

Probabilistic Modelling of Liquid Chromatography Time-of-Flight Mass Spectrometry

Andreas Ipsen

For the degree of Doctor of Philosophy

Department of Surgery & Cancer
Imperial College London

Acknowledgements

This work would not have been possible without the help of many wonderful people whose support it is my pleasure to acknowledge here.

First of all I am deeply grateful to Dr. Tim Ebbels for his invaluable support and friendly advice over the course of this project. Many thanks are due to Dr. Elizabeth Want for carrying out experimental work that was critical to this thesis and for getting me involved in mass spectrometry in the first place. I would also like to thank Prof. John Lindon for his helpful comments on article manuscripts.

The past three years would not have been the same without the support of personal friends and acquaintances and especially of my student colleagues who provided a fun and lively environment. I also give my special thanks to Natalja Strelkowa for countless invigorating discussions about life, science, politics and everything.

I must also thank Prof. Frank Smith and Prof. Mike Gillan whose kind encouragement over the course of my undergraduate degree helped steer me towards academia. The financial support of the Wellcome Trust is gratefully acknowledged.

Finally, I would like to thank Marianne, Birgitte and Bjørn for their continuing support.

Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Andreas Ipsen

Abstract

Liquid Chromatography Time-of-Flight Mass Spectrometry (LC-TOFMS) is an analytical platform that is widely used in the study of biological mixtures in the rapidly growing fields of proteomics and metabolomics. The development of statistical methods for the analysis of the very large data-sets that are typically produced in LC-TOFMS experiments is a very active area of research. However, the theoretical basis on which these methods are built is currently rather thin and as a result, inferences regarding the samples analysed are generally drawn in a somewhat qualitative fashion.

This thesis concerns the development of a statistical formalism that can be used to describe and analyse the data produced in an LC-TOFMS experiment. This is done through the derivation of a number of probability distributions, each corresponding to a different level of approximation of the distribution of the empirically obtained data. Using such probabilistic models, statistically rigorous methods are developed and validated which are designed to address some of the central problems encountered in the practical analysis of LC-TOFMS data, most notably those related to the identification of unknown metabolites.

Unlike most existing bioinformatics techniques, this work aims for rigour rather than generality. Consequently the methods developed are closely tailored to a particular type of TOF mass spectrometer, although they do carry over to other TOF instruments, albeit with important restrictions. And while the algorithms presented may constitute useful analytical tools for the mass spectrometers to which they can be applied, the broader implications of the general methodological approach that is taken are also of central importance. In particular, it is arguable that the main value of this work lies in its role as a proof-of-concept that detailed probabilistic modelling of TOFMS data is possible and can be used in practice to address important data analytical problems in a statistically rigorous manner.

Table of contents

LIST OF FIGURES	7
LIST OF TABLES	9
1 INTRODUCTION AND AIMS	10
2 THE FUNDAMENTALS OF LC-TOFMS DATA	14
2.1 LITERATURE OVERVIEW	14
2.2 A BRIEF OUTLINE OF LC-TOFMS	17
2.3 PRE-PROCESSING	20
2.4 THE DISTRIBUTION OF ION COUNTS	21
2.5 ION DIGITISATION	23
2.6 HISTOGRAMMING AND DETECTOR SATURATION	25
2.7 MAPPING TIME-OF-FLIGHT TO MASS	29
3 A TEST FOR THE IDENTIFICATION OF PARENT-FRAGMENT PAIRS	35
3.1 ELUTION PROFILE	35
3.2 CURRENT SOLUTIONS	37
3.3 THEORY	40
3.3.1 <i>Statistical Model</i>	40
3.3.2 <i>A test of hypothesis for exact coelution</i>	41
3.3.3 <i>Illustrating the coelution test with simulated and real data</i>	44
3.4 EXPERIMENTAL DETAILS	46
3.5 ANALYSIS	48
3.5.1 <i>Selection of test data-sets</i>	48
3.5.2 <i>Validation</i>	51
3.5.3 <i>Performance comparison</i>	58
3.6 DISCUSSION	60
4 CONFIDENCE REGIONS FOR ISOTOPIC ABUNDANCE PATTERNS	62
4.1 BACKGROUND	63
4.2 THEORY	64
4.2.1 <i>Statistical Model</i>	64
4.2.2 <i>Confidence Regions</i>	65
4.2.3 <i>Pooling information</i>	67
4.2.4 <i>Robustness</i>	69
4.3 EXPERIMENTAL DETAILS	71
4.4 ANALYSIS	72
4.4.1 <i>Selection of test data-sets</i>	72
4.4.2 <i>Validation</i>	73
4.5 RESULTS	77
4.6 FUTURE PROSPECTS	79
4.7 DISCUSSION	81
5 PROSPECTS FOR A STATISTICAL THEORY OF LC-TOFMS DATA	83
5.1 COATES' CORRECTION	83
5.2 THEORY	86
5.2.1 <i>The chromatographic dimension</i>	86
5.2.2 <i>The time-of-flight dimension</i>	87
5.2.3 <i>Histogramming of binary TDC data</i>	88
5.2.4 <i>Dead time</i>	90
5.3 MODEL LIMITATIONS	92
5.4 APPLICATIONS OF THE BASIC MODEL	93
5.5 VALIDATION	95
5.6 EXTENSION OF DYNAMIC RANGE	100
5.7 FUTURE APPLICATIONS	106
5.7.1 <i>Deconvolution</i>	106
5.7.2 <i>Isotopic abundance patterns</i>	107

5.7.3	<i>Fine structure isotopic abundance patterns.....</i>	<i>108</i>
5.7.4	<i>Test of hypothesis for mass</i>	<i>109</i>
5.8	DISCUSSION	110
6	DETAILED MODELLING OF DEAD TIME EFFECTS.....	112
6.1	COMPLETING THE COATESIAN MODEL	113
6.2	MAXIMUM LIKELIHOOD ESTIMATES FOR THE COATESIAN MODEL.....	118
6.2.1	<i>The EM algorithm.....</i>	<i>119</i>
6.2.2	<i>Rejection sampling for the Coatesian model</i>	<i>120</i>
6.3	AN GENERALISED MODEL FOR ARBITRARILY TIMED ION ARRIVALS.....	124
6.4	MAXIMUM LIKELIHOOD ESTIMATES FOR THE GENERALISED MODEL	127
6.4.1	<i>Rejection sampling for the generalised model.....</i>	<i>127</i>
6.4.2	<i>Practical implementation issues</i>	<i>129</i>
6.5	APPLYING THE GENERALISED MODEL.....	130
6.6	DISCUSSION	133
7	CONCLUSION.....	136
7.1	SUMMARY	136
7.2	FUTURE WORK	137
7.3	BROADER IMPLICATIONS.....	138
	BIBLIOGRAPHY	141
	APPENDIX A	147
	APPENDIX B.....	148

List of Figures

Figure 1.1 – Schematic comparing the rigorous and heuristic approaches	13
Figure 2.1 – Example mass spectrum.....	19
Figure 2.2 – Example chromatographic peaks.....	20
Figure 2.3 – The standard deviation of the Poisson distribution relative to its rate.....	23
Figure 2.4 – Illustration of the workings of a Microchannel Plate	24
Figure 2.5 – Illustration of the construction of a scan.....	26
Figure 2.6 – Illustration of dead time distortion	27
Figure 2.7 – Comparison of the true distribution of ions and the one sampled from when the Z-focus lens is applied	28
Figure 2.8 – The full set of masses detected in a single scan	30
Figure 2.9 – The full set of mass differences detected in a single scan	31
Figure 2.10 – The smallest set of mass differences	32
Figure 2.11 – The smallest set of mass differences and their doubled values.....	33
Figure 2.12 – The inferred shape of the mapping from time-of-flight to mass	34
Figure 3.1 – Chromatographic peaks and the corresponding ratios illustrating the scenarios of exact coelution, partial coelution, and exact coelution but with distinct widths	36
Figure 3.2 – Chromatographic peaks of two related fragments.....	38
Figure 3.3 – Scatterplot of the ion counts of the matching mass pairs	39
Figure 3.4 – Simulated chromatographic peaks exhibiting exact coelution and partial coelution as well as the corresponding scatterplots.....	45
Figure 3.5 – Real chromatographic peaks exhibiting exact coelution and partial coelution as well as the corresponding scatterplots.....	46
Figure 3.6 – Continuum plots of a pair of isotopologues	49
Figure 3.7 – Scatterplots for the three datasets used.....	51
Figure 3.8 – Histograms of the p-values and quantile-quantile plots of the χ^2 -statistics obtained for the three data-sets examined.....	52
Figure 3.9 – Histograms of the p-values and quantile-quantile plots of the χ^2 statistics obtained for the three data-sets examined following the application of Coates' correction	54
Figure 3.10 – The ion counts of 4-aminohippuric acid a fragment along with a partially coeluting compound.....	57
Figure 3.11 – Histogram of the p-values and a quantile-quantile plot of the χ^2 -statistics obtained from the data-set involving partial coelution	58
Figure 3.12 – Plots of the percentage of the isotopologue pairs that are classified as exhibiting partial coelution by the coelution test and the correlation as a function of increasingly partial coelution.....	60
Figure 4.1 – Quantile-quantile plots of the χ^2 -statistics obtained from the three compounds analysed.....	74
Figure 4.2 – Continuum plot of the two lowest-mass isotopologues of Nitrotyrosine, illustrating the heavy tails of mass peaks.....	75
Figure 4.3 – Quantile-quantile plots of the χ^2 -statistics obtained from the three compounds, after the most extreme 10% have been trimmed	76
Figure 4.4 – The p-values obtained for the data derived from Hippurate, Nitrotyrosine and Chenodeoxycholic Acid, as well as for a series of candidate formulae, and for both individual scans and pooled data	78

Figure 4.5 – The false negatives obtained from the simulated isotopic abundance patterns when a significance level of 0.05 is used for the goodness-of-fit test....	80
Figure 5.1 – Histograms of the p-values and q-q plots of the X^2 statistics obtained from the three datasets by using the Poisson-based model for continuum data to construct the likelihood ratio test	98
Figure 5.2 – Histograms of the p-values and quantile-quantile plots of the X^2 statistics obtained from the three datasets by using the basic model for continuum data to construct the likelihood ratio test	100
Figure 5.3 – The Poisson-to-Binomial mapping.....	103
Figure 5.4 – Illustration of the dead time corrections obtained provided by the basic model for two simulated data-sets when various levels of prior information are assumed.....	105
Figure 6.1 – Validation of the rejection-sampling algorithm for the conditional expectation of the Coatesian model.....	124
Figure 6.2 – Validation of the rejection-sampling algorithm for the conditional expectation of the generalised model	129
Figure 6.3 – Estimates of the entries of θ relative to their true value for the first 100 iteration of the deterministic EM algorithm (the basic model) followed by 100 iterations of the MCEM algorithm (the generalised model).....	132
Figure 6.4 – As above, but focusing on the means of the mass and chromatographic peaks	133

List of Tables

Table 3.1 – Scan numbers and m/z values of the peaks used in the evaluation of the coelution test	50
Table 3.2 – Performance characteristics for the coelution test for the three datasets examined	53
Table 3.3 – Performance characteristics for the coelution test for the three datasets examined following the application of Coates' correction.....	55
Table 3.4 – The parent-fragment pairs used to test for the presence of ionisation suppression effects on the coelution test.....	56
Table 4.1 – The three compounds used in the validation of the confidence regions...	72
Table 6.1 – Comparison of the closure matrices for the Coatesian model and the generalised one	127

1 Introduction and Aims

Over recent years, biological research has become increasingly driven by high-throughput technologies. This has given rise to the so-called “omics” disciplines which aim for a more comprehensive understanding of biological systems than has previously been possible. Two important fields in this category are proteomics, the large scale study of proteins, and metabolomics, the large scale study of metabolites. In both these fields, LC-TOFMS is used extensively in the analysis of biological mixtures, both for the identification of chemical and structural formulae of unknown compounds and in quantitative studies that seek to determine the concentrations of known compounds.

The increasing use of LC-TOFMS has prompted the development of a large number of bioinformatics techniques to facilitate the analysis of the resulting data. Several comprehensive software packages are now available in both proteomics [1, 2] and metabolomics [3-5], which provide extensive tools for the pre-processing and analysis of LC-TOFMS data, and LC-MS data in general. Yet despite these efforts, the task of extracting useful information from the large data-sets produced through LC-TOFMS analysis of complex biological mixtures such as blood or urine remains a central bottleneck to much of the work being carried out in these fields.

There are a large number of pre-processing techniques that are routinely applied to LC-TOFMS data as part of their analysis and these come in roughly two classes: those applied by the manufacturer’s software prior to the data being output to file, and those applied subsequently by the analyst, often through software packages such as those cited above. The former class includes fundamental methods of data compression [6], as well as algorithms that compensate for detector saturation, and which may effectively be part of the physical measurement process [7]. The latter include baseline subtraction, smoothing and feature extraction, as well as methods for standardising data from different experiments, such as normalisation and chromatographic retention time alignment. The pre-processing is followed by the inferential stage of the analysis which may involve the identification of unknown compounds, estimation of the concentration of known ones, and, if multiple biological

samples are involved, finding biologically relevant differences between different groups of samples.

It is clear that the pre-processing will have substantial effects on the data and consequently on all the inferences drawn in downstream analyses [8-10]. Therefore the choice of pre-processing techniques and the manner in which they are applied is extremely important. However, as is evidenced by the broad diversity of techniques that have been developed to address what are effectively the same set of pre-processing problems, there is no consensus as to how best to approach them. It is generally very difficult to provide convincing theoretical arguments for choosing one pre-processing method over others, as essentially none of them are derived from the first principles of the LC-TOFMS operations. Rather, they are heuristic methods, which are constructed based on an intuitive but rather qualitative understanding of the system to which they are applied.

These heuristic methods are often validated by means of direct demonstrations that they produce “reasonable” results when applied to real data, or by arguments that they approximate the steps that would be taken by a trained expert through a more manual analysis [4]. Evaluation of the relative performance of these techniques is extremely difficult as it can be dependent on user experience and the choice of parameters [11, 12]. While there have been calls for the use of “standard data-sets” to compare methods specifically for retention time alignment [12], a meaningful performance comparison based on this approach is likely to remain problematic. And even if the performance of one method could be established as being consistently better than that of others, its effects on the data and on downstream analyses would remain unclear.

The rationale for employing heuristic techniques in the first place is very rarely articulated, possibly because it is taken for granted that the fundamental operations of mass spectrometers are too complex and involve too many unknown variables to allow for a manageable mathematical description. This concern is not unreasonable in view of the elaborate engineering featured in modern mass spectrometers and the rather limited efforts that have so far been made at developing detailed mathematical models to describe the data they produce. The problem is further compounded by the rapidly evolving design of mass spectrometers and the fact that the vast majority of

those used for biological research are commercial models, which renders their design details somewhat inaccessible to many researchers in academia. However a statistical model developed from first principles does not have to account exactly for every aspect of the underlying mass spectrometer design, since approximate models often form a sufficient basis from which to draw the relevant inferences - in the words of George E. P. Box “essentially, all models are wrong, but some are useful”.

The aims of this project are:

1. To evaluate the prospects for adopting a statistically rigorous approach to the interpretation of LC-TOFMS data wherein the detailed architecture of the mass spectrometer is accounted for and data pre-processing is, as far as possible, avoided.
2. To develop specific techniques for improved data analysis based on this rigorous approach, with particular emphasis on methods that facilitate the identification of unknown metabolites in complex mixtures.

The methodological distinction between this “rigorous approach” and the standard “heuristic approach” in which the measurement process and the pre-processing steps are essentially treated as a “black box” is illustrated on Figure 1.1 below.

As will be shown, the rigorous modelling of LC-TOFMS data appears to be feasible and opens up the possibility of applying a new range of statistically rigorous techniques to LC-TOFMS data that better exploit the information contained therein. In the following chapters, a series of mathematical models will be developed that approximate the underlying probability distribution that governs the raw data produced in an LC-TOFMS experiment. A few applications are discussed in detail, but such models may in principle be used to address a very wide range of problems central to the analysis of LC-TOFMS data, by means of the traditional tools of classical frequentist statistics. Thus, inferences can be made by means of tests of hypotheses, which relate directly to the researcher’s aversion to false positives, and parameter estimates can be obtained by means of the method of maximum likelihood. Due to uncertainty regarding the nature of mass and chromatographic peaks and of the detector system used, the models break down at high ion counts. However, it will be

argued that further refinements in our ability to characterise these fundamental features of the data mathematically could allow us to overcome this restriction and lead to substantial improvements in our ability to analyse and interpret LC-TOFMS data. While this thesis focuses on LC-TOFMS data, the methodological approach used applies to other types of mass spectrometers, and indeed to any analytical instrument that provides inexact measurements.

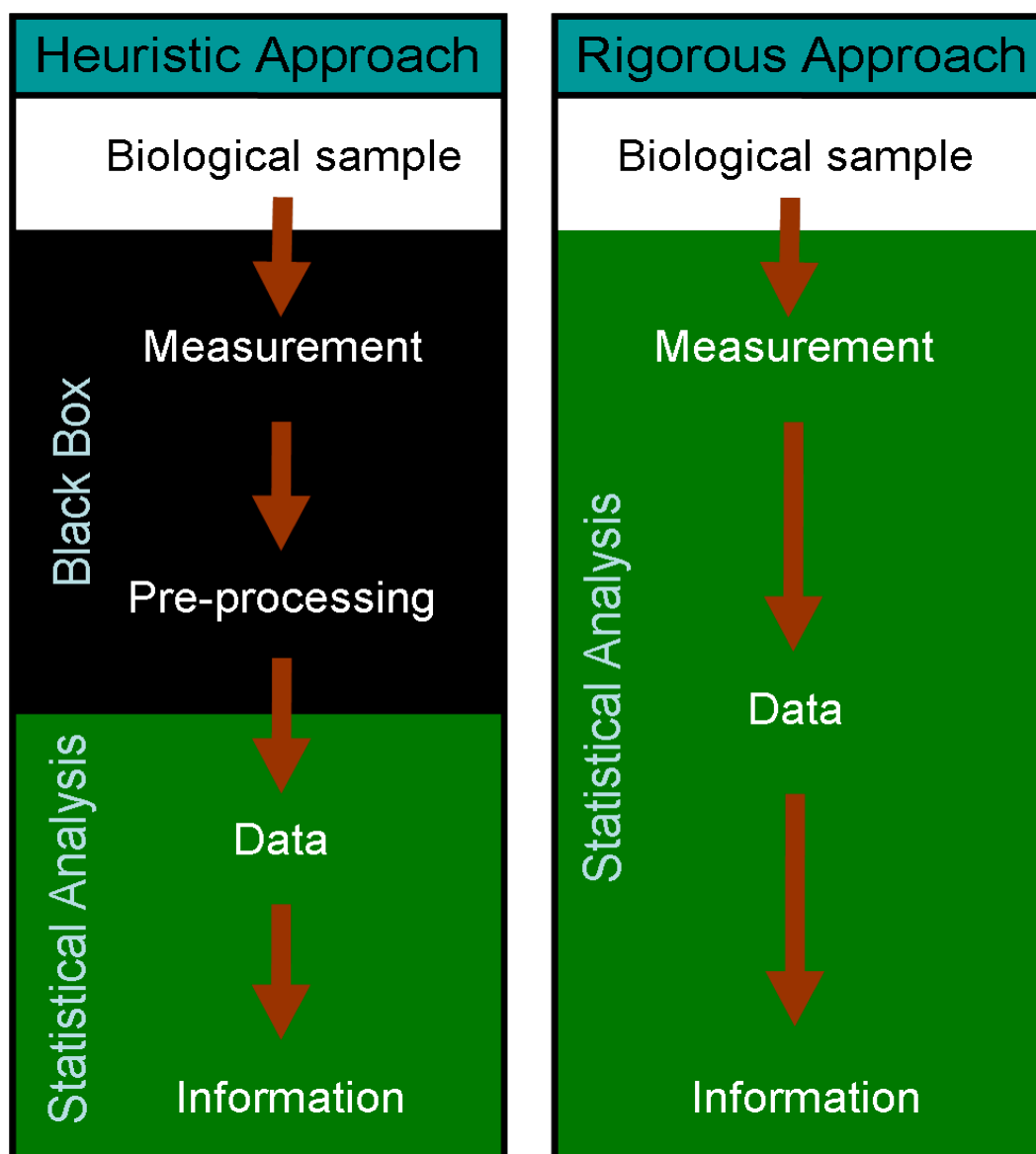


Figure 1.1 - Illustration of the widely used “heuristic approach” to data analysis (left) in which no attempt is made at accounting for the effects of the “black box”. This should be contrasted with the “rigorous approach” (right) which attempts to account for all aspects of the data generation process that are relevant to the eventual analysis.

2 The fundamentals of LC-TOFMS data

2.1 Literature overview

A brief overview of relevant background literature is given below. However, for this thesis relatively limited use has been made of academic articles, as they do not provide many details on the distribution of the data produced by the mass spectrometer used in this project (a Waters Q-ToF Premier). Instead quite extensive efforts have been made at investigating the workings of this mass spectrometer based on Waters patent applications, application notes and manuals, as well as inferences drawn from analyses of acquired data. However, the most useful and reliable information has come from direct communications with engineers at companies that produce mass spectrometers. All of these sources are cited and discussed in further detail in later sections.

It has in fact already been argued that research into the statistical distribution of mass spectrometry data has been somewhat neglected [13]. Several academic articles and books do provide in-depth reviews of the TOF-MS technology [14-20] but these deal mainly with the general instrumental architecture and give only relatively qualitative discussions of the distribution of the resulting data. However an important point that is noted in [14] is that ion arrivals at the detector plate are governed by the Poisson distribution. Therefore, if an ion counting device is employed as part of the detector system (as is the case for the Waters Q-ToF Premier) and detector saturation is negligible, the acquired data will also be governed by the Poisson distribution.

The implications of the data being Poisson distributed, which are critical to this thesis, do not appear to be very widely studied in the scientific literature either, except with respect to the effect that this distribution has on the accuracy of mass estimates. In one such study [21], the author distinguishes between two components of mass errors: those inherent to a given instrument, which should remain roughly constant over time, and the statistical error caused by insufficient measurements, which can be characterised through the Poisson distribution. Interestingly, the mass accuracy is shown to depend strongly on the ion count of the reference compound used to

calibrate the mass spectrometer. It is found that for optimal mass accuracy, this count should be as similar as possible to that of the analyte whose mass is to be measured, the reason being that the detector saturation effects are then approximately the same for both compounds, and can be corrected for through the calibration procedure. Several other studies [22-24] corroborate this finding and (sometimes in rather indirect terms) make note of the Poisson distribution of the moderate ion counts.

One of the few articles that directly studies the distribution of the ion counts [25], investigates the variance of processed proteomics LC-TOFMS data and proposes a two-component error model to describe it. Consistent with other studies, the authors find that for moderate ion counts, the data are consistent with a Poisson distribution, but exhibit variance too great to be Poissonian for high counts – hence the two-component model. It is very possible that this excess variance is a side-effect of the “Z-focus lens” employed by the instrument used in the study (a Waters LCT mass spectrometer) which corrects for detector saturation and which will be discussed in Section 2.6. The author of the article does not appear to be aware of this.

The above study, as well as those dealing with mass accuracy, all adopt a top down approach in that they investigate and characterise the output data, but do not attempt to fit a detailed mathematical model to the processes that are behind it. One article that does present research in this direction is by Du et al. [13] who model the data derived from a given compound as a combination of a Poisson and a Multinomial distribution, the latter being used to account for the different isotopic variants of the compound. For high ion counts, the saturation effects render this model inapplicable. This is accounted for by means of a “semi-empirical” relationship, postulated by the authors, though regrettably little in the way of rationale or evidence is presented to substantiate its validity.

As mentioned in the introduction, an important motivation for developing a mathematical model to describe LC-TOFMS data is to find improved methods for identifying unknown compounds in complex mixtures. Using only the estimated mass from a mass spectrum, one can compile a list of elements consistent with that mass by solving the corresponding linear Diophantine equation for a given set of elements [26]. The task of structural elucidation can however be greatly facilitated by making

use of complimentary information derived from heuristic rules of chemistry or from databases of known compounds. The Dendral project [27] was an early attempt at using computers to identify unknown compounds based on their mass spectra as well as on prior knowledge of chemistry. Several programs have since emerged [28-32], which employ a variety of methods to solve the problem, some of them incorporating spectral information other than MS. A recently developed program for formula elucidation of small molecules [33] claims to assign the correct chemical formula to a mass spectrum with a probability of 98%, so long as the formula exists in the compound database that they use.

One type of complementary chemical information that is particularly relevant to this thesis is based on the fact that metabolites passing through a mass spectrometer often break up into smaller pieces, which are also detected. The observed masses of these fragments can be used to place further constraints on the possible identity of the metabolite from which they are derived. This is the idea behind tandem mass spectrometry, where metabolites of predetermined masses are deliberately fragmented, and the resulting pattern analysed [34]. Various studies have presented methods for improved chemical formula elucidation based on variations of this technique, [35-37]. More modern and higher throughput fragmentation procedures do not restrict the analysis to metabolites of pre-specified masses, but attempt to simultaneously analyse the fragmentation patterns of all metabolites across a wide range of masses. A recent technology known as MS^E [38] attempts to do so by rapidly alternating between a high and a low energy fragmentation process in the hope that the low energy spectrum contains few or no fragments so that the original metabolite can easily be identified and its corresponding fragments read off from the high energy spectrum. But whatever method is used, the analysis is often confounded by the fact that distinct compounds may coelute at roughly the same time, so that it is difficult to determine which fragments are derived from which metabolite. A statistically rigorous method for doing so will be presented in Chapter 3 and is published in [39].

Another important piece of complementary information comes from the isotopic abundance pattern that is characteristic of any given compound. Kind & Fiehn have highlighted the importance of exploiting the information therein [40], pointing out that in terms of its ability to correctly assign chemical formulae, a high resolution

mass spectrometer can be outperformed by a low resolution one, if the latter also makes use of the information in the isotopic abundance pattern. The mathematical description of the exact isotope abundance patterns resulting from a given molecular formula can be expressed using expansions of products of polynomials and has been understood for some time [41]. Various proposals have been made for exploiting the information in isotopic abundance patterns [42-44], including for complex mixtures [45, 46], but none have been found that explicitly account for the statistical distribution of the measured data. However, such a method will be presented in Chapter 4 and is published in [47].

As mentioned above, TOFMS data adhere to the Poisson distribution if an ion counting device is used and if the rate of ion arrivals is low enough that detector saturation is negligible. However, as ion counting detectors are quite easily saturated, numerous statistical methods have been developed [48-54] that aim to enhance their dynamic range by estimating the true number of ion counts, from the recorded counts and using knowledge of the saturation process. One of these correction methods will be described in detail in Chapter 5 and new ones will be proposed in both Chapters 5 and 6 after all of the prerequisite technical concepts have been discussed.

2.2 A brief outline of LC-TOFMS

A characteristic feature of biological mixtures is that they contain a vast number of distinct compounds. In an instrument as sensitive as a time-of-flight mass spectrometer their physical separation is therefore important in order for the analyst to obtain meaningful data. In LC-TOFMS two distinct separation steps are applied to the mixture; a brief account of the two processes is given below. There are in fact several variants of the time-of-flight mass spectrometer but in the following we will restrict ourselves to the widely used “quadrupole orthogonal acceleration time-of-flight mass spectrometer”.

As a mixture passes through the chromatographic column, the compounds within it separate based on their polarity. This is because less polar compounds tend to be held back by a non-polar substance known as the “stationary phase”, which is fixed in place inside the column. Alternatively, a polar stationary phase may be used so that it

is the polar compounds that are retained, as is the case for Hydrophilic interaction chromatography (HILIC) [55], although this approach is less common. While chromatography has improved considerably in recent years, the technology is still not perfect and distinct compounds may coelute at roughly the same time. The time it takes a compound to traverse the chromatographic column is referred to as its “retention time”. While this measure can be useful for metabolite identification, it generally has quite limited reproducibility.

After traversing the column, the analyte enters the gas phase and is ionised. In metabolomics, this is often done through a process known as “electrospray ionisation” due to the method’s high sensitivity even for low mass compounds and the fact that the ionisation process usually does not fragment the analyte excessively. Different compounds ionise with different propensities, which means that compounds that are present in low concentrations can give rise to strong signals and compounds that are present in high concentrations can give rise to weak signals. It is therefore difficult to infer much about the concentrations of the various compounds in the mixture unless isotopic variants of known concentration are spiked into the sample. Once the analyte is ionised, a series of quadrupoles, which can be used for tandem mass spectrometry, guide them towards the orthogonal accelerator while focusing the ion beam and reducing its velocity through collisional damping. One of the quadrupoles will typically contain an inert gas, with which the analyte can collide and this may cause it to fragment. Such “collision induced dissociation” occurs because the analyte’s repeated collisions with the inert gas builds up the potential energy of the analyte molecule so that it may eventually exceed the level of some of its chemical bonds. After the analyte reaches the orthogonal accelerator, an electric field is applied perpendicularly to its direction of travel, giving it a perpendicular velocity that is greater the smaller its mass. The analyte then enters a free drift region, which the lighter and therefore faster compounds will traverse faster than the heavier ones. Many instruments employ one or more “reflectrons”, which use static electric fields to reverse the direction of travel of the analyte and thereby increase the effective length of the free drift region, which improves the instrument’s resolution. At the end of the free drift region the analyte hits a detector plate where its time-of-flight is recorded and from which its m/z (mass per charge) can be estimated.

The data is output in a series of “scans” across the chromatographic dimension and for each scan the number of detected ions of each m/z is listed. An example of a scan as output in Waters’ proprietary MassLynx software is shown below. The sample analysed is a form of synthetic urine, which will be described in detail later.

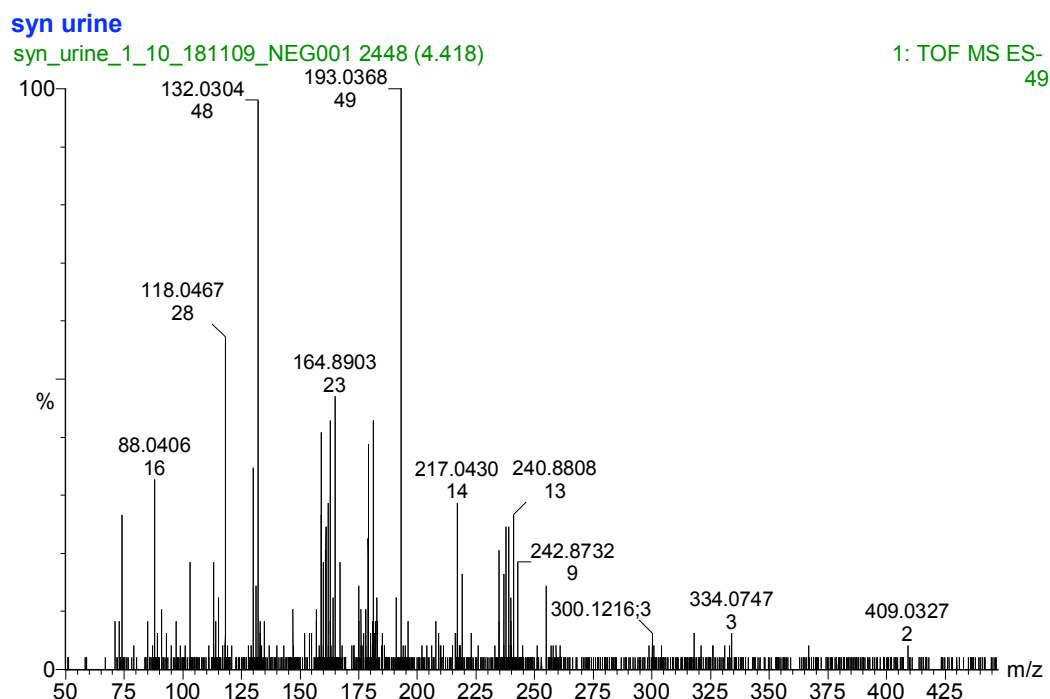


Figure 2.1 - A single scan as seen in MassLynx. For the most prominent peaks the m/z values (top) and ion counts (bottom) are listed. The mass deficits of the m/z values 164.8903, 240.8808, 242.8732 are unusual and may be due to interfering compounds.

Rather than looking at a series of mass peaks at a given scan, we can observe the “chromatographic peaks” by plotting the ion counts at a given m/z value, over a series of scans, as shown on Figure 2.2 below for the three largest mass peaks on the above plot.

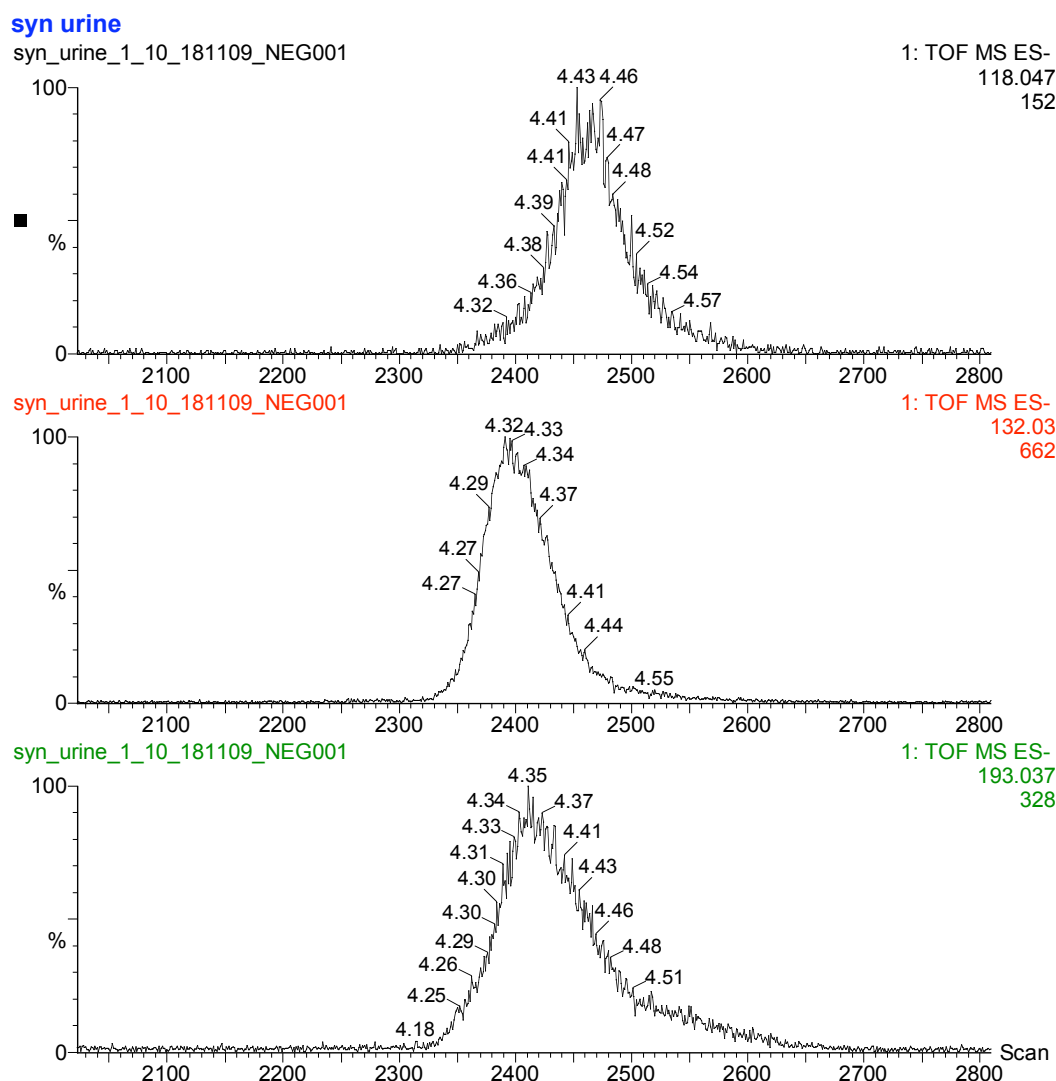


Figure 2.2 – The chromatographic peaks corresponding to the three largest mass peaks on Figure 2.1. The x-axes indicate the scan number and the retention time (in minutes) is listed at various points across the peaks.

2.3 Pre-processing

As stated earlier, LC-TOFMS data as used in routine analyses by experimentalists will have gone through a number of pre-processing steps, which few analysts are likely to know in any detail. The pre-processing steps that are discussed in this thesis are all applied to the data prior to it being seen by the analyst. Although it can in some cases be rather difficult to distinguish between what constitutes pre-processing and what should be regarded as being part of the fundamental architecture of the mass

spectrometer, the following five steps, which are all applied by the Waters Q-Tof Premier certainly require some thought:

- **Centroiding**: reducing all of the data points making up a mass peak to a single mass with a corresponding intensity.
- **Histogramming**: summing the data points across a segment of a chromatographic peak.
- **Dead time correction**: correcting for detector saturation when ion counts are high.
- **Z-focus lens (also known as “DRE lens”)**: reducing the number of ions transmitted when the dead time correction does not suffice.
- **Calibration**: altering the fundamental mapping used to infer the m/z from the time-of-flight, in order to improve mass accuracy.

These pre-processing steps will be discussed in further detail below. It should be emphasised that their use is generally helpful to the analyst: centroiding and histogramming reduce the data size by a large factor, calibration improves mass accuracy and the dead time correction and Z-focus lens improve both mass and ion count accuracy. However, the application of some of the above steps also renders the prospects for proper rigour very remote and their use was therefore avoided or modified for the experiments conducted as part of this project.

The centroiding is easily switched off so that the data can be acquired in what is known as “continuum mode” or “profile mode”. Similarly the dead time correction and Z-focus lens can be switched off as an additional precautionary measure. The calibration was not deactivated, as mass estimates are not central to this project, although a discussion of its likely effects is provided in Section 2.7 below. Similarly, no attempt was made at foregoing the histogramming as this would result in unmanageably large data-sets.

2.4 The distribution of ion counts

The process responsible for the most fundamental variation in the observed peak intensity is the ionisation of a relatively small number of molecules out of the vast

number that actually passes through the mass spectrometer. Barring a comprehensive model that takes into account the position and velocity of all the particles involved, this process must be regarded as inherently random. In fact, because of the large number of molecules passing through the mass spectrometer and the very small probability that any given one is ionised, the resulting number of ions follows a Poisson distribution.

This fact is occasionally pointed out in the scientific literature although the implications seem not to be very widely studied. Part of the reason for this omission may be that due to the limitations of current technology, not all ions are counted when the rate of ion arrivals is large, which in turn distorts the Poisson distribution and diminishes the mass and ion count accuracy. This is “rectified” by some of the pre-processing steps, which improve mass and ion count accuracy, but do little to restore the Poisson distribution.

Putting aside for now the technological limitations, there are some immediate implications of having Poisson distributed data. For one, the pre-processing step of centroiding, which summarises the ion counts of a mass peak by their sum, may not be all that detrimental to statistical rigour. This is because the sum of independent Poisson distributed random variables is itself Poisson distributed. The information that is lost in the procedure primarily regards the velocity and spatial distributions of ions at the time the electric field was applied, which is of often of limited interest. Thus, if the technological limitations do not significantly distort the Poisson nature of the data and if there is no significant overlap of distinct mass peaks, centroiding may be a relatively harmless pre-processing step and the statistical models that will be used in Chapters 3 and 4 are in fact designed for use on centroided data.

Another important implication concerns our ability to estimate the Poisson rate parameter based on the observed ion count. It can be shown that the standard deviation of the Poisson distribution, relative to its rate, decreases with the square root of the rate, so that when the rate is higher it can be more accurately estimated. The relationship is illustrated on the plots below.

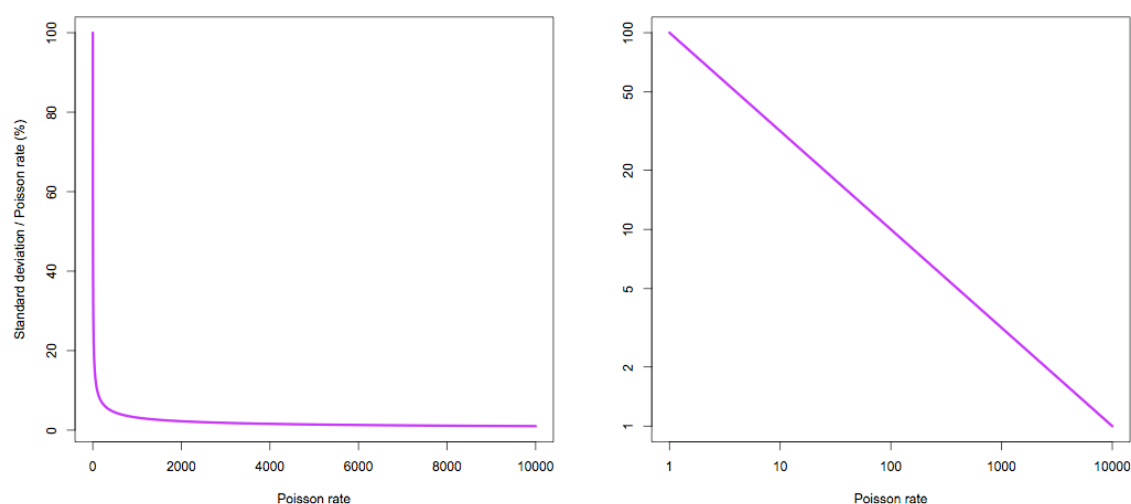


Figure 2.3 - The standard deviation of the Poisson distribution relative to its rate, as seen on a standard scale (left) and a log-log scale (right).

Perhaps the most obvious motivation for estimating the rate parameter is that it may be used to narrow down the possible identities of unknown metabolites, based on the observed isotopic abundance pattern. The Poisson rate associated with an unknown metabolite may to a first approximation be thought of as the product of the concentration of the corresponding metabolite, and a measure of its ionisation propensity. Since the latter is the same for all isotopic variants, the relative heights of the distinct isotopic mass peaks correspond to the relative isotopic abundances in nature. Because each possible chemical formula gives rise to a unique isotopic abundance pattern [41], the accuracy with which we can estimate the Poisson rate parameter, determines the degree to which we can narrow down the possible identities of the unknown metabolites. This will be discussed in detail in Chapter 4.

2.5 Ion digitisation

If a highly advanced mass spectrometer were available – one capable of measuring to an arbitrary degree of accuracy the numbers and arrival times of all incoming ions – then the development of a fairly comprehensive statistical framework for the analysis of LC-TOFMS data based on a simple Poisson distribution might be within reach. At present however, such a model is not appropriate, since the limitations of current

technology places numerous additional constraints on the accuracy of the data obtained.

The first step in the ion detection process lies in amplifying the electronic signal induced by the arrival of a single ion. For this task many mass spectrometers, including the Q-Tof Premier, make use of a Microchannel Plate (MCP) [56], which is an array containing a large number of miniature electron multipliers, positioned parallel to one another (see Figure 2.4). When an ion strikes the side of one of these multipliers it sets off a cascade of electrons, resulting in a strongly amplified signal. A single ion typically induces an output from the MCP of several million electrons [57] although the precise gain, and its distribution in time, can be quite variable [58, 59].

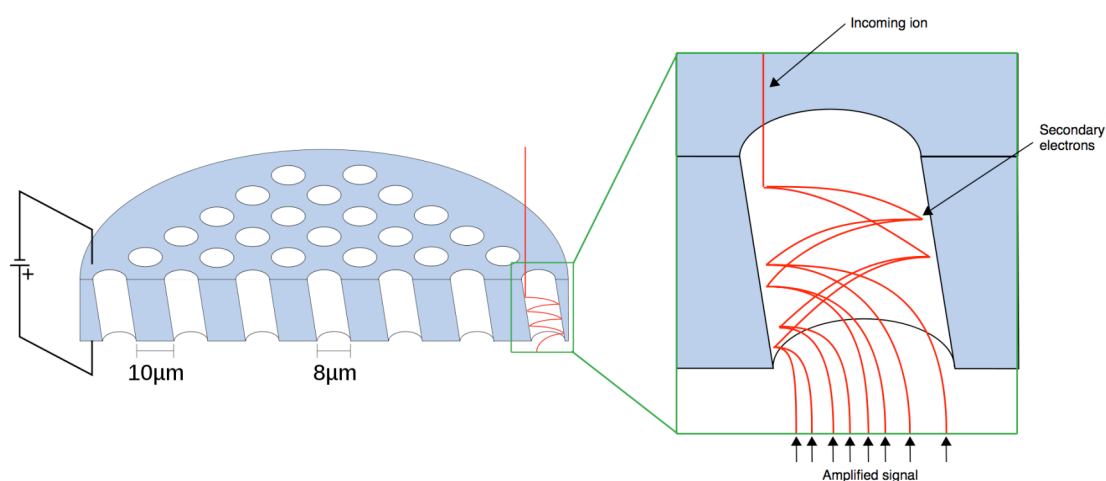


Figure 2.4 – Illustration of the workings of a Microchannel Plate. Note that the actual number of electron multipliers on an MCP is much larger than that indicated here. In addition it is common to employ two MCPs consecutively in a “chevron” formation.

Following its amplification, the signal must be digitised and for this task the Q-Tof Premier, makes use of a time-to-digital converter (TDC). The TDC monitors the strength of the signal passed to it and if it crosses a certain pre-specified threshold [60] the time of this “event” is recorded. TDCs are therefore fundamentally binary in that they do not record the strength of the signal but only determine whether it is greater than the threshold or not. For this reason they are also incapable of recording multiple (near) simultaneous ion arrivals - multiple ion arrivals will generally cause

the MCP to output a stronger signal, but the TDC will only record the one crossing of the threshold [61]. More specifically, in order for the TDC to resolve two ion arrivals, the voltage spike induced by the MCP after the arrival of the first ion must have receded to a level lower than the threshold. The time during which the TDC is incapable of recording further ion arrivals is referred to as the “dead time” and is on the order of 5 nanoseconds for the Q-ToF Premier. The smallest possible “tick” of the TDC clock used in this study is 278 picoseconds and any registered ion arrivals will be grouped into one of these discrete increments.

A widely used alternative to TDCs is the analogue-to-digital converter (ADC). This has a much wider dynamic range, as it records the strength of the raw signal passed to it by the electron multiplier. However this makes ADCs sensitive to the variable gain of the electron multiplier [59] and since they do not use thresholding like TDCs, they are constantly recording what amounts to electronic noise, which makes it difficult to distinguish single ion arrivals and tends to make the resulting data-sets rather large [62]. Additionally their time resolution is generally poorer than for TDCs so that they do not sample the mass peaks as finely as the latter do.

It is not possible to declare one type of digitiser as being generally superior to the other - ADCs work better at high rates of ion arrivals and TDCs work better at low rates [19]. However, the fact that TDCs can effectively block out electronic noise and generally are not affected by detailed characteristics of the MCP makes them much easier to model as they are capable of preserving the fundamental Poisson distribution of the ion counts *if the rate of ion arrivals is not too high*. At high rates of ion arrivals a pure Poisson model will break down and much of this thesis will be devoted to the development of probability distributions that can describe the data reasonably well in this scenario. Data digitised by an ADC will not be modelled or analysed in this thesis, although it is arguable that many of the concepts developed can be carried over to a system employing ADCs.

2.6 Histogramming and detector saturation

Because the length of the dead time is on the same order as the time-of-flight “width” of a mass peak [14], the mass spectrometer will often only detect one ion per mass

peak for each “pulse” of ions that is released into the free drift region. Nevertheless, a single scan as observed in the output data may well display ion counts in excess of 100. This is because a scan is in fact produced by histogramming hundreds or thousands of individual pulses [6] as shown on Figure 2.5 below.

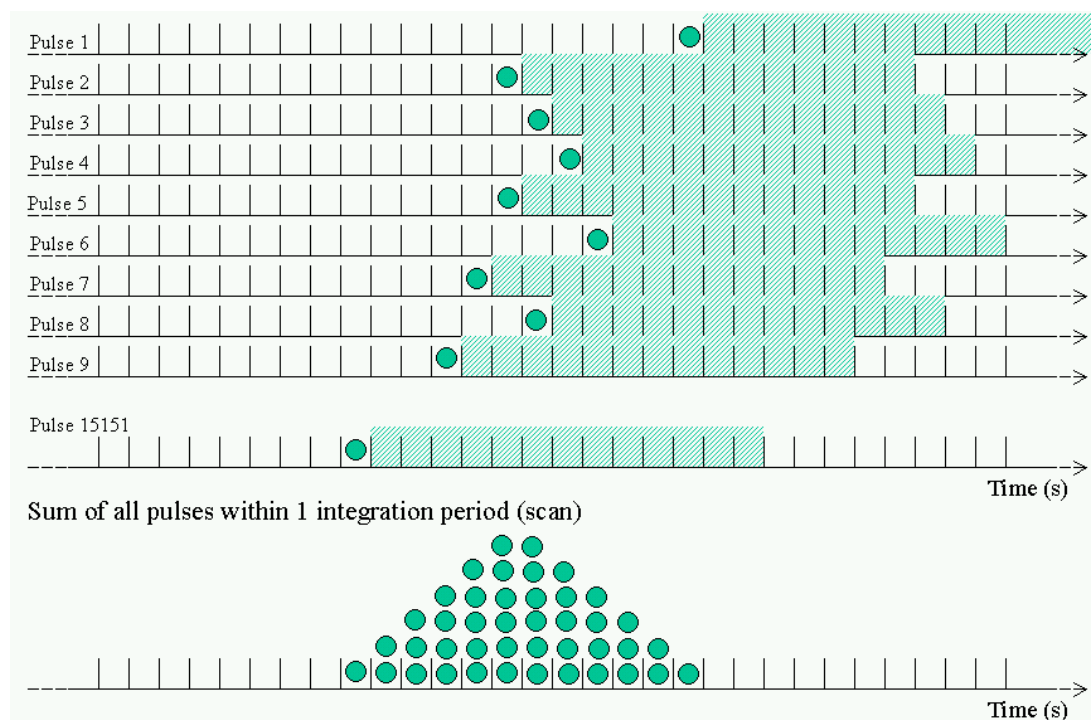


Figure 2.5 – Illustration of the construction of a scan, taken from [6]. The green dots denote the arrival of ions. The green rectangles denote the dead time, during which the detector is unable to record any further incoming ions. Note also the finite time resolution indicated by the division of each pulse into a series of boxes, each of which corresponds to a tick of the TDC clock that lasts 278 picoseconds.

While this pre-processing step does result in the loss of potentially useful information, it should be noted that the Poisson nature of the data is not directly distorted, since we are once again simply summing independent Poisson distributed random variables. As will be discussed in Chapter 5, if it can be assumed that the Poisson rate parameter is approximately constant over the segment of the chromatographic peak constituting a scan, the effects of the histogramming can easily be accounted for.

According to a Waters document [6] the sample concentrations normally associated with TOFMS analysis are such that the total number of ions arising due to a single

mass peak in a single pulse is usually 0 or 1. If this is the case then the finite time resolution and dead time should not cause serious distortions to the Poisson distribution. But if the sample concentration is very high, or a compound ionises very efficiently, then a significant number of ions will not get recorded. This not only worsens the ion count accuracy but also shifts the centre of the observed peak to a lower mass, as only the first ions to hit the detector tend to get recorded. The situation is illustrated on Figure 2.6 below.

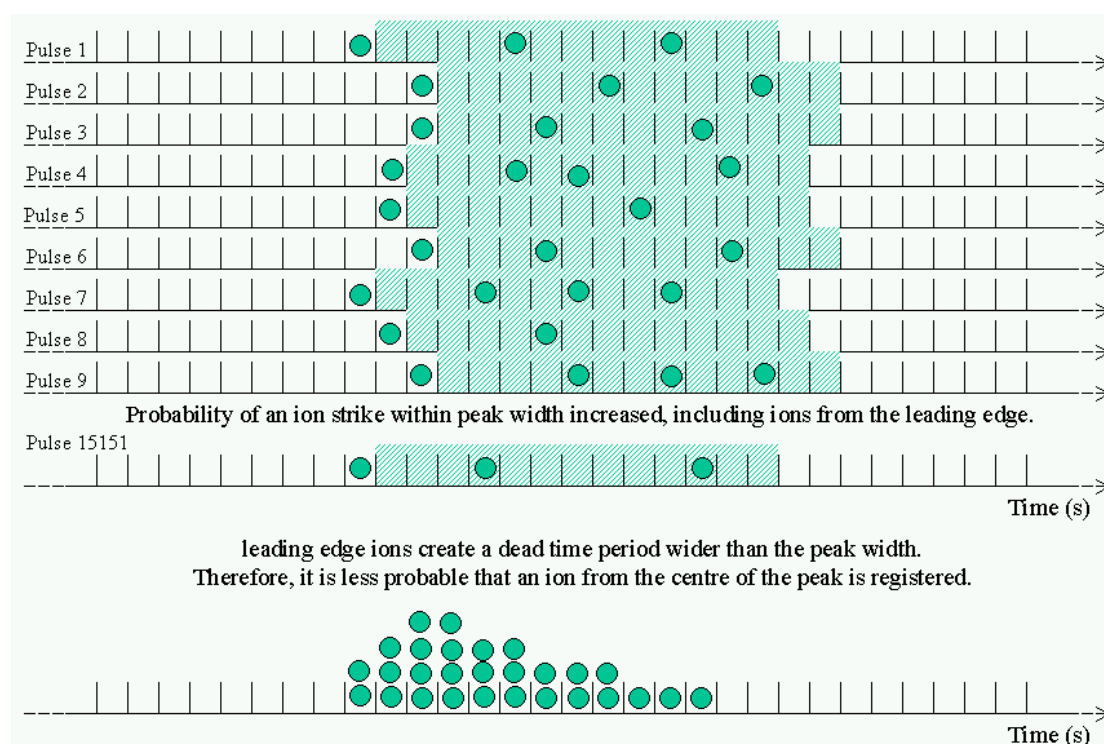


Figure 2.6 – Illustration of dead time distortion, taken from [6]. When more than one ion arrives per mass peak, it is often only the first ion that will be counted.

It is possible to correct for this type of distortion up to a point, via statistical methods and these will be discussed in detail in Chapter 5. But for sufficiently high ion counts it is very difficult to estimate the underlying mass peak with any meaningful degree of accuracy. For such mass peaks it is sometimes possible to work with less abundant isotopic variants or with the tails of the chromatographic peaks.

In addition to using statistical methods, Waters uses a device known as a “Z-focus lens” or “DRE lens” to address the dead time problem. The Z-focus lens works by

periodically attenuating the ion beam, so that only a small fraction (2.5% by default) of ions are transmitted to the detector. The observed ion count is then multiplied by 40, so as to make it consistent with the rest of the acquired data [63, 64].

While the Z-focus lens does reduce the dead time distortion, it is highly detrimental to the use of Poisson statistics. To illustrate why, suppose we had a mass peak with an associated Poisson rate of 100. If the resulting ions were attenuated to 5% by the Z-focus lens and the ion count then multiplied by 20, the distribution actually sampled from would be radically different from the original one, as shown on Figure 2.7 below. In order to avoid such complications, the use of the Z-focus lens was avoided in all experiments conducted as part of this thesis.

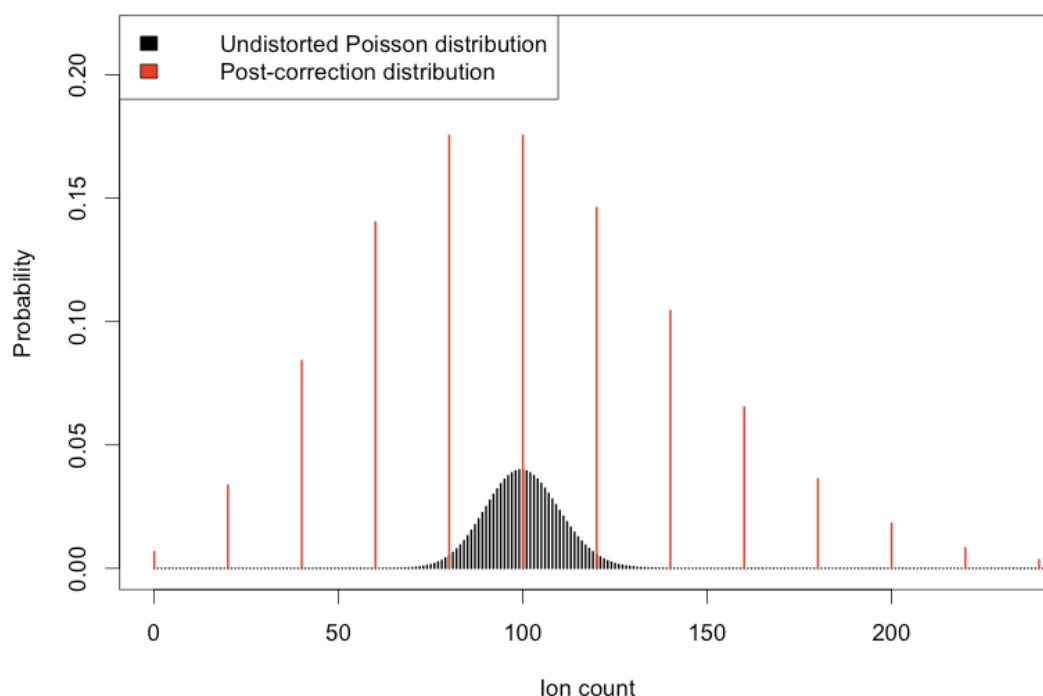


Figure 2.7 – Comparison of the true distribution of ions (black) and the one sampled from when the Z-focus lens is applied (red).

There appears to be limited general awareness of the implications of using a Z-focus lens, or indeed of the fact that it even exists. In one study of the reproducibility of LC-MS in which a Waters LCT Premier is used (which also employs the Z-focus lens) [25], the authors describe Poisson-like variations for lower ion counts, but variation

too great to be Poissonian for higher counts. The Z-focus lens is never mentioned, though it is clearly relevant and could very well contribute to the observed non-Poisson effects.

2.7 Mapping time-of-flight to mass

We will conclude the discussion of LC-TOFMS data with a review of the mechanism for mass, rather than ion count measurement. The study of mass estimates will not be critical to this thesis, but given that any pre-processing may potentially have confounding effects on other downstream analyses, it is prudent to undertake an assessment of the method used to estimate the m/z value from the time-of-flight by the Waters software.

Most textbooks on mass spectrometry will state that the equation used to relate time-of-flight to mass is

$$m/z = \frac{2eU}{d^2} t^2$$

where d denotes the length of the free drift region, e the elementary charge, and U the voltage [65]. The small molecules that are encountered in metabolomics rarely take multiple charges and so z can generally be assumed to be 1.

In reality the relationship between mass and time is considerably more complex [66] and calibration involving a compound of known mass is often undertaken continuously throughout an experiment. The manual for the Q-ToF Premier [67] states only that it uses a “polynomial equation in order to calibrate precisely over a wide mass range”. The Q-ToF Micro manual gives a rather more detailed account, identifying the relationship between the nominal mass (M_n), calculated via the classical TOF equation, and the calibrated mass (M_c) as

$$\sqrt{M_c} = A + B\sqrt{M_n + CM_n + DM_n^{3/2}}$$

where higher order terms can be added if required [68]. The values of all constants except B are set by the analyst through a calibration procedure described in the

manual. The value of B is changed repeatedly throughout an experiment by means of an automated calibration procedure known as “lockmass”, which works by periodically sending a substance of known mass through the mass spectrometer and picking the value of B that results in the correct value of M_c for this substance.

There is no guarantee that the calibration procedure for the Q-ToF Premier works in exactly the same manner as for the Q-ToF micro, but an analysis of empirical data acquired in continuum mode provides convincing evidence that the appropriate mapping is in fact approximately quadratic. The data acquired in continuum mode consists of the ion counts and masses of all detected ions. Plotting all detected m/z values for a single scan in the order in which they were acquired, results in the following curve:

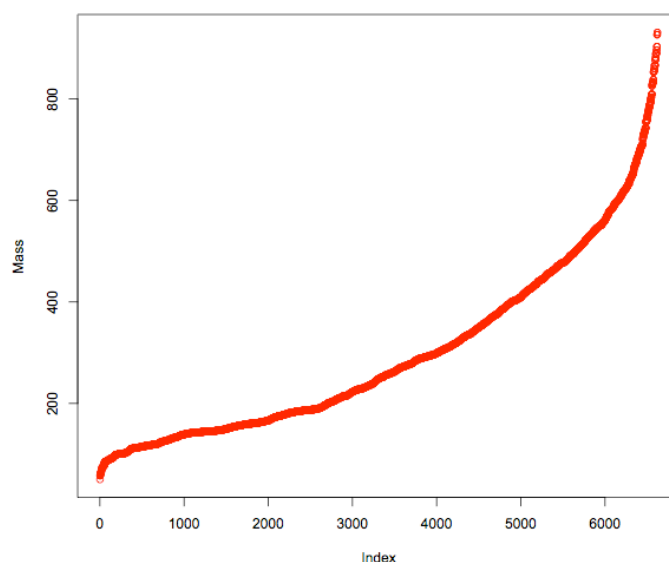


Figure 2.8 – The full set of masses detected in a single scan, arranged chronologically by time-of-flight. The index denotes the order in which the associated m/z value was observed.

While we know that a higher index corresponds to a longer time-of-flight, we do not know how many ticks of the TDC clock pass between two consecutive indices. If we had a way of stretching the x-axis according to how many ticks passed in between any two indices, we would immediately see the shape of the mass-time mapping.

It will be helpful to work with mass differences rather than with the masses. If we difference the series of detected masses – subtract the first from the second, the second from the third, etc – we obtain the plot shown below.

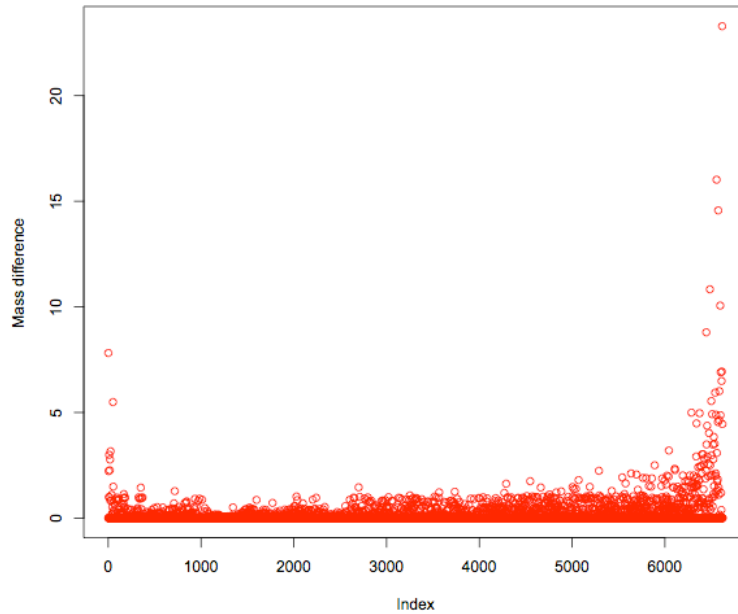


Figure 2.9 – The full set of mass differences detected in a single scan, arranged chronologically by time-of-flight.

By zooming in on the above plot to study the smallest detected mass differences, we see a set of distinct lines:

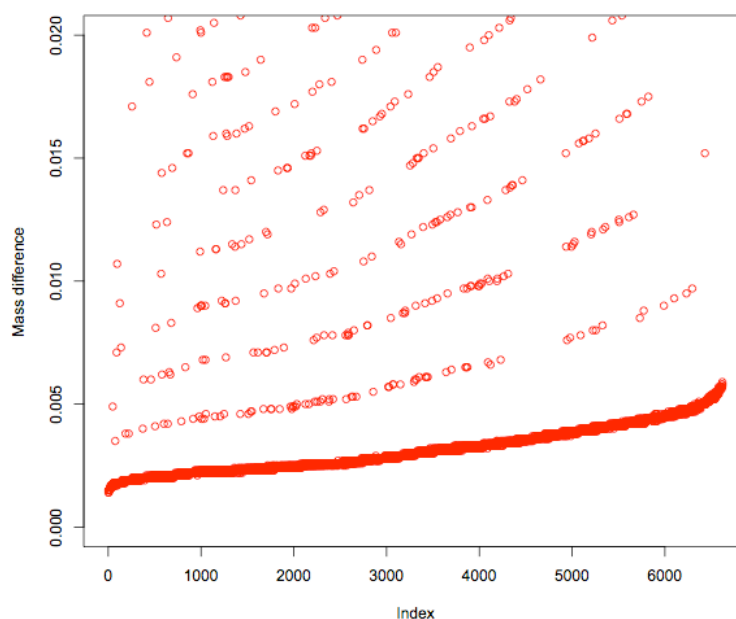


Figure 2.10 – The smallest detected mass differences.

The presence of the first line (from the bottom) could be explained if it represents all pairs of detected ions that are exactly one tick of the clock apart in the time dimension – a reasonable interpretation, since the line obviously represents the smallest detected mass differences over the full range of masses. Similarly the second line may represent all detected ions two ticks apart, and so on. If this interpretation is correct then the doubling of the mass differences should result in a new set of lines such that the first coincides with the second in the original set, the second with the fourth, and so on. As the plot below illustrates, this is indeed the case and the agreement between the two sets of lines is excellent:

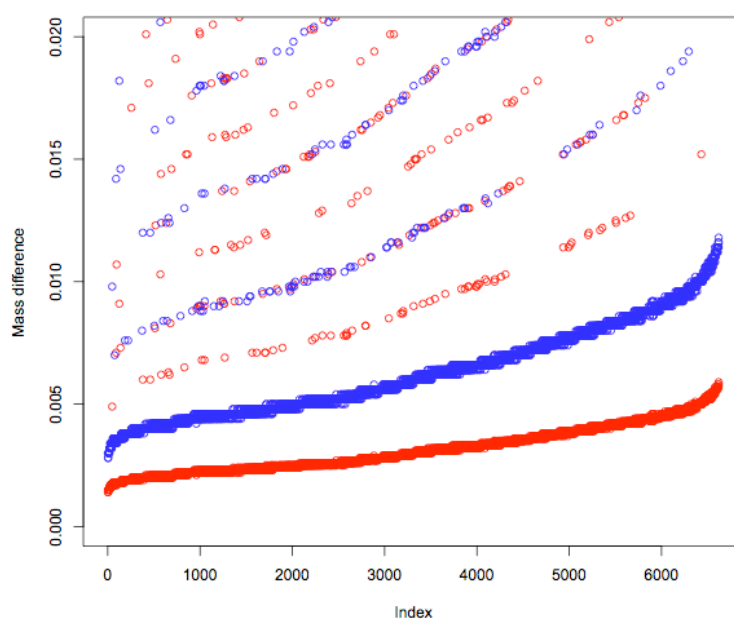


Figure 2.11 – The smallest detected mass differences (red) and their doubled values (blue).

We can now infer the true number of ticks in between all observed ions by going through the full list of mass differences and working out how many multiples of the nearest ‘minimal’ mass difference each one comprises. The detected masses are plotted against the ticks of the clock so inferred on Figure 2.12 below:

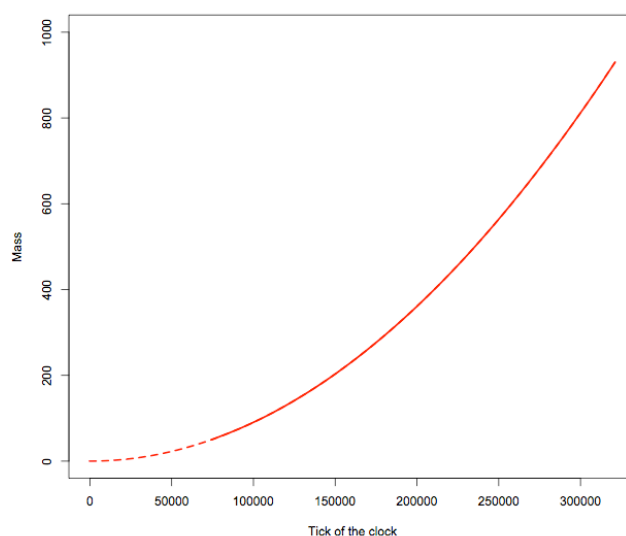


Figure 2.12 – The inferred shape of the mapping from time-of-flight to mass. Since the continuum data ranges from 50 to 1000 Da, the first segment (the dashed line) is inferred from the regression. The rest is the re-processed continuum data overlaid on the regression, however the two are almost indistinguishable.

A regression line fitted to the resulting data points shows excellent fit, indicating that the mass-time mapping is certainly extremely close to its classical quadratic form. This suggests that only rather limited pre-processing could have been applied to the recorded m/z values.

3 A Test for the Identification of Parent-Fragment Pairs

The means by which metabolites are identified in LC-TOFMS is in practice, primarily based on estimates of their masses as well as their observed chromatographic retention time. But as was mentioned in Chapter 2, estimates of the isotopic ratios can provide an entirely separate source of information that may enable us to better narrow down the identities of unknown metabolites. Yet another source of information comes from the so-called “fragmentation pattern” of a metabolite, that is, from making use of the information provided by those metabolites that break up into smaller pieces whose masses and isotopic ratios can also be estimated. A key challenge to the reliable use of this approach lies in establishing which fragments are derived from which metabolite when several distinct metabolites are coeluting from the column at roughly the same time. A statistically rigorous method for doing so, based on the Poisson statistics characteristic of mass spectrometers employing TDCs is presented in this chapter.

3.1 Elution profile

The fundamental property of LC-TOFMS data that allows us to identify related fragment ions in a rigorous manner is the similarity of their elution profiles in the chromatographic dimension. Clearly, the ion counts of related fragment ions will vary across their chromatographic peaks, but since the fragmentation of a compound takes place subsequent to its elution from the chromatographic column, it follows that the shapes of their chromatographic peaks are essentially the same up to a constant. It is noted that this relationship should also hold for distinct isotopic variants of any given compound and possibly also for other “derivatives” such as adducts and dimers.

Related publication:

Andreas Ipsen, Elizabeth J. Want, John C. Lindon and Timothy M. D. Ebbels. (2010). *A Statistically Rigorous Test for the Identification of Parent-Fragment Pairs in LC-MS Datasets*. Analytical Chemistry, **82**(5), 1766-1778.

Although some overlap in the elution times of distinct metabolites is quite common, only highly similar molecules (e.g. stereoisomers) would likely exhibit elution time differences small enough to make the constancy of the ratio of their chromatographic peaks comparable to those observed for related fragment pairs (see Figure 3.1). If a hypothesis test for the constancy of ion count ratios (referred to as “coelution test” in the following) could be designed, we would have a rigorous method to test for the presence of isotope/fragment-relationships. It might also help us to identify adducts and dimers.

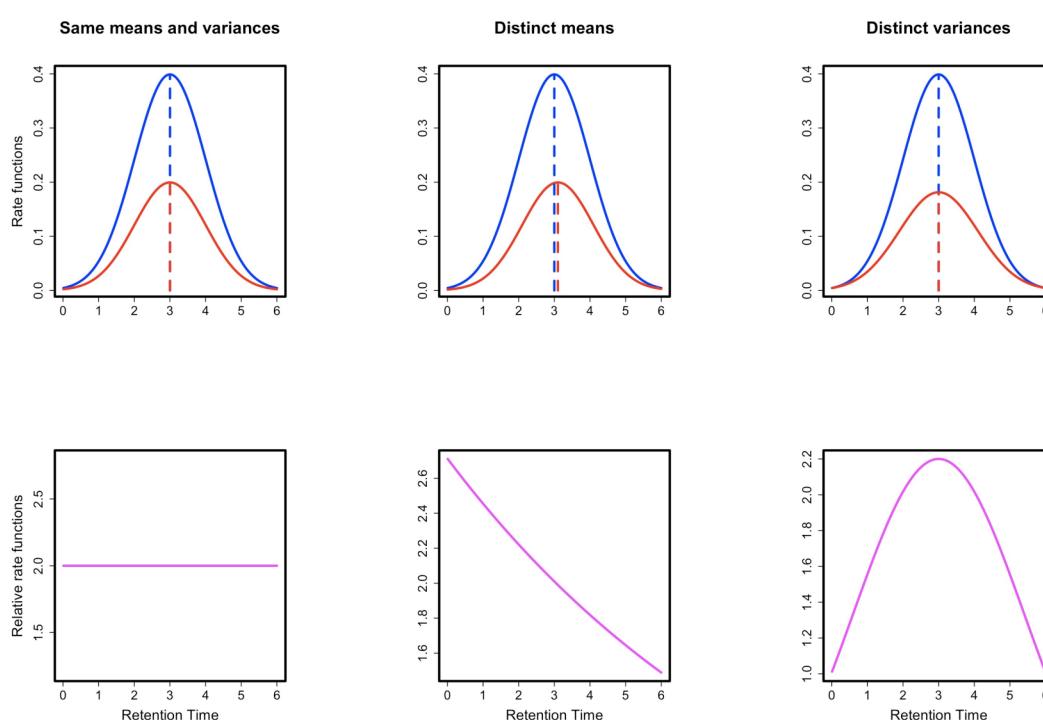


Figure 3.1 – Chromatographic peaks (modelled here as Gaussian) in the retention time dimension of two coeluting compounds (top) and the corresponding ratios (bottom). The different columns illustrate the scenarios of exact coelution (left) partial coelution (middle) and exact coelution but with distinct variances (right).

Waters are well-aware of the chromatographic relationship between related ions, as is seen in a patent application of theirs [69] where they write that

”The chromatographic retention time exhibited by each of [the fragment] ions and the chromatographic peak profile exhibited by each of these ions must exactly reflect both the retention time and peak shape of the originating molecule. [...]. The measured retention times and peak shapes of ions can deviate from the retention time and peak shape of their originating molecule, but these deviations must result only from the irreducible measurement error or from interference due to unrelated ions. Any such differences cannot arise from an intrinsic difference in retention time or peak shape between an ion and its originating molecule. Thus, the retention time and peak shape of the precursor ion and its fragment ions are intrinsically identical to each other and to the retention time and peak shape of the originating molecule”

The “irreducible measurement error” refers to the Poisson error, which we will aim to account for. “Interference due to unrelated ions” can arguably come in only two forms: from overlap in both the m/z and retention time dimensions of peaks associated with distinct molecular species and from ionisation suppression. The former cannot be accounted for by the coelution test, however overlap in both m/z and retention time dimensions is relatively rare. The latter will be tested for and discussed below. To the best of the author’s knowledge, these forms of interference would also confound all other methods currently available.

3.2 Current solutions

A number of heuristic methods have been developed to address the problem of identifying related fragment pairs. The one proposed by Waters in the abovementioned patent application works by comparing chromatographic peak profiles using neural networks, while another algorithm of theirs works rather more simply, by testing whether the apices of coeluting chromatographic peaks share the same retention time [70]. However, the most widely used approach [71, 72] is based on the Pearson correlation and its use can easily be illustrated with real data. The plots on Figure 3.2 below show the ion counts across the chromatographic peaks of two compounds that we suspect of being related fragments.

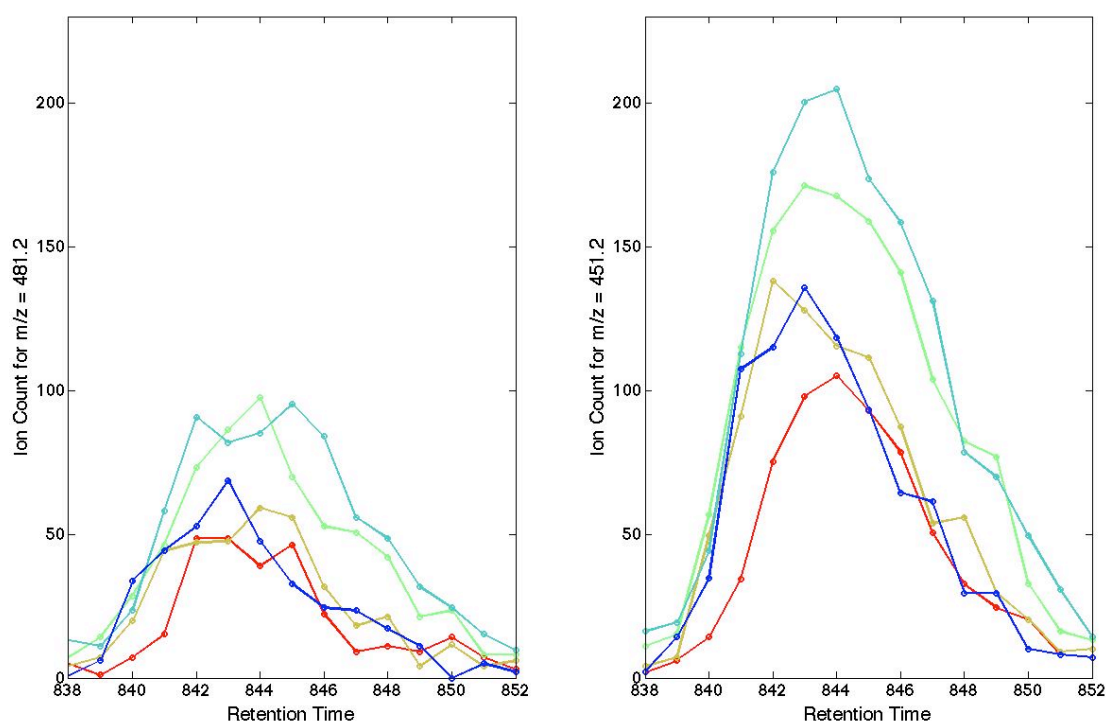


Figure 3.2 – Chromatographic peaks of two compounds that we suspect of being related. The different colours indicate different experimental samples.

We can pair up all matching ion counts for these two compounds within each sample, and plot them against each other. Since the ion count ratio is approximately constant we obtain a clear linear trend and accordingly a very high Pearson correlation coefficient, which can be regarded as evidence that the compounds are indeed related:

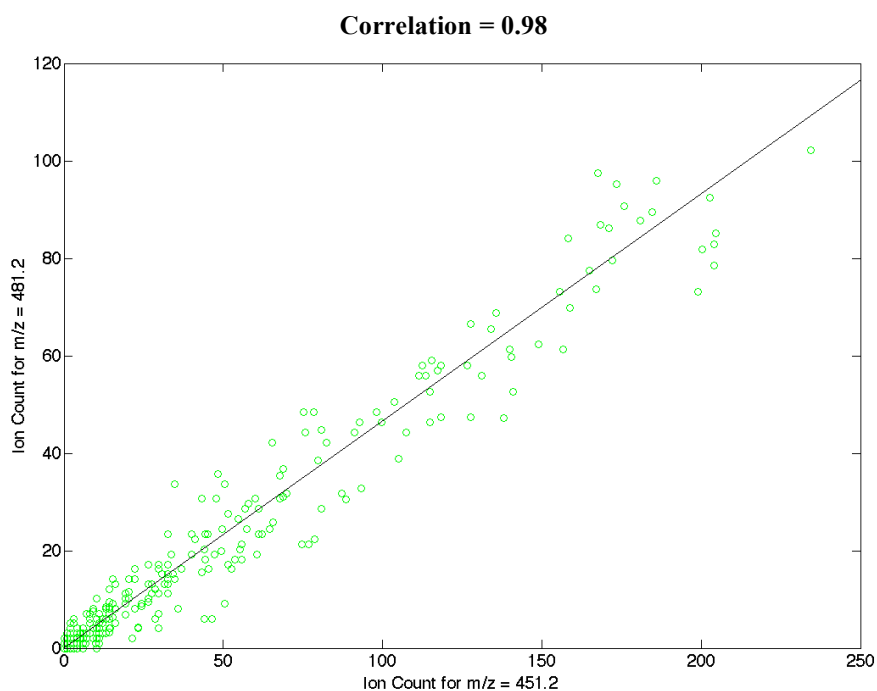


Figure 3.3 – Scatterplot of the ion counts of the matching mass pairs. Note that despite clear differences in the amplitudes of the different experimental samples, the ion count ratios are approximately the same.

But while all of the abovementioned methods are highly valuable to researchers, they are fundamentally heuristic in that they provide intuitive but poorly understood rules of thumb to data analysis. There are a number of reasons for avoiding the use of heuristic methods. Since they are not built on a comprehensive understanding of the system to which they are applied, they may, without warning, perform poorly under “unusual circumstances” (e.g. for very high ion counts). Moreover it is difficult to establish the “optimality” of such heuristic methods – they may be designed to maximise a score function (such as the correlation) that makes some intuitive sense but whose propriety for the system studied is difficult to ascertain since the system is not understood. Finally, if a score function is used to quantify the degree of peak similarity then some “acceptance threshold” for this score must be chosen, above which the peaks are deemed to stem from the same underlying metabolite, and below which they are deemed to be unrelated. The choice of this threshold is generally quite arbitrary, as the distribution of the score function is rarely understood.

3.3 Theory

3.3.1 Statistical Model

In the following it will be assumed that the chromatographic peaks studied are pure, that is, they are derived from only one molecular species. So long as the mass peaks are pure they may be centroided without the loss of important information, since the shape of the mass peak is largely determined by factors such as the velocity and spatial distribution of the ions at the time the electric field is applied [14], which are of little interest here. Moreover, as noted earlier, the count of a centroided mass peak will remain Poissonian as the sum of independent Poisson distributed random variables is itself Poissonian.

The probability of obtaining the count k from a Poisson distributed random variable with rate λ is given by

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

For LC-TOFMS data the rate of ion arrivals of a particular molecular species will be a function of its elution time, t , so that we may write

$$P(k(t)) = \frac{\lambda(t)^k e^{-\lambda(t)}}{k!}.$$

For a Poisson distribution, the rate parameter, λ , is equal to the mean. The *centroided rate function*, $\lambda(t)$, therefore describes the mean number of ion arrivals of a particular molecular species within one scan, as a function of retention time. The rate function may be regarded as the product of the concentration and the ionisation propensity of the compound and consequently may be written

$$\lambda(t) = \pi Q(t)$$

where π is a measure of the compound's ionisation propensity, and $Q(t)$ is a measure of its concentration in the retention time dimension. Supposing a metabolite were to

fragment into multiple ions after eluting from the chromatographic column, these would all share the same $Q(t)$ as the original metabolite.

3.3.2 A test of hypothesis for exact coelution

Suppose we wish to test whether a proposed fragment pair is “legitimate”, rather than the result of close but partial coelution. The two sets of ion counts will be referred to as $k_0(t)$ and $k_1(t)$ and their rate functions can be written $\lambda_0(t) = \pi_0 Q(t)$ and $\lambda_1(t) = \pi_1 Q(t)$. A matching pair of ion counts, $(k_0(t), k_1(t))$, will be referred to as a “data-point”. The joint distribution of the counts at a given retention time is then given by

$$P(k_0, k_1) = \frac{e^{-(\lambda_0 + \lambda_1)} \lambda_0^{k_0} \lambda_1^{k_1}}{k_0! k_1!}$$

where the dependence on t has been omitted for conciseness. But if

$$n = k_0 + k_1, \mu = \lambda_0 + \lambda_1 \text{ and } \rho = \lambda_0 / (\lambda_0 + \lambda_1)$$

we may, following Przyborowski and Wilenski [73], rewrite the joint probability as:

$$P(k_0, n) = \frac{\mu^n e^{-\mu}}{n!} \frac{n!}{k_0! (n - k_0)!} \rho^{k_0} (1 - \rho)^{n - k_0}$$

which is the joint probability of a Poisson distribution with mean μ (which determines the sum of the ion counts, n) and a Binomial distribution of n trials with probability ρ (which determines what portion of the sum is due to k_0 in particular). If the two ions under investigation exhibit exact coelution and hence share the same $Q(t)$ then this term cancels out from the expression for the binomial probability, which, reinstating the dependence on t , can be written

$$\rho(t) = \frac{\lambda_0(t)}{\lambda_0(t) + \lambda_1(t)} = \frac{\pi_0 Q(t)}{\pi_0 Q(t) + \pi_1 Q(t)} = \frac{\pi_0}{\pi_0 + \pi_1}$$

which will therefore be constant across retention time. Under the null hypothesis of constant Binomial probabilities, Pearson's chi-squared goodness-of-fit (GOF) statistic

$$\begin{aligned} x^2 &= \sum_{i=0}^1 \frac{(\text{Expected value of } k_i - \text{Observed value of } k_i)^2}{\text{Expected value of } k_i} \\ &= \frac{(k_0(t) - \rho n(t))^2}{n\rho(1-\rho)} \end{aligned}$$

approximates a χ^2 distribution with one degree of freedom (see for example Wackerly et al. [74], p682). We can evaluate this statistic for all N data-points across the chromatographic peak and sum them to obtain a pooled statistic, $X^2 = \sum x^2$, which approximates a χ^2 distribution with $N - 1$ degrees of freedom, since ρ must be estimated from the data. The approximation to the χ^2 distribution works best when n is large and ρ is moderate. A standard test of validity is to require that $n\rho \geq 5$ and $n(1-\rho) \geq 5$; data-points for which this is not the case should be left out or pooled together.

If there is partial coelution, so that the binomial probability varies from data-point to data-point, then the value of the X^2 statistic will generally be considerably greater than what would be expected by chance, a property that can easily be quantified by calculating the corresponding p-value. Therefore, given a set of ion counts across two coeluting chromatographic peaks, the coelution test indicates the probability of obtaining deviations from the estimated Binomial probability that are as, or more extreme than those observed, under the null hypothesis that the chromatographic peaks examined exhibit exact coelution. The fact that the test quantifies the uncertainty of this assignment with a relevant p-value is its main advantage over the alternative heuristic techniques that measure peak similarity by means of poorly understood score functions and consequently define their acceptance thresholds in an essentially arbitrary manner. The acceptance threshold of the coelution test may simply be given by the chosen significance level, which is easily interpretable and relates straightforwardly to the researchers aversion to false positives.

Statistics other than the one proposed above may be used to detect deviations from exact coelution. Just as the X^2 -statistic, the likelihood ratio test statistic (see for

example Wackerly et al. [74] p517) is approximately χ^2 -distributed and in addition has certain asymptotically optimal properties. However, it has been argued that the likelihood ratio test is less reliable than the X^2 -statistic for small sample sizes [75] so that considerably higher $n(t)$'s would be required by the former in order for the resulting p-value to be accurate. This is consistent with the author's findings from applying the two statistics to the counts of simulated chromatographic peaks. The X^2 -statistic is therefore preferable for low counts and since detector saturation becomes increasingly severe with higher counts, the use of the likelihood ratio test statistic is generally not recommended.

In order to calculate X^2 , the binomial probability must be estimated. A very simple estimator may be constructed by dividing the sum of all the counts of one ion by the sum of all the counts of both ions:

$$\hat{p} = \frac{\sum_i k_0(t_i)}{\sum_i k_0(t_i) + \sum_i k_1(t_i)}$$

If this estimator is used, then the overall computational requirements of the test will be very low, and generally comparable to those of the Pearson correlation. Computational efficiency is an important property, considering the size of typical LC-MS data-sets.

The coelution test can easily be generalised so that it simultaneously tests for the exact coelution of an arbitrary number of chromatographic peaks. When doing so it becomes necessary to work with multinomials rather than binomials, but the general approach, including the estimation of multinomial probabilities is closely analogous to the procedure just described. It should also be noted that confidence intervals can be constructed for the estimated probabilities. As will be discussed in the next chapter, this can be useful when the ions investigated are isotopologues, in which case the multinomial probabilities describe the isotopic abundance pattern.

It should nevertheless be noted that the coelution test is not a clustering algorithm – it does not identify each set of exactly coeluting peaks over a given range of retention

times. Like the Pearson correlation it can provide a similarity score (the p-value) to measure the degree of coelution of each pair of peaks, and on the basis of these scores a clustering algorithm might be constructed. The fact that the coelution test can be used to simultaneously test for the coelution of multiple peaks might well make the associated clustering algorithm faster than for methods that are based exclusively on pair-wise comparisons.

3.3.3 Illustrating the coelution test with simulated and real data

It may be instructive to illustrate the use of the coelution test on simulated chromatographic peaks. The coelution test makes no assumptions regarding shape of the chromatographic peaks under investigation, so we may for simplicity assume a Gaussian peak shape. The centroided rate function at the i th scan (lasting from t_{i-1} to t_i) can then be written

$$\lambda(i) = \int_{t_{i-1}}^{t_i} \frac{I}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

where I is the mean number of ion arrivals over the entire chromatographic peak. For given values of μ , σ and I , we may then simulate Poisson distributed random variables according to this model for each of the N scans over which the chromatographic peak is to be investigated. If two simulated chromatographic peaks share the same μ and σ then the result of applying the coelution test to their counts will be a p-value that is approximately uniformly distributed. A discrepancy in the μ 's, for instance, would tend to inflate the X^2 statistic and result in a correspondingly low p-value.

The data thus simulated in Figure 3.4 illustrates the ability of the coelution test to detect discrepancies from exact coelution ($p = 0.0079$) that are so small that the resulting Pearson correlation (0.9555) is essentially the same as what would be expected for exact coelution. In addition to the highly significant GOF p-value, the excessive deviation from the estimated binomial probability under partial coelution can be seen on the scatterplot from the comparatively large number of data-points with very low p-values - a feature that would be difficult to spot by eye, without the colour-code.

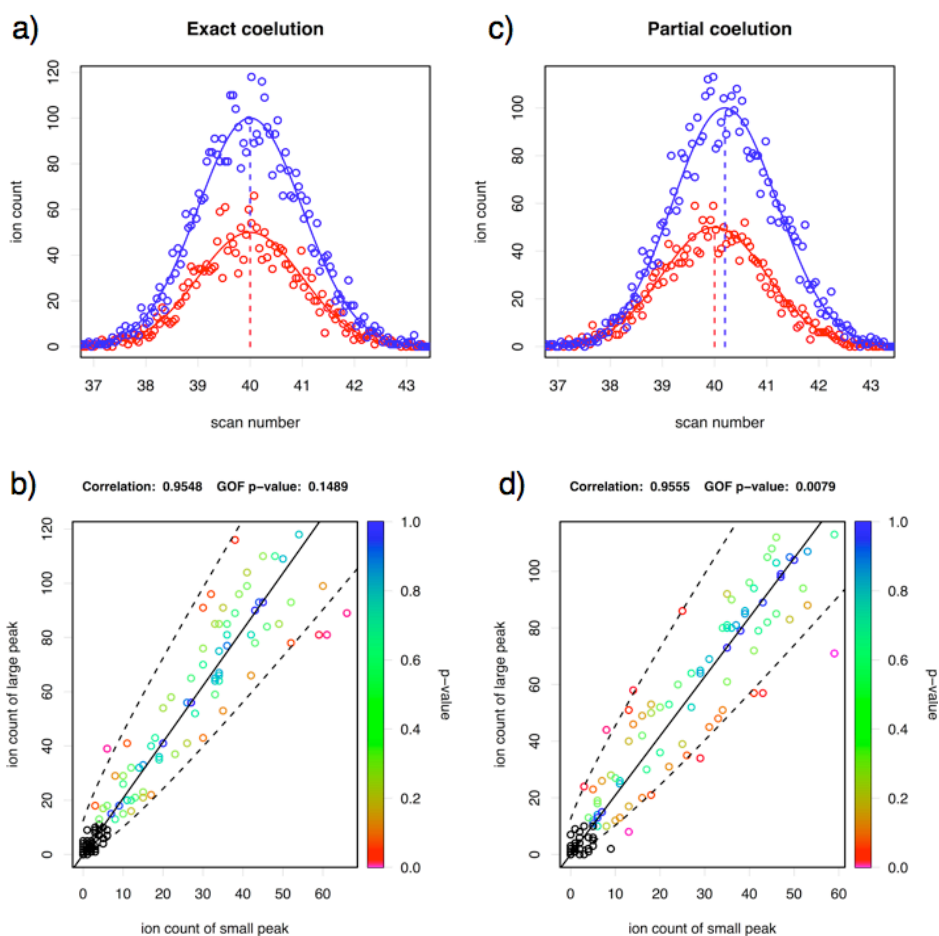


Figure 3.4 – Top: two simulated chromatographic peaks exhibiting exact coelution (a) and two simulated chromatographic peaks exhibiting very close but partial coelution (c) as indicated by the shifted means (10% of the standard deviation of the peaks). Bottom: The corresponding scatterplots with the p-values of the χ^2 -statistics of each data-point indicated by colour-code. Low counts for which the distribution of the χ^2 -statistics may deviate substantially from the χ_1^2 -distribution are excluded and these data-points are indicated in black. While the correlations are about the same in either scenario, the p-value of the pooled X^2 -statistic is highly significant under partial coelution ($p = 0.0079$), but quite moderate under exact coelution ($p = 0.1489$).

Figure 3.5 illustrates the same scenarios of exact and partial coelution using real metabolic data. As with the simulated data, the X^2 -statistic results in a moderate p-value under exact coelution ($p = 0.4426$) but a highly significant one under partial

coelution ($p < 10^{-7}$). In practice, when analysing metabolic samples, the total number of partially coeluting peaks, and the closeness of their coelution, may vary considerably depending on the nature of the sample and on the experimental setup. The coelution test is likely to be most valuable when the coelution is very close.

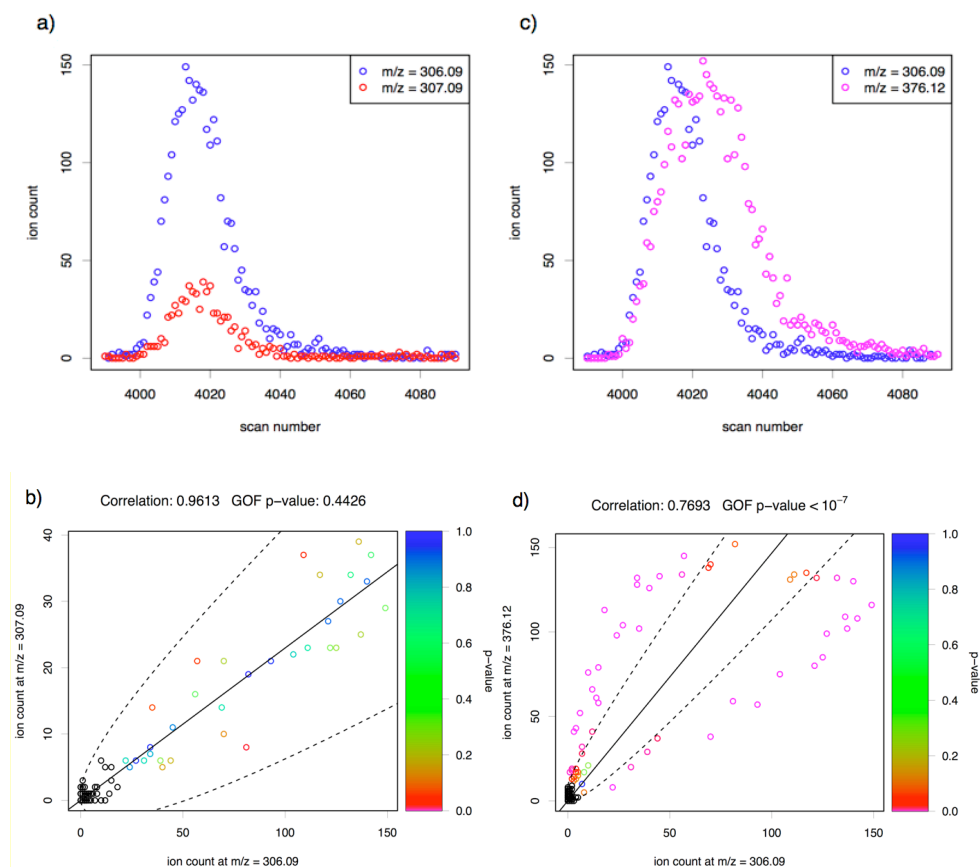


Figure 3.5 – Similar to Figure 3.4, but in this case using real LC-TOFMS data derived from a sample of synthetic urine, which will be described below. On the left is shown the chromatographic peaks (a) and scatterplot (b) of a pair of isotopologues, which, like related fragments, may be expected to exhibit exact coelution. On the right is shown the chromatographic peaks (c) and scatterplot (d) of two presumably unrelated compounds. The difference in the estimated means is 6.34 times the estimated standard deviation.

3.4 Experimental details

The validity of the theory described above rests on two fundamental assumptions:

1. That the recorded ion counts are approximately Poisson distributed.
2. That the ratio of the rate functions at the two m/z values investigated is approximately constant when there is exact coelution.

It is only if these two assumptions hold that we would expect the χ^2 - and the X^2 -statistics obtained from the observed ion counts of exactly coeluting compounds to adhere to the distributions predicted under the null hypothesis. Any departure from the two assumptions, due to detector saturation, interference from unrelated ions, or indeed, errors in the theoretical framework, would likely result in inflated statistics. Therefore the validity of the test may be evaluated by applying it to pairs of compounds known to exhibit exact coelution and by comparing the distributions of the resulting test statistics to the predicted ones. On that basis, the validity of the coelution test was evaluated on real metabolomic data, derived from synthetic urine.

The validity of the test was examined under varying ranges of ion counts, in order to evaluate the effects of detector saturation. Here it should be noted that in LC-TOFMS experiments the number of ion counts obtained will depend on the duration of the scan time - the longer the scan time is, the more pulses are histogrammed, and the higher the final count. However, the only factor affecting detector saturation is the rate of ion arrivals, so that a longer scan time does not induce greater detector saturation despite the higher count. Consequently ion counts will in the following be classified as 'low', 'moderate' or 'high', with these categories corresponding roughly to the tertiles of the full ion count range.

Preparation of synthetic urine. Eighty-three of the most abundant endogenous mammalian metabolites ranging in molecular weight from 30-625Da were weighed into a 1 L bottle and then dissolved in 1 L HPLC grade water (Sigma-Aldrich, St Louis, MO). Any remaining solids were removed by vacuum filtration. Approximate final metabolite concentrations were targeted to fall between 1 and 20 mM, with sodium azide added at 0.05% v/v as a preservative. The normally high levels of inorganic salts found in urine were not added in order to eliminate the effect of salt suppression in the various sample introduction interfaces. The stock solution was stored at -80°C.

Instrumentation. Synthetic urine samples (5µl) were injected onto a 2.1 x 100mm (1.7µm) HSS T3 Acquity column (Waters Corporation, Milford, USA) and eluted using a 18min gradient of 100% A to 100% B (A = water, 0.1% formic acid, B = acetonitrile, 0.1% formic acid). The flow rate was 500µl/min, the column temperature 40°C and sample temperature 4°C. Samples were analysed using a UPLC system (UPLC Acquity, Waters Ltd. Elstree, U.K.) coupled online to a Q-ToF Premier mass spectrometer (Waters MS Technologies, Ltd., Manchester, U.K.) in positive and negative ion electrospray mode with a scan range of 50-1000 m/z and a scan time of 0.08s. Three technical replicates were run. In order to obtain data that were as raw as possible, the spectrometer was run in continuum mode and the detector saturation correction was switched off. The Z-focus lens was also switched off.

All of this experimental work was carried out by Elizabeth Want.

3.5 Analysis

3.5.1 Selection of test data-sets

Clusters of isotopologues provide convenient test sets since they can be expected to exhibit exact coelution [69]. Eleven prominent clusters of isotopologues were investigated in this analysis. In all cases, continuum plots of the mass peaks (Figure 3.6) were closely inspected in order to reduce the risk of “contamination” from closely coeluting compounds with similar mass. While this procedure cannot guarantee the resulting data-set to be one made up exclusively of pure chromatographic peaks, it is quite conservative since any contamination would tend to inflate the resulting χ^2 -statistics which in turn would lead us to reject the validity of the coelution test. Heavier isotopologues were excluded if their signal was weak enough to be comparable to the background noise. A total of 16 peak pairs were included in the final analysis. The masses and scan numbers of their apices are listed in Table 1.

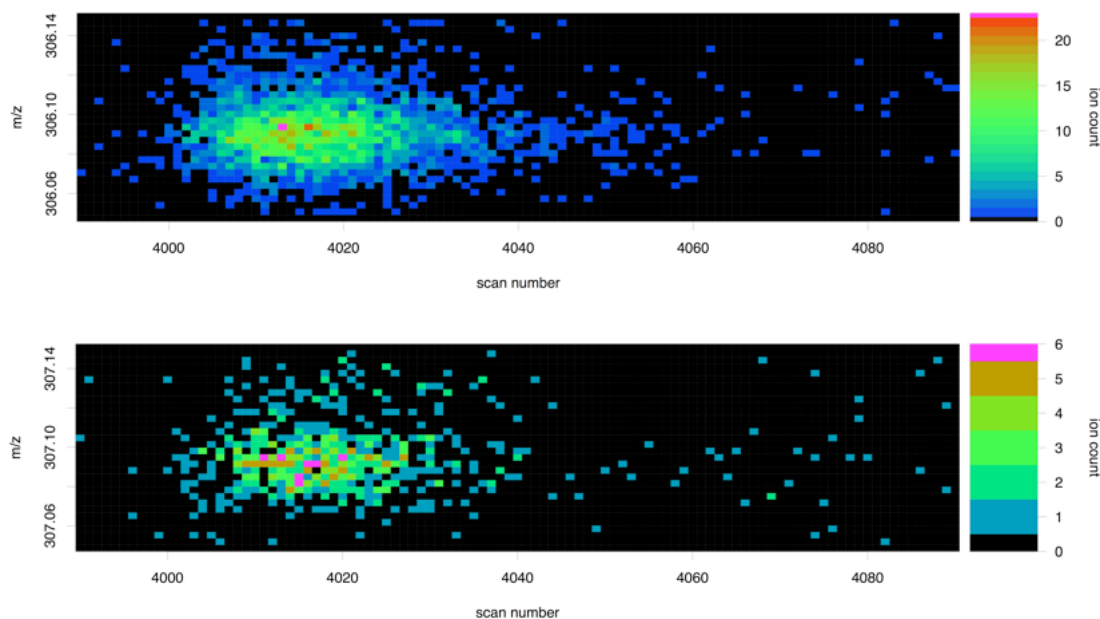


Figure 3.6 – Continuum plots of a pair of isotopologues. The x-axis indicates the scan number while the y-axis indicates each of the individual ticks of the clock that measures the time-of-flight of the ions, along with the corresponding m/z values. The number of ions counted at each tick is indicated by the colour-code. In these two cases there are no apparent signs of interference from other compounds of similar masses.

In order to allow for an approximately constant ratio of counts, all the pairs of isotopologue mass peaks that were accepted as “pure” were centroided over a matching number of mass bins. The binomial probabilities were then estimated and the resulting χ^2 -statistics calculated. Given the large number of data-points available for each pair of isotopologues, the distribution of each statistic should approximate a χ^2 -distribution with one degree of freedom. Thus, the validity of the algorithm may be tested by evaluating whether or not the empirically calculated χ^2 -statistics conform to this distribution.

Cluster	Compound	Scan number / Retention time (mins)	m/z	Isotopologue
1	N-Acetyl-L-glutamic acid	857/1.652 868/1.672 877/1.690	188.0426 189.0572 190.0615	[M-H] ⁻ [M+1-H] ⁻ [M+2-H] ⁻
2	Uridine	1096/2.114 1104/2.129	243.0537 244.0672	[M-H] ⁻ [M+1-H] ⁻
3	4-aminohippuric acid	1677/3.234 1681/3.241 1681/3.241	193.0507 194.0653 195.0645	[M-H] ⁻ [M+1-H] ⁻ [M+2-H] ⁻
4	Glutaric acid	1724/3.323 1755/3.382	131.0251 132.0381	[M-H] ⁻ [M+1-H] ⁻
5	Methylsuccinic acid	2471/4.763 2414/4.654	132.0384 133.0376	[M+1-H] ⁻ [M+2-H] ⁻
6	3-nitro tyrosine	2877/5.546 2871/5.535	225.0399 226.0579	[M-H] ⁻ [M+1-H] ⁻
7	Adipic acid	2952/5.689 2951/5.687	145.0464 146.0554	[M-H] ⁻ [M+1-H] ⁻
8	Indoxyl sulphate	2971/5.725 2965/5.714 2975/5.733 2972/5.727	211.9924 213.0030 213.9970 214.9968	[M-H] ⁻ [M+1-H] ⁻ [M+2-H] ⁻ [M+3-H] ⁻
9	Suberic acid	3635/7.007 3617/6.973 3623/6.985	173.0707 174.0842 175.0832	[M-H] ⁻ [M+1-H] ⁻ [M+2-H] ⁻
10	Salicylic acid	4096/7.895 4100/7.903	138.0239 139.0298	[M+1-H] ⁻ [M+2-H] ⁻
11	Sebacic acid	4615/8.893 4610/8.884	202.1097 203.1155	[M+1-H] ⁻ [M+2-H] ⁻

Table 3.1 – Scan numbers and m/z values of the peaks used in the evaluation of the coelution test at their apices. This may not correspond to the global maximum of the peak, since for many clusters parts of the chromatographic peaks were left out in order to avoid “contamination” from distinct compounds of similar masses. The type of isotopologue and the ionisation mode is also indicated with “[M-H]⁻” denoting the negatively ionised lowest mass isotopologue of the metabolite in question.

In order to evaluate the effects of detector saturation the coelution test was applied to the data at various cut-offs. The cut-off was applied to the sum of the paired ion counts rather than to each one individually, as this reduces the bias caused to the resulting distribution of χ^2 -statistics. In this way, three data-sets were constructed: the full data-set (6090 data-points), a data-set consisting of low and moderate paired ion counts, namely those of less than 600 (4029 data-points) and one consisting only of

low paired ion counts, namely those of less than 300 (2986 data-points). In all cases, low ion counts for which the χ^2 -statistics might be unreliable (those with $n\hat{p} < 5$ or $n(1-\hat{p}) < 5$) were excluded. Figure 3.7 shows the resulting GOF scatterplots for one pair of mass peaks (derived from 4-aminohippuric acid), with the approximate p-values of the χ^2 -statistics indicated by colour-code. The effects of detector saturation are very clear for the full dataset, where there is a very strong deviation from linearity and a correspondingly low GOF p-value. The deviation is difficult to see by eye for the data-set of low and moderate counts, though the GOF p-value remains significant. The dataset of low counts results in a moderate p-value and only a few data-points show possible signs of detector saturation.

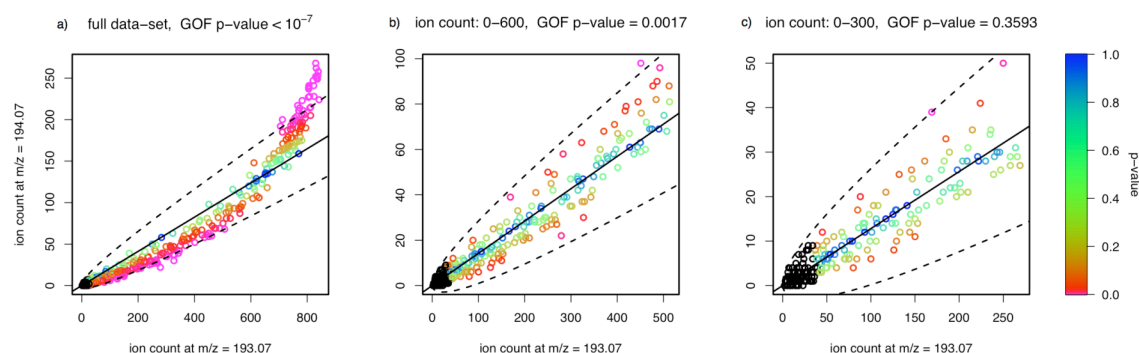


Figure 3.7 – Scatterplots for the three datasets derived from 4-aminohippuric acid: the full data-set (a), the one of low and moderate counts (b) and the one of only low counts (c). The approximate p-values of the χ^2 -statistics are indicated by colour-code and the p-values of the pooled χ^2 -statistics are listed.

3.5.2 Validation

To determine whether the χ^2 -statistics of the three datasets adhere to the χ_1^2 -distribution, quantile-quantile plots were drawn along with histograms of the corresponding p-values (Figure 3.8). The percentage of the χ^2 -statistics that fell within the 5% and 1% critical regions of the χ_1^2 -distribution were also calculated (Table 2). The full dataset clearly shows very strong deviation from the predicted distribution, with far more high values than would be expected from a χ_1^2 -distribution and consequently a distribution of p-values that is strongly biased towards lower values. The same is true of the dataset with the cut-off at 600, though here the deviation is more moderate. The final dataset however, appears to be consistent with a χ_1^2 -distribution.

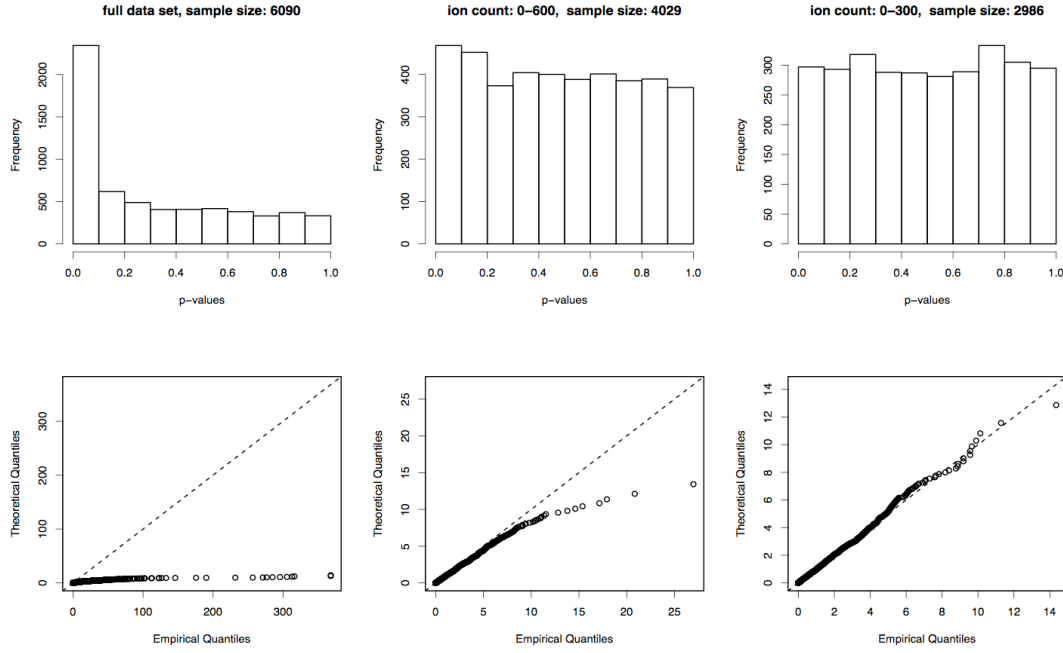


Figure 3.8 – Histograms of the p-values corresponding to the χ^2 -statistics derived from the three datasets (top) and quantile-quantile plots of the χ^2 -statistics themselves as compared with the theoretical χ^2_1 -distribution (bottom). Only the dataset of low counts appears to closely approximate the χ^2_1 -distribution.

To test the validity of the X^2 -statistics of the three datasets, all of the χ^2 -statistics of each group were summed, providing three X^2 -statistics. Their distributions should adhere to the χ^2 -distribution, with a degree of freedom equal to the total sample size minus 16 (the number of parameters estimated). As shown in Table 3.2, the two larger groups yield extremely low p-values, but the group with a cut-off at 300 yields a p-value consistent with a uniform distribution.

	Full data-set	Ion count: 0-600	Ion count: 0-300
Percentage of χ^2 -statistics in 5% critical region	31.13% (1896/6090)	6.45% (260/4029)	4.99% (149/2986)
Percentage of χ^2 -statistics in 1% critical region	21.51% (1310/6090)	1.51% (61/4029)	0.87% (23/2986)
GOF p-value for χ^2 -statistics	$<10^{-7}$	$<10^{-7}$	0.5642

Table 3.2 – The first two rows indicate the percentages of the χ^2 -statistics derived from the three datasets that fall within the 5% and the 1% critical regions. The last row shows the p-value for the χ^2 -statistic pooled from all of the χ^2 -statistics. Again, only the dataset of low counts returns values consistent with the expected theoretical distributions.

As was mentioned earlier there are statistical methods for correcting for the effects of detector saturation. While these do not restore the Poisson distribution of the data, they may extend the range over which the Poisson approximation is valid. In order to investigate such an effect, Coates' correction algorithm [52], which will be discussed in Chapter 5, was applied to the continuum data, after which it was centroided and the coelution test once more applied. The results are shown on Figure 3.9 and on Table 3.3.

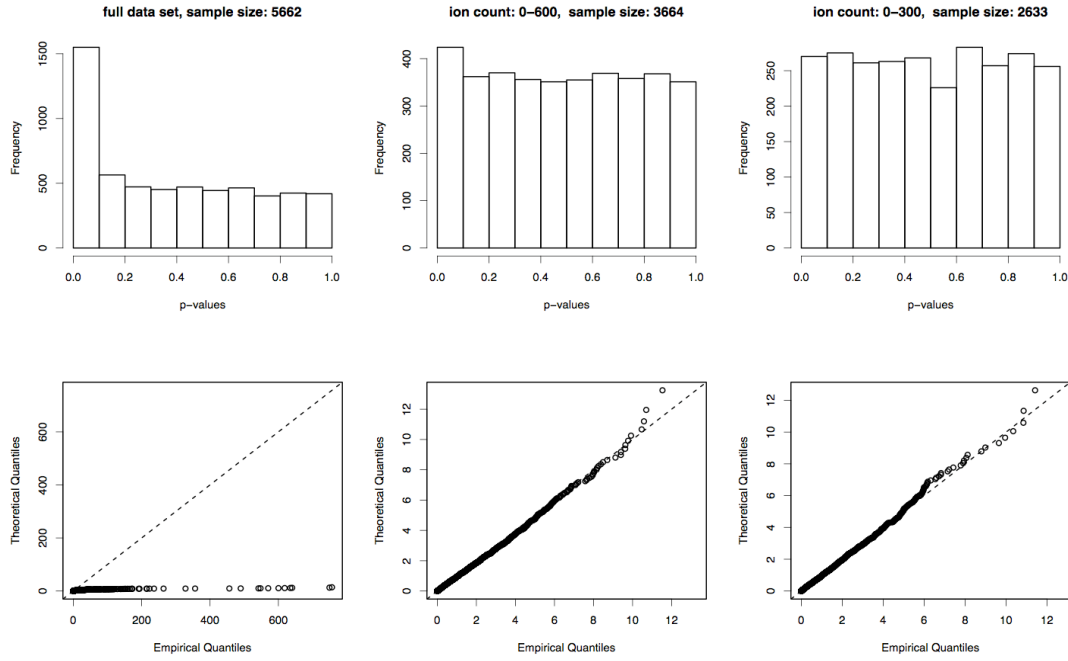


Figure 3.9 – Histograms of the p-values corresponding to the x^2 statistics derived from the three datasets after they had been corrected for detector saturation (top) and quantile-quantile plots of the x^2 statistics themselves as compared with the theoretical χ^2_1 -distribution (bottom). Only for the dataset of low and moderate counts does the correction appear to cause the distribution of the x^2 statistics to be substantially closer to the χ^2_1 -distribution than it was for the raw data, though some deviations remain.

Despite the correction, the distributions of the x^2 - and X^2 - statistics for the full dataset remain significantly different from the appropriate χ^2 -distributions. For the dataset of low and moderate counts the distribution of x^2 -statistics is made considerably closer to acceptable though the p-value of the X^2 statistic remains significant. There is no indication that the distributions of the statistics change substantially for the dataset of low counts.

	Full data-set	Ion count: 0-600	Ion count: 0-300
Percentage of χ^2 -statistics in 5% critical region	20.89% (1183/5662)	5.70% (209/3664)	5.24% (138/2633)
Percentage of χ^2 -statistics in 1% critical region	13.49% (764/5662)	1.06% (39/3664)	0.72% (19/2633)
GOF p-value for χ^2 -statistics	$<10^{-7}$	0.0074	0.3190

Table 3.3 – The first two rows indicate the percentages of the χ^2 -statistics derived from the saturation-corrected datasets that fall within the 5% and the 1% critical regions. The last row shows the p-value for the χ^2 -statistic pooled from all of the χ^2 -statistics.

It is noted that in cases where the ion count ratio is very close to one, the degree of detector saturation will be about the same for both counts, which will have the effect of preserving the approximate constancy of the ion count ratio even for high counts. This can potentially induce a lower than expected χ^2 -statistic for high counts and a correspondingly inflated p-value, however the application of Coates' correction algorithm increases the variance of the counts beyond what would be expected for Poissonian data, so that this form of bias does not arise.

An important issue regarding the validity of the coelution test is the manner in which the rate function of a given fragment may be influenced by the coelution of a distinct metabolite through ionisation suppression. If ionisation suppression were to reduce the rate functions of all the fragments studied by the same proportion, then the coelution test would remain valid, as the linear relationship between the fragments would be unaffected¹. However if the nature of ionisation suppression is such that the rate functions of different fragments are reduced by different proportions then even very mild suppression effects might induce the coelution test to return a low p-value when it should not. If this is the case and if such suppression effects are common under coelution, then the coelution test should only be expected to return a uniformly distributed p-value when a single metabolite is eluting, which would limit its value as an analytical tool.

¹ In fact it is for this reason that instabilities in the electrospray itself would likely not confound the coelution test – the rate functions of the various fragments would be altered by this, but very likely by the same proportion.

In order to test for the presence of such suppression effects, the coelution test was applied to the chromatographic peaks of a number of known parent-fragment pairs for which there was partial coelution with a distinct metabolite. Validating the coelution test in this setting is more problematic than for isotopologues because it requires the identification of the fragment and because of the risk that the fragment investigated may in fact be derived from the coeluting compound. As with the isotopologues, any contamination of the mass peaks from unrelated compounds would tend to inflate the resulting χ^2 -statistics, which in turn would lead us to reject the validity of the coelution test. In addition to validating the coelution test under partial coelution, this procedure illustrates how the test would in practice be applied to pairs of ions.

The parent-fragment pairs that were investigated are listed in Table 3.1. As with the isotopologues the continuum data were inspected, Coates' correction was applied and three data-sets of differing ion counts constructed. The chromatographic peaks of an example parent-fragment pair are shown on Figure 3.10 along with that of the coeluting compound.

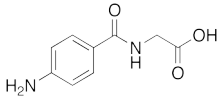
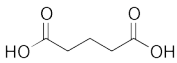
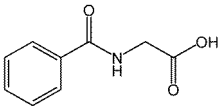
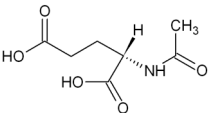
Compound	Chemical structure	Isotopologue (parent)	Neutral loss	Isotopologue (fragment)	Scan number
4-aminohippuric acid		$[M+2-H]^-$	Carbon dioxide	$[M-H]^-$	1741
Glutaric acid		$[M+1-H]^-$	Carbon dioxide	$[M-H]^-$	1785
			Water	$[M-H]^-$	
Hippurate		$[M-H]^-$	Carbon dioxide	$[M-H]^-$	3246
		$[M+H]^+$	Glycine	$[M+H]^+$	3266
N-Acetyl-L-glutamic acid		$[M-H]^-$	Water	$[M-H]^-$	4219

Table 3.4 – The parent-fragment pairs used to test for the presence of ionisation suppression effects on the coelution test. In all cases a distinct metabolite was coeluting with the above pairs. The scan numbers of the apices of the chromatographic peaks used in the evaluation of the coelution test are listed.

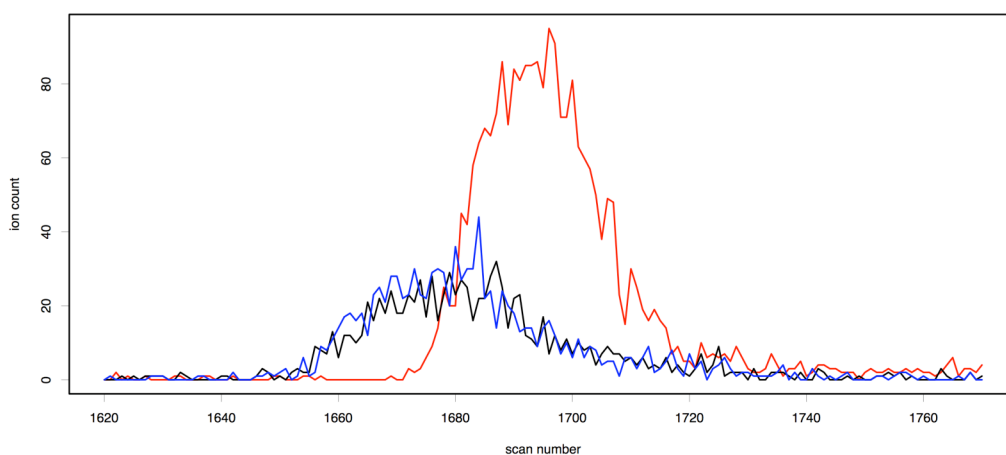


Figure 3.10 – The ion counts of 4-aminohippuric acid (blue), the fragment formed by the loss of Carbon dioxide (black) and a partially coeluting compound (red) used in the ionisation suppression test. If the rate functions of 4-aminohippuric acid and its fragment were reduced by significantly differing factors by the partially coeluting compound, we would expect their ratio to start shifting at around scan number 1680, but no such effect is observed.

The results are similar to those obtained for the isotopologues. As shown on Figure 3.11, the resulting χ^2 -statistics for the “low” ion counts below the first tertile are consistent with a χ^2 -distribution and the overall GOF p-value for the χ^2 -statistics is consistent with a uniform distribution at 0.3705. Thus there is no evidence that the coelution of distinct compounds affects the validity of the coelution test. Nevertheless, given that the detailed mechanics of ionisation suppression remain poorly understood we cannot decisively exclude the possibility that coeluting compounds could interfere in such a manner that the coelution test might be rendered biased. Any technique that identifies related fragments based on the degree of linearity of their ion count ratios would be adversely affected by such interference.

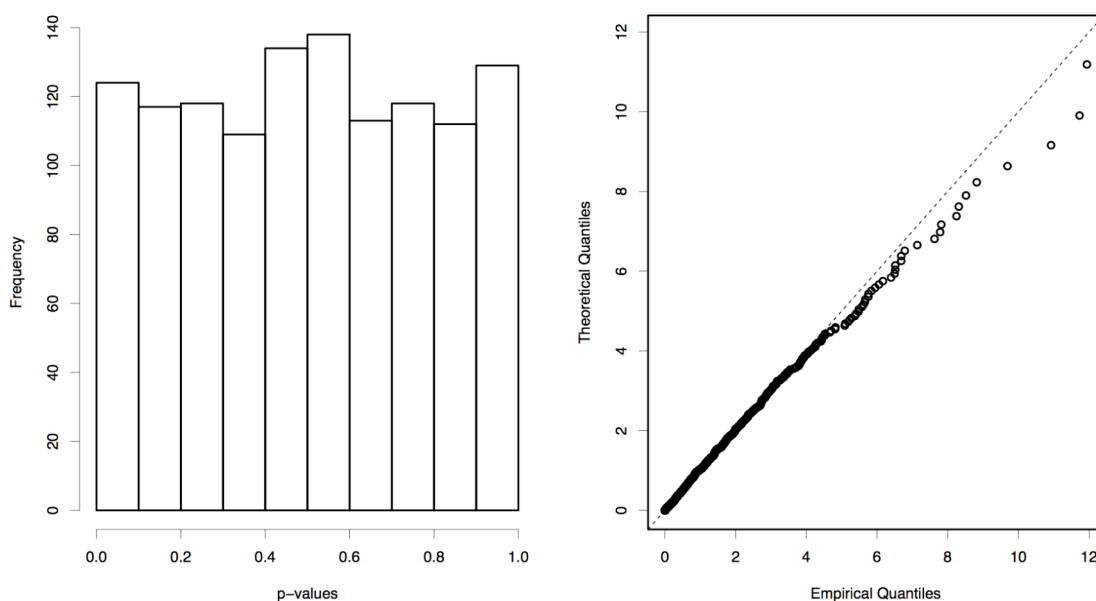


Figure 3.11 – Histogram of the p-values returned by the coelution test when applied to the low ion counts of the 6 parent-fragment pairs (left) and quantile-quantile plot of the corresponding χ^2 -statistics. The results are consistent with those obtained for the isotopologues and there is no evidence that the coelution of distinct compounds affects the validity of the coelution test.

3.5.3 Performance comparison

The false positive and false negative rates of the coelution test and of the Pearson correlation were compared using the data derived from the pairs of coeluting isotopologues. Since the null hypothesis of the coelution test is that the compounds investigated exhibit exact coelution, a false positive will occur when two exactly coeluting peaks are deemed to be partially coeluting whereas a false negative is the failure to classify two partially coeluting peaks as such. The set of isotopologues contains only exactly coeluting peaks, and so allows only for an evaluation of the false positive rate. However artificial coelution can be introduced by shifting the paired peaks away from each other scan by scan, thus producing semi-empirical data to imitate partial coelution, and allowing for the false negative rate of the two tests to be evaluated as a function of “increasingly partial” coelution. Although the coelution is introduced artificially, this procedure allows for a level of control over the degree of coelution that would be impossible to achieve through the use of purely raw data.

The significance level of the coelution test was set to 0.05, which is by definition also the theoretical false positive rate. Unlike the coelution test the theoretical false positive rate of the correlation approach very likely depends on factors such as the peak heights and shapes and cannot be straightforwardly linked to its “acceptance threshold”. In order to allow for a fair comparison, two values for the acceptance threshold of the correlation were investigated: one which resulted in an identical number of false positives for the two tests and another which resulted in a closely comparable number of false negatives across the retention time shifts. Coates’ correction was applied to the data and the cut-off of 300 ion counts was used. The retention times of the 16 pairs of exactly coeluting peaks were shifted by up to ten scans, giving a maximum retention time difference of one second, which is approximately a third of the median full width at half maximum of these chromatographic peaks.

The percentages of peak pairs deemed to exhibit partial coelution are shown on Figure 3.12 as a function of the retention time shift. The empirical false positive rate of the coelution test is 6.25% (1/16) and the empirical false negative rate reaches zero after a shift of just four scans. When the empirical false positive rate of the correlation is fixed to 6.25% its empirical false negative rate is substantially higher than that of the coelution test throughout the retention time range investigated and takes over ten scans to reach zero. When the false negative rates of the two tests are matched, the correlation has a false positive rate of 50% as opposed to the 6.25% for the coelution test. Clearly, the results suggest the performance of the coelution test to be considerably better than that of the correlation at these low counts. This is in addition to the inherent advantages that come with using a test of hypothesis.

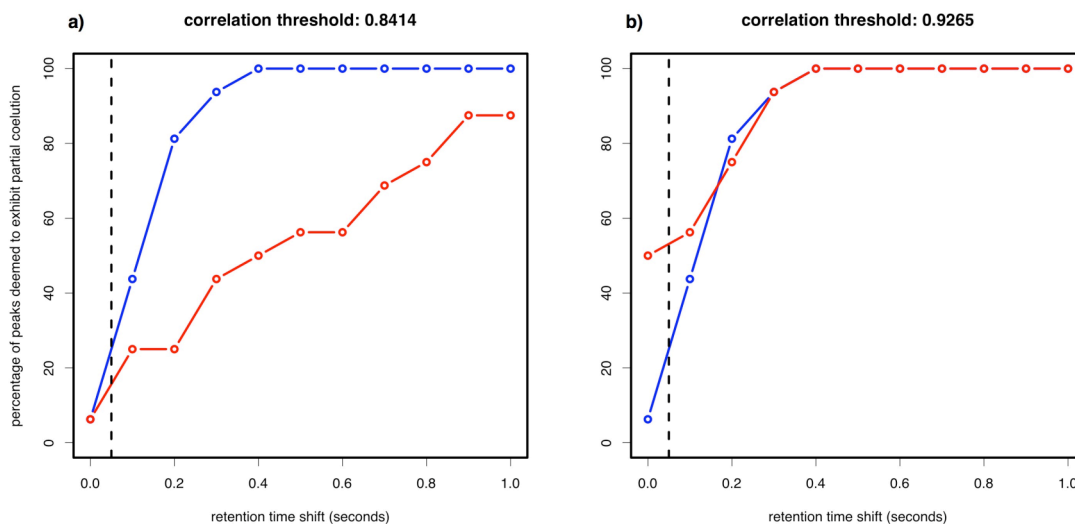


Figure 3.12 – Plots of the percentage of the isotopologue pairs that are classified as exhibiting partial coelution by the coelution test (blue) and the correlation (red), as a function of “increasingly partial” coelution. Only the leftmost point corresponds to exactly coeluting peaks and thereby indicates the false positive rate. False negative rates correspond to 100 minus the ordinate for non-zero retention time shifts. Plot (a) standardises the two tests by matching their false positive rates, while plot (b) matches their false negative rates. Clearly, the performance of the coelution test is considerably better than that of the correlation.

3.6 Discussion

The results suggest that when the analysis is restricted to low ion counts the coelution test provides a good approximation to a genuine test of hypothesis. Owing to detector saturation, the distribution of the test statistic used does not adhere exactly to the predicted distribution, but for low ion counts the approximation appears to be good enough for the p-values produced to be of satisfactory quality and here the test compares very favourably to a test based on the Pearson correlation. There is evidence that the range of ion counts over which the test is valid may be extended by applying Coates’ correction algorithm to the continuum data.

Given the requirement of undistorted Poisson distributed data, the coelution test is only expected to work with mass spectrometers employing a TDC to measure the

time-of-flight of the ions. It seems likely that a similar type of test based on the magnitude of the deviations from the estimated chromatographic peak ratio, might be designed for other types of detectors. The fact that low ion counts are required in order for the test to be valid represents another important constraint on its use. It is however one that can often be overcome by making use of the ion counts of mass peaks of less abundant isotopologues, or by restricting the sampling to the edges of the chromatographic peaks. Moreover, this constraint is likely to become less severe as the technology advances. It should also be noted that the coelution test is *a priori* just as valid as heuristic methods such as the Pearson correlation at higher ion counts. The p-value of the coelution test can still be used as a similarity score, only it can no longer be assumed to adhere to a uniform distribution under the null hypothesis and so its acceptance threshold will have to be chosen in a manner that is as arbitrary as for the heuristic methods.

The coelution test addresses one very specific problem in the analysis of LC-MS data. There are several software packages, such as XCMS [5], MetAlign [4], and Mzmine [3] that provide much more extensive tools for the analysis of LC-MS data. These are based on a rather different philosophy than the coelution test as they are generally designed to work with any type of LC-MS data, irrespective of factors such as the type of detector employed by the mass spectrometer. Nevertheless, the coelution test could quite easily be incorporated into software packages such as these, though given its requirements it should only be used when detailed knowledge of the instrument is available. While the coelution test is expected to become a useful analytical tool, it does also, in the opinion of the author, have considerable value as a proof-of-concept that the data analytical tasks encountered in the analysis of LC-TOFMS data can, in at least some cases, be addressed using a more rigorous mathematical approach.

4 Confidence Regions for Isotopic Abundance Patterns

It has long been recognised that estimates of isotopic abundance patterns may be instrumental in identifying the many unknown compounds encountered when conducting untargeted metabolic profiling using mass spectrometry. This measure is especially useful for detecting the presence of bromine or chlorine due to the highly characteristic isotopic distributions of those atoms, but even for compounds comprised solely of the most biologically abundant elements it provides information that can be crucial for effective formula identification.

However, while numerous methods have been developed for assigning heuristic scores to rank the degree of fit of the observed abundance patterns with theoretical ones, little work has been done to quantify the errors that are associated with the measurements made. Thus, it is generally not possible to determine, in a statistically meaningful manner, whether a given chemical formula would likely be capable of producing the observed data.

In this chapter, a method is presented for constructing confidence regions for isotopic abundance patterns that are measured with mass spectrometers employing TDCs and which therefore can make use of Poisson statistics. Moreover, a method for doing so is developed that makes use of the information from the measurements obtained across an entire chromatographic peak, as well as from any adducts, dimers and fragments observed in the mass spectra. This greatly increases the statistical power, thus enabling the analyst to rule out a potentially much larger number of candidate formulae while explicitly guarding against false positives.

Related publication:

Andreas Ipsen, Elizabeth J. Want and Timothy M. D. Ebbels (2010). *Construction of Confidence Regions for Isotopic Abundance Patterns in LC/MS Data Sets for Rigorous Determination of Molecular Formulas*. *Analytical Chemistry* **82**(17): 7319-7328.

In practice, small departures from the model assumptions are possible due to detector saturation, and interferences between adjacent isotopologues. While these factors form impediments to statistical rigour they can to a large extent be overcome by restricting the analysis to moderate ion counts and by applying robust statistical methods. Using real data, it is demonstrated that the method is capable of significantly reducing the number of candidate formulae, even when no bromine or chlorine atoms are present.

4.1 Background

The ability of a mass spectrometer to provide reliable estimates of isotopic abundance patterns is often regarded as being less critical than its mass accuracy. However a highly cited study by Kind & Fiehn [40] has demonstrated that if a hypothetical mass spectrometer with an accuracy of 0.1 ppm were available, it would be less successful at identifying unknown metabolites than a mass spectrometer capable of only 3ppm accuracy, but which were also capable of estimating isotopic ratios with a fixed accuracy of 2%. This would suggest that the extensive efforts put into improving the mass accuracy of mass spectrometers might be somewhat misplaced if good estimates of isotopic abundance patterns could be obtained instead.

Since all mass spectrometers produce errors in their spectral intensity measurements, a fundamental question that must be asked when exploiting isotopic abundance patterns is whether the deviation of a given theoretical isotopic abundance pattern from the observed abundance pattern is sufficiently small that it may realistically be attributed to the measurement error. If not, then the chemical formula to which the theoretical isotopic abundance pattern corresponds may be deemed to be inconsistent with the observed data and excluded from the list of candidate formulae. However, rather than addressing this question, most available methods attempt only to *rank* the degree of fit of all the feasible molecular formulae by means of various heuristic scores [76-78]. Other procedures simply assume that the observed isotopic ratios are accurate to within a few percent [40], but this is somewhat imprecise, as the accuracy depends on numerous factors, including the spectral intensity and the type of detector system used.

Therefore, while these heuristic methods can be extremely useful analytical tools, they do not enable the analyst to quantify, in a statistically meaningful manner, the range of molecular formulae that could realistically have produced the observed isotopic abundance pattern. The preferred method for doing so, according to classical frequentist statistical theory, would be through the construction of a confidence region, which, by definition, would cover the true parameter values, say, 95% of the time. However, the construction of such intervals requires a detailed understanding of the fundamental distribution of the data, which will in turn be dependent on the type of mass spectrometer used as well as the forms of pre-processing that are applied to the data. These issues have not, to the knowledge of the author, been considered in detail by any existing algorithms.

4.2 Theory

4.2.1 Statistical Model

The model that will be used is essentially the same as the one developed in the previous chapter, but adapted to a different problem. As before, we will work with centroided mass peaks and it will be assumed that the peaks studied are comprised of only one metabolite, which may be referred to as M . Let us suppose that there are $s + 1$ isotopologues of M , so that we may refer to them as $M_0, M_1, \dots, M_s$.

According to the Poisson distribution, the probability of obtaining the count k_i , for the isotopologue M_i is given by

$$P(k_i) = \frac{\lambda_i^{k_i} e^{-\lambda_i}}{k_i!}$$

where the parameter, λ_i , denotes the mean number of ion arrivals of the i th isotopologue, M_i , within one scan. Consequently, the probability of obtaining the sequence of counts $k_0, k_1, \dots, k_s$, from the full set of isotopologues can be written

$$P(k_0, k_1, \dots, k_s) = \prod_{i=0}^s \frac{\lambda_i^{k_i} e^{-\lambda_i}}{k_i!}.$$

Each of the λ_i in the above expression governs the absolute number of ion arrivals of the corresponding isotopologue, so that a total of $s + 1$ parameters are required. However, when investigating isotopic abundance patterns, we are interested in the relative, rather than the absolute numbers of ion arrivals. We may therefore work with the distribution of the ion counts at the various isotopologues, *conditional* on the total number of ion arrivals. If

$$n = \sum_{i=0}^s k_i \quad \text{and} \quad \rho_j = \frac{\lambda_j}{\sum_{i=0}^s \lambda_i}$$

then the conditional distribution that we seek may be written:

$$P(k_0, k_1, \dots, k_s | n) = \frac{P(k_0, k_1, \dots, k_s)}{P(n)} = \frac{n!}{\prod_{i=0}^s k_i!} \prod_{i=0}^s \rho_i^{k_i}$$

which is a multinomial distribution with n trials and probabilities $\rho_0, \rho_1, \dots, \rho_s$ where ρ_i is the isotopic abundance of M_i . Since electrospray ionisation involves the addition or removal of a proton from the underlying metabolite, a hydrogen ion should be added to or removed from the formula from which the ρ_i are calculated. Adducts may also be produced as will be discussed below.

4.2.2 Confidence Regions

Confidence regions may be constructed by exploiting the fundamental duality between tests of hypotheses and confidence regions, whereby the confidence regions is defined as the set of parameter values that are *not* rejected by the corresponding test of hypothesis. Several methods are available for constructing confidence regions around multinomial proportions and while no one method is universally accepted as being optimal in all circumstances, the one based on Pearson's χ^2 test, also used in the previous chapter, is arguably an uncontroversial choice. The statistic which in this case must be “inverted” can be written:

$$\begin{aligned}
 x_{(M)}^2 &= \sum_{i=0}^s \frac{(\text{Expected value of } k_i - \text{Observed value of } k_i)^2}{\text{Expected value of } k_i} \\
 &= \sum_{i=0}^s \frac{(np_i - k_i)^2}{np_i}
 \end{aligned}$$

where the p_i indicate the multinomial parameters that are being tested. If, for all i , $p_i = \rho_i$, then $x_{(M)}^2$ approximates the χ^2 -distribution with s degrees of freedom. Thus, given the counts $k_0, k_1, \dots, k_s$, a 95% confidence region, can be defined as the set of p_i for which $x_{(M)}^2$ is less than or equal to the 95th percentile of the χ_s^2 -distribution. As before, the procedure may be unreliable when the counts are very low, but scans for which $np_i \leq 5$ for any i can be pooled together.

Note that owing to the dependence between the p_i , the confidence region defined above cannot be expressed as a set of intervals around each of the estimated probabilities. Rather, the shape of the confidence region is ellipsoidal, which can make its interpretation rather awkward, depending on the physical nature of the multinomial probabilities. A number of procedures have been developed for constructing “simultaneous confidence intervals” which can be expressed as a simple set of intervals around each of the estimated probabilities [79]. But while this can indeed facilitate the interpretation, it also makes the resulting confidence region larger than it needs to be, reducing the statistical power of the test. Moreover, when the purpose of the study is formula elucidation, where there are a finite number of possible multinomial probabilities and the aim is simply to narrow them down as far as possible, any extension of the confidence region seems difficult to justify.

It is therefore arguable that the most appropriate method for constructing confidence regions for isotopic abundance patterns is the one based directly on the ellipsoid described above. In practice, this will entail conducting a test of hypothesis based on the $x_{(M)}^2$ statistic for all chemically realistic formulae that are consistent with the observed mass estimate. The $x_{(M)}^2$ statistics must be calculated using the multinomial probabilities that correspond to the known isotopic abundance patterns of the candidate formulae. While the total number of formulae for which the statistic must

be calculated may be large, depending on the mass accuracy, each individual calculation requires very little computational power.

4.2.3 Pooling information

It has long been understood that improved estimates of both masses and isotopic abundance patterns may be obtained by combining the measurements obtained across a compound's chromatographic peak. However, the procedure by which the data are pooled must be chosen carefully if a valid confidence region is to be constructed for the combined data-set. Moreover, in order to make full use of the information in the acquired data-set, the pooling procedure should ideally be generalised to incorporate the observed isotopic abundance patterns of any adducts, fragments or dimers of the compound of interest.

Since the power of Pearson's χ^2 test increases with the sample size, a higher value of n will reduce the volume of the confidence region and allow us to exclude a larger number of chemical formulae. However, owing to the risk of detector saturation, we cannot apply the test to scans with high counts, as these generally do not adhere to the Poisson distribution. Fortunately there are a number of ways of reducing the volume of the confidence region without using high counts.

The χ^2 -distribution has the very useful property that if the statistic X adheres to the χ^2_A -distribution and the statistic Y adheres to the χ^2_B -distribution, then $X + Y$ adheres to the χ^2_{A+B} -distribution. We may therefore calculate the $x^2_{(M)}$ statistic for each of the scans, obtained from the metabolite M , and sum the resulting $x^2_{(M)}$ statistics, to obtain a pooled statistic, $X^2_{(M)}$. If we have a total of $N_{(M)}$ $x^2_{(M)}$ statistics, then $X^2_{(M)}$ approximates the χ^2 -distribution with $N_{(M)}s$ degrees of freedom, under the null hypothesis that the multinomial probabilities $p_0, p_1, \dots, p_s$ used in calculating the x^2 reflect the true isotopic abundance pattern of M . Scans for which at least one isotopologue produces counts that are high enough to induce substantial detector saturation, should be left out. The more counts pooled in this manner, the greater the power of the test, so this is a rare scenario in which broader chromatographic peaks are desirable, although of course this is entirely dependent on them not having any overlap with other peaks.

There is in fact a rather more straightforward way to pool the data. The multinomial interpretation of the ion counts of the isotopologues applies to all of the scans that comprise a chromatographic peak. These multinomials differ in the number of trials, n , but they all share the same probabilities, which are governed by the same isotopic abundance pattern. Therefore the counts derived from each isotopologue may simply be summed, reducing the entire data-set to the outcome of a single multinomial distribution with a potentially very large number of trials. While this method of pooling the data is simpler and has greater statistical power than the one based on summing the $x^2_{(M)}$ statistics, the latter method has the advantage of being capable of providing a p-value associated with each scan. As will be shown below this turns out to be very useful when constructing confidence regions that are robust to small departures from the model assumptions, as are often encountered in practice. We will therefore use this latter procedure in the following.

It is possible to further constrain the confidence region by exploiting the information that is contained in the isotopic abundance patterns of “derivatives” of the compound being investigated, such as adducts, fragments and polymers, which are frequently observed in LC-MS experiments. Consider a derivative, D , which has been definitively identified in this manner and which has the isotopologues $D_0, D_1, \dots, D_t$. As with the underlying metabolite, M , we may calculate the $x^2_{(D)}$ statistic associated with a proposed set of multinomial probabilities, $q_0, q_1, \dots, q_t$ for a given scan:

$$x^2_{(D)} = \sum_{i=0}^t \frac{(nq_i - k_i)^2}{nq_i}$$

and we may sum the $x^2_{(D)}$ statistics obtained over, say, $N_{(D)}$ scans to obtain $X^2_{(D)}$. Again, if the q_i correspond to the true isotopic abundance pattern of the derivative, the distribution of $X^2_{(D)}$ will approximate a χ^2 -distribution with $N_{(D)}t$ degrees of freedom. We can therefore easily combine it with the $X^2_{(M)}$ statistic to obtain a single final statistic:

$$X^2 = X^2_{(M)} + X^2_{(D)}$$

which approximates the χ^2 -distribution with $N_{(M)}s + N_{(D)}t$ degrees of freedom, under the null hypothesis that all of the multinomial probabilities used were correct. Therefore, information from a given derivative may easily be pooled by using the X^2 statistic, which may be calculated for all chemically realistic formulae that are consistent with the mass estimates of M and D , and which are consistent with the neutral loss. It is trivial to generalise the procedure to include an arbitrary number of derivatives.

The above theory has assumed that the multinomial probabilities reflect the isotopic abundance patterns, but in practice it is rarely possible to make use of the full set of isotopologues. This may be because of interference from coeluting compounds, or because the observed ion counts are too low. It is straightforward to exclude any subset of the isotopologues $M_0, M_1, \dots, M_s$ from the analysis, so long as two or more remain. Whichever isotopologues are excluded, the degree of freedom of the $x^2_{(M)}$ statistic will equal the total number of remaining isotopologues minus 1. The theoretical isotopic abundance patterns of putative formulae must be normalised when evaluating the $x^2_{(M)}$ statistic.

4.2.4 Robustness

A critical issue that arises when applying this procedure to groups of isotopologues stems from the requirement that the centroiding of the mass peaks must in principle be carried out over wide enough mass intervals that essentially all ions of each species are included. However, as will be demonstrated below, there is evidence to suggest that mass peaks have very heavy tails, so that a significant number of ions may be detected over mass ranges very distant from the peak apices, and even beyond 1 Da of the true mass. Consequently a mild mixture of adjacent isotopologues can arise when peak centroiding is applied, which has the effect of inducing an observed isotopic abundance pattern that, in general, differs from the theoretical one, somewhat beyond what may be attributed to the Poisson statistics. While this contamination appears to be very slight, and largely undetectable based on the $x^2_{(M)}$ statistics obtained from the individual scans, it will inevitably lead us to reject the true chemical formula more often than the chosen significance level would indicate. This is a trait that is highly

undesirable in a test of hypothesis, as it severely weakens the statistical argument on the basis of which a candidate formula is rejected as being “inconsistent with the observed data”. Moreover, the larger the sample size, the higher the probability will be of falsely rejecting the true chemical formula, so that the pooling of data becomes highly problematic.

It may therefore be advisable to employ a more robust version of the test of hypothesis described above, that is, a version which is not disproportionately affected by small departures from the model assumptions. This may relatively easily be accomplished by discarding, or “trimming”, a sufficiently high proportion of the largest $\chi^2_{(M)}$ statistics obtained from the individual scans, so that the nominal significance level is higher than the actual false positive rate. In other words, we ensure that we falsely reject the correct chemical formula *less often* than is specified by the chosen significance level. Therefore, the robust test produces p-values which, if they are very low, allow us to reject putative chemical formulae using the argument that:

“Assuming the proposed chemical formula is true, the probability of obtaining a deviation from its theoretical isotopic abundance pattern that is at least as extreme as the one observed, *is at most* p. The proposed formula is therefore not plausible.”

Thus, in rejecting a given chemical formula we have *at least* the degree of confidence that we would for a test whose nominal significance level is exactly equal to the false positive rate. The robust nature of this procedure comes at the cost of reduced statistical power – the test will be somewhat less effective at rejecting false candidate formulae. But as the failure to reject a false chemical formula is arguably a lesser concern than falsely rejecting the true chemical formula, such a trade-off will in most cases be warranted.

An issue that arises when applying the robust procedure regards the choice of the specific proportion of $\chi^2_{(M)}$ statistics that should be “trimmed”, T . Ideally, T , should be set as low as possible while ensuring that the false positive rate is consistently less than the chosen significance level. In practice it will be advisable to inspect the

distributions of the $x^2_{(M)}$ statistics, after trimming, for a series of known compounds, so as to ensure that their tails are consistently substantially lighter than the appropriate χ^2 -distribution. Clearly, this is not ideal, as it will not guarantee with absolute certainty that the test is conservative for the full dataset, although, qualitatively, it may be regarded as “very likely” that it is, assuming the sensitivity to these interference effects is reasonably uniform. The development of a test of hypothesis with a known null distribution would be highly desirable, but for want of a detailed mathematical model which can rigorously account for the mixture of isotopologues, the procedure outlined above may be close to the best that can currently be achieved.

Note also that it has so far been assumed that the isotopic abundance patterns of the elements included in the analysis do not significantly deviate from the standard values [80]. While the deviations are usually so slight that they will not be noticeable for the measurements made at individual scans, the greater statistical power obtained by pooling the data, could potentially make the test sensitive to them. However, any substantial deviations from the standard natural abundances would be detectable through the inspection of the x^2 statistics derived from known compounds, and the value of T might be increased accordingly. It has also been assumed that distinct isotopologues have the same underlying retention time profiles and that their ionisation propensities are identical, although in practice extremely small deviations are possible. Again, unless a very large data-set is used and T is very close to zero, this is not likely to confound the analysis.

4.3 Experimental details

The validity of the methods described may be evaluated by investigating the distribution of the $x^2_{(M)}$ statistics of known compounds for which the theoretical isotopic abundance patterns are known. If these $x^2_{(M)}$ statistics were to approximate the appropriate χ^2 -distribution, then the results relating to the construction of the simple multinomial confidence region follow immediately. However, owing to the distorting effects of the heavy tails of the mass peaks, this is not generally the case, and the distribution of the $x^2_{(M)}$ statistics has a somewhat heavier tail than the appropriate χ^2 -distribution. It therefore remains to determine whether the robust confidence region is sufficiently small to be useful in excluding candidate formulae.

The data used for the validation are extracted from the same data-set used in the previous chapter. Thus Section 3.4 can be consulted for the instrumental settings and the details of the sample preparation.

4.4 Analysis

4.4.1 Selection of test data-sets

The distribution of the $\chi^2_{(M)}$ statistics was examined for Hippurate, Nitrotyrosine and Chenodeoxycholic acid, as well as their respective derivatives (see Table 4.1). For Chenodeoxycholic acid and its dimer, the three lowest-mass isotopologues produced signals of sufficient strength for them to be included in the analysis, for the remaining compounds only the two lowest-mass isotopologues were included.

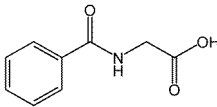
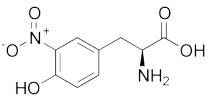
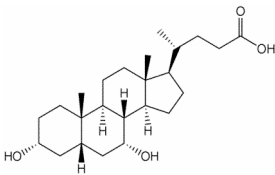
Compound	Chemical structure	Isotopologues	Isotopic abundance	Derivatives
Hippurate		$[M+H]^+$, $[M+1+H]^+$	90.63% 9.37%	Loss of Glycine
Nitrotyrosine		$[M-H]^-$, $[M+1-H]^-$	90.28% 9.72%	Dimer
Chenodeoxycholic acid		$[M-H]^-$, $[M+1-H]^-$, $[M+2-H]^-$	76.47% 20.31% 3.22%	Dimer

Table 4.1 – The three compounds used in the validation of the confidence regions.

Since the construction of the confidence regions require that the chromatographic peaks used be pure (or comprised only of isomers), continuum plots of all the peaks used were closely inspected. No evidence of contamination was found, and while this cannot guarantee that the peaks are pure, any interference from compounds that are

not isomers would tend to inflate the resulting x^2 statistics, which would lead us to trim a larger proportion of the x^2 statistics, thus reducing the statistical power of the test. As before, the validation procedure is therefore quite conservative.

In order to reduce the effects of detector saturation, Coates' dead time correction algorithm was applied to the continuum data. In addition, the scans for which the sum of the corrected ion counts were greater than 300 were removed. Scans for which the ion counts were too low, that is $n\rho_i < 5$ for some i , were pooled together before the x^2 statistics were calculated. In order to obtain a relatively unbiased sampling from the multinomials, all related isotopologues were centroided over an identical number of mass bins.

4.4.2 Validation

As for the coelution test, quantile-quantile plots may be used to evaluate the degree to which the x^2 statistics adhere to the appropriate χ^2 -distributions. The x^2 statistics derived from Hippurate, Nitrotyrosine and their derivatives should all adhere to the χ^2 -distribution with *one* degree of freedom, since they were derived from *two* isotopologues. Similarly the statistics derived from Chenodeoxycholic acid and its dimer should all adhere to the χ^2 -distribution with *two* degrees of freedom, since they were derived from *three* isotopologues.

The quantile-quantile plots of the x^2 statistics obtained for the three compounds are shown in Figure 4.1. The distribution of the statistics obtained from Chenodeoxycholic acid appears to be consistent with the χ^2_2 -distribution. The distribution of the statistics obtained from both Hippurate and Nitrotyrosine closely approximate the χ^2_1 -distribution over much of its central range, but have substantially heavier tails as evidenced by the most extreme x^2 statistics, which render the quantile-quantile plots slightly “flatter” than would be expected for χ^2_1 -distributed data.

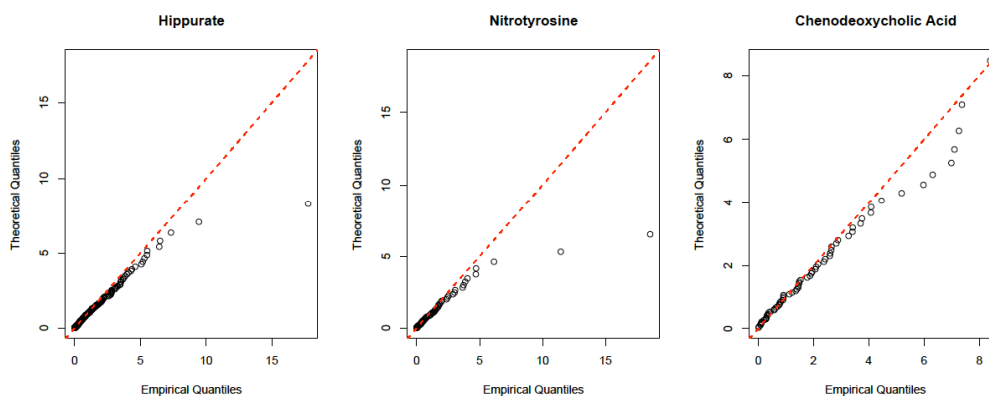


Figure 4.1 – Quantile-quantile plots of the χ^2 -statistics obtained from the three compounds, against the appropriate χ^2 -distributions. While the observed fit is very good for low quantiles, it is clear that the tails of the distributions obtained for Hippurate and Nitrotyrosine are too heavy to be consistent with the χ^2_1 -distribution.

It is possible that the deviations from the χ^2_1 -distributions could be explained by mild contaminations from unrelated compounds that were not visible on the continuum plots or by deviations from the standard natural isotopic abundances. However, a more likely explanation is that the tails of the mass peaks of adjacent isotopologues of the same molecular species are heavy enough to have been included in the centroiding thus distorting the isotopic ratios. Figure 4.2, shows a continuum plot of the two lowest mass isotopologues of Nitrotyrosine, where this phenomenon is clearly visible. It is not clear whether these heavy tails are due to high variability of the arrival times of the ions at the detector plate or whether they are artifacts induced by the digitisation electronics.

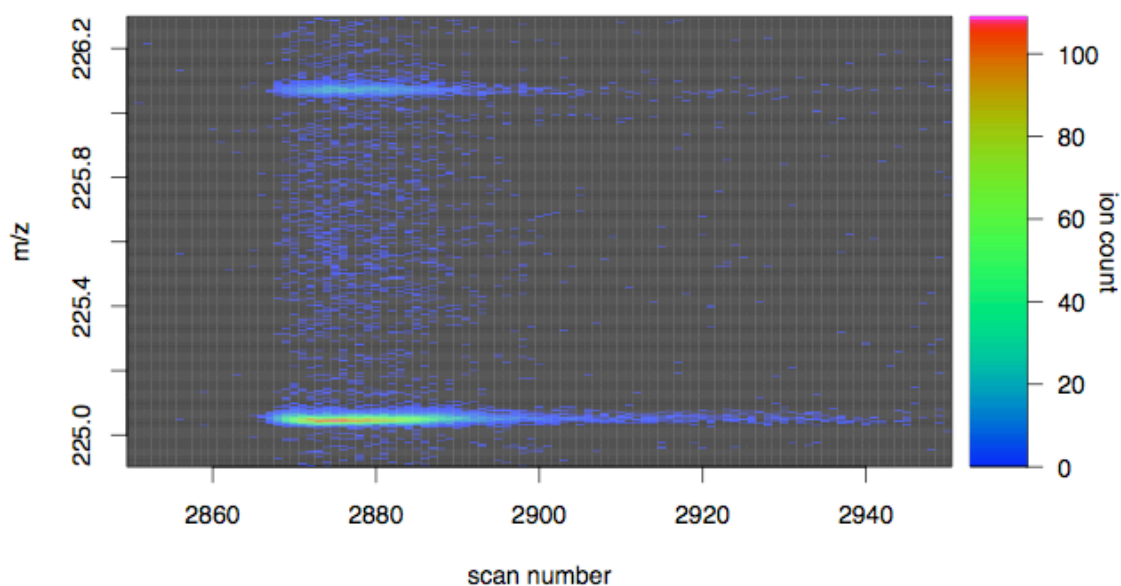


Figure 4.2 – Continuum plot of the two lowest-mass isotopologues of Nitrotyrosine. The tails of the mass peaks are heavy enough to reach the apices of the mass peaks of adjacent isotopologues, so that it is not possible to construct a centroid that is comprised of only one species of isotopologue. While the effect is less apparent for scans where the total ion count is lower, the mass-peaks at these scans will be all the more sensitive to any contamination.

In order to account for the effects of the heavy tails, the robust procedure described in the Section 4.2.4 was applied to the data. When $T = 0.05$ so that the largest 5% of the χ^2 statistics obtained from the individual scans were removed, the quantile-quantile plots of the resulting distributions displayed tails that were slightly lighter than the χ^2_1 -distribution. However, in order ensure that a cautious approach was taken, the value of $T = 0.10$ was used. The quantile-quantile plots of the resulting distributions are shown in Figure 4.3.

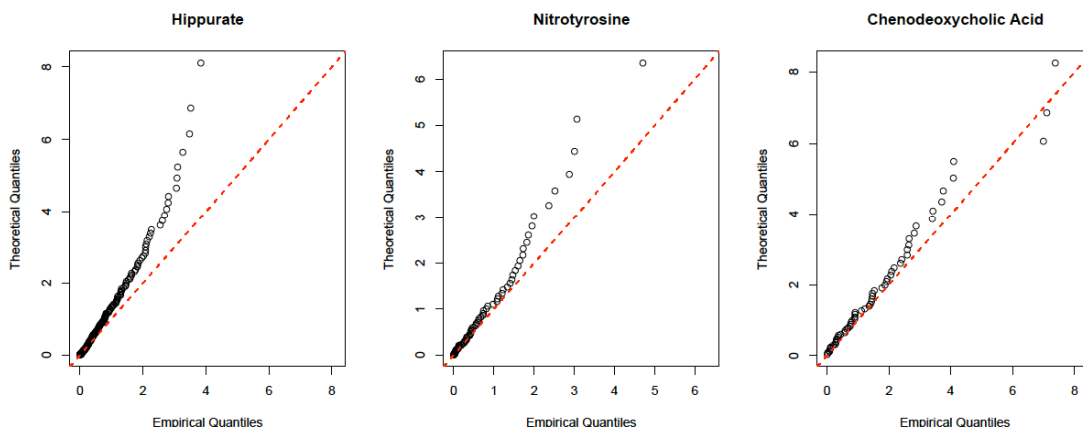


Figure 4.3 – Quantile-quantile plots of the x^2 statistics obtained from the three compounds, after the most extreme 10% have been trimmed. The quantiles obtained for Hippurate and Nitrotyrosine are now consistently smaller than those of the χ^2_1 -distribution, as required. The effects are more moderate for the x^2 statistics obtained from Chenodeoxycholic acid due to the smaller sample size.

As evidenced by the steep trends on their plots, the tails of the distributions of x^2 statistics obtained from Hippurate and Nitrotyrosine are now considerably lighter than that of the χ^2_1 -distribution. While the value of $T = 0.10$ is more than sufficient for all of the compounds that we have investigated, different mass spectrometers operating under different conditions and with different settings, might produce mass peaks with heavier tails than have been encountered here. Thus, any analyst employing the technique should apply it to known compounds to ensure that the chosen value of T makes the test sufficiently conservative.

Considering the many similarities between the test presented in this chapter and the coelution test, it is interesting that no trimming was required for the latter. However a key difference between the two methods is that the coelution test required the ρ_i in the expression for x^2 to be estimated from the acquired data, rather than calculated from a theoretical model. It therefore has a degree of flexibility that the current technique does not, and this is likely the reason why the latter shows greater sensitivity to the heavy tails of the mass peaks.

4.5 Results

As mentioned earlier, the practical procedure for formula elucidation, using the confidence regions described above, involves calculating the X^2 statistic for all chemically realistic formulae that are consistent with the mass error of the mass spectrometer. This was done for Hippurate, Nitrotyrosine and Chenodeoxycholic Acid. The robust procedure for which the 10% most extreme statistics were discarded was applied. The set of chemically realistic formulae was extracted from a list [40] compiled by the Fiehn group, which includes all formulae comprised of C, H, S, N, O, and P, which are consistent with the LEWIS rule.

It is difficult to determine the range of chemical formulae that are consistent with a mass estimate obtained through TOFMS since the uncertainty associated with such estimates is not very well understood. Modern TOF mass spectrometers are often said to have an accuracy of around 5 ppm, however, it appears that no serious attempt has been made at devising a method for constructing proper confidence intervals for them, although such a procedure would clearly be extremely valuable. While it is true that TOF mass spectrometers are capable of routinely producing mass estimates within 5 ppm of the theoretical mass, this is dependent on having carefully controlled operating conditions, which, in practice, cannot be ensured for all of the compounds encountered in high-throughput LC-MS experiments. Thus, mass errors substantially higher than 5 ppm are possible.

Therefore, in order to obtain a quite conservative list of candidate formulae, all chemically realistic compounds within 30 ppm of the theoretical masses of the compounds investigated were regarded as being consistent with the mass estimate of the mass spectrometer. In order to provide a broader illustration of the ability of the isotopic confidence regions to rule out putative formulae, a second list of all realistic chemical formulae within 0.1 Da of the theoretical masses was also compiled.

In an effort to assess the degree to which a standard scan provides information regarding the isotopic abundance pattern, the p-values associated with the median x^2 statistics, after trimming, of each of the candidate formulae was calculated. Similarly, the median X^2 statistics derived from the full chromatographic peaks of both the

parent compounds and their respective derivatives were calculated. The results are shown in Figure 4.4.

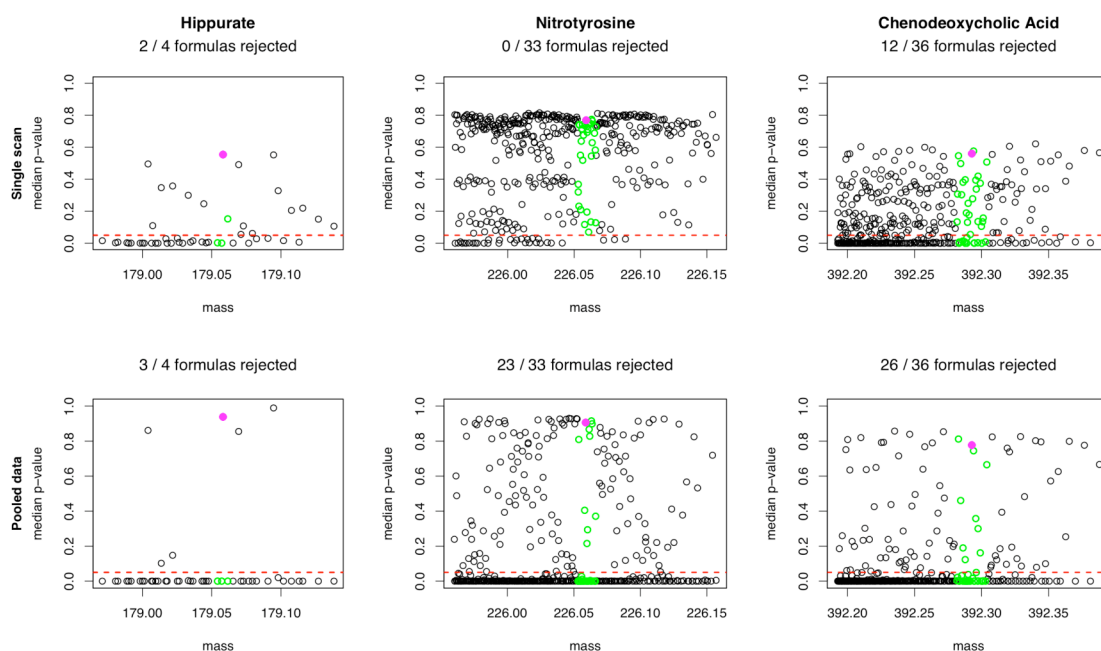


Figure 4.4 – Using the robust approach, the median x^2 and X^2 statistics were evaluated for the data obtained from Hippurate, Nitrotyrosine and Chenodeoxycholic Acid. The statistics were calculated for all formulae within 0.1 Da of the theoretical mass (black), for all formulae within 30 ppm of the theoretical mass (green) and for the true formula (magenta). Above each plot is listed the number of formulae that may be rejected at the 5% significance level (red line) out of the list of formulae within 30 ppm of the theoretical mass.

It is very clear that despite the conservative nature of the robust confidence region, it remains a powerful tool for excluding candidate formulae. While the confidence regions constructed from a single scan range from being incapable of rejecting a single formula, in the case of Nitrotyrosine, to being capable of rejecting 12, for Chenodeoxycholic Acid, the confidence regions constructed from the pooled data-sets all exclude a substantial number of formulae. Especially in the case of Nitrotyrosine, where the proportion of candidate formulae that can be rejected rises from zero to around two thirds, the benefit of pooling the data is impressive. For the wider mass window of ± 0.1 Da the percentage of false candidate formulae that are rejected for all three compounds is 26.79% for the single scan and 70.27% for the pooled data.

4.6 Future prospects

It may be worth investigating the upper limits of what might be achieved if instrumental developments allowed us to sample from undistorted multinomials corresponding to the isotopic abundance patterns. In this scenario we may pool the multinomial counts across the chromatographic peaks, as described in the Theory section, so that we can construct confidence regions, based on the outcome of a single multinomial with a very large number of trials. Chromatographic peaks for which detector saturation effects are relatively minor may easily be comprised of 10,000 ion counts, under standard operational settings. More intense peaks for which the high ion counts induce significant detector saturation may be comprised of over 100,000 ion counts.

401 compounds ranging in nominal mass from 100 to 500 and all spaced close to 1 Da apart were extracted from the list of chemically realistic compounds. For each of these, all compounds within 30ppm of the theoretical masses were considered to be consistent with the mass estimate. 10,000 multinomials corresponding to the isotopic abundance patterns of the selected compounds were simulated. Confidence regions were constructed for each of these simulations and the mean number of false candidate formulae within these regions was calculated, when a significance level of 0.05 was used.

The scenario in which a total of 10,000 counts were obtained was investigated when using either the two or the three lowest-mass isotopologues. A more idealised scenario in which a count of 100,000 was obtained was also investigated for the three lowest-mass isotopologues. In addition, the number of false negatives obtained when using only the mass estimate was calculated. The results, shown in Figure 4.5, demonstrate that, as anticipated, the strong statistical power achieved through the high ion counts, allows for a very substantial reduction in the number of false candidate formulae, when isotopic information is exploited. The statistical power achieved in the scenario in which 100,000 ions are counted is especially impressive, and it should be noted that at such high counts, it will usually be possible to use more than 3 isotopologues.

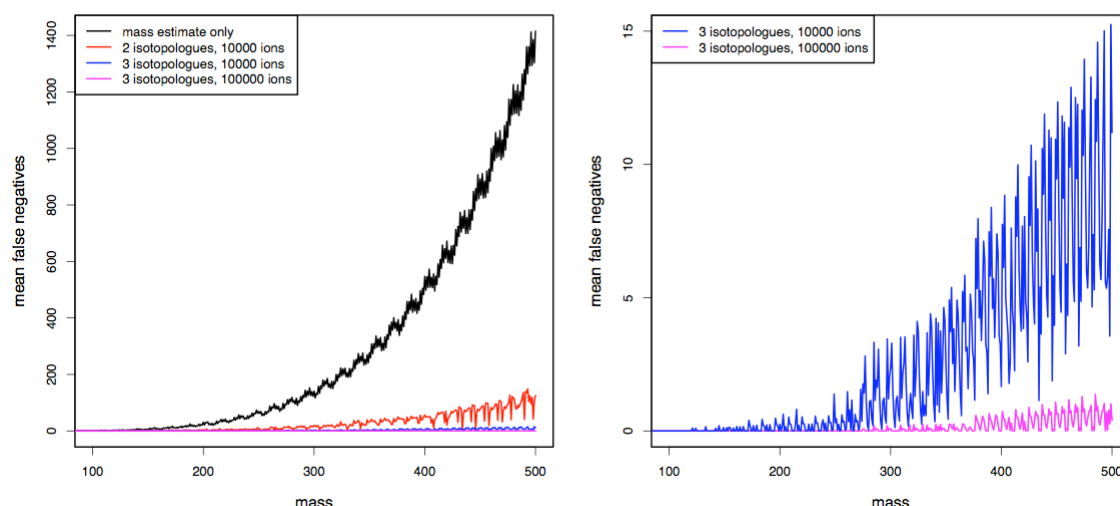


Figure 4.5 – The mean number of false candidate formulae within the confidence regions (false negatives) obtained from the simulated isotopic abundance patterns. The probability that a true candidate formula lies outside a given confidence region (a false positive) is given by the chosen significance level, which was set to 0.05 for these simulations.

Undoubtedly, the assumptions on which the simulations are based are currently highly idealised. However, they clearly suggest that the potential utility of isotopic abundance estimates could be very considerable. Moreover, even without further instrumental developments, it is entirely possible that careful modelling of the detailed characteristics of the mass peaks and of the detection system might allow us to better account for some of the phenomena that currently impede the analysis, and thereby obtain substantially improved estimates of the isotopic abundance patterns.

At the high counts used in the above simulations, it is quite possible that the deviations from the standard values of the natural isotopic abundances could confound the analysis. However, we may assume, for simplicity, that the standard abundances had been confirmed in advance, through separate measurements. This supposes a relatively uniform distribution of abundances across the entire sample, but if this assumption is false, the results might be even more interesting. Since different biological reactions can occur at different rates for different isotopologues [81] they tend to leave a weak isotopic signature on the compounds involved. It is conceivable

that potentially very interesting lines of research might be opened if isotopic abundance patterns could be estimated with sufficient accuracy to allow for the detection of these signatures for individual species of molecules.

4.7 Discussion

The above analysis suggests that Pearson's χ^2 test provides a reliable method for constructing conservative confidence regions for the isotopic abundance patterns observed in LC-TOFMS experiments. Thus, it is possible to determine, in a statistically rigorous manner, whether or not the theoretical isotopic abundance pattern of a given chemical formula is consistent with the observed data, and thereby reduce the number of candidate formulae for unknown compounds. This is a substantial improvement over alternative methods which attempt only to rank the fit of candidate formulae [76-78], or assume, rather imprecisely, that isotopic abundance estimates are accurate to within a few percent [40]. The method allows for information to be pooled from distinct scans and from distinct derivatives of the same underlying metabolite, which significantly increases the statistical power of the procedure.

The method is based on the assumption that the ion counts are Poisson distributed, and therefore does not apply to scans for which the ion counts are high enough to induce significant detector saturation. This constraint reduces the power of the test, but it does not affect its validity since even very large chromatographic peaks, which are severely saturated near their apices, will have low ion counts near their edges, to which the test can be applied. Moreover, the fact that information from distinct scans and distinct derivatives of the same underlying metabolite may be pooled has the effect of increasing the power of the test.

A more serious constraint stems from the fact that there appears to be a certain degree of mixture of the mass peaks of adjacent isotopologues. While the effect is often minor, it necessitates the use of robust methods, if a sound statistical argument is to be used in declaring candidate formulae to be inconsistent with the observed data. Again, the consequence is reduced statistical power, although, as was demonstrated, the test remains capable of excluding a substantial number of false candidate formulae.

A fundamental requirement of the test is that the detector used must employ a TDC. While it seems quite possible that confidence regions may also be constructed for mass spectrometers employing ADCs, the procedure may not prove to be as straightforward as for TDCs as it is the ability of the latter to block out electronic noise and preserve the Poisson distribution of incoming ions that makes the procedure particularly simple. Thus, while TDCs are criticised for their relatively limited dynamic range, their ability to produce data that approximate a simple and well-understood distribution constitutes an important advantage.

The application of the test to the three compounds investigated suggests that the information contained in the observed isotopic abundance patterns may be extremely valuable in identifying unknown metabolites, even when these do not contain bromine or chlorine. While we have outlined methods for reducing the size of the confidence regions, it is likely that these might be reduced much further if the information from the scans with high ion counts could be included in the analysis, or if the mixture of the mass peaks of adjacent isotopologues did not arise. Thus it is clear that there is scope for improvements in the accuracy with which isotopic abundance patterns can be estimated, and such improvements may be as just as important as improvements in mass accuracy. Considering the very high cost of mass spectrometers capable of high mass accuracy, it is arguable that this line of research is somewhat neglected.

5 Prospects for a statistical theory of LC-TOFMS data

The model used in the previous two chapters was based on the assumption that the ion counts were Poisson distributed and that the data were centroided. In this chapter these constraints will be lifted – we will attempt to define a probability distribution that describes continuum data, including the saturation effects due to the dead time and the finite time resolution of the TDC. The model that will be derived in the following approximates the true probability distribution (which is extremely complex) by means of a series of Binomial distributions. In doing so, it builds on the work of P. B. Coates who developed somewhat similar methods for the correction of detector saturation, first in the context of radiative lifetime measurements [51] but later also applied to TOFMS data [52].

While the model used by Coates is reminiscent of a Binomial distribution it was never explicitly defined as such, and the assumptions required for its validity were not defined. Moreover it was used strictly to enhance dynamic range – no attempt was made at relating the model to the mass and chromatographic peak shapes, or to use it to draw broader inferences about the data. While numerous other models of mass spectrometry data have been proposed [13, 25, 82, 83] none have been found that are developed from the fundamental characteristics of the instrumentation employed. Moreover, the author is not aware of any other models whose predictions have been validated in a statistically rigorous manner - model validity is typically argued by means of simple qualitative comparisons to real data.

5.1 Coates' correction

Although the assumption is not explicitly formulated, Coates assumes that chromatographic peaks are approximately constant in the period over which they are histogrammed. Thus, all of the pulses constituting the histogram are considered to be independent and identically distributed (i.i.d.). In addition, each of the ions hitting the detector is assumed to arrive in the centre of a tick of the TDC clock and the dead time is assumed to last for an integer number of ticks. This introduces an error in the correction method, which, according to Coates, is negligible if the ticks are sufficiently narrow that the Poisson rate function varies linearly across each of them.

As discussed in Section 2.5, owing to the finite time resolution of the TDC, multiple ions that hit the detector within the same tick of the clock cannot be distinguished. Thus for a given pulse, the correction algorithm is based on the probabilities of obtaining either “at least one ion count” (p) or “zero ion counts” ($1 - p$), and these must be estimated at every tick. However as discussed below, it is straightforward to relate these probabilities to the Poisson rate function.

Suppose that the total number of pulses that are histogrammed over is N_p and that the dead time lasts for exactly D ticks of the clock. Then an ion arrival in the i th tick will only be recorded if it has not been preceded by an ion arrival in the previous D ticks, that is, with probability $\prod_{j=i-D+1}^{j=i-1} (1 - p_j)$. We may therefore expect that the number of pulses that are “valid” at the i th tick, that is, that are capable of recording any new ion arrivals is given by

$$V_i = N_p \prod_{j=i-D+1}^{j=i-1} (1 - p_j).$$

Consequently, if the observed ion count at the i th tick is k_i , we may estimate the probability of detecting at least one ion in the i th tick of a given pulse as²

$$\hat{p}_i = \frac{k_i}{V_i} = \frac{k_i}{N_p \prod_{j=i-D+1}^{j=i-1} (1 - \hat{p}_j)}.$$

The product is only taken over positive j and if the correction is applied chronologically it will only involve estimates of p_j that have already been calculated.

Suppose the expected number of ion arrivals in the i th tick of a given pulse is h_i . Then, based on Poisson statistics, the probability of recording at least one ion arrival is

² Coates in fact lets the product run from $j = i - D - 1$. This is altered in this discussion because a deadtime of length D is interpreted to mean that a pulse blocked at the first tick will stay blocked throughout the D th tick and open at the $D + 1$ th tick assuming it is not hit by further ions.

$$p_i = P(k_i > 0) = 1 - P(k_i = 0) = 1 - \exp(-h_i)$$

so that

$$h_i = -\log(1 - p_i)$$

and since all N_p pulses are assumed to be i.i.d. we may then use

$$\text{estimated total rate of ion arrivals} = -N_p \log(1 - \hat{p}_i) = -N_p \log\left(1 - \frac{k_{i,j}}{V_{i,j}}\right)$$

We will henceforth refer to this algorithm as “Coates’ correction”. A somewhat different formalism to that used by Coates has been used here in order to make the equations more easily comparable to those that will be presented later in this chapter.

The reasoning behind Coates’ correction seems very sensible, but the algorithm is derived through what amounts to a plausibility argument and without further theoretical backing there is no reason to suppose that the correction is in any way “optimal”. It is in fact not difficult to see that Coates’ correction does not exploit all of the information available in the system. Note that in estimating p_i , it only makes use of the current and previous ion counts. But if the ion count at the i th tick were higher than the estimated number of valid pulses, we would obtain an estimate of p_i larger than one, which should serve as an indication that earlier probability estimates were too high. Moreover, a method for dead time correction is not inherently useful for drawing inferences regarding the data, except at a very qualitative level.

However, statistical theory can provide powerful and very general methods for constructing statistical tests for many of the questions that naturally arise in the analysis of LC-TOFMS data *if the probability distribution of the acquired data is known*. Additionally, it provides methods for parameter estimation, which in a certain sense can be considered “optimal”. Thus, the first step towards developing a comprehensive framework for the statistical analysis of LC-TOFMS data lies in formulating a probability distribution that approximates that of the acquired data, based on our understanding of the detailed workings of the instrument.

5.2 Theory

We consider a molecular species, S , which passes through the chromatographic column, is ionised and accelerated onto a detector plate. Our aim is to express the probability of recording a given set of ion counts over the specified mass and retention time ranges.

5.2.1 The chromatographic dimension

Owing to the chromatographic separation, the concentration of S at the end of the chromatographic column will vary as a function of the retention time, τ , corresponding to the familiar chromatographic peak. It is useful to describe this varying concentration by means of a normalised function, $\Gamma(\tau)$, that integrates to 1. Thus, if $n^{(S)}$ is the total number of molecules of S in the mixture, the number of molecules of S that elute between the retention times τ_a and τ_b can be expressed as

$$n^{(S)} \int_{\tau_a}^{\tau_b} \Gamma(\tau) d\tau$$

If p^{ionise} is the probability that a given molecule of S is ionised, then the probability that a total of k such ions are produced between τ_a and τ_b , can be written:

$$P(k) = \frac{\left(p^{ionise} n^{(S)} \int_{\tau_a}^{\tau_b} \Gamma(\tau) d\tau \right)^k e^{-p^{ionise} n^{(S)} \int_{\tau_a}^{\tau_b} \Gamma(\tau) d\tau}}{k!}$$

where we make use of the standard result that a Binomial distribution with a very large sample size and a very low probability approximates a Poisson distribution. Since Γ varies over time, the technical name for the above distribution is “a non-homogeneous Poisson process”. It is noted that although p^{ionise} can typically be assumed to be independent of Γ , this will not be the case for very high concentrations of S , or if a coeluting molecular species causes ionisation suppression.

A large number of the ions generated are not detected as they are lost on their way to the orthogonal accelerator (oa), lost following the orthogonal acceleration, or fail to get registered by the detector. However if the process by which the ions are lost can be regarded as the outcome of a Binomial distribution, then the distribution of the remaining ion count remains Poissonian [84].

If it is assumed that the ion optics do not significantly distort the distribution of ions, so that $\Gamma(\tau)$ may be used to describe the “concentration” of the ions of S in the oa, a short time after their formation, then we can express the number of ions in the oa as:

$$P(k) = \frac{\left(p^{oa} p^{ionise} n^{(S)} \int_{\tau_a}^{\tau_b} \Gamma(\tau) d\tau \right)^k e^{-p^{oa} p^{ionise} n^{(S)} \int_{\tau_a}^{\tau_b} \Gamma(\tau) d\tau}}{k!} \quad (5.1)$$

where p^{oa} is the (Binomial) probability that a given ion of S eventually enters the oa, and τ_a now denotes the time at which the ion beam first enters the oa and τ_b is the time at which the electric field is applied. It is reasonable to suppose that p^{oa} will be independent of Γ unless the ion count is so high that the ions interact significantly with each other.

5.2.2 The time-of-flight dimension

While we continue to describe the distribution of the ions in time, it is useful to regard the time-of-flight as a separate dimension to the retention time, as it relates to the mass of S rather than its polarity. Thus, while Γ describes the relative concentration of S as a function of retention time, we now require a function, $\Omega(t)$ to describe the variation in the relative “concentration” of ions at the detector plate as a function of time-of-flight, t . In addition to the mass of S , the shape of Ω reflects factors such as the initial velocity and spatial distributions of the ions at the time the electric field is applied, as well as the strength of the applied field and the length of the flight path. However, for the sake of conciseness, it is written only as a function of the time-of-flight.

As discussed in 2.5, the clock that measures the time-of-flight has limited time resolution and measures finite time increments of 10s to 100s of picoseconds for modern TOFMS systems. If a given such interval runs from t_a to t_b , then the number of ions that arrive at the detector plate over this period remains Poissonian, as described below. The absolute number of ions in the oa at $t = 0$, when the electric field is first applied is given by equation 5.1. If p^{tof} is the Binomial probability that a given one of these ions strikes an active area of the detector plate, then the probability of obtaining k ion arrivals in the interval $[t_a, t_b]$ can be expressed as

$$P(k) = \frac{\left(p^{tof} p^{oa} p^{ionise} n^{(s)} \int_{\tau_a}^{\tau_b} \Gamma(\tau) d\tau \int_{t_a}^{t_b} \Omega(t) dt \right)^k e^{-p^{tof} p^{oa} p^{ionise} n^{(s)} \int_{\tau_a}^{\tau_b} \Gamma(\tau) d\tau \int_{t_a}^{t_b} \Omega(t) dt}}{k!}$$

where we disregard detector saturation for now. We have again assumed that p^{tof} is independent of Γ and also of Ω although this requirement would break down at very high ion counts owing to space charge effects [85].

5.2.3 Histogramming of binary TDC data

As explained earlier, once a pulse has been digitised by a TDC it is represented as a binary sequence, indicating only whether zero, or one or more ions were detected in each of the ticks of the TDC clock. Consequently, these data are not Poissonian, but may be regarded as the outcome of Bernoulli trials where the probability of success is the probability that one or more ions are detected in the tick. Thus we must make use of the mapping that was also involved in Coates' correction:

$$P(k \geq 1) = 1 - P(k = 0) = 1 - e^{-p^{tof} p^{oa} p^{ionise} n^{(s)} \int_{\tau_a}^{\tau_b} \Gamma(\tau) d\tau \int_{t_a}^{t_b} \Omega(t) dt}$$

where t_a and t_b are chosen such that they define a time-of-flight interval corresponding to a tick of the TDC clock.

If Γ is approximately constant across the pulses that are histogrammed, and if the length of the flight path and power supply output are sufficiently stable over the corresponding period of time that Ω remains approximately constant over matching

ticks in distinct pulses, then the Bernoulli trials can be considered to be i.i.d. Consequently, the counts obtained by histogramming the pulses may be regarded as the outcome of a Binomial distribution. In view of the short period of time involved and the comparatively modest slope of Γ , the assumptions required are not unreasonable. These assumptions are also required for Coates' correction.

The relationship between the rate function and the Binomial probability is not entirely trivial. The highest duty cycle is achieved by “over-pulsing” the ion beam in the oa, that is, by applying the accelerating electric field before the fastest ion has traversed the oa [86]. However, this is generally not an attractive option, as a substantial delay is required following the application of each pulse in order to ensure that the heaviest (and therefore slowest) of the ions has reached the detector plate prior to the application of the next pulse. Consequently, Γ is not sampled over contiguous time intervals over the course of a scan. However, given the high sampling frequency, the ions lost due to the pulse delay may be regarded as a fixed fraction of Γ , which is independent of its precise shape. If N_p is the total number of pulses histogrammed, ε is the period between the application of consecutive pulses, and p^{scan} is the mean proportion of ions of S that are accelerated rather than pass through the oa, then this assumption can be expressed as

$$\sum_{i=1}^{N_p} \left(\int_{\tau_A + (i-1)\varepsilon}^{\tau_A + (i-1 + p^{scan})\varepsilon} \Gamma(\tau) d\tau \right) \approx p^{scan} \int_{\tau_A}^{\tau_A + N_p \varepsilon} \Gamma(\tau) d\tau$$

where the summation of the integrals is chosen so that p^{scan} corresponds to the proportion of the domain of Γ that we integrate over. In other words, rather than summing the integrals of all the segments of Γ that are accelerated, we integrate Γ over the entire scan and reduce the result by a factor of p^{scan} .

We can now write the total expected ion count, I , over the entire peak in both dimensions as

$$I = n^{(S)} p^{ionise} p^{tof} p^{oa} p^{scan}$$

and we can write the Poisson rate function more concisely as

$$\lambda(\tau, t) = I\Gamma(\tau)\Omega(t)$$

though it is stressed that regarding I as independent from Γ and Ω restricts the range of ion counts over which the model can be applied somewhat.

It is furthermore useful to define a *discretised rate function*, $\Lambda_{i,j}$, that specifies the mean ion count over all the pulses associated with the i th scan at the j th tick:

$$\Lambda_{i,j} = \int_{\tau_a}^{\tau_b} \int_{t_a}^{t_b} \lambda(\tau, t) dt d\tau = I \int_{\tau_i}^{\tau_{i+1}} \Gamma(\tau) d\tau \int_{t_j}^{t_{j+1}} \Omega(t) dt$$

We can then express the Binomial probability, ρ , of obtaining one or more ion counts in a given pulse of the i th scan, at the j th time-of-flight tick. Since there are N_p such pulses and Γ has been assumed to be approximately constant over them, this is written:

$$\rho_{i,j} = P(k \geq 1) = 1 - P(k = 0) = 1 - e^{-\Lambda_{i,j}/N_p} \quad (5.2)$$

So that the probability of obtaining a count of k , at the i th scan and j th time-of-flight tick, is given by

$$P(k) = \binom{N_p}{k} \rho_{i,j}^k (1 - \rho_{i,j})^{N_p - k}$$

5.2.4 Dead time

As explained in Section 2.6 most pulses will only register the very first ion to reach the detector at a given mass peak due to the ensuing dead time. This greatly facilitates the modelling problem as it makes it easier for us to account for how many of the N_p pulses are valid and capable of registering further ion arrivals in each of the time-of-flight ticks. If $k_{i,j}$ denotes the ion count observed in the i th scan and at the j th tick of the TDC, lasting from t_j to t_{j+1} , then assuming all N_p pulses are valid at time-of-flight

t_a where the mass peak “starts”, we can write the number of valid pulses at the j th tick, $V_{i,j}$, as

$$V_{i,j} = N_p - \sum_{x=a}^{x=j-1} k_{i,x}$$

Note that this expression is different to the one used by Coates, as his expression attempts to account for the possibility that a pulse that is closed may become valid again, so that it can register multiple ions over a single mass peak. A more rigorous treatment of this scenario will be given in Chapter 6.

If $V_{i,j}$ is defined as above, then the recorded ion count at the i th interval adheres to the following Binomial distribution

$$P(k_{i,j}) = \binom{V_{i,j}}{k_{i,j}} \rho_{i,j}^{k_{i,j}} (1 - \rho_{i,j})^{V_{i,j} - k_{i,j}}$$

In practice, when applying the above model to real data it is best to do so over a short time-of-flight period of similar length to the dead time period, to help ensure that pulses that are rendered “invalid” by ion arrivals remain so. Owing to the rather heavy tails of mass peaks (in other words, of Ω), it is not always possible to ensure that the value of V_a is exactly equal to N_p . However, as will be shown, the approximation works well for mass peaks of moderate ion counts.

It is straightforward to extend the model to describe multiple scans and time-of-flight ticks. Suppose that the ions of S are obtained over N scans and M time-of-flight ticks. If $\mathbf{k}$ is a matrix³ such that $k_{i,j}$ is the observed ion count in the i th scan and the j th tick in which S is observed, that is

³ Strictly speaking this does not have to be a matrix as we might use different numbers of ticks at different scans.

$$\mathbf{k} = \begin{pmatrix} k_{1,1} & \cdots & k_{1,M} \\ \vdots & \ddots & \vdots \\ k_{N,1} & \cdots & k_{N,M} \end{pmatrix}$$

then taking these counts to be independent, the probability distribution for the full set of ion counts can be written:

$$P(\mathbf{k}) = \prod_{i=1}^N \prod_{j=1}^M \left\{ \begin{pmatrix} V_{i,j} \\ k_{i,j} \end{pmatrix} \rho_{i,j}^{k_{i,j}} (1 - \rho_{i,j})^{V_{i,j} - k_{i,j}} \right\} \quad (5.3)$$

with $V_{i,j}$ and $\rho_{i,j}$ defined as before. This probability distribution will in the following be referred to as the “basic model”.

5.3 Model limitations

The most demanding assumptions of the basic model are the requirements that the number of valid pulses is equal to N_p at the start of the time-of-flight range and that the length of the dead time invariably exceeds the remaining time-of-flight range. Small deviations from these assumptions do not render the model inapplicable, but it is nevertheless an important respect in which it is incomplete and a key reason why it breaks down at high ion counts. The construction of a more comprehensive model for LC-TOFMS data would require detailed knowledge of the workings of the detector system and in particular of the statistical distribution of the dead time. It is also highly likely that other components of the detector system would require more attentive modelling at extreme ion counts.

In addition, the basic model is incomplete in the sense that the functional forms of Γ and Ω have not been specified and neither has the nature of their dependence on I . Several papers and patent applications have modelled mass peaks based on a Gaussian shape [49, 83, 87-89], however, significant deviations from this functional form have been noted at the tails of mass peaks [47, 89]. A number of models have been developed for chromatographic peaks [90], but again, no single model has been found to be satisfactory under all circumstances [91]. Modeling Γ might be especially

difficult since any distortions resulting from the ion optics or the electrospray would have to be accounted for. Until these fundamental questions in the theory of chromatography and the theory of time-of-flight mass spectrometry have been more comprehensively answered, important constraints on the statistical modelling of LC-TOFMS data are inevitable.

Nonetheless, knowledge of Γ and Ω is not required for the basic model to be of use. By working with the discretised rate function, $\Lambda_{i,j}$, it is possible to obtain estimates of the true rate of ion arrivals, irrespective of the functional forms of Γ and Ω . However, this requires a total of NM estimates to fully characterise the rate function. For the mass spectrometer used in this study, M is typically between ten and twenty, while N can range from around ten to several hundred for chromatographic peaks with heavy tailing. Whatever the true functional forms of Γ and Ω are, they will undoubtedly require far fewer parameters, so that there are effectively more data available for each parameter that must be estimated to fully describe the rate function. Note that if either Γ or Ω are known, but not both, two further models could be envisaged, which estimate the parameters of the known functional form as well as the overall ion count at either every scan or at every time-of-flight tick.

5.4 Applications of the basic model

Although Γ and Ω are described as being functions of the time-of-flight and the retention time, they depend on a larger number of parameters. Some of these are likely to be nuisance parameters that provide little information on S , but clearly μ_Ω - the location parameter of Ω - is of great interest as it relates to the mass of S , and we may for the moment assume, rather naively, that it does so through the standard time-of-flight equation

$$m/z = \frac{2eU}{d^2} \mu_\Omega^2$$

where U is the voltage, e is the elementary charge, and d the length of the flight path. Other parameters must include, at a minimum, a scale parameter for Ω , σ_Ω , and location, scale and skewness parameters for Γ - μ_Γ , σ_Γ and γ_Γ respectively. These may

be affected by the intensity of the peaks, but this dependence may be close to negligible for moderate ion counts.

When the probability distribution of the acquired data is known, the problem of parameter estimation may be addressed by means of the widely used *method of maximum likelihood*. Note that the left-hand side of equation 5.3 should strictly speaking be written $P(\mathbf{k} \mid \mu_\Omega, \sigma_\Omega, \mu_\Gamma, \sigma_\Gamma, \gamma_\Gamma, I, N_p)$ in order to make explicit the dependence of the probability distribution on the full set of parameters of the system. However, we can reinterpret this probability distribution as the *likelihood function* $L(\mu_\Omega, \sigma_\Omega, \mu_\Gamma, \sigma_\Gamma, \gamma_\Gamma, I \mid \mathbf{k}, N_p)$, for which we allow the parameters to vary but consider $\mathbf{k}$ (and N_p which is always known) to be fixed. We then find the parameters that maximise the likelihood function and use these as our estimates as they are the ones that would give rise to the observed data with the highest possible probability. Clearly, this approach has a stronger theoretical appeal than a simple centroid, or even finding the parameters through a least-squares approach [87]. However, the basic model may be used to draw a much broader range of inferences, by means of tests of hypotheses, constructed using the *likelihood ratio test*.

As will be discussed below, the basic model can be used to describe a wide range of features of the data acquired in LC-TOFMS experiments, by expanding the likelihood function accordingly. Certain hypotheses that the analyst may have regarding the acquired data can be expressed very naturally by placing specific constraints on the likelihood function, and thereby reducing the total number of parameters of the model. The likelihood ratio test can be used to assess whether such hypotheses are plausible, by determining whether or not the unconstrained model is significantly better at describing the acquired data than the constrained one is.

More specifically, suppose L_0 is the supremum of the likelihood function for the constrained model, and L_A is the supremum of the unconstrained one, and let d be the difference in dimensionality of the two models. If the hypothesis is true, and the constraint is appropriate, then under certain regularity conditions for the likelihood functions, it can be shown [92] that for large sample sizes:

$$X^2 = -2 \log \left(\frac{L_0}{L_A} \right) \sim \chi_d^2$$

That is, the X^2 statistic adheres to a χ^2 -distribution with d degrees of freedom. Thus, by comparing X^2 to the cumulative distribution function of the appropriate χ^2 -distribution, we can determine whether the data are consistent with the hypothesis associated with the constrained model, at a given significance level. Since the above result is asymptotic, care must be taken to ensure that it applies in practice.

An important practical difficulty in applying the likelihood ratio test lies in finding L_0 and L_A in the first place. Since the likelihood functions encountered in this study are quite complex, analytical solutions are not generally available and consequently numerical methods must be employed. Aside from the inevitable computational demands this entails, caution must be exercised to ensure that the errors associated with the final approximations are very small relative to values typical of the χ_d^2 distribution.

5.5 Validation

The basic model may be validated by determining whether its predictions are borne out in real LC-TOFMS data. We will assume that Ω is Gaussian, so that

$$\int_{t_j}^{t_{j+1}} \Omega(t) dt = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{t_i}^{t_{i+1}} e^{-\frac{(t-\mu_\Omega)^2}{2\sigma_\Omega^2}} dt$$

but leave Γ unspecified, so that

$$I \int_{\tau_i}^{\tau_{i+1}} \Gamma(\tau) d\tau = I_i$$

where the “intensity factors”, I_i , must be estimated independently for each scan. For the purpose of validating the model we will also fit μ_Ω and σ_Ω independently at each scan. This is primarily to simplify the maximisation of the likelihood function, which in turn must be done independently for each scan. However, it also accounts for potentially confounding effects, for example that the values of μ_Ω and σ_Ω might

fluctuate depending on the intensity of the mass peak (so that Ω and I are not independent) or that μ_Ω might drift over time due to temperature fluctuations. The discretised rate function of S is then written

$$\Lambda_{i,j}^{(S)} = I_i^{(S)} \frac{1}{\sqrt{2\pi\sigma_{\Omega,i}^2}} \int_{t_i}^{t_{i+1}} e^{-\frac{(t-\mu_{\Omega,i})^2}{2\sigma_{\Omega,i}^2}} dt$$

In fitting the likelihood function to a given chromatographic peak, two parameters are required for the Gaussian mass peak and one for the intensity factors. Consequently, fitting the model to the ion counts obtained over N scans requires $3N$ parameters.

As a key test of model validity, we consider again the phenomenon of fragmentation, where some molecules of S (the parent molecule) break up into a smaller molecular species, R (the fragment) and recall that such peak pairs can be identified by determining whether they exhibit exact coelution or partial coelution. If the two peaks exhibit exact coelution then $I^{(S)} = I^{(R)}$ while no such relationship would be expected under partial coelution.

In the framework of the likelihood ratio test, the unconstrained model corresponds to partial coelution, so that the basic model with the discretised rate function given above, must be fitted to S and R independently. Therefore $3 + 3 = 6$ parameters are required for a single scan and $6N$ for the full data-set. For exact coelution, the basic model must be fitted to S and R with the constraint that the chromatographic peak shapes are identical. Consequently, the ratio of their intensity factors can be written

$$\frac{I_i^{(S)}}{I_i^{(R)}} = \frac{I^{(S)} \int_{\tau_i}^{\tau_{i+1}} \Gamma(\tau) d\tau}{I^{(R)} \int_{\tau_i}^{\tau_{i+1}} \Gamma(\tau) d\tau} = \frac{I^{(S)}}{I^{(R)}} = b$$

which is constant across all scans if, as has been argued in the derivation of the basic model, the intensity of the mass peaks is largely independent of the shape of Γ . Therefore, the basic model must be fitted to S and R simultaneously, using the constraint that $I_i^{(S)} = bI_i^{(R)}$ for each scan. Consequently, we require effectively only 5

parameters to fit the likelihood function to a single scan, and $5N + 1$ for the full data-set, the extra parameter being b . Provided N is sufficiently large, a satisfactory estimate of b can be obtained by taking the median of the ratios provided by the estimates of the intensity factors obtained from the unconstrained model.

For a given mass peak, the difference in the number of parameters for the constrained and the unconstrained models is effectively 1, and for N scans the difference is $N - 1$. Thus *if* the basic model provides a good approximation to the true distribution of the empirical data then applying the likelihood ratio test to known parent-fragment pairs, should give rise to X^2 statistics that are distributed according to the χ^2_I -distribution for individual scans and the χ^2_{N-I} -distribution for the full data-set. By comparing the empirical values of the X^2 statistics to the cumulative distribution function of the appropriate χ^2 -distribution, a p-value can be obtained, indicating whether or not the data are consistent with this null hypothesis.

In addition to testing the validity of the basic model it is worth investigating whether the quite considerable level of detail that it includes is even necessary. Therefore the same likelihood ratio test was also constructed for the model which assumes the ion counts to be purely Poissonian. As in previous chapters, the two tests were applied to the data over varying ion count ranges constituting the tertiles of the full ion count range.

The data used for the validation were extracted from the same data-set used in the previous chapters (see Section 3.4). The compounds investigated were Nitrotyrosine, Glutaric acid and Hippurate. The fragments of Nitrotyrosine and Glutaric acid correspond to a loss of CO_2 and the fragment of Hippurate corresponds to a loss of Glycine.

The likelihood functions were maximised by means of a Newton-type algorithm [93] implemented by Charles J. Geyer [94] in the R statistical programming language [95]. This method requires knowledge of the likelihood function, its gradient (the vector of first-order partial derivatives) and its Hessian (the matrix of second-order partial derivatives). The gradient of the likelihood function was calculated analytically and

was in turn used to obtain numerical estimates of the Hessian. After convergence, several of the likelihood functions were visually inspected in all dimensions near the maximum likelihood estimates in order to help ensure that a maximum had indeed been attained. The results for the pure Poisson model are shown on Figure 5.1 below.

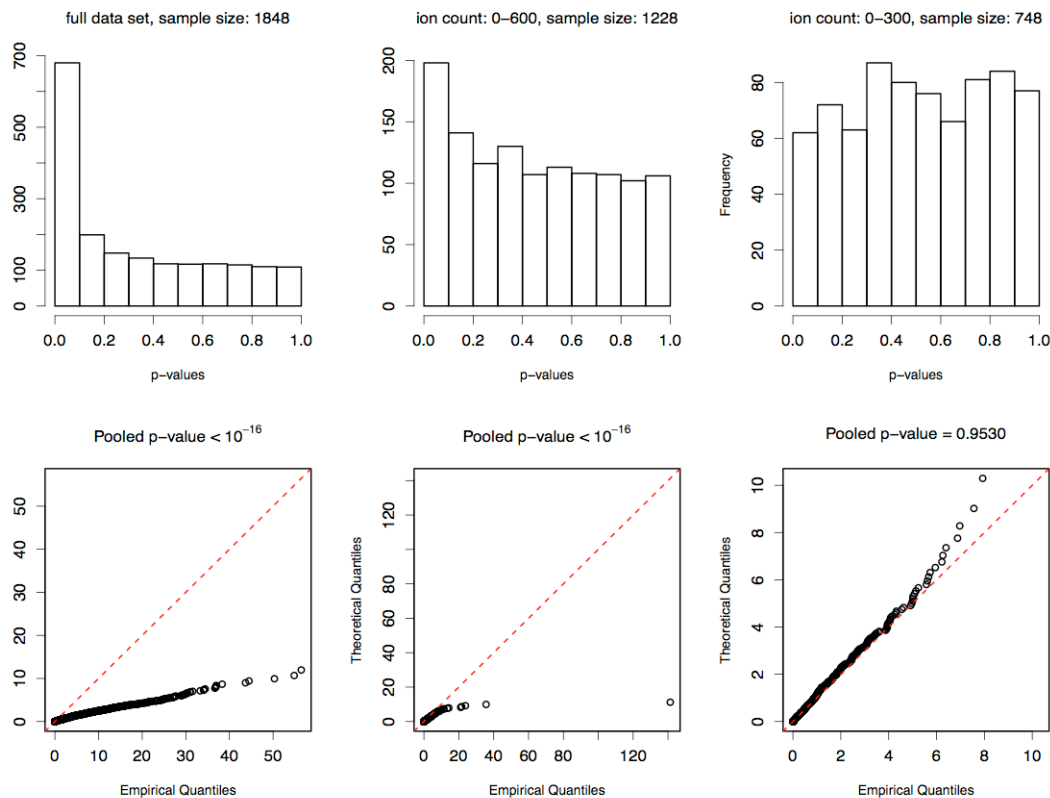


Figure 5.1 - Histograms of the p-values associated with the X^2 statistics obtained from the individuals scans of the three datasets by using the pure Poisson model to construct the likelihood ratio test (top) and quantile-quantile plots of the X^2 statistics themselves as compared with the theoretical χ^2_1 -distribution (bottom). The p-values obtained for the full data-sets are listed above the quantile-quantile plots. Only for low ion counts do the statistics conform reasonably well to the χ^2_1 -distribution predicted by the likelihood ratio test.

The results are reminiscent of those obtained for the coelution test in Chapter 3. At low ion counts where detector saturation is minimal, the pure Poisson model does a reasonable job of explaining the observed data and consequently the X^2 statistics conform fairly closely to the distribution predicted by the likelihood ratio test. But at

higher counts the saturation effects become more substantial and significant deviations from the predicted distribution are evident.

The results for the basic model are shown on Figure 5.2 below. It is clear that the fit is significantly better than that of the pure Poisson model for these data as the X^2 statistics are consistent with the predicted distributions over the two lower tertiles of the ion count range. Thus we have strong evidence that the basic model approximates the probability distribution of the acquired data and that the mathematical modelling of the saturation effects is not superfluous. In addition we have a direct demonstration that the likelihood ratio test can be used in practice to make inferences about the sample being analysed. In view of the quite considerable detail with which the basic model has been formulated, and the very specific predictions made, these results are very encouraging.

It is noted however that the likelihood ratio test for exact coelution based on the basic model does not appear to be valid over an ion count range that is significantly greater than that obtained with the coelution test from Chapter 3 – so long as Coates' correction is applied as part of the latter test. Therefore, owing to its substantial computational demands, the main value of this likelihood ratio test is its demonstration of the validity of the basic model rather than its use as a routine analytical tool.

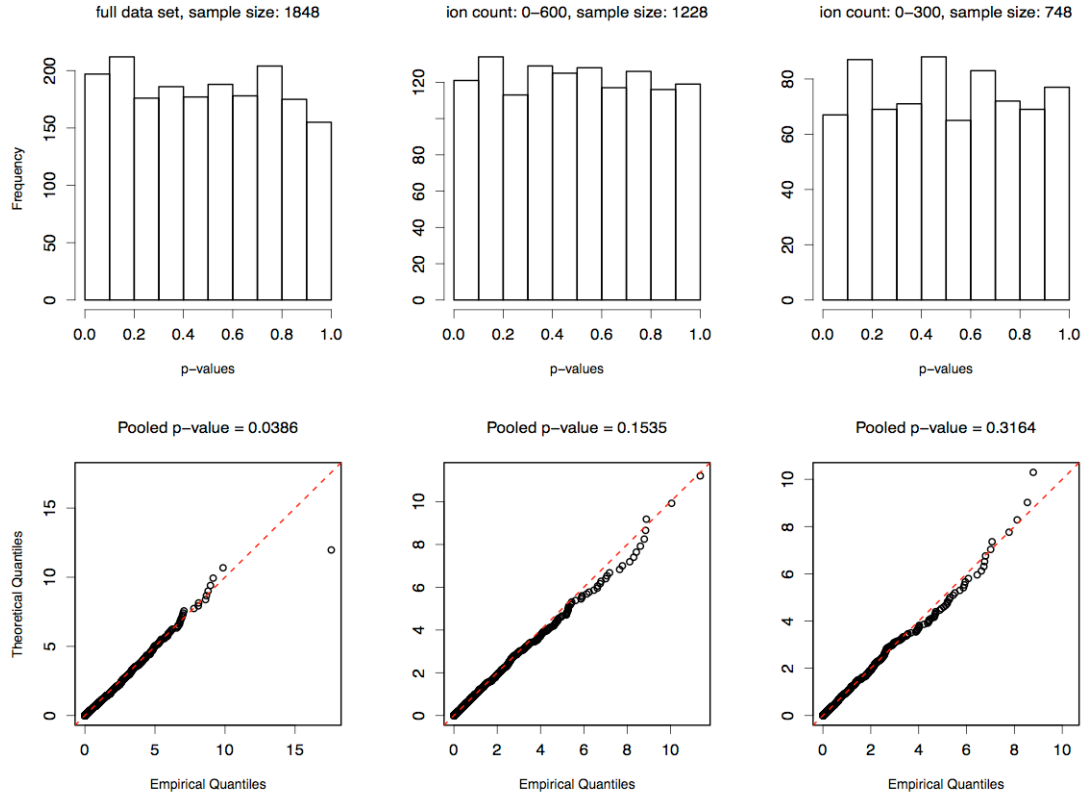


Figure 5.2 - Histograms of the p-values associated with the X^2 statistics obtained from the individual scans of the three datasets by using the basic model to construct the likelihood ratio test (top) and quantile-quantile plots of the X^2 statistics themselves as compared with the theoretical χ^2_1 -distribution (bottom). The p-values obtained for the full data-sets are listed above the quantile-quantile plots. The fit is substantially better than that obtained for the pure Poisson model although the basic model does break down at high ion counts.

5.6 Extension of Dynamic Range

One of the most important potential applications of the basic model is the estimation of the true ion count, when there is substantial detector saturation. However, this is a task that by its definition is most important in the region of very high ion counts, when the requirements that Ω is Gaussian, and that V_{ij} is known, are not met. Thus, until further progress is made at refining these aspects of the model, it will not be possible to obtain the maximal extension of dynamic range. We nevertheless believe it is important to illustrate the significant improvements in dynamic range that are

possible for the basic model in order to demonstrate the potential value of such research.

Whether or not Ω and Γ are known, the method of maximum likelihood can be used to obtain estimates of the true ion counts. It is when Ω and Γ are known that the greatest improvements in dynamic range is possible, since this allows us to incorporate extensive prior knowledge into the estimators. However, maximising the likelihood turns out to be particularly simple if neither functional form is known, as it is then possible to obtain a very simple analytical expression for the maximum likelihood estimate. In fact, for systems which meet the requirement that the dead time period always exceeds the time-of-flight range investigated, it turns out that this maximum likelihood estimate is equivalent to the dead time correction algorithm proposed by Coates in [52], although Coates' derivation was based solely on heuristic arguments.

If Ω and Γ are unknown, each $k_{i,j}$ is the outcome of a Binomial distribution with sample size $V_{i,j}$. and each binomial probability, $\rho_{i,j}$, must be estimated independently. The maximum likelihood estimator is

$$\hat{\rho}_{i,j} = \frac{k_{i,j}}{V_{i,j}}$$

and since

$$\rho_{i,j} = 1 - e^{-\Lambda_{i,j}/N_p}$$

we have

$$\begin{aligned}\hat{\Lambda}_{i,j} &= -N_p \log(1 - \hat{\rho}_{i,j}) \\ &= -N_p \log\left(1 - \frac{k_{i,j}}{V_{i,j}}\right)\end{aligned}$$

which is equivalent to the correction proposed in [51] and generalised in [52], so long as all the pulses that have been closed remain so over the time-of-flight range investigated.

While this correction algorithm provides substantial improvements in dynamic range over the raw recorded counts, it is limited by the fact that any inferences made regarding the Binomial probability must be mapped to the Poisson dimension in order to provide information on the true rate of ion arrivals. At low ion counts this is irrelevant, since the two quantities approximate each other as is seen by taking the first two terms of the Taylor expansion of equation 5.2:

$$\rho_{i,j} = 1 - e^{-\Lambda_{i,j}/N_p} \simeq 1 - \left(1 - \Lambda_{i,j}/N_p\right) = \Lambda_{i,j}/N_p$$

However, as is shown on Figure 5.3 below, for a fixed sample size the uncertainty associated with an estimate of the Binomial probability increases dramatically at high ion counts, following its mapping to the Poisson dimension. This effect will be more severe on the high-mass side of a mass peak when many of the pulses have been closed so that the sample size of the Binomial distribution is small. If the mass peak being examined is large enough that all pulses are closed, then the estimate of the Binomial probability in the time-of-flight tick in which the last remaining pulses are closed will be exactly 1 and the estimate of the Poisson rate will be infinity. This deficiency highlights the importance of incorporating prior knowledge of the data into the estimator used.

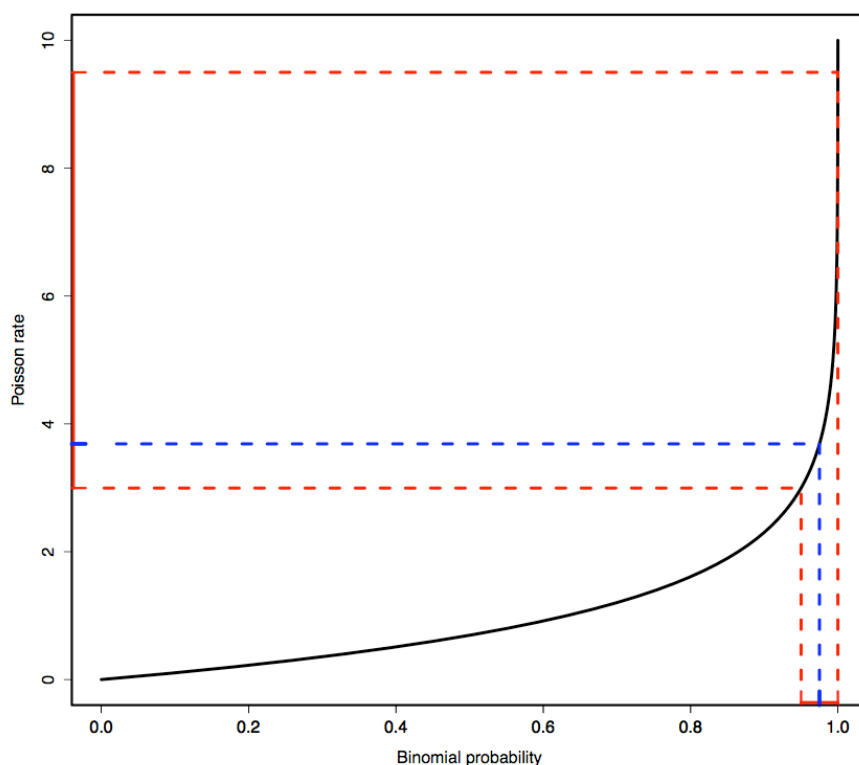


Figure 5.3 – The black line indicates the mapping that relates the Binomial probability to the Poisson rate. If an estimate is made of the Binomial probability (blue) and an associated confidence interval constructed (red) then these must be mapped to the Poisson-dimension in order to be properly interpretable. Therefore, if the Binomial probability is high, a very high sample size may be required to estimate the Poisson rate with any meaningful accuracy.

The performance of the various possible correction methods depends a great deal on the parameters of the system. Since the basic model is only a rough approximation to the true probability distribution and since the true functional forms of Γ and Ω remain unknown, a detailed performance comparison under a wide variety of settings would not be very instructive. However, the general characteristics of the correction methods are not surprising and can be illustrated with a few examples.

The plots on Figure 5.3 below show the true and the observed ion counts of two simulated, heavily saturated, mass peaks. To allow for greater realism, the simulations were not based directly on the basic model, but rather all pulses were simulated

independently (with distinct rate functions for their mass peaks corresponding to their distinct chromatographic retention times) and histogrammed with $N_p = 915$ and a TDC time resolution of 278 picoseconds. This is consistent with the settings of the mass spectrometer used in this study, which are fairly standard. The mass peak shown on the first plot has a total expected ion count of 1.5×10^6 over the entire peak, including other scans. The total expected ion count of the second peak is 3×10^6 but here σ_Ω is increased by a factor of 3 so that the ions are observed over a larger number of ticks meaning that the individual counts tend to be lower and, by the nature of the Poisson distribution, have a lower signal-to-noise ratio.

For these simulations, Ω was taken to be Gaussian and Γ skew-normal [96], the latter having greater computational stability than other models, such as the exponentially modified Gaussian [97], which is important when maximising the likelihood, although the exact choice makes little difference for the purposes of these simulations. Three correction methods were employed: one with Γ and Ω unknown (Coates' correction), one with only Ω known (the model also used in the likelihood ratio test for exact coelution), and one with both Γ and Ω known. For Coates' correction, N_p was set to 914 in order to avoid the infinities that would inevitably be obtained otherwise. For the two other methods, the resulting likelihood functions were maximised using the Newton-based method described above.

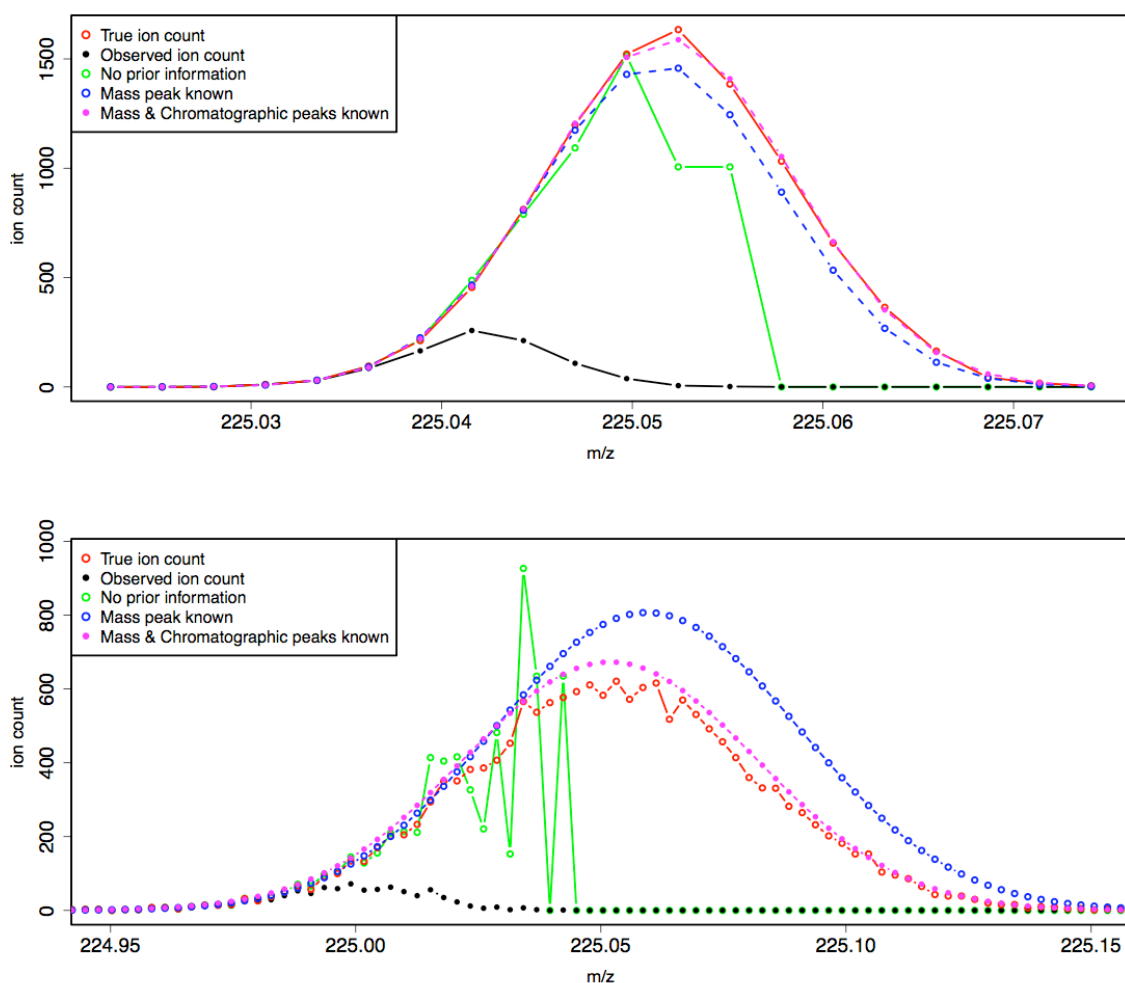


Figure 5.4 - True (red) and observed (black) ion counts, along with the statistical corrections obtained when both the functional forms of the mass and chromatographic peaks are known (magenta), when only the mass peak is known (blue) and when no such prior information is available (green). The rate of ion arrivals of the peak on the bottom plot is twice that of the peak on the top plot, but partly due to the heavier tail of the former, the correction methods that do incorporate prior information nevertheless provide reasonable estimates. The peaks shown are representative examples of mass peaks near the modes of the chromatographic peaks.

Despite the very heavy saturation, the correction methods that make use of prior knowledge provide sensible estimates for both simulations. As anticipated, Coates' correction method performs well at the low-mass side of the peak, where a large number of pulses are valid, but poorly at the high-mass end, where most of these have

been closed. Clearly this problem is especially pronounced for the second simulation. The relatively long tail of the mass peak of the second simulation allows for the methods that incorporate knowledge of the peak shapes to sample it quite extensively over time-of-flight ticks in which many of the pulses remain valid and thereby provide decent estimates despite the overall high ion count. It is of course unsurprising that the model which assumes knowledge of both peak shapes would perform better than the other methods, as it can pool information from the other scans. However, the method for pooling together information from distinct scans that this correction procedure implicitly provides is an important and legitimate advantage.

5.7 Future Applications

As was demonstrated in the validation section, the expansion of the likelihood function may be used to model, and test for, phenomena such as exact coelution, through the likelihood ratio test. Numerous other applications may be conceived of, but given the very general nature of the basic model, and the fact that it is only partially complete, a detailed discussion of all of these would be premature. However, we provide a brief outline of what we would regard as some of the most important potential applications of a more complete model, along with rough outlines for their use. Since the likelihood ratio test is asymptotic, care would in all cases have to be taken to ensure that the null distribution does approximate the predicted χ^2 distribution for all of these applications.

5.7.1 Deconvolution

An important practical problem in LC-MS concerns the scenario in which two or more compounds elute from the chromatographic column at roughly the same time *and* have very similar masses. If two compounds, R and S , are present, and it is assumed that there is no significant ionisation suppression or other interference, the resulting discretised rate function may be written

$$\Lambda_{i,j} = I^{(S)} \int_{\tau_i}^{\tau_{i+1}} \Gamma^{(S)}(\tau) d\tau \int_{t_j}^{t_j+t_\epsilon} \Omega^{(S)}(t) dt + I^{(R)} \int_{\tau_i}^{\tau_{i+1}} \Gamma^{(R)}(\tau) d\tau \int_{t_j}^{t_j+t_\epsilon} \Omega^{(R)}(t) dt$$

If Γ and Ω are known then this rate function may be used in equation 5.3, and all parameters of both S and R may be estimated by maximising the resulting likelihood function. It is trivial to generalise the procedure to three or more compounds. Of course, it is unlikely that the number of compounds, n , to be deconvoluted would be known in advance, but it might be estimated by applying the likelihood ratio test sequentially to the likelihood functions obtained for consecutive values of n .

5.7.2 Isotopic abundance patterns

A procedure very similar to the likelihood ratio test for exact coelution can be used to construct a test of hypothesis to determine whether the data derived from a given compound are consistent with a putative isotopic abundance pattern. We may use the same model used in the validation section, although a model in which Γ is known could also easily be designed.

Rather than applying the likelihood ratio test to a molecular species and one of its fragments, it is applied to the lowest-mass isotopologue of a molecular species, M , and to the adjacent isotopologue, containing one additional neutron, $M + 1$. If it can be assumed that the chromatographic peaks, the ionisation propensity and various other Binomial factors of I are identical for the two isotopologues then the ratio of the intensity factors can be written

$$\frac{I_i^{(M+1)}}{I_i^{(M)}} = \frac{I^{(M+1)} \int_{\tau_i}^{\tau_{i+1}} \Gamma(\tau) d\tau}{I^{(M)} \int_{\tau_i}^{\tau_{i+1}} \Gamma(\tau) d\tau} = \frac{n^{(M+1)}}{n^{(M)}}$$

where the final fraction is simply the natural abundance ratio of the isotopologues for the compound in question. This fraction can of course easily be calculated theoretically for a putative compound, and used as a constraint when fitting the basic model simultaneously to both M and $M + 1$.

Applying the likelihood ratio test to this constrained model and to the one obtained by fitting the basic model to the two isotopologues independently, should yield X^2 statistics that are distributed according to the χ^2_I -distribution for individual scans and

the χ^2_N -distribution for the full data-set. A p-value can then be obtained to indicate whether the data are consistent with a putative theoretical isotopic abundance pattern. Again, the procedure can easily be generalised to accommodate three or more isotopologues.

It is quite conceivable that this procedure could be generalised to account for the overlapping tails of the mass peaks. This would be particularly straightforward if these can be described via a simple superposition of the individual mass peaks. Such a procedure would in fact resolve the difficulties encountered in Chapter 4 and could thus allow for significantly improved estimates of isotopic abundance patterns, as there would be few remaining constraints on the amount of data that could be pooled. However, the basic model cannot be used for this task, as it is clear that the time-of-flight range over which the model would have to be applied would far exceed the duration of the dead time. Consequently, a more careful method of accounting for how many pulses are valid at a given tick would have to be developed.

5.7.3 Fine structure isotopic abundance patterns

More ambitiously, if the mass- and TDC time resolution of TOF mass spectrometers were improved, inferences might be made regarding fine structure isotopic abundance patterns. Since the mass of the +1 isotopologue will vary slightly depending on which element the additional neutron is associated with, all observed isotopologues, aside from the lowest-mass one will be mixtures of these species. Fine structure isotopic abundance patterns have been observed for FTICR mass spectrometers [98], but currently cannot be properly resolved for standard commercial TOF mass spectrometers, due to their more limited mass resolution.

However, if Ω were known, a model similar to the one used for deconvolution, but with the constraint that all Γ 's are identical, could be fitted to such a mixture to obtain estimates of the fine-structure abundances. As part of this study, attempts were made at fitting such a model to a +1 isotopologue comprised of C, H and O, however, owing to the limited mass- and TDC time resolution of the mass spectrometer used, a meaningful result could not be obtained. While a better mass resolution (i.e. a lower value of σ_Ω) would certainly be helpful to this approach, it is not inconceivable that

improvements in the TDC time resolution alone might enable reasonable estimates if data were pooled together from a large number of chromatographic peaks. Today, TDC-based TOF mass spectrometers are available that have a time resolution roughly ten times higher than the one used in this study [99] and while our ignorance of Ω would remain a severe impediment to the use of this approach, it is a line of research that could yield interesting results.

5.7.4 Test of hypothesis for mass

The errors associated with the mass estimates produced by current high-throughput mass spectrometers are not very well understood. They are typically quantified with descriptive statistics such as the root mean square error, and under varying conditions, such as distinct mass ranges and ion intensities. Aside from issues of reproducibility, these measures are not very satisfactory, as they do not provide statistically rigorous arguments for determining whether or not the theoretical mass of a putative compound is consistent with the observed time-of-flight. A test of hypothesis for mass, analogous to the test of hypothesis for isotope patterns outlined above, would address this problem.

If the standard equation relating time-of-flight to mass applied exactly, that is if

$$m/z = \frac{2eU}{d^2} \mu_{\Omega}^2$$

it would in principle be straightforward to construct a likelihood ratio test in which the unconstrained model involves estimating μ_{Ω} while the constrained model has the μ_{Ω} specified in advance according to the theoretical mass of a putative compound. However, the true relationship between time-of-flight and mass is far more complicated and involves numerous additional parameters to those listed above [66]. Furthermore, since U and d are known only to a rather limited degree of accuracy due to nuisance effects such as temperature fluctuations [14], frequent calibration is needed.

It is conceivable that a model could be constructed which accommodates for further terms in the time-of-flight equation and for internal calibration by expanding the likelihood function accordingly, but its development would undoubtedly be a highly ambitious task. Even so, it is difficult to understate the potential utility of such a procedure, as it could essentially resolve the longstanding problem of quantifying the uncertainty of mass estimates. In addition, it could enable significantly improved mass estimates since we could very likely pool together measurements across scans (and derivatives) as was done for isotopic abundance patterns in Chapter 4.

5.8 Discussion

This chapter has attempted to establish the core formalism required to conduct a rigorous statistical analysis of LC-TOFMS data for mass spectrometers employing TDCs. Although demanding simplifying assumptions were made in formulating the basic model, the χ^2 statistics obtained from its application to related fragment pairs through the likelihood ratio test conform closely to the predicted distribution so long as the ion count is not too high. The basic model's rather high level of detail does not appear to be unwarranted since the fit of the corresponding statistics obtained for the more parsimonious pure Poisson model of continuum data deteriorates much faster.

While the basic model can in fact already be applied usefully to LC-TOFMS data, it is clear that its uses would be greatly expanded if the functional forms of the mass and chromatographic peaks were known, including any dependence they might have on the overall ion count. Not only would this allow us to address a potentially very wide range of data analytical problems in a statistically rigorous manner, but it could enable significant improvements in mass and isotopic abundance estimates through the pooling of measurements across distinct scans.

However, even if the shapes of the mass and chromatographic peaks were fully understood, the basic model would not be applicable to large mass peaks owing to the model's assumption that pulses that are closed remain so over the "duration" of a mass peak. We might, as in previous chapters, restrict ourselves to working with heavier isotopologues or with the tails of chromatographic peaks, but this entails the loss of potentially valuable information. Any model aiming to describe mass peaks of

arbitrary sizes will have to relax the assumption that closed pulses always remain so, which in turn necessitates careful accounting of how many pulses are valid at a given tick of the TDC and that is the topic of the following chapter.

6 Detailed modelling of dead time effects

While Coates' correction is not a rigorous algorithm, it has one important advantage over the basic model in that it attempts to account for the possibility that a pulse that is closed by an incoming ion may reopen over the course of a mass peak. The fact that the basic model doesn't consider this scenario is likely a key reason why it breaks down at high counts when mass peaks are sufficiently wide that they may exceed the dead time. In this chapter, we take a strongly theoretical approach. We generalise the basic model to accommodate for the possibility that pulses reopen.

The models presented in this chapter are considerably more complicated than the ones previously encountered and their mathematical formulation requires some thought and effort. In addition the task of obtaining maximum likelihood estimates is much more challenging as the models contain a large number of unobserved, discrete-valued parameters. For this reason the Newton-type algorithm described in the previous chapter is not capable of maximising the likelihood function. Instead, a method for doing so by means of the Expectation-Maximisation (EM) algorithm [100] will be presented.

The model initially developed will make the exact same assumptions made by Coates' correction, namely that pulses are i.i.d., that incoming ions arrive in middle of a tick, and that the dead time is constant and lasts for an integer number of ticks. While these assumptions are somewhat demanding, they are also made by a correction method proposed in a Waters patent application [48], which is taken as evidence that they are not entirely unwarranted. Nevertheless, this model is not sufficiently accurate that we can use it to extend the range of ion counts over which we can construct rigorous statistical tests, and it is clear that a more complete treatment would rely on a weaker set of assumptions.

To illustrate the plausibility of such an approach, we extend the model further to describe the scenario where ions may arrive at any point throughout the tick of the TDC clock. However, as will be explained, the detailed characteristics of the detector system do not yet appear to be sufficiently well understood that we can extend the

model to the scenario where the dead time can be of variable duration. This is undoubtedly an important constraint, but it is one that can only be overcome through careful investigations of the behaviour of the various components of the detector system and that is a task that goes beyond the scope of this thesis.

6.1 Completing the Coatesian model

For simplicity, we will in the following change our time scale to match the units of the TDC clock so that we can refer to the time period corresponding to the i th tick by writing $t = i$. We will begin by describing a single mass peak and introduce the chromatographic dimension later.

On removing the assumption that the dead time always exceeds the width of the mass peak, it becomes necessary to carefully account for how many of the N_p pulses are valid at a given tick. We recall that the dead time lasts D ticks and that it is “extending” so that if a pulse has been closed by an incoming ion and it is struck by an additional ion prior to recovering, its period of dead time will be extended by a further D ticks. It is clear that the number of valid pulses at time $t = i$ may be split into three components:

- The number of valid pulses at $t = i - 1$ (V_{i-1})
- The number of pulses closed at $t = i - 1$ (k_{i-1})
- The number of closed pulses that recovered just prior to $t = i$ (R_{i-D})

That is,

$$V_i = V_{i-1} - k_{i-1} + R_{i-D}$$

The third term may be written

$$\begin{aligned} R_{i-D} &= [\text{the total number of pulses that have been scheduled to open at } t = i] \\ &\quad - [\text{the number of these that had their dead time further extended prior to } t = i] \\ &= [\text{the number of pulses closed or extended at } t = i - D] \\ &\quad - [\text{the number of these that had their dead time further extended prior to } t = i] \end{aligned}$$

In order to work with these quantities it is helpful to define a “closure matrix”, $\mathbf{C}$, which describes the unobserved events of the system, namely the extension of the dead time of closed pulses by incoming ions. Define the entry at the i th row and j th column ($C_{i,j}$) to be the number of pulses, closed or extended at j , that were further extended at i . For convenience, we may in addition write the k_i on the diagonal of the closure matrix, since there is no need for the $C_{i,i}$ of the diagonal when all ions are assumed to arrive exactly in the middle of the ticks. Below is shown the form of the matrix with $D = 4$, and a total of seven ticks, that is $M = 7$:

k_1	0	0	0	0	0	0
$C_{2,1}$	k_2	0	0	0	0	0
$C_{3,1}$	$C_{3,2}$	k_3	0	0	0	0
$C_{4,1}$	$C_{4,2}$	$C_{4,3}$	k_4	0	0	0
0	$C_{5,2}$	$C_{5,3}$	$C_{5,4}$	k_5	0	0
0	0	$C_{6,3}$	$C_{6,4}$	$C_{6,5}$	k_6	0
0	0	0	$C_{7,4}$	$C_{7,5}$	$C_{7,6}$	k_7

Suppose we wish to evaluate the number of pulses that were either closed or extended at $t = 3$ and then extended at $t = 5$, namely $C_{5,3}$ (highlighted in blue). We first need to add up all of the pulses closed or extended at $t = 3$ (highlighted in green). Next, subtract from this number all of the pulses closed or extended at $t = 3$ that were further extended at $t = 4$ (red). We now know exactly how many of the pulses closed or extended at $t = 3$ were left undisturbed by incoming ions until the start of $t = 5$. Let ρ_5 be the probability that a given one of these pulses is struck by at least one ion in the 5th tick, i.e. the familiar Binomial probability defined in equation 5.2. We can then finally obtain $C_{5,3}$ through the random variable $\text{Bin}[C_{3,1} + C_{3,2} + k_3 - C_{4,3}, \rho_5]$.

The i th row of the closure matrix lists all of the pulses that were closed (k_i) or extended ($C_{i,i-D+1}, \dots, C_{i,i-1}$) at time $t = i$. Below the diagonal, the i th column lists the times at which pulses, closed or extended at $t = i$, are further extended. Thus, more generally, for all meaningful values of i and j we have

$$C_{i,j} | \phi \sim \text{Bin} \left[\left(k_j + \sum_{k=j-D+1}^{j-1} C_{j,k} - \sum_{k=j+1}^{i-1} C_{k,j} \right), \rho_i \right],$$

where ϕ denotes the values of all other $C_{i,j}$ and k_j in the expression.

Those pulses that are *not* further extended, but recover (the R_i), can be expressed as the sum of the i th row, minus the sum of the i th column below the diagonal:

$$R_i = k_i + \sum_{k=i-D+1}^{i-1} C_{i,k} - \sum_{k=i+1}^{i+D-1} C_{k,i}$$

It can also be useful to write the R_i in their respective columns of the closure matrix:

k_1	0	0	0	0	0	0
$C_{2,1}$	k_2	0	0	0	0	0
$C_{3,1}$	$C_{3,2}$	k_3	0	0	0	0
$C_{4,1}$	$C_{4,2}$	$C_{4,3}$	k_4	0	0	0
R_1	$C_{5,2}$	$C_{5,3}$	$C_{5,4}$	k_5	0	0
0	R_2	$C_{6,3}$	$C_{6,4}$	$C_{6,5}$	k_6	0
0	0	R_3	$C_{7,4}$	$C_{7,5}$	$C_{7,6}$	k_7

Going back to the expression for V_i :

$$V_i = V_{i-1} - k_{i-1} + R_{i-D}$$

we see that, by iteratively rewriting V_i in terms of V_{i-1} , k_{i-1} and R_{i-D} , we obtain

$$V_i = N_p - \sum_{j=1}^{i-1} k_j + \sum_{j=1}^{i-D} R_j.$$

or to put it more plainly, the number of valid pulses at the i th tick is equal to the total number of pulses, minus all of the valid pulses that have been closed at any point so far, plus all of the closed pulses that have recovered, at any point so far.

In evaluating the last sum it is helpful to consider $R_1 + R_2 + R_3$:

k_1	0	0	0	0	0	0
$C_{2,1}$	k_2	0	0	0	0	0
$C_{3,1}$	$C_{3,2}$	k_3	0	0	0	0
$C_{4,1}$	$C_{4,2}$	$C_{4,3}$	k_4	0	0	0
R_1	$C_{5,2}$	$C_{5,3}$	$C_{5,4}$	k_5	0	0
0	R_2	$C_{6,3}$	$C_{6,4}$	$C_{6,5}$	k_6	0
0	0	R_3	$C_{7,4}$	$C_{7,5}$	$C_{7,6}$	k_7

where we note that six terms cancel out ($\pm C_{2,1}$, $\pm C_{3,1}$ and $\pm C_{3,2}$). In general we see that most $C_{i,j}$ will cancel out when evaluating this sum so that:

$$\sum_{j=1}^{i-D} R_j = \sum_{j=1}^{i-D} k_j - \sum_{j=i-2D+2}^{i-D} \sum_{k=i-D+1}^{D-1+j} C_{k,j}$$

and therefore

$$V_i = N_p - \sum_{j=i-D+1}^{i-1} k_j - \sum_{j=i-2D+2}^{i-D} \sum_{k=i-D+1}^{D-1+j} C_{k,j}$$

Recall that for the basic model, the distribution of a given k_i was given by

$$P(k_i) = \binom{V_i}{k_i} \rho_i^{k_i} (1 - \rho_i)^{V_i - k_i}$$

We will in fact continue to work with this distribution, but will now use our new expression for V_i , which gives us

$$P(k_i | \phi, \rho_i) = \binom{N_p - \hat{\mathbf{A}}_{j=i-D+1}^{i-1} k_j - \sum_{j=i-2D+2}^{i-D} \sum_{k=i-D+1}^{D-1+j} C_{k,j}}{k_i} \rho_i^{k_i} (1 - \rho_i)^{N_p - \sum_{j=i-D+1}^i k_j - \sum_{j=i-2D+2}^{i-D} \sum_{k=i-D+1}^{D-1+j} C_{k,j}}$$

Both the k_i and the $C_{i,j}$ depend on the values of prior entries of $\mathbf{C}$, but we can write out the joint distribution of all of these entries by simply taking the product of their probability distributions⁴

$$P(\mathbf{k}, \mathbf{C} | \theta) = \prod_{i=1}^M \left\{ \binom{N_p - \sum_{j=i-D+1}^{i-1} k_j - \sum_{j=i-2D+2}^{i-D} \sum_{k=i-D+1}^{D-1+j} C_{k,j}}{k_i} \rho_i^{k_i} (1 - \rho_i)^{N_p - \sum_{j=i-D+1}^i k_j - \sum_{j=i-2D+2}^{i-D} \sum_{k=i-D+1}^{D-1+j} C_{k,j}} \right\} \prod_{i=2}^M \prod_{j=i-D+1}^{i-1} \left\{ \binom{k_j + \sum_{k=j-D+1}^{j-1} C_{j,k} - \sum_{k=j+1}^{i-1} C_{k,j}}{C_{i,j}} \rho_i^{C_{i,j}} (1 - \rho_i)^{k_j + \sum_{k=j-D+1}^{j-1} C_{j,k} - \sum_{k=j+1}^i C_{k,j}} \right\}$$

where θ denotes all parameters required to specify the rate function, that is, the total rate of ion arrivals and the parameters required to define the chromatographic and mass peak shapes. To write this more concisely we define a “cumulative closure matrix”, $\mathbb{C}$, whose entries are defined as

$$\mathbb{C}_{i,j} = k_j + \sum_{k=j-D+1}^{j-1} C_{j,k} - \sum_{k=j+1}^i C_{k,j}$$

so that $\mathbb{C}_{i,j}$ denotes the sample size of the Binomial distribution from which $C_{i+1,j}$ is drawn. Note that $R_i = \mathbb{C}_{i+D-1,i}$. Our probability distribution can then be expressed much more simply as

$$P(\mathbf{k}, \mathbf{C} | \theta) = \prod_{i=1}^M \left\{ \binom{V_i}{k_i} \rho_i^{k_i} (1 - \rho_i)^{V_i - k_i} \right\} \prod_{i=2}^M \prod_{j=i-D+1}^{i-1} \left\{ \binom{\mathbb{C}_{i-1,j}}{C_{i,j}} \rho_i^{C_{i,j}} (1 - \rho_i)^{\mathbb{C}_{i,j}} \right\}$$

And since scans are taken to be independent, it is straightforward to generalise the model to describe N scans:

⁴ $\mathbf{k}$ is strictly speaking included in $\mathbf{C}$, but we nevertheless write $P(\mathbf{k}, \mathbf{C})$ for clarity. Any $C_{i,j}$ for which the indices are non-positive may be left out in the following.

$$P(\mathbf{k}, \mathbf{C} | \theta) = \prod_{h=1}^N \left\{ \prod_{i=1}^M \left\{ \begin{pmatrix} V_{h,i} \\ k_{h,i} \end{pmatrix} \rho_{h,i}^{k_{h,i}} (1 - \rho_{h,i})^{V_{h,i} - k_{h,i}} \right\} \prod_{j=i-D+1}^{i-1} \left\{ \begin{pmatrix} \mathbb{C}_{h,i-1,j} \\ C_{h,i,j} \end{pmatrix} \rho_{h,i}^{C_{h,i,j}} (1 - \rho_{h,i})^{\mathbb{C}_{h,i,j} - C_{h,i,j}} \right\} \right\}$$

where $\mathbf{C}$ is now regarded as a three-dimensional array.

Of course, the distribution that we are primarily interested in is not $P(\mathbf{k}, \mathbf{C} | \theta)$ but rather $P(\mathbf{k} | \theta)$, and by the law of total probability this can be expressed as

$$P(\mathbf{k}) = \sum P(\mathbf{k}, \mathbf{C})$$

where the sum goes over all possible permutations of the closure matrix. But given the large number of possible permutations, this expression is far too complex to be of much practical use. Fortunately it is possible to use $P(\mathbf{k}, \mathbf{C} | \theta)$ to obtain maximum likelihood estimates for $P(\mathbf{k} | \theta)$.

6.2 Maximum likelihood estimates for the Coatesian model

We may interpret the expression for $P(\mathbf{k}, \mathbf{C} | \theta)$ as its likelihood function

$$L(\theta | \mathbf{k}, \mathbf{C}) = \prod_{h=1}^N \left\{ \prod_{i=1}^M \left\{ \begin{pmatrix} V_{h,i} \\ k_{h,i} \end{pmatrix} \rho_{h,i}^{k_{h,i}} (1 - \rho_{h,i})^{V_{h,i} - k_{h,i}} \right\} \prod_{j=i-D+1}^{i-1} \left\{ \begin{pmatrix} \mathbb{C}_{h,i-1,j} \\ C_{h,i,j} \end{pmatrix} \rho_{h,i}^{C_{h,i,j}} (1 - \rho_{h,i})^{\mathbb{C}_{h,i,j} - C_{h,i,j}} \right\} \right\}$$

where θ denotes the parameters to be estimated and

$$\begin{aligned}
\mathbb{C}_{h,i,j} &= k_{h,j} + \sum_{k=j-D+1}^{j-1} C_{h,j,k} - \sum_{k=j+1}^i C_{h,k,j} \\
V_{h,i} &= N_p - \sum_{j=i-D+1}^{i-1} k_j - \sum_{j=i-2D+2}^{i-D} \sum_{k=i-D+1}^{D-1+j} C_{k,j} \\
\rho_{h,i} &= 1 - e^{-\Lambda_{h,i}/N_p}
\end{aligned}$$

If the C_{ij} were known, then finding the maximum likelihood estimates would not be difficult. In the case where Γ and Ω are known we would be able to apply the same Newton-based optimisation method used in the previous chapter. If they are not, and all of the $\rho_{h,i}$ have to be estimated independently, it turns out that, as was the case with the basic model, there is an analytical expression for the maximum likelihood estimator which turns out to be quite simple and intuitive:

$$\hat{\rho}_{h,i} = \frac{k_{h,i} + \sum_{j=i-D+1}^{i-1} C_{h,i,j}}{N_p} \quad (6.1)$$

This can then be related to $\Lambda_{h,i}$ by the usual mapping. The derivation is given in Appendix A. However, since we cannot observe the $C_{h,i,j}$, and since we do not have the computational resources to calculate the desired likelihood function, $L(\theta|\mathbf{k})$ directly, a rather more complicated method must be used to obtain the maximum likelihood estimates.

6.2.1 The EM algorithm

A method that is often used when discrete unobserved variables are required to write the likelihood in a simple form is the EM algorithm. The algorithm is applied iteratively and in this case would entail the following two steps

Expectation: calculate $\mathbf{C}^i = E[\mathbf{C} | \theta^i, \mathbf{k}]$, where θ^i is the parameter estimate at the i th iteration.

Maximisation: let $\theta^{i+1} = \arg_{\theta} \max L(\theta | \mathbf{C}^i)$

If these two steps are applied until convergence, θ^i will reach a local maximum or saddle point [100]. For this particular system the maximisation step can either be tackled through the Newton-based method discussed earlier if Γ and Ω are known, or simply by means of equation 6.1 if they are not. However, calculating the expectation is considerably more challenging, since a simple analytical expression has not been found. Note that it would be trivial to calculate $E[\mathbf{C}, \mathbf{k} | \theta]$ as this only involves standard, well defined Binomial distributions, but $E[\mathbf{C} | \mathbf{k}, \theta]$ is a much more complicated quantity.

The exact value of the expectation is given by

$$E[\mathbf{C} | \theta^i, \mathbf{k}] = \sum P(\mathbf{C} | \mathbf{k}, \theta^i) \mathbf{C}$$

where, again, the sum goes over all possible permutations of $\mathbf{C}$ that are consistent with the observed diagonal $\mathbf{k}$. In principle, this sum could be evaluated exactly since it is straightforward to calculate the value of any given $P(\mathbf{k}, \mathbf{C} | \theta)$. However, except for very small systems, the number of permutations to go through would be prohibitively large.

6.2.2 Rejection sampling for the Coatesian model

Monte Carlo simulations can provide a satisfactory approximation to the required expectation. If we can draw a reasonable number of samples from $\mathbf{C} | \theta^i, \mathbf{k}$ we can average them to obtain an approximation to $E[\mathbf{C} | \theta^i, \mathbf{k}]$. This approach is referred to as the Monte Carlo Expectation-Maximisation (MCEM) algorithm and it is not original [101]. Note that since the various scans are independent, we need only find a method for sampling from the 2-dimensional closure matrix and apply it to each of the scans in turn. A Markov Chain Monte Carlo sampler [102] was investigated, but the burn-in period (the period of time required for it to converge to the target distribution) was found to be far too long for the approach to be practical. Instead a form of rejection sampling [103] was used.

Suppose we wish to draw samples from a given probability distribution, $f(\mathbf{x}|\boldsymbol{\theta})$, referred to as the “target distribution”. If we can draw samples from a “proxy distribution”, $g(\mathbf{x}|\boldsymbol{\theta})$, such that

$$Hg(\mathbf{x}|\boldsymbol{\theta}) \geq f(\mathbf{x}|\boldsymbol{\theta}) \quad \forall \mathbf{x}$$

where H is a constant, then we can sample from $f(\mathbf{x}|\boldsymbol{\theta})$ by the following procedure:

1. Draw a sample, $\mathbf{v}$, from $g(\mathbf{x}|\boldsymbol{\theta})$
2. Draw a sample, u , from $U[0, Hg(\mathbf{v}|\boldsymbol{\theta})]$ (the uniform distribution)
3. If $u < f(\mathbf{v}|\boldsymbol{\theta})$ we accept $\mathbf{v}$
4. If not, we reject it and start again

Our expression for the target distribution, $f(\mathbf{x}|\boldsymbol{\theta})$, does not have to be properly normalised in order for this approach to be valid. So long as $f(\mathbf{x}|\boldsymbol{\theta})$ can be evaluated everywhere up to a fixed constant of proportionality, for which the above inequality holds, the procedure will produce random samples for the proper target distribution.

For the Coatesian system, our target distribution is $P(\mathbf{C}|\boldsymbol{\theta}, \mathbf{k})$, but since

$$P(\mathbf{C}|\mathbf{k}, \boldsymbol{\theta}) = \frac{P(\mathbf{C}, \mathbf{k}|\boldsymbol{\theta})}{P(\mathbf{k}|\boldsymbol{\theta})}$$

we see that the much simpler distribution, $P(\mathbf{C}, \mathbf{k}|\boldsymbol{\theta})$, is proportional to our target distribution and we may therefore use it for the rejection sampling.

We will use the proxy distribution

$$Q(\mathbf{C}|\mathbf{k}, \boldsymbol{\theta}) = \prod_{i=1}^M \prod_{j=i-D+1}^{i-1} \left\{ \begin{pmatrix} \mathbb{C}_{i-1,j} \\ C_{i,j} \end{pmatrix} \rho_i^{C_{i,j}} (1 - \rho_i)^{\mathbb{C}_{i,j}} \right\}$$

where the diagonal of $\mathbf{C}$ is fixed to $\mathbf{k}$, regardless of whether this is consistent with the corresponding V_i (the sample will always be rejected if it isn't). We choose this distribution because of its similarity to $P(\mathbf{C}|\mathbf{k},\boldsymbol{\theta})$ and its relatively simple Binomial nature, which reduces the computational cost of sampling from it.

We now require a constant H , such that

$$H \geq \frac{P(\mathbf{C},\mathbf{k}|\boldsymbol{\theta})}{Q(\mathbf{C}|\mathbf{k},\boldsymbol{\theta})}$$

for all possible $\mathbf{C}$. Clearly

$$\frac{P(\mathbf{C},\mathbf{k}|\boldsymbol{\theta})}{Q(\mathbf{C}|\mathbf{k},\boldsymbol{\theta})} = \prod_{i=1}^M \left\{ \binom{V_i}{k_i} \rho_i^{k_i} (1-\rho_i)^{V_i-k_i} \right\}$$

Ideally we would let H equal the global maximum of this expression, as that would result in the highest possible acceptance rate. Finding this maximum without enumerating every possible $\mathbf{C}$ is not an easy task however, so we will restrict ourselves to finding an upper bound for it.

Since we know the V_i for the first D bins (before the dead time which started at the first tick ends), we need only maximise the product

$$\prod_{i=D+1}^M \left\{ \binom{V_i}{k_i} \rho_i^{k_i} (1-\rho_i)^{V_i-k_i} \right\}$$

We may find our upper bound by choosing the V_i that maximise each of the terms in the above product, with only limited consideration for their consistency with the k_i .

We know that $V_i \geq k_i$ and that $V_i \geq V_{i-1} - k_{i-1}$ which establishes the lower bound, and that $V_i \leq N_p$ and $V_i \leq V_{i-1} + k_{i-D}$ which establishes the upper bound.

We have

$$\begin{aligned}
V_{D+1} &\in \left[\max(k_{D+1}, V_D - k_D), \min(N_p, V_D + k_1) \right] \\
V_{D+2} &\in \left[\max(k_{D+2}, \min(V_{D+1}) - k_{D+1}), \min(N_p, \max(V_{D+1}) + k_2) \right] \\
&\vdots \\
V_{D+i} &\in \left[\max(k_{D+i}, \min(V_{D+i-1}) - k_{D+i-1}), \min(N_p, \max(V_{D+i-1}) + k_i) \right]
\end{aligned}$$

and we may simply choose the V_i that maximise the binomial product, subject to the above constraints. It is very likely that a tighter bound could be found, but since high efficiency will not be essential until the algorithm is applied to large data-sets, this task will be not be considered further here.

The above simulation technique was validated by applying it to the system which has $\mathbf{k} = (4 \ 2 \ 3 \ 3)^T$, $\mathbf{p} = (0.5 \ 0.5 \ 0.5 \ 0.5)^T$ and $D = 2$, and which is sufficiently simple that the theoretical mean of $C_{2,l}$ can be calculated without too much difficulty. While this system is very small, it requires all of the features of the full Coatesian model and it is highly unlikely that the cumulative mean of the simulated values would converge to the theoretical mean if the simulation procedure were not valid generally. The result is shown on Figure 6.1 below.

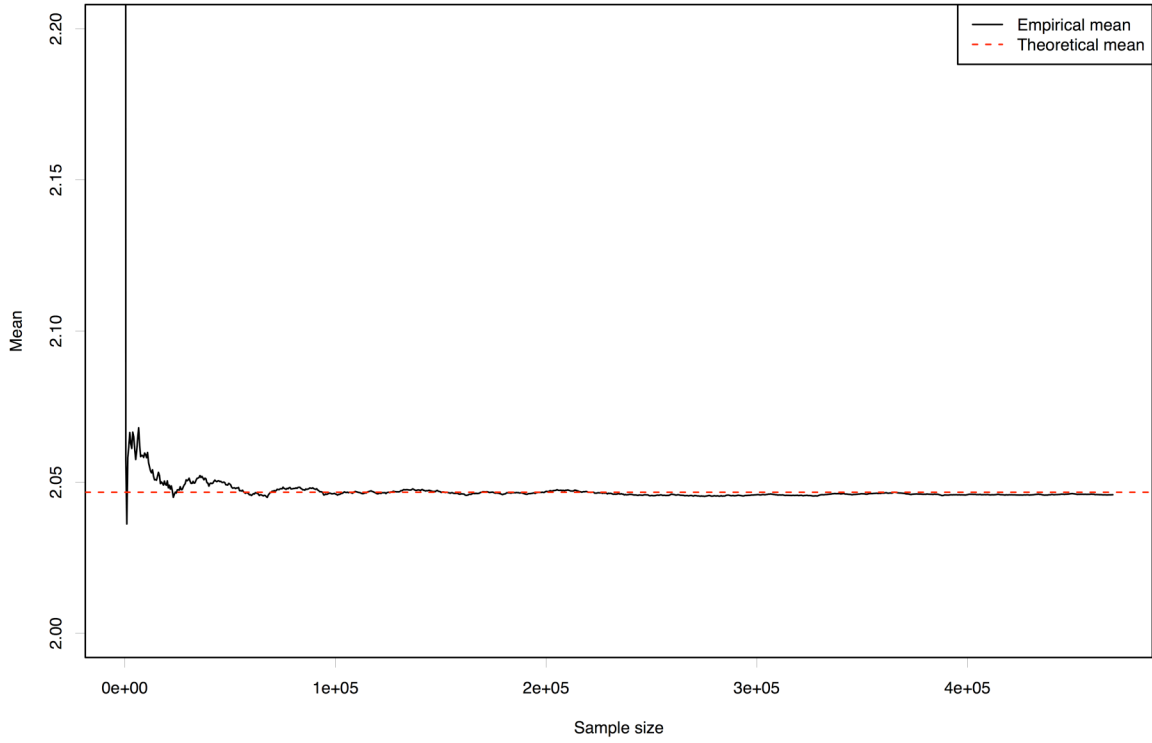


Figure 6.1 – Validation of the rejection-sampling algorithm described above. Samples are generated from $C_{2,l}$ for the system with $k = (4 \ 2 \ 3 \ 3)^T$, $\rho = (0.5 \ 0.5 \ 0.5 \ 0.5)^T$, $D = 2$ and the empirical mean is calculated as a function of sample size. The theoretical mean is also indicated.

Clearly, the mean of the simulated samples appears to be consistent with the theoretical value, and as a result all the steps required for the MCEM algorithm to be applied are established.

6.3 A generalised model for arbitrarily timed ion arrivals

As mentioned earlier, the assumptions required for the Coatesian model are very demanding. One could imagine a detector system for which the distribution of the dead time has sufficiently low variance that it might be considered roughly constant and whose mean value corresponds to an integer number of ticks. However, it is difficult to conceive of a system for which the assumption that ions always arrive at the centre of a tick can be justified.

If it is assumed that ions can arrive at any point in a tick, we must account for the possibility that a pulse closed at tick $i - D$, is reopened during the course of tick i , prior to it being struck by another ion in that same tick. Since the probability of such an event taking place depends on the shape of the rate function at both the $i - D$ th and i th ticks, it must be modelled rather differently to those recorded ion counts that result from an ion striking the detector in a pulse that is valid throughout the i th tick. In the following we must regard the observed ion count at the i th tick (k_i) as the sum of those pulses closed in the latter, conventional fashion (w_i) and of those pulses closed in the former fashion (f_i), that is $k_i = w_i + f_i$.

The probability distribution of the recorded ion counts derived from the conventional legitimate pulses will remain essentially unchanged as

$$P(\mathbf{w} | \mathbf{V}, \boldsymbol{\theta}) = \prod_{i=1}^M \left\{ \binom{V_i}{w_i} \rho_i^{w_i} (1 - \rho_i)^{V_i - w_i} \right\}$$

although we now have

$$V_i = V_{i-1} + \mathbb{C}_{i-1, i-D-1} - w_{i-1}$$

The closure matrix will also be modified slightly. The diagonal will be set to $\mathbf{w}$ rather than $\mathbf{k}$ in order to avoid double-counting when calculating the entries of the cumulative closure matrix. Furthermore the number of non-zero entries below the diagonal will be increased by one for each column, from $D - 1$ to D . This is to account for the pulses that were closed in the $i - D$ th tick and closed again at the i th tick, either before or after they had reopened. If it was after, then the ion would be observed, and so this event would count towards f_i . In fact the probability distribution of $\mathbf{f}$ can be written as

$$P(\mathbf{f} | \mathbf{C}, \boldsymbol{\theta}) = \prod_{i=1}^{M-D} \left\{ \binom{C_{i+D, i}}{f_{i+D}} q_i^{f_{i+D}} (1 - q_i)^{C_{i+D, i} - f_{i+D}} \right\}$$

where the $C_{i+D, i}$ are the newly added entries to the closure matrix. The q_i are defined as follows:

$$q_i = P \left(\begin{array}{l} \text{the time from the last ion arrival in } [t_{i-D-1}, t_{i-D}] \text{ to the first in } [t_{i-1}, t_i] \text{ is greater than } D \\ \text{at least one ion arrival in both } [t_{i-D-1}, t_{i-D}] \text{ and } [t_{i-1}, t_i] \end{array} \right)$$

If $\lambda(t)$ is the rate function for a single pulse in the time-of-flight dimension, it can be shown that

$$q_i = \frac{1}{\left(e^{\int_{t_{i-D-1}}^{t_{i-D}} \lambda(t) dt} - 1 \right) \left(e^{-\int_{t_{i-1}}^{t_i} \lambda(t) dt} - 1 \right)} \int_0^\varepsilon \left\{ \lambda(t_{i-D-1} + \tau_\alpha) e^{\int_{t_{i-D-1}}^{t_{i-D-1} + \tau_\alpha} \lambda(t) dt} \left[e^{-\int_{t_{i-1}}^{t_i} \lambda(t) dt} - e^{-\int_{t_{i-1}}^{t_{i-1} + \tau_\alpha} \lambda(t) dt} \right] \right\} d\tau_\alpha$$

where ε indicates the length of one tick. The derivation is given in Appendix B. A closed form solution to the above integral has not been found and so it must be evaluated numerically at present. Note that working with this generalised model is rather more difficult than before if Ω is not known, as we must then estimate all of the q_i in addition to the ρ_i .

The full likelihood of this new generalised system can then be written⁵

$$L(\mathbf{w}, \mathbf{f}, \mathbf{C} | \boldsymbol{\theta}) = \prod_{h=1}^N \left\{ \prod_{i=1}^M \left\{ \begin{pmatrix} V_{h,i} \\ w_{h,i} \end{pmatrix} \rho_{h,i}^{w_{h,i}} (1 - \rho_{h,i})^{V_{h,i} - w_{h,i}} \right\} \prod_{i=1}^{M-D} \left\{ \begin{pmatrix} C_{h,i+D,i} \\ f_{h,i+D} \end{pmatrix} q_{h,i+D}^{f_{h,i+D}} (1 - q_{h,i+D})^{C_{h,i+D,i} - f_{h,i+D}} \right\} \prod_{i=2}^M \prod_{j=i-D}^{i-1} \left\{ \begin{pmatrix} \mathbb{C}_{h,i-1,j} \\ C_{h,i,j} \end{pmatrix} \rho_{h,i}^{C_{h,i,j}} (1 - \rho_{h,i})^{\mathbb{C}_{h,i,j}} \right\} \right\}$$

The closure matrices for the Coatesian and the generalised models are shown on Table 6.1 below.

⁵ It would suffice to write $L(\mathbf{f}, \mathbf{C} | \boldsymbol{\theta})$ rather than $L(\mathbf{w}, \mathbf{f}, \mathbf{C} | \boldsymbol{\theta})$ as $\mathbf{w}$ is given by the diagonal of $\mathbf{C}$, however it will be written explicitly here and in the following for the sake of clarity.

k_1	0	0	0	0	0	0
$C_{2,1}$	k_2	0	0	0	0	0
$C_{3,1}$	$C_{3,2}$	k_3	0	0	0	0
$C_{4,1}$	$C_{4,2}$	$C_{4,3}$	k_4	0	0	0
0	$C_{5,2}$	$C_{5,3}$	$C_{5,4}$	k_5	0	0
0	0	$C_{6,3}$	$C_{6,4}$	$C_{6,5}$	k_6	0
0	0	0	$C_{7,4}$	$C_{7,5}$	$C_{7,6}$	k_7

w_1	0	0	0	0	0	0
$C_{2,1}$	w_2	0	0	0	0	0
$C_{3,1}$	$C_{3,2}$	w_3	0	0	0	0
$C_{4,1}$	$C_{4,2}$	$C_{4,3}$	w_4	0	0	0
$C_{5,1}$	$C_{5,2}$	$C_{5,3}$	$C_{5,4}$	w_5	0	0
0	$C_{6,2}$	$C_{6,3}$	$C_{6,4}$	$C_{6,5}$	w_6	0
0	0	$C_{7,3}$	$C_{7,4}$	$C_{7,5}$	$C_{7,6}$	w_7

Table 6.1 - Comparison of the closure matrix for $M = 7$, $D = 4$, for the Coatesian model (left) and the generalised one (right). The entries highlighted in green, indicate the binomial trials that will, with probabilities q_5 , q_6 , and q_7 result in ion counts being observed and which are required only for the generalised model.

6.4 Maximum likelihood estimates for the generalised model

As with the Coatesian system, the MCEM algorithm can be used to find the maximum likelihood estimates, and as before, the main difficulty lies in estimating the conditional expectation through rejection sampling.

6.4.1 Rejection sampling for the generalised model

Our target distribution is

$$\begin{aligned}
P(\mathbf{C}, \mathbf{f}, \mathbf{w} | \mathbf{k}, \boldsymbol{\theta}) &= \frac{P(\mathbf{C}, \mathbf{f}, \mathbf{w}, \mathbf{k} | \boldsymbol{\theta})}{P(\mathbf{k} | \boldsymbol{\theta})} \\
&= \frac{P(\mathbf{C}, \mathbf{f}, \mathbf{w} | \boldsymbol{\theta})}{P(\mathbf{k} | \boldsymbol{\theta})} \\
&\propto P(\mathbf{C}, \mathbf{f}, \mathbf{w} | \boldsymbol{\theta}) \\
&= \prod_{i=1}^M \left\{ \binom{V_i}{w_i} \rho_i^{w_i} (1 - \rho_i)^{V_i - w_i} \right\} \prod_{i=1}^{M-D} \left\{ \binom{C_{i+D,i}}{f_{i+D}} q_{i+D}^{f_{i+D}} (1 - q_{i+D})^{C_{i+D,i} - f_{i+D}} \right\} \\
&\quad \prod_{i=2}^M \prod_{j=i-D}^{i-1} \left\{ \binom{C_{i-1,j}}{C_{i,j}} \rho_i^{C_{i,j}} (1 - \rho_i)^{C_{i,j}} \right\}
\end{aligned}$$

We will use the proxy distribution

$$\mathcal{Q}(\mathbf{C}, \mathbf{f} | \mathbf{k}, \boldsymbol{\theta}) = \prod_{i=1}^{M-D} \left\{ \binom{C_{i+D,i}}{f_{i+D}} q_{i+D}^{f_{i+D}} (1 - q_{i+D})^{C_{i+D,i} - f_{i+D}} \right\} \\ \prod_{i=2}^M \prod_{j=i-D}^{i-1} \left\{ \binom{\mathbb{C}_{i-1,j}}{C_{i,j}} \rho_i^{C_{i,j}} (1 - \rho_i)^{\mathbb{C}_{i-1,j} - C_{i,j}} \right\}$$

where the diagonal of $\mathbf{C}$ is fixed to $\mathbf{w} = \mathbf{k} - \mathbf{f}$, regardless of whether this is consistent with the corresponding V_i . If the simulation results in an $f_i > k_i$, then it is rejected as this would otherwise result in negative w_i . As before, the main reason for choosing this proxy distribution is its simple nature and its similarity to the target distribution.

We have

$$\frac{P(\mathbf{C}, \mathbf{f}, \mathbf{w} | \boldsymbol{\theta})}{\mathcal{Q}(\mathbf{C}, \mathbf{f} | \mathbf{k}, \boldsymbol{\theta})} = \prod_{i=1}^M \left\{ \binom{V_i}{w_i} \rho_i^{w_i} (1 - \rho_i)^{V_i - w_i} \right\}$$

and as before we must find an upper bound for this expression. For the generalised system, we know the V_i for the first $D + 1$ rather than D bins, so that in order to find our upper bound, H , we need to maximise the product

$$\prod_{i=D+1}^M \left\{ \binom{V_i}{w_i} \rho_i^{w_i} (1 - \rho_i)^{V_i - w_i} \right\}$$

This task is rather more difficult for the generalised system as we know neither w_i nor V_i exactly. Of course we can fix w_i to the mode of the binomial distribution which is $\text{floor}(p_i V_i)$. V_i may be set to the lowest feasible value for each binomial, which maximises the above product. Again, though it is very likely that a tighter bound might be found, that task will be not be addressed further here.

As for the Coatesian system, the rejection-sampler was validated by comparing the empirical mean of the simulated samples with the theoretical mean of $C_{2,l}$, in this case for the system with $\mathbf{k} = (4 \ 2 \ 3 \ 3)^T$, $\boldsymbol{\rho} = (0.5 \ 0.5 \ 0.5 \ 0.5)^T$, $\mathbf{q} = (0 \ 0 \ 0.1 \ 0.1)^T$, $D = 2$.

Again, though the system is very simple, it includes all of the features of the generalised model. The results are shown on Figure 6.2 below. Since the cumulative mean appears to converge to the theoretical value, the sampling procedure is very likely valid and we can apply the MCEM algorithm essentially as before.

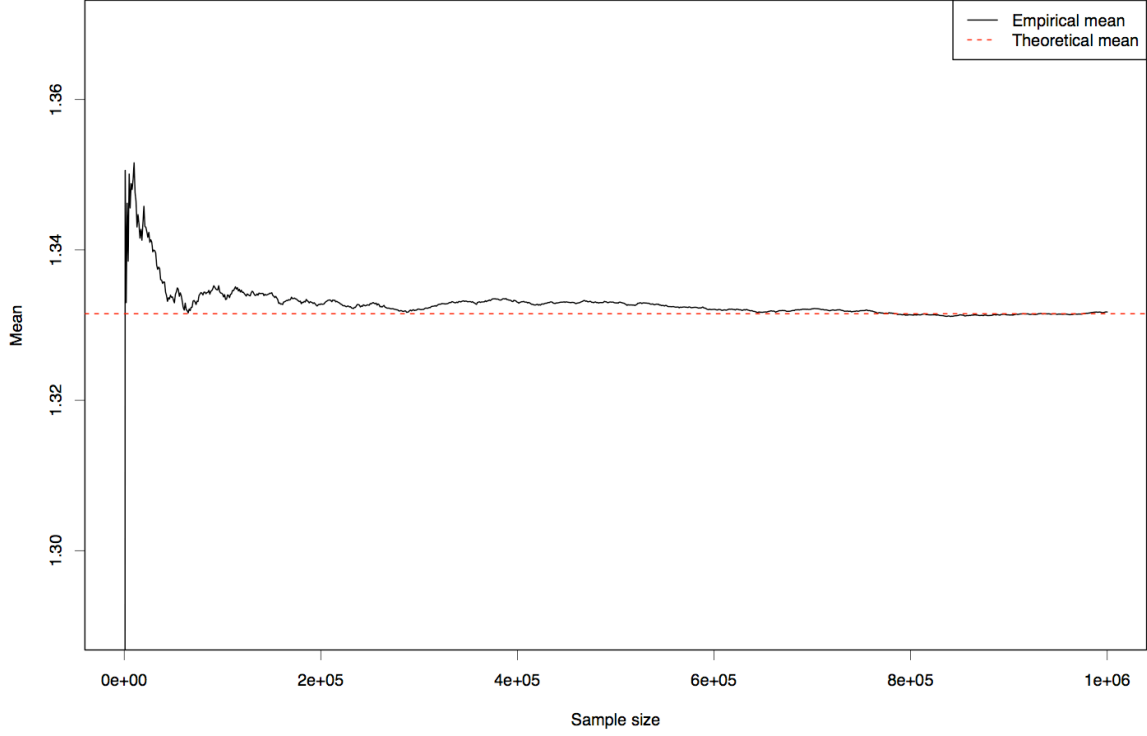


Figure 6.2 - Validation of the rejection-sampling algorithm as described above. Samples are generated from the variable $C_{2,l}$ for the system which has the parameters $\mathbf{k} = (4 \ 2 \ 3 \ 3)^T$, $\mathbf{\rho} = (0.5 \ 0.5 \ 0.5 \ 0.5)^T$, $\mathbf{q} = (0 \ 0 \ 0.1 \ 0.1)^T$, $D = 2$ and the empirical mean is calculated as a function of sample size. The theoretical mean is also indicated.

6.4.2 Practical implementation issues

It has been found that the acceptance rate of the sampler is often higher if $\boldsymbol{\theta}^i$ is close to $\boldsymbol{\theta}$. If the two are very far apart, the acceptance rate may be so low that it takes a long time to generate even a single sample. If this is the case and if Γ is known, it may be sensible to exclude from the likelihood those scans that are close to the mean of the peak and which are therefore more distorted by saturation effects. The version of the algorithm that was implemented would generate up to 100000 samples from the proxy

distribution for each scan, and if none were accepted the scan was excluded from the likelihood for that iteration of the EM algorithm.

The algorithm may also be sped up by initially disregarding the last $M - D$ ticks, so that none of the pulses have time to reopen in the data-set used. We can then obtain an initial and often quite good estimate of θ by applying the basic model from the previous chapter. It is in fact also possible to use the EM algorithm to maximise the likelihood of the basic model. When this is done, the expectation step of the EM algorithm is extremely simple as we have, for all h, i, j :

$$E[C_{h,i,j} | \mathbf{k}, \theta] = \rho_i C_{h,i-1,j}$$

Thus, there is no need to apply the computationally demanding Monte Carlo simulations.

6.5 Applying the generalised model

While the generalised model contains features, such as the closure matrix, which would very likely have to be included in a truly comprehensive model of the TOFMS TDC data, it is not sufficiently detailed that it can extend the ion count range over which we can construct statistical tests for real experimental data. In part this is because the assumption of Gaussian mass peaks breaks down at high counts, but even if the true functional form of the peaks were known, the assumption of constant dead time would severely limit the model's applicability. It nevertheless constitutes an important conceptual improvement over the basic model, and as will now be illustrated with simulated data, it can also provide better parameter estimates than that model when the assumptions on which it relies are met.

Given the already high degree of complexity of the model, a Gaussian shape was used for the simulated chromatographic peaks, rather than a skew-normal or exponentially modified Gaussian. This is not a major restriction since we are not aiming for complete realism with this model and chromatographic peaks can in any case be Gaussian under certain conditions.

For these simulations N_p was set to 1440, the duration of a tick of the TDC clock to 278 picoseconds, the dead time to 18 ticks and a total 21 ticks were included in the analysis, for each of the 30 scans. The parameters of the mass peaks, referred to as μ_x and σ_x below were set to 50 and 1 respectively and the parameters of the chromatographic peak, μ_y and σ_y were set to 100 and 0.5. The mean ion count over the entire peak, I , was set to 10000. As with the simulations carried out in Chapter 5 these parameters are representative of real data.

The θ^i/θ , are plotted on Figure 6.3 and Figure 6.4 below, so that convergence to 1 would indicate convergence to the true values. Also indicated are the maximum likelihood estimates that would be obtained if the true entries of $\mathbf{C}$ were known. The degree of convergence to these estimates is perhaps more telling, as it would be unrealistic to expect any estimator based solely on $\mathbf{k}$ to perform better. The θ^i are plotted for 100 iterations of the deterministic EM-algorithm applied only to the first D data-points, followed by 100 iterations of the MCEM algorithm applied to the full system. Thus the estimates obtained for $i = 100$ indicate the result we obtain by applying the basic model from the previous chapter.

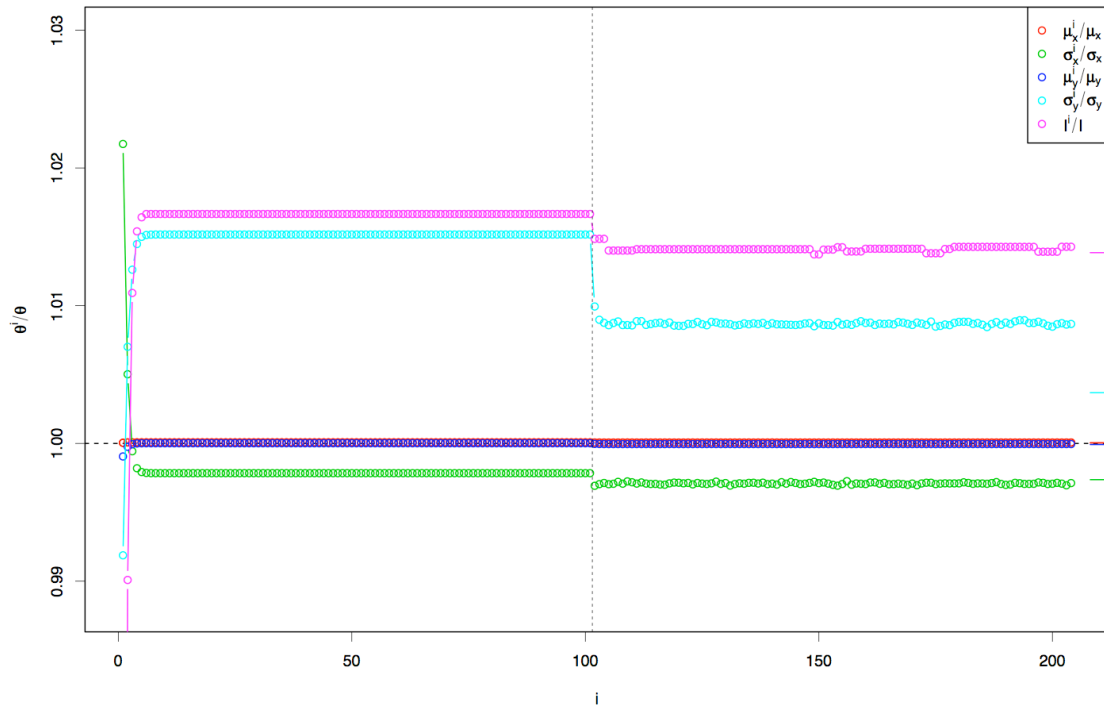


Figure 6.3 – Estimates of the entries of θ relative to their true value for the first 100 iteration of the deterministic EM algorithm (the basic model) followed by 100 iterations of the MCCEM algorithm (the generalised model). Convergence to 1 would indicate convergence to the true θ , but the degree of convergence to the maximum likelihood estimates that would be obtained if the true C was known is arguably more relevant. These estimates are indicated by the short horizontal bars on the right-hand side of the figure.

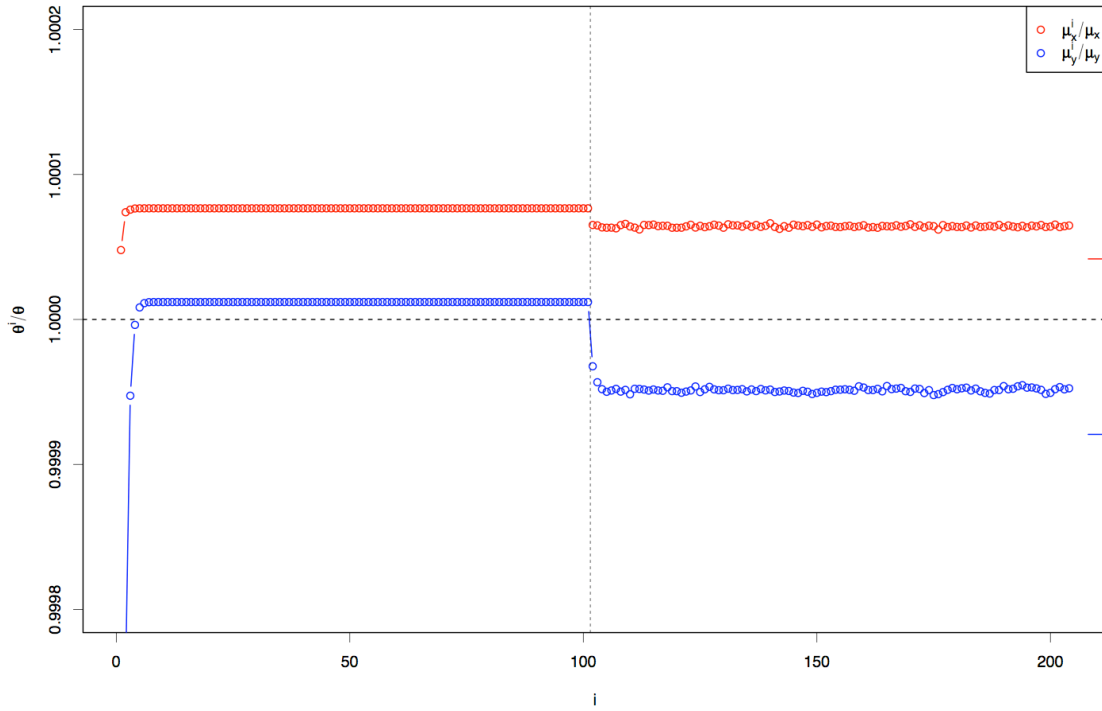


Figure 6.4 – As above, but focusing on the means of the mass and chromatographic peaks. While the estimate of μ_y (the mean of the chromatographic peak) strictly speaking deteriorates when the MCEM algorithm is applied, it moves closer to the estimates that would be obtained if $\mathbf{C}$ were known.

For all parameters we see some improvement in convergence towards $\hat{\theta}^{MLE} | \mathbf{C}$ when the MCEM algorithm is applied. In general the degree of improvement obtained will vary depending on the parameters of the system and is likely to be greatest when the dead time is short relative to the width of the mass peak. In practice the decision of whether or not to apply a method as computationally demanding as the MCEM algorithm would likely depend on the quality of the estimates required and the computing power available.

6.6 Discussion

The generalised model presented in this chapter accounts for the possibility that a pulse that is closed by an incoming ion may reopen over the course of a single mass peak, and it does so in a manner that is more rigorous than the method proposed by

Coates. However, the generalised model is not sufficiently detailed that it can extend the ion count range over which we can address standard analytical problems through the likelihood ratio test, as was proposed in the previous chapter.

The model's primary shortcoming is that for real detector systems the dead time is not constant, but rather can be of somewhat variable duration. Based on communications with engineers from several different mass spectrometry manufacturers, it appears that the distribution of the dead time is not known beyond basic qualitative features. It is not sufficient to model it by means of an arbitrary distribution as was done for mass and chromatographic peaks in the previous chapter, since this distribution might be altered somewhat if a pulse, which is already closed, is struck by an additional ion. We might attempt to construct a model for which the distribution of the dead time depends in some manner on the time elapsed since each of the previous ion arrivals, but it is arguably premature to attempt to model such a high level of detail without a better understanding of the system.

But while it is clear that the generalised model is incomplete, it does establish the basic framework from which further aspects of the system such as the variable dead time can be modelled. For example, it is very likely that the closure matrix would be used in such a model, although the $C_{i,j}$ would have to be extended to the entire lower triangle to accommodate for the possibility of the dead time terminating in other ticks. Similarly the concept of the q_i would be required in much the same way as described above to account for closed pulses being struck by further ions in the same tick in which their dead time terminates. If a complete model were developed it could be used to construct likelihood ratio tests in scenarios where the basic model breaks down. This would be extremely valuable if the applications discussed in section 5.7 were available, though these would of course also require knowledge of the shapes of mass peaks and preferably also of chromatographic peaks.

The very heavy computational demands of the generalised model also pose a severe problem. Given the size of LC-TOFMS data-sets, it is clearly not realistic to apply such a complex model to the entire data-set. However it is conceivable that a less rigorous but faster technique might be applied to identify regions of the data that are

of high interest to the researcher, at which point the more rigorous but computationally demanding algorithms could be used.

This chapter also illustrates how a seemingly minor relaxation in the assumptions required by the basic model necessitates a vastly more mathematically and computationally demanding treatment. That being the case, it is arguable that, if indeed we are to aim for a statistically rigorous analysis of LC-TOFMS data, the detector systems of future mass spectrometers might have to be more consciously designed to accommodate the statistical analysis of the output data. One might even argue that counter to intuition, a detector system with a longer dead time might be desirable if it would broaden the applicability of the basic model, as this would make it easier for the analyst to appeal to statistically meaningful arguments when seeking to draw inferences regarding the data.

7 Conclusion

7.1 Summary

An investigation was undertaken of the detailed workings of LC-TOFMS in order to evaluate the prospects for adopting a rigorous approach to the analysis of the output data. A “rigorous approach” is one wherein, ideally, all the aspects of the data generation process that are relevant to the eventual analysis are understood and accounted for and it should be contrasted with the “heuristic approach”, which provides intuitive but poorly understood rules of thumb to data analysis. A second aim was to develop specific data analytical techniques based on the rigorous approach, with particular emphasis on methods that facilitate the identification of unknown metabolites in complex mixtures. Focus was placed on TOF mass spectrometers employing TDCs as part of their detector systems.

It was found that if the rate of ion arrivals at the detector plate is relatively low, the ion counts could be described as approximately Poissonian in nature, with this approximation being best for low or moderate ion counts. Using a Poisson model, a “coelution test” for identifying parent-fragment pairs based on the similarity of their chromatographic peak shapes was constructed. The same model was used to construct a test of hypothesis to determine whether the theoretical isotopic abundance pattern of a candidate formula could realistically have produced the observed abundance pattern. A method was proposed for increasing the statistical power of this test by pooling together measurements from multiple scans and from any observed adducts, dimers or fragments. However, owing to the overlapping tails of the mass peaks of adjacent isotopologues, the level of statistical rigour attained was not as high as for the coelution test.

A much more general “basic model” was developed, which can, to some extent, account for detector saturation, although it requires some strong simplifying assumptions regarding the duration of the detector dead time relative to the width of mass peaks, which therefore renders it inapplicable to large peaks. Through the likelihood ratio test, methods were proposed for addressing a potentially very wide

range of data analytical problems in a statistically rigorous manner. However the tests require knowledge of the functional forms of the mass and preferably also the chromatographic peaks and these are currently not very well understood.

An attempt was made at formulating a detailed mathematical description of the data when the requirement that the dead time must be longer than the mass peaks was removed. One model was proposed which required the same assumptions as the method by Coates. This was further generalised by removing the rather implausible requirement that ions always arrive in the middle of a tick of the TDC clock. While this generalised model is probably the most detailed description of LC-TOFMS data that has so far been formulated, it is not sufficiently detailed that it can be used to construct valid tests of hypotheses via the likelihood ratio test when applied to real data. It nevertheless establishes a basis from which such a model could be developed.

Thus, it clearly is possible to take a more rigorous approach to the analysis of LC-TOFMS data and given the wide variety of problems that could potentially be quite comprehensively addressed by means of likelihood ratio tests it is a line of research that warrants further study. Of the data analytical techniques that have been developed as part of this thesis, it is arguable that the most practically useful ones are the coelution test and the goodness-of-fit test for the fit of isotope patterns, owing to their rather minimal computational demands.

7.2 Future work

The findings of this thesis suggest several lines of further research. Undoubtedly, one of the most important remaining tasks is to develop more accurate models to describe the mass and chromatographic peak shapes. In addition to improving the quality of dead time corrections this would allow for the application of the likelihood ratio tests discussed in Section 5.7, and could furthermore lead to more accurate estimates through data pooling.

However, even if we knew the exact functional forms of the peak shapes, the heavy tails of larger mass peaks would place substantial constraints on the use of the basic model, owing to the increasing likelihood of closed pulses recovering over the course

of a mass peak. Unless we resign ourselves to addressing this problem by working with heavier isotopologues or with the tails of chromatographic peaks, efforts will have to be made at completing the generalised model, which will require more attentive modelling of the dead time.

But given the findings of chapter 6, it is all but certain that even if an accurate model for the distribution of the data were available, obtaining maximum likelihood estimates would be a computationally very demanding task. This is a problem that would be greatly compounded by the large size of LC-TOFMS data-sets. Thus, substantial efforts would have to be made at reducing the computational requirements of fitting such a model as much as possible. In addition to implementing the optimisation procedure in a faster programming language, this might be done by developing an optimisation algorithm that is more closely tailored to the likelihood function and by applying various heuristic tricks as was done in Section 6.4.2.

It would also be instructive to develop a model to describe the data obtained from mass spectrometers employing ADCs. Dead time would not be an issue for such devices, but the electronic noise would have to be accounted for. In addition, efforts would have to be made to characterise the gain of the electron multiplier through some form of probability distribution, which would have to be linked to the Poisson distribution of the incoming ions. This distribution would in turn have to be linked to the distribution obtained following digitisation by the ADC, which discretises the electronic signal to one of 256 levels, assuming an 8 bit ADC is used. Given these obstacles it may not be possible to obtain a simple expression for the approximate distribution of the acquired data, analogous to the pure Poisson model or even the basic model.

7.3 Broader implications

The critical importance of employing sound statistical arguments when seeking to draw inferences from inexact measurements is decidedly well-established throughout the sciences. It is then perhaps somewhat surprising that more extensive efforts have not previously been made at extending statistical rigour to the analysis of LCTOF-MS data. However it is plainly the case that most of the statistical methods developed for

LC-TOFMS take the heuristic approach, which leaves the details of data generation and pre-processing in the black box. This lack of rigour may stem in part from the highly interdisciplinary requirements of the rigorous approach, bringing together engineering, statistics, and chemistry, as well as the fact that much of the engineering is currently consigned to industry and therefore not very accessible to scientists in academia.

While the potential benefits of adopting the rigorous approach could be substantial it is not yet known whether the required modelling can be properly completed or whether the computational demands can be sufficiently reduced that the resulting methods can be applied in a routine manner. The only way of definitively determining the end utility of this line of research is by pursuing it further, but it is not yet clear whether the necessary resources are likely to get allocated to this task in the short term, as high statistical rigour remains a somewhat esoteric concept in analytical chemistry. It may therefore be constructive to more vigorously communicate to analysts the fact that the data-analytical tools that are currently used are not of a very high statistical standard and that it is perfectly possible that much better alternatives could be developed.

Similarly, it is important to stress to engineers that it is a mistake to regard the statistical analysis of the acquired data as a wholly distinct task to the instrumental design. In fact if the basic objective of the engineer is to design an instrument that can be used by the analyst to draw sound and reliable inferences regarding the sample being analysed, it would seem that particular care should be taken to ensure that the output data are amenable to an informative statistical analysis. While the convention is for the statistician to develop a statistical methodology that can accommodate the data output by an instrument that has been designed independently by the engineer, taking a more integrated approach is not at all unreasonable and could prove to be highly beneficial. Thus, mass spectrometers might be deliberately designed so that the data produced can more easily be described by a probability distribution and so that the maximum likelihood estimators, or rather, the relevant test statistics can more easily be obtained. This applies to mass spectrometers other than time-of-flight and indeed to any analytical instruments that produce inexact measurements.

This integrated approach may require a slight shift in our conception of what constitutes a good mass spectrometer. Currently, heavy emphasis is placed on developing mass spectrometers with improved mass accuracy, the improvement sometimes being quantified through the mean square error of the mass estimates. Similarly, considerable efforts are made at increasing the resolution and dynamic range of the instruments. But we are of course not interested in improving these measures for their own sake – we care about them only to the extent that they help us draw inferences about the sample being analysed. It is therefore important to bear in mind that the fundamental measure by which we must judge the quality of mass spectrometers is the range of inferences that we can draw from the data they produce and the ease with which we can do so. By this measure it is perfectly possible that a mass spectrometer which is poor by the conventional measures of mass accuracy, dynamic range and resolution, but which has well-defined mass peaks and a mathematically well-characterised detector system could be considered superior to a mass spectrometer that performs extremely well by the conventional measures, but which cannot be properly characterised mathematically.

Bibliography

1. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification*. Nat Biotech, 2008. **26**(12): p. 1367-1372.
2. Kohlbacher, O., et al., *TOPP-the OpenMS proteomics pipeline*. Bioinformatics, 2007. **23**(2): p. e191-e197.
3. Katajamaa, M., J. Miettinen, and M. Oresic, *MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data* Bioinformatics, 2006. **22**(5): p. 634.
4. Lommen, A., *MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing* Anal. Chem., 2009. **81**(8): p. 3079-3086.
5. Smith, C.A., et al., *XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification*. Anal. Chem., 2006. **78**: p. 779.
6. *Dead time theory 29508_2*, in *Internal Waters document*.
7. Green, M. and M. Jackson, *Mass spectrometer and methods of mass spectrometry*, E.P. Office, Editor. 2007, Micromass UK Limited: EP 1 768 164 A1.
8. Katajamaa, M. and M. Oresic, *Data processing for mass spectrometry-based metabolomics*. Journal of Chromatography A, 2007. **1158**(1-2): p. 318-328.
9. Brodsky, L., et al., *Evaluation of Peak Picking Quality in LC-MS Metabolomics Data*. Analytical Chemistry, 2010. **82**(22): p. 9177-9187.
10. Theodoridis, G., H.G. Gika, and I.D. Wilson, *LC-MS-based methodology for global metabolite profiling in metabonomics/metabolomics*. TrAC Trends in Analytical Chemistry, 2008. **27**(3): p. 251-260.
11. Kind, T., et al., *A comprehensive urinary metabolomic approach for identifying kidney cancer*. Analytical Biochemistry, 2007. **363**(2): p. 185-195.
12. Lange, E., et al., *Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements*. BMC Bioinformatics, 2008. **9**(1): p. 375.
13. Du, P., et al., *A noise model for mass spectrometry based proteomics*. Bioinformatics, 2008. **24**(8): p. 1070-1077.
14. Chernushevich, I., A. Loboda, and B. Thomson, *An introduction to quadrupole-time-of-flight mass spectrometry*. Journal of Mass Spectrometry, 2001. **36**(8): p. 849-865.
15. Cotter, R., *Time-of-Flight Mass Spectrometry: Instrumentation and Applications in Biological Research*. 1997: American Chemical Society.
16. Cotter, R.J., *Time-of-flight mass spectrometry*. Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. 2004: John Wiley & Sons, Ltd.
17. Ens, W., K.G. Standing, and A.L. Burlingame, *Hybrid Quadrupole/Time-of-Flight Mass Spectrometers for Analysis of Biomolecules*, in *Methods in Enzymology*. 2005, Academic Press. p. 49-78.
18. Guilhaus, M., *Special feature: Tutorial. Principles and instrumentation in time-of-flight mass spectrometry. Physical and instrumental concepts*. Journal of Mass Spectrometry, 1995. **30**(11): p. 1519-1532.

19. Guilhaus, M., D. Selby, and V. Mlynski, *Orthogonal acceleration time-of-flight mass spectrometry*. Mass Spectrometry Reviews, 2000. **19**(2): p. 65-107.
20. Wiley, W.C. and I.H. McLaren, *Time-of-Flight Mass Spectrometer with Improved Resolution*. Review of Scientific Instruments, 1955. **26**(12): p. 1150-1157.
21. Blom, K., *Estimating the Precision of Exact Mass Measurements on an Orthogonal Time-of-Flight Mass Spectrometer*. Analytical Chemistry, 2001. **73**(3): p. 715-719.
22. Colombo, M., F.R. Sirtori, and V. Rizzo, *A fully automated method for accurate mass determination using high-performance liquid chromatography with a quadrupole/orthogonal acceleration time-of-flight mass spectrometer*. Rapid Communications in Mass Spectrometry, 2004. **18**(4): p. 511-517.
23. Wolf, J.-C., T.R. Fuentes, and J. Taylor, *Investigations into the accuracy and precision obtainable on accurate mass measurements on a quadrupole orthogonal acceleration time-of-flight mass spectrometer using liquid chromatography as sample introduction*. Rapid Communications in Mass Spectrometry, 2003. **17**(11): p. 1216-1219.
24. Laures, A., et al., *Investigation into the factors affecting accuracy of mass measurements on a time-of-flight mass spectrometer using Design of Experiment*. Rapid Communications in Mass Spectrometry, 2007. **21**(4): p. 529-535.
25. Anderle, M., et al., *Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum*. Bioinformatics, 2004. **20**(18): p. 3575-3582.
26. Lederberg, L., *How DENDRAL was conceived and born*. Proceedings of ACM conference on History of medical informatics, 1987: p. 5-19.
27. Djerassi, C., et al., *The DENDRAL project: computational aids to natural products structure elucidation*. Pure and Applied Chemistry, 1982. **54**(12): p. 2425-2442.
28. Sasaki, S. and Y. Kudo, *Structure Elucidation System Using Structural Information from Multisources: CHEMICS*. Journal of Chemical Information and Computer Sciences, 1985. **25**: p. 252-257.
29. Shelley, C. and M. Munk, *CASE: A computer model of the structure elucidation process*. Analytica Chimica Acta, 1981. **133**: p. 507-516.
30. Cadisch, M., et al., *SpecTool: A hypermedia toolkit for structure elucidation*. Journal of Chemical Information and Computer Sciences, 1992. **32**: p. 286-290.
31. Heuerding, S. and T. Clerc, *Simple Tools for the Computer-Aided Interpretation of Mass Spectra*. Chemometrics and intelligent laboratory systems, 1993. **20**: p. 57-69.
32. Seebass, B. and E. Pretsch, *Automated Compatibility Tests of the Molecular Formulas or Structures of Organic Compounds with Their Mass Spectra*. Journal of Chemical Information and Computer Sciences, 1999. **39**(4): p. 713-717.
33. Kind, T. and O. Fiehn, *Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry*. BMC Bioinformatics, 2007. **8**.
34. Griffiths, W., et al., *Electrospray and tandem mass spectrometry in biochemistry*. Journal of Biochemistry, 2001. **355**: p. 545-561.

35. Wu, Q., *Multistage Accurate Mass Spectrometry: A "Basket in a Basket" Approach for Structure Elucidation and Its Application to a Compound from Combinatorial Synthesis*. Analytical Chemistry, 1998. **70**: p. 865-872.
36. O'Connor, P., D.P. Little, and F.W. McLafferty, *Isotopic Assignment in Large-Molecule Mass Spectra by Fragmentation of a Selected Isotopic Peak*. Analytical Chemistry, 1996. **68**: p. 542-545.
37. Rockwood, A., M.M. Kushnir, and G.J. Nelson, *Dissociation of individual isotopic peaks: predicting isotopic distributions of product ions in MS_n*. Journal of the American Society for Mass Spectrometry, 2003. **14**(4): p. 311-322.
38. Plumb, R., et al., *UPLC/MS^E; a new approach for generating molecular fragment information for biomarker structure elucidation*. Rapid Commun. Mass Spectrom., 2006. **20**(13): p. 1989.
39. Ipsen, A., et al., *A Statistically Rigorous Test for the Identification of Parent-Fragment Pairs in LC-MS Datasets*. Analytical Chemistry, 2010. **82**(5): p. 1766-1778.
40. Kind, T. and O. Fiehn, *Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm*. BMC Bioinformatics, 2006. **7**.
41. Yervey, J.A., *A general approach to calculating isotopic distributions for mass spectrometry*. International Journal of Mass Spectrometry and Ion Physics, 1983. **52**(2-3): p. 337-349.
42. Roussis, S. and R. Proulx, *Reduction of Chemical Formulas from the Isotopic Peak Distributions of High-Resolution Mass Spectra*. Analytical Chemistry, 2003. **75**(6): p. 1470-1482.
43. Evans, J. and N. Jurinski, *Program ELAL: An Interactive Minicomputer Based Elemental Analysis of Low and Medium Resolution Mass Spectra*. Analytical Chemistry, 1975. **47**(6): p. 961-963.
44. Do Lago, C. and C. Kascheres, *New Method of Isotope Pattern Analysis*. Journal of Computational Chemistry, 1991. **15**: p. 149-155.
45. Katwijk, J., *Improved evaluation of high-resolution mass spectra through the investigation of peak shapes and isotopic contributions*. International Journal of Mass Spectrometry and Ion Physics, 1981. **39**: p. 287-310.
46. Roussis, S., *Automated Tandem Mass Spectrometry by Orthogonal Acceleration TOF Data Acquisition and Simultaneous Magnet Scanning for the Characterization of Petroleum Mixtures*. Analytical Chemistry, 2001. **73**: p. 3611-3623.
47. Ipsen, A., E.J. Want, and T.M.D. Ebbels, *Construction of Confidence Regions for Isotopic Abundance Patterns in LC/MS Data Sets for Rigorous Determination of Molecular Formulas*. Analytical Chemistry, 2010. **82**(17): p. 7319-7328.
48. Green, M., J. Wildgoose, and M. Gorenstein, *Mass Spectrometer*. 2006.
49. Hoyes, J. and J. Cottrel, *Method and apparatus for the correction of mass errors in time-of-flight mass spectrometry*. 2002: International Patent.
50. Esposito, F., N. Spinelli, and R. Velotta, *Dead time correction of time distribution measurements*. Review of Scientific Instruments, 1991. **62**(11): p. 2822-2827.
51. Coates, P., *The correction for photon 'pile-up' in the measurement of radiative lifetimes*. Journal of Physics E, 1968. **1**: p. 878-879.

52. Coates, P., *Analytical corrections for dead time effects in the measurements of time-interval distributions*. Review of Scientific Instruments, 1991. **63**(3): p. 2084-2088.
53. Donohue, D. and R. Stern, *Correction of Single Photon or Particle Timing Measurements for Multiparticle Events*. Review of Scientific Instruments, 1972. **43**(5): p. 791-796.
54. Schiff, L., *Statistical Analysis of Counter Data*. Physical Review, 1936. **50**: p. 88-96.
55. Alpert, A.J., *Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds*. Journal of Chromatography A, 1990. **499**: p. 177-196.
56. Ladislav Wiza, J., *Microchannel plate detectors*. Nuclear Instruments and Methods, 1979. **162**(1-3): p. 587-601.
57. Bateman, R.H., J. Brown, and D.J. Kenny, *Ion Detector*. 2004, Micromass UK Limited: US 2004/0173741 A1.
58. Mullock, S., *Personal Communication*. 2010.
59. Bateman, R.H., et al., *Time Of Flight Mass Spectrometer And Dual Gain Detector Therefor*, E.P. Office, Editor. 2004: EP 0 970 504 B1.
60. Bateman, R.H., J. Brown, and M. Green, *Ion Detector*. 2004, Micromass UK Limited: US 2004/0227070 A1.
61. Green, M., *Mass spectrometer*. 2004, Micromass UK Limited: US 2005/0061968 A1.
62. Bateman, R.H., et al., *Mass Spectrometer*. 2006, Micromass Limited: PCT/GB2006/001996.
63. *Waters Micromass LCT Premier Operator's Guide, 71500067902 / Revision B*.
64. Bateman, R.H., M. Green, and M. Jackson, *Mass spectrometer and method of mass spectrometry* 2006, Micromass Limited 7038197.
65. Hoffmann, E. and V. Stroobant, *Mass Spectrometry: Principles and Applications*. Third Edition. ed. 2007: Wiley-Interscience.
66. Cotter, R.J., *Time-of-flight mass spectrometry for the structural analysis of biological molecules*. Analytical Chemistry, 1992. **64**(21): p. 1027A-1039A.
67. *Waters Micromass Q-ToF Premier, Operator's Guide, 71500089402 / Revision B*.
68. *Micromass Q-ToF micro, Operator's Guide, 71500058502 / Revision A*.
69. Geromanos, S.J., et al., *System and Method for Grouping Precursor and Fragment Ions Using Selected Ion Chromatograms*. 2008: US 2008/0272292.
70. Chakraborty, A., S.J. Berger, and J.C. Gebler, *Use of an integrated MS – multiplexed MS/MS data acquisition strategy for high-coverage peptide mapping studies*. Rapid Communications in Mass Spectrometry, 2007. **21**(5): p. 730-744.
71. Tautenhahn, R., C. Bottcher, and S. Neumann, *Lecture Notes in Computer Science: Bioinformatics Research and Development*. 2007. 371.
72. Luedemann, A., et al., *TagFinder for the quantitative analysis of gas chromatography--mass spectrometry (GC-MS)-based metabolite profiling experiments*. Bioinformatics, 2008. **24**(5): p. 732.
73. Przyborowski, J. and H. Wilenski, *Homogeneity of results in testing samples from poisson series*. Biometrika, 1940. **31**(3-4): p. 313-323.
74. Wackerly, D.D., W. Mendenhall, and R.L. Scheaffer, *Mathematical Statistics with Applications*. 2002.

75. Larntz, K., *Small-Sample Comparisons of Exact Levels for Chi-Squared Goodness-of-Fit Statistics*. Journal of the American Statistical Association, 1978. **73**(362): p. 253-263.
76. Böcker, S., et al., *SIRIUS: decomposing isotope patterns for metabolite identification*. Bioinformatics, 2009. **25**(2): p. 218.
77. Tong, H., et al., *Automated data massaging, interpretation, and e-mailing modules for high throughput open access mass spectrometry*. J. Am. Soc. Mass Spectrom., 1999. **10**: p. 1174.
78. Zhang, J.F., et al., *Predicting molecular formulas of fragment ions with isotope patterns in tandem mass spectra*. IEEE/ACM Trans. Comput. Biol. Bioinf., 2005. **2**(3): p. 217.
79. May, W.L. and W.D. Johnson, *Properties of simultaneous confidence intervals for multinomial proportions*. Commun. Stat. Simul. Comput., 1997. **26**(2): p. 495.
80. Bohlke, J.K., et al., *Isotopic Compositions of the Elements*. J. Phys. Chem. Ref. Data, 2005. **34**(1): p. 57.
81. Gannes, L.Z., C.M. del Rio, and P. Koch, *Natural Abundance Variations in Stable Isotopes and their Potential Uses in Animal Physiological Ecology*. Comp. Biochem. Physiol., 1998. **119A**(3): p. 725.
82. Coombes, K.R., et al., *Understanding the Characteristics of Mass Spectrometry Data Through the Use of Simulation*. Cancer Informatics, 2005. **1**.
83. Schulz-Trieglaff, O., et al., *LC-MSsim - a simulation software for liquid chromatography mass spectrometry data*. BMC Bioinformatics, 2008. **9**(1): p. 423.
84. Kelbert, M., I. Sazonov, and A.G. Wright, *Exact expression for the variance of the photon emission process in scintillation counters*. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 2006. **564**(1): p. 185-189.
85. Guilhaus, M., *Essential elements of time-of-flight mass spectrometry in combination with the inductively coupled plasma ion source*. Spectrochimica Acta Part B, 2000. **55**: p. 1511-1525.
86. Brenton, A.G., et al., *Improvement of the duty cycle of an orthogonal acceleration time-of-flight mass spectrometer using ion gates*. Rapid Communications in Mass Spectrometry, 2007. **21**(18): p. 3093-3102.
87. Feng, R., Y. Konishi, and A.W. Bell, *High accuracy molecular weight determination and variation characterization of proteins up to 80 ku by ionspray mass spectrometry*. Journal of the American Society for Mass Spectrometry, 1991. **2**(5): p. 387-401.
88. Rather, O., *High resolution detection for time-of-flight mass*. 2003, Bruker Daltonik GMBH: United States.
89. Strittmatter, E.F., N. Rodriguez, and R.D. Smith, *High Mass Measurement Accuracy Determination for Proteomics Using Multivariate Regression Fitting: Application to Electrospray Ionization Time-Of-Flight Mass Spectrometry*. Analytical Chemistry, 2003. **75**(3): p. 460-468.
90. Li, J., *Comparison of the capability of peak functions in describing real chromatographic peaks*. Journal of Chromatography A, 2002. **952**(1-2): p. 63-70.

91. Li, J., *Development and Evaluation of Flexible Empirical Peak Functions for Processing Chromatographic Peaks*. Analytical Chemistry, 1997. **69**(21): p. 4452-4462.
92. Casella, G. and R. Berger, *Statistical Inference*. 2nd ed. 2002: Duxbury.
93. Fletcher, R., *Practical Methods of Optimization*. 2nd ed. 1987: John Wiley.
94. Geyer, C.J., *Trust*. 2009, <http://www.stat.umn.edu/geyer/trust/>.
95. Team, R.D.C., *R: A Language and Environment for Statistical Computing*. 2010.
96. Azzalini, A., *A Class of Distributions Which Includes the Normal Ones*. Scandinavian Journal of Statistics, 1985. **12**(2): p. 171-178.
97. Lan, K. and J.W. Jorgenson, *A hybrid of exponential and gaussian functions as a simple model of asymmetric chromatographic peaks*. Journal of Chromatography A, 2001. **915**(1-2): p. 1-13.
98. Shi, S.D.H., C.L. Hendrickson, and A.G. Marshall, *Counting individual sulfur atoms in a protein by ultrahighresolution Fourier transform ion cyclotron resonance mass spectrometry: Experimental resolution of isotopic fine structure in proteins*. Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(20): p. 11532-11537.
99. *In-Depth Qualitative Analysis of Complex Proteomic Samples Using High Quality MS/MS at Fast Acquisition Rates*. 2010, AB Sciex.
100. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society. Series B (Methodological), 1977. **39**(1): p. 1-38.
101. Wei, G.C.G. and M.A. Tanner, *A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms*. Journal of the American Statistical Association, 1990. **85**: p. 699-704.
102. Chib, S. and E. Greenberg, *Understanding the Metropolis-Hastings Algorithm*. The American Statistician, 1995. **49**(4): p. 327-335.
103. MacKay, D.J.C., *Information Theory, Inference, and Learning Algorithms*. 2003: Cambridge University Press.

Appendix A

Taking the logarithm to work with the log-likelihood:

$$\begin{aligned}
 l = \log(L) = & \sum_{i=1}^M \left\{ N_p - \sum_{j=i-D+1}^{i-1} k_j + \sum_{j=1}^{i-D} \left\{ \sum_{k=j-D+1}^{j-1} C_{j,k} - \sum_{k=j+1}^{j+D-1} C_{k,j} \right\} \right. \\
 & \left. \log(h) - \sum_{h=1}^{k_j} \log(h) + k_i \log(\rho_i) + \left(N_p - \sum_{j=i-D+1}^i k_j + \sum_{j=1}^{i-D} \left\{ \sum_{k=j-D+1}^{j-1} C_{j,k} - \sum_{k=j+1}^{j+D-1} C_{k,j} \right\} \right) \log(1-\rho_i) \right\} \\
 & + \sum_{i=2}^M \sum_{j=i-D+1}^{i-1} \left\{ \begin{aligned} & k_j + \sum_{k=j-D+1}^{j-1} C_{j,k} - \sum_{k=j+1}^{j+D-1} C_{k,j} \\ & \log(h) - \sum_{h=1}^{C_{i,j}} \log(h) + C_{i,j} \log(\rho_i) + \left(k_j + \sum_{k=j-D+1}^{j-1} C_{j,k} - \sum_{k=j+1}^i C_{k,j} \right) \log(1-\rho_i) \end{aligned} \right\}
 \end{aligned}$$

We can then maximise for ρ_α to obtain the desired result:

$$\begin{aligned}
 \frac{\partial l}{\partial \rho_\alpha} &= 0 \\
 \Rightarrow 0 &= \rho_\alpha^{-1} k_\alpha + (\rho_\alpha - 1)^{-1} \left(N_p - \sum_{j=\alpha-D+1}^\alpha k_j + \sum_{j=1}^{\alpha-D} \left\{ \sum_{k=j-D+1}^{j-1} C_{j,k} - \sum_{k=j+1}^{j+D-1} C_{k,j} \right\} \right) + \rho_\alpha^{-1} \sum_{j=\alpha-D+1}^{\alpha-1} C_{\alpha,j} + (\rho_\alpha - 1)^{-1} \sum_{j=\alpha-D+1}^{\alpha-1} \left\{ k_j + \sum_{k=j-D+1}^{j-1} C_{j,k} - \sum_{k=j+1}^\alpha C_{k,j} \right\} \\
 \Rightarrow 0 &= (\rho_\alpha - 1) k_\alpha + \rho_\alpha \left(N_p - \sum_{j=\alpha-D+1}^\alpha k_j + \sum_{j=1}^{\alpha-D} \left\{ \sum_{k=j-D+1}^{j-1} C_{j,k} - \sum_{k=j+1}^{j+D-1} C_{k,j} \right\} \right) + (\rho_\alpha - 1) \sum_{j=\alpha-D+1}^{\alpha-1} C_{\alpha,j} + \rho_\alpha \sum_{j=\alpha-D+1}^{\alpha-1} \left\{ k_j + \sum_{k=j-D+1}^{j-1} C_{j,k} - \sum_{k=j+1}^\alpha C_{k,j} \right\} \\
 \Rightarrow 0 &= \rho_\alpha \left(N_p + \sum_{j=1}^{\alpha-D} \left\{ \sum_{k=j-D+1}^{j-1} C_{j,k} - \sum_{k=j+1}^{j+D-1} C_{k,j} \right\} + \sum_{j=\alpha-D+1}^{\alpha-1} C_{\alpha,j} + \sum_{j=\alpha-D+1}^{\alpha-1} \left\{ \sum_{k=j-D+1}^{j-1} C_{j,k} - \sum_{k=j+1}^\alpha C_{k,j} \right\} \right) - k_\alpha - \sum_{j=\alpha-D+1}^{\alpha-1} C_{\alpha,j} \\
 \Rightarrow 0 &= \rho_\alpha \left(N_p - \sum_{j=\alpha-D+1}^{\alpha-1} \left\{ \sum_{k=j-D+1}^{\alpha-D} C_{j,k} \right\} + \sum_{j=\alpha-D+1}^{\alpha-1} C_{\alpha,j} + \sum_{j=\alpha-D+1}^{\alpha-1} \left\{ \sum_{k=j-D+1}^{\alpha-D} C_{j,k} \right\} - \sum_{j=\alpha-D+1}^{\alpha-1} C_{\alpha,j} \right) - k_\alpha - \sum_{j=\alpha-D+1}^{\alpha-1} C_{\alpha,j} \\
 \Rightarrow \rho_\alpha &= \frac{k_\alpha + \sum_{j=\alpha-D+1}^{\alpha-1} C_{\alpha,j}}{N_p}
 \end{aligned}$$

Appendix B

We wish to calculate the expression

$$q_{i+D} = P \left(\begin{array}{l} \text{the time from the last ion arrival in } [t_{i-1}, t_i] \text{ to the first in } [t_{i+D-1}, t_{i+D}] \text{ is greater than } D \\ \text{at least one ion arrival in both } [t_{i-1}, t_i] \text{ and } [t_{i+D-1}, t_{i+D}] \end{array} \right)$$

We start by defining two probability density functions, Ψ_1 and Ψ_2 , that relate to the ion arrivals in the two ticks involved.

$$\begin{aligned} \Psi_1(\tau_\alpha) &= P(\text{last ion arrival time is at } t_{i-1} + \tau_\alpha, \text{ conditional on at least one ion arrival in } [t_{i-1}, t_i]) \\ &= \frac{P(\text{last ion arrival time is } t_{i-1} + \tau_\alpha \cap \text{at least one ion arrival in } [t_{i-1}, t_i])}{P(\text{at least one ion arrival in } [t_{i-1}, t_i])} \\ &= \frac{\sum_{h=1}^{\infty} \{P(\text{last of } h \text{ ion arrivals is } t_{i-1} + \tau_\alpha) P(h \text{ ion arrivals in } [t_{i-1}, t_i])\}}{1 - P(\text{no ion arrivals in } [t_{i-1}, t_i])} \\ &= \frac{\sum_{h=1}^{\infty} \{h P(\text{one ion arrival at } t_{i-1} + \tau_\alpha) P(h-1 \text{ ion arrivals in } [t_{i-1}, t_{i-1} + \tau_\alpha]) P(h \text{ ion arrivals in } [t_{i-1}, t_i])\}}{1 - P(\text{no ion arrivals in } [t_{i-1}, t_i])} \\ &= \frac{\sum_{h=1}^{\infty} \left\{ h \left(\frac{\lambda(t_{i-1} + \tau_\alpha)}{\int_{t_{i-1}}^{t_i} \lambda(t) dt} \right) \left(\frac{\int_{t_{i-1}}^{t_{i-1} + \tau_\alpha} \lambda(t) dt}{\int_{t_{i-1}}^{t_i} \lambda(t) dt} \right)^{h-1} \left(\frac{e^{-\int_{t_{i-1}}^{t_i} \lambda(t) dt} \left(\int_{t_{i-1}}^{t_i} \lambda(t) dt \right)^h}{h!} \right) \right\}}{1 - e^{-\int_{t_{i-1}}^{t_i} \lambda(t) dt}} \\ &= \frac{\lambda(t_{i-1} + \tau_\alpha) e^{-\int_{t_{i-1}}^{t_i} \lambda(t) dt}}{1 - e^{-\int_{t_{i-1}}^{t_i} \lambda(t) dt}} \sum_{h=1}^{\infty} \left\{ \left(\frac{\int_{t_{i-1}}^{t_{i-1} + \tau_\alpha} \lambda(t) dt}{(h-1)!} \right)^{h-1} \right\} \\ &= \frac{\lambda(t_{i-1} + \tau_\alpha) e^{-\int_{t_{i-1}}^{t_i} \lambda(t) dt}}{1 - e^{-\int_{t_{i-1}}^{t_i} \lambda(t) dt}} e^{\int_{t_{i-1}}^{t_{i-1} + \tau_\alpha} \lambda(t) dt} \\ &= \frac{\lambda(t_{i-1} + \tau_\alpha) e^{-\int_{t_{i-1} + \tau_\alpha}^{t_i} \lambda(t) dt}}{1 - e^{-\int_{t_{i-1}}^{t_i} \lambda(t) dt}} \end{aligned}$$

$$\begin{aligned}
\Psi_2(\tau_\beta) &= P(\text{first ion arrival is at } t_{i+D-1} + \tau_\beta, \text{ conditional on at least one ion arrival in } [t_{i+D-1}, t_{i+D}]) \\
&= \frac{P(\text{first ion arrival is at } t_{i+D-1} + \tau_\beta \cap \text{at least one ion arrival in } [t_{i+D-1}, t_{i+D}])}{P(\text{at least one ion arrival in } [t_{i+D-1}, t_{i+D}])} \\
&= \frac{\sum_{h=1}^{\infty} \left\{ P(\text{first of } h \text{ ion arrivals is at } t_{i+D-1} + \tau_\beta) P(h \text{ ion arrivals in } [t_{i+D-1}, t_{i+D}]) \right\}}{1 - P(\text{no ion arrivals in } [t_{i+D-1}, t_{i+D}])} \\
&= \frac{\sum_{h=1}^{\infty} \left\{ h P(\text{one ion arrival at } t_{i+D-1} + \tau_\beta) P(h-1 \text{ ion arrivals in } [t_{i+D-1} + \tau_\beta, t_{i+D}]) P(h \text{ ion arrivals in } [t_{i+D-1}, t_{i+D}]) \right\}}{1 - P(\text{no ion arrivals in } [t_{i+D-1}, t_{i+D}])} \\
&= \frac{\sum_{h=1}^{\infty} \left\{ h \left(\frac{\lambda(t_{i+D-1} + \tau_\beta)}{\int_{t_{i+D-1}}^{t_{i+D}} \lambda(t) dt} \right) \left(\frac{\int_{t_{i+D-1} + \tau_\beta}^{t_{i+D}} \lambda(t) dt}{\int_{t_{i+D-1}}^{t_{i+D}} \lambda(t) dt} \right)^{h-1} \left(\frac{e^{-\int_{t_{i+D-1}}^{t_{i+D}} \lambda(t) dt} \left(\int_{t_{i+D-1}}^{t_{i+D}} \lambda(t) dt \right)^h}{h!} \right) \right\}}{1 - e^{-\int_{t_{i+D-1}}^{t_{i+D}} \lambda(t) dt}} \\
&= \frac{\lambda(t_{i+D-1} + \tau_\beta) e^{-\int_{t_{i+D-1}}^{t_{i+D}} \lambda(t) dt}}{1 - e^{-\int_{t_{i+D-1}}^{t_{i+D}} \lambda(t) dt}} \sum_{h=1}^{\infty} \left\{ \left(\frac{\left(\int_{t_{i+D-1} + \tau_\beta}^{t_{i+D}} \lambda(t) dt \right)^{h-1}}{(h-1)!} \right) \right\} \\
&= \frac{\lambda(t_{i+D-1} + \tau_\beta) e^{-\int_{t_{i+D-1}}^{t_{i+D}} \lambda(t) dt}}{1 - e^{-\int_{t_{i+D-1}}^{t_{i+D}} \lambda(t) dt}} e^{\int_{t_{i+D-1} + \tau_\beta}^{t_{i+D}} \lambda(t) dt} \\
&= \frac{\lambda(t_{i+D-1} + \tau_\beta) e^{-\int_{t_{i+D-1}}^{t_{i+D-1} + \tau_\beta} \lambda(t) dt}}{1 - e^{-\int_{t_{i+D-1}}^{t_{i+D}} \lambda(t) dt}}
\end{aligned}$$

Note that

$$\begin{aligned}
&P\left(\begin{array}{l} \text{the last ion arrival in } [t_{i-1}, t_i] \text{ is } t_{i-1} + \tau_\alpha \mid \text{at least one ion arrival in } [t_{i-1}, t_i] \cap \\ \text{the first in } [t_{i+D-1}, t_{i+D}] \text{ is after } t_{i+D-1} + \tau_\alpha \mid \text{at least one ion arrival in } [t_{i+D-1}, t_{i+D}] \end{array}\right) \\
&= \Psi_1(\tau_\alpha) \left(\int_{\tau_\alpha}^{\varepsilon} \Psi_2(\tau_\beta) d\tau_\beta \right)
\end{aligned}$$

so that the desired expression can be written

$$\begin{aligned}
& P \left(\begin{array}{l} \text{the time from the last ion arrival in } [t_{i-1}, t_i] \text{ to the first in } [t_{i+D-1}, t_{i+D}] \text{ is greater than } D \\ \text{and at least one ion arrival in both } [t_{i-1}, t_i] \text{ and } [t_{i+D-1}, t_{i+D}] \end{array} \right) \\
&= \int_0^\varepsilon \Psi_1(\tau_\alpha) \left(\int_{\tau_\alpha}^\varepsilon \Psi_2(\tau_\beta) d\tau_\beta \right) d\tau_\alpha \\
&= \int_0^\varepsilon \left\{ \frac{\lambda(t_{i-1} + \tau_\alpha) e^{-\int_{t_{i-1}+\tau_\alpha}^{t_i} \lambda(t) dt}}{1 - e^{-\int_{t_{i-1}}^{t_i} \lambda(t) dt}} \left(\int_{\tau_\alpha}^\varepsilon \frac{\lambda(t_{i+D-1} + \tau_\beta) e^{-\int_{t_{i+D-1}}^{t_{i+D-1}+\tau_\beta} \lambda(t) dt}}{1 - e^{-\int_{t_{i+D-1}}^{t_{i+D}} \lambda(t) dt}} d\tau_\beta \right) \right\} d\tau_\alpha \\
&= \int_0^\varepsilon \left\{ \frac{\lambda(t_{i-1} + \tau_\alpha) e^{-\int_{t_{i-1}+\tau_\alpha}^{t_i} \lambda(t) dt}}{1 - e^{-\int_{t_{i-1}}^{t_i} \lambda(t) dt}} \left(\int_{\tau_\alpha}^\varepsilon \frac{\lambda(t_{i+D-1} + \tau_\beta) e^{-\int_{t_{i+D-1}}^{t_{i+D-1}+\tau_\beta} \lambda(t) dt}}{1 - e^{-\int_{t_{i+D-1}}^{t_{i+D}} \lambda(t) dt}} d\tau_\beta \right) \right\} d\tau_\alpha \\
&= \frac{1}{\left(e^{\int_{t_{i-1}}^{t_i} \lambda(t) dt} - 1 \right) \left(e^{-\int_{t_{i+D-1}}^{t_{i+D}} \lambda(t) dt} - 1 \right)} \int_0^\varepsilon \lambda(t_{i-1} + \tau_\alpha) e^{\int_{t_{i-1}}^{t_{i-1}+\tau_\alpha} \lambda(t) dt} \left[e^{-\int_{t_{i+D-1}}^{t_{i+D}} \lambda(t) dt} - e^{-\int_{t_{i+D-1}}^{t_{i+D-1}+\tau_\alpha} \lambda(t) dt} \right] d\tau_\alpha
\end{aligned}$$