flat and the intensity varies around a constant. Hence, it can be viewed as a histogram for a noisy uniform distribution. In the region where a peak is present, the intensity profile is considered as a histogram of a more centralized distribution. An intensity weighted variance (IWV) is defined as:

$$IWV = \frac{\sum_{j<N} I_j (x_j - \bar{x})^2}{\sum_{j<N} I_j} \tag{1.7}$$

With the assumption that I_j is white and independent, identically distributed (i.i.d) with mean μ and variance σ, one may test the null hypothesis:

$$H_0: \quad IWV / s_U^2 = 1$$

Where s_U^2 is the variance of a discrete uniform distribution. With the above assumption on I_j, a critical value for rejection H_0 can be computed theoretically. A peak is considered present in the window if the null hypothesis is rejected.

Efforts have also been made to smooth the spectrum to increase the signal-to-noise ratio before attempting peak detection. For example, Morris *et al.* developed an algorithm to detect peaks based on the mean spectrum of the spectra from an ensemble of similar samples [19]. By averaging spectra of similar samples, the noise is reduced. The mean spectrum is then further smoothed by wavelet denoising. Local maxima in the smoothed spectrum are labeled as candidate peaks. They then search regions around the candidate peaks in the spectrum of each sample which has also been wavelet denoised and detrend to identify peaks in the individual spectrum.

Andreev *et al.* smooth the spectrum by first characterizing the noise in the frequency domain [20]. The spectra they analyzed are from MALDI that is coupled to liquid chromatography. They analyzed the spectra without chromatographic peaks and find that the noise characteristics depend on *m/z*. They then characterize the frequency property of the noise for different mass regions. This allows them to build separate filters for each mass region and to filter out noise from the spectrum.

Though all of the above peak picking procedures work to some degree in terms of finding peaks in the spectrum, none of them addresses the confidence level in peak position and amplitude assignments, except that minimum average risk method could. Because of the noise in the spectrum, there are always uncertainties associated with the estimates made. These uncertainties represent the confidence level about the peak positions and amplitude assignments and occur in addition to the instrument precision. A peak picking procedure that leads to low peak position confidence level would degrade the instrument performance that researchers spent vast amount of money and effort to improve.

When a peak is detected in a spectrum and compared with a database, it is very rare to find an exact match. It is almost certain that a search will return a list of possible chemical IDs around the detected peak. Knowing the position uncertainty will help us to determine how many possible chemical IDs we should seriously consider. The uncertainty would also help to determine whether peaks that appear at slightly different positions in different spectra are in fact the same.

In this thesis, a new peak picking algorithm that includes a physically valid noise model will be presented. The algorithm utilizes Bayes' theorem to test if there is a peak in a window and, if a conclusion of positive peak presence is arrived, peak positions and amplitudes are estimated via a maximum likelihood method. We will introduce the general idea about this peak picking procedure and then specifically focus on finding peaks in a time-of-flight secondary ion mass spectrum. We will first discuss how a Time-of-Flight Mass Spectrometry (TOF-MS) instrument works and

discuss why a Poisson process is involved in TOF-SIMS. Next, the necessary formulas for peak detection will be derived in such a way that Poisson (counting) noise is included. This approach will produce maximum likelihood estimates of peak positions and amplitudes, and simultaneously develop estimates of the uncertainties in each of these quantities.

Having developed the peak detection algorithm, we will apply it to spectra that we collected from a carefully designed experiment to test our ability to measure relative abundance of biomolecules in a sample. In the experiment, solutions containing three peptides at various concentration ratios were deposited onto etched silver and spectra were acquired by a TOF-SIMS instrument. After using our peak picking algorithm, multivariate analysis is applied to estimate the mixture ratios. We will demonstrate that it is possible to infer the concentration ratio from the spectrum.

The primary results reported in the thesis are:

1. Development of an automated peak picking algorithm that can be applied to a variety of TOF-MS spectra, from counting experiments (TOF-SIMS) to instruments like MALDI-TOF which integrate the ion signal rather than counting individual ions. The noise characteristics are different in these devices.

2. For the TOF-SIMS, the new algorithm improves the precision (*i.e.*, repeatability of estimates of peak positions) by almost an order of magnitude. A similar improvement has been found for surface enhanced

laser desorption ionization mass spectrometry (SELDI-MS), which is a version of MALDI.

3. The algorithm automatically provides estimates of uncertainties in the peak positions and amplitudes. This information is used to verify that aligned spectra collected at different spatial positions or at different times are properly aligned.

4. We tested these algorithms on a set of peptide mixtures and showed that it is possible to estimate three-way mixture ratios, though there is much room for further improvement.

Fig. 1.1 A mass spectrum contains hundreds of peaks
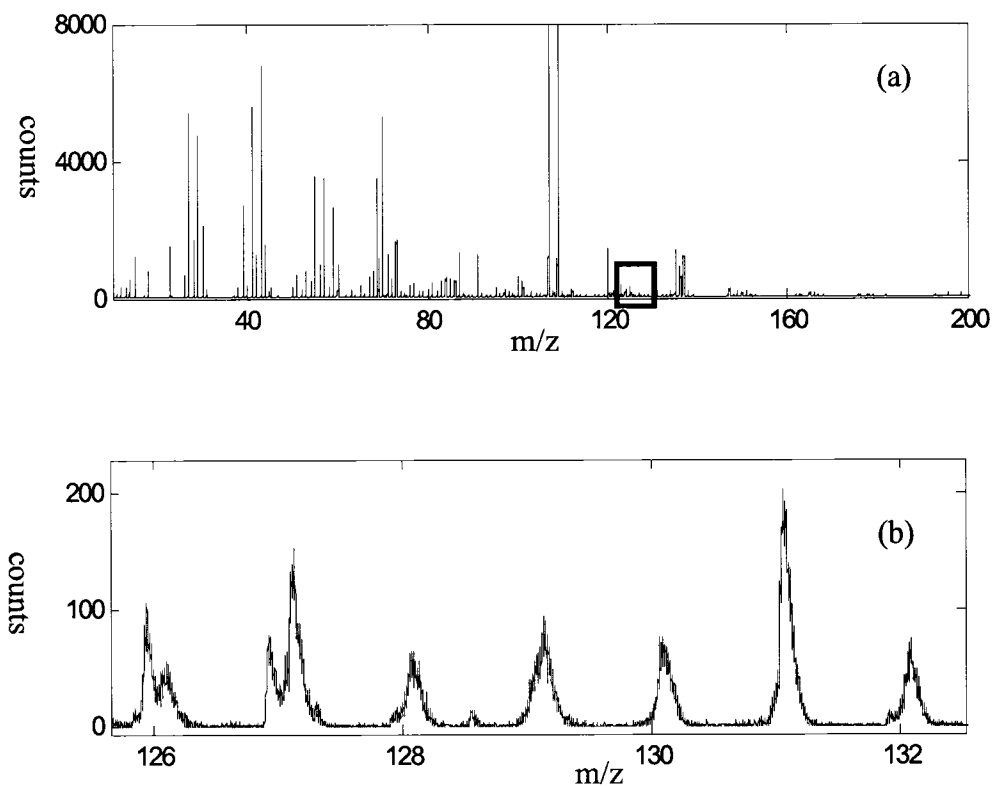
(a) The low mass region (0~200Da) of a typical TOF-SIMS spectrum of a biological sample deposited on silver;

(b) An expanded view of the mass region inside the square in part (a)

Fig. 1.2 Finding peaks by segmentation
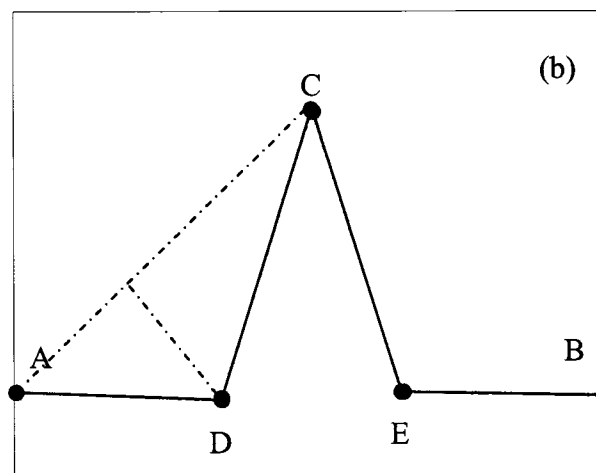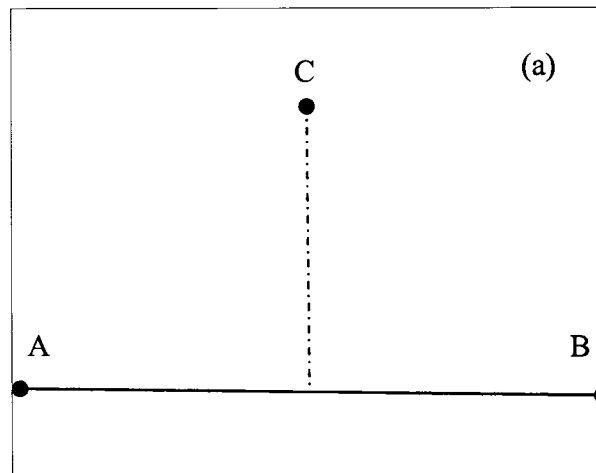
(a) Begin with two ending points A and B, find the point C which has the largest orthogonal distance to line AB.

(b) Next, find point D that has the largest orthogonal distance to AC, similarly find point E. This process iterates until the largest orthogonal distance falls below a predetermined threshold.

Chapter 2

Maximum Likelihood

Methods in Peak Picking

In this chapter we describe how to use Bayesian methods to automatically identify peaks in a TOF-MS spectrum. We will begin with an introduction to TOF-MS, followed by a brief introduction to Bayes' theorem and maximum likelihood method.

Having in mind these basic concepts, we start with an overview of the logic. The first step is to put an observation window of carefully chosen width on the spectrum and thus isolate a small portion of data. We then compare the hypothesis H_1 that there is a peak in the observation window versus hypothesis H_0 that there is no peak in the window, i.e., just background noise. This step is essentially calculating what is called the odds ratio, a definition of which will be given below. Once the comparison concludes that there is a peak inside the window, we then estimate its position and amplitude via parameter fitting by maximum likelihood method.

It is possible sometimes that there could be more that one peak in an observation window. This can be addressed in several ways. An easy fix is to choose an

17

appropriate observation window width such that the window is wide enough to conclude whether or not there is a peak in the window while it is too narrow for more than one peak to be in the window. This works when the instrument is of very high resolving power and peaks are not severely overlapped. If the resolving power of the instrument is not very high and peaks of nearby masses overlap resulting in a broader, fat peak, a test on whether there are multiple peaks in the window will be necessary. However, exploring this subject is beyond the scope of current work and readers interested in this subject may refer to [21].

Once the logical approach is set up, an example of finding peaks in the presence of Gaussian white noise will be given. Gaussian white noise case is particularly chosen to be presented here because it is usually assumed from the perspective of maximum entropy, see Appendix C for details. A nice property of the Gaussian distribution is that all results can be arrived at in closed form because of its smoothness and differentiability. We will show that the outcome of the Gaussian white noise case is what is called matched filter for a known peak shape that is embedded in additive Gaussian white noise of unknown amplitude.

This chapter is arranged as the following. In section 2.1, we will introduce the TOF-MS and Bayes' theorem. We then describe the logic of using Bayesian methods to find peaks in a spectrum in section 2.2. In the last section, section 2.3, we give an example of how to find peaks embedded in Gaussian white noise.

§ 2.1 Introduction to TOF-MS and Bayes' theorem

In this section, we will first introduce some needed concepts for TOF-MS, Bayes' Theorem, and the maximum likelihood method.

§ 2.1.1 Introduction of TOF-MS Instruments

A mass spectrometer is an instrument which separates ions according their mass-to-charge ratio (m/z). The ions are generated from samples of interest, which are usually initially neutral. The history of mass spectrometer goes back to the pioneering work of J. J. Thomson, Physics Nobel laureate of 1906, the discoverer of the electron who investigated the action of electrostatic and magnetic fields on anode rays and canal rays. His work led to the invention of mass spectrograph by F. W. Aston, which was used to measure the mass of isotopes of elements and also won Aston the Nobel Prize in Chemistry in 1922. It was called a mass spectrograph rather than mass spectrometer because photographic plates were used to record ions dispersed by electromagnetic fields.

Roughly speaking, a mass spectrometer consists of three important components: ion source, mass analyzer and ion detector. The ion source generates ions from the sample; the mass analyzer separates ions based on their mass-to-charge ratio; the detector records the separated ions.

"Time-of-Flight" refers to the way that ions are separated, *i.e.*, the mass analyzer. The concept of a TOF mass analyzer is quite simple. A handful of ions start with the same kinetic energy, *e.g.*, after falling through a fixed electrostatic field Φ, fly through

a field-free tube, usually in vacuum, towards an ion detector at the end of the tube. It is easily shown that the time that ions take to fly through the tube of length D is proportional to the square root of mass:

$$E_k = Ze\Phi = \frac{1}{2}mv^2; \quad v = \left(\frac{2Ze\Phi}{m}\right)^{\frac{1}{2}} \tag{2.1}$$

Using $vt=D$, we find:

$$t = \left(\frac{m}{2Ze\Phi}\right)^{\frac{1}{2}} D \tag{2.2}$$

However, developing instrumentation for a TOF mass analyzer in the early days was not as easy as its concept. There were two major reasons. First, there was not an efficient method to create ions from neutral samples, especially from solid samples, only volatile sample could be analyzed. Second, for an ion around 1000Da that is brought to 1keV energy, it would have a speed of $1.4 \times 10^4 m/s$, if the flight distance is $1m$, then the flight time is only about $72\mu s$. This requires a TOF instrument to have a recording device which can work at least at microsecond frame which was not easy 40 years ago. Thus the development of Time-of-Flight type of instrument was limited in early development and was soon displaced by magnetic and quadrupole instruments. However, breakthroughs have been made in recent years. These breakthroughs, on the ionization side, are new ionization methods like Laser Desorption Ionization (LDI), Electro-Spray Ionization (ESI), *etc.*, which enable ions to be generated efficiently from liquid or solid samples. The consequences are that increasingly heavier ions like peptides and proteins can now be generated and that detectable ions can be generated

from a very small amount of sample. For example, it has been reported that MALDI may achieve a detection limit as low as zeptomoles [22]. Fast electronics that can work at nanosecond or sub-nanosecond rates make recording a mass spectrum no longer a problem and give great resolving power. These advantages, plus that a TOF instrument conceptually puts no limit on the mass range, leads to rapid developments in instrumentation and applications. TOF-MS is now widely used in chemistry, biochemistry, biology and biomedical science.

A cartoon of a TOF-MS instrument is shown in Fig. 2.1. Though there are many kinds of TOF-MS instruments, they all work conceptually the same: The sample of interest is placed upon a carefully prepared surface and acted upon by an energy source, which could be a laser, ion beam, or some other energy forms to produce ions. These ions are extracted and accelerated by a static electronic potential to provide the same kinetic energy. To help focus the ions in time and space, 'ion optics' are usually used. These ions then fly freely in a vacuum tube during which they are separated according to their mass-to-charge ratio (m/z), as shown in equation (2.1). Heavier ions will take more time to arrive at the ion detector that sits at a fixed spatial location, a distance D from the ion source.

Figure 2.2 shows a sample spectrum, in which the intensities of ions were plotted versus the time that ions take to fly to the detector. The time can be converted to the mass-to-charge ratio (m/z) through the calibration equaiton:

$$t = am^{1/2} + b \qquad (2.3)$$

It can then be used to deduce chemical, biological and other information of interest.

Ideally, ions of a specific *m/z* would hit the detector after the same time of flight, resulting in a sharp peak lineshape like a delta function of certain height. In reality, however, because of the finite time during which the energy source acts on the sample, sample surface morphology and complex physical and chemical reactions that occur when energy is deposited onto the sample, ions of the same *m/z* are formed at different times and positions according to some initial time distribution and spatial distribution. They also come off the sample surface with an initial velocity distribution of finite width. Though there are ion optics strategies that attempt to correct for these effects, such as time-lag extraction or reflectrons, ions still enter the free drift region with velocity and time distributions of finite width, which results in a finite peak width for ions of a specific *m/z*. We refer to the total sum of ions of a given mass peak (integrated intensity) as the intensity of that peak.

Very often, in a TOF instrument, such as MALDI-MS and SIMS, laser/primary ions which are very well focused can be rastered over the sample surface for a number of irradiations in a certain pattern. For each irradiation, only a small portion of the scanned area is irradiated. The sum of the output of each irradiation gives the final spectrum.

The final spectrum can thus be viewed as consisting of measurements repeated many times under nominally identical conditions to gradually build up a portrait of the probability distribution of arrival times for each *m/z*. This means that the final time

series $s(t)$, that is to be analyzed after all the data has been gathered, is proportional to the conditional probability distribution

$$s(t_k) \propto n(m)p(t_k \mid f_0(v_*(m),t_*))\Delta t, \qquad (2.4)$$

where $s(t_k)$ is the observed signal amplitude at time t_k, Δt is the sampling interval, $n(m)$ is the total number of ions of m/z and $p(t_k \mid f_0(v_*(m),t_*))$ is the probability that an ion of m/z will strike the detector between t_k and $t_k + \Delta t$ given that it entered the drift region at time t_* with a velocity close to $v_*(m)$. If only one ion species is present, the velocity distribution when the ion enters the drift region is a convolution of the ion formation time distribution, initial spatial distribution, initial velocity distribution, etc. Of all mentioned distributions, none of them is fully understood, though some experiments and simulations have resolved some special cases. Thus, we assume that the velocity distribution, $f_0(v_*(m),t_*)$ is just a sharply peaked function of v of the form

$$f_0(v_*(m),t_*) = g\left(\frac{v - v_*(m)}{\sigma(m)}\right), \qquad \int ds g(s) = 1. \qquad (2.5)$$

If a variety of different ions is present, f_0 will be a superposition of terms like (2.5), properly normalized so that f_0 will have unit total area. This assumes that the shape of the velocity spread is universal, up to translation and rescaling. We will assume $v_*(m)$ and $\sigma(m)$ to be weak functions of the mass, and to scale like $1/\sqrt{m}$. In practice, the velocity distributions for each peak will contain information about the dynamics of the ion formation process, which depend upon the ion species. However,

if we assume that the ion optics are designed so that it will only pass a narrow range of most possible velocities for any given mass, then maximizing the entropy of the probability distribution f_0 for a single peak, given v_* and σ, leads to the choice of a Gaussian for g:

$$g\left(\frac{v-v_*}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(v-v_*)^2}{2\sigma^2}\right) \qquad (2.6)$$

The choice of the maximum-entropy distribution is founded upon the principle that it maximizes the number of possible outcomes of repeated observations that are consistent with this assignment of the probability [23]. Hence, it is the least biased assignment of the probability that is consistent with our limited knowledge of the initial distribution. Transforming from the initial velocity distribution g to the temporal peak shape requires solution of a simple Fokker-Planck equation, and is described in a later chapter.

It is important to realize that ions that arrive between time t_k and $t_k + \Delta t$ are independent of those that arrive at any other time t_j, even when t_k and t_j are associated with the same ion peak. This is because, as stated above, the resulting spectrum is an accumulation of many repeated independent measurements under the same experimental condition. This independence will be a crucial assumption that underlies the entire analysis we pursue, and we will discuss this assumption more in the next chapter when we focus more specifically on SIMS. Any correlations in the

signal are assumed to be due to the electronics and should be taken into account as a part of the model used.

§ 2.1.2 Introduction to Bayes' theorem

As we are dealing with finding peaks in the mass spectrum, let two propositions A and B be:

A=There is a particle species of m/z equal to 100 in the sample;

B=We see a bump in the mass spectrum;

If we are told that "if A is true then B is true", then, anytime we knew A true, we could say for sure that B would be true, or, if we learned B was false, we could say for sure A should be false too. This is deductive reasoning, it is certain.

However, in real life, this kind of deductive logic, which deduces the truth about a proposition with a full certainty, does not happen very often. In many cases, we are facing questions like "B is true, what can we say about A"? From the perspective of deductive reasoning, we could find both 'A is true' and 'A is false' are consistent with 'B is true'. The fact that seeing a bump in the spectrum plot does not tell for sure that there is a particle species which has a m/z of 100: the bump could be present because there is such a particle species, but it is also possible due to other reasons like noise from external disturbances, for example, an eighteen-wheeled truck passed nearby when the data is collected.

On the other side, though 'B is true' does not lead to the definite conclusion that A is true, it rules out the possibility of both A and B being false, and makes A more plausible. This is plausible or inductive reasoning, it deals with uncertainty.

As Laplace said in 1819 that, "Probability theory is nothing but common sense reduced to calculations", the next question is then to set up quantitative rules for performing the inductive reasoning.

Before setting out to look for the quantitative rules for inductive reasoning, we want to remind the reader that when we talk about probabilities, we are going to always use the notations introduced in Chapter One.

Second, let us formulate that the derived rules have to satisfy the following desired properties:

    1. Representation of degrees of plausibility by real numbers

    2. Qualitative correspondence with common sense

    3. Consistency

Of these three desiderata, the first two are intuitive, the third one needs a little bit of explanation. Consistency means that, first, if there are several ways to do the reasoning on the same problem, they should arrive at the same result. Second, all evidence that is related to the problem is taken into account, no information is arbitrarily ignored. Third, if the state of knowledge about two problems is the same, then the plausibilities assigned to these two problems are the same.

Having these desiderata set up, it is then just a matter to work the mathematics out to find these rules. However, the detailed mathematical derivation is not the purpose of

this work, we simply point out that, based on the above desiderata, Cox showed that

the product rule and sum rule can be derived [24]:

$$p(X,Y \mid I) = p(X \mid Y,I)p(Y \mid I)$$
$$p(X \mid I) + p(\overline{X} \mid I) = 1$$

(2.7)

Bayes' theorem directly follows from the product rule, if we interchange X and

Y:

$$p(Y,X \mid I) = p(Y \mid X,I)p(X \mid I)$$

(2.8)

Since the probability of 'X and Y' must be the same as 'Y and X':

$$p(X,Y \mid I) = p(X \mid Y,I)p(Y \mid I) = p(Y \mid X,I)p(X \mid I)$$

(2.9)

This gives rise to the Bayes' theorem [25]:

$$p(X \mid Y,I) = \frac{p(Y \mid X,I)p(X \mid I)}{p(Y \mid I)}$$

(2.10)

Now let X be an element of a set of hypotheses. In our peak finding problem, the

hypothesis set is:

$$\{hypothesis_k : there\ are\ k\ peaks\ in\ the\ data, k = 0,1,2 \cdots N\}$$

where N may be a large but finite integer; let Y be the collected data. As a simple

example, let Y be the outdoor temperature measurements that are taken hourly during

successively 24 hours. Then Bayes' theorem in this specific case becomes:

$$p(hypothesis_k \mid data, I)$$
$$= \frac{p(data \mid hypothesis_k, I)p(hypothesis_k \mid I)}{p(data \mid I)}$$

(2.11)

Terms in above equation have their special names:

$p(hypothesis_k \mid I)$ is the *prior probability*, it states our current knowledge of the plausibility of the kth hypothesis, without taking into account the measured data. Common sense tells that it is cooler at night and is warmer during the day, thus, even without looking at the data, we would know that it would be very much more plausible that 'there is a peak in the data' than that 'there is no peak in the data'.

$p(data \mid hypothesis_k, I)$ is the *likelihood function*, sometimes just called *likelihood*. It represents the chance of getting the observed data given the kth hypothesis. For example, if the hypothesis is that 'there is no peak in the data', then the probability of getting data Y would be small as there is peak in Y and hence inconsistent with the hypothesis, or, in other words, less likely.

$p(data \mid I)$ is the chance of getting the observed data ignoring all hypotheses. It appears in the denominator in the right hand side of equation (2.11) as a normalization (scale) factor because we should have:

$$\sum_k p(hypothesis_k \mid data, I) = 1 \qquad (2.12)$$

since the jth hypothesis and k^{th} hypothesis are disjoint, *i.e.*, $hypothesis_j \cap hypothesis_k = \varnothing$. It is often canceled out if we are only interested in comparing two posteriors which is defined next.

$p(hypothesis_k \mid data, I)$, the term in the left hand side of equation (2.11), is the *posterior*, the term we are after in Bayes' theorem. It represents the UPDATED state

of knowledge about the plausibility of the hypothesis having the measured data taken into account.

One should notice that the posterior in one problem could be, say, prior in another problem with independently observed data.

The likelihood function has a special use in parameter fitting, where it is referred to as the maximum likelihood method. To illustrate it, let the hypothesis be 'there is no peak in the data', or better, let us be more specific by saying 'there is no peak in the data, only the background noise which can be characterized by parameter λ.' If the noise is white noise, then the parameters are the mean and the variance. In the usual case, λ is unknown. To estimate λ using the maximum likelihood methods, first draw a number of samples $(x_1, x_2, \cdots x_n)$, write out the likelihood function $p(x_1, x_2, \cdots x_n \mid \lambda, I)$. We will show in the next section how to estimate $\hat{\lambda}$ by maximizing $p(x_1, x_2, \cdots x_n \mid \lambda, I)$, assuming that the noise is white Gaussian.

§ 2.2 Finding peaks in a spectrum----Overview of the logic

We have briefly introduced TOF-MS. The important idea is that the final spectrum is an accumulation of many independent measurements under the same conditions. The ion count at time t_i is independent from that at time t_j, even if they are from the same m/z. We also introduced Bayes' Theorem, which can be derived from the product rule. Now let us put things together to establish the logic of finding peaks in a

spectrum. That is, we are going to put a window of appropriate width on the spectrum. For the data within the window, we are going to compare two models, one model is that there is a peak in the window; the other is that there is no peak present. After the model comparison, if we are confident that there is a peak present, we will then find parameters such as peak amplitude via the maximum likelihood method.

§ 2.2.1 Model comparisons

A mass spectrum, as we can see in Fig. 2.2 (an example spectrum), usually consists of a large number of peaks. To identify them, let us introduce a window at t_0 that includes only N points in the time series, as illustrated in Fig. 2.3.

Before proceeding to further analysis, one should notice that in a mass spectrum, because all ions are subjected to the same instrumental function, peaks at different m/z share the same characteristic shape, the peak lineshape at one m/z and another m/z are similar up to a shift and rescaling. Let us assume that we have some peak model, $x = f(t - t_0)$, which maximizes at $t = t_0$, and describes what the peak lineshape would be. This function could be obtained either empirically or derived from laws of physics/chemistry, but the bottom line is that it captures most of the characteristics of a typical peak in the spectrum. We are going to use t_0 to label the position of the window. The window width N is chosen such that the window covers the region from the left half-max to the right half-max of $f(t - t_0)$. This is a (rough-and-ready) compromise between the desire to include as much data as possible in the window to

improve the sampling statistics and the realization that nearby peaks may overlap and that our peak shape model is probably not very good out on the tails of the peak.

Thus, for the window, we have N isolated data points $(s_1, s_2 \cdots s_N)$ from the spectrum, and we have an N-component vector that describes the peak line shape:

$$x = (x_1, \ldots x_N) \equiv (x(t_1 - t_0), \ldots x(t_N - t_0)) \qquad (2.13)$$

For convenience, let x be normalized to have unit area:

$$\sum_{k=1}^{N} x_k = 1 \qquad (2.14)$$

This will only introduce a constant correction for the peak amplitude computed later.

The first thing we want to determine is whether or not there is a peak in the window. This is a comparison between two hypotheses:

- H_1 = There is a single peak in the window around t_0 with the shape x but of unknown amplitude, embedded in noise of assumed type. Deviations from this shape in the data are due to noise. Let us call the associated peak-plus-noise model M_1;

- H_0 = There is no peak in the widow t_0. The data are noise of the assumed type. Let us call the associated pure-noise model M_0

If more than one peak is present in the window, this complicates the analysis, one needs to compare among hypotheses that 'there are two peaks', 'there are three peaks',

etc., and more parameters must be fit. It will be algebraically more involved, but the logic is similar to what we describe below [21].

For hypothesis H_1, if the noise is additive, it is equivalent to assume that the observed signal $s(t)$ within the window is given by

$$s_k = ax_k + \eta_k, \qquad k = 1, 2, \ldots N \tag{2.15}$$

where a is an unknown amplitude and $\eta = (\eta_1, \eta_2 \cdots \eta_N)$ is a random process of some appropriate type. For example, η may come from a white Gaussian noise. Similarly, for hypothesis H_0, we have:

$$s_k = \eta_k \qquad k = 1, 2 \cdots N \tag{2.16}$$

Under the assumption that the ion counts at different times are independent, the likelihood function, *i.e.*, the probability of observing the particular count sequence $s = (s_1, s_2 \cdots s_N)$ is simply

$$
\begin{aligned}
p(s \mid a, \lambda, t_0, M_k) \\
= p(\{s_1, s_2, \ldots s_N\} \mid a, \lambda, t_0, M_k) \\
= \prod_{i=1}^{N} p(s_i \mid a, \lambda, t_0, M_k)
\end{aligned}
\tag{2.17}
$$

By the notation λ we indicate parameter(s) that characterize(s) the noise process (*e.g.*, the variance σ and mean μ for a Gaussian process or the 'dark' current rate r_0 for a Poisson process). By M_k we mean to emphasize a particular choice of model class, including a peak shape x and a noise model. As in the introduction, the

probability (2.17) is the likelihood function of observing the data in the window labeled t_0 given the model class M_k and the parameters λ.

We wish to compute the probability of which each of these hypotheses was true, given the data, and then take the ratio of two probabilities $p(H_1 \mid s)$ and $p(H_0 \mid s)$:

$$\frac{p(H_1 \mid s)}{p(H_0 \mid s)} = \frac{p(M_1 \mid s, t_0)}{p(M_0 \mid s, t_0)} \tag{2.18}$$

Since at this point for each window position we are only considering two possibilities, i.e., H_1 (there is one peak in the window) and H_0 (there is no peak in the window), these two events are mutually exclusive:

$$p(H_0) + p(H_1) = 1 \tag{2.19}$$

One can then immediately recognize that equation (2.18) is what is called an odds ratio in statistics. If this ratio is large compared to one, we can be confident that there is a peak in the window, while if it is approximately one we interpret that as saying there is only weak evidence of a peak in the window (because $a=0$ is a possible estimate of the peak amplitude, which we interpret as 'no peak'). Therefore, one may set a threshold for peak detection.

In order to compute the odds ratio in (2.18), let us invoke Bayes' theorem (2.11), identifying *hypothesis*$_k$ as the model class M_k, and 'data' as the observed data $s = (s_1, s_2 \cdots s_N)$ in the window t_0:

$$p(M_k \mid s, t_0) = \frac{p(s \mid M_k, t_0) p(M_k \mid t_0)}{p(s \mid t_0)} \tag{2.20}$$

The denominator in (2.20) is the normalization factor we encountered before, it may be computed:

$$p(s \mid t_0) = p(s \mid M_0, t_0)p(M_0 \mid t_0) + p(s \mid M_1, t_0)p(M_1 \mid t_0) \qquad (2.21)$$

However, our interest here is to compare two posteriors as in (2.18), it is clear that this normalization factor cancels.

If we have no reason to prefer one model class over another, we should assign them all equal *prior* probabilities. For example, if we are comparing two types of models (a single peak *vs.* no peak), then $p(M_0 \mid t_0) = p(M_1 \mid t_0) = 1/2$. Therefore, we have the simple result that

$$p(M_k \mid s, t_0) = \frac{p(s \mid M_k, t_0)}{2p(s \mid t_0)} \qquad (2.22)$$

Thus, the odds ratio in (2.18) becomes:

$$\frac{p(H_1 \mid s)}{p(H_0 \mid s)} = \frac{p(M_1 \mid s, t_0)}{p(M_0 \mid s, t_0)} = \frac{p(s \mid M_1, t_0)}{p(s \mid M_0, t_0)} \qquad (2.23)$$

Hence, we need to calculate the likelihood that we observed the data given the model class, $p(s \mid M_k, t_0)$, a quantity called the evidence, notice that it is not conditional on parameters (a, λ). We get this by marginalizing the likelihood function (2.17) over the model parameters using an appropriate *prior* for the parameters:

$$p(s \mid M_k, t_0) = \int da \, d\lambda \, p(s \mid a, \lambda, M_k, t_0) \, p(a, \lambda \mid M_k, t_0) \qquad (2.24)$$

where $p(a,\lambda\,|\,M_k,t_0)$ is the prior distribution of parameters (a,λ). For each model class, there will be a *prior* probability distribution for the parameters. For example, when no peak is present we choose the prior:

$$p(a,\lambda\,|\,t_0,M_0) = \frac{1}{2}\delta(a)p(\lambda\,|\,t_0,M_0) \tag{2.25}$$

where the ½ factor appears because the δ-function will be integrated only over the positive values of a. If we know nothing about the values of λ, then we choose a uniform prior, or some other prior that is very broad in $\lambda-$space on the grounds that when we integrate against (2.17) only the neighborhood of the maximum likelihood value of λ will contribute.

Up to this point, we have set up the concept of doing a model comparison for the data in a window located at t_0, all necessary terms have been computed and the odds ratio is ready to be computed. The window will then slide across the spectrum point by point. When the window is sliding, the window width will eventually get wider because, for instrumental reasons, peaks of heavier mass will become broader. For each window, the odds ratio is computed. As the window comes across a peak, the odds ratio will increase, and will decrease again when the window passes a peak. One can then set a threshold for the confidence we need to have to declare a peak to be detected.

Once we have detected that a peak lies within a certain region of the time axis, we then fix the position and amplitude of the peak by maximizing the likelihood over the

parameters (a, λ, t_0), as discussed in the next section. Before doing so, we summarize that to compare the hypotheses that there is, or is not, a peak in the window, we need to compute the ratio (2.18), which requires computation of (2.24) for both model M_0 and model M_1.

§ 2.2.2 Parameter fitting

As the window slides across the spectrum one point at a time, the odds ratio is calculated for each window position. We can then justify in which region in the spectrum we are confident that there is a peak. We then look into particular regions of interest, to fit parameters (a, λ) by maximizing the likelihood function (2.17). This requires solving the following equations, for a window located at t_0 with isolated data $s = (s_1, s_2 \cdots s_N)$:

$$\frac{\partial L(a, \lambda, t_0)}{\partial a} = 0$$

$$\frac{\partial L(a, \lambda, t_0)}{\partial \lambda} = 0$$
(2.26)

where $L(a, \lambda, t_0)$ is the natural logarithm of likelihood function (2.17):

$$L(a, \lambda, t_0) = \ln(p(s \mid a, \lambda, M_k, t_0))$$
(2.27)

Solving equations (2.26) gives the maximum likelihood estimations of parameters (a^*, λ^*) for window with a fixed t_0 and data s. If the data is informative, then the likelihood would sharply peak around the point (a^*, λ^*) in the parameter space and

die off quickly as we move away from (a^*, λ^*). It is natural to Taylor expand (2.27)

around (a^*, λ^*):

$$L(a, \lambda, t_0)$$

$$\approx L(a^*, \lambda^*, t_0)$$

$$+ \frac{1}{2} \left[\frac{\partial^2 L}{\partial a^2} \Big|_{a^*, \lambda^*} (a - a^*)^2 + \frac{\partial^2 L}{\partial \lambda^2} \Big|_{a^*, \lambda^*} (\lambda - \lambda^*)^2 \right] \qquad (2.28)$$

$$+ \frac{\partial^2 L}{\partial a \partial \lambda} \Big|_{a^*, \lambda^*} (a - a^*)(\lambda - \lambda^*)$$

$$= L(a^*, \lambda^*, t_0) + \frac{1}{2}(X - X^*)'\nabla\nabla L(a^*, \lambda^*, t_0)(X - X^*)$$

where $X = \begin{bmatrix} a \\ \lambda \end{bmatrix}$, and

$$\nabla\nabla L(a^*, \lambda^*, t_0) = \begin{bmatrix} \dfrac{\partial^2 L}{\partial a^2}\Big|_{a^*, \lambda^*} & \dfrac{\partial^2 L}{\partial a \partial \lambda}\Big|_{a^*, \lambda^*} \\[2ex] \dfrac{\partial^2 L}{\partial a \partial \lambda}\Big|_{a^*, \lambda^*} & \dfrac{\partial^2 L}{\partial \lambda^2}\Big|_{a^*, \lambda^*} \end{bmatrix} \qquad (2.29)$$

is the Hessian matrix evaluated at (a^*, λ^*).

It follows from (2.28) that the leading term of the likelihood function in (2.17) is

approximately:

$$p(s \mid a, \lambda, M_k, t_0) = \exp\left[L(a, \lambda, t_0)\right]$$

$$\approx \exp\left(L(a^*, \lambda^*, t_0) + \frac{1}{2}(X - X^*)'\nabla\nabla L(a^*, \lambda^*, t_0)(X - X^*) \right) \qquad (2.30)$$

$$= e^{L(a^*, \lambda^*, t_0)} e^{\frac{1}{2}(X - X^*)'\nabla\nabla L(a^*, \lambda^*, t_0)(X - X^*)}$$

This implies that the likelihood function looks like a multivariate normal distribution in parameter space, centered at (a^*, λ^*) with the following uncertainties, if a and λ are not coupled:

$$\sigma_a = \left(-\frac{\partial^2 L}{\partial a^2} \bigg|_{a^*, \lambda^*} \right)^{-1/2}$$

$$\sigma_\lambda = \left(-\frac{\partial^2 L}{\partial \lambda^2} \bigg|_{a^*, \lambda^*} \right)^{-1/2}$$

(2.31)

Moreover, the approximation in equation (2.30) provides a possibly easy way to compute the evidence in equation (2.24) in the sense that if the prior is independent of (a, λ), for example, a and λ are uniformly distributed in some region (a_{min}, a_{max}) and $(\lambda_{min}, \lambda_{max})$, with substitution of (2.30) into (2.24), the integration is readily carried out:

$$p(s \mid M_k, t_0) = \int da\, d\lambda\, p(s \mid a, \lambda, M_k, t_0) p(a, \lambda \mid M_k, t_0)$$

$$= \int_{a_{min}}^{a_{max}} \int_{\lambda_{min}}^{\lambda_{max}} da\, d\lambda\, e^{L(a^*, \lambda^*, t_0)} e^{\frac{1}{2}(X - X^*)' \nabla\nabla L(a^*, \lambda^*, t_0)(X - X^*)} \frac{1}{a_{max} - a_{min}} \frac{1}{\lambda_{max} - \lambda_{min}} \quad (2.32)$$

$$= \frac{1}{a_{max} - a_{min}} \frac{1}{\lambda_{max} - \lambda_{min}} e^{L(a^*, \lambda^*, t_0)} \frac{(2\pi)^{m/2}}{\sqrt{\left| \det\left[\nabla\nabla L(a^*, \lambda^*, t_0) \right] \right|}}$$

where m is the dimension of parameter space. In the last step of integration, lower and upper boundaries of integration are extended to infinity. This is valid if the likelihood function is sharply peaked around (a^*, λ^*) and (a_{min}, a_{max}) and $(\lambda_{min}, \lambda_{max})$ are large enough such that contributions from outside these regions are negligible. Otherwise, the integral will result in an error function. Note $\det\left[\nabla\nabla L(a^*, \lambda^*, t_0) \right]$ is the

determinate of Hessian matrix evaluated at (a^*, λ^*), and $1/\sqrt{\left|\det\left[\nabla\nabla L(a^*, \lambda^*, t_0)\right]\right|}$ is

proportion to the 'volume' within σ_a and σ_λ around (a^*, λ^*) in parameter space, *i.e.*

$$p(n \mid M_k, t_0) \sim p(s \mid a^*, \lambda^*, M_k, t_0) \frac{\sigma_a}{a_{\max} - a_{\min}} \frac{\sigma_\lambda}{\lambda_{\max} - \lambda_{\min}} \qquad (2.33)$$

Notice that solving equation (2.26) only maximizes the likelihood with respect to

(a, λ), the maximizing of likelihood with respect to t_0 is done by computing the

likelihood for each window position at (a^*, λ^*), *i.e.*, $p(n \mid a^*, \lambda^*, M_k, t_0)$ and then find

the maximum point of $p(n \mid a^*, \lambda^*, M_k, t_0)$ with respect to t_0. However, maximizing

over t_0 has a different logical character than the other parameters, because we are

comparing *different* data sets as we slide the window across the peak. The

justification is based upon the physical reasonableness of the approach: the width of

the window is large compared to the uncertainty in the position of the peak, hence near

the maximum of the likelihood, most of the data being compared comes from

overlapping windows. An alternative way of looking this is that by comparing

$p(n \mid a^*, \lambda^*, M_k, t_0)$ at different t_0 we are actually looking for a window in which the

data best support the assumption that there is a peak in the window.

Before we go any further, let us summarize that, in the normal way, in order to

compare model M_1 versus M_0 we need to evaluate the odds ratio (2.18) in which we

need to compute the evidence in (2.24) by marginalizing the likelihood function (2.17)

over the prior distribution $p(a, \lambda \mid M_k, t_0)$ of parameters (a, λ). If the odds ratio is

large compare to one, it strongly suggests that there is a peak in the window; otherwise, it is more likely that there is no peak. Once the odds ratio concludes that a peak is present, parameters may be fit by maximizing the likelihood function (2.17), via solving equation (2.26). Equation (2.32) provides an alternative way of computing the evidence if that the likelihood is strongly peaked in parameter space is satisfied.

§ 2.3 Finding a peak embedded in Gaussian noise

We now consider a specific example, *i.e.*, detection of a peak in white Gaussian noise. We show that the approach described above leads to what is called the 'matched filter' in the engineering literature. The Gaussian white noise case is of particular importance because in many situations the noise characteristics are not clear or are very complex and Gaussian white noise is often assumed based on a maximum entropy argument. In this section, we do not specify the peak shape but simply use x_i to denote peak shape in general.

For convenience, from now on, except for t_0 , when referring to parameters associated with model M_0, a subscript 0 will be used, a subscript 1 will be used for parameters associated with model M_1. For both models, a superscript of star (*) will be used for best estimations of parameters.

Since we assume the noise process is Gaussian and white, then for the noise $\eta = (\eta_1, \eta_2 \cdots \eta_N)$ in the window t_0 we have:

$$p(\eta_k) = \frac{1}{\sqrt{2\pi}\sigma_\eta} \exp\left(-\frac{(\eta_k - \mu)^2}{2\sigma_\eta^2}\right) \quad k = 1,2\cdots N \tag{2.34}$$

$$\langle \eta \rangle = \mu \tag{2.35}$$

$$\langle \eta_j \eta_k \rangle = \sigma_\eta^2 \delta_{jk} \tag{2.36}$$

Let us first consider model M_1, which says there is a peak in the observation window of known shape x but unknown amplitude a_1, in the presence of noise:

$$s_i = a_1 x_i + \eta_i \quad i = 1\ldots N \tag{2.37}$$

It can be rewritten as:

$$\eta_i = s_i - a_1 x_i \quad i = 1\ldots N \tag{2.38}$$

Under the assumption of independence between two data points, one can write the likelihood function (2.17) explicitly:

$$p_N(\eta \,|\, a_1, \mu_1, \sigma_{\eta 1}, t_0, M_1) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{\eta 1}} e^{-\frac{(s_i - a_1 x_i - \mu_1)^2}{2\sigma_{\eta 1}^2}}$$
$$= \frac{1}{(\sqrt{2\pi}\sigma_{\eta 1})^N} \exp\left(-\frac{1}{2\sigma_{\eta 1}^2}\sum_{i=1}^{N}(s_i - a_1 x_i - \mu_1)^2\right) \tag{2.39}$$

where a subscript N is used to emphasize that $p_N(\eta \,|\, a_1, \mu_1, \sigma_{\eta 1}, t_0, M_1)$ is a likelihood function with N data points considered.

Following the summary made at the end of previous section, the next step is to compute the evidence. One can compute it by marginalizing the likelihood function (2.39) over parameters $(a_1, \mu_1, \sigma_{\eta 1})$ to get the evidence, $p(\eta \,|\, M_1, t_0)$. We now assume

that we know nothing about the prior probabilities for all parameters, the amplitude of

the peak a_1, the mean of noise μ_1 and the variance of the noise $\sigma_{\eta 1}$, aside from the

requirement that they be probabilities in the parameter space $(a_1, \mu_1, \sigma_{\eta 1})$. The

simplest choice is to assume that they are constant over some range $(0, a_{max}]$, $(0, \mu_{max}]$,

$(0, \sigma_{\eta_{max}}]$:

$$p(a_1, \mu_1, \sigma_{\eta 1} \mid t_0, M_1) = \frac{1}{a_{max}} \frac{1}{\mu_{max}} \frac{1}{\sigma_{\eta_{max}}} \qquad (2.40)$$

However, this allows us to take a short cut by utilizing equation (2.32) to get the

evidence, instead of carrying out the integral. We will do it that way in the next

chapter when finding peaks in a SIMS spectrum. So, let us find the maximum

likelihood estimation of parameters $(a_1^*, \mu_1^*, \sigma_{\eta 1}^*)$ and the Hessian matrix at

$(a_1^*, \mu_1^*, \sigma_{\eta 1}^*)$. These are easily computed, we will only quote the results here, one may

find detailed calculations in the Appendix A:

$$a_1^* = \frac{\overline{(sx)} - \overline{s}\,\overline{x}}{\overline{x^2} - \overline{x}^2} = \frac{\overline{(s - \overline{s})(x - \overline{x})}}{\overline{(x - \overline{x})^2}} = \frac{\overline{(s - \overline{s})(x - \overline{x})}}{\sigma_x^2}$$

$$\mu_1^* = \overline{s} - a_1^* \overline{x} \qquad (2.41)$$

$$\sigma_{\eta 1}^* = \overline{(s_i - \overline{s} - a_1^*(x_i - \overline{x}))^2} = \sigma_s^2 - a_1^{*2}\sigma_x^2$$

$$\nabla\nabla L(a_1^*,\mu_1^*,\sigma_{\eta 1}^*,t_0) = \begin{bmatrix} L_{aa} & L_{a\mu} & L_{a\sigma} \\ L_{\mu a} & L_{\mu\mu} & L_{\mu\sigma} \\ L_{\sigma a} & L_{\sigma\mu} & L_{\sigma\sigma} \end{bmatrix}$$

$$= -\begin{bmatrix} \dfrac{N(\sigma_x^2 + \overline{x}^2)}{\sigma_{\eta 1}^{*\,2}} & \dfrac{\sum\limits_{i=1}^{N} x_i}{\sigma_{\eta 1}^{*\,2}} & 0 \\[4mm] \dfrac{\sum\limits_{i=1}^{N} x_i}{\sigma_{\eta 1}^{*\,2}} & \dfrac{N}{\sigma_{\eta 1}^{*\,2}} & 0 \\[4mm] 0 & 0 & \dfrac{2N}{\sigma_{\eta 1}^{*\,2}} \end{bmatrix}$$

(2.42)

where $\overline{x}$ is arithmetic mean of x, σ_x^2 is the variance of x:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

(2.43)

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

(2.44)

$\overline{s}$ and σ_s^2 are defined similarly.

Thus, we have our uncertainties about our best estimation on parameters and the evidence:

$$\Delta a_1 = \frac{\sigma_{\eta 1}^*}{\sqrt{N(\sigma_x^2 + \overline{x}^2)}}$$

$$\Delta\mu_1 = \frac{\sigma_{\eta 1}^*}{\sqrt{N}}$$

(2.45)

$$\Delta\sigma_{\eta 1} = \frac{\sigma_{\eta 1}^*}{\sqrt{2N}}$$

$$p(\eta \mid M_1, t_0) \approx \frac{1}{a_{1_{max}}} \frac{1}{\mu_{1_{max}}} \frac{1}{\sigma_{\eta 1_{max}}} \frac{(2\pi)^{3/2}}{\sqrt{\left| \det(\nabla \nabla L(a_1^*, \mu_1^*, \sigma_{\eta 1}^*, t_0)) \right|}} e^{L(a_1^*, \mu_1^*, \sigma_{\eta 1}^*, t_0)} \quad (2.46)$$

A little more algebra shows (see Appendix A):

$$p(\eta \mid M_1, t_0)$$

$$\approx \frac{1}{a_{1_{max}}} \frac{1}{\mu_{1_{max}}} \frac{1}{\sigma_{\eta 1_{max}}} \frac{\sigma_{\eta 1}^*}{\sqrt{N \sigma_x^2}} \frac{\sigma_{\eta 1}^*}{\sqrt{N}} \frac{\sigma_{\eta 1}^*}{\sqrt{2N}} (2\pi)^{3/2} p^*(\eta \mid a_1^*, \mu_1^*, \sigma_{\eta 1}^*, M_1, t_0) \quad (2.47)$$

$$\propto \frac{\Delta a_1}{a_{1_{max}}} \frac{\Delta \mu_1}{\mu_{1_{max}}} \frac{\Delta \sigma_{\eta 1}}{\sigma_{\eta 1_{max}}} (2\pi)^{3/2} p_N^*(\eta \mid a_1^*, \mu_1^*, \sigma_{\eta 1}^*, M_1, t_0)$$

This is consistent with the general results stated in previous section.

For the case of model M_0, i.e. there is no peak in the observation window, the signal observed is considered as all coming from noise, i.e. $s_i = \eta_i$, the calculation is similar to model M_1. In fact one can get all results simply by eliminating amplitude component from model M_1 and find the evidence is:

$$p(\eta \mid M_0, t_0) \approx \frac{1}{\mu_{0_{max}}} \frac{1}{\sigma_{\eta 0_{max}}} \frac{\sigma_{\eta 0}^*}{\sqrt{N}} \frac{\sigma_{\eta 0}^*}{\sqrt{2N}} 2\pi p_N^*(\eta \mid \mu_0^*, \sigma_{\eta 0}^*, M_0, t_0)$$

$$= \frac{\Delta \mu_0}{\mu_{0_{max}}} \frac{\Delta \sigma_{\eta 0}}{\sigma_{\eta 0_{max}}} 2\pi p_N^*(\eta \mid \mu_0^*, \sigma_{\eta 0}^*, M_0, t_0) \quad (2.48)$$

Having the evidences for both models computed, the odds ratio is easily computed:

$$\frac{p(H_1 \mid n)}{p(H_0 \mid n)} = \frac{p(M_1 \mid n, t_0)}{p(M_0 \mid n, t_0)} = \frac{p(\eta \mid M_1, t_0)}{p(\eta \mid M_0, t_0)}$$

$$= \frac{1}{a_{1_{max}}} \frac{\sigma_{\eta 1}^*}{\sqrt{N\sigma_x^2}} \sqrt{2\pi} \left(\frac{\sigma_{\eta 0}^*}{\sigma_{\eta 1}^*} \right)^{N-2} \tag{2.49}$$

$$\propto \frac{\Delta a_1}{a_{1_{max}}} \sqrt{2\pi} \left(\frac{\sigma_{\eta 0}^{*\,2}}{\sigma_{\eta 1}^{*\,2}} \right)^{\frac{N-2}{2}}$$

One must notice that the above calculation is done only on ONE window that is located at t_0. Perhaps a better way to describe the best estimation of $(a_1, \mu_1, \sigma_{\eta 1})$ is to write their dependence on t_0 explicitly, i.e. $(a_1^*(t_0), \mu_1^*(t_0), \sigma_{\eta 1}^*(t_0))$. In the region that the odds ratio (2.49) indicates a peak, one will need to compute $p_N^*(\eta \mid a_1^*, \mu_1^*, \sigma_{\eta 1}^*, M_1, t_0)$ for each window position, find the t_0^* that maximizes $p_N^*(\eta \mid a_1^*, \mu_1^*, \sigma_{\eta 1}^*, M_1, t_0)$ and report t_0^* as peak position and $a_1^*(t_0^*)$ as peak amplitude.

If one looks closely at the maximum likelihood estimation of the peak amplitude:

$$a_1^* = \frac{\overline{(sx)} - \overline{s}\,\overline{x}}{\overline{x^2} - \overline{x}^2} = \frac{\overline{(s - \overline{s})(x - \overline{x})}}{\overline{(x - \overline{x})^2}} = \frac{\overline{(s - \overline{s})(x - \overline{x})}}{\sigma_x^2} \tag{2.50}$$

one would see that it is in fact doing a correlation between data in the window and the known peak lineshape, the same as a matched filter does.

However, we do not use $a_1^*(t_0)$ to estimate peak position as is typically done with matched filters, but instead use the likelihood function p_N^*. Matched filter, as can be seen in (2.50), involves correlation between the signal and the known peak line shape, which would result in a broad peak $a_1^*(t_0)$, while p_N^* is narrower. This is showed in

Fig. 2.4 with simulated data. In Fig. 2.4, from the top to bottom, a simulated peak,

$a_1^*(t_0)$ and the natural log of p_N^*, $\log(p_N^*)$, are plotted in turn. The location of the

simulated peak is marked by a dash line. The natural log of p_N^* has its characteristic

shape, which we will elaborate in next chapter. The thing to notice here is that

$\log(p_N^*)$ is much narrower than $a_1^*(t_0)$, it would be even shaper if we exponentiate it

to get p_N^*.

Fig. 2.1 Illustration of the concept of TOF-MS.

(3)

(5)

(5)

D

(2)

(1)

(4)

(6)
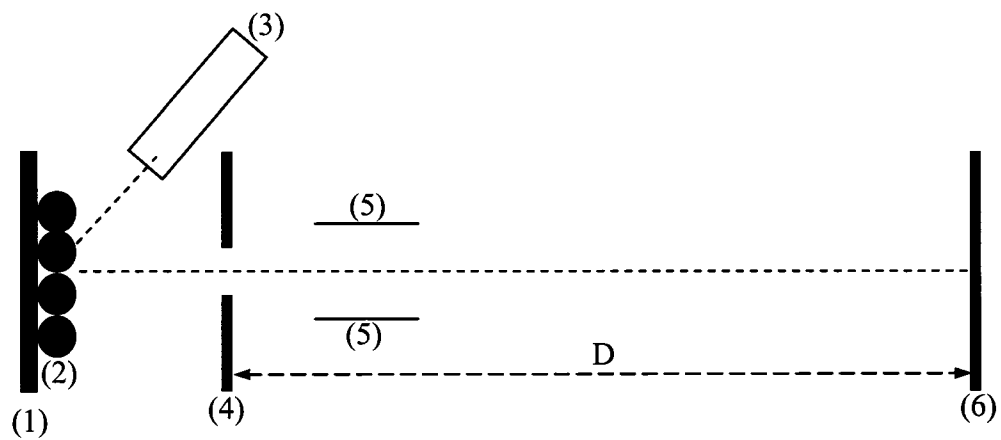
(1) Substrate
(2) Sample
(3) Energy source
(4) Extracting voltage
(5) Ion optics
(6) Detector

Fig. 2.2 An example spectrum
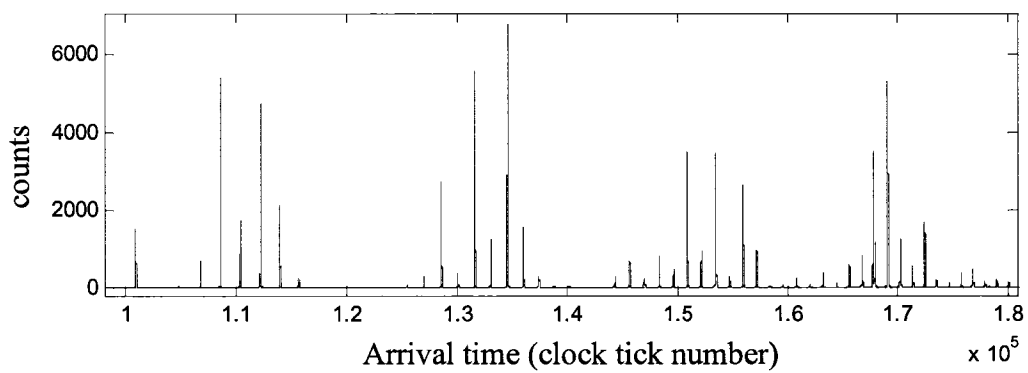
The spectrum is plotted as counts versus arrival time (in clock tick number)

Fig. 2.3 An observation window of width N
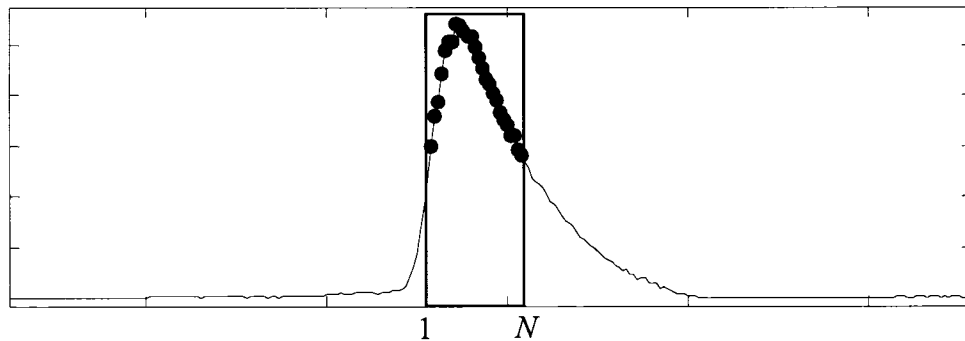
A window of width N is superimposed on the spectrum. It isolates N data points (the black dots). For these N data points, we compare the hypotheses that there is a peak versus that there is no peak. The window width is chosen in such a way that when the window is right on top of the peak, the window runs from the left half maximum to the right half maximum of the peak. The window slides one clock tick a time to the right and the two hypotheses are compared for each window.

Fig. 2.4 Comparison of sharpness
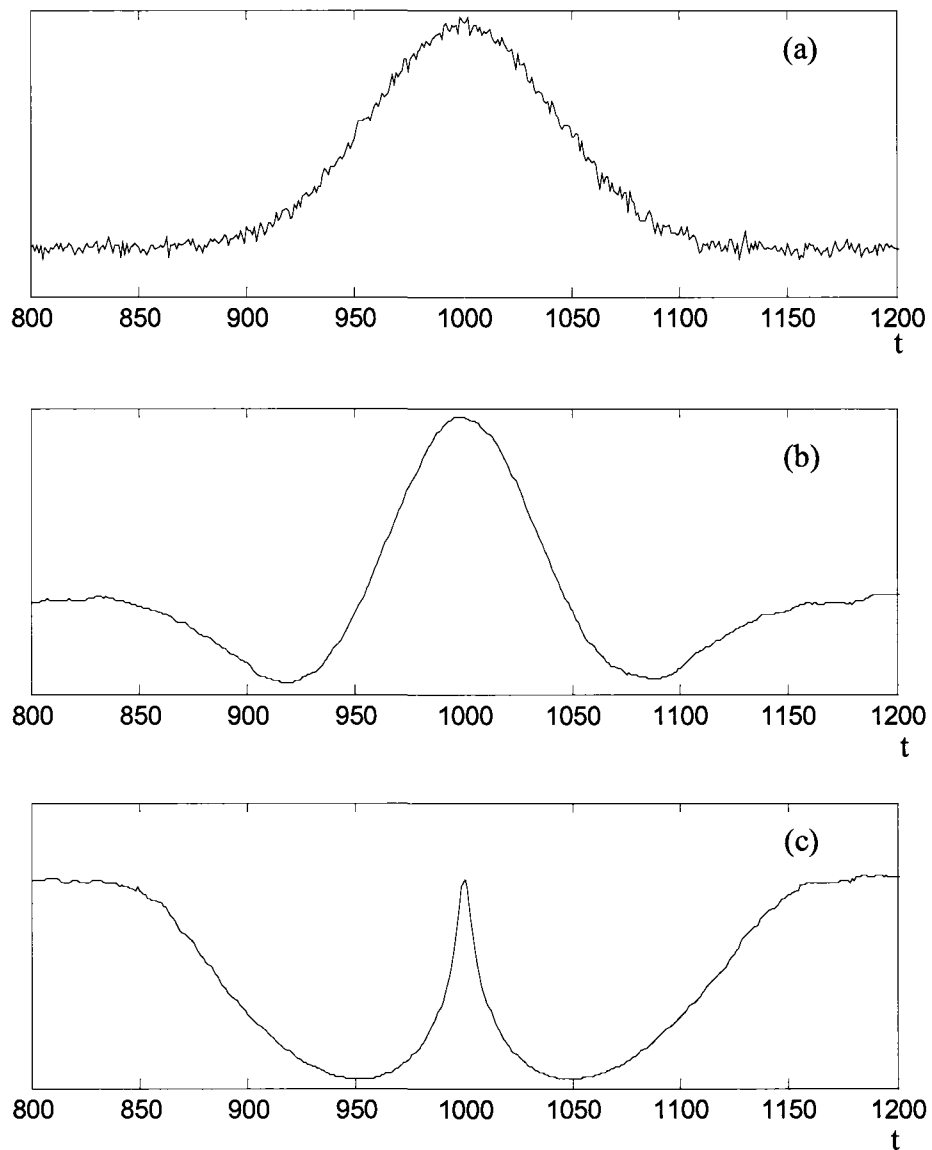
For a simulated peak, $\log(p_N^*)$ is much sharper than $a_1^*(t_0)$

   (a) A simulated peak locates at t=1000;

   (b) The corresponding $a_1^*(t_0)$;

   (c) The corresponding $\log(p_N^*)$

Chapter 3

TOF-SIMS Data Analysis

In this chapter, we first introduce basic principles of static TOF-SIMS. We will show that a Poisson process is involved and this would be our basis for finding peaks in a TOF-SIMS spectrum. Then, formulas for identifying peaks in a SIMS spectrum will be derived and examples with simulated data will be given.

§ 3.1 Introduction to TOF-SIMS

In 1931, Woodstock first observed secondary ion formation when a sample was bombarded with ions and SIMS came to the world thereafter. SIMS, by itself, is a kind of desorption mass spectrometry. It uses an energetic primary ion beam, usually with keV energy, to probe the sample surface. As a result of the primary ion impact and subsequent energy transfer, secondary particles are generated. Of these secondary particles, some are ionized and then mass analyzed.

Generally, SIMS works in two different regimes, dynamic SIMS and static SIMS. The difference between dynamic SIMS and static SIMS is in the primary ion flux.

51

Dynamic SIMS uses a continuous primary ion current to erode the sample surface, consecutive sample layers are removed very fast, generating primarily elemental ions and some small cluster ions. Because of the fast erosion of the sample, depth profiling can be easily achieved. When combined with scanning of the primary ion beam that rasters the surface, one can build 3-D characterization of a sample's composition. However, because of the high primary ion current, organic samples will be fragmented too much to generate useful information.

It was Benninghoven who introduced static SIMS in 1970, and a detectability of less than 1ppm was reported [26]. In static SIMS, a small current of primary ion beam is used to probe the sample surface such that less than 1% of the top layer receives an ion impact. The idea is that no area should receive more than one impact. This puts an upper limit on the primary ion dose. It is generally accepted that for the static SIMS regime, the primary ion dose should not exceed 10^{13} ions/ cm^2, which is approximately 1% of the atomic surface density of a silicon wafer. The actual static limit depends on properties of the analyte under investigation, the primary ion being used, primary ion incident angle, etc. For example, an organic sample would have a smaller static limit, as do primary-ion beams of heaver mass and larger geometrical size.

There are many kinds of primary-ion beam. For example, inert gas ions like Ar^+, Xe^+ are commonly used as primary ions because of their inertness. Liquid metal ion guns (LMIG), like Ga^+, are also widely used because they provide submicron lateral resolution and are thus ideal for imaging purposes. Recent research has shown that the use of polyatomic primary ions, such as C_{60}^+ and Au_n^+, leads to a nonlinear increase in

secondary-ion yield and thus is promising for large organic and biological samples [27, 28, 29].

When the sample is bombarded by the primary ions, secondary particles, including neutrals, positive and negative ions, and electrons, are emitted from the sample surface. The majority of these particles come from the top one or two monolayers, and over 99% of them are neutral. The exact mechanisms by which ions are formed during the emission are not fully understood and seem to depend on the type of analyte, as well as the local chemical and physical environment. Various models have been proposed, either from a physical perspective or from chemical point of view. Vaeck *et al.* provide a good summary in their review on static TOF-SIMS [30]. However, it is clear that the energy needed for ejecting these secondary particles comes from the energetic primary ion. Around the primary-ion impact point, energy and momentum is transferred into the neighboring area via a collision cascade, creating an excited area [31, 32]. The energy and momentum distributions within the excited area depends on many factors, such as the lattice structure (if any) of the sample/substrate, the energy, mass and physical size of the primary ion, the incident angle of the primary ion, and so on. But, qualitatively, the closer to the impact point, the higher the energy transfer. Those molecules that are directly impacted by the primary ions will break into pieces (fragments) as the energy of primary ions, in keV range, is far larger than the bond energy. Even those molecules close to the impact point will also get so much energy that they will fragment. Only molecules that are remote from the impact point might gain sufficient energy to escape the surface as intact molecules yet remain

unfragmented. For organic samples that are adsorbed as thin film on a metal surface, the primary ion penetrates into the substrate and initiates a collision cascade in the substrate atoms. Molecular dynamics simulations have shown that this collision cascade in the substrate is responsible for the ejection of intact molecules [33]. It has also been shown that most of the fragments come from a region close to the impact point, less than 5 angstroms, while intact molecules come from more distant areas up to 30 angstroms away [34].

As the bottleneck of fast electronics is no longer a problem, a TOF mass analyzer becomes a favorable one. The ratio of the number of ions that leave a region of a mass spectrometer to the number of ions that enter that region is often called transmission. It is the high transmission that makes TOF mass analyzer superior to some other scanning mass analyzers, like quadrupole, which have been commonly used on SIMS apparatus for a long time. The quadrupole mass filter has a mass limit and runs in a scanning mode. Only ions of a specifically selected m/z can pass through at a time. Moreover, a quadrupole has an energy band limit about 5~20eV, but the secondary ions generated on the surface usually have an initial energy distribution spread over tens of eV's, especially for elemental ions, which means that even ions of the specifically selected m/z can not all pass through the quadrupole. These considerations suggest that quadrupoles have a very low transmission. In the case of static SIMS, where the ion yield is very low, this requires a sample to be exposed to the primary-ion beam for a long time to get a survey spectrum, and this exceeds the static limit. The importance of keeping to the static limit is that the damage caused by primary-ion

erosion is permanent. It changes the chemical and physical composition in the excited area. Bombarding in such area again will generate information of limited usefulness. The advantage of a TOF mass analyzer is then obvious: all ions generated on the surface will be transmitted to the detector and all ions of the full range of m/z are detected in parallel. The transmission of a TOF mass analyzer overwhelms that of quadrupole. This also results in a much shorter acquisition time and thus avoiding violating the static limit. Another advantage is that, theoretically, a TOF mass analyzer puts no upper limit on mass. This becomes crucial as organic and biological samples are analyzed. For static SIMS, the heaviest detectable analyte goes up to 10,000Da [35].

The use of a TOF mass analyzer poses a special requirement. As a TOF mass analyzer differentiates ions of different m/z by measuring their flight time, it is important to have an accurate measurement of the flight start time. A way to achieve this is to use a pulsed primary-ion beam. A pulsed primary-ion beam is also essential to maintain the static limit. The pulsed primary-ion beam is repeated at an appropriate repetition rate which is determined by the detection mass range set up such that the detector will not confuse the heaviest (slowest) secondary ion of one pulse with the lightest (fastest) secondary ion of the next pulse. The final spectrum is the sum of the detector output of each primary-ion pulse.

As mentioned above, of all secondary particles from the primary ion impact, only a few are ionized, i.e., the secondary-ion yield is very low. For example, in a typical TOF-SIMS spectrum we collected on a peptide sample, about 1.4×10^8 primary ions

were delivered onto the sample surface ($4 \times 10^{-4} cm^2$) in about 1.1×10^6 pulses, and about 10×10^6 secondary ions were detected. This means that for each primary ion pulse, we detected only about 10 secondary ions. These secondary ions cover the mass range up to 2000Da, and are well separated in arrival time. It is thus possible to count each of them. And, in fact, it is a counting detector system that most TOF-SIMS are equipped with. Modern developments of fast electronics have facilitated such counting systems. For example, a microchannel plate (MCP), which is often used in such counting systems, generates an electron pulse of width from a sub-nanosecond to a few nanoseconds for a single particle impact event. Working together with a multistop time-to-digital converter (TDC), ion impact events occurring less than a nanosecond apart can be resolved. For example, the TRIFT II TOF-SIMS has a maximum time resolution of .13ns. It is also shown that, when optimized carefully, MCP offers good ion detection efficiency for ions up to 10,000Da [36].

§ 3.2 Applications of TOF-SIMS

As TOF-SIMS gathers information only from the top few monolayers and within a very small area (usually in micron scale) on sample surface, it is very sensitive to sample composition. The high sensitivity, together with the high resolution, has made TOF-SIMS a powerful surface analytical technique, from research to production

control [37]. It has been widely used from isotope ratio measurements, polymer analysis, biological surface analysis to biological tissue imaging [38, 39, 40, 41].

§ 3.3 Poisson processes and Independence

As stated in the Introduction of this chapter, the essence of the static limit is that each primary ion impinges on fresh sample area that has not been impacted and is unaffected by impacts that have happened elsewhere. It is then clear that each impact is an independent measurement of the sample under the same conditions. For each measurement, the chance of seeing an ion of m/z arriving at the detector at t_j, $p_{m/z,t_j}$, is determined by the sputtering yield, the ionization probability, transmission, detection probability, initial velocity distribution, etc. The combination of these factors makes $p_{m/z,t_j}$ very small but stable under the static regime. During a TOF-SIMS experiment, often millions of primary ions are delivered to the sample surface to get a good signal. Thus, the chance of observing n m/z ions in N impacts (n "good" outcomes in N "observations") is given by the Bernoulli distribution:

$$P(n \mid p_{m/z,t_j}, N) = \frac{N!}{n!(N-n)!} p_{m/z,t_j}^n (1 - p_{m/z,t_j})^{N-n} \qquad (3.1)$$

Now, let:

$$p_{m/z,t_j} = \frac{r}{N} \qquad (3.2)$$

where r, called the 'rate', is the expected the number of m/z ions in N observations. Let N become large, holding r fixed:

$$P(n \mid r, N) = \lim_{N \to \infty} \frac{N!}{n!(N-n)!} \left(\frac{r}{N}\right)^n \left(1 - \frac{r}{N}\right)^{N-n} \tag{3.3}$$

A little algebra and using Stirling's approximation, note that $(1 - r/N)^N \sim e^{-r}$, leads us to the Poisson distribution:

$$P(n \mid r, N) = \frac{r^n e^{-r}}{n!} \tag{3.4}$$

This says that, at a specific time t_j in a spectrum of some sample, if the "IDEAL" (expected) number of counts after N pulses is r, then the probability of actually observing n counts follows the Poisson distribution and can be calculated by (3.4).

Let us consider an analogy to this TOF-SIMS counting process. Suppose we want to make a survey of the automobile market share. We do this by standing beside a road next to a traffic signal and for each green light, we categorize each passing vehicle according its brand and model. Let us now build a conceptual mapping that connects the TOF-SIMS process and market share survey:

Table 3.1 Conceptual mapping between TOF-SIMS process and market share survey

	TOF-SIMS	**Automobile market share survey**
Event	A primary ion impact	A green light
Observed quantity	Ions of a specific m/z, say m/z=100	Vehicles from a specific maker, say Toyota
Possible Outcome 1	An ion of m/z=100 arrives detector at t_j	A Toyota passes, and the model is 4Runner
Possible Outcome 2	Another ion of m/z=100 arrives detector at t_{j+1}	Another Toyota passes, and the model is Camry
Outcome 1 Total	After N primary ions, seeing n counts at t_j	After N green lights, seeing k 4Runners
Outcome 2 Total	After N primary ions, seeing n' counts at t_{j+1}	After N green lights, seeing k' Camrys
Expected Outcome 1 and Outcome 2	The expected counts at t_j and t_{j+1} are r and r'	The expected number for 4Runners and Camries are $\bar{r}$ and $\bar{r}'$

The automobile survey is clearly a counting problem, the probability of seeing k 4Runners given the rate is $\bar{r}$ also follows Poisson distribution:

$$P(k \mid \bar{r},N) = \frac{\bar{r}^k e^{-\bar{r}}}{k!} \qquad (3.5)$$

Note that the likelihood of seeing k 4Runners does not depend on seeing k' Camries, it only depends on the actual market share of 4Runner, *i.e.*, the rate of 4Runner $\bar{r}$. Thus, in N green lights, the joint probability of seeing k 4Runners and k' Camrys is:

$$P(k,k' \mid \bar{r},\bar{r}',N) = \frac{\bar{r}^k e^{-\bar{r}}}{k!} \frac{\bar{r}'^{k'} e^{-\bar{r}'}}{k'!} \qquad (3.6)$$

Similarly, in TOF-SIMS, observing n counts at t_j is independent of seeing n' counts at t_{j+1}, the joint probability of getting (n, n') given (r, r') is simply:

$$P(n,n' \mid r,r',N) = \frac{r^n e^{-r}}{n!} \frac{r'^{n'} e^{-r'}}{n'!}$$ (3.7)

This general result for likelihood of independent counting experiments will allow us to find peak in a TOF-SIMS spectrum.

§ 3.4 Finding peaks in TOF-SIMS spectrum

Finding peaks in a TOF-SIMS spectrum follows the same logic as finding peaks in a Gaussian white noise which we demonstrated in last chapter. One needs to put a window on the spectrum. Within the window, compare the hypotheses H_1 and H_0 as in previous chapter. If there is a peak, then do a parameter fitting via the maximum likelihood method.

§ 3.4.1 Model comparison

First, we summarize that the noise is still white because the essence of the Poisson process is that we are counting discrete events that are uncorrelated from one time to another. The probability that we observe n events in the time interval $[t,t+\Delta t]$ is given by the single-step Poisson distribution:

$$p_1\big(n \mid r(t)\big) = \frac{\big(r(t)\big)^n e^{-r(t)}}{n!}$$ (3.8)

where $r(t)$ is the *rate* at time $[t,t+\Delta t]$. We note that the expectation value of n is $\langle n \rangle = \sum np(n \mid r(t)) = r(t)$. The rate will depend upon the local signal strength. If no

signal is present, *i.e.*, there is only dark current, then the rate will be denoted r_0. The signal rides on top of the dark current and it is assumed that the local count rate is directly proportional to the signal:

$$r_i \equiv r(t_i) = r_0 + ax(t_i) \tag{3.9}$$

where $x(t_i)$ describes the known peak lineshape. We can try to estimate both r_0 and a from the same data, or we can estimate r_0 either from a separate time series, or a region of the time series that is far from any peak.

As counts at t_j are independent of counts at any other time, the likelihood function of N data points that we observed, the count sequence $(n_1, n_2, \cdots n_N)$ in the window located at time t_0, is then:

$$p_N(\{n_1, n_2 \cdots n_N\} \mid a, r_0, M_k, t_0) = \prod_{i=1}^{N} \frac{r_i^{n_i} e^{-r_i}}{n_i!}$$
$$= e^{-Nr_0} e^{-a} \prod_{i=1}^{N} \frac{(ax_i + r_0)^{n_i}}{n_i!} \tag{3.10}$$

where it is assumed that x_i has unit area: $\sum_{i=1}^{N} x_1 = 1$.

In order to compute the odds ratio, we need to compute the evidence for both M_0 and M_1, which requires marginalizing the likelihood function over parameters, as shown in Chapter Two. First, let us consider the case where there is no peak in the window, only the dark current due to electronic fluctuation, *i.e.* M_0. We may use a prior distribution

$$p(a, r_0 \mid M_0) = \frac{\delta(a)}{2r_{0\max}}, \qquad 0 < r_0 \le r_{0\max} \tag{3.11}$$

as we know nothing about the dark current, so we choose a uniform distribution $\dfrac{1}{r_{0\max}}$

for r_0. We choose $\dfrac{\delta(a)}{2}$ to be the prior of the amplitude because in M_0 the peak

amplitude is zero and the integration only performed on the positive axis when

marginalizing.

Substituting (3.10) and (3.11) into equation (2.24), one may find the evidence for

M_0. As in Chapter Two, the detailed calculation may be found in the Appendix B, here

we only quote the result:

$$
\begin{aligned}
p(n \mid M_0, t_0) &= \iint p(n \mid a, r_0, M_0, t_0) p(a, r_0 \mid M_0, t_0) \, da \, dr_0 \\
&= \iint e^{-Nr_0} e^{-a} \prod_{i=1}^{N} \frac{(ax_i + r_0)^{n_i}}{n_i!} \frac{\delta(a)}{2 r_{0\max}} \, da \, dr_0 \\
&= \frac{\left(\sum\limits_{i=1}^{N} n_i \right)!}{2 r_{0\max} N^{1 + \sum\limits_{i=1}^{N} n_i} \prod n_i!}
\end{aligned}
\tag{3.12}
$$

We now consider model M_1, where there is a peak in the presence of dark current

in the window. We choose the prior distribution of a and r_0 to be uniformly

distributed:

$$
p(a, r_0 \mid M_1) = \frac{1}{a_{\max} r_{0\max}}, \quad 0 < a \le a_{\max}, 0 < r_0 \le r_{0\max}
\tag{3.13}
$$

Then, according to equation (2.24), the evidence for M_1 is:

$$p(n \mid M_1, t_0) = \iint p(n \mid a, r_0, M_1, t_0) p(a, r_0 \mid M_1, t_0) \, da \, dr_0$$

$$= \frac{1}{r_{max} a_{max} \prod_{i=1}^{N} n_i!} \iint e^{-Nr_0} e^{-a} \prod_{i=1}^{N} (ax_i + r_0)^{n_i} \, da \, dr_0 \qquad (3.14)$$

The integral in (3.14) is over the parameter plane (a, r_0), the term $\prod_{i=1}^{N} (ax_i + r_0)^{n_i}$ is

problematic when the peak amplitude is comparable to the dark current. However, as we are looking for strong evidence of peaks in the spectrum, the product

$\prod_{i=1}^{N} (ax_i + r_0)^{n_i}$ will be dominated by the peak even if ax_i is just slightly larger than r_0

due the power of n_i, and vice versa. Thus we will neglect a very narrow region close to the diagonal of ar_0-plane, and approximate (3.14) by:

$$p(n \mid M_1, t_0) \approx p(n \mid M_1', t_0)$$
$$= p(n \mid only\ a\ peak\ in\ the\ window\ at\ t_0)$$
$$+ p(n \mid only\ dark\ current\ in\ the\ window\ at\ t_0) \qquad (3.15)$$
$$= p(n \mid only\ a\ peak\ in\ the\ window\ at\ t_0) + p(n \mid M_0, t_0)$$

where model M_1' is an approximation of model M_1: there is either only dark current or only a peak (no dark current) in the window. Hence the ratio in equation (2.23) becomes:

$$\frac{p(H_1 \mid n)}{p(H_0 \mid n)} \approx \frac{p(n \mid M'_1, t_0)}{p(n \mid M_0, t_0)} = 1 + \frac{p(n \mid only\ a\ peak\ in\ the\ window\ at\ t_0)}{p(n \mid M_0, t_0)} \qquad (3.16)$$

where $p(n \mid only\ a\ peak\ in\ the\ window\ at\ t_0)$ can be computed with the same fashion as $p(n \mid M_0)$, but with a different prior:

$$p(a, r_0 \mid only\ peak) = \frac{\delta(r_0)}{2a_{max}} \qquad (3.17)$$

This results in (see details in Appendix B):

$$p(n \mid only\ a\ peak\ in\ the\ window\ at\ t_0) = \frac{\left(\prod x_i^{n_i}\right)\left(\sum n_i\right)!}{2a_{max} \prod n_i!} \qquad (3.18)$$

Substitute (3.12) and (3.18) into (3.16), we find the odds ratio to be:

$$\frac{p(H_1 \mid n)}{p(H_0 \mid n)} \approx \frac{Nr_{0max}}{a_{max}} \prod_{i=1}^{N} (Nx_i)^{n_i} + 1 \qquad (3.19)$$

In (3.19), only the first term changes as data $\{n\}$ changes. As we will see soon, the first term dominates when there is a peak and becomes negligible where there is no peak.

§ 3.4.2 Parameter fitting

With above ratio (3.19) computed, if the odds ratio is above some threshold, we infer there is a peak in the window, and the parameter, *i.e.*, the amplitude, can be estimated by maximizing the likelihood function (3.10), using the dark current estimated from the tail of the spectrum where there is no peak. The natural logarithm of the likelihood function (3.10) is:

$$\begin{aligned} L(n \mid a, \hat{r}_0, M_1, t_0) &= \log\left[p(n \mid a, \hat{r}_0, M_1, t_0)\right] \\ &= -N\hat{r}_0 - a + \sum n_i \log(\hat{r}_0 + ax_i) - \sum \log(n_i!) \end{aligned} \qquad (3.20)$$

To maximize it, we set:

$$\frac{\partial L}{\partial a} = -1 + \sum_i n_i \frac{x_i}{\hat{r}_0 + ax_i} = 0 \tag{3.21}$$

Note $ax_i \gg \hat{r}_0$ if there is a peak, the $\hat{r}_0$ in the denominator is then negligible (recall $x_i > 0$, the window lies near the center of the peak and does not extend to the tails):

$$\frac{\partial L}{\partial a} \approx -1 + \sum_i \frac{n_i}{a} = 0 \tag{3.22}$$

which results in:

$$a^* \approx \sum_{k=1}^{N} n_k \tag{3.23}$$

This is our best estimate of the amplitude of the peak. This approximation is *not* good for very small peaks. The second derivative of (3.20) gives the uncertainty of the estimated a^*:

$$\Delta a = \sqrt{\sum_{i=1}^{N} n_i} \tag{3.24}$$

As in the previous chapter, one has to notice that the above calculations have been carried out only for ONE window that is located at some t_0. One has to let the window now slide point by point and carry out above necessary calculations for each window. The best estimation of t_0^* is found by looking for local maximum of

$$L^*(n \mid a^*, \hat{r}_0, M_1, t_0) = L^*(t_0).$$

§ 3.5 Results on simulated data and their interpretation

As we have derived necessary formulas for finding peaks in a TOF-SIMS spectrum, let us first try it on simulated data to see what the results look like and how to interpret them.

In Fig. 3.1 we show a simulated peak with Poisson noise in the first pane. The ideal peak without Poisson noise is a Gaussian with a variance equal to 100, centered at t=1000. In this case, the assumed known peak lineshape x_j is the same Gaussian but runs only from the left half maximum to the right half maximum. This is an ideal case where we know *exactly* what the peak lineshape is. In the next two panes, the log of the odds ratio and the log of the maximized likelihood for each window position are plotted:

$$R(n \mid t_0) = \log\left(\frac{p(H_1 \mid n)}{p(H_0 \mid n)} \right)$$

$$\approx \sum_{i=1}^{N} n_i \log(Nx_i) + \log(N) + \log(\frac{r_{0max}}{a_{max}})$$

(3.25)

$$L^*(t_0) = L^*(n \mid a^*, \hat{r}_0, M_1, t_0) = \log\left[p^*(n \mid a^*, \hat{r}_0, M_1, t_0) \right]$$

$$= -N\hat{r}_0 - a^* + \sum n_i \log(\hat{r}_0 + a^* x_i) - \sum \log(n_i!)$$

(3.26)

As the moving window approaches a peak, three distinct regions can be identified. The behavior of the log of the odds ratio , $R(n \mid t_0)$, and the log of the likelihood, $L^*(n \mid a^*, \hat{r}_0, M_1, t_0)$ for each window, and their interpretation is listed in the following table:

Table 3.2 Interpretations of behavior of $R(n \mid t_0)$ and $L^*(n \mid a^*, \hat{r}_0, M_1, t_0)$

Region	$R(n \mid t_0)$	Interpretation	$L^*(n \mid a^*, \hat{r}_0, M_1, t_0)$	Interpretation
I	Close to zero[1]	No peak	High[2]	It is highly likely that there is only dark current occurs in this region
II	First remains close zero, and then begins to increase.	As the window first encounters the rising edge of the peak, there is not sufficient evidence of a peak, but as the window keeps moving towards the peak, it is more evident there is a peak, so $R(n \mid t_0)$ begins to increase.	Decreases first, but eventually begins to increase.	In this region, first it looks like there is neither a peak, nor only dark current, so $L^*(n \mid a^*, \hat{r}_0, M_1, t_0)$ initially decreases. As the window continues to move to the right, it eventually looks like there is a peak, so $L^*(n \mid a^*, \hat{r}_0, M_1, t_0)$ begins to increase.
III	Reaching a local max	There is a peak in this region	Forms a local spike	The maximum occurs when the window is right on top of the peak

---

[1] H_I has one more parameter to fit than H_0, generally this will improve H_I. The worst situation is that this parameter does not bring in any improvement, thus $p(H_1 \mid n)$ will be always no less than $p(H_0 \mid n)$, so the smallest value of $R(n \mid t_0)$ is zero.

[2] the highest possible of value of $L^*(n \mid a^*, \hat{r}_0, M_1, t_0)$ is also zero, corresponding to the likelihood having a value of 1, i.e., certainty.

To find the best estimate of t_0, fit the peak of $L^*(n\,|\,a^*,\hat{r}_0,M_1,t_0)$ in region III to a parabola:

$$L^*(t_0) = L^*(t_0^*) - \frac{1}{2}L^{*\prime\prime}(t_0 - t_0^*)^2 \tag{3.27}$$

Because $L^*(n\,|\,a^*,\hat{r}_0,M_1,t_0)$ is sharply peaked, and we exponentiate $L^*(n\,|\,a^*,\hat{r}_0,M_1,t_0)$ to get the likelihood, the likelihood should look like an even sharper Gaussian, the center of the parabola is then the center of the Gaussian and its curvature gives our local estimate of the uncertainty in the peak position:

$$L^{*\prime\prime} = \frac{1}{\sigma_{t_0}^2} \tag{3.28}$$

For the simulated peak in Fig. 3.1, we find that $t_0^* = 999.803$, and $\sigma_{t_0^*} = 1.336$. At first glance, one may think that $R(n\,|\,t_0)$ is more sharply peaked than $L^*(n\,|\,a^*,\hat{r}_0,M_1,t_0)$. In fact this is not the case. $R(n\,|\,t_0)$ seems to be sharper simply because its y-axis range is much smaller. It becomes clear that $L^*(n\,|\,a^*,\hat{r}_0,M_1,t_0)$ is actually more sharply peaked in Fig. 3.2 where $R(n\,|\,t_0)$ and $L^*(n\,|\,a^*,\hat{r}_0,M_1,t_0)$ are plotted such that y-axes have the same range. Actually, for $R(n\,|\,t_0)$, we can calculate the "uncertainty" as we calculate σ_{t_0} and it turns out to be 3.287, much larger that $\sigma_{t_0^*}$.

Having seen the $L^*(n\,|\,a^*,\hat{r}_0,M_1,t_0)$ in Fig. 3.1, one might think that since $L^*(n\,|\,a^*,\hat{r}_0,M_1,t_0)$ has this characteristic shape, we can look for the spike in $L^*(n\,|\,a^*,\hat{r}_0,M_1,t_0)$ to find the peak, why should we bother with the odds ratio? The

answer is that we do need the odds ratio for several reasons. First, the odds ratio, which depends on the peak shape, but is independent of parameters, gives a better sense in which region there might be a peak. This can be done by setting a thresholds on $R(n \mid t_0)$ which we will discuss in detail in next chapter. Second, as in Fig. 3.3, when there are two peaks close to each other (in pane a), the characteristic shape of $L^*(n \mid a^*, \hat{r}_0, M_1, t_0)$ will be distorted. It forms a spike between the two peaks (in pane b). It would be wrong if one sees this spike and considers there is a peak, the odds ratio can help us to avoid this mistake (in pane c). Third, as in (3.26), $L^*(n \mid a^*, \hat{r}_0, M_1, t_0)$ depends on the signal amplitude. The spike will become less obvious for small peaks, then finding peaks could rely on the odds ratio by setting an appropriate threshold.

Fig. 3.1 Behavior of log of the odds ratio and log
of the maximized likelihood of a simulated peak.
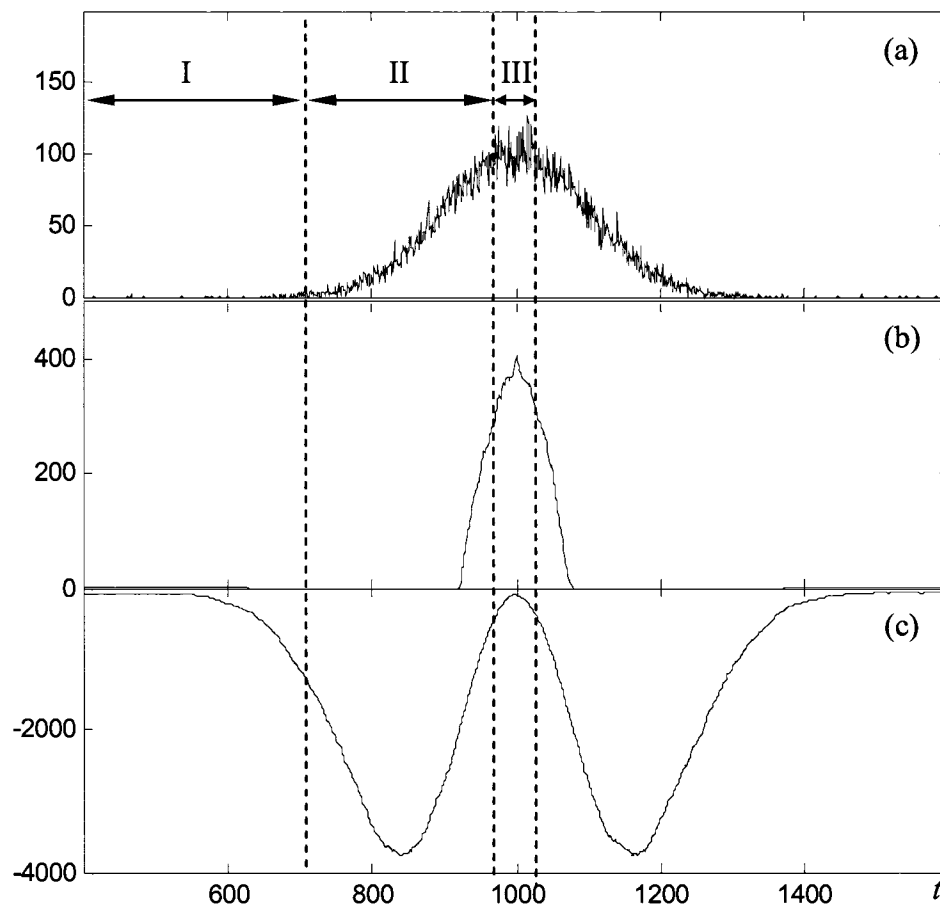
(a) A simulated peak with Poisson noise. The true peak (without
noise) is a Gaussian curve of variance 100 centered at t=1000.
(b) The log of the odds ratio for each window.
(c) The log of the maximized likelihood for each window.

Fig. 3.2 Graphical comparison of the sharpness of
$R(n \mid t_0)$ and the sharpness of $L^*(n \mid a^*, \hat{r}_0, M_1, t_0)$
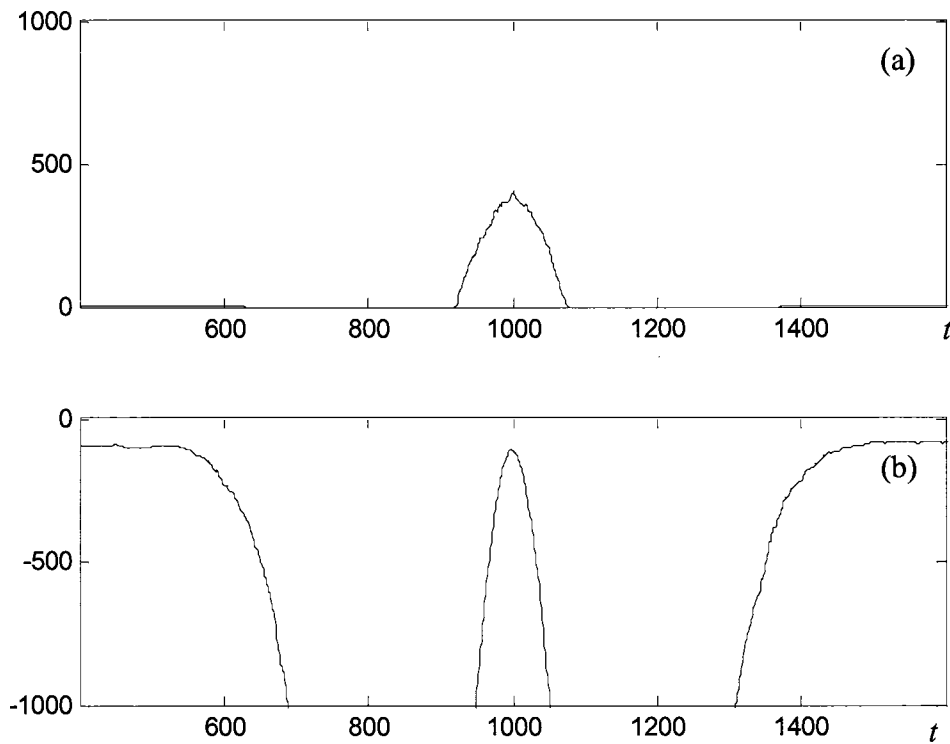
In (a) and (b), $R(n \mid t_0)$ and $L^*(n \mid a^*, \hat{r}_0, M_1, t_0)$ are plotted in a way the y-axes cover the same range. It is evident that $L^*(n \mid a^*, \hat{r}_0, M_1, t_0)$ is sharper, and more sensitive to peak position.

Fig. 3.3 'Phony peak' in the log of likelihood
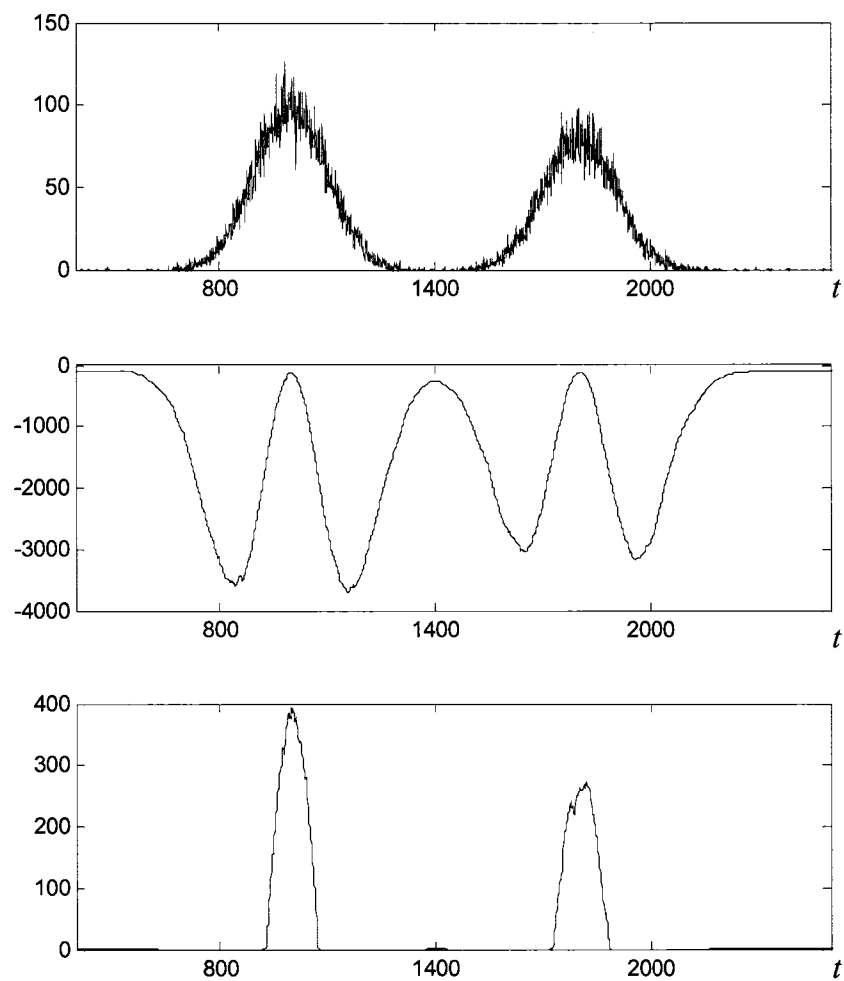
When two peaks close to each other as in (a), $L^*(n \mid a^*, \hat{r}_0, M_1, t_0)$ forms a 'phony peak' between the two actual peaks as in (b), which may mislead someone to think there is peak in the spectrum corresponding to the 'phony peak' in $L^*(n \mid a^*, \hat{r}_0, M_1, t_0)$. The odds ratio clear eliminates this possibility. As in (c), $R(n \mid t_0)$ remains close to zero in the region between two peaks.

Chapter 4

Automated TOF-SIMS Peak Picking

In this chapter, we are going to apply the peak picking method that has been developed in the previous chapter to real TOF-SIMS data. In order to do so, we will need to first solve some problems, namely, derivation of the peak lineshape and threshold setting strategy. We will begin with a description of the TOF-SIMS instrument we used, which will facilitate the derivation of the peak lineshape.

§ 4.1 TOF-SIMS apparatus

The instrument we used for TOF-SIMS experiments is a TRIFT II spectrometer from Physical Electronics. The configuration of the instrument is sketched in Fig. 4.1. It uses a Ga^+ (or Au_n^+) liquid metal ion gun (LMIG) as the primary-ion gun to probe the sample surface and has three quasi-hemispherical electrostatic-sector analyzers (ESA) as part of the secondary ion optics.

The secondary ions are extracted by an immersion lens and fly in a field-free region for some distance and then transfer into the ESA through a transfer lens. These

73

lenses are important for secondary ion detection but the details about how they work are of less interest here and will not be discussed.

Each ESA uses an electric field (E) that is perpendicular to the secondary ion velocity to bend the secondary ion by 90° according to:

$$qE = \frac{mv^2}{r} \qquad (4.1)$$

Let E_k denote the kinetic energy of the secondary ion, it is easy to see that:

$$r = \frac{2E_k}{qE} \qquad (4.2)$$

From (4.2), it is clear that ESA provides energy focusing in the sense that faster ions of the same mass have larger radii and hence follow longer trajectories. More specifically, let E_0 and v_0 be the nominal kinetic energy and velocity of a secondary ion species, it will take L_0/v_0 for the particle to fly a distance of L_0. For a secondary ion of kinetic energy $E=E_0(1+\delta)$, $(\delta<<1)$, in the time L_0/v_0, it will fly a distance of:

$$\sqrt{\frac{E}{2m}}\frac{L_0}{v_0} = \sqrt{\frac{E_0(1+\delta)}{2m}}\frac{L_0}{v_0} = L_0\sqrt{1+\delta} \approx L_0 + \frac{1}{2}L_0\delta \qquad (4.3)$$

That is, to a first order approximation, this ion will fly an extra distance $L_0\delta/2$. The principle for ESA to achieve energy compensation is to let the secondary ion of energy E to travel $L_0\delta/2$ longer by following a loop with a greater radius.

In a TRIFT II system, as shown in Fig. 4.1, three consecutive ESA provide triple focusing and bend the secondary ions by 270°. After the secondary ions leave the third ESA, they fly in a field free region for distance L before they hit the detector.

Sometimes a post acceleration right in front of the detector is needed to boost the kinetic energy and momentum of the secondary ions to increase detection efficiency.

It is clear from Fig. 4.1 that throughout the flight, only electrostatic fields act on the secondary ion, so the Hamiltonian is time independent and conserved:

$$H = \frac{p^2}{2m} + \Phi(r) \tag{4.4}$$

In fact, only when in the acceleration region and in ESA (region I and III in Fig. 4.2) the secondary ion is affected by the electrostatic fields. A graphical sketch of the phase flow that illustrates the ion dynamics is shown in Fig. 4.2. The axial velocity and spatial distributions are illustrated by an ellipse. The ellipse is used just for convenience, one should not interpret the ellipse as multivariate normal distribution.

In region I, secondary ions come off the sample surface with initial axial velocity and spatial distributions, which are represented by the ellipse at the sample surface. The secondary ions are accelerated by an electrostatic potential. Because of the initial spatial distributions, the ellipse will be elongated in velocity after the acceleration. The combination of initial axial velocity and spatial distribution makes the ellipse tilt to the right when the ions enter region II, a field free drift region. In this region, the ellipse will be further tilted to the right as higher speed ions travel more distance in the same time. After the ions enter the ESA (region III), because higher speed results in larger radius, the tilt is eventually corrected. As a matter of fact, the ideal situation is that when the ions exit the ESA, the ellipse would be slightly tilted in the reversed direction so that when the ions fly through region IV, they would hit detector at the

same time, as illustrated in region IV in Fig. 4.2. That is, the ESA does not compensate the velocity distribution of the secondary ions, but compensates for the spatial spread caused by the velocity distribution.

Another thing to note is that, Fig. 4.2 happens in a two dimensional space, only the axial velocity and distance is considered, but ions move in a three dimensional space. The angular distribution of the secondary ion velocity is another source that degrades the mass resolution. In order to achieve high mass resolution, an angular filter can be brought into place in front of the entrance of the first ESA. However, for biological samples, the entrance is often left fully open because of the low secondary ion yield.

Above, we have discussed the TOF-SIMS instrument we used in experiments. The detailed information about how TOF-SIMS experiments were conducted will be given in the next chapter, since here we want to focus on finding peaks in a TOF-SIMS spectrum. We will just "borrow" a typical spectrum from the next chapter, as shown in Fig. 4.3.

Pane (a) in Fig. 4.3 is a typical TOF-SIMS spectrum of pure peptide, Vasopressin, deposited on etched silver foil. The instrument runs at a time resolution of 138ps. The spectrum is plotted as counts $vs.$ time, with about 8.6×10^5 time steps are plotted corresponding to a mass range of 0~2000Da. A close look at the parent peak of Vasopressin is plotted in pane (b). The sequential peaks, marked by I, II...V, are the isotopic pattern of the parent ion due to natural abundance of the constituent peptides atoms, H, C, N, S, $etc.$ Two successive peaks are separated by 1Da, easily resolved by

the high resolution of TOF-SIMS. In mass spectrometry, *mass resolution* is used to describe the power of resolving nearby peaks. It is defined as:

$$mass\ resolution = \frac{m}{\Delta m} \qquad (4.5)$$

where m is the mass of a peak, Δm is the full width at half maximum (FWHM) of the peak.

If the resolution was lower, each of these peaks would become broader and overlap with neighbor peaks resulting in one big, fat peak.

The majority of this work is done in time domain, because in time domain, the arrival time record is equally sampled, *i.e.*, each time step is the same, $138ps$. If we worked in mass domain, because of the quadratic relationship between mass and time, a non-linearity would be introduced; *i.e.*, the mass spectrum would not be equally sampled in mass domain. It is easy to find that the resolution in time domain has the following relationship with the *mass resolution*:

$$\frac{m}{\Delta m} = \frac{t}{2\Delta t} \qquad (4.6)$$

where t is the peak position, Δt is the FWHM in time. From now on, when we mention *resolution* which will be denoted by R, we mean the resolution in time domain:

$$R = \frac{t}{\Delta t} = \frac{2m}{\Delta m} \qquad (4.7)$$

It can be seen in pane (b) in Fig. 4.3 that Δt for peak I, II...V is about 150 time steps, which yields a R of about 4000. One thing to notice is that the R is roughly a

constant across the spectrum. This can be seen in pane (c) in Fig. 4.3, which is a expansion of the early part of the spectrum.

§ 4.2 Application of automated peak picking methods to TOF-SIMS spectra

We have TOF-SIMS data in our left hand and the necessary formulas for finding peaks in a TOF-SIMS spectrum (Chapter Three) in our right hand. Now let us put them together.

There are two issues that need to be dealt with. First, a useful peak lineshape, x_j, which is assumed to be known up to this point, needs to be derived. Second, we said in previous chapters that once the odds ratio was computed, a threshold could be set so that we could identify regions which we are confident contain peaks. We must choose the strategy to set the threshold.

Having resolved these two problems, we then coded our peak picking method in MATLAB. The algorithm takes the TOF-SIMS spectrum as input and gives peak positions, intensities and their uncertainties as output.

§ 4.2.1 Derivation of a peak lineshape

Let us consider ions of a specific m/z. As we mentioned in the previous chapter, the TOF-SIMS is a counting experiment and typically counts the arrival time of one ion at a time. For any given primary ion pulse, there will typically be, at most, one ion of m/z that reaches the detector and the eventual peak shape at m/z spectrum is an

accumulation of many such independent measurements. Since ions generated at the sample surface have their initial distributions in velocity and time, ion optics are used to compensate for these initial spreads to get high mass resolution. An example is the ESA used in the TRIFT II apparatus that we discussed above.

Consider ions of *m/z* after they exit the third ESA and now fly *freely* towards the detector. Had we known the exact velocity and spatial distributions of the ions as they enter region IV in Fig. 4.2, we would be able to derive an exact peak lineshape, which is the distribution of the arrival time. Ultimately, the spreads in velocity and space as ions begin to fly freely come from initial distributions during the secondary ion formation. These distributions include, for example, spatial distribution due to reasons like the surface morphology and finite width of primary-ion pulse, the initial axial velocity distribution, angular distribution, *etc*. If we knew all of them, we would be able to convolute them together and derive an exact peak lineshape since the following on ion dynamics is pretty well defined. However, in reality, these distributions are not well understood.

Though part of these distributions still remains a mystery, some studies have revealed certain properties. For example, it has been shown that the axial velocity distribution of the secondary organic ion peaks at less than 5eV and has a width of a few eV [42, 43]. The distribution due to finite width of the primary ions can be minimized by optimizing the primary-ion gun such that the pulse is short and only minimal degradation of resolution is caused. For example, the primary-ion pulse in our experiment is about 10*ns*.

The important thing to keep in mind is that, because of the presence of ion optics, by the time the secondary ions enter region IV, the distributions in velocity and space are *sharply* peaked around the nominal value.

We have pointed out in the previous section that the ideal situation occurs when secondary ions leave the third ESA, the ellipse is tilted a little bit backwards, which causes the time at which the secondary ion hits the detector to be:

$$t_{total} = t_{enter} + t_{free} \qquad (4.8)$$

where t_{enter} is the time at which the ion enters free flight region IV, t_{free} is the time that the ion takes to fly through region IV.

In Chapter Two we suggested a Gaussian for the velocity distribution $g(v)$ from maximum entropy arguments: when assigning a probability density function (PDF) for the outcome of what will be a prior of a repeated series of identical and independent experiments, one should choose that PDF which maximizes the number of possible outcomes that are consistent with the probability assignment [44]. In other words, the choice of a Gaussian here is based on the desire to minimize bias in the outcome, not on physical arguments requiring 'thermalization' in the ionization process or the nature of the ion optics.

Thus, let us assume that as a secondary ion of a given *m/z* leaves the third ESA and begins to fly freely to the detector, it has a Gaussian velocity distribution which centers at $v_0(m)$ and has a width $\sigma(m)$, the distance of the free flight is *L*. The problem to be solved is the arrival time distribution at the detector. This can be done if we go to

the phase space *(z, v)*, as in Fig. 4.4, we are going to let $z=0$ be the point where the secondary ion begins to fly freely, let the detector be located at $z=L$. Let *f(z,v,t)* be the PDF. Then *f(z,v,t)dvdz* is the probability at time *t* that the ion lies in an infinitesimal neighborhood of the point *(z,v)* in phase space. At the instant *t*=0, the secondary ion enters region IV, we have:

$$f(z,v,t) = f(z = 0, v, t = 0) = \delta(z)g(v) \qquad (4.9)$$

where *g(v)* is the Gaussian distribution:

$$g(v) = \frac{1}{\sqrt{2\pi}\sigma(m)} e^{-\frac{(v-v_0(m))^2}{2(\sigma(m))^2}} \qquad (4.10)$$

This PDF *f(z,v,t)* evolves according to the Fokker-Planck equation, which in free flight is simply:

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial z} = 0 \qquad (4.11)$$

implying that:

$$f(z,v;t) = \delta(z - vt)g(v) \qquad (4.12)$$

In the *(z, v)* phase space, the motion of the particle is represented by a straight line that crosses the origin and has a slope of *1/t*. The cumulative probability *P(t)* that the secondary ions will fly across the detector in time interval *(0,t]* is:

$$P(t) = \int_{L/t}^{\infty} g(v)dv = 1 - \int_{0}^{L/t} g(v)dv \qquad (4.13)$$

Then, the PDF of arrival time is simply:

$$p(t) = \frac{dP(t)}{dt} = -\frac{d}{dt} \int_0^{L/t} g(v) dv$$

$$= -g(L/t) \frac{d}{dt} \left(\frac{L}{t} \right) \tag{4.14}$$

$$= \frac{1}{\sqrt{2\pi}\sigma(m)} \frac{L}{t^2} e^{-\frac{\left(\frac{L}{t} - v_0(m)\right)^2}{2(\sigma(m))^2}}$$

We pointed out previously in equation (4.8) that the ion enters the final free flight region at time t_{enter} which will be different from one ion to another even if both ions are of the same m/z. This means that t_{enter} itself has a distribution and we should convolute it with (4.14). But, it is not clear what the t_{enter} distribution looks like. We find that, as a simple approximation, let t_{enter} be a constant, the resulting peak lineshape still fits well with observed data in the region above the half maximum which be illustrated momentarily. The reason for this may because that tilt of the ellipse at the entrance of free flight region IV is small compared to the spread in space due to the finite pulse width and surface morphology. Thus, we may set t_{enter} to be zero.

However, we do not use (4.14) directly in our calculation as we do not know $v_0(m)$ and $\sigma(m)$ in advance. Instead, we do the following time transformation for the N data points included in the window at t_0:

$$u_i = 1 + \frac{t_j - t_0}{t_0} \times R \times S \tag{4.15}$$

and let:

$$x_i = N \frac{1}{u_i^2} \exp\left[-\frac{1}{u_i^2}\right]$$ (4.16)

The N in (4.16) is a normalization factor such that $\sum x_i = 1$.

The R in (4.15) is the *resolution* in time as in (4.7). As we have already pointed out several times, in our TOF-SIMS spectra, R is roughly the same over the observed mass range. For other types of instruments, such as MALDI, R is roughly constant over the mass "focusing" range where the resolution has been optimized. Outside this region, for MALDI, R begins to decrease.

The S in (4.15) is a scale factor, it is the FWHM of the function $f(u) = \exp(-1/u^2)/u^2$.

There are two remarks about (4.14) and (4.16). First, (4.16) shares the same functional form as (4.14), the peak lineshape described by (4.14) and (4.16) has the same asymmetry property. Second, as shown in Fig. 4. 5, the transform $u_i = 1 + \frac{t_i - t_0}{t_0} \times R$ shifts a peak with maximum near t_0 to a peak with maximum near $u=1$ and compresses it to have a FWHM=1. It is then stretched by the multiplication of S such that it has the same FWHM as the peak lineshape x_i.

§ 4.2.2 Optimizing peak lineshape parameter

First, suppose we 'magically' have a perfect sequence of counts $\{n\}$ that has no noise on it. In other words, $\{n\}$ represents the *true* peak lineshape. The set of numbers $\{x_i\}$ that would exactly be proportional to the *true* peak lineshape would maximize the

likelihood function (3.10) or, equivalently, the natural log of the likelihood function (3.20). This can be shown as following. We wish to find the set of $\{x_i\}$ that maximizes (3.20) subject to the constraint that $\sum x_i - 1 = 0$. Let us use Lagrange multiplier:

$$h = L(n \mid a, \hat{r}_0, M_1, t_0) - \lambda \left(\sum_{i=1}^{N} x_i - 1 \right)$$
$$= -N\hat{r}_0 - a + \sum n_i \log(\hat{r}_0 + ax_i) - \sum \log(n_i!) - \lambda \left(\sum_{i=1}^{N} x_i - 1 \right)$$

(4.17)

Take the derivative of h with respect to x_j, note ax_j is far larger than r_0 when there is a peak, and that x_j only goes from the left half-max to the right half-max of the peak. We then have:

$$\frac{\partial h}{\partial x_i} \simeq \frac{n_i}{x_i} - \lambda = 0$$

(4.18)

Equation (4.18) has to be true for all $i=1,2...N$, which implies that the best set of $\{x_i\}$ is the one that is proportional to $\{n_i\}$. Such a set of $\{x_i\}$ would maximize the likelihood. This makes sense as in Chapter Three, where the local rate is assumed to be $r_i = ax_i + r_0$.

The peak lineshape in (4.15) and (4.16) is derived from a simple approximation, it is not the exact peak lineshape, but Fig. 4.5 (and later in Fig. 4.8) shows it is a good approximation in the sense that it fits well with the data above the half maximum. The peak lineshape will deviate from data as we go farther away from the center. Thus, in the following calculation, when we use (4.16) as our peak lineshape to identify peaks

in a spectrum, we restrict ourselves to only use $\{x_i\}$ that are above the half maximum

of $f(u) = \exp(-1/u^2)/u^2$, as in Fig. 4.5.

The $\{x_i\}$ that we derived is determined by the parameter R. One should use an

optimal R which leads to a set of $\{x_j\}$ such that, when the window is right on top of a

peak, the likelihood function is maximized, or, in other words, $\{x_j\}$ is made to be

proportional to observed data as closely as possible. However, as we do not know in

advance at which window position t_0^* the window is 'at the peak', we sum the

likelihood in the region around t_0^*, as in Fig. 4.6. In doing so, we are, in some sense,

marginalizing the likelihood function against t_0, which makes us less insensitive to t_0^*.

Strictly speaking, we are not marginalizing since the data varies as the window shifts,

however, in the region around t_0^*, data only changes a few points, this introduces only

a small effect. Also, since around t_0^*, $n_1 \approx n_N$, the changes are not great.

As we set out to find an optimal R, we notice that, within a TOF-SIMS spectrum,

the R is roughly the same but not exactly a constant, which we have demonstrated in

previous section (Fig. 4.3). So, we manually picked about one hundred "obvious"

peaks in a spectrum. By "obvious" peaks, we mean peaks that are well above noise

level, i.e. are of large intensities. For one selected peak, we let R vary from 3100 to

6000 in steps of 100, and found the one value of R that maximizes above the sum of

likelihood. This was treated as the 'individual optimal' for that peak. This was done

again for all selected "obvious" peaks and we then took the mean of all 'individual

optima' as the 'spectrum optimal' R for the spectrum. Second, for similar samples, the

R does not change much from sample to sample, especially for samples that are taken only days apart, as long as the machine is running stably. Thus, in principle, one can use the 'spectrum optimal' for all similar samples. Yet, to be careful, we took an extra step. Our samples were taken over three successive days. We performed the above optimization on two spectra from each day, and got six 'spectrum optimal' R, the mean of the six 'spectrum optimal' R was then used in (4.15) and (4.16) to compute the peak lineshape.

§ 4.2.3 Threshold setting strategy

Having settled on the peak lineshape model parameter, we are then left to fix a strategy for choosing the threshold. We first note that the actual data observed in a window $\{n_1, n_2...n_N\}$ is just one realization of a process with rates $\{r_1, r_2...r_N\}$, where r_i is the assumed local rate for some peak amplitude a, and shape x_i, as in equation (3.9). Thus, the calculated log of the odds ratio in equation (3.25), $R(n|t_0)$, will deviate from its expected value $< R(n|t_0)>$ due to sampling fluctuations. If we can find $< R(n|t_0)>$, we can then set a threshold by comparing $R(n|t_0)$ with $< R(n|t_0)>$. To find $< R(n|t_0)>$, we will need an ensemble of $\{n\}$. Just like in the univariate case, we need to draw a series of samples to estimate a random variable's mean. However, only one realization of $\{n\}$ was observed in the spectrum, so, let us try a mental experiment. Imagine we have an ensemble of M realizations of $\{r_1, r_2...r_N\}$. We use n_i^j to denote the counts at t_i in j^{th} realization. We may, approximately, write:

$$n_i^j \cong ax_j + \xi_i^j \sqrt{ax_i} \quad i = 1, 2 \cdots N; \ j = 1, 2 \cdots M \qquad (4.19)$$

where ξ_i^j is a random variable from Normal distribution with mean zero and variance one. In writing n_i^j in the form of (4.19), we preserve two important properties of n_i^j as a Poisson random variable. Its mean and variance, when averaged over j, we would have:

$$\langle n_i \rangle = ax_i$$
$$\mathrm{var}(n_i) = \langle (n_i - ax_i)^2 \rangle = ax_i \qquad (4.20)$$

Substitute (4.19) into (3.25), and then average over j. We would have:

$$\langle R(n \mid t_0) \rangle = \sum_{i=1}^{N} ax_i \log(Nx_i) + \log(N) + \log(\frac{r_{0\max}}{a_{\max}}) \qquad (4.21)$$

and the corresponding variance:

$$\mathrm{var}\left(R(n \mid t_0)\right) = \sum_{i=1}^{N} ax_i \left(\log(Nx_i)\right)^2 \qquad (4.22)$$

A careful look at (4.21) and the log of odds ratio (3.25), shows that they all are proportional to the peak amplitude. Higher peak amplitudes would lead to larger odds ratios, which means stronger evidence of the presence of a peak. A threshold on odds ratio implies a requirement on the minimum peak amplitude. In order to be confident there is a peak in the window, there must be sufficient counts in the window. If there is a total of a counts observed in the window, the $\{n\}$ observed in the window may come from two distinct local rates, one is $r_i = ax_i + r_0$, i.e. associated with a peak; the other

one is $r_i' = \dfrac{a}{N} + r_0$, *i.e.* associated with a larger dark current (pure noise). For the latter,

we can similarly compute $\langle R'(n \mid t_0) \rangle$ and $\mathrm{var}\big(R'(n \mid t_0)\big)$:

$$\langle R'(n \mid t_0) \rangle = \frac{a}{N} \sum_{i=1}^{N} \log(Nx_i) + \log(N) + \log\left(\frac{r_{0\max}}{a_{\max}}\right) \qquad (4.23)$$

$$\mathrm{var}\big(R'(n \mid t_0)\big) = \sum_{i=1}^{N} \frac{a}{N}\big(\log(Nx_i)\big)^2 \qquad (4.24)$$

The requirement on the amplitude is then that, a has to be large enough such that $\langle R(n \mid t_0) \rangle$ and $\langle R'(n \mid t_0) \rangle$ have to be well separated compare to $\mathrm{var}\big(R(n \mid t_0)\big)$ and $\mathrm{var}\big(R'(n \mid t_0)\big)$.

§ 4.3 Results

Applying our peak picking algorithm allows accurate and automatic peak finding in a TOF-SIMS spectrum. The calculation for one spectrum of $\sim 10^6$ time series takes about 3 to 4 minutes on a 650 Hz Sun Fire V120 server. Peak positions and amplitudes, together with their uncertainties are reported. In Fig. 4.7, the work flow of peak finding is shown. In Fig. 4.8, a few example peaks are overlapped with estimated peak amplitudes times our peak lineshape that locate at the estimated peak positions. One may see that stand-alone peaks, peaks with satellites, and peaks that are partially overlapped are detected. In the next chapter, we discuss an automatic algorithm for aligning the peaks that have been identified in multiple spectra. This "auto-alignment"
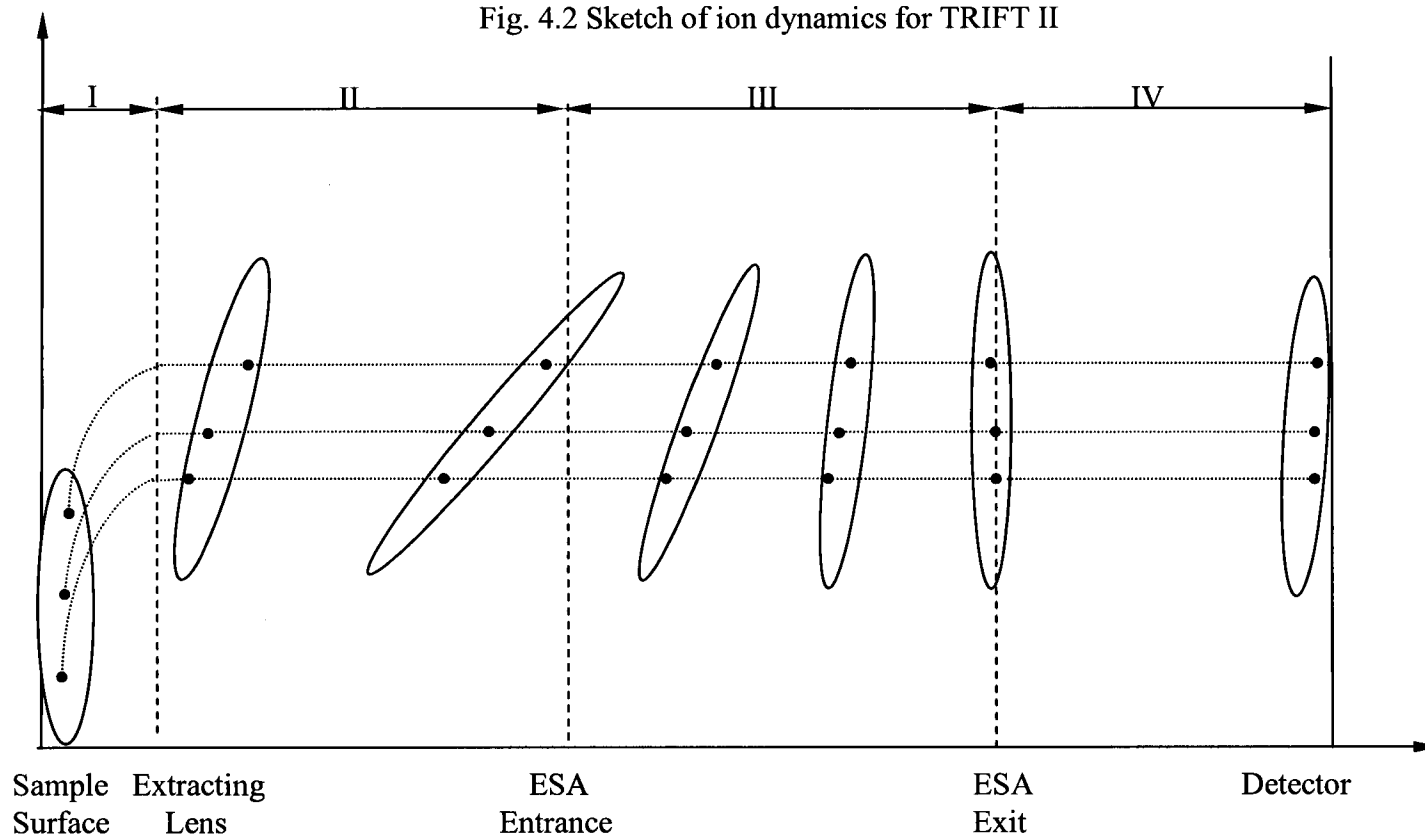
step is required before we can compare two different spectra to look for similarities or

dissimilarities.

Fig. 4.1 Layout of TRIFT II configuration

Immersion
Lens

Transfer
Lens

Ga^+

L
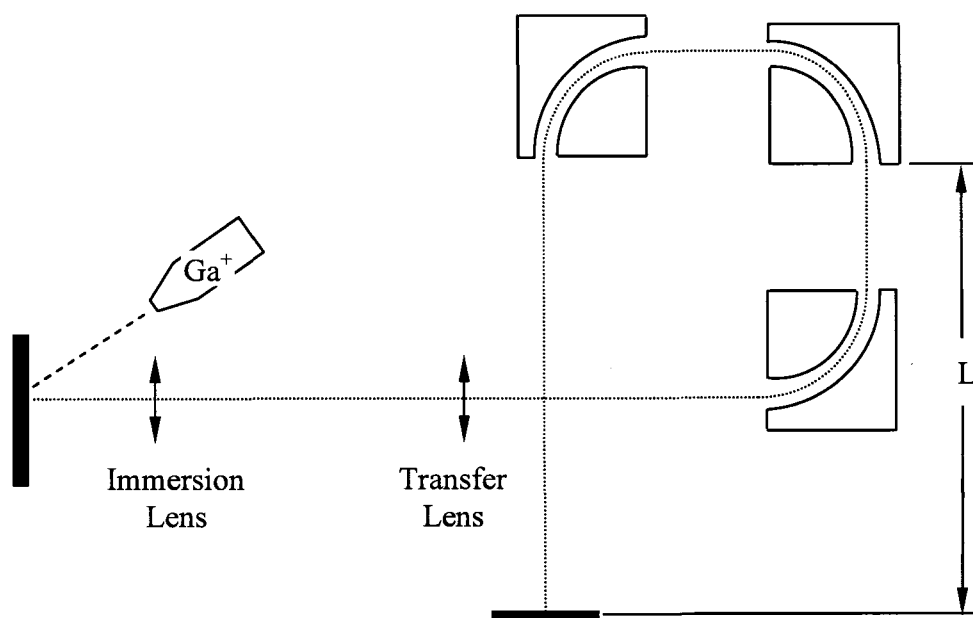
A TRIFT II system has three ESA's and is equipped with a Ga^+ primary ion gun.

Fig. 4.2 Sketch of ion dynamics for TRIFT II

Ions of the same m/z are generated on the sample surface with initial velocity and spatial distributions. They spread even more after the acceleration region (I) and the first free flight region (II). ESA (III) compensate the spread so that ions exit the ESA at the same time and then fly freely (IV) to the detector

Fig. 4.3 A typical TOF-SIMS spectrum of Vasopressin
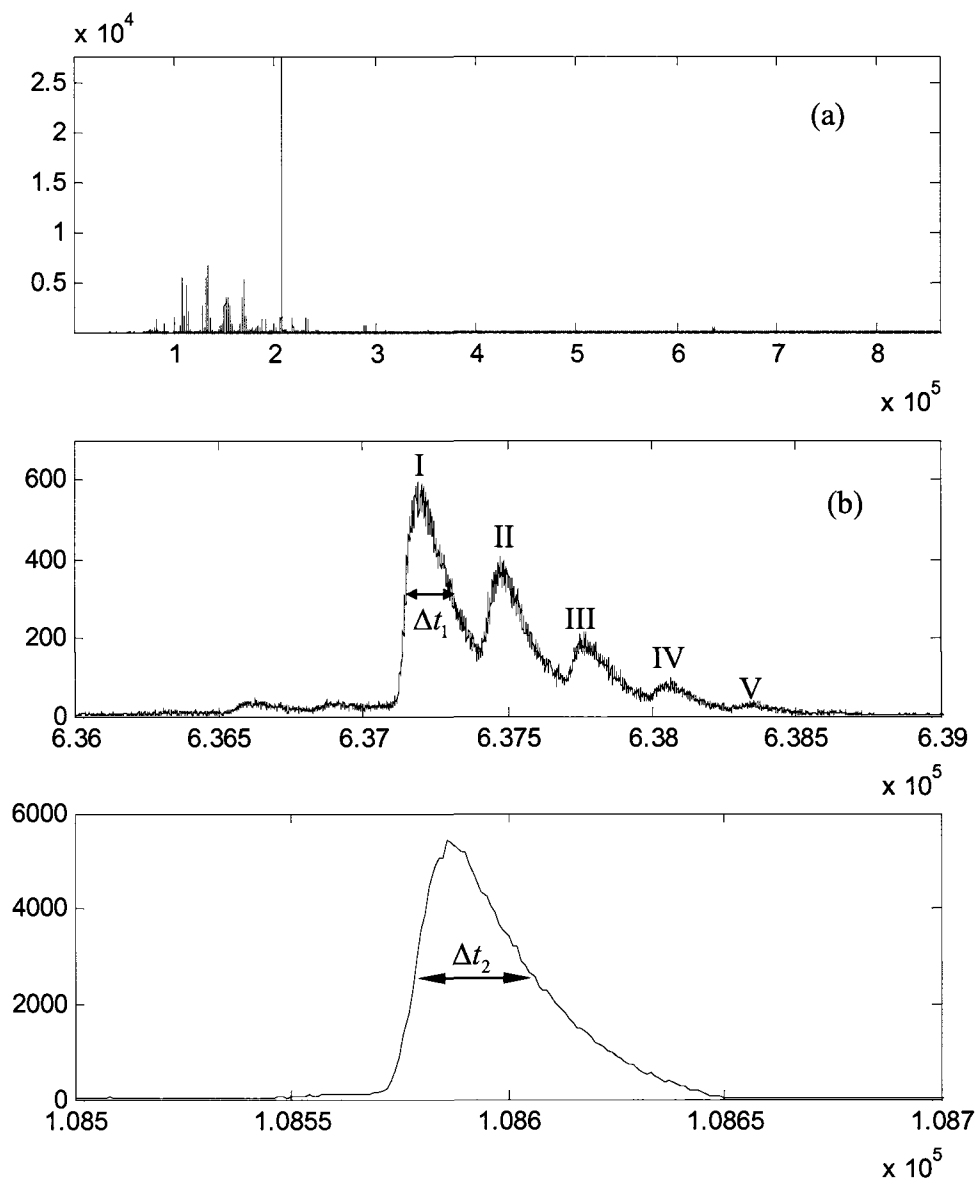
Δt_1 in (b) is about 150, Δt_2 in (c) is about 25, both give $R \approx 4000$.

Fig. 4.4 Ions that can hit the detector at time t
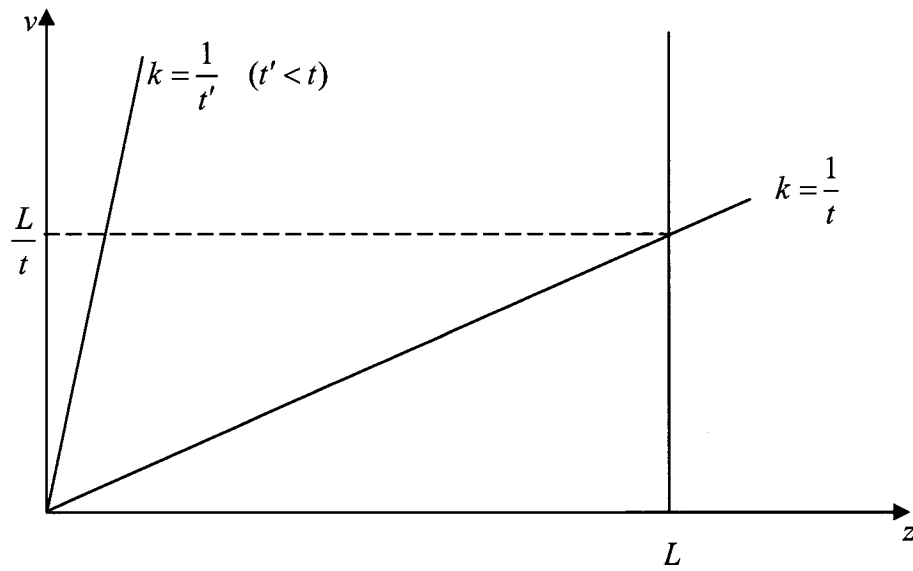
Only the portion with velocity above L/t could pass the line $z=L$ at time t

Fig. 4.5 Illustration of transformation of equation (4.15).
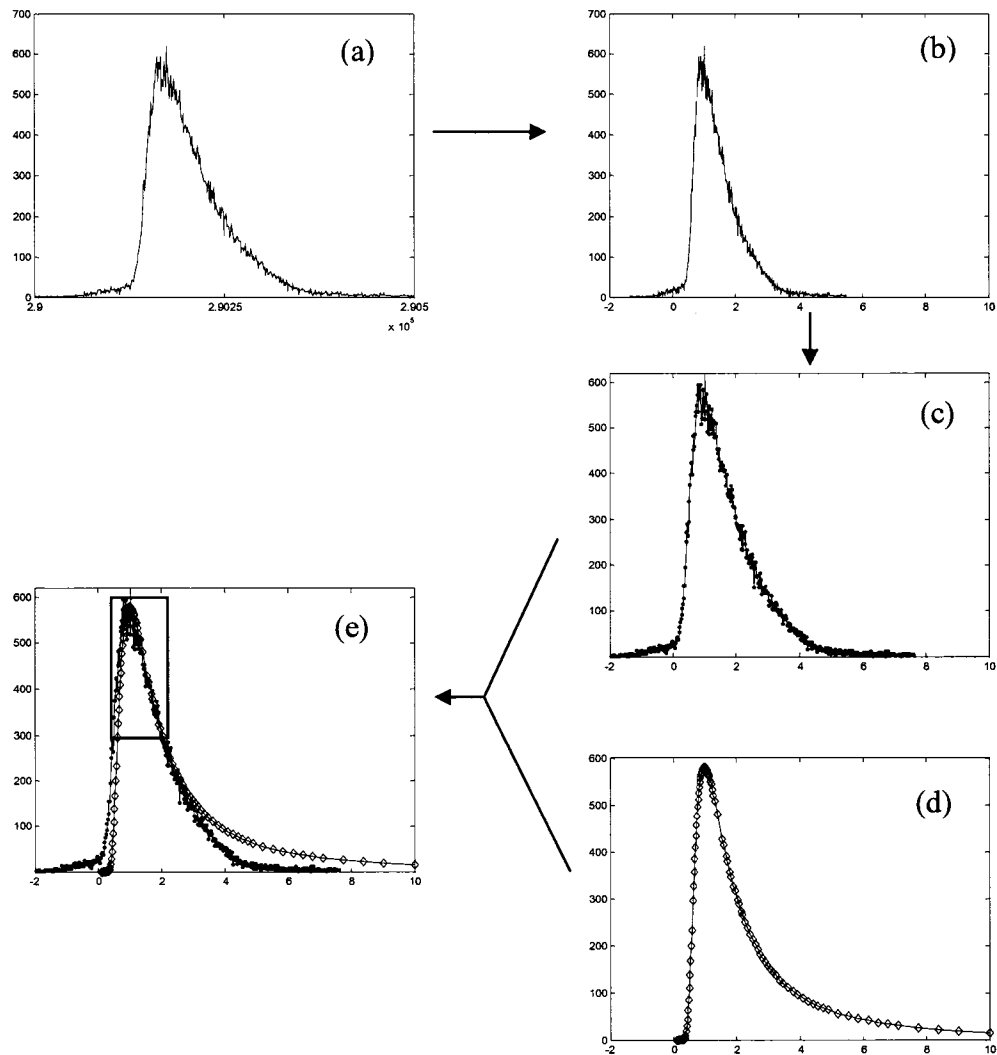
(a) a peak;

(b) transform time t to $u_i = 1 + \dfrac{t_i - t_0}{t_0} \times R$, now the peak maximizes

around 1 and has FWHM=1;

(c) stress (b) by multiply S;

(d) peak lineshape;

(e) overlap of (c) and (d)

Fig. 4.6 Optimize the resolution R
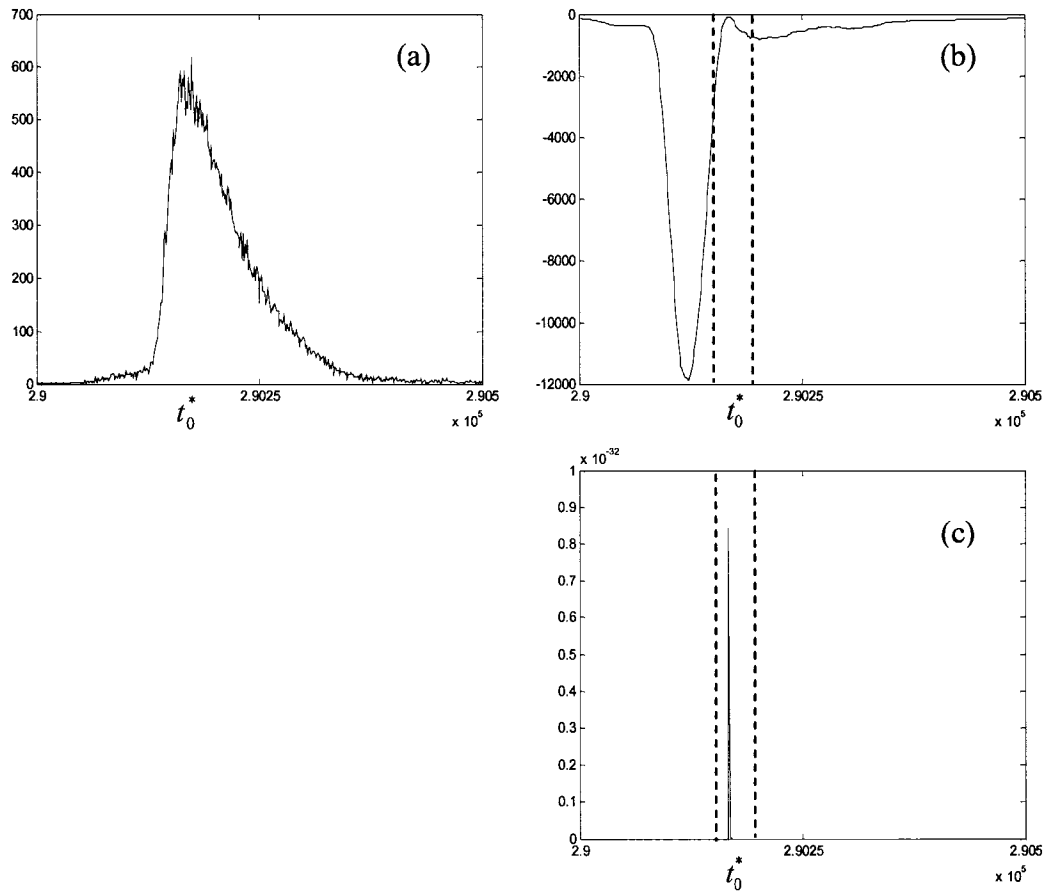
(a). A peak that maximized around t_0^*;

(b). Corresponding log(Likelihood) around t_0^*;

(c). Likelihood (take exponential of (b))around t_0^*, if sum over the region within the vertical bars, major contribution come from points around t_0^*.

Fig. 4.7 Work flow of finding peaks in a TOF-SIMS spectrum

(1) Read in spectrum

↓

(2). Put a window of width N which isolates N data points

↓

(3). For these N data points, compute the log of odds ratio and log of likelihood, $i.e.$, equation (3.25) and (3.26)

↓

(4). Let the window slide one data point by one point. Window width will increase so that it always runs from left half max of a peak to the right half max of the peak

↓

(5). Do calculation in (3) for each of these windows

↓

(6). Find the regions where the odds ratio is above the threshold

↓

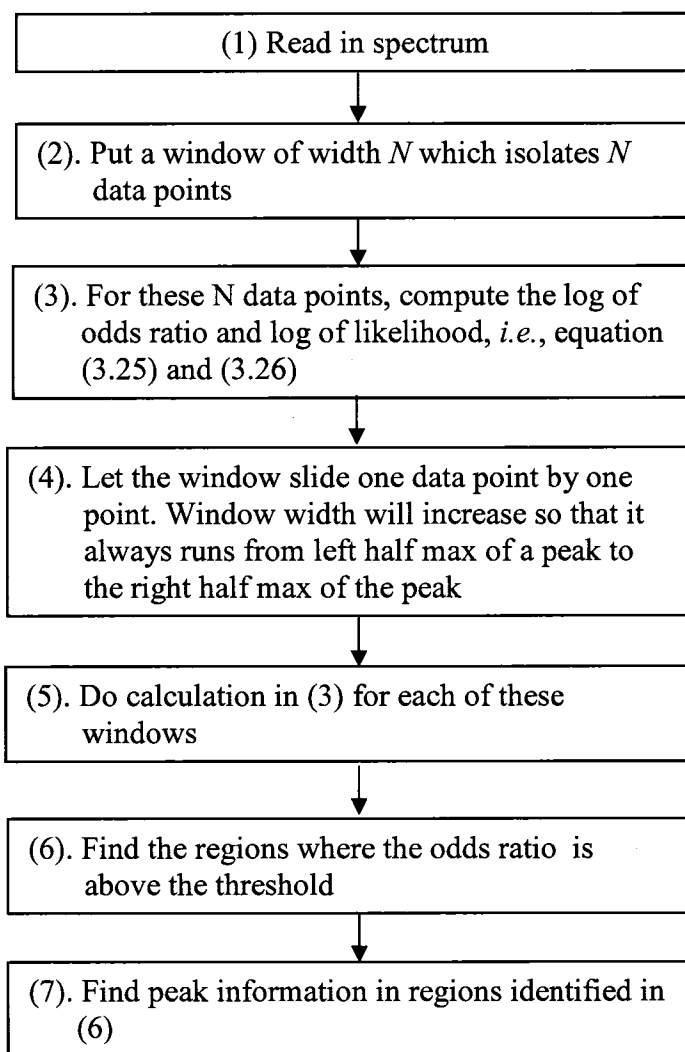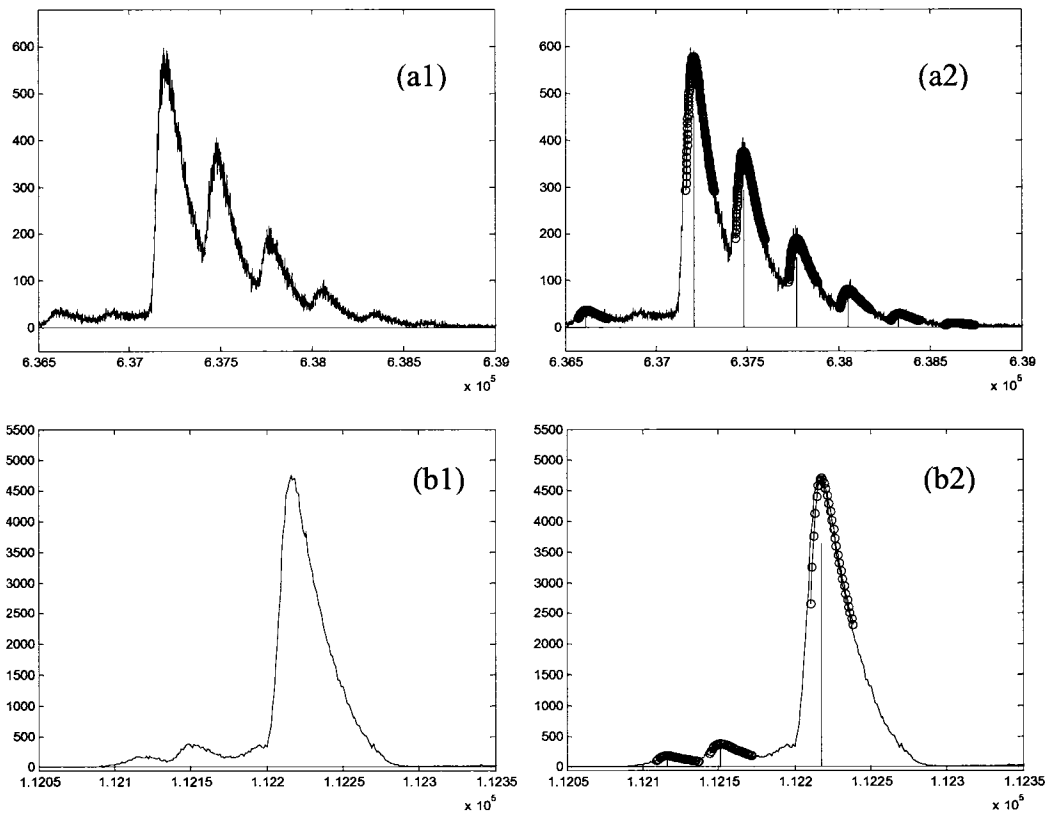| (7). Find peak information in regions identified in (6) |
|---|

Fig. 4.8 Detected peak and the fitted curve overlapped with spectrum

(a1) the isotopic pattern peaks of Vasopressin;

(a2) overlap with estimated peak amplitudes times peak lineshape and locate at estimated peak positions (circle);

(b1) and (b2) are similar to (a1) and (a2) but in a much low mass region.

Chapter 5

Searching for Patterns in

TOF-SIMS Mass Spectra Part I:

Peak Alignment and Feature Selection

In this chapter, we first describe the collection of some TOF-SIMS spectra upon which we apply the method developed in previous chapters. Instead of collecting spectra from arbitrary samples, we carefully designed an experiment which would be a model for the use of TOF-SIMS as a biomarker discovery tool. Specifically, we created various mixtures of the peptides angiotensin (Ang), somatostatin (Som) and vasopressin (Vas), the goal being to infer the concentration ratio from the TOF-SIMS spectra. We will achieve this by introducing multivariate analysis. We will first find patterns in the spectra of each individual peptide that provides discriminant ability and then infer the concentration ratio by examining the strength of patterns in the mixture spectra. In this chapter, we will only describe how the spectra were collected and how to convert the spectra into appropriate data format so we can apply multivariate

98

analysis. That is, we will introduce a peak alignment strategy and a feature selection procedure. The discriminant analysis and inference of concentration ratio from the spectra will be left to next chapter.

§ 5.1 Experiment details

Motivation:

This work was done under more challenging conditions in which tools for the discovery of biomarkers associated with disease are under development. The idea is that disease processes would lead to unusual protein levels in the body which can be screened by high throughput profiling of tissue or blood by means of a mass spectrometer. The spectra would be collected from both sick and healthy people, the goal is then to find "biomarkers" that discriminate between these groups [3, 4, 5, 6, 7, 8, 45]. The biomarkers are characteristic peaks in the mass spectra that provide discrimination capability and are likely to be a combination of a few peaks. It is important in the biomarker discovery process that one needs to first identify significant peaks that are above the noise level and to reliably find their positions and amplitudes. In doing so, a spectrum that contains tens or hundreds of thousands of data points is compressed to far fewer data points that represent only meaningful peaks and hence reduces the dimensionality. It is the information of peak position and peak intensity that is used for finding different patterns between healthy and diseased

patients and our peak-picking algorithm is designed to automatically extract this information from the spectra. Simply using the raw spectrum may lead to "false discovery". False discovery, in turn, can lead to a significant waste of resource as any potential biomarkers are tested in further clinical trials or drug developments.

Another potential application lies in the imaging of biomaterial surfaces, where the surface is rastered and each rastering position would be a pixel in the final image. At each pixel, a full mass range spectrum is acquired. By detecting the peak positions and intensities in each spectrum, one can select a specific peak (a chemical species, for example, cholesterol) and build an image of surface abundance of that peak [46, 47, 48, 49, 50, 51, 52].

In previous work, we have been able to register peptides that were deposited on etched silver. Because of the destructive nature of SIMS, both fragments and molecular ions were observed. As each different peptide has its own specific amino acid composition, it would not surprise us that each peptide would have its own fragmentation pattern, which would appear in the spectrum as repeatable peak distributions at different m/z with different intensities. Finding these patterns would enable us to discriminate between different parent peptides, just as finding patterns in a sick person's serum spectrum would help us diagnose disease. It would also be interesting if we can quantitatively infer the concentration ratio of peptides by checking the relative strength of patterns in a spectrum from a mixture sample. This is a model system for the more complicated biomarker discovery problem.

Experimental design:

The experiment was designed with help from Dr. Michael Trosset of the William and Mary Mathematics Department. A series of ten samples of three different peptides (Ang, Som, Vas), consisting of individual peptides (pure samples), binary mixtures and ternary mixtures, were made according the concentration ratio shown in Fig. 5.1. This design helped us to test if there was a linear relationship between the relative pattern strength and concentration ratio, which will be discussed in next chapter.

Materials:

Three peptides were used for the experiment. They were Angiotensin II Human, Arg^8-Vasopressin and Somatostatin. Vasopressin was purchased from Sigma Aldrich and the other two were purchased from American Peptide Company. A diagram of the amino acid sequence of each peptide is shown in Fig. 5.2. All samples came in sealed glass containers, kept in a refrigerator, and were used without further purification. Silver substrate foils were purchased from Alfa Aesar.

Sample preparation:

The silver substrate was cut into small pieces ($0.6 \sim 1 \ cm^2$) then etched/cleaned by immersion into 25% nitric acid. Previous experiments had shown that etching for 4 minutes would expose fresh silver without introducing significant surface roughness, which causes a decrease in the secondary ion yield. During etching, a stirring bar was

used to disperse gas bubbles forming on the surface. These substrates were then rinsed with deionized water once, ultrasonicated for 5 minutes, then rinsed with deionized water 3 times.

Three peptides were dissolved in separate beakers in deionized water and each had a concentration of $0.15 mg/ml$. They were shaken for 5 minutes to completely suspend or dissolve them. They were then mixed together according to table 5.1 to get mixture samples of the desired concentration ratios (Fig. 5.1).

Table 5.1 Volume of pure peptide solutions used for mixture sample preparation

Concentration ratio (A:S:V)	Angiotensin (ml)	Somatostatin (ml)	Vasopressin (ml)
1:0:0	3	0	0
0:1:0	0	3	0
0:0:1	0	0	3
1/2:1/2:0	1.5	1.5	0
1/2:0:1/2	1.5	0	1.5
0:1/2:1/2	0	1.5	1.5
1/3:1/3:1/3	1	1	1
2/3:1/6:1/6	2	0.5	0.5
1/6:2/3:1/6	0.5	2	0.5
1/6:1/6:2/3	0.5	0.5	2

Each mixture solution was then incubated with one etched silver foil at room temperature for 40-60 minutes. After that, the silver foil was taken out of solution, shaken to get rid of extra liquid droplets on the surface and blow-dried with nitrogen gas. They were then stored under nitrogen in dessicators before the acquisition of

SIMS spectra. All preparation was done exclusively using glassware that was cleaned with nitric acid and deionized water to minimize contamination by organic polymers from plastic containers.

Spectra acquisition:

A TRIFT II spectrometer (Physical Electronics) with a Ga^+ primary-ion gun was used to acquire spectra. The spectrometer was operated at a pressure of $(1.5-3.0) \times 10^{-10}$ *Torr*. The primary ion had an energy of 15 keV. The scanning area was $200 \times 200 \mu m^2$. The primary-ion dose was computed to be *3.47 x 10^{11}* ions/cm^2, within the static regime. For each piece of silver, thirty areas that spread over the silver were scanned, *i.e.*, thirty spectra for each mixture solution.

§ 5.2 Multivariate analysis

The great analytical power has made TOF-SIMS a powerful surface characterization tool. It gives a wealth of chemical information and structural information about the analyte. Its capability of detecting ions in parallel over a large mass range generates spectra with a large amount of data rapidly. This is particularly so when organic materials are analyzed. For heavy mass polymers and biological samples such as proteins, it is always the case that due to the inherent destructive property of SIMS, many fragment ions are generated as a result of energetic primary

ion impact, resulting in from a few hundreds to thousands of peaks. Sometimes, the intact molecular ions are not observed at all, in which case all information is embedded in the large number of fragment peaks. It is then a real challenge to transform these data into useful information.

Multivariate analysis, which has been commonly used in statistical pattern recognition and chemometrics, has been increasingly used to help to interpret SIMS spectra and SIMS imaging to provide insights into the spectra [41, 53, 54].

Multivariate analysis, as opposite to univariate analysis, deals with the situation where more than one measurement is performed on one sample. In multivariate analysis language, each measurement is called a variable. Each multivariate object naturally lives in a multi-dimensional space. The "dimensionality" is the number of variables. For example, in a peptide TOF-SIMS spectrum, there are a few hundreds peaks, each of these peaks would be a variable. The dimensionality is the number of peaks. It is usually true that some of these variables are correlated. In a TOF-SIMS spectrum, this is demonstrated by the fact that many ions are fragments of the same parent ions but from different fragmentation pathways and thus carry related information. The advantage of multivariate analysis is that it is potentially capable of untangling this information and thus simplifying the spectrum.

§ 5.3 Peak alignment

In general, if we have p measurements on each of n objects, multivariate analysis

usually begins with a $n \times p$ data matrix X:

$$X = \begin{pmatrix} x_{11} & x_{12} \cdots x_{1p} \\ x_{21} & x_{22} \cdots x_{2p} \\ \cdots\cdots\cdots\cdots \\ x_{n1} & x_{n2} \cdots x_{np} \end{pmatrix} \tag{5.1}$$

in which each row is an object (or sample, or patient), and each column is a variable.

In order to perform any multivariable analysis on TOF-SIMS spectra, we have to fill the detected peaks into such a data matrix in such a way that each row corresponds to a spectrum and each column corresponds to a peak of certain m/z. However, the spectra collected are not naturally aligned. Peaks of the same m/z will arrive at the detector at slightly different m/z positions due to factors such as fluctuations in extraction voltage, as shown in Fig. 5.3, the dot and dashed lines are two spectra of the same sample in the same mass range. It is clear that there are small shifts between the two spectra. The same peak arrives at the detector at slightly different time, but they should be treated as peaks of the same m/z. This indicates a peak alignment is necessary.

Comparing two TOF-SIMS spectra, we observed that there was a linear trend in the relative shifts of peaks of approximately the same m/z in two spectra with respect to peak position, as shown in Fig. 5.4. Let p_1 be the peak position of a peak of m/z in the first spectrum, p_2 be the position of the peak of the same m/z in the second spectrum, we may fit the linear trend into the form:

$$p_1 - p_2 = ap_2 + b \tag{5.2}$$

This constant shift b is likely due to the surface morphology or triggering error while the linear trend could be caused by a small change in acceleration voltage. This can be shown as follows. The velocity of an ion after acceleration is:

$$e\Phi = \frac{1}{2}mv^2 \qquad (5.3)$$

substitute that $v = L/t$, where L is flight length after acceleration, t is the time of flight:

$$e\Phi = \frac{mL^2}{2t^2} \qquad (5.4)$$

take the derivative:

$$ed\Phi = -\frac{mL^2}{t^3}dt \qquad (5.5)$$

this implies that:

$$\frac{\Delta\Phi}{\Phi} = -2\frac{\Delta t}{t} \qquad (5.6)$$

It is easy to see that the shift is proportional to the fluctuation in the voltage:

$$\Delta t = -\frac{t\Delta\Phi}{2\Phi} \qquad (5.7)$$

Thus, we can shift and scale p_2 to match it to p_1:

$$p_2' = (1+a)p_2 + b \qquad (5.8)$$

This is done globally for all peaks that lie "close", with a and b optimized by the least square minimization of $\sum(\Delta p)^2$. The absolute value of residue shifts after match p_2 to p_1 is shown in Fig. 5.5, in which, one may see that the residue shift is of

the order of one in 10^5. The result of this transformation is shown in Fig. 5.6, where the same region as in Fig. 5.3 after this transformation is plotted. One can see that, though the transformation is very simple, peaks are aligned well. Having aligned the two spectra, we then take the mean of p_1 and p_2' as the peak position. We then use it as p_1 to align a third spectrum, and so on. This is done for thirty spectra in each sample, and a peak list for the sample is generated after this procedure. Figure 5.7 shows a heat map of aligned peaks. In the figure, there are thirty horizontal lines, each line corresponds to a spectrum, and each vertical line represents a peak position, the intensity is plotted in log scale and is color code according the color bar on the right. A simple check of overall alignment effects of this procedure is to look at peaks of two silver isotopes, which are of the highest intensities in a spectrum, i.e., the brightest line in the heat map. It is no doubt that they are aligned across these thirty spectra after the alignment.

We first applied this alignment procedure on the spectra collected from ten samples on samplewise base. Then, the sample peak lists of three individual peptides (Ang, Som, and Vas) were used to generate a "master peak list." All other sample peak lists were aligned to this master peak list.

Having finished the peak alignment procedure, TOF-SIMS spectra are converted into the data matrices of the form (5.1). Before this data matrix is used for further analysis, those peaks that only occasionally appear in a few spectra are discarded. More specifically, we look into the thirty spectra of one individual peptide sample, say

Ang, find those peaks that appear in at least twenty spectra. These peaks are candidates to be retained. We do the same for the other two pure peptides samples, the union of all candidates are the peaks retained for further analysis.

§ 5.4 Feature selection

As stated in the beginning of this chapter, after constructing the data matrix, the first thing we need to do is to find patterns that separate the three individual peptide samples. These patterns can then be used to either classify new spectra, or, in our experiment, to infer the concentration ratios of mixture samples. If one thinks of the spectra of these individual samples as vectors in a multi-dimensional space, then the question is, do they cluster on a sample bases? That is, do the Ang vectors cluster separately from Som, *etc.*

The data matrix generated after previous peak alignment still has quite a number of variables. Even after discarding rare peaks, there are still over 400 peaks. One might welcome such a large number of variables because, intuitively, one would think that adding a new variable will provide additional information about the samples. With this additional information, one could build a better classifier. By better, we mean the misclassification rate is smaller. The worst case is that the additional information that the new variable provided has nothing to do with the sample, then the misclassification rate would remain unchanged. However, in practice, it is not the case. Adding in new variables will initially improve the performance of the classifier,

if these variables are not from noise. But at some point, adding in variables will then lead to a degradation in the performance when the sample size is finite. This is the so-called "small sample size" problem, and it is an active research topic in statistical pattern recognition. This problem is beyond the scope of current project. For more information, one may refer to [55, 56, 57, 58]. However, it is important to realize here that, as we have only thirty spectra for each individual sample, a step of choosing an optimal variable set with significant fewer than 30 peaks is required.

In order to find such variable set, a criteria is needed. What do we mean by the "optimal set"? We chose Wilks' Λ test [59], which measures the ratio of the within-group covariance to between-group covariance for each choice of variable combinations.

Suppose we have data matrices $X_1, X_2...X_k$ from k groups of p variables, in the ith group, there are n_i objects, *i.e.* X_i is a $n_i \times p$ matrix. Let $x_{i,j}$ be the jth objects in the ith group, $\bar{x}_i$ be the mean of ith group:

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i,j} = \begin{pmatrix} \bar{x}_{i,1} \\ \bar{x}_{i,1} \\ \vdots \\ \bar{x}_{i,p} \end{pmatrix} \tag{5.9}$$

Let $\bar{\bar{x}}$ be the over all mean of all objects from all groups:

$$\bar{\bar{x}} = \frac{\sum_{i,j} x_{i,j}}{\sum_{i=1}^{k} n_i} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} \tag{5.10}$$

Let W and B be the within group sum of squares and between group sum of squares, respectively:

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x}_i)'(x_{i,j} - \overline{x}_i) \tag{5.11}$$

$$B = \sum_{i=1}^{k} n_i (\overline{x}_i - \overline{\overline{x}})'(\overline{x}_i - \overline{\overline{x}}) \tag{5.12}$$

It can be shown that $W + B$ equals the total sum of squares:

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \overline{\overline{x}})'(x_{i,j} - \overline{\overline{x}}) = W + B \tag{5.13}$$

The Wilks' Λ is testing the hypothesis: $H_0 : \overline{x}_1 = \overline{x}_2 = \cdots \overline{x}_k$:

$$\Lambda = \frac{|W|}{|W + B|} \tag{5.14}$$

If the k groups do not show any clustering in p-dimensional space, then $\overline{x}_1 = \overline{x}_2 = \cdots \overline{x}_k = \overline{\overline{x}}$, and $\Lambda = 1$. On the other hand, if k groups cluster in distinct regions, then Λ would be smaller than 1. Thus, the goal of selecting features with discriminating capability is to find a set of features that results in the smallest Λ.

If we desire d variables, one can, in theory, do an exhaustive search in all possible $\dfrac{p!}{(p-d)!d!}$ combinations to find the best set of d. However, this is often computationally expensive. To avoid this difficulty, we implement the first part of what is called McHenry's variable selection [60]. The strategy is to find a subset of d
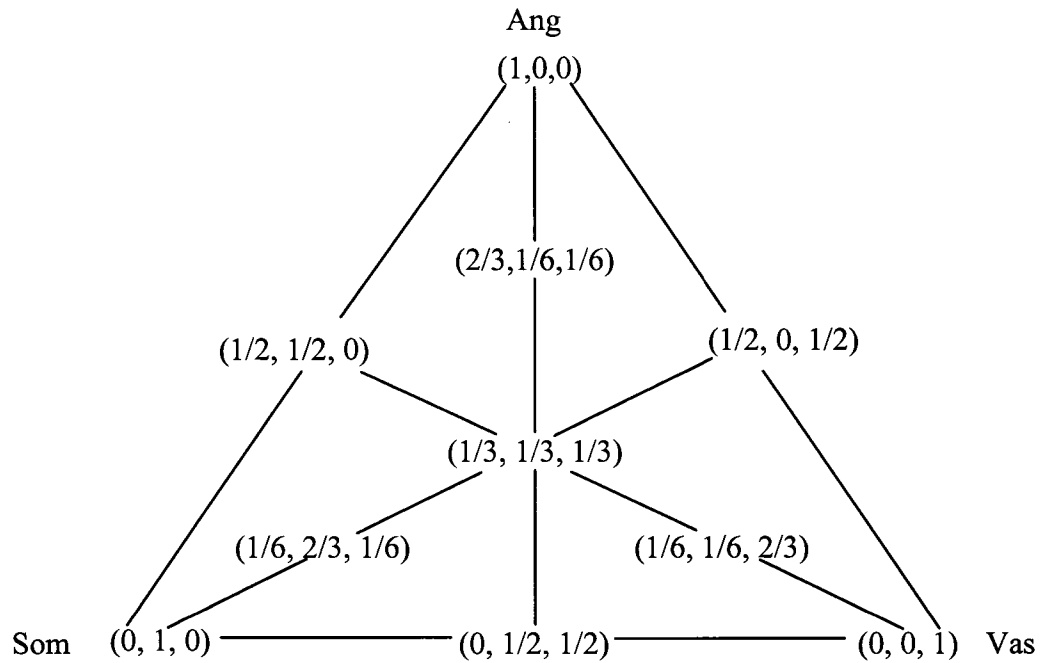
variables such that it is impossible to replace any of the d variables to decrease Wilks' Λ. The work flow is demonstrated in the flow chart in Fig. 5.8.

To test the discriminant capability of selected features, a randomization test is performed. The randomization tests the hypothesis that a selected feature (a combination of peaks) provides no discriminating capability, $i.e.$, spectra do not cluster in the space. For the interested reader who wants more detailed descriptions about randomization tests, see [61]; for an example of randomization test, see [62]. Here we only describe how we do this test. First, we pool spectra of individual peptide samples together in the order that 1-30 are Ang, 31-60 are Som, 61-90 are Vas. We reassign the chemical label by randomly permuting the order, then treat the first random thirty as 'Ang', the second random thirty as 'Som', and the last random thirty as 'Vas'. If the original spectra cluster, this random permutation will break the clustering, resulting in a much larger Λ compared to the Λ before shuffling. This random permutation is done 1000 times, and the distribution of resulting Λ's is shown in Fig. 5.9. The original Λ is also indicated in Fig. 5.9. Clearly the original Λ is far smaller then those after permutation, indicating a good discrimination capability.

This summarized how our variables for classification are selected. In the next chapter, we describe, finally, the results of the classification on our mixtures.

Fig. 5.1 The mixture map

Ang
(1,0,0)

(2/3,1/6,1/6)

(1/2, 1/2, 0)   (1/2, 0, 1/2)

(1/3, 1/3, 1/3)

(1/6, 2/3, 1/6)   (1/6, 1/6, 2/3)

Som   (0, 1, 0) ——————— (0, 1/2, 1/2) ——————— (0, 0, 1) Vas

The three vertices represent three pure ingredients Ang, Som and Vas respectively. The vertical distances from vertices to their opposite edges, represent the abundance of the ingredients.
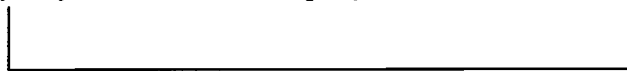
Fig. 5.2 Sequences of three peptides used in the experiment

Angiotensin:

Asp-Arg-Val-Tyr-Ile-His-Pro-Phe

Somatostatin:

Ala-Gly-Cys-Lys-Asn-Phe-Phe-Trp-Lys-Thr-Phe-Thr-Ser-Cys
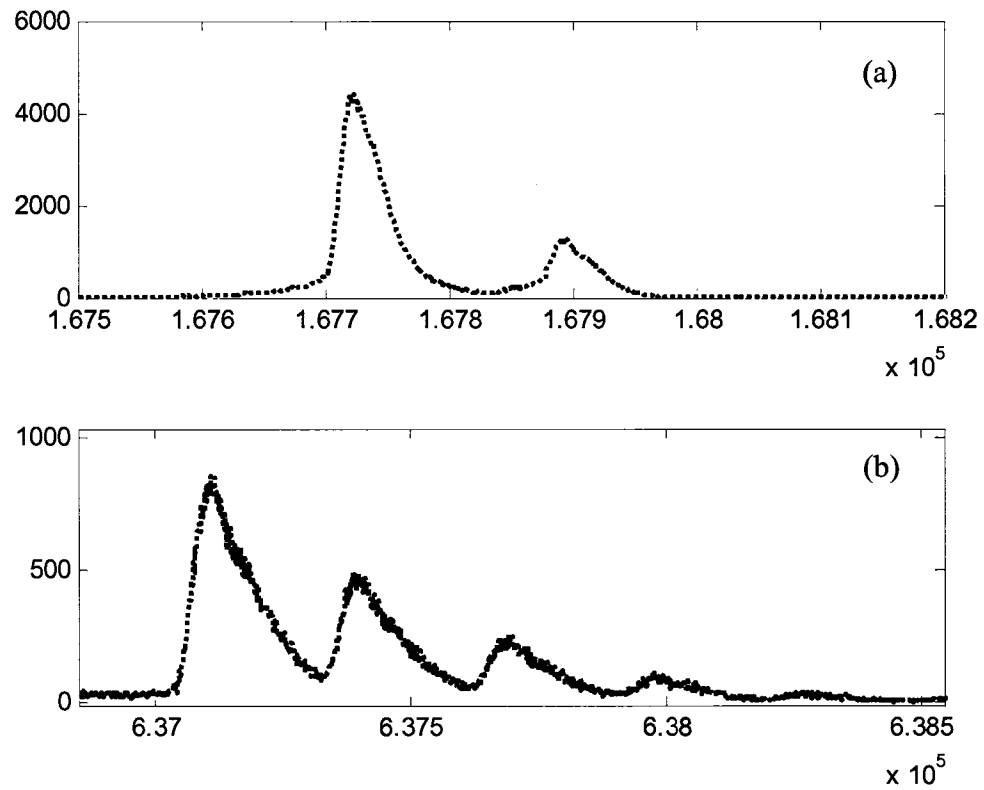
Vasopressin:

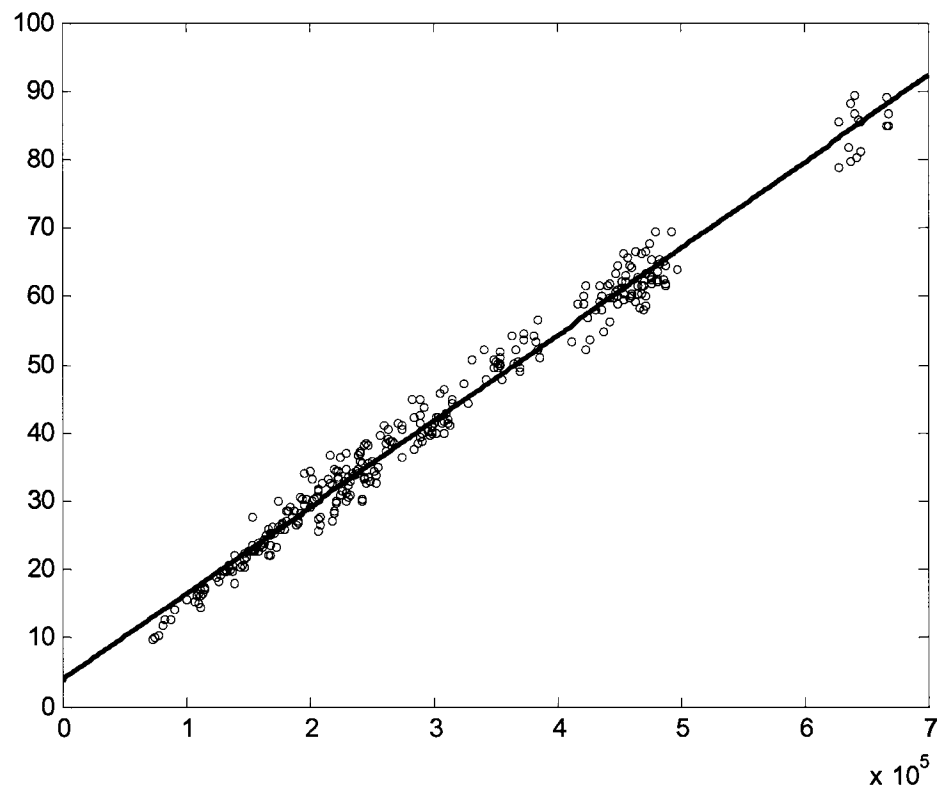Cys-Tyr-Phe-Gln-Asn-Cys-Pro-Arg-Gly

Fig. 5.3 The shift of the same peak in different spectra

The same peak arrives at the detector at different times in two different spectra (black dots and gray dashes). The shift in early part of the time series (a) is smaller than the shift in the late part of the time series (b).

Fig. 5.4 Linear trend of the peak shift

The shift between the same peak but in two different spectra gets larger as the arrival time of the peak gets later. The relationship between the shifts and arrival times can be approximated by a linear relationship.

Fig. 5.5 The absolute value of residue shift
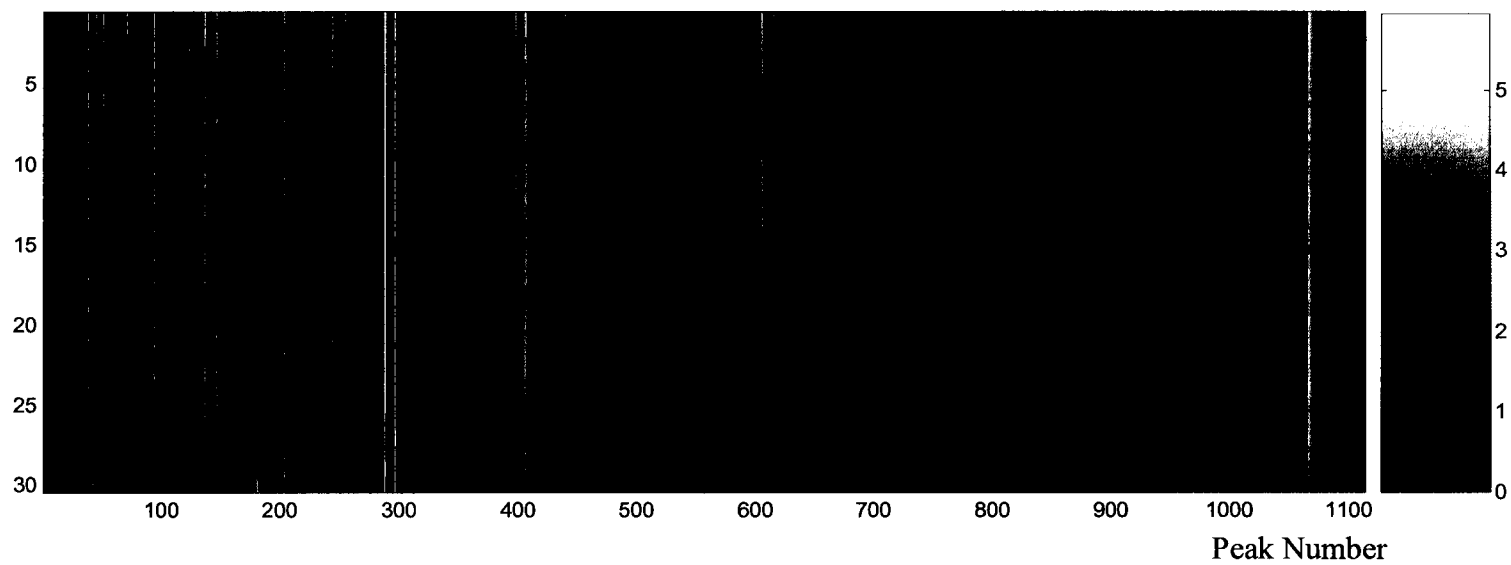after correcting the linear trend
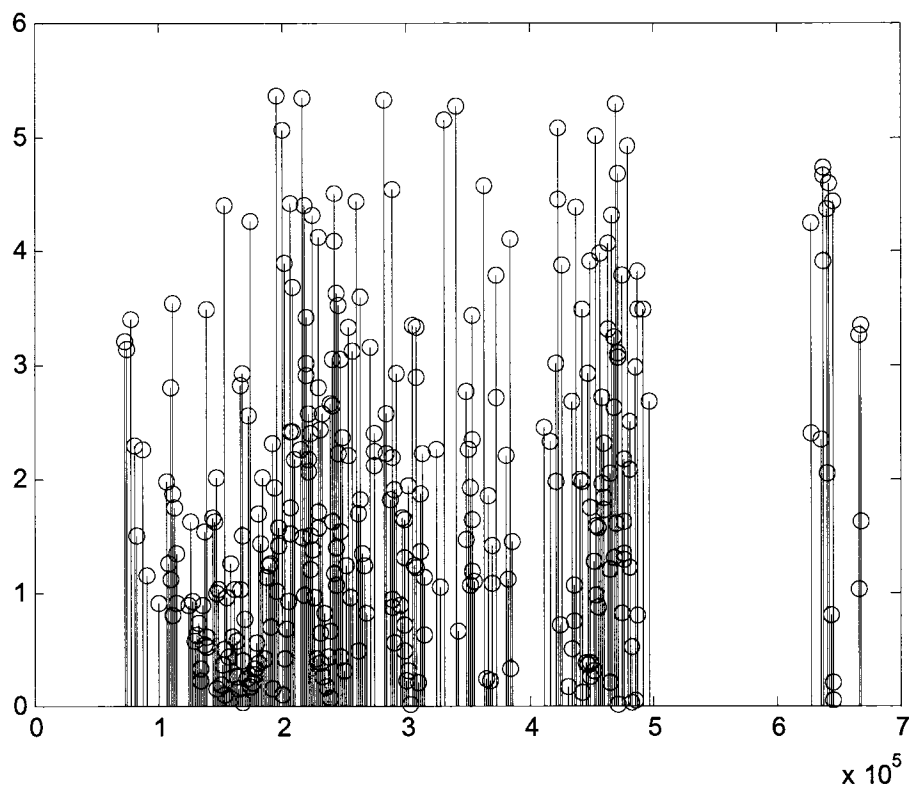
After correcting the linear trend in the shift between the same peak
in two spectra, the residue is only about 1 in 10^4 to 1 in 10^5

Fig. 5.6 Effects of correcting the linear trend on spectra
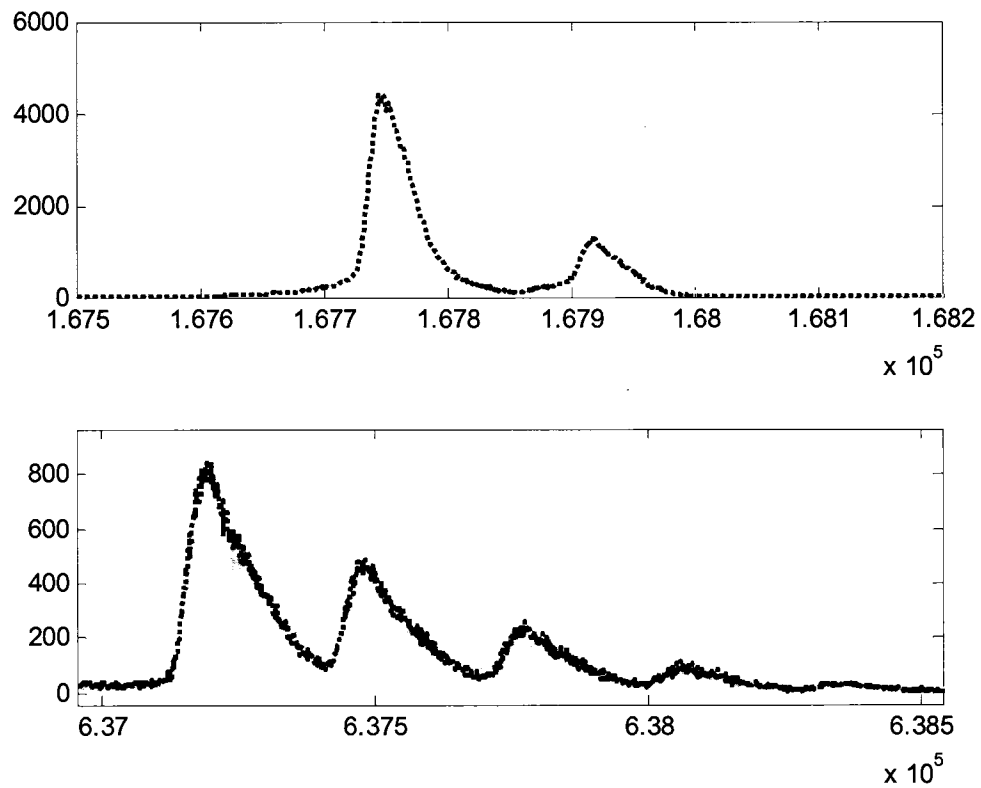
The same region as in Fig. 5.3, peaks are matched much better after correcting the linear trend.

Fig. 5.7 Heat map of 30 aligned spectra

Thirty spectra are plotted as thirty horizontal lines. Each vertical line corresponds to a peak. The log10 of the intensities are color coded according the color bar on the right.

Fig. 5.8 The work flow of McHenry's variable selection

General terminology: a subset is a set that contains a number of variables that are considered as candidates which lead to the smallest Wilks' Λ. The remaining set contains all other variables that are not in the subset just mentioned.
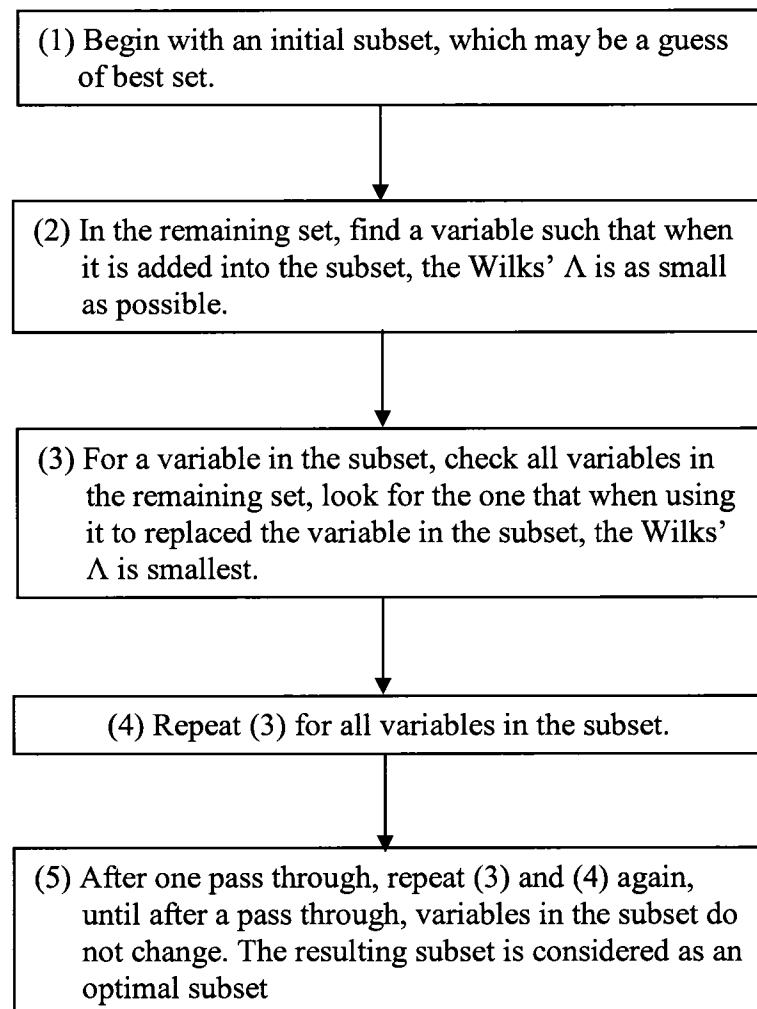
(1) Begin with an initial subset, which may be a guess of best set.

(2) In the remaining set, find a variable such that when it is added into the subset, the Wilks' Λ is as small as possible.

(3) For a variable in the subset, check all variables in the remaining set, look for the one that when using it to replaced the variable in the subset, the Wilks' Λ is smallest.

(4) Repeat (3) for all variables in the subset.

(5) After one pass through, repeat (3) and (4) again, until after a pass through, variables in the subset do not change. The resulting subset is considered as an optimal subset
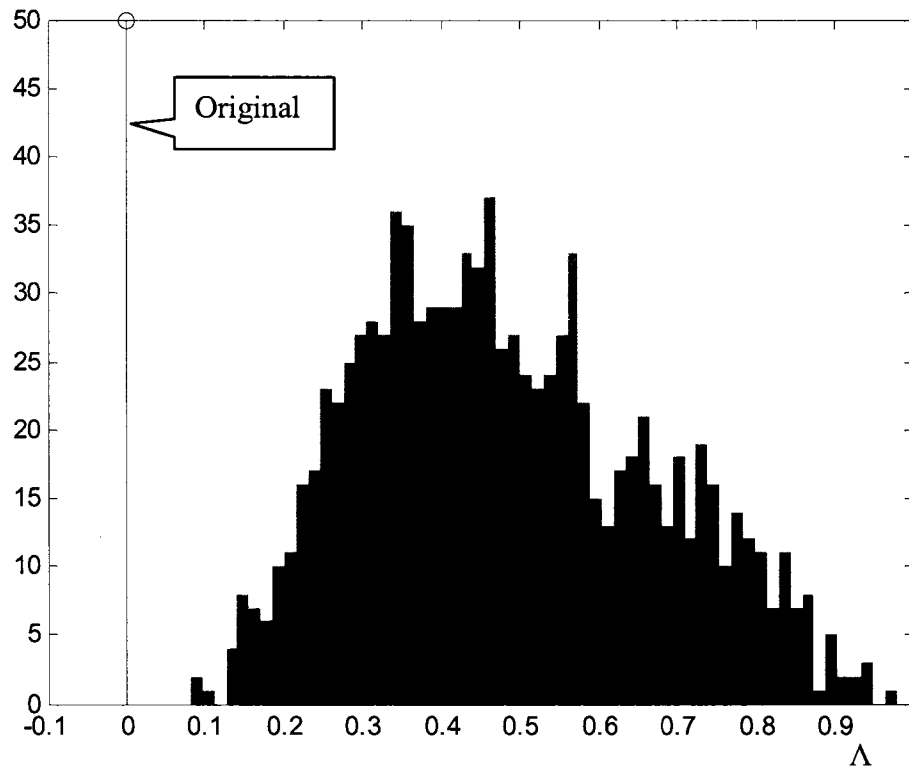
Fig. 5.9 Histogram of Wilks' Λ from randomization test

After randomization, spectra do not cluster, Wilks' Λ is much larger than before randomization, the spike marked 'Original'.

Chapter 6

Searching for Patterns in

TOF-SIMS Mass Spectra Part II:

Classification and Mixture

In this chapter, we will first introduce linear discriminant analysis (LDA) which projects high-dimensional data into a low-dimensional sub-space such that different groups are separated as much as possible. With our TOF-SIMS data, we first apply LDA to three individual peptide samples, $i.e.$, looking for the projecting directions, and show the discrimination among them. Once the projection directions were determined, we then applied it to all mixture samples.

§ 6.1 Introduction to linear discriminant analysis

Linear discriminant analysis is one of the most common methods for discrimination. Suppose we have sample sets X_1, X_2...X_g (note that X_i is an $n_i \times p$

matrix as in the previous chapter) that are drawn from g ($g \geq 2$) different populations, *i.e.*, g groups. The purpose of discrimination is to allocate an individual object x to one of these g groups. There are many ways to do this. For example, if the probability density for each of these groups is known, this can be done by the maximum likelihood approach. Furthermore, if the prior probability of each group is known, then a Bayesian discriminant that incorporates this prior would be a good choice [63]. However, in practice, these probabilities are not always known. LDA provides a way of performing discrimination without these probabilities. The assumption it makes is that all groups share a common covariance structure, but have different mean values.

The idea of LDA is easy to illustrate in the case of two groups. Let X_1 ($n_1 \times p$) and X_2 ($n_2 \times p$) be the data matrices for group 1 and group 2, as illustrated in Fig. 6.1:

$$X_1 = \begin{bmatrix} x'_{1,1} \\ x'_{1,2} \\ \vdots \\ x'_{1,n1} \end{bmatrix} \quad X_2 = \begin{bmatrix} x'_{2,1} \\ x'_{2,2} \\ \vdots \\ x'_{2,n1} \end{bmatrix} \tag{6.1}$$

LDA looks for a linear transformation from x to $a'x$ (where a is vector of p elements) such that the ratio of the between-group sum of squares to the within-group sum of squares is maximized.

The linear transformation:

$$z_{g,i} = a'x_{g,i} = \sum_{j=1}^{p} a_j x_{g,i,j} \quad g = 1, 2 \quad i = 1, 2 \ldots n_g \tag{6.2}$$

gives a scalar $z_{g,i}$. The goal is to find an array (projector) a such that the squared standardized difference is maximized:

$$\frac{(\bar{z}_1 - \bar{z}_2)^2}{s_z^2} = \frac{\left[a'(\bar{x}_1 - \bar{x}_2)\right]^2}{a'S_{pl}a} \tag{6.3}$$

where S_{pl} is the pooled covariance.

The solution to this problem is:

$$a = S_{pl}^{-1}(\bar{x}_1 - \bar{x}_2) \tag{6.4}$$

An object x will be assigned to group 1 if:

$$a'(x - \mu) > 0 \tag{6.5}$$

where $\mu = (\bar{x}_1 + \bar{x}_2)/2$. Otherwise, it will be assigned to group 2.

This result is exactly the same as the maximum likelihood approach, assuming a multivariate normal distribution for the probability density function for each group.

In the cases of several groups, Fisher suggested to look for a vector a that maximizes:

$$\frac{a'Ba}{a'Wa} \tag{6.6}$$

where B and W are the between group sum of squares and within group sum of squares defined in the previous chapter.

It turns out that such a vector a is the eigenvector associated with the largest eigenvalue of the matrix $W^{-1}B$ [59, 63]. It defines the direction (the first discriminate coordinate) on which different groups are most separated. In general, matrix $W^{-1}B$

has min(p, g-1) non-zero eigenvalues, the eigenvector corresponding to the second largest eigenvalue gives the second discriminant coordinate, and so on. One should notice that since $W^{-1}B$ is in general nonsymmetric, its eigenvectors are not necessarily orthogonal, which means that the discriminant coordinates may be correlated.

LDA is naturally related to Wilks' Λ in the sense that Wilks' Λ may be written as:

$$\Lambda = \frac{|W|}{|W+B|} = \prod \frac{1}{1+\lambda_i} \tag{6.7}$$

where, λ_i is the ith eigenvalue of $W^{-1}B$. That is, the feature selected according Wilk's Λ criteria automatically gives best discrimination capability when LDA is used.

§ 6.2 Other work

Before we apply LDA to the TOF-SIMS data that we collected, we want to mention that Castner's group at the University of Washington has done extensive work using TOF-SIMS as an analytical tool to characterize proteins adsorbed onto different surfaces. Part of their work involves using TOF-SIMS to analyze binary and ternary mixtures of proteins and using the "partial least squares" method to predict the relative concentration ratio[54, 64]. The result is then compared with the concentration ratio measured by other means such as ^{125}I-radiolabel method and shows good consistency. The proteins used in their work are bovine serum albumin, bovine immunoglobulin C and human fibrinogen.

Most experiments they performed are binary mixtures where the total mass concentrations were maintained but the relative concentration ratio of two components varied. In their ternary mixture experiment, the mass concentration of one component was fixed but the relative concentration ratio of the other two varied. It is effectively a binary mixture but with a background having a fixed concentration component. In our work, the relative concentration ratio of all three components is varied.

The variables Castner *et al.* used for their data analysis are selected based on a public literature search. The most intense fragment peaks, which are presumed to be originated from an amino acid, are chosen for further analysis. This might bias the data analysis because the various experiments in the open literatures were carried out under very different conditions.

The method Castner *et al.* used for data analysis is partial least squares (PLS) regression. It is a variation of ordinary linear least squares regression:

$$Y = X\beta + E \qquad (6.8)$$

where X are independent variables Y are dependent variables and E are residuals.

The least square solution to (6.8) is:

$$\hat{\beta} = (X'X)^{-1} X'y \qquad (6.9)$$

The problem with (6.9) is that it will break down when X is not of full rank, which often happens with TOF-SIMS spectra analysis where many of variables (peaks) are present. PLS regression provides a way out of this situation. To understand PLS, we have to mention principal component regression (PCR). In PCR, X is replaced by its

principal component scores and Y is regressed on these scores. One may also replace Y by its principal component scores. However, in above, the principal component scores of Y and principal component scores of X are calculated separately and may have only weak connections. PLS decomposes X and Y into a small number of latent variables ("scores") with the constraint that the covariance between X and Y is preserved as much as possible. This is usually done by nonlinear iterative algorithm. For more details, readers may refer to [65, 66].

§ 6.3 Apply LDA to TOF-SIMS mixture data

We first applied LDA to three pure peptide samples. Since the group number is three, we get two nonzero eigenvalues, $i.e.$, two discriminant coordinate axes, which determines a plane. The projection of samples onto this plane is shown in Fig. 6.2. It is clear that they are well separated.

We then project our mixture samples onto this discriminant plane. If there is linear relationship between the peak intensities and the relative concentration ratio, then the projection of mixture samples on this plane would have the same configuration as Fig. 5.1, with three pure samples as three vertices of the simplex. For simplicity, in the following projections for mixtures, only the means of individual samples will be superimposed.

In Fig. 6.3, the scatter plots of projections of three binary mixture samples are shown. The expected position is also shown. As we can see, they lie on the edges of the simplex, close to the middle point.

In Fig. 6.4, ternary mixture of concentration 1:1:1 is plotted, they are in the middle of the simplex, but locate to the left of expected location (gray diamond).

In Fig. 6.5, ternary mixtures of concentration ratios 1:1:4, 1:4:1 and 4:1:1 are plotted. They are relatively more spread, but not far away from their expected value (gray diamonds).

In Fig 6.6, the mean of all samples are plotted. The expected values of all samples are also plotted. Arrows are used to illustrate the deviation of sample means from expected values.

As we can see in above figures, projections of all samples are qualitatively in the positions we expected. Deviations from the expected positions may be due to many reasons. For example, the relatively large spread of each sample suggests a better normalization strategy may be needed. It is also possible that the intensities of peaks may not depend linearly on the concentrations, thus some higher order correction may be needed.

Fig. 6.1 Illustration of LDA in two-group case
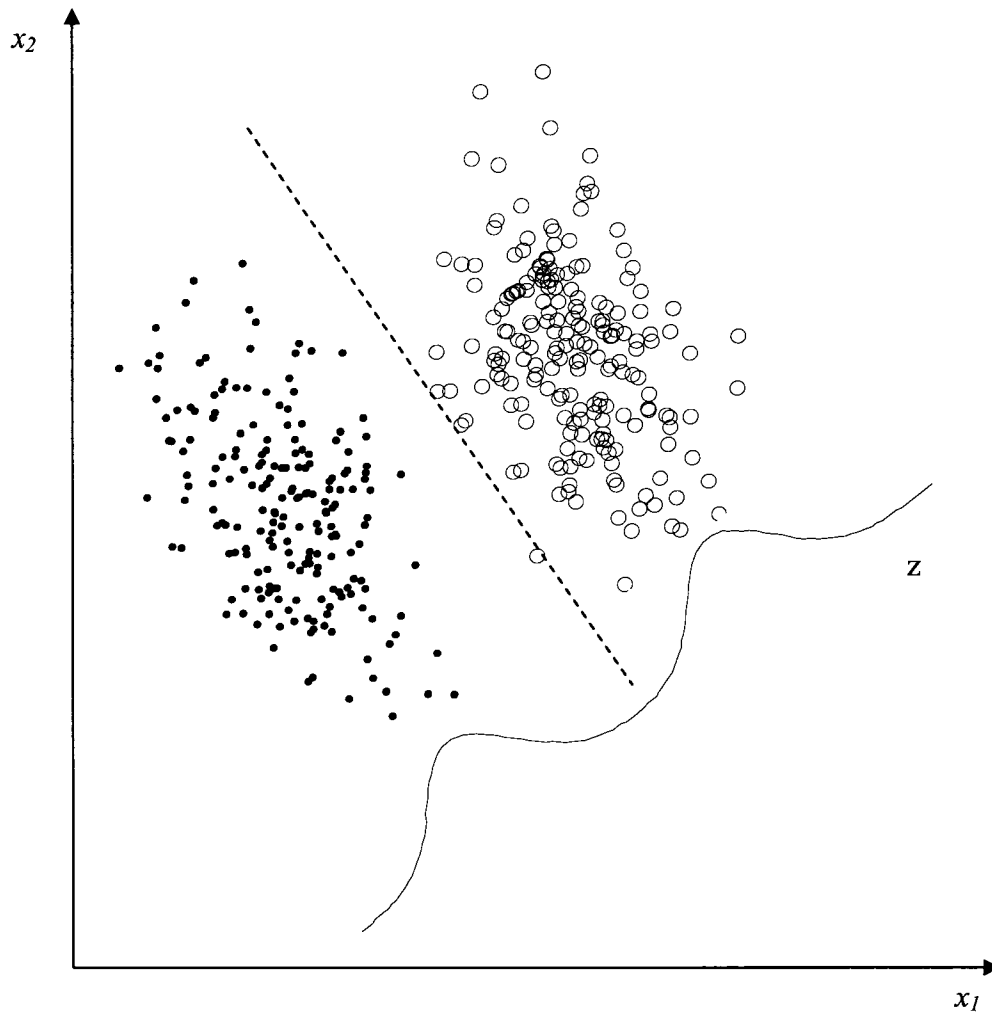
An illustration of LDA in the case of two groups (dots and circles), two variables (x_1 and x_2). The two groups, the dots and circles are not well separated in either x_1 or x_2 directions. LDA looks for a direction z, which is a linear combination of x_1 and x_2, such that when the groups are projected onto the z direction, the two groups are well separated. The two Gaussian-like curves represent the distribution of the projections of two groups onto z direction.

Fig. 6.2 LDA projections of pure samples
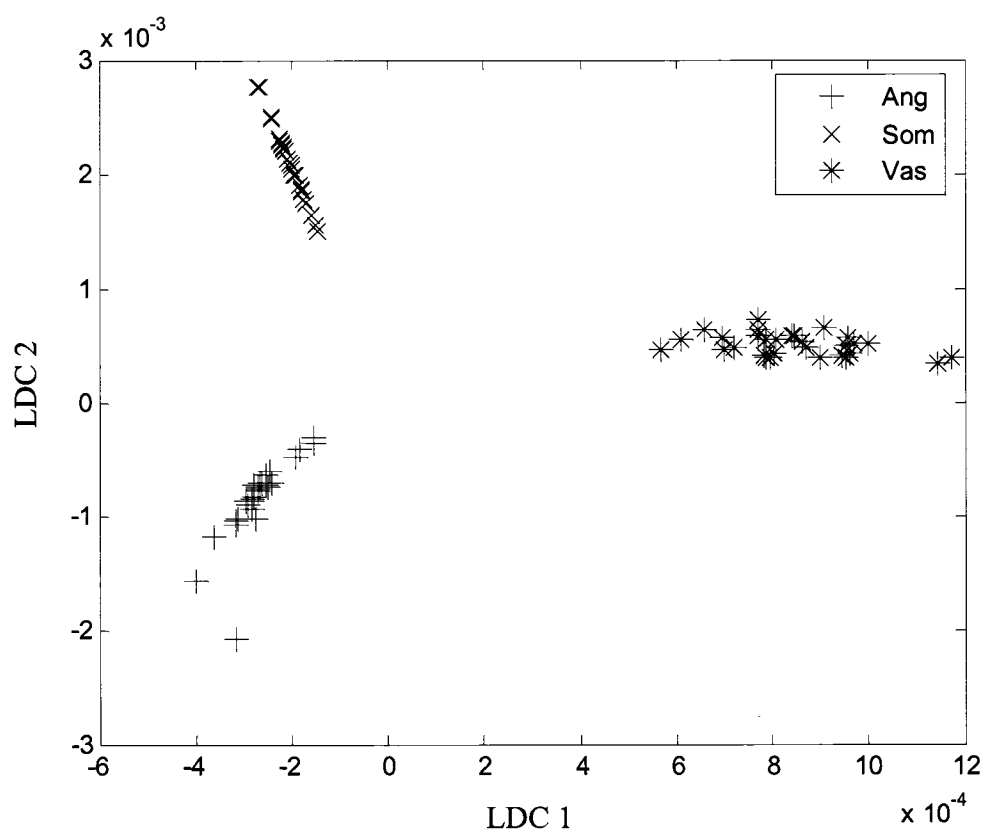
LDA projection of the spectra of three pure peptides samples onto a two dimensional plane. Spectra of three pure peptides form three distinct clusters.
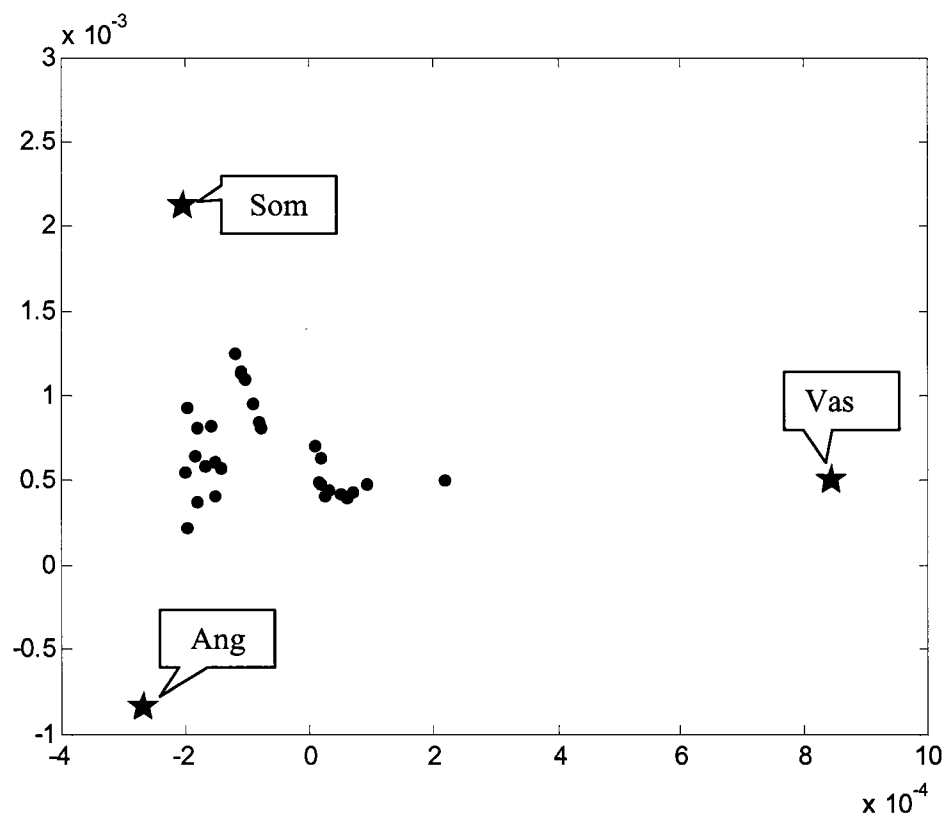
Fig 6.3 Projections of binary mixture samples
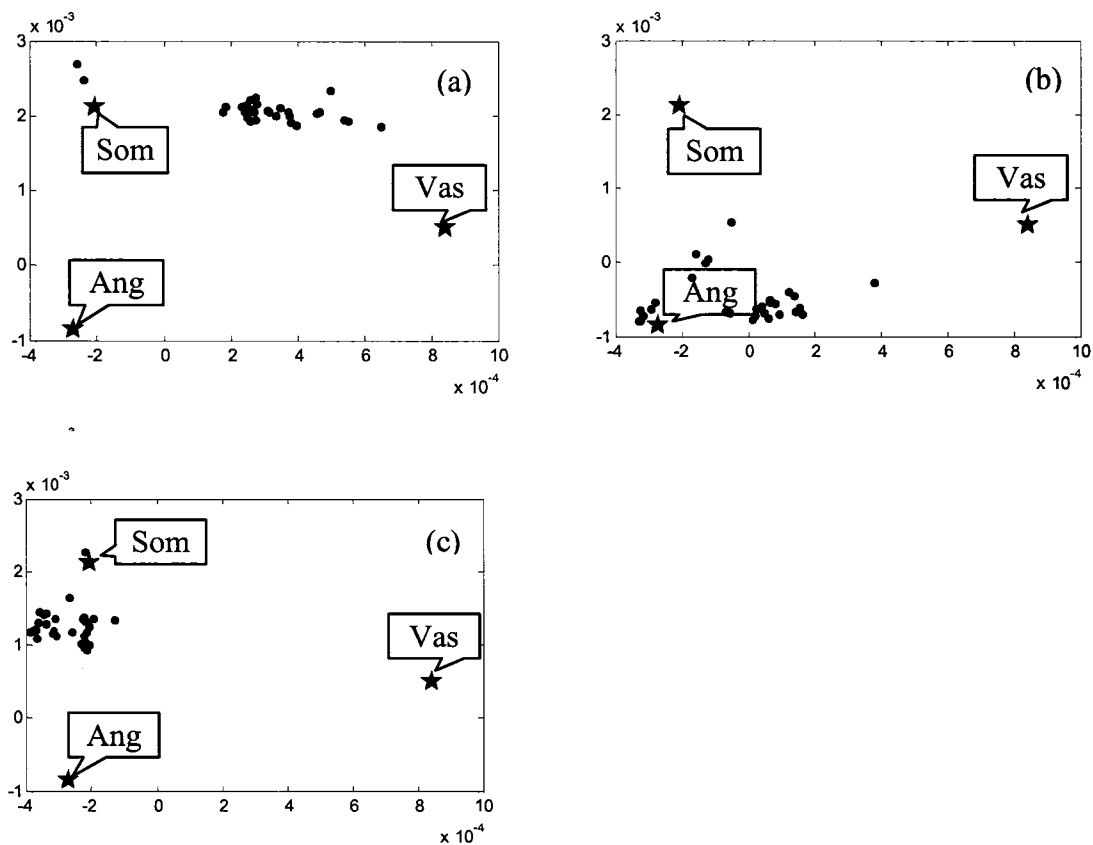
(a): (Ang, Som, Vas)=(0,½,½);   (b): (Ang, Som, Vas)=(½,0,½);
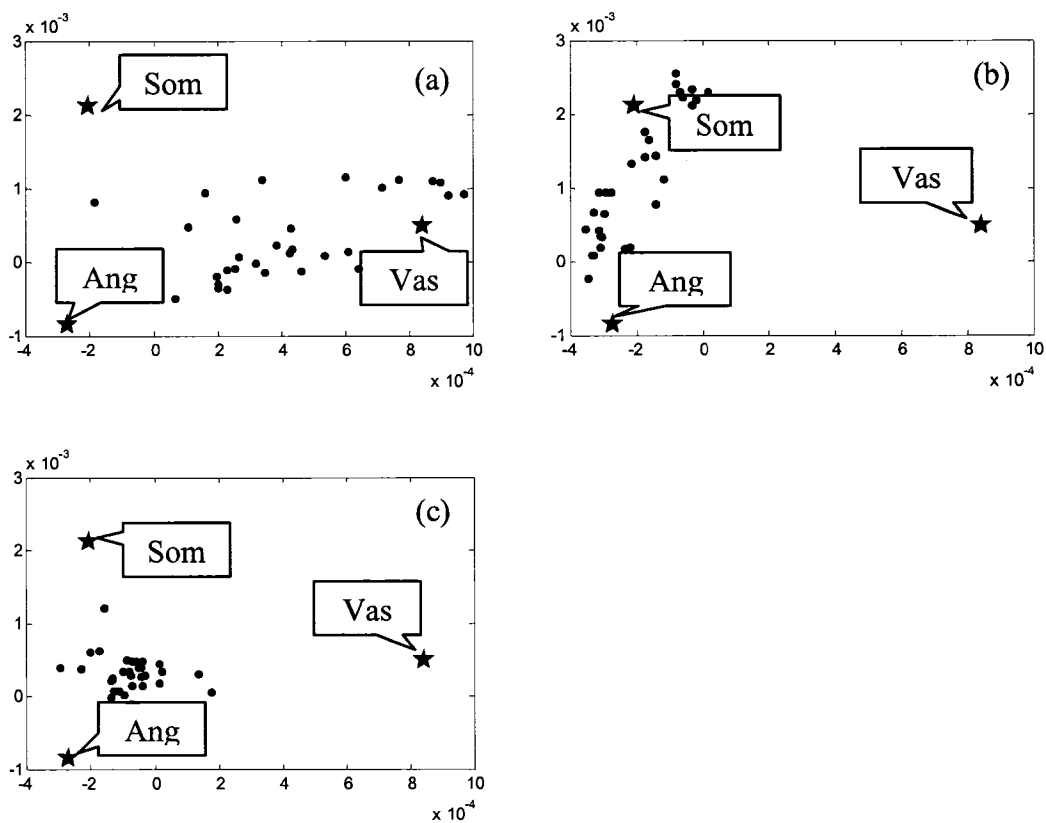(c): (Ang, Som, Vas)=(½,½,0);

The means of three pure samples are illustrated by three stars. The gray diamonds are expected locations, respectively. The projections of binary samples (black dots) are located close to the expected places (gray diamonds). For example, the binary mixture of Som and Vas in pane (a) are located between the means of Som and Vas, except a few outliers.

Fig. 6.4 Projections of ternary mixture
with equal concentration

The projections (dots) of ternary mixture with concentration ratio (Ang, Som, Vas)=(1/3, 1/3, 1/3) locate inside the triangle formed by the stars (the means of pure samples).

Fig. 6.5 Projections of other ternary mixtures

The projections of ternary mixtures with concentration ratio:
(a): (Ang, Som, Vas)=(1/6, 1/6, 2/3); (b): (Ang, Som, Vas)=(1/6, 2/3, 1/6)
(c): (Ang, Som, Vas)=(2/3, 1/6, 1/6)

The means of pure peptides are represented by three stars and for each mixture, the expected positions are represented by the gray diamond. Though the projections are more scattered, they are still close to the expected positions.

Fig. 6.6 The mean of all samples
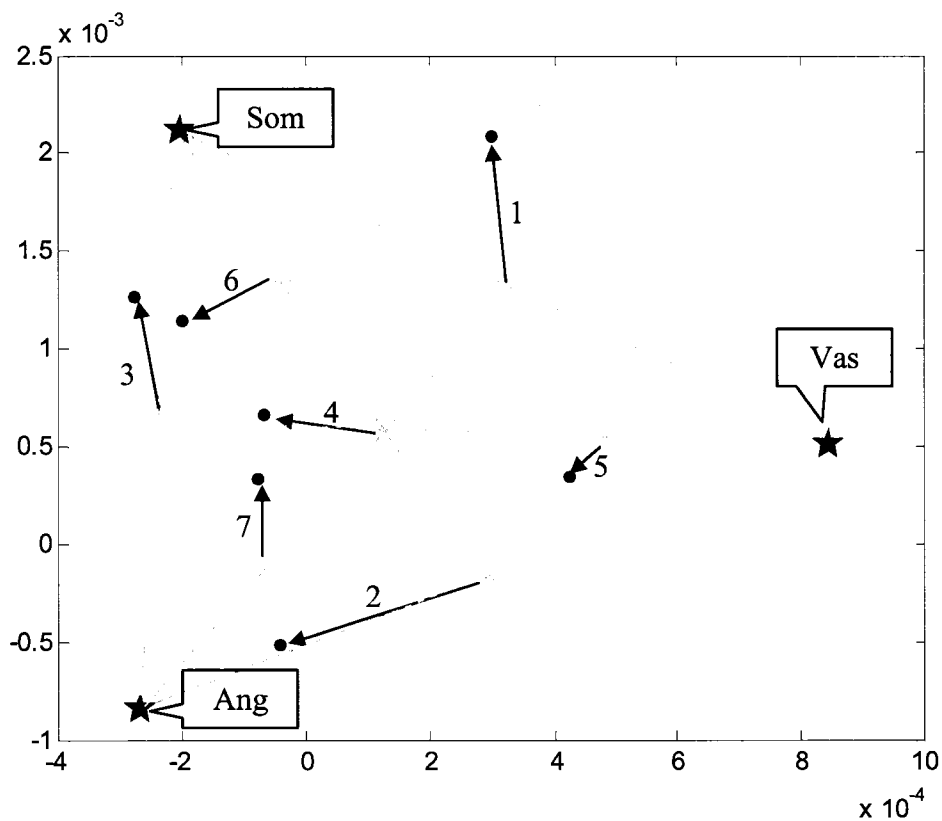
1: (0, 1/2, 1/2); 2: (1/2, 0, 1/2); 3: (1/2, 1/2, 0); 4: (1/3, 1/3, 1/3);
5: (1/6,, 1/6, 2/3); 6: (1/6, 2/3, 1/6); 7: (2/3, 1/6, 1/6)

Three stars are the means of three pure sample, they formed the vertices
of the simplex. The expected locations of all mixture samples are
represented by gray diamonds. The actual means of mixture samples
are plotted as black dots. An arrow from the expected location to actual
mean is used to shown the deviation from expected value.

Chapter 7

Summary and future work

§ 7.1 Conclusion

A new peak identification procedure for static TOF-SIMS spectra has been developed. The algorithm is based on the understanding of the nature of static TOF-SIMS, a statistical test of peak presence and maximum likelihood parameter fitting.

The understanding of the nature of TOF-SIMS is one of the most basic aspects of this thesis. Without it, it would be impossible to accommodate the correct noise model and to formulate the appropriate statistical test. Three essential points about TOF-SIMS are: 1]. it is a counting problem; 2] the count at time t_i is independent of the count at t_j, even when t_i and t_j are from the same peak; and 3] a theoretical understanding of TOF-SIMS as a physical device enables us to derive a reasonable one-parameter peak lineshape function that works well for a wide mass range.

The first point is obvious. By the nature of time of flight and the instrument configuration, the primary-ion beam has to be run in a pulsed mode. For each pulse containing a given number of ions, there are many fewer secondary ions that reach the

134

detector. The detector, which consists of two microchannel plates, has such a high time resolution (138 ps) that it actually detects every ion impact event as a single event on the detector. This means that a Poisson process is involved. If at a given time t_i the ideal rate is r_i, then the probability of an actual observation n_i is described by Poisson distribution function, as in equation (3.4).

The second point is less evident, but is due to the fact that a count at time t_i and a count at t_j may result from different primary-ion pulses, even if they represent the same peak. The static TOF-SIMS regime requires that any part of the sample should only be impacted no more than once during the acquisition. This means that each primary-ion pulse is an independent probe of the surface under nominally the same condition. One secondary ion that happens at time t_i in one probe is independent of another secondary ion that happens at time t_j in another probe, which in turn is independent of all other probes. The final spectrum is a summation of the outputs of millions of independent probes. Thus the count at time t_i is independent of the count at t_j. The importance of this is that it allows us to write out the likelihood function for N successive data points.

The third point is also crucial. By studying the ion dynamics and an argument from maximum entropy, a peak lineshape is derived. Though it is based on simple approximation, the derived peak lineshape captures the shape of the top half of a peak very well, including its asymmetry. The derived peak lineshape gives us the local rate in the Poisson distribution function (3.4). Without this peak model, no calculation can be done.

With the above understanding of TOF-SIMS, we have developed a novel procedure to detect peaks in a TOF-SIMS spectrum.

The first step in peak detection is to put an observation window of width N on the spectrum and thus isolate N data points $n = \{n_1, n_2 \cdots n_N\}$. Since our peak lineshape only fits well to the top half of the peak, N is chosen such that the window runs from the left half maximum of a peak to the right half maximum. For these N data points, we compute the odds ratio that compares the hypothesis H_1 that there is a peak in the observation window versus the hypothesis H_0 that there is no peak in the window (i.e., just dark current). In doing so, we need to compute two posteriors, $p(H_1 \mid n)$, the probability that there is a peak given the data $n = \{n_1, n_2 \cdots n_N\}$ in the window, and $p(H_0 \mid n)$ the probability that there is not a peak in the window given the data $n = \{n_1, n_2 \cdots n_N\}$. These two terms are calculated by invoking Bayes' theorem. Having finished the calculation in this window, the window is shifted one time step to the right and the same comparison is made for the new window. This is performed continuously until the window hits the end of the spectrum. One thing to keep in mind is that, as the mass resolution of a TOF-SIMS spectrum under investigation is roughly the same, mass peaks get wider as the mass gets larger, and correspondingly, the window width must get larger to always cover the top half of a peak.

Having computed the odds ratio, a threshold value is set according the expected value of the odds ratio. Region where the odds ratio rises above the threshold yields confidence that there is a peak present. We then estimate its position and amplitude via

parameter fitting by maximum likelihood methods. The maximum likelihood also gives quantitative estimations of uncertainties in peak position and peak amplitudes. To our knowledge, our procedure is the only peak finder to yield this information.

The above peak detection procedure can detect peaks in a TOF-SIMS spectrum automatically and efficiently. It greatly reduces a TOF-SIMS spectrum for a couple million data points to a couple of thousand numbers representing four entries of each of a few hundred peaks. That is, the number of data points of a spectrum is reduced by 1000 times without losing any essential information.

We found a shift between the same peak in different spectra taken at different times or positions. The shift gets larger at later times (higher mass) in a spectrum and shows a linear trend between the time shift and arrival time. This shift is likely due to surface morphology or electric voltage fluctuations, both of which may cause a slight change in the kinetic energy. Correcting this linear trend, the same peaks in different spectra are brought into alignment within a range of one or two time points out of more than hundred thousand time points. This provides us a way to align peaks among spectra.

We then applied the method to the spectra from a mixture ratio experiment. The experiment is designed to be a model for the use of TOF-SIMS as a supplementary biomarker discovery tool. In the experiment, mixtures of the peptides angiotensin (Ang), somatostatin (Som) and vasopressin (Vas) of known solution concentration ratios were deposited onto etched silver. The purpose of the work is to deduce the

mixture ratio on the surfaces (which we have assumed to be nominally the same as the

solution ratios) from the TOF-SIMS spectra.

In order to infer the concentration ratio from the spectra, multivariate analysis is

introduced. The above alignment allows us to prepare spectra into a data matrix format

for multivariate analysis. Since there are hundreds of peaks in a spectrum, and only a

small number of spectra, a variable selection must be performed. We do so by

implementing an algorithm proposed by McHenry which uses Wilks' Λ as a criteria to

select a set of d peaks that provides the best discriminating ability. We then applied

linear discriminant analysis (LDA) to three individual peptide samples using the

selected d peaks. The inherent connection between Wilks' Λ and LDA ensured that

the selected peaks give the best discriminant capability when LDA is applied. With the

selected d peaks, LDA projects the data from d-dimensions onto a 2-dimensional

plane, since we have only three groups. In this 2-dimensional plane, the three different

groups (peptides) are separated as much as possible. Once the projection direction is

determined, we then applied it to all mixture samples of various concentration ratios.

The projections of mixture samples in the 2-dimensional plane distribute as expected,

having the 3 pure peptides sit in the vertices, and mixtures close to the expected

nominal locations as in Fig. 5.1.

§ 7.2 Future work

Though the peak detection algorithm developed here specifically focused on

identifying peaks in a TOF-SIMS spectrum, the logic is general and can be modified

to accommodate other types of analytical instrument, not necessarily just mass spectrometers. In Chapter Two, we gave an example of finding peaks in Gaussian noise.

Our peak lineshape is derived based on a simple approximation. A better peak lineshape would lead to better results because we would be able to include more data points in the window, yielding greater statistical reliability. The peak lineshape has a fixed resolution parameter which is estimated from the peak in the spectrum. We have found that resolution varies from peak to peak a little bit, thus using a fixed resolution may not be a good assumption. A more appropriate method would let the resolution be a parameter that varies along with the peak amplitude.

It is possible that sometimes nearby masses overlap, resulting in a broader peak. With TOF-SIMS, this may not be as serious an issue because of its high mass resolution even isotopes are generally resolved very well. But for other mass spectrometers, or other instruments whose resolutions are low, this may lead to difficulties in accurate peak position and amplitude assignments. It is then necessary to test a series of hypotheses that in the window there is no peak, one peak, two peaks, and so on, so that, we may choose the most likely one.

Appendix A

Derivation of Equations

(2.41), (2.42) and (2.47)

1. Derivation of equation (2.41)

When the noise is white Gaussian, for the N data points that are isolated by the window at t_0, for the hypothesis that there is a peak present, the likelihood function, as in equation (2.39):

$$p_N(\eta \mid a_1, \mu_1, \sigma_{\eta 1}, t_0, M_1) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{\eta 1}} e^{-\frac{(s_i - a_1 x_i - \mu_1)^2}{2\sigma_{\eta 1}^2}}$$

$$= \frac{1}{(\sqrt{2\pi}\sigma_{\eta 1})^N} \exp(-\frac{1}{2\sigma_{\eta 1}^2} \sum_{i=1}^{N} (s_i - a_1 x_i - \mu_1)^2)$$

(A.1)

The natural log of the likelihood function is:

$$L = \log\left[p_N(\eta \mid a_1, \mu_1, \sigma_{\eta 1}, t_0, M_1) \right]$$

$$= -\frac{N}{2}\log(2\pi) - N\log(\sigma_{\eta 1}) - \frac{1}{2\sigma_{\eta 1}^2} \sum_{i=1}^{N} (s_i - a_1 x_i - \mu_1)^2$$

(A.2)

To maximize it, let:

140

$$\frac{\partial L}{\partial a} = -\frac{1}{\sigma_\eta^2} \sum_{i=1}^{N} (s_i - ax_i - \mu)(-x_i) = 0$$

$$\frac{\partial L}{\partial \mu} = -\frac{1}{\sigma_\eta^2} \sum_{i=1}^{N} (s_i - ax_i - \mu)(-1) = 0 \qquad\qquad (A.3)$$

$$\frac{\partial L}{\partial \sigma_\eta} = -\frac{N}{\sigma_\eta} + \frac{1}{\sigma_\eta^3} \sum_{i=1}^{N} (s_i - ax_i - \mu)^2 = 0$$

Solving these equations, we have:

$$a_1^* = \frac{\overline{(sx)} - \overline{s}\,\overline{x}}{\overline{x^2} - \overline{x}^2} = \frac{\overline{(s-\overline{s})(x-\overline{x})}}{\overline{(x-\overline{x})^2}} = \frac{\overline{(s-\overline{s})(x-\overline{x})}}{\sigma_x^2}$$

$$\mu_1^* = \overline{s} - a_1^* \overline{x} \qquad\qquad (A.4)$$

$$\sigma_{\eta 1}^* = \overline{(s_i - \overline{s} - a_1^*(x_i - \overline{x}))^2} = \sigma_s^2 - a_1^{*2}\sigma_x^2$$

This is the result in equation (2.41).

2. Derivation of equation (2.42)

The elements in Hessian matrix $\nabla\nabla L(a_1, \mu_1, \sigma_{\eta 1})$ evaluated at $(a_1^*, \mu_1^*, \sigma_{\eta 1}^*)$ are:

$$-\frac{\partial^2 L}{\partial a_1^2}\bigg|_{(a_1^*,\mu_1^*,\sigma_{\eta 1}^*)} = \frac{1}{\sigma_{\eta 1}^{*\,2}}\sum_{i=1}^{N} x_i^2 = \frac{N\left(\sigma_x^2 + \overline{x}^2\right)}{\sigma_{\eta 1}^{*\,2}}$$

$$-\frac{\partial^2 L}{\partial \mu_1^2}\bigg|_{(a_1^*,\mu_1^*,\sigma_{\eta 1}^*)} = \frac{1}{\sigma_{\eta 1}^{*\,2}}\sum_{i=1}^{N}(1) = \frac{N}{\sigma_{\eta 1}^{*\,2}}$$

$$-\frac{\partial^2 L}{\partial \sigma_{\eta 1}^2}\bigg|_{(a_1^*,\mu_1^*,\sigma_{\eta 1}^*)} = -\frac{N}{\sigma_{\eta 1}^{*\,2}} + \frac{3}{\sigma_{\eta 1}^{*\,4}}\sum_{i=1}^{N}(s_i - a_1^* x_i - \mu_1^*)^2 = \frac{2N}{\sigma_{\eta 1}^{*\,2}}$$

$$-\frac{\partial^2 L}{\partial a_1 \partial \mu_1}\bigg|_{(a_1^*,\mu_1^*,\sigma_{\eta 1}^*)} = \frac{\sum_{i=1}^{N} x_i}{\sigma_{\eta 1}^{*\,2}}$$

$$-\frac{\partial^2 L}{\partial \mu_1 \partial \sigma_{\eta 1}}\bigg|_{(a_1^*,\mu_1^*,\sigma_{\eta 1}^*)} = \frac{2}{\sigma_{\eta 1}^{*\,3}}\sum_{i=1}^{N}(s_i - a_1^* x_i - \mu_1^*) = 0$$

$$-\frac{\partial^2 L}{\partial a_1 \partial \sigma_{\eta 1}}\bigg|_{(a_1^*,\mu_1^*,\sigma_{\eta 1}^*)} = \frac{2}{\sigma_{\eta 1}^{*\,3}}\sum_{i=1}^{N}(s_i - a_1^* x_i - \mu_1^*)x_i = 0 \tag{A.5}$$

The last 2nd term is zero obviously because $\mu_1^* = \overline{s} - a_1^* \overline{x}$. The last term is also because it is proportional to the correlation between the residue, which is presumably white noise with zero mean, and the target x_i, which is zero. Nevertheless, it can be shown by following:

$$\sum_{i=1}^{N}(s_i - a_1^* x_i - \mu_1^*)x_i$$

$$= \sum_{i=1}^{N}(s_i - a_1^* x_i - \overline{s} + a_1^* \overline{x})x_i$$

$$= \sum_{i=1}^{N} s_i x_i - a_1^* \sum_{i=1}^{N} x_i^2 - \overline{s}\sum_{i=1}^{N} x_i + a_1^* \overline{x}\sum_{i=1}^{N} x_i$$

$$= N(\overline{sx} - a_1^* \overline{x^2} - \overline{s}\,\overline{x} + a_1^* \overline{x}^2) \tag{A.6}$$

$$= N(\overline{sx} - \overline{s}\,\overline{x} - a_1^*(\overline{x^2} - \overline{x}^2))$$

$$= N(\overline{(s-\overline{s})(x-\overline{x})} - a_1^* \sigma_x^2)$$

$$= 0$$

Thus, we have the result in equation (2.42):

$$\nabla\nabla L(a_1^*, \mu_1^*, \sigma_{\eta1}^*) = \begin{bmatrix} L_{aa} & L_{a\mu} & L_{a\sigma} \\ L_{\mu a} & L_{\mu\mu} & L_{\mu\sigma} \\ L_{\sigma a} & L_{\sigma\mu} & L_{\sigma\sigma} \end{bmatrix}$$

$$= -\begin{bmatrix} \dfrac{N\left(\sigma_x^2 + \bar{x}^2\right)}{\sigma_{\eta1}^{*2}} & \dfrac{\sum\limits_{i=1}^{N} x_i}{\sigma_{\eta1}^{*2}} & 0 \\[3ex] \dfrac{\sum\limits_{i=1}^{N} x_i}{\sigma_{\eta1}^{*2}} & \dfrac{N}{\sigma_{\eta1}^{*2}} & 0 \\[3ex] 0 & 0 & \dfrac{2N}{\sigma_{\eta1}^{*2}} \end{bmatrix} \qquad \text{(A.7)}$$

3. Derivation of equation (2.47)

To get (2.47), as we can see in equation (2.46):

$$p(\eta \mid M_1, t_0) \approx \frac{1}{a_{1\max}} \frac{1}{\mu_{1\max}} \frac{1}{\sigma_{\eta1\max}} \frac{(2\pi)^{3/2}}{\sqrt{\left|\det(\nabla\nabla L(a_1^*, \mu_1^*, \sigma_{\eta1}^*, t_0))\right|}} e^{L(a_1^*, \mu_1^*, \sigma_{\eta1}^*, t_0)} \qquad \text{(A.8)}$$

we need the determinant of the Hessian matrix:

$$\left|\det\left[\nabla\nabla L(a_1^*,\mu_1^*,\sigma_{\eta1}^*)\right]\right| = \frac{2N^3\left(\sigma_x^2+\bar{x}^2\right)}{\sigma_{\eta1}^{*\,6}} - \frac{2N\left(\displaystyle\sum_{i=1}^{N}x_i\right)^2}{\sigma_{\eta1}^{*\,6}}$$

$$= \frac{2N^3\left(\sigma_x^2+\bar{x}^2\right)-2N^3\bar{x}^2}{\sigma_{\eta1}^{*\,6}} \tag{A.9}$$

$$= \frac{2N^3\sigma_x^2}{\sigma_{\eta1}^{*\,6}}$$

Inverse it,

$$\frac{1}{\sqrt{\left|\det\left[\nabla\nabla L(a_1^*,\mu_1^*,\sigma_{\eta1}^*)\right]\right|}}$$

$$= \frac{\sigma_{\eta1}^*}{\sqrt{N\sigma_x^2}}\,\frac{\sigma_{\eta1}^*}{\sqrt{N}}\,\frac{\sigma_{\eta1}^*}{\sqrt{2N}} \tag{A.10}$$

$$\propto \Delta a_1 \Delta \mu_1 \Delta \sigma_{\eta1}$$

Substitute into (2.46), we have:

$$p(\eta\,|\,M_1,t_0) \approx \frac{1}{a_{1_{max}}}\frac{1}{\mu_{1_{max}}}\frac{1}{\sigma_{\eta1_{max}}}\frac{(2\pi)^{3/2}\sigma_{\eta1}^{*\,3}}{(2N^3)^{1/2}\sigma_x}\frac{1}{(2\pi)^{N/2}\sigma_{\eta1}^{*\,N}}e^{-N/2}$$

$$= \frac{1}{a_{1_{max}}}\frac{1}{\mu_{1_{max}}}\frac{1}{\sigma_{\eta1_{max}}}\frac{\sigma_{\eta1}^*}{\sqrt{N\sigma_x^2}}\frac{\sigma_{\eta1}^*}{\sqrt{N}}\frac{\sigma_{\eta1}^*}{\sqrt{2N}}(2\pi)^{3/2}\,p^*(\eta\,|\,a_1^*,\mu_1^*,\sigma_{\eta1}^*,M_1) \tag{A.11}$$

$$\propto \frac{\Delta a_1}{a_{1_{max}}}\frac{\Delta \mu_1}{\mu_{1_{max}}}\frac{\Delta \sigma_{\eta1}}{\sigma_{\eta1_{max}}}(2\pi)^{3/2}\,p^*(\eta\,|\,a_1^*,\mu_1^*,\sigma_{\eta1}^*,M_1)$$

This is the equation (2.47)

Appendix B

Derivation of Equation (3.12) and (3.18)

1. Derivation of equation 3.12

To evaluate the following integral:

$$p(n \mid M_0) = \iint p(n \mid a, r_0, M_0) p(a, r_0 \mid M_0) da dr_0$$

$$= \iint e^{-Nr_0} e^{-a} \prod_{i=1}^{N} \frac{(ax_i + r_0)^{n_i}}{n_i!} \frac{\delta(a)}{2r_{0\max}} da dr_0 \tag{B.1}$$

$$= \frac{1}{2r_{0\max} \prod n_i!} \int_0^{r_{0\max}} e^{-Nr_0} r_0^{\sum_{i=1}^{N} n_i} dr_0$$

Let us change variables:

$$Nr_0 = t$$

$$dr_0 = \frac{dt}{N} \tag{B.2}$$

$$r_0^{\sum n_i} = \left(\frac{t}{N}\right)^{\sum n_i}$$

Note that:

$$z! = \int_0^{\infty} t^z e^{-t} dt \tag{B.3}$$

Substitute (B.2) and (B.3) into (B.1), we have:

145

$$p(n \mid M_0) = \frac{1}{2r_{0\max} \prod n_i!} \int e^{-t} \left(\frac{t}{N}\right)^{\sum_{i=1}^{N} n_i} \frac{dt}{N}$$

$$= \frac{1}{2r_{0\max} N^{1+\sum_{i=1}^{N} n_i} \prod n_i!} \int e^{-t} t^{\sum_{i=1}^{N} n_i} \, dt \qquad (B.4)$$

$$= \frac{\left(\sum_{i=1}^{N} n_i\right)!}{2r_{0\max} N^{1+\sum_{i=1}^{N} n_i} \prod n_i!}$$

We get the formula in equation (3.12).

2. Derivation of equation (3.18)

The assumption is that there is only peak, no dark current, the integral need to be evaluated is:

$$p(n \mid only \ a \ peak \ in \ the \ window \ at \ t_0)$$

$$= \iint e^{-Nr_0} e^{-a} \prod_{i=1}^{N} \frac{(ax_i + r_0)^{n_i}}{n_i!} \frac{\delta(r_0)}{2a_{\max}} \, da \, dr_0$$

$$= \frac{1}{2a_{\max} \prod n_i!} \int e^{-a} \prod_{i=1}^{N} (ax_i)^{n_i} \, da \qquad (B.5)$$

$$= \frac{\prod_{i=1}^{N} x_i^{n_i}}{2a_{\max} \prod n_i!} \int e^{-a} a^{\sum_{i=1}^{N} n_i} \, da$$

Use (B.3) again, we get equation (3.18):

$$p(n \mid only \ a \ peak \ in \ the \ window \ at \ t_0) = \frac{\left(\prod x_i^{n_i}\right)\left(\sum n_i\right)!}{2a_{\max} \prod n_i!} \qquad (B.6)$$

Appendix C

Maximum Entropy Method

The principle of maximum entropy is a method to determine a unique probability distribution that is consistent with available information.

In discrete case, for example, a random variable has m possible discrete values, $\{A_1, A_2...A_m\}$, and one wants to assign a probability p_i to each possible value A_i, he may draw independently from $\{A_1, A_2...A_m\}$ for n times. The most possible outcome is the one that maximize:

$$w(n) = \frac{n!}{m^n \prod_{i=1}^{m} n_i!}$$

(C.1)

where n_i represents how many times he gets A_i and

$$\sum_{i=1}^{m} n_i = n$$

(C.2)

He would conclude that:

$$p_i = n_i / n$$

(C.3)

147

if the outcome consists with available information, for example, he knows the mean $\bar{A}$:

$$\bar{A} = \sum_{i=1}^{m} A_i p_i \qquad (C.4)$$

That is, he wants to maximize (C.1) subject to the constraint (C.4) and $\sum p_i = 1$.

To maximize (C.1) is equivalent to maximize $\dfrac{\ln\left(w(n)\right)}{n}$. When $n \to \infty$, by using Stirling's approximation and substitute n_i with np_i, we have:

$$\frac{\ln\left(w(n)\right)}{n} = -\sum_{i=1}^{m} p_i \ln(p_i) \qquad (C.5)$$

where,

$$H = -\sum_{i=1}^{m} p_i \ln(p_i) \qquad (C.6)$$

is called entropy.

In continuous case, entropy is defined as:

$$H(x) = -\int p(x) \ln\left(p(x)\right) \qquad (C.7)$$

If the available information is the mean and variance:

$$\bar{x} = \int x p(x) dx \qquad (C.8)$$

$$\sigma^2 = \int (x - \bar{x})^2 p(x) dx \qquad (C.9)$$

We then want to maximize (C.7) with the constraint (C.8), (C.9) and $\int p(x) dx = 1$, in which case, we need to use Lagrange multiplier, let:

$$\Phi\big(p(x)\big) = -\int p(x)\ln\big(p(x)\big)dx + \lambda_0\left(\int p(x)dx - 1\right)$$
$$+ \lambda_1\left(\int xp(x)dx - \overline{x}\right) + \lambda_2\left(\int (x-\overline{x})^2\, p(x)dx - \sigma^2\right) \tag{C.10}$$

We need to functional derivative of (C.10) with respect to $p(x)$:

$$\frac{\partial\Phi\big(p(x)\big)}{\partial p} = \lim_{\varepsilon\to 0}\frac{\Phi\big(p(x)+\varepsilon\zeta(x)\big)-\Phi\big(p(x)\big)}{\varepsilon} = 0 \tag{C.11}$$

First, notice that:

$$\ln\big(p(x)+\varepsilon\zeta(x)\big) = \ln\big(p(x)\big)+\ln\left(1+\frac{\varepsilon\zeta(x)}{p(x)}\right)$$
$$= \ln\big(p(x)\big)+\frac{\varepsilon\zeta(x)}{p(x)}+O(\varepsilon^2) \tag{C.12}$$

and

$$\big(p(x)+\varepsilon\zeta(x)\big)\ln\big(p(x)+\varepsilon\zeta(x)\big)$$
$$= p(x)\ln\big(p(x)\big)+\varepsilon\zeta(x)+\varepsilon\zeta(x)\ln\big(p(x)\big)+O(\varepsilon^2) \tag{C.13}$$

thus, we have:

$$\Phi\big(p(x)+\varepsilon\zeta(x)\big)-\Phi\big(p(x)\big)$$
$$= \varepsilon\int\big(\zeta(x)\ln\big(p(x)\big)+\zeta(x)\big)dx + \varepsilon\lambda_0\int\zeta(x)dx \tag{C.14}$$
$$+ \varepsilon\lambda_1\int x\zeta(x)dx + \varepsilon\lambda_2\int(x-\overline{x})^2\zeta(x)dx$$

and

$$\frac{\partial\Phi\big(p(x)\big)}{\partial p} = \int dx\,\zeta(x)\big(\ln\big(p(x)\big)+1+\lambda_0+\lambda_1 x+\lambda_2(x-\overline{x})^2\big) \tag{C.15}$$
$$= 0$$

Equation (C.15) has to be true for any $\zeta(x)$, thus, we have to have:

$$\ln\big(p(x)\big)+1+\lambda_0+\lambda_1 x+\lambda_2(x-\overline{x})^2 = 0 \tag{C.16}$$

which implies that

$$p(x) = e^{-(1+\lambda_0)} e^{-\lambda_1 x} e^{-\lambda_2 (x-\bar{x})^2} \tag{C.17}$$

Now, consider that $p(x)$ has to satisfy the constraints,

$$\int p(x)dx = 1$$
$$\int xp(x)dx = \bar{x} \tag{C.18}$$
$$\int (x-\bar{x})^2 p(x) = \sigma^2$$

we get:

$$e^{-(1+\lambda_0)} = \frac{1}{\sqrt{2\pi}\sigma}$$
$$\lambda_1 = 0 \tag{C.19}$$
$$\lambda_2 = -\frac{1}{2\sigma^2}$$

This leads $p(x)$ to the Gaussian distribution:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} \tag{C.20}$$

BIBLIOGRAPHY

[1] M. Mann, R. C. Hendrickson, and A. Pandey, Annu. Rev. Biochem. **70**, 437 (2001)

[2] P. L. Ferguson and R. D. Smith, Annu. Rev. Biophys. Biomol. Struct. **32**, 399 (2003)

[3] B. Adam, A. Vlahou, O. J. Semmes and G. L. Wright, Jr., Proteomics **1**, 1264 (2001)

[4] L. H. Cazares, *et al.*, Clin. Cancer Res. **8**, 2541 (2002)

[5] B. Adam, *et al.*, Cancer Res. **62**, 3609 (2002)

[6] Y. Qu, *et al.*, Clin. Chem. **48**, 1835 (2002)

[7] Y.F. Wong, *et al.*, Cancer Lett. **211**, 227 (2004)

[8] Y. Hu, *et al.*, The Breast **14**, 250 (2005)

[9] J. N. Adkins, *et al.*, Mol. Cell. Proteomics **1**, 947 (2002)

[10] C. Wrotnowski, Genet. Eng. News **18**, 14 (1998)

[11] S. Vorderwubecke, S. Cleverley, S. R Weinberger and A. Wiesner, Nat. Methods **2**, 393 (2005)

[12] W. E. Wallace, A. J. Kearsley, and C. M. Guttman, Anal. Chem. **76**, 2446 (2004)

[13] C. M. Guttman, *et al.*, Anal. Chem. **73**, 1252 (2001)

[14] Y. Yasui, *et al.*, Biostatistics **4**, 449 (2003)

[15] Wm. F. Bryant, *et al.*, Anal. Chem. **52**, 38 (1980)

[16] R. Gras, *et al.*, Electrophoresis **20**, 3535(1999)

[17] S. Y. Sokolov, *et al.*, Comput. Methods Programs in Biomed. **72**, 21 (2003)

[18] K. H. Jarman, D. S. Daly, K. K. Anderson and K. L. Wahl, Chemom. Intell. Lab. Syst. **69**, 61 (2003)

[19] J. S. Morris, *et al.*, Bioinformatics, **21**, 1764 (2005)

[20] V. P. Andreev, *et al.*, Anal. Chem. **75**, 6314 (2003)

[21] D. S. Sivia and C. J. Carlile, J. Chem. Phys. **96**, 170 (1992)

[22] B. O. Keller and L. Li, J. Am. Soc. Mass Spectrom. **12**, 1055 (2001)

[23] E. T. Jaynes, *Probability Theory* (Cambridge University Press, Cambridge, UK, 2003)

[24] R. T. Cox, *The Algebra of Probable Inference* (Johns Hopkins Press, Baltimore, 1961)

[25] W. M. Bolstad, *Introduction to Bayesian Statistics* (John Wiley & Sons, Inc., Hoboken, New Jersey, 2004)

[26] A. Benninghoven, Zeitschrift fuer Physik **230**, 403 (1970)

[27] N. Davies, *et al.*, Appl. Surf. Sci. **203-204**, 223 (2003)

[28] R. Hill and P.W.M. Blenkinsopp, Appl. Surf. Sci. **231–232**, 936 (2004)

[29] S.R. Bryan, A.M. Belu, T. Hoshi and R. Oiwa, Appl. Surf. Sci. **231–232**, 201, (2004)

[30] L. V. Vaeck, A. Adriaens, and R. Gijbels, Mass Spectrom. Rev. **18**, 1 (1999)

[31] A. Benninghoven, Int. J. Mass Spectrom. Ion Physica **53**, 85 (1983)

[32] D. Rading, R. Kersting, and A. Benninghoven, J. Vac. Sci. Technol. A **18**, 312 (2000)

[33] A. Delcorte, *et al.*, J. Phys. Chem. B **104**, 2673 (2000)

[34] A. Delcorte and B. J. Garrison, J. Phys. Chem. B **104**, 6785 (2000)

[35] B. J. Garrison, *et al.*, Appl. Surf. Sci. **203-204**, 69 (2003)

[36] I. S. Gilmore and M. P. Seah, Int. J. Mass Spectrom. **202**, 217 (2000)

[37] H. W. Werner, Surf. Interface Anal. **35**, 859 (2003)

[38] M. Betti, Int. J. Mass Spectrom. **242**, 169 (2005)

[39] K. J. Coakley, D. D. Simons, A. M. Leifer, Int. J. Mass Spectrom. **240**, 107 (2005)

[40] A. M. Belu, D. J. Graham, and D. G. Castner, Biomaterials **24**, 3635 (2003)

[41] M. S. Wagner, D. J. Graham, B. D. Ratner, and D. G. Castner, Sur. Sci. **570**, 87 (2004)

[42] S. Widdiyaskekera, P. Hakansson and B.U.R. Sundqvist, Nucl. Instrum. Methods Phys. Res. Sect. B **33**, 836 (1988)

[43] N. Winograd, B. J. Garrison, Int. J. Mass Spectrom. **212**, 467 (2001)

[44] E. T. Jaynes, IEEE Transactions on Systems Science and Cybernetics, sec-4, 227 (1968)

[45] D. I. Malyarenko, *et al.*, Clin. Chem. **51**, 65 (2005)

[46] N.Winograd, Appl. Surf. Sci. **203-204**, 13 (2003)

[47] S.P.H.T. Freeman, Nucl. Instrum. Methods Phys. Res., Sect. B **79**, 627 (1993)

[48] C. A. McCandlish, J. M. McMahon, and P. J. Todd, J. Am. Soc. Mass Spectrom. **11**, 191 (2000)

[49] J. M. McMahon, N. N. Dookeran and P. J. Todd, J. Am. Soc. Mass Spectrom. **6**, 1047 (1995)

[50] P. J. Todd, J. M. McMahon, and C. A. McCandlish, Jr., J. Am. Soc. Mass Spectrom. **15**, 1116 (2004)

[51] H. Nygren *et al.,* Biochim. Biophy. Acta, in press, (2005)

[52] J. Guerquin-Kern, T. Wu, C. Quintana and A. Croisy, Biochim. Biophys. Acta **1724**, 228 (2005)

[53] M. S. Wagner and D. G. Castner, Appl. Surf. Sci. **231–232**, 366 (2004)

[54] M. S. Wagner, T. A. Horbett, and D. G. Castner, Langmuir **19**, 1708 (2003)

[55] A. K. Jain and B. Chandrasekaran, in *Handbook of Statistics 2*, edited by P. R. Krishnaiah and L. N. Kanal (North-Holland Publishing Company—Amsterdam·New York· Oxford, 1982)

[56] L. Kanal and B. Chandrasekaran, Pattern Recog. **3**, 225 (1971)

[57] S. J. Raudys and A. K. Jain, IEEE Trans. Pattern Anal. Mach. Intell. **13**, 252, (1991)

[58] H. Schulerud and F. Albregtsenb, Comput. Methods Programs in Biomed. **73**, 91 (2004)

[59] A. C Rencher, *Methods of Multivariate Analysis, 2nd Ed.* (John Wiley & Sons, Inc., 2002)

[60] C. E. McHenry, Appl. Statist. **27**, 291 (1978)

[61] E. S. Edgington, *Randomization tests* (Marcel Dekker, New York, 1980)

[62] A. R. Solow, Ecology **71**, 2379 (1990)

[63] K. V. Mardia, J. T. Kent and J. M. Bibby, *Multivariate Analysis* (Academic Press Inc., New York, 1979)

[64] M. S. Wagner, M. Shen, T. A. Horbett, and D. G. Castner, J. Biomed. Mater. Res. A **64**(A), 1 (2003)

[65] P. Geladi and B. R. Kowalski, Anal. Chim. Acta **185**, 1 (1986)

[66] A. Phatak and S. De Jong, Journal of Chemometrics **11**, 311 (1997)

VITA

Haijian Chen

Haijian Chen was born in Rugao, Jiangsu Province, China, on 28 January 1976. He graduated from Rugao High School in 1994. In 1998, he graduated from Nankai University in Tianjin, China, with Bachelor of Science degree in Applied Physics. He then entered graduate program at Nankai University and received Master of Science in Biophysics in 2001. In fall 2001, he began his graduate studies at the College of William and Mary. This dissertation was defended on 12 December 2005.

156