# DOCTORAATSPROEFSCHRIFT

2010 | School voor Informatietechnologie
Kennistechnologie, Informatica, Wiskunde, ICT

# Statistical Modeling for the Analysis of High-resolution Mass Spectrometry Data

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, Wiskunde, te verdedigen door:

Qi ZHU

Promotor: prof. dr. Tomasz Burzykowski

universiteit hasselt

I-BioStat

KATHOLIEKE UNIVERSITEIT LEUVEN

D/2010/2451/31

# Acknowledgements

Five years ago, Bruce and I left our families and everything we possessed in Shanghai (our hometown), and came to Belgium for the studying opportunities. Ever since we stepped onto this piece of European land, we have received countless helps and support, for which I owe deeply my gratitude.

I remember getting off the airplane, nearly a dozen of people, including a German couple over seventy years old, helped us carrying our 150 kilograms of luggage from airport to the train station, and from one platform to another.

Not knowing what my destiny would be, I started my first year of studying at CenStat. With no previous statistical background, I was impressed by the passion and devoted spirits of the CenStat teaching group of Master programs, and my classmates. I am thus grateful for the inspiring studying, teaching and research atmosphere, which charmed me when taking a first glimpse at the world of statistics. I would especially thank Nöel (Veraverbeke), my Master thesis supervisor, for guiding me through the concluding part of the program.

My first year of the Master program finally concluded with a happy end. In the mean time, I applied for a PhD position available at the department. Being a freshman in the area of statistics and knowing the competition would be severe, I didn't even expect an interview. As soon as I mailed my application, I forgot about it. The time when Tomasz informed me for an interview, I told him I was in Greece. I felt it a pity to miss the interview. Immediately after I returned from my holiday, I got a strong inspiration that I should write Tomasz an email, thanking him for granting me an interview and expressing my interest in future vacancies.

Miraculously, around two months later, I received a phone call from Tomasz. He told me that the position was open again and if I would be available for an interview. I can still recall my trembling voice when answering the phone, due to excitement. To my mentor, Tomasz, I owe a deep sense of gratitude. Your presence made me feel as though holding a compass when walking through a strange forest of the scientific world – a compass that could always guide me through when I felt lost in a mist of darkness. More particularly, the creativity I had learnt from you when discussing my research will be a gift for me.

During the four-year work, I received help from other people at CenStat as well. I am especially grateful for the Bayesian knowledge Ziv has shared with me, despite his extremely busy schedule. The group of Bioinformatics showed me also a cohesion of helping each other at work. Dirk, Kasim, Dan, Ivy, Suzy, to name but a few, have patiently provided me the resources and knowledge, without which the work would never have been done. I would also thank An, my office-mate, for creating a loving working atmosphere by sharing this small office together with me in the four years. I am grateful for your always patient explanation about innumerous Dutch documents I received, however busy you were. Many thanks to Niel and Marc, for providing the simulation and software resources, and to Jürgen, who helped translate the Dutch summary.

When trying to seek for work-life balance, I got to know many people during my spare time. I enjoyed every minute I spent with the families of Patrick and Diana, Kris and Marlinda, Jo and Esther, Guido and Pascale, Myriam, Albert, and many more. You are my family in Belgium. It was your love and support that raised me up when my soul was down-drifted by the difficulties I had met. Many thanks also go to JingShu Du (杜景姝), and the family of YongJun Shen (沈永俊) and Qiong Bao (鲍琼), as well as Mathew (Karren), Aaron (Stark), Nathen (Shepherd), and Logan (Gurr). Knowing you is a treasure in my life.

Most importantly, I owe a profound gratitude, which my words are insufficient to express, to my family. My husband, Bruce (薛峰), has sacrificed his time to take care of the family and planned journeys to help me shrug off the pressures of my study. I thank my parents YongLin Zhu (祝永林) and WenHua Zhou (周文华), and my mother-in-law YuJuan Gu (顾玉娟), for the material and spiritual support in these years. Being thousands of miles away from home, your longings of family reunion, which Chinese people treasure most, is the most luxurious thing for us. However, we believe that today's separation is for a more cherished reunion, in a longer term, even for eternity. Finally, this book is also dedicated to the memory of our deceased family members: JinSong Zhu (祝金松), XiuYing Ni (倪秀英), Hui Zhou (周辉), and LinQing Xue (薛林庆).

Qi Zhu （祝琦）
05 November 2010
Diepenbeek

# Samenvatting

In dit proefschrift ligt de nadruk op de analyse van massa spectrometrie (MS) data. Het gebruiken van dergelijke data voor de studie van proteïnen, kan leiden tot een nieuw inzicht in de moleculaire en cellulaire eigenschappen van biologische processen en mogelijk tot nieuwe biomerkers. *Proteomica*, de studie van alle aanwezige proteïnen in een cel of weefsel, wordt vaak als een vervolg op genomica (studie van genen en hun expressieniveau) beschouwd. In vergelijking tot genomica, is dit ingewikkelder omdat de hoeveelheid en het aantal unieke eiwitten varieert in tegenstelling tot de genen. Het proteoom, alle aanwezige proteïnen, kan verschillen van weefsel tot weefsel en van moment tot moment. In het begin van de proteomica, werd er vooral onderzoek gedaan naar boodschapper RNA, maar men vond dat er geen of weinig correlatie bestond tussen de hoeveelheid boodschapper RNA en het bijhorende proteïne (Rogers *et al.* 2008, Dhingraa *et al.* 2005). Men weet nu dat de aanwezigheid van boodschapper RNA niet altijd leidt tot de productie van proteïnen en dat de hoeveelheid eiwit afhankelijker is van het coderend gen en de fysiologische toestand van de cel dan van het boodschapper RNA. Recente ontwikkelingen in MS gebaseerde proteomica hebben geleid tot de mogelijkheid om op een geautomatiseerde manier duizenden proteïnen en peptiden gelijktijdig te kwantificeren en te identificeren. Deze mogelijkheden hebben voor nieuwe uitdagingen voor het statistisch onderzoek gezorgd. Eén van deze uitdagingen is het op een snelle, efficiënte en correcte manier conclusies trekken voor deze grote en complexe datasets. Hiernaast kan statistiek ook helpen bij het correct opzetten van massa spectrometrie gerelateerde experimenten. In dit proefwerk stellen we enkele statistische modellen voor die gebruikt kunnen worden voor de interpretatie van MS experimenten van gelabelde en niet-gelabelde peptiden. Het feit dat we niet weten welke peptiden aanwezig zijn en we dus hét onderscheidende kenmerk, nl. de massa van een peptide, niet kennen maakt het analyseren van MS data gecompliceerd. Het schatten van de massa zorgt voor een mate van onzekerheid

waarmee rekening moet gehouden worden omdat alle andere schatters afhankelijk zijn van de geschatte massa.

### $^{18}$O-labeled MS

In dit deel stellen we verscheidene modellen voor die het analyseren van gelabelde peptiden mogelijk maken. Peptidenstalen van verschillende biologische oorsprong worden vaak met *Liquid Chromatography Mass Spectrometry* geanalyseerd. Om geen rekening te moeten houden met de niet-biologische variaties tussen de verschillende spectra kan men deze stalen, na het labelen, samen verwerken in één spectrum. Dankzij het merken kunnen de peptiden van verschillende stalen maar met dezelfde massa in hetzelfde spectrum van elkaar onderscheiden worden. Er bestaan verschillende manieren om peptiden te labelen. Het enzymatisch labelen met $^{18}$O van peptiden is zo'n techniek die nieuw is en zeer veel potentie heeft. We stellen modellen voor die rekeningen houden met de vorm van de pieken, *stick* en *shape* en die ofwel op een Bayesiaanse ofwel op een frequentistische manier toegepast kunnen worden. In vergelijking met de methoden van (Mirgorodskaya *et al.* 2000, Rao *et al.* 2005, López-Ferrer *et al.* 2006, Eckel-Passow *et al.* 2006, Ramos-Fernández *et al.* 2007), is er geen nood aan extra experimentele stappen. Hiernaast zijn er nog enkele andere verschillen ten opzichte van de bestaande technieken:

- Er wordt rekening gehouden met de mogelijke aanwezigheid van de drie zuurstof isotopen.
- De isotoop distributie wordt geschat in plaats van een gemiddelde verdeling te veronderstellen.
- Alle parameters kunnen op hetzelfde ogenblik geschat worden en de precisie van de schatters wordt eveneens berekend.
- Het heteroscedastisch karakter van de *mean-variance* functie wordt niet genegeerd.
- Dankzij *random effects* kan de technische en biologische variabiliteit van de spectra bepaald worden.

### Kwantificatie van overlappende peptide in niet-gelabelde MS experimenten

Het bepalen of er peptide pieken zijn die overlappen en het scheiden van dergelijke pieken is al tientallen jaren een probleem. Verschillende onderzoekers met verschillende achtergronden hebben voor dit probleem oplossingen voorgesteld, variërend van aanpassingen maken aan het experiment en of het uitbreiden van het experiment tot het ontwikkelen van statistische modellen. In het tweede deel van deze scriptie worden twee modellen voorgesteld voor MS data met een ongekend aantal peptiden. We beschouwen opnieuw twee voorstellingswijzen van de pieken: *stick* en *shape*. Wanneer men gebruik maakt van de *stick representation* moet men rekening houden met een extra afwijking op de isotopen ratio's. De oorzaken van deze afwijking zijn het gebruiken van samenvattende statistieken en de veronderstellingen die gemaakt worden. In het geval van de *shape representation* zijn deze veronderstellingen niet nodig. Maar het *Bayesian mixture model* levert ook foutieve schatters op en overschat eveneens de precisie. *Bayesian model averaging* is een oplossing voor dit probleem, wanneer het massaverschil tussen twee overlappende peptiden groot genoeg is, dit wil zeggen wanneer het verschil minstens gelijk is aan de helft van de breedte van de piek. Wanneer men de massa's van de aanwezige peptiden kent, kan men dit model voor enkelvoudig geladen deeltjes aanpassen voor spectra met meervoudig geladen peptiden. Hiervoor volstaat het aanpassen van de *mean structure* en de *prior distribution*. In dergelijke gevallen wordt het model eenvoudiger, doordat men de massa van de peptiden niet meer hoeft te schatten en dus kan veranderen in vaststaande en gekende waarden. De resultaten van deze modellen zijn weliswaar afhankelijk van de kwaliteit van het gebruikte *preprocessing* algoritme. We steunen namelijk op het feit dat de gevonden pieken correct zijn. In het geval van het algoritme dat voorgesteld is door Valkenborg *et al.* (2009) betekent dit dat de verhouding tussen een piek en ruis groot genoeg moet zijn om gedetecteerd te kunnen worden.

# List of publications and reports

- Zhu, Q., Valkenborg, D. and Burzykowski, T. (2010) A Markov-chain-based heteroscedastic regression model for the analysis of high-resolution enzymatically $^{18}$O-labeled mass spectra. *Journal of Proteome Research,*, **9(5)**, 2669-2677.
- Zhu, Q., and Burzykowski, T. A Markov-chain-based regression model with random effects for the analysis of $^{18}$O-labeled mass spectra. *Journal of Statistical Modeling*, Manuscript submitted for publication.
- Zhu, Q., and Burzykowski, T. A Bayesian Markov-chain-based heteroscedastic regression model for the analysis of $^{18}$O-labelled mass spectra. *Journal of the American Society for Mass Spectrometry*, Manuscript submitted for publication.
- Zhu, Q., Kasim, A., Valkenborg, D. and Burzykowski, T. A Bayesian Model Averaging Approach to the Quantification of Overlapping Peptides in a MALDI-TOF Mass Spectrum. *Journal of Computational Biology*, Manuscript submitted for publication.
- Zhu, Q., Kasim A., Valkenborg, D., Jansen, I. and Burzykowski, T. A Bayesian approach to the quantification of overlapping peptides in a MALDI-TOF mass spectrum. *Technical report.*
- Zhu, Q., and Burzykowski, T. A Markov-chain-based shape model with heteroscedasticity for the analysis of $^{18}$O-labelled mass spectra. *Working paper.*
- Valkenborg, D., Van Sanden, S., Lin, D., Kasim, A., Zhu, Q., Haldermans, P., Jansen, I., Shkedy, Z. and Burzykowski, T. (2008) A Cross-Validation Study to Select a Classification Procedure for Clinical Diagnosis Based on Proteomic Mass Spectrometry. *Statistical Applications in Genetics and Molecular Biology:* Vol. 7, Iss. 2, Article 12.

- Valkenborg, D., Zhu, Q., Jansen, I., and Burzykowski, T. Identification issues for the assessment of peptide ion ratios with proteolytic $^{18}$O stable-isotopic labeling. *Technical report.*

# Contents

# Part I

# Introductory materials

# Chapter 1

# The focus and content of the dissertation

The main focus of the dissertation is the analysis and quantification of mass spectrometry-based proteomics to study molecular and cellular biological characteristics for the search of, e.g., new protein biomarkers, surrogate endpoints or markers of classification of diseases. *Proteomics* is the large-scale study of proteins, particularly their structures and functions. Proteomics confirms the presence of the protein and provides a direct measure of the quantity present. After genomics, proteomics is often considered the next step in the study of biological systems. It is much more complicated than genomics because, while an organism's genome is more or less constant, the proteome differs from cell to cell and from time to time. In the past research was focused on the mRNA analysis, but this was found not to correlate with protein content (Rogers *et al.* 2008, Dhingraa *et al.* 2005). It is now known that mRNA is not always translated into protein, and the amount of protein produced for a given amount of mRNA depends on the gene it is transcribed from and on the current physiological state of the cell.

Recent advances in mass spectrometry-based proteomics has led to the ability of quantifying and identifying thousands of proteins and peptides from complex biological samples in an automated and high-throughput fashion. Typically, such techniques extensively use liquid chromatography (LC) combined with mass spectrometry (MS) for protein-expression profiling. MS allows to separate peptides, present in a sample, according to their mass and charges. The LC step is used to reduce the complexity of the peptide mixture, which needs to be analyzed by MS, by separating the peptides based on their physico-chemical properties and by selecting only a subset for further processing in a mass spectrometer.

The advent of technological advances in proteomics has brought about a new area of statistical research. However, it is only until recently that the research has started to be devoted to the identification and quantification of peptides and proteins. The challenge of application of statistical methods lies in several aspects. First, mass-spectrometry data are often highly complex and with huge dimensionality. Thus, they require elaborate statistical analyses that extract and maintain most of the information from the data in a proper way. On the other hand, the large-scale nature of such data implies that the suited approaches should be a fast tool for the feasibility of real-life applications. In addition, a statistical analysis is said to be valid based on a well-conducted experiment. This indicates that the experimental design should also be statistically involved.

In the dissertation, some statistical modeling approaches will be introduced, both for the labeled and label-free MS experiments. It will be assumed that the sequences (function of the chemical elements that consist a peptide) of the peptides in the mass spectrometry data are unknown. Had the peptide sequences been known, the peptides in a mass spectrum could have been easily separated by their masses (as a known function of sequences) and the quantification would have become straightforward. Most of the existing methods are based on such assumption, which becomes an important limitation when in reality the sequences of the peptides are unknown. The assumption of unknown sequences of the peptides implies that peptides' masses are unknown and need to be estimated using an elaborate statistical approach. This becomes an additional source of complication, because an extra uncertainty related to the estimation of masses should be accounted for. Moreover, this uncertainty should be carefully analyzed, because all the other estimations are performed conditional on the estimation of masses. This is because, in a mass spectrum, masses of the peptides are the typical characteristics used to separate peptides. It is worth noting that the methods proposed in this dissertation require high-resolution mass spectra.

The dissertation consists of three parts. The first part gives a fundamental introduction to the mass-spectrometry data analysis. **Chapter 2** introduces the basic principles and terminology used for the analysis of mass spectrometry. **Chapter 3** describes the case studies that will be considered in the dissertation.

The second part focuses on the modeling approach that quantifies the peptides by using data from the enzymatic $^{18}$O-labeling. We implement the modeling approach in both the frequentist and Bayesian framework. **Chapter 4** gives the definitions related to the analysis and discusses the previous approaches considered for the modeling of the $^{18}$O-labeled mass spectra. In **Chapter 5**, **Chapter 6**, and **Chapter 7**, respectively, models assuming homoscedasticity, heteroscedasticity, and heteroscedasticity with random effects to account for the between-spectra variability, are presented. The models are formulated using the frequentist approach. The same types of the models in the Bayesian framework are implemented and their applications are discussed in **Chapter 8** and **Chapter 9**. Finally, in **Chapter 10**, we consider an extended model for the shape representation of the MS data, implemented in the frequentist framework, and we discuss its advantages over the models, presented in the previous chapters.

In the third part, a statistical model dealing with overlapping peptides with unknown sequences in mass spectra is introduced. This model is a solution for the separation and quantification of peptides processed in a label-free experiment. More specifically, **Chapter 11** gives an introduction to the existing methods, dealing with the estimation and quantification of overlapping peptides, and illustrates briefly the use of prior information, which underlies our estimation approaches. In **Chapter 12** and **Chapter 13**, respectively, we propose models for the stick and shape representations of a mass spectrum. We then present an improved method – Bayesian model averaging– in **Chapter 14**, based on the model described in **Chapter 13**.

# Chapter 2

# Introduction to mass spectrometry-based proteomics

This chapter provides a brief presentation of mass spectrometry and the notation, which will be used later for the analyses. Section 2.1 gives a general introduction to proteins, peptides, and proteomics. In Section 2.2, the principles of liquid chromatography-mass spectrometry (LC-MS), applied to proteomics, are described. Section 2.3 introduces the concept of the isotopic distribution. Section 2.4 gives a brief description of the pre-processing steps prior to the analyses of mass spectrometry data. In Section 2.5, some definitions of the shape and stick representations of the MS data, which will be used for the analyses in the dissertation, are provided. Part of the dissertation deals with the analyses of labeled mass spectrometry, therefore the basic principle of the enzymatic $^{18}$O-labeling is explained in Section 2.6.

## 2.1 Proteins, peptides, and proteomics

Proteins (also known as polypeptides) are organic compounds made of amino acids arranged in a linear chain by using information encoded in genes. Each protein has its own unique amino acid sequence that is specified by the nucleotide sequence of the gene encoding this protein. The genetic code is a set of three-nucleotide sets called codons. Each three-nucleotide combination designates an amino acid, for example AUG (adenine-uracil-guanine) is the code for methionine. Genes encoded in DNA are first transcribed into pre-messenger RNA (mRNA) by proteins such as RNA polymerase. Most organisms then process the pre-mRNA (also known as a primary transcript) using various forms of post-transcriptional modification to form the mature mRNA, which is then used as a template for protein synthesis (Mathews *et al.* 1999). Figure 2.1 depicts the schematic form of the biological information flow from DNA to protein.

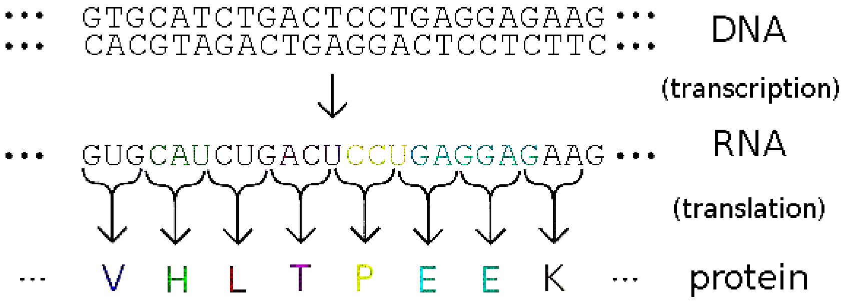

··· GTGCATCTGACTCCTGAGGAGAAG ···
··· CACGTAGACTGAGGACTCCTCTTC ···
DNA
(transcription)
··· GUGCAUCUGACUCCUGAGGAGAAG ···
RNA
(translation)
··· V H L T P E E K ···
protein

**Figure 2.1:** The DNA sequence of a gene encodes the amino acid sequence of a protein (http://en.wikipedia.org/wiki/Protein).

Overall, there is a total of 20 different amino acids that can occur in a peptide. Although each of the standard amino acids has a distinct structure, they do share a general set up. As displayed in Figure 2.2, all amino acids consist of an amino end, a carbon end, and a side chain. The only chemical elements occurring in these standard amino acids - and thus in proteins - are hydrogen (H), carbon (C), nitrogen (N), oxygen (O), and sulfur (S). Despite their identical general structure, amino acids differ considerably in the chemical structure of the side chain, ranging from very short chains to large sub-molecules and having varying chemical properties.

Proteins are also called the “polypeptide molecules”, distinguished from peptides only by the number of amino acid units. Thus, peptides are short polymers, composed of a smaller number of amino acids and thus their structures are less complex than proteins. The identification of proteins are allowed based on peptides’ masses and sequences when the technique of mass spectrometry is applied. In this case, the peptides are most often generated by in-gel digestion after electrophoretic separation of the proteins.

Proteomics, which studies the structures and functions of large-scaled proteins

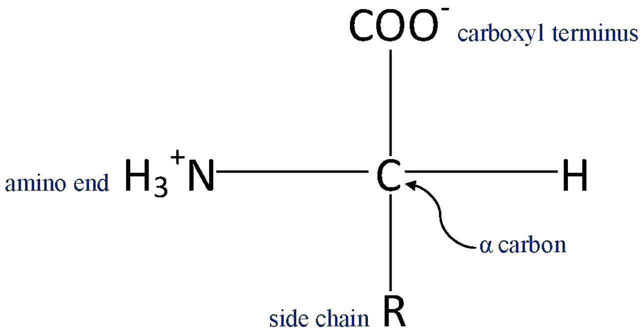


**Figure 2.2:** General chemical structure of an amino acid, which consists of an amino end, a carbon end and a side chain.

and peptides, has received a lot of interest in the recent years. It has been found that proteomics plays an irreplaceable role in the process leading to a better understanding of an organism. This is reflected in several aspects. First, the level of transcription of a gene gives only a rough estimate of its level of expression into a protein. An mRNA produced in abundance may be degraded rapidly or translated inefficiently, resulting in a small amount of protein. Second, many proteins experience post-translational modifications that profoundly affect their activities by, e.g., the current physiological state of the cell. Third, many transcripts give rise to more than one protein, through alternative splicing or alternative post-translational modifications. Fourth, many proteins form complexes with other proteins or RNA molecules, and only function in the presence of these other molecules (Belle *et al.* 2006). Hence, proteomics helps to obtain a better understanding of the biological characteristics of living organisms and cells.

The technique of mass spectrometry is a valuable tool in the field of proteomics. It can be used to identify proteins through variations of peptides' masses in a mass spectrum. Another use of mass spectrometry in proteomics is protein quantification. By labeling proteins with stable heavier isotopes, the relative abundance of proteins can be determined.

## 2.2 Liquid chromatography-mass spectrometry (LC-MS)

Liquid chromatography-mass spectrometry (abbreviated as “LC-MS”) is an analytical chemistry technique that combines the physical separation capabilities of liquid chromatography with the mass analysis capabilities of mass spectrometry. It is generally applied for the specific detection and potential identification of chemicals in a complex mixture. Typically, such technique works in two steps. The LC step is used to reduce the complexity of the peptide mixture by separating the peptides based on their physico-chemical properties. The “so-processed” peptide mixtures are then allowed to be separated according to their masses in the MS step.

### 2.2.1 Liquid chromatography

Chromatography is a term for a set of laboratory techniques for the separation of mixtures. It involves passing a mixture, dissolved in a “mobile phase”, through a stationary phase, which separates the analyte to be measured from other molecules in the mixture based on differential partitioning between the mobile and stationary phases. Subtle differences in a compound’s partition coefficient result in differential retention on the stationary phase and thus changing the separation.

Chromatography may be preparative or analytical. The purpose of preparative chromatography is to separate the components of a mixture for further use (and is thus a form of purification). Analytical chromatography is done normally with smaller amounts of material and is for measuring the relative proportions of analytes in a mixture. The two are not mutually exclusive.

Liquid chromatography (LC) is a separation technique in which the mobile phase is a liquid. Liquid chromatography can be carried out either in a column or a plane. Present day LC that generally utilizes very small packing particles and a relatively high pressure is referred to as high performance liquid chromatography (HPLC).

In the HPLC technique, the sample is forced through a column that is packed with irregularly or spherically shaped particles or a porous monolithic layer (stationary phase) by a liquid (mobile phase) at high pressure. The sample to be analyzed is introduced in small volume to the stream of mobile phase. The analyte’s motion through the column is slowed by specific chemical or physical interactions with the stationary phase as the analyte traverses the length of the column. How much the

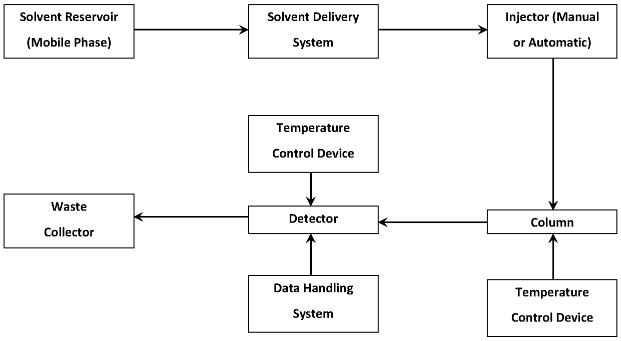

Solvent Reservoir (Mobile Phase)
Solvent Delivery System
Injector (Manual or Automatic)
Temperature Control Device
Waste Collector
Detector
Column
Data Handling System
Temperature Control Device

**Figure 2.3:** Schematic plot of the HPLC instruments and the working principle.

analyte is slowed down depends on the nature of the analyte and on the compositions of the stationary and mobile phases (http://en.wikipedia.org/wiki/Chromatography). The time, at which a specific analyte elutes (comes out of the end of the column), is called the retention time; the retention time under particular conditions is considered a reasonably unique identifying characteristic of a given analyte.

The basic instruments involved in an HPLC system can be represented schematically in Figure 2.3. Solvent from the reservoir is pumped at a certain selectable flow rate to an injector, at which point the sample is introduced and carried to the column. It is important that the flow be maintained slowly, continuously, and without pulsations. The column is the heart of the system where the separation of various components take place. The resolved components are then monitored by the detector and subsequently analyzed qualitatively and quantitatively through an integrator/plotter.

After chromatographic fractionation of the analyte mixture, the fractions of these analytes (peptides) are then processed in different mass spectra. As a result, a peptide can be present in several spectra due to the chromatographic fractionation.

### 2.2.2 Mass spectrometry

Generally, mass spectrometers are devices, which rely on separating charged ions by their mass-to-charge ratios. When applied in proteomics, it allows for separating peptide molecules by their different masses. Thanks to mass spectrometry, the peptide content of a biological sample can be visualized, which eases the identification and quantification of the peptides.

In the dissertation, we will focus on the application of Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry (MALDI-TOF-MS). A MALDI-TOF mass spectrometer is composed of several different parts: a source that ionizes the sample, the analyzer that separates the ions based on mass-to-charge ratio, and a detector that "sees" the ions.

### Matrix-Assisted Laser Desorption/Ionization(MALDI)

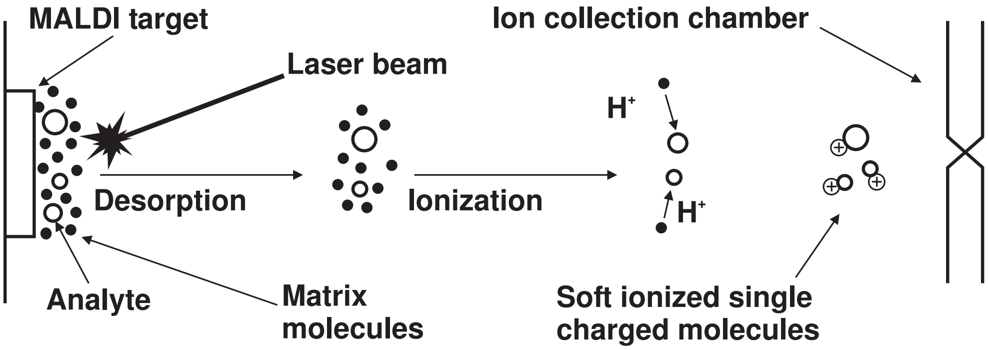


**Figure 2.4:** Basic principle of Matrix-Assisted Laster Desorption/Ionization (MALDI), reproduced from Valkenborg (2008).

MALDI is a soft ionization technique, which can be used to volatize entire molecules to gas phase (desorption) and to transfer a proton to the molecule (ionization). In MALDI analysis, the analyte molecules are first mixed with a high amount of matrix molecules in solution (usually a UV-absorbing weak organic acid), and spotted on a metallic plate. Next, the solution evaporates and the matrix molecules crystallize around the analyte molecules, serving as a protecting shield of the fragile analyte molecules. When irradiated by a laser, the matrix molecules efficiently absorb the laser energy directed towards the spot on the metallic target plate. Energy from the laser is converted into kinetic energy of the irradiated molecules, leading to the vaporization of a small amount of the spotted sample. In the gas phase, the protonation reactions occur, by which a proton is transferred between the acidic matrix ion and the molecules from the sample, leading to a charged vaporized molecule. After ionization, the ions (of the analyte and matrix) are colletected in the ion collection chamber (also called the "acceleration chamber"), waiting for the process of the Time-Of-Fligt (TOF) step.

The matrix plays a key role by strongly absorbing photon energy from the laser beam and transfering it into excitation energy of the solid system. On the other hand, the matrix serves as a solvent for the analyte, to reduce the intermolecular forces and the aggregation of the analyte molecules.

Figure 2.4 depicts the basic principle of the MALDI step.

### 2.2.3 Time-Of-Flight (TOF)

In this work, the linear TOF will be considered.

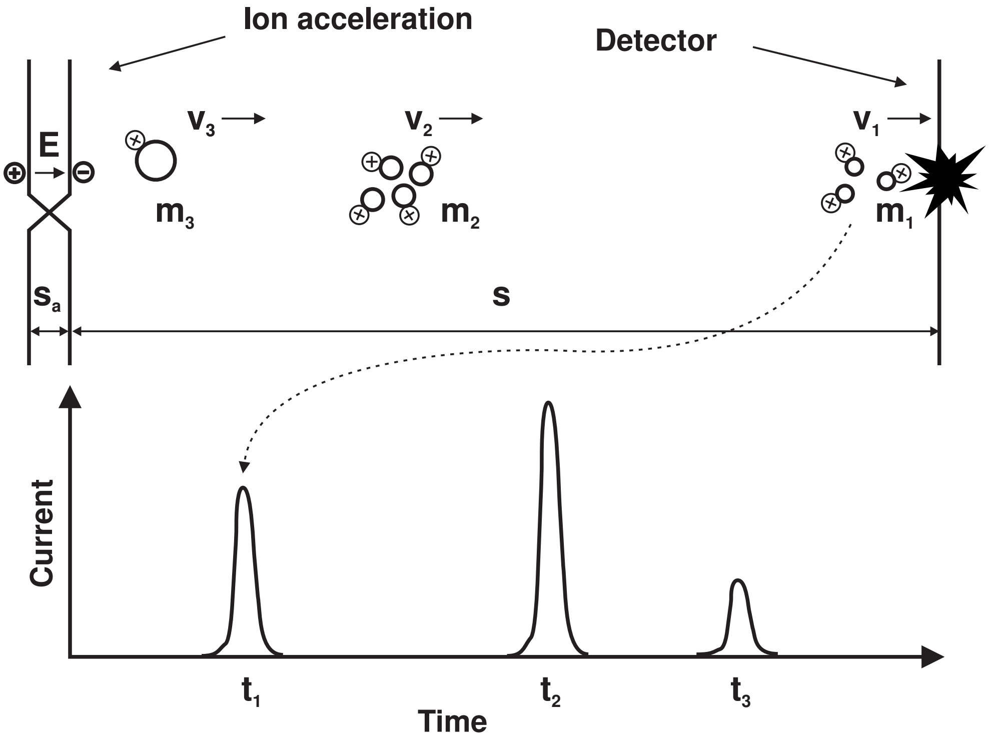


**Figure 2.5:** Basic principle of linear Time-of-flight Mass Spectrometer (TOF-MS), reproduced from Valkenborg (2008).

The basic principle of the linear TOF mass spectrometer is depicted in Figure 2.5. Essentially, a linear TOF mass analyzer is a long vacuum tube with a collection chamber at the inlet, collecting the ions generated from the ion source. At the end of the tube there is an ion detector that records the time of the arrival of the accelerated ions and counts the number of ions by measuring the electric current when the ions hit the detector.

At the last step of MALDI, the ions from the collection chamber are shortly exposed to a known electrical field, denoted as $E$, for which the potential energy in the chamber is transferred to kinetic energy of the ions. Suppose an ion with charge $q$ is exposed to an electrical filed $E$. As a result, a force $F$ with

$$F = Eq, \tag{2.1}$$

will be imposed on the charged ion. According to Newton's second law, this force can also be written as

$$F = ma, \tag{2.2}$$

where $m$ is the ion's mass. This means that the force that is imposed onto the charged ion, results in an acceleration $a = \frac{Eq}{m}$ of the ion. The ion with charge $q$, still in the collection chamber, is exposed to the electrical field and accelerates during a short time period $t_a$, until it leaves the acceleration chamber and enters the field-free vacuum tube with a certain velocity, $v$. Given a known distance $s_a$ in the acceleration chamber, again by applying Newton's laws, the time that the ion arrives at the vacuum tube $t_a$, and the corresponding velocity $v$ can be calculated. By assuming that the initial velocity $v_0$, i.e., the velocity of the ion before it enters the electrical field, is zero,

$$t_a = \sqrt{2\frac{s_a}{a}} = \sqrt{2\frac{m}{Eq}s_a}, \tag{2.3}$$

$$v = at_a = \sqrt{2\frac{Eq}{m}s_a}. \tag{2.4}$$

Subsequently, when the ion reaches the vacuum tube, the velocity $v$ will be kept constant. As the length of the vacuum tube $s$ is known, the time the ion takes to travel through the vacuum tube $t$ can be calculated:

$$t = \frac{s}{v} = \frac{s}{\sqrt{2\frac{Eq}{m}s_a}}. \tag{2.5}$$

The ratio of $m/q$ is defined as the mass-to-charge ratio. As the electrical field $E$, the acceleration distance $s_a$, and the length of the vacuum tube $s$ are known and kept constant, the travelling time $t$ for the ion in the vacuum tube depends only on the mass-to-charge ratio. It should be noted that the ratio $m/q$ , which is expressed in units of dalton-per-coulomb ($Da/C$), is difficult to interpret. Alternatively, instead

| $m/z$ | Intensity |
|---|---|
| 499.866376 | 26.1438 |
| 499.879673 | 24.9673 |
| 499.892970 | 20.6536 |
| 499.906267 | 21.5686 |
| 499.919565 | 24.5752 |
| 499.932863 | 27.4510 |
| 499.946160 | 28.7582 |
| 499.959458 | 27.8431 |
| 499.972757 | 25.8824 |
| 499.986055 | 30.5882 |
| 499.999353 | 37.3856 |
| 500.012652 | 32.2876 |
| 500.025951 | 27.4510 |
| 500.039250 | 32.4183 |
| 500.052549 | 38.8235 |
| ⋮ | ⋮ |

**Figure 2.6:** Output data from the TOF mass spectrometry.

of the absolute charge, a dimensionless variable $z$ can be introduced, representing the relative charge state of an ion. In other words, $z$ indicates the number of protons a molecule is carrying. In this respect, a peptide, which has a mass of 1000 Da and carries one proton ($H^+$) (called singly charged), corresponds roughly to a mass-to-charge ratio of 1001 $m/z$. This is because the molecules are protonated by the MALDI-procedure ($mH^+$). Therefore, the $m/z$ should be corrected by adding 1.00783 $Da$. Thus, a singly charged peptide will approximately have its mass-to-charge ratio equal to its mass value plus one. Similarly, the mass of a doubly charged peptide will be roughly two times its mass-to-charge ratio plus one. In the dissertation, we will only focus on the singly charged peptides. However, the developed modeling approaches can be adapted for the multiply charged peptides.

Finally, when the ions hit the detector, the detector records the time of the arrival of these ions and counts the number of ions arriving at distinct times by measuring the electric current, which represents proportionally the number of ions. The mass-to-charge ratios are then calculated from the arrival times through (2.5). As a result, the mass spectrometer produces an output with typically two variables, the mass-to-charge ratio ($m/z$) and the corresponding intensity value, as a measure of the number of ions (ion counts). Figure 2.6 gives an example of the output. To visualize the output, the data can be plotted as a line plot, as shown in Figure 2.7.

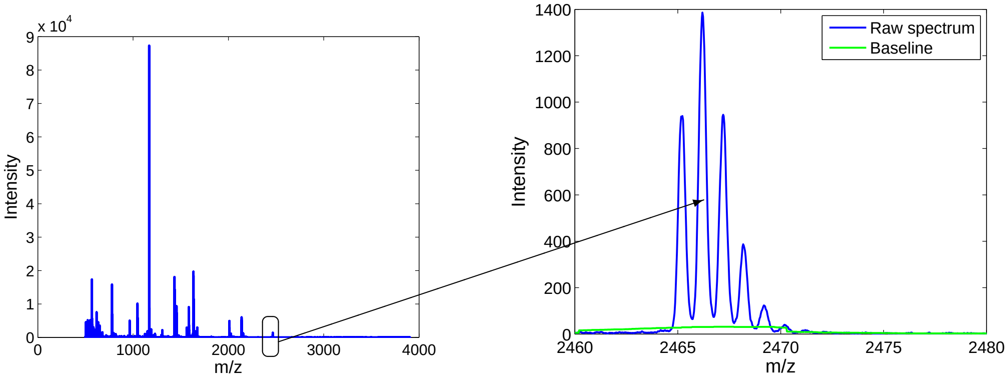


**Figure 2.7:** Graphical representation of the MS data.

## 2.3 Isotopic distribution

Because the five chemical elements that compose a peptide: carbon (C), hydrogen (H), nitrogen (N), oxygen (O), and sulphur (S), have different isotopes, peptides can have different isotopic variants, which differ with respect to their weights. For a peptide of a known chemical composition, the probability of occurrence of these variants is called the *isotopic distribution*. It follows that, in a high-resolution mass spectrum, a peptide produces a series of peaks that are separated by roughly 1 Da (for singly-charged peptides) and that correspond to different isotopic variants of the peptide. These peaks are called the *isotopic peaks*. Their relative heights are related to the probabilities of the isotopic distribution of the peptide. This leads to the idea of searching for peptide-related peaks in a mass spectrum by using an average isotopic distribution. By zooming in at around 2500 Da in Figure 2.7, an example of the "isotopic peaks" can be observed. Typically, these peaks appear in the vicinity of the mass range, with mass difference of roughly multiples of 1 Da. The first isotopic peak corresponds to the so-called *monoisotopic variant* of the peptide and is called the *monoisotopic peak*. The monoisotopic variant contains atoms of only the lightest isotopes of the chemical elements that contribute to the molecule.

To quantify the isotopic distribution, we define two sets of isotopic ratios by referring to Figure 2.8: the common reference ratios and the consecutive ratios. The common reference ratios are defined as ratios of the intensity values of each isotopic peak with respect to the intensity of the monoisotopic peak. More specifically, let

$h_j$ denote the probability of occurrence of the $j$th isotopic variant (see Figure 2.8). Given $l$ isotopic variants for a peptide, the isotopic ratio of the $j$th isotopic variant can be defined as

$$R_j = h_j/h_1, \text{ with } j = 1, ..., l. \tag{2.6}$$

The consecutive ratio $C_j$ is defined as a ratio of the intensity value of each isotopic peak with respect to that of its previous isotopic peak, i.e.,

$$\begin{cases} C_1 = h_j/h_j = 1 & \text{if } j = 1. \\ C_j = h_j/h_{j-1} & \text{if } j = 2, ..., l, \end{cases} \tag{2.7}$$

It can be easily seen that $C_1 = R_1 = 1$ and $R_j = C_1 C_2 \dots C_j$. We will later use both definitions of isotopic ratios, for different purposes.

To compute the isotopic distribution, the information about the chemical composition of the peptide is needed. Given the known chemical composition, the isotopic distribution can then be computed, e.g., by using a Fourier transform as proposed by Rockwood (1995). In reality, however, the chemical composition of a peptide is often not available. As an alternative, the average isotopic distribution can be predicted as a function of the mass. Several approaches (Breen *et al.* 2000, Gay *et al.* 1999, Senko *et al.* 1995, Valkenborg *et al.* 2007, Valkenborg *et al.* 2008) have been proposed to this aim.

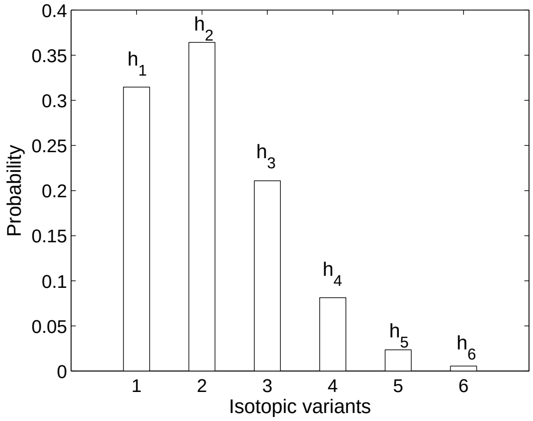


**Figure 2.8:** Stick representation of the isotopic peaks in a spectrum.

We consider two approaches to predict the distribution from the information about the monoisotopic mass of the peptide. Breen *et al.* (2000) suggested the use of an average distribution, obtained by using a Poisson approximation. In the approximation, the probability of the $l$th isotopic peak variant takes the form:

$$P(l;\mu) = \begin{cases} \frac{e^{-\mu}\mu^{l-1}}{(l-1)!} & \text{if } l = 1, 2, ...; \mu > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, Breen *et al.* (2000) discovered an empirical linear relationship between the Poisson mean $\mu$ and the monoisotopic mass $m$ of a peptide: $\mu = 0.000594m - 0.03091$. This linear relationship allows to predict the isotopic distribution of a peptide based on peptide's monoisotopic mass.

Another method to model the isotopic distribution, as suggested by Gay *et al.* (1999) and Valkenborg *et al.* (2008), is to use a polynomial model. The polynomial model either treats the abundance of each isotopic peak or the isotopic ratio regarding each peak as the response variable, and models it as a function of either the monoisotopic mass $m$ or a transformation of it. Valkenborg *et al.* (2008) suggested that a fourth order polynomial is sufficient as a result of assessing the improvement in the adjusted coefficient of determination with respect to the addition of an extra parameter. Valkenborg *et al.* (2008) also suggested that models fitted to the consecutive ratios produce smaller errors than ratios with the monoisotopic peak as the common reference. This is because the monoisotopic peak is always among the most abundant peaks which would result in larger errors for the ratio estimation if it is taken as the common reference for these ratios. More specifically, the model takes the following form:

$$C_l = \beta_{0l} + \beta_{1l}\left(\frac{m}{1000}\right) + \beta_{2l}\left(\frac{m}{1000}\right)^2 + \beta_{3l}\left(\frac{m}{1000}\right)^3 + \beta_{4l}\left(\frac{m}{1000}\right)^4 + \varepsilon_l, \qquad (2.8)$$

## 2.4 Pre-processing of mass spectrometry data

Prior to the analyses of mass spectrometry data, the data need to be pre-processed. In general, the pre-processing of the MS data comprises mainly four steps: *baseline correction*, *noise filtering*, *feature finding*, and *mass calibration*. For the details of the pre-processing algorithms, we refer to Valkenborg *et al.* (2009). The following sections explain briefly the algorithms, developed by Valkenborg *et al.* (2008) and Valkenborg *et al.* (2009).

### 2.4.1 Baseline correction

The intensity (ion count) is used as a measure for the abundance of a peptide in a spectrum. The green line shown in the right panel of Figure 2.7 indicates the baseline (i.e., an offset of the MS data) that varies along the mass coordinate. Although, the baseline can be viewed as constant (approximately horizontal) in a few Da range, for real applications, it is mandatory to subtract the baseline from the spectrum. In this way, the baseline variability does not influence the measure of abundance. Moreover, for distinguishing peptide-related peaks from the noises, the height of the peaks is used. The baseline would influence the height of the observed peaks and, consequently, would complicate the assessment of valid peptide peaks. Baseline is found by applying a moving window of around 10 Da and by smoothing the local minima of the intensity values using a linear extrapolation.

### 2.4.2 Dimensionality reduction

The data from one mass spectrum contain approximately 150,000 data points. Some of the measurements are likely to be noise-generated ones. We are only interested in finding the group of peaks corresponding to the composite isotopic distribution of a peptide. To reduce the dimensionality of the problem, we first select all the local maxima in a spectrum. However, many of these local maxima are likely due to noise and thus need to be filtered out. For this purpose, either a signal to noise ratio or a threshold intensity value, which is assumed to be the smallest intensity value for the peptide peaks, can be used.

### 2.4.3 Feature finding

After dimensionality reduction, the real features can be found by applying the typical characteristics of the peptide peaks. As has been mentioned in Section 2.3, a peptide produces a series of peaks, which correspond to different isotopes that compose the peptide, separated by approximately 1 Da. This feature can be used to distinguish the features from the noise. More specifically, the local maximum having mass difference with a threshold of 1.00235 Da can be selected.

### 2.4.4 Mass calibration

The monoisotopic mass is an appropriate measure for the location of the peptide in the mass dimension. But the monoisotopic mass measurement with a MALDI-TOF mass spectrometer is often affected by an error. First, the calibration is performed via a quadratic transformation by using the information about the internal standards, available in the data sets:

$$\text{TOF (time-of-flight) } = \beta_1 m/z + \beta_2 \sqrt{m/z} + \beta_3. \tag{2.9}$$

Next, the mass calibration is done based on the obtained values of $\beta_1$, $\beta_2$, and $beta_3$, from (2.9) by computing:

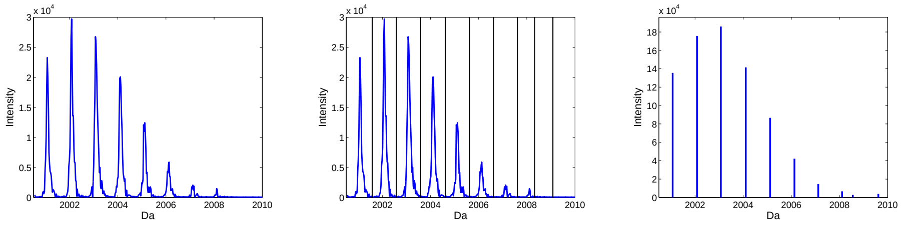
$$m/z_{cal.} = -\frac{\beta_2}{2\beta_1} \pm \sqrt{\left(\frac{\beta_2}{2\beta_1}\right)^2 - \frac{\beta_3 - TOF}{\beta_1}}$$

## 2.5 Mass spectrometry data representation

By looking at a particular cluster of the peptide peaks, shown in Figure 2.9a, one can clearly see that each peak exhibits a peak envelope composed of multiple data points. These peak envelopes follow similar shape. We term the original setting of the data the *shape representation.* Alternatively, each of the peaks can be represented by one data point, as a summary statistic for the peak. In this way, each of the peaks can be represented by a single data point, which can be reflected by a stick in the graphical representation, as shown in Figure 2.9c. For such type of setting, we term it the *stick representation.* More specifically, to obtain data for the stick representation, the spectrum is first binned with a bin width of approximately 1 Da. The binning points are defined to be the mid points of two neighboring local maxima (shown in Figure 2.9b. The intensity of a certain observed peak can then be taken either as a sum or a maximum of intensities of all the data points within a bin, corresponding to the observed peak. The mass location of the peak can be defined as the mass of the mid point within the bin.

## 2.6 Labeling techniques

In LC-MS applications, peptide samples from different biological conditions are usually analyzed. To avoid the between-spectra variability, these peptide samples can be



(a) Shape representation (b) Binned spectrum (c) Stick representation

**Figure 2.9:** Shape and stick representations of the mass spectrum data.

pooled and processed in the same spectrum. To do so, a labeling approach can be considered. The idea is similar to, e.g., two-channel cDNA microarrays, where mRNA from one sample is labeled with a green fluorescent dye and mRNA from another sample is labeled with a red fluorescent dye. Afterwards, the samples are pooled together and processed simultaneously. As a result, gene-expression measurements are subject to the same sources of variability. However, labeling with a dye is not an option in the mass spectrometry context. Instead, peptides are labeled with a stable isotope, which results in an increase of peptide's mass. The mass spectrometry based on stable isotope labeling method provides quantitative information of the peptides.

### 2.6.1 Enzymatic $^{18}$O-lableling

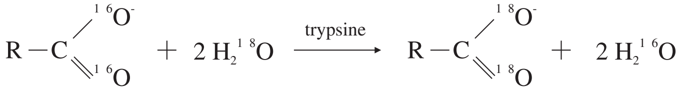
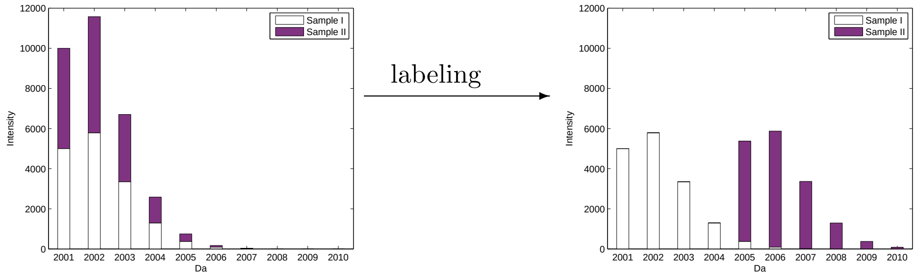


**Figure 2.10:** Chemical reaction scheme for the enzymatic $^{18}$O-labeling procedure.

A relatively new and powerful technique for stable isotope labeling is the enzymatic $^{18}$O-labeling (see Figure 2.10), which is a two-step labeling approach, as described by Miyagi and Rao (2007). In the first step, peptide samples from two cell states, or biological conditions are both digested in normal water ($H_2\,^{16}O$) with trypsine.

During the second step, the peptide sample to be labeled undergoes proteolysis, i.e., digest, in a heavy-oxygen-water reagent ($H_2\,^{18}O$). This step involves two reactions. In the first reaction, a $^{16}$O-carboxyl-oxygen atom is replaced by an $^{18}$O-oxygen from the heavy-oxygen-water. This hydrolysis reaction is fast and happens immediately upon the proteolytic digest, resulting in the incorporation of one $^{18}$O-oxygen atom from the heavy-oxygen-water into the carboxyl terminus of the peptide sample. As a result, this oxygen-exchange introduces a mass shift of two Da for the peptide sample.

The second reaction is much slower, and is in principle the reverse of the protease-catalyzed peptide-bond (Miyagi and Rao 2007). The oxygen-replacement of the carboxyl-terminus continues and is enzymatically catalyzed by a proteolytic reagent, in this case, trypsine. In this respect, the two-step labeling protocol is also referred to as enzymatic labeling. The reaction speed depends on multiple unobserved factors and therefore can be different for different peptides.



(a) Unlabeled

(b) Labeled

**Figure 2.11:** Effect of enzymatic $^{18}$O-labeling in a mass spectrum in the stick representation. Left panel: "sticks" can be seen as a representation of the distribution the isotopic variants of the peptide. Right panel: labeling causes accumulation of different isotopic variants in a joint spectrum.

Because there are two reaction sites, which can incorporate $^{18}$O-atoms, the labeling should lead, in ideal circumstances, to an increase of the mass of the peptide molecule by 4 Da, as two $^{18}$O-atoms are roughly 4 mass units heavier than two $^{16}$O-atoms. The resulting mass spectrum for peptide Sample I and Sample II are symbolically depicted in Figure 2.11.

Before the enzymatic $^{18}$O-labeling, the pooled peptide samples would appear at the same mass location in a mass spectrum, as illustrated on the left panel of Figure 2.11. In such case, the two samples would not be distinguishable from each other and therefore the quantification of the two peptide samples would have been impossible. After the enzymatic $^{18}$O-labeling, the isotopic peaks corresponding to the labeled peptide will shift four Da to the right, as shown in Figure 2.11b. A clear distinction between the peptide samples can now be made and the relative abundance of the peptide in the two samples can be calculated.

In practice, however, there are additional problems related to the use of the enzymatic $^{18}$O-labeling strategy. First, the heavy-oxygen water does not contain 100% pure $^{18}$O-water. It can also contain $^{16}$O- and $^{17}$O-atoms. We term these *water impurities*. Note that, if the two carboxyl-terminus oxygen atoms are replaced by, e.g., $^{17}$O-atoms, the peptide molecule becomes heavier by only 2, and not 4 Da, as it ideally would be the case in 100% pure $^{18}$O-water. Second, the speed of the enzymatic reaction, that is, the oxygen incorporation rate, depends on multiple unobserved factors and therefore can differ for different peptides. As a result, at the end of the enzymatic reaction, not all peptide molecules from the labeled peptide sample may have been actually labeled. The isotopic peaks for these molecules will overlap with the peaks from unlabeled sample. We term this situation the *incomplete labeling*.

Denote the proportions of $^{16}$O, $^{17}$O, and $^{18}$O atoms in the heavy-oxygen water by $p_{16}$, $p_{17}$, and $p_{18}$, respectively, with $p_{16} + p_{17} + p_{18} = 1$. Due to water impurities, the carboxyl-terminus of a peptide can contain different isotopes of oxygen. Let us consider the triplet $(n_{16}, n_{17}, n_{18})$, where $n_{16}$, $n_{17}$, and $n_{18}$ denote the number of $^{16}$O, $^{17}$O, and $^{18}$O atoms in a carboxyl-terminus, respectively. Clearly, $n_{16}+n_{17}+n_{18} = 2$. The possible isotope combinations can now be expressed as follows:

$$\begin{aligned} X(1) &= (2,0,0), \quad X(3) = (1,0,1), \quad X(5) = (0,1,1), \\ X(2) &= (1,1,0), \quad X(4) = (0,2,0), \quad X(6) = (0,0,2), \end{aligned} \tag{2.10}$$

For example, configuration $X(3) = (1, 0, 1)$ indicates that one of the carboxyl-terminus oxygen atoms was replaced by a $^{16}$O-atom, while the other was replaced by an $^{18}$O-atom. Note that the numbers of atoms of different isotopes of oxygen in the triplet sum to two, because the amount of oxygen atoms in the carboxyl-terminus cannot exceed two.

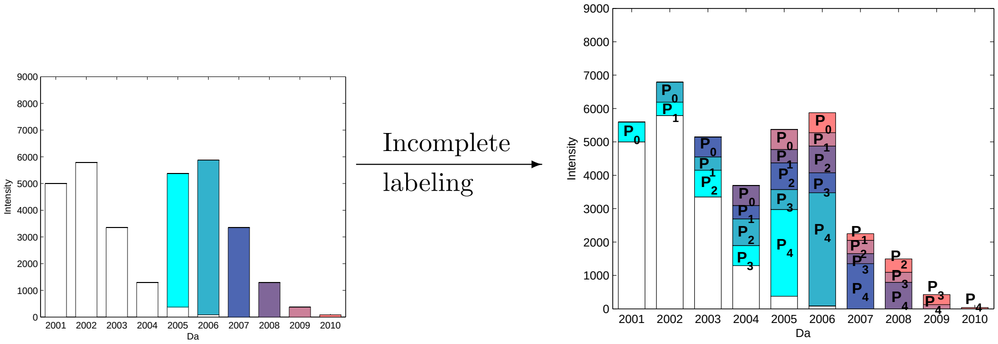
For different configurations $X(i)$, peaks corresponding to the isotopic distribution of a labeled peptide sample will shift with multiples of 1 Da. The mass shift depends on the configuration. The probability of a particular shift follows from the probability distribution of the six possible configurations of the carboxyl-terminus:

$$\begin{aligned} P_0 = P\{X(1)\}, \quad P_2 = P\{X(3)\} + P\{X(4)\}, \\ P_1 = P\{X(2)\}, \quad P_3 = P\{X(5)\}, \quad P_4 = P\{X(6)\}, \end{aligned} \tag{2.11}$$

where $P_k$ indicates the probability of the mass shift of $k$ Da ($k = 0, \ldots, 4$). It should be noted that we define the mass shifts relative to a carboxyl-terminus, which contains two $^{16}$O-atoms.

These problems imply that the peaks, observed for a peptide in a joint spectrum, will correspond to a complex mixture of shifted and overlapping isotopic peaks that are related to the isotopic distributions of the peptide molecules from the unlabeled and labeled samples. The corresponding isotopic peak heights of the two peptide samples are therefore distorted. Neglecting the distortion caused by the incomplete labeling will result in biased estimation for the quantification of the two peptide samples.

Figure 2.12 shows the effect of the incomplete labeling. In ideal circumstances, the enzymatic $^{18}$O-labeling should result in exactly four Da shifted for the labeled peptide sample, as depicted on the left panel of Figure 2.12. However, in reality, due to incomplete labeling, the isotopic peaks of the labeled sample exhibit multiple shifts of one Da. These shifted isotopic peaks get overlapped with those of the unlabeled peptide sample, distorting the peak heights of both samples. The amount of the labeled peptide molecules exhibiting multiple shifts correspond to the shift probabilities $P_k$, as demonstrated on the right panel of Figure 2.12.



(a) Ideal

(b) Real

**Figure 2.12:** Effect of incomplete labeling in a mass spectrum in stick representation. Left panel: ideal circumstance of enzymatic $^{18}$O-labeling, which results in four Da shift for the labeled sample. Right panel: due to incomplete labeling, isotopic peaks of the labeled sample show multiple shifts of one Da with the amount of shifted molecules corresponding to the shift probability $P_k$.

In order to estimate the relative abundance of the peptide in the two samples, the incomplete labeling has to be taken into account. In Part II of the dissertation, we introduce a modeling approach to correctly estimate the relative abundance of a peptide in the two samples.

# Chapter 3

# Case studies

The modeling approaches for the analysis of mass spectrometry data, to be introduced in this dissertation, are applied to a number of data sets. These data sets will be briefly described in Section 3.1. The research topics based on these data sets will thereafter be introduced in Section 3.2.

## 3.1 The considered data sets

### 3.1.1 Bovine cytochrome C mass spectra

Bovine cytochrome C is a relatively small protein related to mitochondria in a cell. It is a chain of 105 amino acids: MGDVEKGKKIFVQKCAQCHTVEKGGKHKTG-PNLHGLFGRKTGQAPGFSYTDANKNKGITWGEETLMEYLENPKKYIPGTKM-IFAGIKKKGEREDLIAYLKKATNE.

A peptide mixture of tryptic digested bovine cytochrome C was purchased from LC Packings and mixed with five internal standards from Laser BioLabs used for the calibration of the mass spectrometer. According to the data sheets of the suppliers, the mixture should contain 17 protein fragments. The amino acid sequences and the theoretical monoisotopic masses of these fragments are known. For instance, the sequence of the peptide at mass 1168.61 Da is: TGPNLHGLFGR; the sequence of the peptide with mass 1456.66 Da is: TGQAPGFSYTDANK; the sequence of the peptide with mass 1584.76 Da is: KTGQAPGFSYTDANK.

The peptide mixture was divided into two parts. One part was enzymatically labeled with a stable $^{18}$O-isotope, with trypsine as a catalyst, while the other part remained unlabeled (Miyagi and Rao 2007). In the first case, three units from the unlabeled part where mixed with one unit from the labeled part, what should result in the relative abundance of 1/3. In the second case, three units from the labeled part where mixed with one unit from the unlabeled part, what should result in the relative abundance of 3/1. In both cases, the composed mixture was automatically spotted six times on one stainless steel plate by a robot. The plate was processed by a 4800 MALDI-TOF/TOF analyzer mass spectrometer and yielded six spectra for the 1/3 mixture and six spectra for the 3/1 mixture.

### 3.1.2 NCBI public database

The RefSeq database of the NCBI, available at `http://www.ncbi.nlm.nih.gov/RefSeq`, provides the monoisotopic masses and istopoic distributions of human peptides. When accessed on February 27, 2008, for the human proteome, the database contained amino acid sequences for 132,292 proteins. Performing an *in silico* digest by trypsine results in 2,616,371 peptides with monoisotopic masses between 400 and 4000 Da, with 306,427 unique atomic compositions.

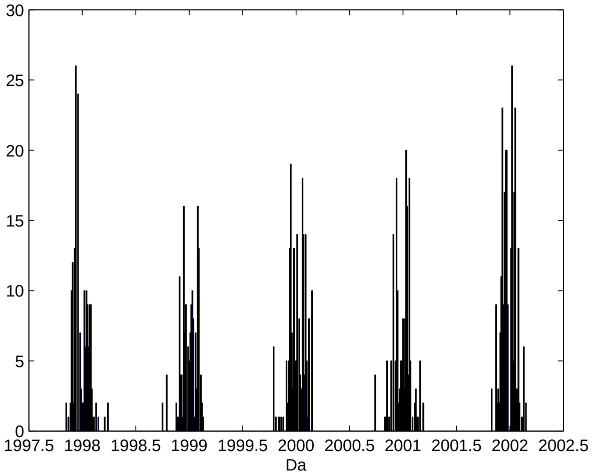


**Figure 3.1:** Histogram of the monoisotopic mass locations of peptides in the mass range of 1997.5-2002.5 Da in the NCBI data set.

The data set contains the isotopic distributions and monoisotopic masses of the 2,616,371 peptides. The information of the isotopic distributions and monoisotopic masses will later be used as supplementary information for the analysis of MS data in the dissertation.

For the isotopic distribution, the modeling approaches, described in Section 2.3, can be applied to the NCBI data. The obtained parameter estimates can then be included as the prior information for the estimation of the isotopic distribution.

Figure 3.1 presents the number of peptides with monoisotopic masses appearing in small intervals of 0.01 Da around the mass range of 2000 Da. It can be observed that the monoisotopic masses vary around integer values. Moreover, there are regions where no peptides can be found. This prior information can be quantified by using an appropriate prior distribution in modeling MS data.

## 3.2 The case studies

### 3.2.1 Modeling of enzymatic $^{18}$O-labeled mass spectra

In this study, a modeling approach for the estimation of relative abundance of the labeled and unlabeled peptides is implemented. The study is based on the bovine cytochrome C data. For each peptide, the six technical replications of spectra, each with two different relative abundances (1/3 or 3/1) of the $^{16}$O and $^{18}$O labeled peptides, are considered simultaneously in the modeling approach. We chose only the three peptides, with optimal data quality, namely peptides with masses 1168.6 Da, 1456.7 Da and 1584.8 Da.

### 3.2.2 The quantification of overlapping peptides in MALDI-TOF mass spectra

For this case study, we consider a label-free MS and assume that the sequences of the peptides, and therefore the masses of these peptides, are unknown. In a MALDI-TOF mass spectrum, peptides result in an overlap when they have similar masses and thus appear in the vicinity of the mass scale. We develop a method to quantify the overlapping peptides. The developed method can be easily modified for peptides with known sequences. In such case, the masses are known and the approach can be simplified.

The case study is, again, based on the bovine cytochrome C data from Section 3.1.1, with each spectrum treated as a source of pairs of overlapping peptides with a four Da difference in the monoisotopic masses. For the analysis purposes, we select two peptides with monoisotopic masses of 1456.66 Da and 1584.76 Da. For each peptide, we considered six spectra for each of the two different relative abundances (1/3 or 3/1).

Prior information, that assists the analysis, is obtained from the NCBI data.

### The prior information for the isotopic distribution

As has been mentioned in Section 2.3, the isotopic distribution can be predicted as a function of the mass. We adapted the method, proposed by Valkenborg *et al.* (2008), by fitting the models to the logarithmic consecutive ratios $C_l^*$ ($C_l^* = \ln C_l = \ln R_l - \ln R_{l-1}$, $(l = 2, ..., L)$, and $C_1^* = 0$) of NCBI data set (see Section 3.1.2). The gain of polynomial models fitted to the logarithmicly transformed consecutive ratios $C_l^*$ is three-fold:

1. Figures 3.2 and 3.3 show the histograms of the residuals of the models fitted to the original and log scales of $C_l$ respectively. It is clear that the log transformation gives more symmetric distributions for the model residuals;

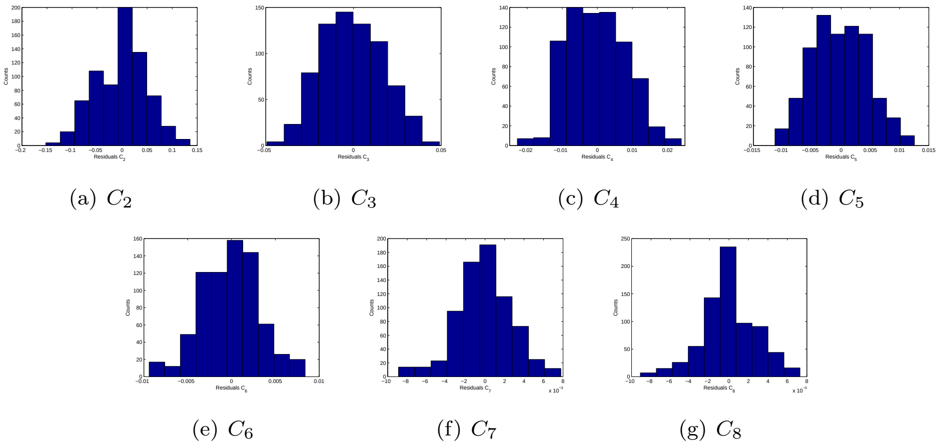
2. Figures 3.4 and 3.5 show the fit of the models to the two scales of $C_l$. It can be clearly observed that the variance (reflected as the 'bandwidth' of the scatters) of the observed ratios around the fitted lines after the log transformation is more constant;

3. When transforming the consecutive ratios to ratios with the monoisotopic peak as a common reference, the errors on the original scale are multiplicative while on the logarithmic scale they become additive, since $R_l = \exp\left(\sum_{i=1}^{l} C_l^*\right)$.

As a result, the model takes the form:

$$C_l^* = \beta_{0l} + \beta_{1l}\left(\frac{m}{1000}\right) + \beta_{2l}\left(\frac{m}{1000}\right)^2 + \beta_{3l}\left(\frac{m}{1000}\right)^3 + \beta_{4l}\left(\frac{m}{1000}\right)^4 + \varepsilon_l, \qquad (3.1)$$

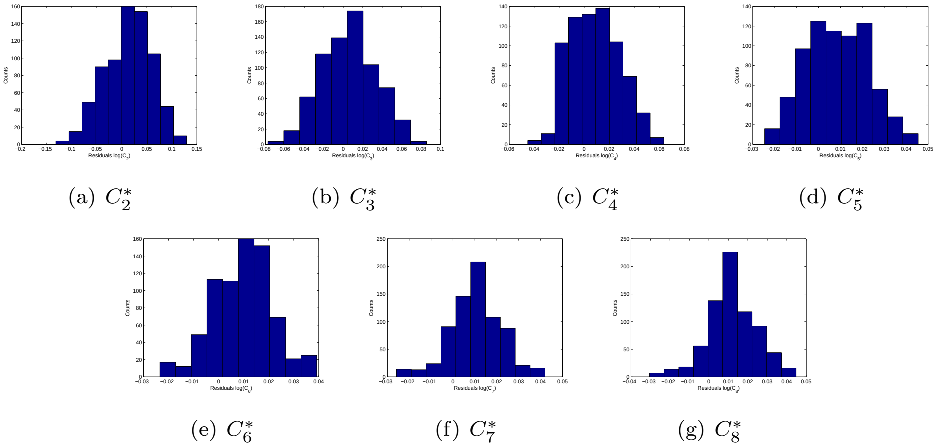
where $\varepsilon_l \sim N(0, \sigma^2_{C_l^*})$. We will use the estimates of this model as the informative prior distributions for the isotopic ratios. The reason of using the polynomial models as the prior, instead of the Poisson approximated one, is that the variability around



(a) $C_2$ (b) $C_3$ (c) $C_4$ (d) $C_5$

(e) $C_6$ (f) $C_7$ (g) $C_8$

**Figure 3.2:** Histograms of residuals for the polynomial model to the original scale of consecutive ratios $C_l$ at around 2000Da.

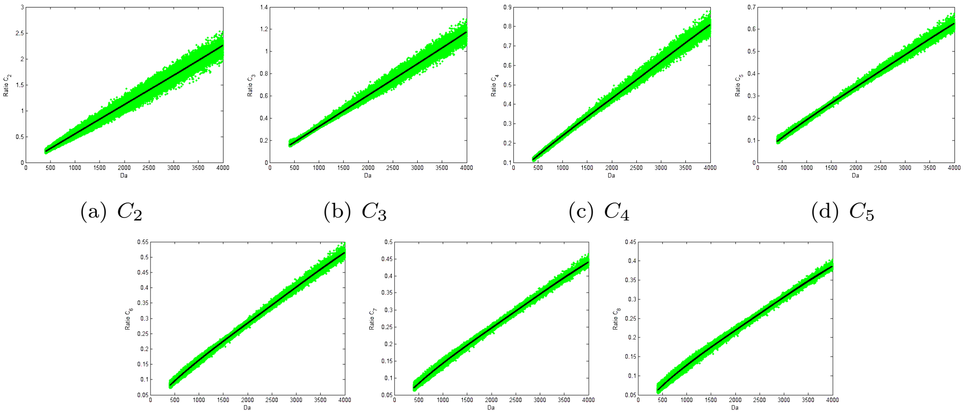


(a) $C_2^*$ (b) $C_3^*$ (c) $C_4^*$ (d) $C_5^*$

(e) $C_6^*$ (f) $C_7^*$ (g) $C_8^*$

**Figure 3.3:** Histograms of residuals for the polynomial model to the log scale of consecutive ratios $C_l$ at around 2000Da.

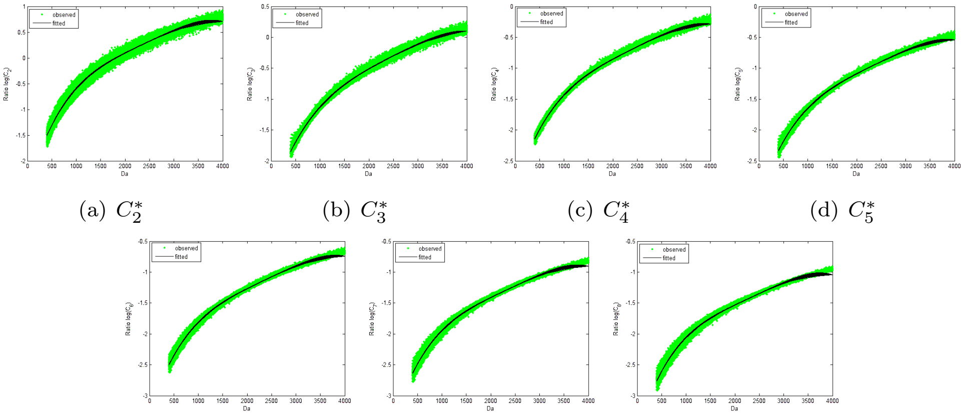
the average ratios can be obtained directly from the models and used in the prior distributions. Details of this will be given Chapter 12.



(a) $C_2$ (b) $C_3$ (c) $C_4$ (d) $C_5$

(e) $C_6$ (f) $C_7$ (g) $C_8$

**Figure 3.4:** The fit of the polynomial model to the original scale of consecutive ratios $C_l$ (green dots represent observed ratios; black line represents the mean of the polynomial models).



(a) $C_2^*$ (b) $C_3^*$ (c) $C_4^*$ (d) $C_5^*$

(e) $C_6^*$ (f) $C_7^*$ (g) $C_8^*$

**Figure 3.5:** The fit of the polynomial model to the log scale of consecutive ratios $C_l$ (green dots represent observed ratios; black line represents the mean of the polynomial models).

# Part II

# Analysis of enzymatically $^{18}$O-labeled mass-spectrometry data

# Chapter 4

# Introduction to the analysis of enzymatically $^{18}$O-labeled mass-spectrometry data

## 4.1 Enzymatic $^{18}$O-labeling and problem statement

As it has been mentioned in Section 2.6.1, the enzymatic $^{18}$O-labeling separates the labeled peptide samples from the unlabeled ones by four Da. As a result, the labeled peptides from, say, Sample II, can be pooled together with the unlabeled peptides from, say, Sample I, and processed simultaneously by LC and MS. Thanks to the enzymatic $^{18}$O-labeling, the isotopic peaks, which correspond to the labeled peptide, shift 4 Da to the right in the mass spectrum. This allows for making a distinction between the peaks related to peptides from different samples. Consequently, a direct comparison of the peptide abundance in the two samples is possible, because the abundance measurements are affected by the same amount of machine noise.

A "naïve" approach to compute the relative abundance of the peptide in the two samples would be to take the ratio of the heights of the first and fifth peak observed for the peptide in the joint mass spectrum (see Figure 4.1), as these peaks would correspond to the monoisotopic variants of the peptide in the unlabeled and labeled sample, respectively. However, as it can be observed from Figure 4.1, some isotopic peaks of the unlabeled peptide will still overlap with the monoisotopic peak

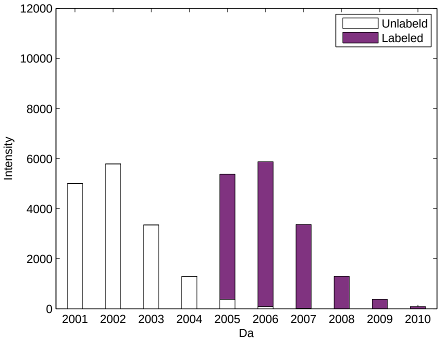


**Figure 4.1:** Peptide samples with enzymatic $^{18}$O-labeling in a mass spectrum in stick representation, under ideal circumstances: labeling causes isotopic variants shifted by 4 $Da$ in a joint spectrum.

of the labeled peptide. Thus, the ratio would yield a biased estimate of the relative abundance, because it does not take into account the overlap of the isotopic peaks.

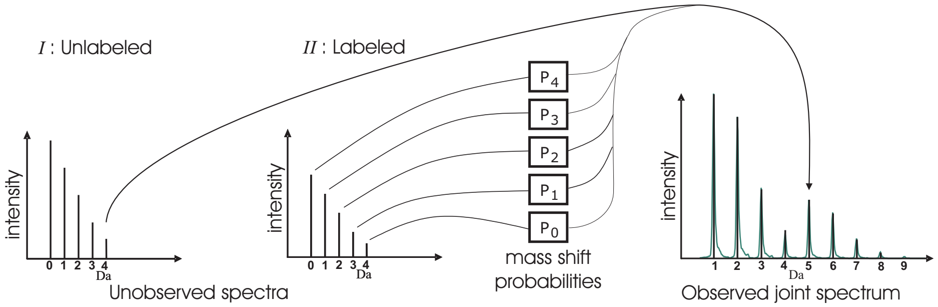
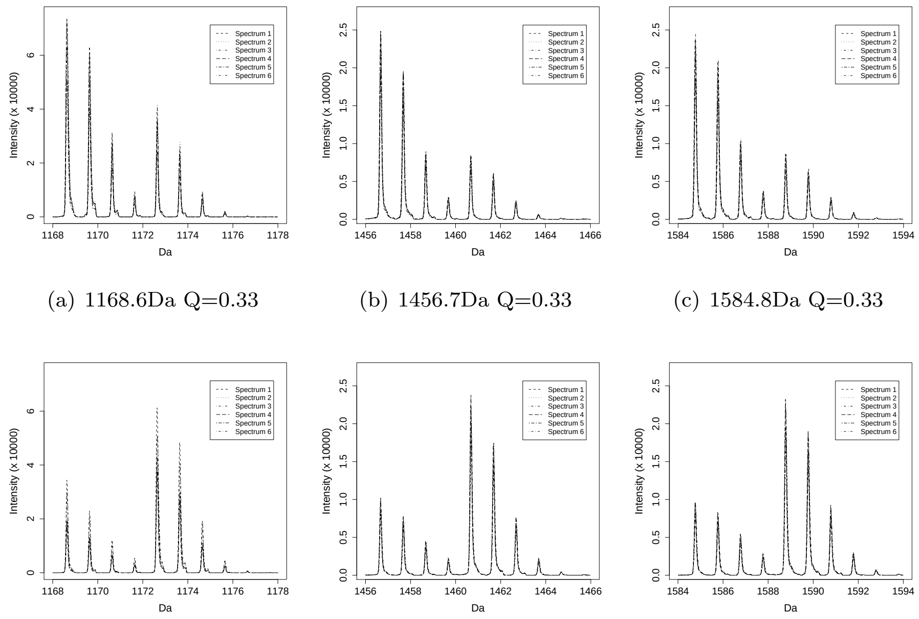


**Figure 4.2:** Fifth observed peak in function of the unobserved peptide peak intensities, reproduced from Valkenborg (2008).

In practice, however, there are additional problems related to the use of the enzymatic $^{18}$O-labeling strategy due to incomplete labeling. As explained in Section 2.6.1, the incomplete labeling results in the isotopic peaks of the labeled peptide sample shifted by multiples of one Da, instead of four Da. The amount of molecules with these multiple shifts correspond to the shift probability $P_k$. Taking the fifth observed peak as an example (Figure 4.2), its intensity value will be a sum of the intensities of the fifth isotopic peak of the unlabeled sample and of the isotopic peaks of the labeled sample with different shift probabilities $P_k$ (refer to Section 2.6.1).

These problems imply that the peaks, observed for a peptide in a joint spectrum, will correspond to a complex mixture of shifted and overlapping isotopic peaks that are related to the isotopic distributions of the peptide molecules from the unlabeled and labeled samples. In order to estimate the relative abundance of the peptide in the two samples, the overlap of the isotopic peaks has to be taken into account.

## 4.2 Data representation



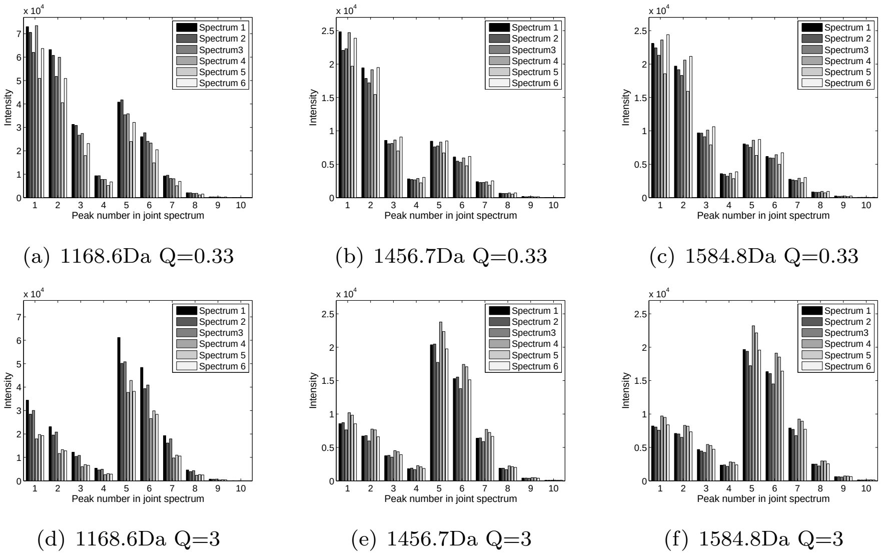
(a) 1168.6Da Q=0.33 (b) 1456.7Da Q=0.33 (c) 1584.8Da Q=0.33

(d) 1168.6Da Q=3 (e) 1456.7Da Q=3 (f) 1584.8Da Q=3

**Figure 4.3:** Graphical representation of the observed spectra for the six replications.

We apply the developed method to the data set of tryptic peptides of bovine cytochrome C from LC Packings.

We assume that, prior to the statistical analysis of a series of peaks observed in a MALDI-TOF spectrum and considered to be corresponding to a peptide, the spectrum was appropriately pre-processed. To this aim, we use the strategy proposed



(a) 1168.6Da Q=0.33 (b) 1456.7Da Q=0.33 (c) 1584.8Da Q=0.33

(d) 1168.6Da Q=3 (e) 1456.7Da Q=3 (f) 1584.8Da Q=3

**Figure 4.4:** Stick representation of the spectra presented in Figure 4.3. Bars of the same shade represent peaks from the same spectrum.

by Valkenborg *et al.* (2009). A summary of the pre-processing algorithm is described in Section 2.4. The pre-processing strategy extracts the information about the mass location and the height (intensity) of peaks, which are most likely due to a peptide.

We restrict the analysis to three bovine cytochrome C peptides, for which joint spectra of acceptable quality were obtained. The three peptides are with masses 1168.61 Da, 1456.66 Da and 1584.76 Da (detailed description of these peptides is available in Section 3.1). Figure 4.3 presents the shape representation, i.e., the original settings of the observed spectra for the three peptides in both mixing experiments. The modeling approach based on the shape representation of the spectra needs to take into account the shape of the envelopes.

Alternatively, a modeling approach can be based on the stick representation (defined in Section 2.5). In this respect, we take the maximum intensity value of a peak as the intensity of that peak. The stick representation of the data is presented in Figure 4.4. To work with the stick representation, several assumptions have to be made. First, the peak envelopes have exactly the same shape. Second, the isotopic peaks of Sample II align with those of Sample I. Or, equivalently, for each observed peak, the maximum intensity values of the isotopic peaks of the two samples are at the same mass location.

## 4.3 Previous approaches

Several methods have been proposed to deal with the analysis of the $^{18}$O-labeled mass spectra. On one hand, efforts aimed at the optimization of the enzymatic labeling process have been undertaken. For instance, methods that prohibit the back-exchange have been investigated (Storms *et al.* 2006). Alternatively, techniques that only allow for the incorporation of a single $^{18}$O-atom have been proposed (Rao *et al.* 2005).

On the other hand, approaches that address the issue at the analysis stage have been developed. Mirgorodskaya *et al.* (2000) have formulated a regression approach, which uses information about the isotopic distribution and about the labeling efficiency of the labeled peptide. The information is extracted from an additional mass spectrum of the labeled peptides, obtained before mixing the unlabeled and labeled sample. This extra MS step complicates the conduct of the experiment. Rao *et al.* (2005), López-Ferrer *et al.* (2006), and Ramos-Fernández *et al.* (2007) have suggested to identify the amino acid sequence of the peptide via an additional MS identification (tandem MS). Consequently, they can calculate the isotopic distribution of the peptide. The extra MS identification and the calculation of the isotopic distribution are computationally involved and require extra mass spectrometer time. Eckel-Passow *et al.* (2006) have proposed a regression approach similar in spirit to the method of Mirgorodskaya *et al.* (2000). They have used the method of Senko *et al.* (1995) to estimate the average isotopic distribution. This method is fast and does not need extra MS steps. However, the approach does not consider the possible presence of $^{17}$O atoms in the heavy-oxygen water. It can also lead to biased relative abundance estimates, as the actual isotopic distribution of a peptide can substantially deviate from the average isotopic distribution when, e.g., the peptide contains sulphur atoms (Valkenborg *et al.* 2007).

In the following chapters, we describe an alternate, model-based approach to estimate the relative abundance of a peptide from enzymatically $^{18}$O-labeled MS data. The approach uses the regression framework, considered by Mirgorodskaya *et al.* (2000) and Eckel-Passow *et al.* (2006). We combine the framework with a stochastic model, which describes the enzymatic $^{18}$O-labeling reaction. The method also allows us to estimate peptide's isotopic distribution from the observed data, which in turn can be used to validate if the peaks are indeed originating from a peptide. The pro-

**Table 4.1:** Overview of the models presented in Chapters 5 to 9.

| | Frequentist | | Bayesian | |
|---|---|---|---|---|
| | Fixed effects | Random effects | Fixed effects | Random effects |
| Homoscedasticity | Chapter 5 | – | – | – |
| Heteroscedasticity | Chapter 6 | Chapter 7 | Chapter 8 | Chapter 9 |

posed method is evaluated by using data from a controlled MS experiment, described in Section 3.1.1. In Chapter 5, we review the model with homoscedastic residual variance and with fixed effects in the frequentist approach, proposed by Valkenborg (2008). In Chapter 6, we extend the model to incorporate heteroscedastic residual variance by an appropriate mean-dependent variance-function. A further extension in the frequentist framework by accounting for the between-spectra variability is described in Chapter 7. A Bayesian modeling approach with and without accounting for the between-spectra variability is explained in Chapters 8 and 9, respectively. Table 4.1 gives an overview of the models presented in these chapters. Finally, the implementation of the model based on the shape representation of a spectrum is presented in Chapter 10.

# Chapter 5

# A frequentist approach to the analysis of $^{18}$O-labeled mass spectra using a homoscedastic fixed-effect discrete-time Markov-chain based model

## 5.1 Introduction

In this chapter, we review the approach of a homoscedastic regression model in the frequentist framework to analyze $^{18}$O-labeled mass spectra data, proposed by Valkenborg (2008). We first present a model to estimate the relative abundance of a peptide from a mixture of overlapping peptide peaks observed in the joint spectrum from an enzymatic $^{18}$O-labeling experiment. The model is overparameterized, so in the next step we describe further modeling steps to reduce the number of the parameters. After formulating the model, we describe methods for its estimation.

## 5.2 Model formulation

### 5.2.1 A model for the joint spectrum

As it has been mentioned in Section 2.6.1, the heavy-oxygen water contains water impurities, with proportions of $^{16}$O, $^{17}$O, and $^{18}$O atoms in the heavy-oxygen water denoted by $p_{16}$, $p_{17}$, and $p_{18}$, respectively, with $p_{16} + p_{17} + p_{18} = 1$. For the triplet $(n_{16}, n_{17}, n_{18})$ with $n_{16} + n_{17} + n_{18} = 2$, the possible isotope combinations are shown in equation (2.10). The probability of a particular shift follows from the probability distribution of the six possible configurations of the carboxyl-terminus and is defined in equation (2.11).

Consider a peptide, which has $l \geq 5$ isotopic variants (including the monoisotopic one). The enzymatic $^{18}$O-labeling and mixing of this peptide with its unlabeled counterpart will result in an observed joint spectrum of $l + 4$ peaks. The observed peak intensities $y_j$ in the joint spectrum, where $j = 1, 2, \ldots,$ denotes the position of the peak in the observed series of peaks in a joint spectrum, with $j = 1$ referring to the first observed peak of the spectrum, will be a function of the abundance of the unobserved isotopic variants of the peptide Samples I and II, i.e., the unlabeled and labeled samples, respectively. The function will depend on the mass shift probabilities, defined in (2.11).

To model the observed peak intensities, we assume that

$$y_j = \mu_j + \varepsilon_j \ , \qquad \varepsilon_j \sim N(0, \sigma^2) \tag{5.1}$$

and that $\varepsilon_j$'s are independent. The mean intensity, $\mu_j$, of the $j$th peak in the joint spectrum is expressed as follows:

$$\begin{aligned}
\mu_1 &= H_1^I + P_0 H_1^{II}, \\
\mu_2 &= H_2^I + P_0 H_2^{II} + P_1 H_1^{II}, \\
&\vdots \\
\mu_l &= H_l^I + P_0 H_l^{II} + P_1 H_{l-1}^{II} + P_2 H_{l-2}^{II} + P_3 H_{l-3}^{II} + P_4 H_{l-4}^{II}, \\
\mu_{l+1} &= P_1 H_l^{II} + P_2 H_{l-1}^{II} + P_3 H_{l-2}^{II} + P_4 H_{l-3}^{II}, \\
\mu_{l+2} &= P_2 H_l^{II} + P_3 H_{l-1}^{II} + P_4 H_{l-2}^{II}, \\
\mu_{l+3} &= P_3 H_l^{II} + P_4 H_{l-1}^{II}, \\
\mu_{l+4} &= P_4 H_l^{II}.
\end{aligned} \tag{5.2}$$

where $H_i^I$ and $H_i^{II}$ are the unobserved abundances of the $i$th isotopic variant (with $i = 1$ corresponding to the monoisotopic variant) in Sample I and Sample II, respectively, and $P_k$ is the mass shift probability (2.11). Now, upon defining relative abundance $Q$ of the peptide in Sample II and the isotopic ratios $R_i$ (refer to Section 2.3):

$$Q = \frac{H_1^{II}}{H_1^I}\ , \quad R_i = \frac{H_i^I}{H_1^I} = \frac{H_i^{II}}{H_1^{II}}\ , \tag{5.3}$$

where $R_1 = 1$. By putting $H \equiv H_1^I$, we can re-write (5.2) as follows:

$$\begin{aligned}
\mu_1 &= HR_1 + HQR_1P_0, \\
\mu_2 &= HR_2 + HQ(P_0R_2 + P_1R_1), \\
&\vdots \\
\mu_l &= HR_l + HQ(P_0R_l + P_1R_{l-1} + P_2R_{l-2} + P_3R_{l-3} + P_4R_{l-4}) \\
\mu_{l+1} &= HQ(P_1R_l + P_2R_{l-1} + P_3R_{l-2} + P_4R_{l-3}), \\
\mu_{l+2} &= HQ(P_2R_l + P_3R_{l-1} + P_4R_{l-2}), \\
\mu_{l+3} &= HQ(P_3R_l + P_4R_{l-1}), \\
\mu_{l+4} &= HQP_4R_l.
\end{aligned} \tag{5.4}$$

A few remarks regarding (5.4) are worth mentioning. The parameter of interest is $Q$, as it captures the relative abundance of the peptide in the two samples. Terms $HQP_kR_j$ denote the contributions to the mean values of the observed peaks from the isotopic variants of the peptide from Sample II (see Figure 4.2). Note that, for peaks $l+1, \ldots, l+4$, there are no contributions from the unlabeled peptide in Sample I. The isotopic ratios (5.3) are used for both the unlabeled and labeled peptide, because the ratios depend on the isotopic distribution, which is the same for both peptides. Finally, to determine the structure of the system of equations (5.4), we need to specify $l$. If a series of, say, $m \geq 9$ one-Da-separated peaks is observed in the joint spectrum, then we assume that the series was generated by $m - 4$ unobserved isotopic variants of a peptide, and we put $l = m - 4$.

Note that (5.4) is a system of equations with $5 + 2 + (m - 5) = m + 2$ parameters. However, this is more than the number of observations (peaks) $m$. Consequently, the model, specified by (5.1)–(5.4), is over-parameterized. Thus, we need to consider additional simplifying assumptions to reduce the number of parameters. These are discussed in the next section.

### 5.2.2 A model for the enzymatic $^{18}$O-labeling

To further reduce the number of parameters in (5.4), we consider a Markov-model (Welton and Ades. 2005) for the enzymatic $^{18}$O-labeling, so that the shift probabilities $P_0, \ldots, P_4$ are replaced by a smaller number of parameters.

In (2.10), we have introduced the configurations $X(i)$, which indicate the combination of oxygen isotopes present at the carboxyl-terminus of a peptide. We will refer to the configurations as states. We assume that, before the enzymatic labeling, the carboxyl-terminus of all isotopic variants of a peptide from Sample II contains two $^{16}$O-atoms, i.e., it is in state $X(1)$. This is depicted in Figure 5.1, where the white circle denotes state $X(1)$. After the first oxygen-atom replacement ($k = 1$), the carboxyl-terminus will stay with certain probability in state $X(1)$ or move to states $X(2)$ or $X(3)$. This is indicated by the light gray color in Figure 5.1, where the arrows indicate the possible direction of transitions. After the second oxygen replacement ($k = 2$), the probabilities for the carboxyl-terminus to remain in states $X(1)$, $X(2)$, or $X(3)$ will change. Moreover, three additional states can be reached, namely, $X(4)$, $X(5)$, and $X(6)$ (see the dark gray color in Figure 5.1). A third oxygen-replacement reaction ($k = 3$) will allow for eight new transitions, indicated by the black arrows in Figure 5.1, and so on. This process can be seen as a discrete-time Markov-chain, with the discrete time steps interpreted as the oxygen replacements. The discrete-time

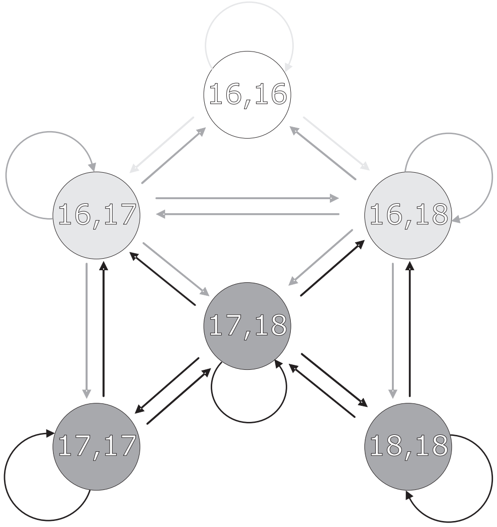


**Figure 5.1:** Transitions between carboxyl-terminus states.

Markov-chain can now be defined more formally. Given the transition probability matrix $\boldsymbol{T}$, the state probabilities are expressed as follows:

$$\boldsymbol{S}_k' \quad = \quad \boldsymbol{S}_0' \boldsymbol{T}^k P(k) \; , \tag{5.5}$$

with $\boldsymbol{S}_k$ denoting a $6 \times 1$ column vector containing the state probabilities after $k$ oxygen replacements and $P(k)$ denoting the probability that $k$ replacement reactions will take place. Under the assumption that at the beginning of the labeling process the isotopic variants of a peptide in Sample II contain 100% $^{16}$O-atoms at the carboxyl-terminus, the $6 \times 1$ initial state vector is given by $\boldsymbol{S}_0 = (1, 0, 0, 0, 0, 0)'$.

We assume that the enzymatic reaction is equally likely on both reaction sites of the carboxyl-terminus. We also assume that the previous oxygen replacements do not influence the enzymatic reaction for future oxygen replacements, i.e., that the transition probabilities are independent of the number oxygen replacements $k$. The transition probability matrix $\boldsymbol{T}$ with transition probabilities $P_{ij}$ is then given by

$$\begin{pmatrix} p_{16} & p_{17} & p_{18} & 0 & 0 & 0 \\ \frac{p_{16}}{2} & \frac{p_{16}+p_{17}}{2} & \frac{p_{18}}{2} & \frac{p_{17}}{2} & \frac{p_{18}}{2} & 0 \\ \frac{p_{16}}{2} & \frac{p_{17}}{2} & \frac{p_{16}+p_{18}}{2} & 0 & \frac{p_{17}}{2} & \frac{p_{18}}{2} \\ 0 & p_{16} & 0 & p_{17} & p_{18} & 0 \\ 0 & \frac{p_{16}}{2} & \frac{p_{16}}{2} & \frac{p_{17}}{2} & \frac{p_{17}+p_{18}}{2} & \frac{p_{18}}{2} \\ 0 & 0 & p_{16} & 0 & p_{17} & p_{18} \end{pmatrix} , \tag{5.6}$$

where $p_{16}$ and $p_{17}$ are the proportions of the heavy-oxygen water impurities (assumed known). Row ($i = 1, \ldots, 6$) and column ($j = 1, \ldots, 6$) indices correspond to states $X(1), \ldots, X(6)$. The transition probabilities $P_{ij}$ give the probability to move from state $X(i)$ to state $X(j)$. For example, the probability to move from state $X(3) = (1, 0, 1)$ to state $X(1) = (2, 0, 0)$ equals $P_{31} = p_{16}/2$, because only if the $^{18}$O-atom in state $X(3)$ is replaced by a $^{16}$O-atom, we reach state $X(1)$.

Term $P(k)$ in (5.5) denotes the probability of $k$ oxygen replacements. The number of oxygen replacements $k$ during the labeling reaction is unknown and depends on the reaction speed and duration. The duration of the enzymatic reaction is usually known and kept constant across multiple labeling experiments. We denote the duration by $\tau$. The reaction speed depends on many factors and is specific for each peptide. We express the speed as the peptide-specific incorporation rate $\lambda$, which gives the number of reactions per time unit. We assume that $\lambda$ is constant over time.

Under these assumptions, the probability for $k$ oxygen replacements can be modeled by a Poisson process with rate $\lambda$ and time $\tau$. As a result, after summing over all possible values of $k$ and rearranging terms, equation (5.5) is expressed as follows:

$$\boldsymbol{S}'(\lambda;\tau,p_{16},p_{17}) \quad = \quad \boldsymbol{S}'_0 e^{-\lambda\tau} e^{\boldsymbol{T}\lambda\tau} \ , \tag{5.7}$$

where $\boldsymbol{S}'(\lambda,\tau,p_{16},p_{17})$ is the vector containing the state probabilities for the isotope combination of the carboxyl-terminus of a peptide with incorporation rate $\lambda$ after a reaction time $\tau$ in heavy-oxygen water with impurities $p_{16}$ and $p_{17}$. Note that, to simplify notation, we will suppress the use of $\tau$, $p_{16}$, and $p_{17}$ in subsequent formulae.

Now, the probabilities of the isotopic distribution mass shifts, defined in (2.11), are computed as follows:

$$\begin{gathered} P_0(\lambda) = S_1(\lambda), \quad P_2(\lambda) = S_3(\lambda) + S_4(\lambda), \\ P_1(\lambda) = S_2(\lambda), \quad P_3(\lambda) = S_5(\lambda), \quad P_4(\lambda) = S_6(\lambda), \end{gathered} \tag{5.8}$$

where $S_i(\lambda)$ denotes the $i$th element of the state probability vector $\boldsymbol{S}(\lambda)$.

Figure 5.2 shows the values of the mass shift probabilities as a function of $\lambda$ for a labeling reaction of $\tau = 120$ in heavy-oxygen water with impurities $p_{16} = 2\%$ an $p_{17} = 1\%$. Note that, for $\lambda \geq 0.09$, the shift probabilities are basically constant. A similar plot would be obtained for the dependence of the probabilities on the reaction duration. It follows that, for a peptide with $\lambda \geq 0.09$, the enzymatic reaction is basically completed after 120 time units, e.g., minutes; extending the duration does not change the mass shift probabilities. This means that, if we consider a peptide with $\lambda = 0.09$, after $\tau = 120$ minutes, only 94.08% of the molecules will receive two $^{18}$O-atoms on their carboxyl group. In other words, isotopic peaks of only 94.08% of the peptide molecules from Sample II will shift by 4 Da to the right in the joint mass spectrum. Further, the peaks of 1.94%, 3.90%, 0.04%, and 0.04% of the labeled molecules will shift by 3, 2, 1, and 0 Da, respectively. The analysis of a labeled mass spectrum should correct for these different shifts and overlaps to avoid biased estimates of the relative peptide abundance. By using (5.7) and (5.8), we replace the five shift probabilities (2.11) by a single parameter, namely, $\lambda$. Consequently, we further reduce the number of parameters in (5.4) to $3 + (m - 5) = m - 2$, which is less then the number of available observations (peaks) $m$. This allows to fit the model, specified by (5.1)–(5.4) and (5.7)–(5.8), to observed data.

### 5.2.3 Estimation and inference

Assume that we have got $n$ joint spectra, each with $m$ observed peaks. The model, specified by (5.1)–(5.4) and (5.7)–(5.8), can be fitted to observed data by maximizing

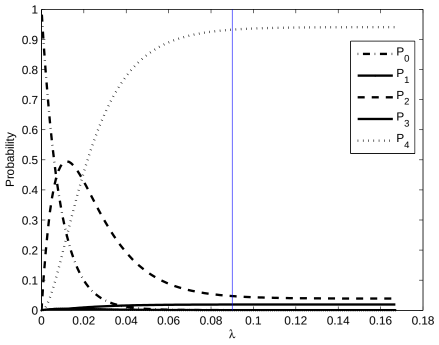


**Figure 5.2:** Shift probabilities $P_0, P_2, P_3$ and $P_4$ in function of $\lambda$ for an enzymatic reaction of 120 minutes with heavy-oxygen water impurities of $p_{16=2\%}$ and $p_{17} = 1\%$.

the log-likelihood, given by

$$l_{\mathrm{ML}}(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} \log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} \left[y_{ij} - \mu_{ij}(\boldsymbol{\beta})\right]^2, \tag{5.9}$$

where $y_{ij}$ is the $j$th observed peak in the $i$th joint spectrum, $\mu_{ij}$ is the corresponding mean value, and $\boldsymbol{\beta} = (H_1, \ldots, H_n, Q, \lambda, R_1, \ldots, R_{m-4})$ is a parameter vector that includes all the parameters used to model the mean value. Note that we use spectrum-specific intensities $H_i$ to adjust for the possible variation of intensity scales between the $n$ joint spectra.

Maximum-likelihood (ML) estimates of $\boldsymbol{\beta}$ and $\sigma^2$ can be obtained by simultaneously maximizing log-likelihood function (5.9) with respect to these parameters. Alternatively, the estimates can also be obtained via the least squares approach, i.e., by minimizing the sum of squared residuals: $\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}\left[y_{ij} - \mu_{ij}(\boldsymbol{\beta})\right]^2$. The REML-estimator for $\sigma^2$ is given by

$$\widehat{\sigma}^2_{\mathrm{REML}} = \frac{1}{nm - p}\sum_{i=1}^{n}\sum_{j=1}^{m} \left[y_{ij} - \mu_{ij}(\boldsymbol{\beta})\right]^2, \tag{5.10}$$

where $p$ is the total number of parameters to be estimated in (5.1)–(5.4) and (5.7)–(5.8).

### 5.2.4 Practical implementation

For practical implementation, it is worth noting that all the parameters were positive.

The analyses were done using Matlab 2009a with functions $fmincon$ (for constrained estimation) and $fminunc$ (for unconstrained estimation) in the optimization toolbox.

When performing the constrained estimation, the upper and lower boundaries should be specified for the parameters using function $fmincon$: the relative abundance $Q$ was constrained to be in the interval $[0, 100]$ and the reference intensities of each technical replicates $H_i$ was constrained to be non-negative. The (common-reference) isotopic ratios were constrained to be positive and not larger than 1.3. The oxygen incorporation rate $\lambda$ was constrained to be $[0, \lambda_0]$. This is because discrimination between large values of $\lambda$, so that the labeling efficiency reaches its maximum, becomes difficult.

The logarithmic transformation can be used for the parameters to work with an unconstrained optimization problem. For $\lambda$ we may want to ensure that it is bounded in a $[0, \lambda_0]$ interval. In this case, we can consider the use of the Box-Cox transformation: $\lambda = \lambda_0 \exp(\lambda')/\{\exp(\lambda') + 1\}$.

Both the constrained and unconstrained estimation approaches were implemented and they showed much similarity in terms of both estimating the mean structure and variance function parameters. Theoretically, the constrained and unconstrained estimation approaches should lead to the same statistical estimates. However, usually in practice, unconstrained estimation is more robust than constrained estimation since constraints often lead to the skewness of the parameter distributions while with a transformation of the bounded parameters, the distributions are often more symmetric. Global optimization becomes easier for the more symmetric distributions. Moreover, it is often unclear in reality what boundaries should be specified for each of the parameters when performing a constrained estimation approach. Hence, all the analyses in the following chapters of the dissertation will be based on the unconstrained estimation approach.

## 5.3 Results

We present results of the application of the model to the controlled experiment of the enzymatic labeling of bovine cytochrome C peptides (see Section 3.1). The model was estimated by minimizing the minus log-likelihood (5.9). The estimation approaches were implemented by using Matlab 2009a. In particular, function *fminunc* for unconstrained optimization problems with Newton-Raphson algorithm was used. To use the Newton-Raphson algorithm for the optimization, the gradient functions of all the parameters have to be specified analytically, which is feasible in our modeling approach. The proportions of water impurities of the heavy-oxygen water were assumed to be equal to $p_{16} = 2\%$ and $p_{17} = 1\%$. The true values of isotopic ratios $R_i$ were calculated from the atomic composition of the peptides by using the convolution method, developed by Rockwood (1995). As the duration of the experiment is not known, we estimate products $\lambda\tau$ instead of $\lambda$.

**Table 5.1:** Results of the analysis of the data for $Q = 1/3$ for the homoscedastic model (Est.: estimate; SE: standard error.)

| | 1584.8 Da | | | 1456.7 Da | | | 1168.6 Da | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | TRUE | Est. | SE | TRUE | Est. | SE | TRUE | Est. | SE |
| $H_1$ | – | 22919.2 | 54.15 | – | 24650.2 | 106.4 | – | 74535.5 | 1009.5 |
| $H_2$ | – | 22331.6 | 53.87 | – | 22255.9 | 104.5 | – | 72945.2 | 1002.5 |
| $H_3$ | – | 21289.5 | 53.44 | – | 22116.7 | 104.6 | – | 63110.5 | 975.8 |
| $H_4$ | – | 23742.0 | 54.40 | – | 24445.3 | 106.3 | – | 71135.1 | 1007.1 |
| $H_5$ | – | 18474.1 | 52.40 | – | 19583.8 | 103.0 | – | 48362.5 | 944.3 |
| $H_6$ | – | 24517.0 | 54.74 | – | 24315.8 | 105.7 | – | 61482.0 | 977.7 |
| $Q$ | 0.3333 | 0.3382 | 0.0060 | 0.3333 | 0.3419 | 0.0112 | 0.3333 | 0.5543 | 0.0255 |
| $\lambda\tau$ | – | 7.1631 | 0.4061 | – | 7.0290 | 0.6894 | – | 4.7178 | 0.3317 |
| $\sigma$ | – | 72.19 | 6.59 | – | 135.95 | 12.41 | – | 1342.16 | 122.52 |
| $R_2$ | 0.8703 | 0.8608 | 0.0017 | 0.7933 | 0.7892 | 0.0031 | 0.6645 | 0.8249 | 0.0105 |
| $R_3$ | 0.4223 | 0.3980 | 0.0039 | 0.3567 | 0.3276 | 0.0071 | 0.2454 | 0.2880 | 0.0170 |
| $R_4$ | 0.1478 | 0.1233 | 0.0033 | 0.1166 | 0.0880 | 0.0056 | 0.0653 | 0.0249 | 0.0143 |
| $R_5$ | 0.0413 | 0.0357 | 0.0037 | 0.0306 | 0.0258 | 0.0067 | 0.0139 | 0.0610 | 0.0130 |
| $R_6$ | 0.0097 | 0.0067 | 0.0027 | 0.0068 | 0.0023 | 0.0049 | 0.0025 | 0.0000 | 0.0000 |

Table 5.1 shows the results of the analysis for the three peptides, for which the intended value of the relative abundance $Q$ in the controlled experiment was 1/3. Several patterns can be observed in Table 5.1. First of all, for each peptide, a considerable between-spectra variability of intensity measurements, as indicated by the estimates of $H_i$, is worth noting. By using the $^{18}$O-labeling strategy, this variability is removed from the comparison of the peptide abundance in the unlabeled and labeled samples.

It is also worth noting that, for the peptides with masses 1584.8 Da and 1456.7

**Table 5.2:** Results of the analysis of the data for $Q = 3/1$ for the homoscedastic model (Est.: estimate; SE: standard error.)

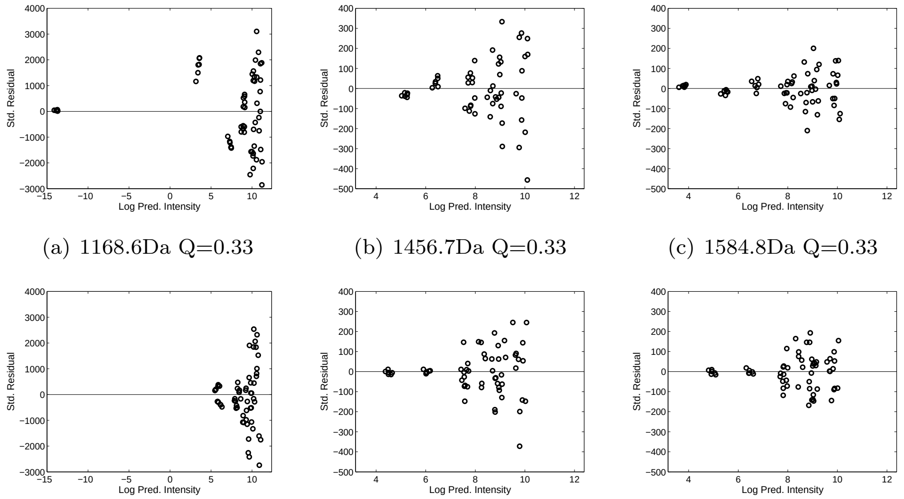
| | 1584.8 Da | | | 1456.7 Da | | | 1168.6 Da | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | TRUE | Est. | SE | TRUE | Est. | SE | TRUE | Est. | SE |
| $H_1$ | – | 8315.4 | 33.68 | – | 8682.8 | 47.58 | – | 32299.5 | 591.2 |
| $H_2$ | – | 8178.9 | 33.39 | – | 8773.4 | 47.84 | – | 26557.5 | 533.7 |
| $H_3$ | – | 7366.2 | 31.75 | – | 7703.2 | 44.87 | – | 27490.0 | 543.6 |
| $H_4$ | – | 9776.5 | 36.99 | – | 10107.3 | 51.84 | – | 18438.3 | 460.0 |
| $H_5$ | – | 9429.3 | 36.22 | – | 9664.3 | 50.52 | – | 20800.6 | 478.3 |
| $H_6$ | – | 8334.4 | 33.77 | – | 8547.2 | 47.21 | – | 19252.5 | 466.8 |
| $Q$ | (2.4) | 2.4122 | 0.0089 | (2.4) | 2.3896 | 0.0119 | (2.4) | 2.0073 | 0.0327 |
| $\lambda\tau$ | – | 9.3945 | 0.1534 | – | 11.7699 | 0.6823 | – | 20.0000 | 0.0001 |
| $\sigma$ | – | 81.26 | 7.42 | – | 109.77 | 10.02 | – | 1086.65 | 99.20 |
| $R_2$ | 0.8703 | 0.8611 | 0.0023 | 0.7933 | 0.7741 | 0.0028 | 0.6645 | 0.7541 | 0.0117 |
| $R_3$ | 0.4223 | 0.4184 | 0.0018 | 0.3567 | 0.3349 | 0.0023 | 0.2454 | 0.3069 | 0.0090 |
| $R_4$ | 0.1478 | 0.1350 | 0.0016 | 0.1166 | 0.0967 | 0.0021 | 0.0653 | 0.0739 | 0.0085 |
| $R_5$ | 0.0413 | 0.0338 | 0.0017 | 0.0306 | 0.0213 | 0.0022 | 0.0139 | 0.0067 | 0.0094 |
| $R_6$ | 0.0097 | 0.0075 | 0.0017 | 0.0068 | 0.0046 | 0.0022 | 0.0025 | 0.0093 | 0.0094 |

Da, the point estimates for $Q$ and for the isotopic ratios, for both models, are very close to the true values. For the peptide with mass 1168.6 Da, the point estimates differ from the true values.

Table 5.2 presents results of the analysis of the three peptides for the spectra, for which the intended value of relative abundance $Q$ was equal to 3/1. Remarkably, for the peptides with masses 1584.8 Da and 1456.7 Da, relative abundance $Q$ is consistently estimated to be equal to about 2.4. This suggests a possible inaccuracy at the stage of mixing the unlabeled and labeled samples when running the experiment.

The results, shown in Table 5.2, exhibit similar trends to those present in Table 5.1. The point estimates for the isotopic ratios for the peptides with masses 1584.8 Da and 1456.7 Da are in agreement with the values presented in Table 5.1 and with the true values. Some differences between the estimated values of $\lambda\tau$ in 5.2 and 5.1 can be observed. This may be due to, e.g., a difference in the duration of the labeling process.

For the peptide with mass 1168.6 Da, the estimates of isotopic ratios are more different from those reported in Table 5.1 and from the true values. Also, the estimation of $\lambda\tau$ is clearly yielding different results. This suggests some problems with the assumed form of the model for this peptide.



(a) 1168.6Da Q=0.33

(b) 1456.7Da Q=0.33

(c) 1584.8Da Q=0.33

(d) 1168.6Da Q=3

(e) 1456.7Da Q=3

(f) 1584.8Da Q=3

**Figure 5.3:** Scatter plots of the residuals versus the logarithm of predicted intensity values.

To check the goodness of fit of the model, scatter plots of the residuals versus the logarithm of the predicted intensity values are presented in Figure 5.3. The symmetry of the clouds of the residuals around the horizontal line at zero indicates the adequacy of the model with resect to its mean structure. However, the variability of the residuals is clearly not constant, but increases as the intensity value increases. This implies that the residual variance for the model is not correctly specified and that an improvement could be made by assuming the variance of the residual errors to be mean-intensity-dependent.

For comparison purposes, the data for $Q = 1/3$ were analyzed by using the method developed by Eckel-Passow *et al.* (2006). We applied it to each of the six mass spectra separately, as the method does not accommodate multiple spectra. Subsequently, we computed the mean values of the estimates of $Q$ and $\lambda\tau$, obtained for the spectra. The mean estimates of $Q$ were equal to 0.347, 0.345, and 0.595 for the peptides with masses 1584.8 Da, 1456.7 Da, and 1168.6 Da, respectively. For the first two peptides, the estimates are close to the corresponding values in Table 5.1 and to the true value of $Q = 1/3$. For the peptide with mass 1168.6 Da, the estimate is larger than the corresponding estimate in Table 5.1. The mean estimates of $\lambda\tau$ were equal to 3.38, 3.685, and 2.30 for the peptides with masses 1584.8 Da, 1456.7 Da, and 1168.6 Da, respectively. These values seem to be halved, as compared to the corresponding estimates, shown in Table 5.1. This is a systematic, expected effect, which can be explained theoretically (refer to Valkenborg 2008, p.185–187).

## 5.4 Discussion

As mentioned in Chapter 4, several methods have already been proposed to analyze data from enzymatic $^{18}$O-labeling experiments (Mirgorodskaya *et al.* 2000, Rao *et al.* 2005, López-Ferrer *et al.* 2006, Eckel-Passow *et al.* 2006, Ramos-Fernández *et al.* 2007). Most of them, however, postulate the use of additional experimental steps, what is an important limitation. The model described in this chapter does not require such steps. It is similar in spirit to the approach developed by Eckel-Passow *et al.* (2006). In fact, it is possible to show that the Markov-model, shown in this chapter, includes the model developed by Eckel-Passow *et al.* (2006) for the probabilities of particular mass shifts of the labeled peptide molecules (see equations (1) and (2) in their paper). However, the model extends in several ways the one used by Eckel-Passow *et al.*. First, it allows to account for the possible presence of $^{17}$O-atoms in the heavy-oxygen water. Second, Eckel-Passow *et al.* (2006) suggest to estimate the isotopic distribution of a peptide by using the average distribution developed by Senko *et al.* (1995). However, the actual isotopic distribution of a peptide can markedly deviate from the average one when, e.g., the peptide contains sulphur atoms (Valkenborg *et al.* 2007). Instead, in the proposed model, isotopic ratios of the isotopic distribution are estimated. The advantage of this solution is that the information about the ratios can be used to check whether the observed series of mass-spectrum peaks is truly generated by a peptide (Valkenborg *et al.* 2009). Note, however, that it is also possible to use the model with a fixed, e.g., predicted (Valkenborg *et al.* 2008), isotopic distribution. Finally, we develop a unified modeling framework, in which all parameters of interest, like the relative abundance and the peptide-specific incorporation rate, are simultaneously estimated from the data. It can easily accommodate different parameterizations, and provide necessary estimates of precision.

The results of the application to the real-life data for relative abundance $Q = 1/3$ were consistent with the true parameter values for two of three analyzed peptides. For one peptide, however, the results were biased both for the proposed model and for the methods of Zhu *et al.* (2010) and Eckel-Passow *et al.* (2006). The bias may be caused by the quality of MS-measurements in the available spectra by some experimental factors unknown to us. The possibility of the effect of such factors is further supported by the results of the analysis of the data for $Q = 3/1$, in which the relative abundance was consistently estimated to equal 2.4. This suggests that, in fact, the mixing of the samples during the conduct of the controlled experiment might have not been executed with a required precision.

In the next chapter, we present an extension of the model by accounting for the heterosecdastic nature of the residual variance, which is an important characteristic of the mass-spectrometry data.

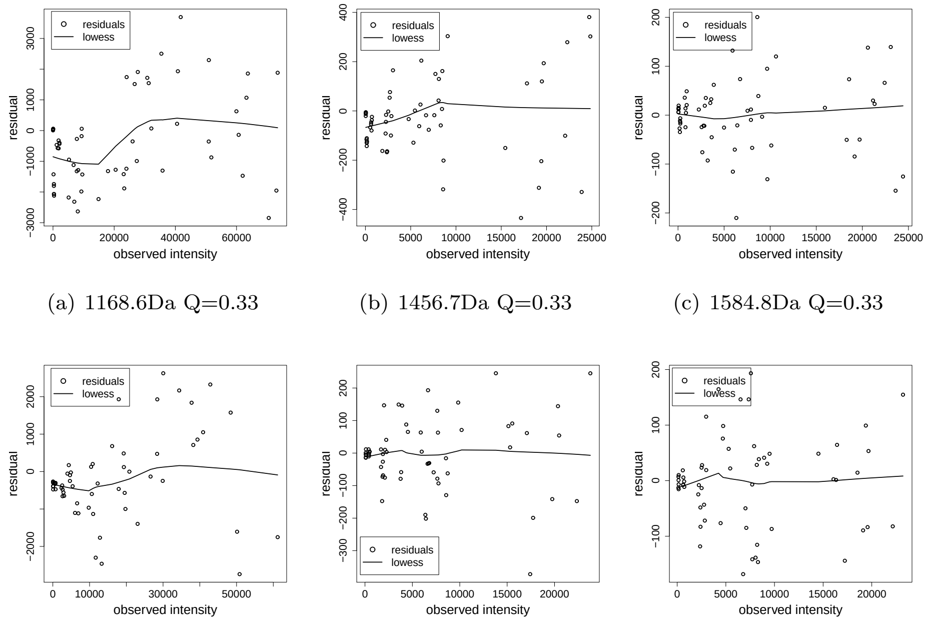
# Chapter 6

# A frequentist approach for the analysis of $^{18}$O-labeled mass spectra using a heteroscedastic fixed-effect discrete-time Markov-chain based model

In this chapter, we present a heteroscedastic regression model in the frequentist framework to analyze $^{18}$O-labeled mass spectra data, based on the model described in Chapter 5. The presented method has been published by Zhu *et al.* (2010).

## 6.1 Data exploration for an appropriate variance function

In Section 5.3, we noted that the variability of residual errors was dependent on the intensity scale. In order to find an appropriate variance function for the residual errors, we considered the residuals from the model assuming homoscedasticity, defined by (5.1)–(5.4) and (5.7)–(5.8). Figure 6.1 shows the scatter plots of the residuals and their observed intensities together with a lowess smoother. The smoothed curves, in general, do not exhibit systematic trends, but oscillate around the horizontal line with zero intercept, which indicates correct specification of the mean structure.



(a) 1168.6Da Q=0.33

(b) 1456.7Da Q=0.33

(c) 1584.8Da Q=0.33

(d) 1168.6Da Q=3

(e) 1456.7Da Q=3

(f) 1584.8Da Q=3

**Figure 6.1:** Model residuals versus the observed intensities for the model with a constant residual variance.

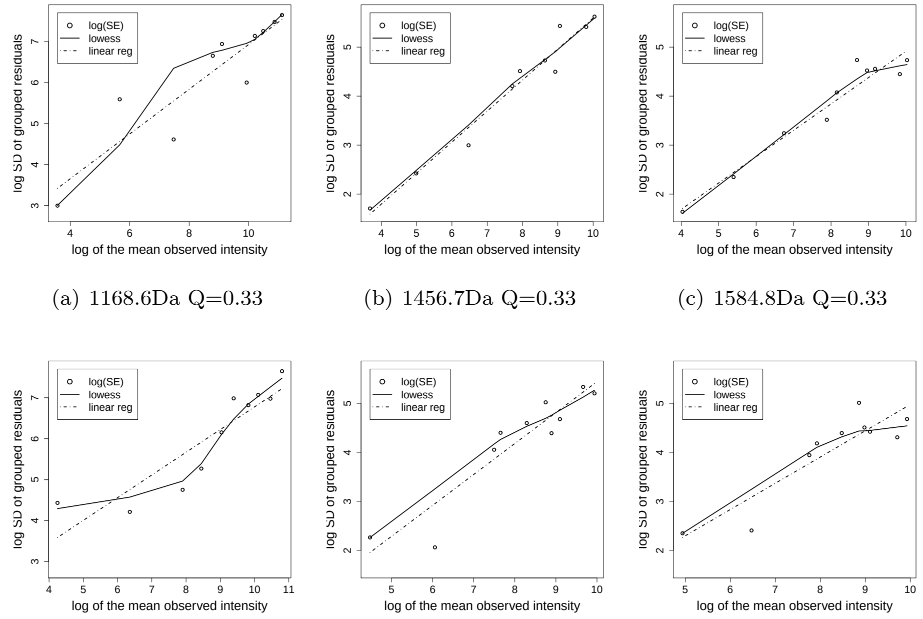
However, the variability of residuals increases with the intensity. This is illustrated in Figure 6.2, which shows the scatter plots of the logarithm of the variance of residuals versus the logarithm of the mean of the corresponding intensities. The variances were computed by grouping the residuals corresponding to the same peak in the six joint spectra. The interpretation of the scatter plots is enhanced by including a linear regression and a lowess curve in the plots. The lowess smoothed curves in the majority of the cases are fairly linear, and the linear regression lines fit the scatter plots quite well. This indicates that the residual variance can be described by a power function of the mean intensity with power parameter $\theta$, i.e.,

$$y_{ij} = \mu_{ij} + \varepsilon_{ij} \ , \qquad \varepsilon_{ij} \sim N(0, \sigma^2 \mu_{ij}^{2\theta}) \tag{6.1}$$

where $y_{ij}$ and $\mu_{ij}$ are, respectively, the observed and mean intensities of the $j$th observed peak in the $i$th spectrum. Based on (5.4), $\mu_{ij}$ can be equivalently written as:

$$\mu_{ij} \equiv \mathrm{E}(y_{ij}) = \begin{cases} H_i R_j + Q H_i \sum_{k=0}^{min(4,j-1)} P_k R_{j-k} & \text{if } 1 \leq j \leq l, \\ Q H_i \sum_{k=j-l}^{4} P_k R_{j-k} & \text{if } l+1 \leq j \leq l+4. \end{cases} \tag{6.2}$$

Consequently, further analysis of the data was based on the model, defined by (6.1)–(6.2) and (5.7)–(5.8).



(a) 1168.6Da Q=0.33

(b) 1456.7Da Q=0.33

(c) 1584.8Da Q=0.33

(d) 1168.6Da Q=3

(e) 1456.7Da Q=3

(f) 1584.8Da Q=3

**Figure 6.2:** The logarithm of the standard deviations of (grouped) model residuals versus the logarithm of the (mean) observed intensity for the model with a constant residual variance.

## 6.2 Estimation and inference

Assume that we have got $n$ joint spectra, each with $m$ observed peaks. The model, specified by (6.1)–(6.2) and (5.7)–(5.8), can be fitted to observed data by using various methods (Carroll and Ruppert 1988, Davidian and Giltinan 1995). The starting point for them is the log-likelihood, given by

$$l_{\text{ML}}(\boldsymbol{\beta}, \theta, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} \log\left\{\sigma^2 \mu_{ij}^{2\theta}(\boldsymbol{\beta})\right\} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\frac{y_{ij} - \mu_{ij}(\boldsymbol{\beta})}{\mu_{ij}^{\theta}(\boldsymbol{\beta})}\right\}^2, \tag{6.3}$$

where $y_{ij}$ is the $j$th observed peak in the $i$th joint spectrum, $\mu_{ij}$ is the corresponding mean value, and $\boldsymbol{\beta} = (H_1, \ldots, H_n, Q, \lambda, R_1, \ldots, R_{m-4})$ is a parameter vector that includes all the parameters used to model the mean value.

Maximum-likelihood (ML) estimates of $\boldsymbol{\beta}$, $\theta$, and $\sigma^2$ can be obtained by simultaneously maximizing log-likelihood function (6.3) with respect to these parameters. In general, however, this is a numerically complex task, which requires finding an optimum in a multidimensional parameter space. This task can be simplified by observing that, if we assume that $\boldsymbol{\beta}$ and $\theta$ are *known*, the estimator for $\sigma^2$ is given by

$$\widehat{\sigma}^2_{\text{ML}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\frac{y_{ij} - \mu_{ij}(\boldsymbol{\beta})}{\mu_{ij}^{\theta}(\boldsymbol{\beta})}\right\}^2. \tag{6.4}$$

By plugging expression (6.4) in (6.3) and omitting constant terms, we obtain the following *log-profile-likelihood* function, which depends only on $\theta$ and $\boldsymbol{\beta}$:

$$\begin{aligned} l^*_{\text{ML}}(\boldsymbol{\beta}, \theta) &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} \log\left\{\mu_{ij}^{2\theta}(\boldsymbol{\beta})\right\} - \frac{nm}{2}\log\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\frac{y_{ij} - \mu_{ij}(\boldsymbol{\beta})}{\mu_{ij}^{\theta}(\boldsymbol{\beta})}\right\}^2\right] \\ &= -\frac{nm}{2}\log\left[\left\{\prod_{i=1}^{n}\prod_{j=1}^{m}\mu_{ij}(\boldsymbol{\beta})\right\}^{\frac{2\theta}{nm}}\right] - \frac{nm}{2}\log\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\frac{y_{ij} - \mu_{ij}(\boldsymbol{\beta})}{\mu_{ij}^{\theta}(\boldsymbol{\beta})}\right\}^2\right]. \end{aligned} \tag{6.5}$$

Maximizing (6.5) with respect to $\theta$ and $\boldsymbol{\beta}$ allows obtaining estimates for these parameters. The estimates can then be used to compute the ML-estimate of $\sigma^2$ from (6.4). However, it is well known that the ML-estimator is biased downwards. Thus, especially when the number of spectra is small, it is better to replace it by the following

REML-estimator:

$$\widehat{\sigma}^2_{\text{REML}} = \frac{1}{nm-p}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\frac{y_{ij}-\mu_{ij}(\boldsymbol{\beta})}{\mu_{ij}^{\theta}(\boldsymbol{\beta})}\right\}^2, \tag{6.6}$$

where $p$ denotes the number of the estimated mean-structure parameters.

The use of log-profile-likelihood (6.5) still requires a simultaneous maximization of the function over $\boldsymbol{\beta}$ and $\theta$. Moreover, the use of log-likelihood (6.3) or of log-profile-likelihood (6.5) assumes that the data fulfill all the assumptions of the model, defined in (6.1), (6.2) and (5.7)–(5.8). If some of the assumptions are not fulfilled, the obtained estimates of the parameters may be incorrect.

An alternative estimation approach is to use a *pseudo-likelihood generalized least squares* (PL-GLS) approach (Davidian and Giltinan 1995), which is more robust to mis-specifications of the model, e.g., when the assumed distribution for the residuals is hampered by some outlying observations. Moreover, it is also much simpler numerically. In the case of the power-of-the-mean variance, as specified in (6.1), the approach is especially straightforward. Namely, log-profile-likelihood (6.5) can be expressed as

$$l^*_{\text{ML}}(\boldsymbol{\beta},\theta) \quad = \quad -\frac{nm}{2}\log\left[\{\widetilde{\mu}(\boldsymbol{\beta})\}^{2\theta}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\frac{y_{ij}-\mu_{ij}(\boldsymbol{\beta})}{\mu_{ij}^{\theta}(\boldsymbol{\beta})}\right\}^2\right], \tag{6.7}$$

where $\widetilde{\mu}(\boldsymbol{\beta}) = \left\{\prod_{i=1}^{n}\prod_{j=1}^{m}\mu_{ij}(\boldsymbol{\beta})\right\}^{\frac{1}{nm}}$. It follows that maximization of (6.7) is equivalent to minimization of

$$l^{**}_{\text{ML}}(\boldsymbol{\beta},\theta) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\{y_{ij}-\mu_{ij}(\boldsymbol{\beta})\}\left\{\frac{\widetilde{\mu}(\boldsymbol{\beta})}{\mu_{ij}(\boldsymbol{\beta})}\right\}^{\theta}\right]^2 \equiv \sum_{i=1}^{n}\sum_{j=1}^{m}\{f_{ij}(\boldsymbol{\beta},\theta)\}^2. \tag{6.8}$$

Thus, minimization of (6.8), either over $\theta$ (while keeping $\boldsymbol{\beta}$ fixed) or over $(\boldsymbol{\beta},\theta)$, can be viewed as an ordinary least squares (OLS) problem for a linear model with all data equal to 0 and $f_{ij}(\boldsymbol{\beta},\theta)$ as the fitted mean structure. It can be also viewed as a weighted least squares (WLS) problem for estimating $\boldsymbol{\beta}$, with weights

$$w_{ij}(\boldsymbol{\beta},\theta) = \left\{\frac{\widetilde{\mu}(\boldsymbol{\beta})}{\mu_{ij}(\boldsymbol{\beta})}\right\}^{\theta}.$$

As a result, the following algorithm can be used to estimate $\boldsymbol{\beta}$, $\theta$, and $\sigma^2$:

1. Set $k = 0$. Use an initial estimate $\widehat{\boldsymbol{\beta}}^{(0)}$ of $\boldsymbol{\beta}$.

2. Set $k = k + 1$.
3. While keeping $\widehat{\boldsymbol{\beta}}^{(k-1)}$ fixed, compute an estimate $\widehat{\theta}^{(k)}$ of $\theta$ from (6.8) by using OLS.
4. Compute weights $w_{ij}^{(k)}(\widehat{\boldsymbol{\beta}}^{(k-1)}, \widehat{\theta}^{(k)})$. While keeping the weights fixed, obtain estimate $\widehat{\boldsymbol{\beta}}^{(k)}$ of $\boldsymbol{\beta}$ by using WLS.
5. Iterate between steps 2–4 until convergence.
6. Use the obtained estimates of $\boldsymbol{\beta}$ and $\theta$ to compute an estimate of $\sigma^2$ from (6.4) or (6.6).

Irrespectively of the estimation approach used, standard errors of the estimates of $\boldsymbol{\beta}$, $\theta$, and $\sigma^2$ can be obtained from the inverse of the negative Hessian of log-likelihood (6.3), computed at the estimated values of the parameters. Alternatively, if parameters $\boldsymbol{\beta}$ are of the main interest, variance-covariance matrix of the estimate of $\boldsymbol{\beta}$ can be computed from the following formula (Davidian and Giltinan 1995):

$$\sigma^2 \left[ \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{ij}^{-2\theta}(\boldsymbol{\beta}) \frac{\partial \mu_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \left\{ \frac{\partial \mu_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right\}' \right], \tag{6.9}$$

with all unknown parameters replaced by their estimated values.

## 6.3 Application to bovine cytochrome C data

Again, we present results of the application of the model to the controlled experiment of the enzymatic labeling of bovine cytochrome C peptides. The model was estimated by maximizing log-profile-likelihood (6.5) and by using the PL-GLS approach (refer to Section 6.2). The estimation approaches were implemented by using Matlab 2009a. In particular, function *fminunc* for unconstrained optimization problems was used. The proportions of water impurities of the heavy-oxygen water were assumed to be equal to $p_{16} = 2\%$ and $p_{17} = 1\%$.

Table 6.1 presents the PL-GLS estimates of the model for the three analyzed peptides for the controlled experiment with intended relative abundance $Q = 1/3$. The estimates obtained by maximizing log-profile-likelihood (6.5) were very similar and therefore are not shown. Standard errors were obtained from the inverse of the

**Table 6.1:** Results of the analysis of the data for $Q = 1/3$ for the heteroscedastic model (Est.: estimate; SE: standard error.)

| | 1584.8 Da | | | 1456.7 Da | | | 1168.6 Da | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | TRUE | Est. | SE | TRUE | Est. | SE | TRUE | Est. | SE |
| $H_1$ | – | 22919.4 | 116.3 | – | 24585.0 | 221.2 | – | 79829.8 | 2785.3 |
| $H_2$ | – | 22330.7 | 114.2 | – | 22466.2 | 208.2 | – | 79495.0 | 2815.8 |
| $H_3$ | – | 21347.8 | 111.0 | – | 22335.1 | 208.2 | – | 68540.7 | 2480.6 |
| $H_4$ | – | 23857.2 | 121.7 | – | 24461.1 | 221.2 | – | 73386.6 | 2536.7 |
| $H_5$ | – | 18464.0 | 99.7 | – | 19525.5 | 184.9 | – | 48984.7 | 1803.0 |
| $H_6$ | – | 24687.6 | 126.5 | – | 24832.6 | 230.0 | – | 63503.9 | 2251.8 |
| $Q$ | 0.3333 | 0.3369 | 0.0028 | 0.3333 | 0.3384 | 0.0043 | 0.3333 | 0.5215 | 0.0186 |
| $\lambda\tau$ | – | 7.3162 | 0.2295 | – | 7.3941 | 0.3247 | – | 4.7178 | 0.2890 |
| $\sigma$ | – | 0.4394 | 0.2283 | – | 0.3471 | 0.1428 | – | 1.2152 | 0.5235 |
| $\theta$ | – | 0.6041 | 0.0645 | – | 0.6894 | 0.0514 | – | 0.7320 | 0.0461 |
| $R_2$ | 0.8703 | 0.8570 | 0.0038 | 0.7933 | 0.7806 | 0.0061 | 0.6645 | 0.7638 | 0.0226 |
| $R_3$ | 0.4223 | 0.3977 | 0.0025 | 0.3567 | 0.3273 | 0.0035 | 0.2454 | 0.2845 | 0.0121 |
| $R_4$ | 0.1478 | 0.1243 | 0.0015 | 0.1166 | 0.0904 | 0.0017 | 0.0653 | 0.0545 | 0.0036 |
| $R_5$ | 0.0413 | 0.0331 | 0.0008 | 0.0306 | 0.0211 | 0.0007 | 0.0139 | 0.0097 | 0.0011 |
| $R_6$ | 0.0097 | 0.0084 | 0.0003 | 0.0068 | 0.0057 | 0.0003 | 0.0025 | 0.0012 | 0.0002 |

**Table 6.2:** Results of the analysis of the data for $Q = 3/1$ for the heteroscedastic model (Est.: estimate; SE: standard error.)

| | 1584.8 Da | | | 1456.7 Da | | | 1168.6 Da | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | TRUE | Est. | SE | TRUE | Est. | SE | TRUE | Est. | SE |
| $H_1$ | – | 8316.0 | 53.45 | – | 8617.7 | 72.99 | – | 32130.8 | 857.0 |
| $H_2$ | – | 8168.4 | 52.80 | – | 8732.5 | 73.21 | – | 27108.5 | 723.9 |
| $H_3$ | – | 7424.0 | 49.99 | – | 7748.5 | 67.15 | – | 28827.4 | 775.5 |
| $H_4$ | – | 9773.1 | 60.68 | – | 10135.9 | 82.37 | – | 17037.4 | 468.9 |
| $H_5$ | – | 9482.2 | 59.72 | – | 9702.9 | 79.19 | – | 19488.7 | 527.7 |
| $H_6$ | – | 8372.2 | 54.17 | – | 8623.3 | 74.02 | – | 18502.8 | 503.8 |
| $Q$ | (2.4) | 2.3996 | 0.0131 | (2.4) | 2.3785 | 0.0165 | (2.4) | 2.1200 | 0.0467 |
| $\lambda\tau$ | – | 9.6100 | 0.1551 | – | 12.7660 | 0.8446 | – | 20.0000 | 0.0181 |
| $\sigma$ | – | 0.4624 | 0.3165 | – | 0.4743 | 0.2675 | – | 0.1178 | 0.0475 |
| $\theta$ | – | 0.6109 | 0.0823 | – | 0.6362 | 0.0690 | – | 0.9336 | 0.0448 |
| $R_2$ | 0.8703 | 0.8607 | 0.0042 | 0.7933 | 0.7739 | 0.0049 | 0.6645 | 0.7187 | 0.0177 |
| $R_3$ | 0.4223 | 0.4203 | 0.0027 | 0.3567 | 0.3374 | 0.0029 | 0.2454 | 0.2901 | 0.0083 |
| $R_4$ | 0.1478 | 0.1355 | 0.0013 | 0.1166 | 0.0975 | 0.0013 | 0.0653 | 0.0699 | 0.0023 |
| $R_5$ | 0.0413 | 0.0338 | 0.0005 | 0.0306 | 0.0213 | 0.0005 | 0.0139 | 0.0121 | 0.0005 |
| $R_6$ | 0.0097 | 0.0073 | 0.0002 | 0.0068 | 0.0044 | 0.0002 | 0.0025 | 0.0015 | 0.0001 |

negative Hessian of log-likelihood (6.3). The values obtained from (6.9) were very similar.

Several patterns can be observed in Table 6.1. First of all, for the peptides with masses 1584.8 Da and 1456.7 Da, the point estimates are quite similar to those in the homoscedastic model shown in Table 5.1 for all parameters, except for $\sigma$. The latter difference is obvious, because the two models assume different forms of residual variance. Important differences can be seen in the standard errors of the parameters. For instance, for the parameter of interest, $Q$, standard errors for the homoscedastic model (see Tables 5.1 and 5.2) are larger than for the heteroscedastic one. This means that the former model is less efficient in estimating $Q$. As a consequence, the model may more often lead to false negative findings, i.e., to declare differences in peptides' abundances as statistically nonsignificant while, in fact, there are such differences. On the other hand, standard errors for, e.g., the reference intensities $H_i$ are smaller for the homoscedastic model. Thus, they are more “optimistic”, i.e., they suggest a higher precision of estimation of the parameters than it is actually the case, as seen from the heteroscedastic model.

Again, for the peptides with masses 1584.8 Da and 1456.7 Da, the point estimates for $Q$ and for the isotopic ratios, for both models, are very close to the true values. For the peptide with mass 1168.6 Da, the results of the two models differ.

Table 6.2 presents results of the analysis of the three peptides for the spectra, for which the intended value of relative abundance $Q$ was equal to 3/1. The results, shown in Table 6.2, exhibit similar trends to those present in Table 6.1. The estimated values of $\theta$ for the heteroscedastic model are close, taking into account the precision of estimation, to the results from Table 6.1. This suggests that the chosen functional form of dependence of residual variance on intensity was appropriate.

## 6.4 A simulation study

In this section, we show results of a simulation study, undertaken to check the statistical properties of the proposed heteroscedastic model.

For the simulation, we chose three sets of isotopic ratios – the average one (denoted by **A**) obtained by a Poisson approximation with the model developed by Breen *et al.* (2000) (see Section 2.3); the extremely small ratios (denoted by **E1**); and the extremely large ratios (denoted by **E2**) within $2001 \pm 0.5$ Da mass range. **E1** and **E2** are the isotopic distributions with the second isotopic peak being the least and most abundant among all the peptides around 2001 Da from the NCBI data (see Section 3.1). Figure 6.3 illustrates graphically the three sets of isotopic distributions.

The duration of the enzymatic reaction was kept constant at $\tau = 120$ minutes. The

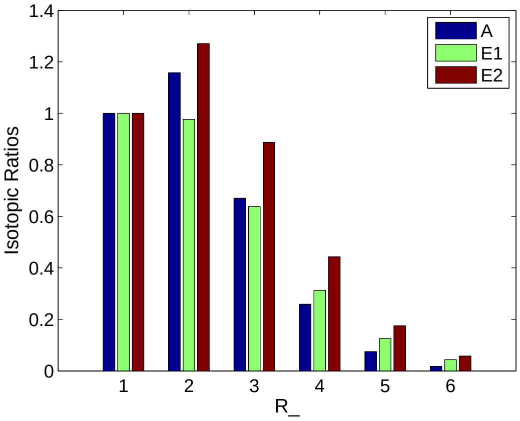


**Figure 6.3:** The three sets of isotopic distributions used in the simulation study.

proportions of heavy-oxygen water impurities were assumed to be equal to $p_{16} = 2\%$ and $p_{17} = 1\%$. For this simulation study, we considered as in the bovine cytochrome C data sets, six technical replicates. Possible variability due to, e.g., laser fluctuations and inefficient crystallization, was simulated by using six different reference intensities, namely, $H_1 = 18000, H_2 = 20000, H_3 = 23000, H_4 = 21000, H_5 = 19000$, and $H_6 = 22500$.

In the simulation study, the data sets were generated with combinations of settings for different parameters shown as below:

$$\begin{array}{ll} Q: & \{0.5 \quad 1 \quad 2\} \\ \lambda: & \{0.02 \quad 0.04 \quad 0.10\} \\ \sigma: & \{0.05 \quad 1.50\} \\ \boldsymbol{R}: & \{\mathbf{A} \quad \mathbf{E1} \quad \mathbf{E2}\} \\ \theta: & 0.6 \end{array}$$

500 data sets were generated for each setting. A graphical representation of the various settings of the simulation for the spectrum with reference intensity $H_1 = 18000$ is shown in Figures 6.4 to 6.6.

For comparison purposes, both heteroscedastic model, defined by (6.1)–(6.2) and (5.7)–(5.8), and homoscedastic model, defined by (5.1)–(5.4) and (5.7)–(5.8), were applied to the simulated data. For comparison purposes, the heteroscedastic model was evaluated by log-profile-likelihood (PLK), by using equation (6.5) and by PL-GLS (see Section 6.2) algorithm. The homoscedastic model, based on equations (5.1)–(5.4) and (5.7)–(5.8), was evaluated by the least squares (LS) and maximum-likelihood (ML).

### 6.4.1 Simulation with fixed reaction time $\tau = 120$(minutes)

For the homoscedastic model, as ML and LS showed very similar results, only the results of LS are shown. As the isotopic ratios are estimated very precisely and close to their true values, these results are not discussed here. Thus, we mainly focus on the results of the estimation of $Q$, $\lambda$, $\theta$ and $\sigma$. The summary statistics of the parameters are shown in Tables A.1 to A.12 (Appendix A). The mean relative bias is defined as the percentage of bias with respect to the scale of the true value for the parameter, i.e., $\bar{b} = \left(\hat{\beta} - \beta\right)/\beta = \hat{\beta}/\beta - 1$. The empirical variance $S^2_{\text{emp}}$ denotes the variance of estimates for the 500 replicated data sets. The model-based variance $S^2_{\text{mb}}$ is the mean of the model-based variances for the 500 data sets. Since most of the parameters were estimated based on transformations, the Delta method was used to transform these model-based variances of the parameter estimates into their original scales. The mean squared error indicates the efficiency of a parameter estimate and is equal to the sum of the squared bias and the empirical variance, i.e., MSE= $(\hat{\beta} - \beta)^2 + S^2_{\text{emp}}$.

Figures 6.7 and 6.8 show the estimation results for $\lambda$. When $\lambda$ was equal to 0.10 and $\sigma$ was equal to 1.50, for most of the data sets, the optimization algorithms didn't converge. Thus, the corresponding results were excluded from the plots. The reason for the non-convergence when $\lambda = 0.1$ will be explained in Section 6.4.2. The figures show that the MSE for $\lambda$ decreases with the increase of $Q$, indicating that a more abundant labeled peptide leads to more precise estimates for $\lambda$. Furthermore, a larger $\lambda$ results in a larger MSE.

Figures 6.9 to 6.10 show the MSE for $Q$ by including only the data sets that converged for $\lambda = 0.10$. The figures indicate that a smaller $Q$ or a larger $\lambda$ leads to a smaller MSE for $Q$. All the figures also show that the MSE is consistently smaller for the heteroscedastic model, fitted by PLK and PL-GLS, than for the homoscedastic model, fitted by LS. The results for the heteroscedastic model, using the two optimization approaches, look almost identical in most of the figures.

Figures 6.11 to 6.12 show that, in general, the MSE for $\theta$ increases as $\lambda$ or $Q$ increases.

### 6.4.2 Non-estimability for $\lambda$

The simulation study shows that bias of $\lambda$ estimates increases as the true value of $\lambda$ increases. This can be explained by Figure 5.2. When $\lambda$ is greater than 0.09, it becomes very difficult to estimate, because any value in the interval (0.09, 0.167) yields almost the same shift probabilities. Consequently, it becomes difficult to distinguish between different values of $\lambda$ based on data. The convergence problem became more serious when the residual variance was larger (when $\sigma = 1.50$) as the information for estimating the parameter would very likely be buried under the noise.

To investigate the influence of the incomplete labeling on the non-estimability, the simulations were repeated by using the same settings for all the other parameters, but with four possible reaction times: 24, 48, 90, and 120, and $\lambda = 0.10$.

The estimates of parameters, other than $\lambda$, were almost identical to the results obtained in the previous simulation study. Figures 6.13 and 6.14 show the MSE for $\lambda$ for the new simulation. Again, due to non-convergence for $\tau = 120$, the corresponding results were not included. The simulation study ascertains that, for shorter reaction times $\tau$, $\lambda$ can be estimated. This means that to solve the problem of non-estimability for $\lambda$, shortening the labeling (reaction) time can be considered.

An alternative solution is to perform a two-stage estimation approach.

To be more specific, the estimation of $\lambda$ can be separated from the other parameters. To perform the analysis, in the first stage, a sufficient number of grid points for $\lambda$ is chosen. The other parameters are estimated by maximizing the likelihood function, given the value of $\lambda$ at a grid point. The value of $\lambda$, which gives the maximum of values of the likelihood function over all the grid points, is chosen as the estimate of $\lambda$. At the second stage, $\lambda$ is treated as a fixed value by using the estimate from the first stage and the other parameters are estimated conditional on the fixed value of $\lambda$.

From Table 6.3, it can be observed that larger $Q$ or smaller $\sigma$ leads to smaller MSE for $\lambda$. Table 6.4 indicates that the MSE for $Q$ increases for larger values of $Q$ or $\sigma$. Moreover, it can be seen from Table 6.5 that the MSE for $\theta$ increases with the increase of $Q$. These findings are all consistent with the results obtained from the one-stage analysis of the simulations (see Section 6.4.1).

Through simulation study based on the settings in which there was non-estimability issue for $\lambda$, two-stage analysis solved the problem. Moreover, parameter estimates with a smaller MSE were obtained (shown in Tables 6.3 to 6.6).

## 6.5 Discussion

We have presented an extension of the model, proposed by Valkenborg (2008), which takes into account the (mean-dependent) heteroscedastic nature of the residual errors. We implemented the model by using a PL-GLS algorithm, which is more robust than the direct likelihood maximization.

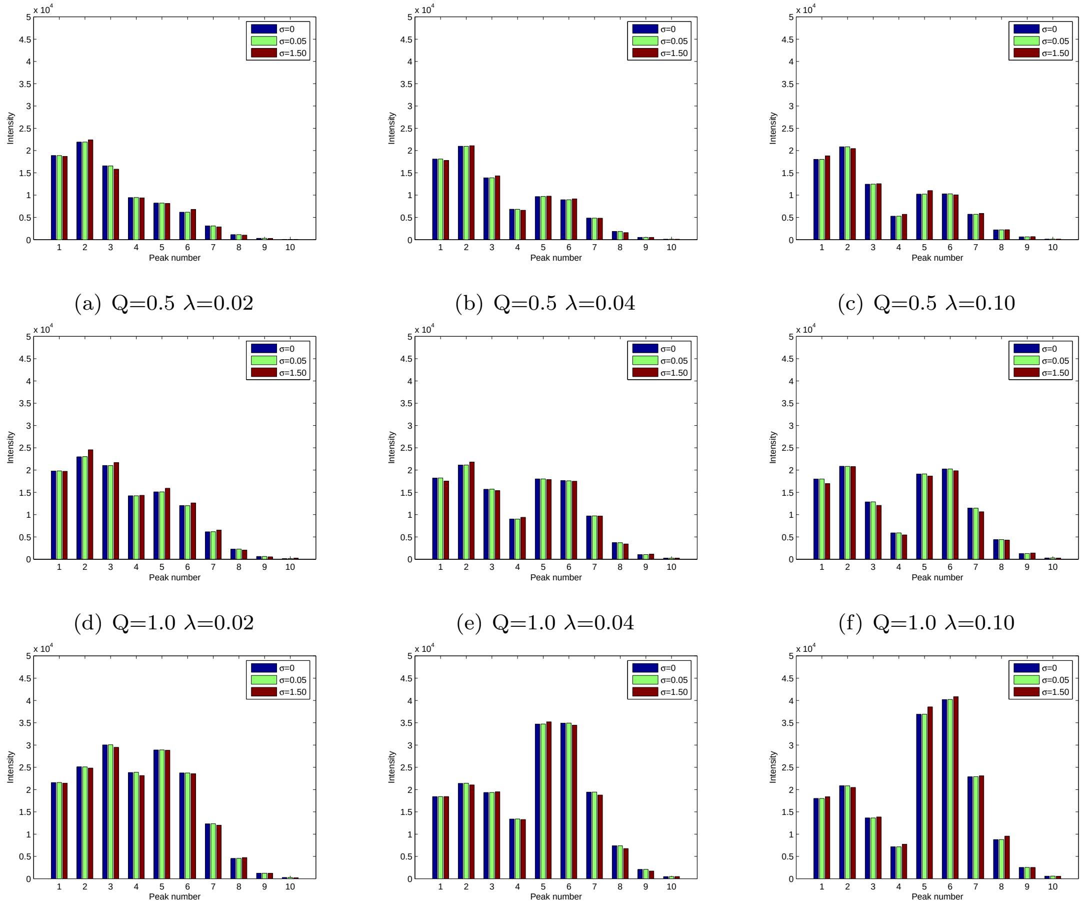
The results of the application to the real-life data were, in general, consistent with the true parameter values for two of three analyzed peptides. The consideration of the heteroscedastic residual variance leads to a precision gain for the parameter of interest.

In the simulation study, the parameters of the heteroscedastic model were well estimated. For the homoscedastic model, although the bias was also very small, the variances were slightly larger, resulting in larger mean squared errors for almost all the settings. This indicates that the misspecification of the variance function will lead to a precision loss. The relative abundance parameter was estimated with a better precision when it was smaller or when the labeling was more complete (for larger $\lambda$). Moreover, a more abundant labeled peptide sample leads to more precise estimates for the oxygen incorporation rate $\lambda$. This is reasonable, because the information of the estimation of $\lambda$ comes mainly from the labeled peptide. All the parameters were more precisely estimated when the amount of random noise was smaller.

It is also worth noting that non-estimability for $\lambda$ may occur when labeling is complete ($\lambda \geq 0.09$ for reaction time $\tau = 120$ minutes), and when the amount of residual error increases. Such problem can be solved either by shortening the reaction time period, i.e., by decreasing $\tau$ when conducting the experiment, or by performing a two-stage analysis (see Section 6.4.2). Note, however, that the more complete the labeling is, the more precise estimates of the relative abundance can be produced. Thus, there is a trade-off between the estimation of $\lambda$ and $Q$.

Numerical complexity of the developed methodology is low. On average, fitting the model for each peptide, presented in Tables 6.1 and 6.2, took about 1.5s on a HP8530p laptop using Matlab 2009a under Windows Vista®. Thus, upon the automation of the selection of peak-clusters for fitting the model, the method can be used in a high-throughput environment.

Several extensions of the proposed methodology are possible. For instance, different residual variance functions can be used (Davidian and Giltinan 1995). Inclusion of random effects, which would allow estimating, e.g., the between-sample biological variability, is possible. This type of extension will be dealt with in the next chapter.



(a) Q=0.5 $\lambda$=0.02

(b) Q=0.5 $\lambda$=0.04

(c) Q=0.5 $\lambda$=0.10

(d) Q=1.0 $\lambda$=0.02

(e) Q=1.0 $\lambda$=0.04

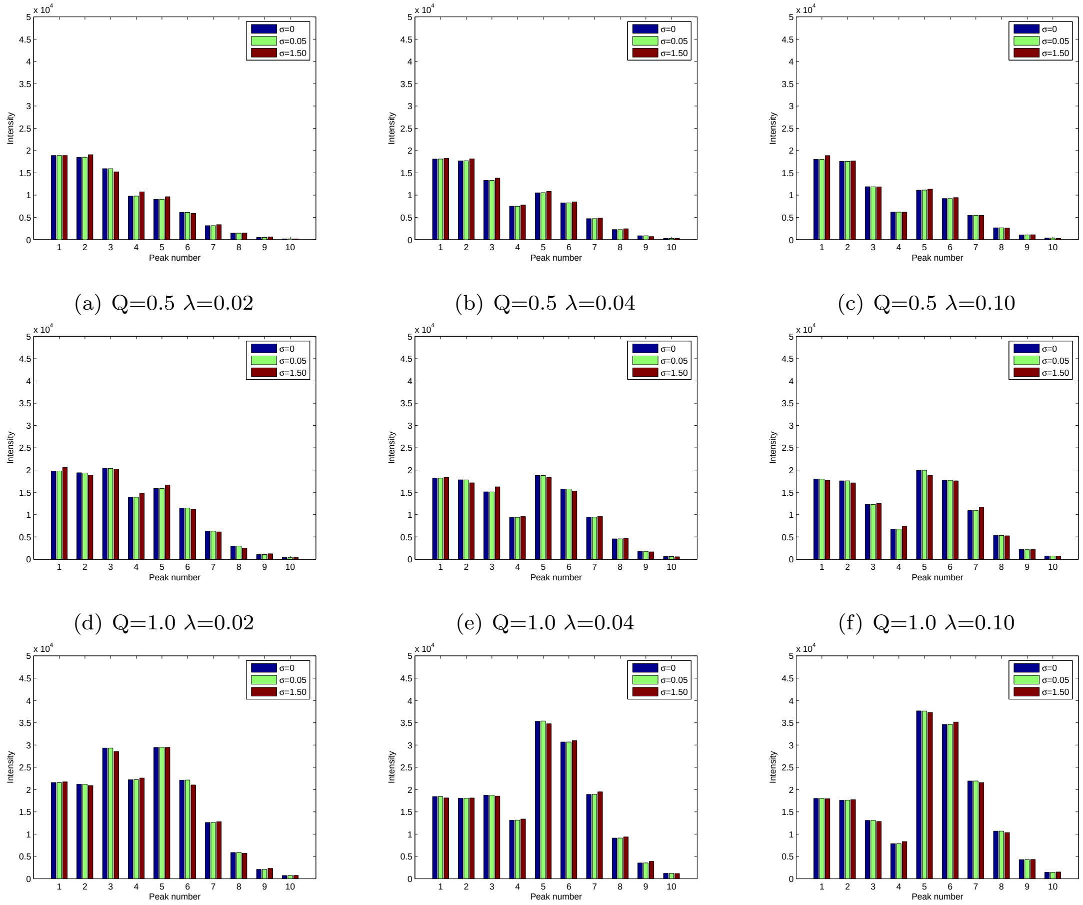
(f) Q=1.0 $\lambda$=0.10

(g) Q=2.0 $\lambda$=0.02

(h) Q=2.0 $\lambda$=0.04

(i) Q=2.0 $\lambda$=0.10

**Figure 6.4:** Graphical representation for simulation settings with **A** ratios and $H_1$ = 18000.

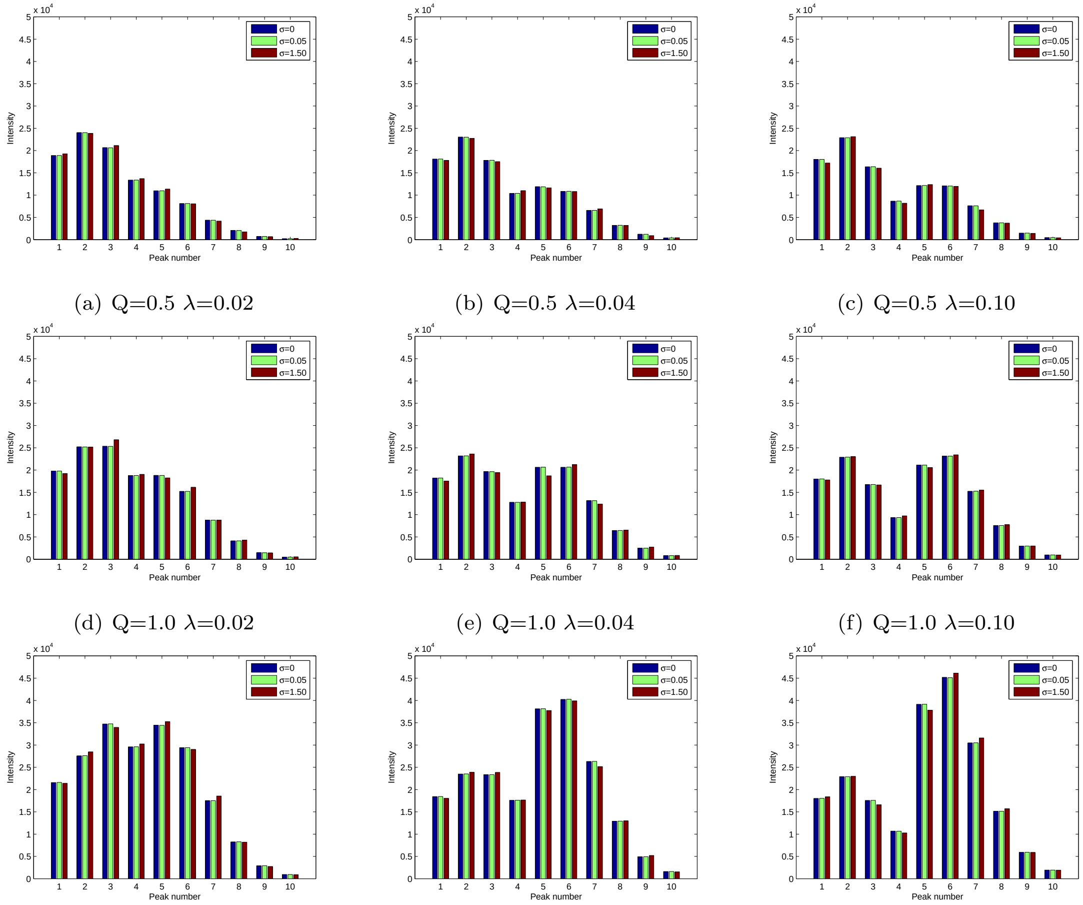


(a) Q=0.5 $\lambda$=0.02 (b) Q=0.5 $\lambda$=0.04 (c) Q=0.5 $\lambda$=0.10

(d) Q=1.0 $\lambda$=0.02 (e) Q=1.0 $\lambda$=0.04 (f) Q=1.0 $\lambda$=0.10

(g) Q=2.0 $\lambda$=0.02 (h) Q=2.0 $\lambda$=0.04 (i) Q=2.0 $\lambda$=0.10

**Figure 6.5:** Graphical representation for simulation settings with **E1** ratios and $H_1$ = 18000.



(a) Q=0.5 $\lambda$=0.02

(b) Q=0.5 $\lambda$=0.04

(c) Q=0.5 $\lambda$=0.10

(d) Q=1.0 $\lambda$=0.02

(e) Q=1.0 $\lambda$=0.04

(f) Q=1.0 $\lambda$=0.10

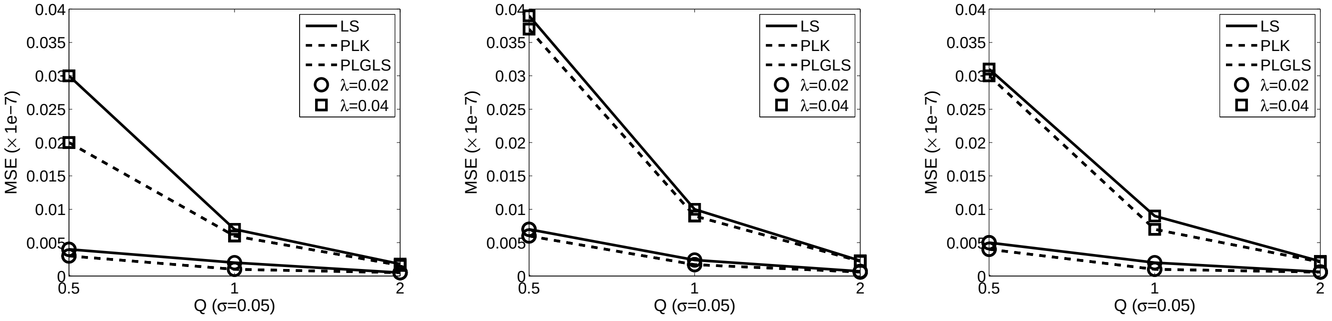
(g) Q=2.0 $\lambda$=0.02

(h) Q=2.0 $\lambda$=0.04

(i) Q=2.0 $\lambda$=0.10

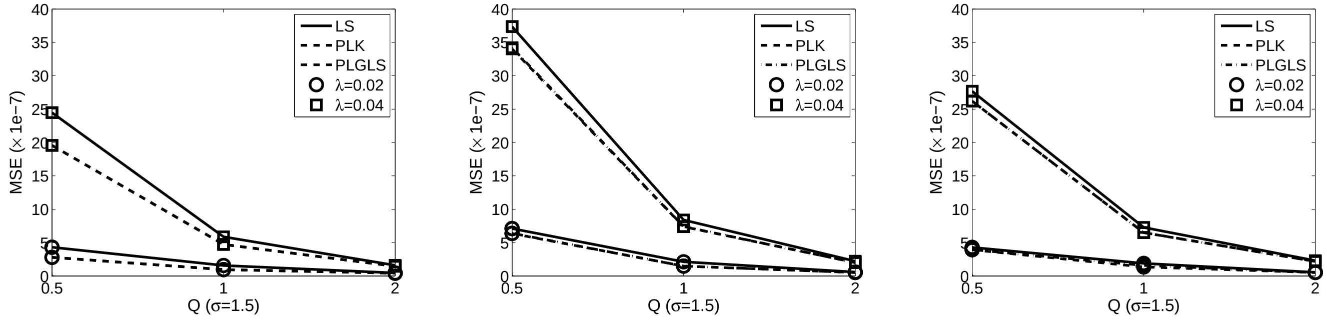
**Figure 6.6:** Graphical representation for simulation settings with **E2** ratios and $H_1$ = 18000.

## Results of the simulation with fixed reaction time $\tau = 120$(minutes)



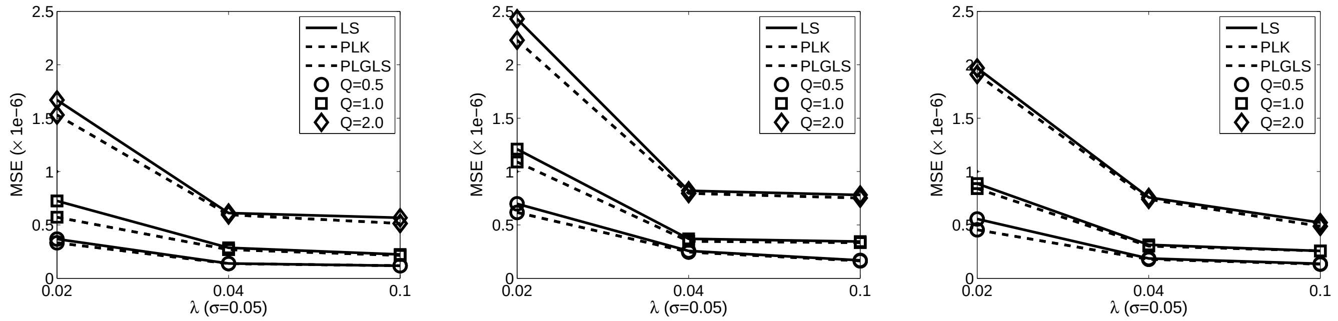
(a) **A** (b) **E1** (c) **E2**

**Figure 6.7:** Graphical representation of the MSE of $\lambda$ for settings with $\sigma = 0.05$ and $\tau = 120$ (excluding $\lambda = 0.10$).



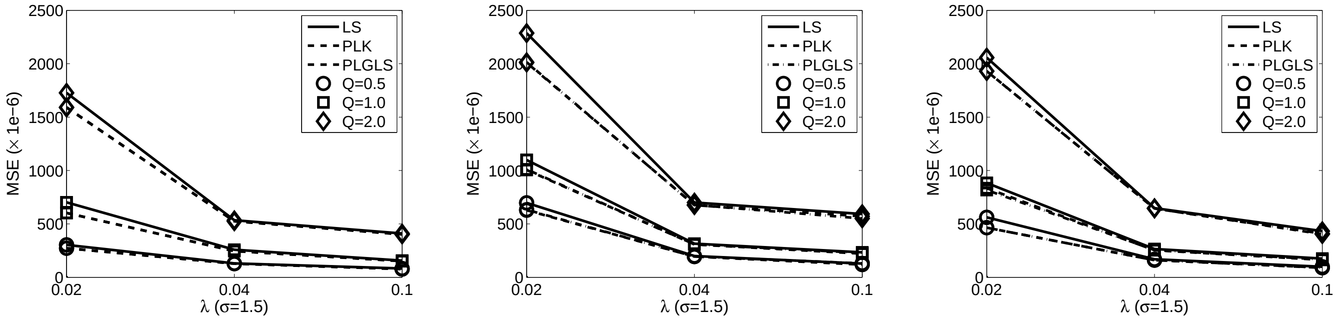
(a) **A** (b) **E1** (c) **E2**

**Figure 6.8:** Graphical representation of the MSE of $\lambda$ for settings with $\sigma = 1.50$ and $\tau = 120$ (excluding $\lambda = 0.10$).



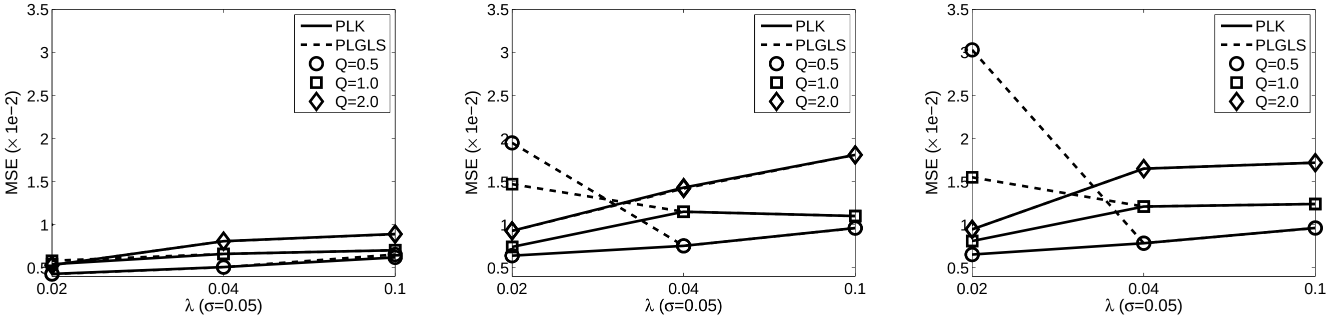
(a) **A** (b) **E1** (c) **E2**

**Figure 6.9:** Graphical representation of the MSE of $Q$ for settings with $\sigma = 0.05$ ($\tau = 120$).



(a) **A** (b) **E1** (c) **E2**

**Figure 6.10:** Graphical representation of the MSE of $Q$ for settings with $\sigma = 1.50$ ($\tau = 120$).

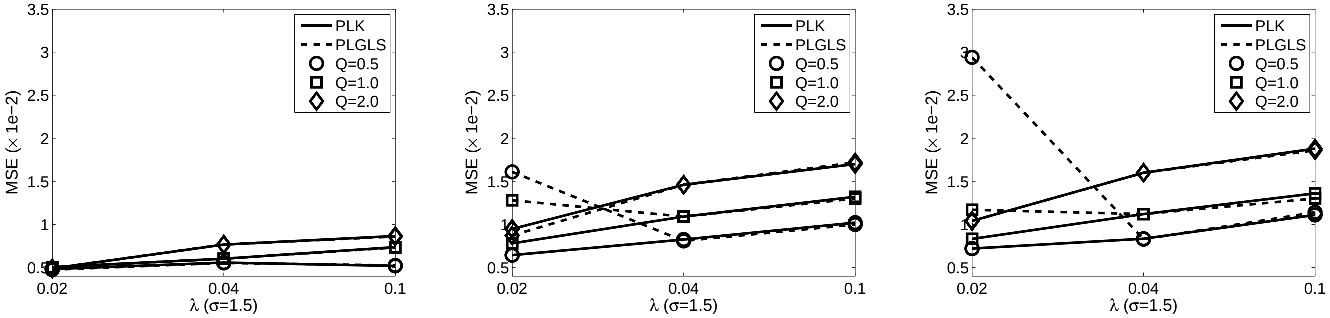


(a) **A**

(b) **E1**

(c) **E2**

**Figure 6.11:** Graphical representation of the MSE of $\theta$ for settings with $\sigma = 0.05$ ($\tau = 120$).



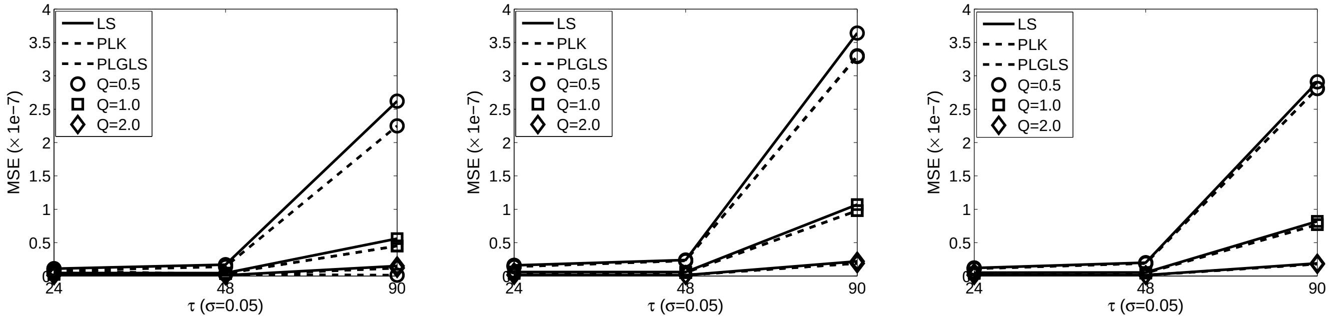
(a) **A**

(b) **E1**

(c) **E2**

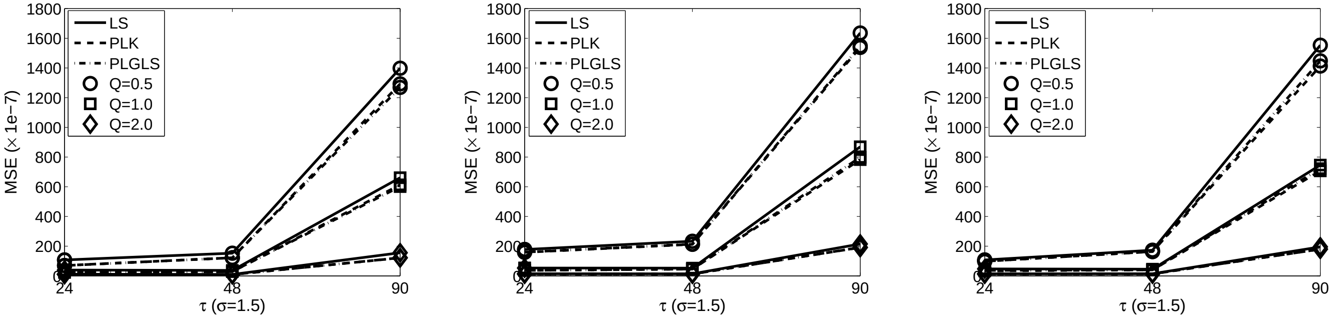
**Figure 6.12:** Graphical representation of the MSE of $\theta$ for settings with $\sigma = 1.50$ ($\tau = 120$).

## Results of the simulation for various values of $\tau$



(a) **A** (b) **E1** (c) **E2**

**Figure 6.13:** Graphical representation of the MSE of $\lambda$ for settings with $\sigma = 0.05$ for various values of $\tau$ (excluding $\tau = 120$).



(a) **A** (b) **E1** (c) **E2**

**Figure 6.14:** Graphical representation of the MSE of $\lambda$ for settings with $\sigma = 1.50$ for various values of $\tau$ (excluding $\tau = 120$).

## Simulation results of the two-stage analysis:

**Table 6.3:** Mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$ and MSE of $\lambda$ (From the first-stage analysis).

| $\boldsymbol{R}$ | $Q$ | $\sigma$ | $\bar{b}$ ($\times 1e-3$) | $S^2_{\text{emp}}$ ($\times 1e-7$) | MSE ($\times 1e-7$) |
|---|---|---|---|---|---|
| **A** | 0.5 | 0.05 | 0.14 | 24.55 | 24.55 |
| | | 1.50 | -109.6 | 2485 | 3685 |
| | 1.0 | 0.05 | 0.58 | 6.04 | 6.07 |
| | | 1.50 | -12.58 | 2510 | 2526 |
| | 2.0 | 0.05 | 0.14 | 1.78 | 1.78 |
| | | 1.50 | 22.87 | 1453 | 1506 |
| **E1** | 0.5 | 0.05 | 1.14 | 37.85 | 37.98 |
| | | 1.50 | -142.4 | 2642 | 4670 |
| | 1.0 | 0.05 | 0.18 | 9.64 | 9.64 |
| | | 1.50 | -44.25 | 2376 | 2572 |
| | 2.0 | 0.05 | -0.02 | 2.34 | 2.34 |
| | | 1.50 | 20.79 | 1900 | 1943 |
| **E2** | 0.5 | 0.05 | 1.70 | 33.28 | 33.57 |
| | | 1.50 | -114.3 | 2798 | 4105 |
| | 1.0 | 0.05 | 0.26 | 8.71 | 8.72 |
| | | 1.50 | -51.55 | 1968 | 2234 |
| | 2.0 | 0.05 | -0.16 | 2.40 | 2.40 |
| | | 1.50 | 12.39 | 1673 | 1688 |

**Table 6.4:** Mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$, mean model based variance $S^2_{\text{mb}}$ and MSE of $Q$ (From the second-stage analysis).

| $\boldsymbol{R}$ | $Q$ | $\sigma$ | $\bar{b}$ ($\times 1e-5$) | $S^2_{\text{emp}}/S^2_{\text{mb}}$ ($\times 1e-6$) | MSE ($\times 1e-6$) |
|---|---|---|---|---|---|
| **A** | 0.5 | 0.05 | 2.14 | 0.107/0.021 | 0.108 |
| | | 1.50 | 442.1 | 63.41/18915 | 68.29 |
| | 1.0 | 0.05 | -2.86 | 0.229/0.016 | 0.230 |
| | | 1.50 | 177.3 | 155.3/13014 | 158.4 |
| | 2.0 | 0.05 | -0.73 | 0.572/0.017 | 0.572 |
| | | 1.50 | 65.18 | 414.8/12841 | 416.5 |
| **E1** | 0.5 | 0.05 | 0.11 | 0.155/0.099 | 0.155 |
| | | 1.50 | 460.8 | 94.20/39921 | 99.51 |
| | 1.0 | 0.05 | -1.78 | 0.325/0.026 | 0.325 |
| | | 1.50 | 221.5 | 186.9/34066 | 191.8 |
| | 2.0 | 0.05 | -1.02 | 0.729/0.117 | 0.730 |
| | | 1.50 | 116.0 | 568.2/64795 | 573.6 |
| **E2** | 0.5 | 0.05 | -1.04 | 0.122/0.058 | 0.122 |
| | | 1.50 | 413.6 | 79.83/43410 | 841.0 |
| | 1.0 | 0.05 | -1.59 | 0.243/0.030 | 0.244 |
| | | 1.50 | 198.3 | 164.7/24655 | 168.7 |
| | 2.0 | 0.05 | -1.07 | 0.616/0.066 | 0.617 |
| | | 1.50 | 96.47 | 422.7/54702 | 426.5 |

**Table 6.5:** Mean estimate $\bar{\theta}$, mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$, mean model based variance $S^2_{\text{mb}}$ and MSE of $\theta$ (From the second-stage analysis).

| $\boldsymbol{R}$ | $Q$ | $\sigma$ | $\bar{\theta}$ | $\bar{b}(\times 1e-2)$ | $S^2_{\text{emp}}/S^2_{\text{mb}}$ $(\times 1e-2)$ | MSE$(\times 1e-2)$ |
|---|---|---|---|---|---|---|
| **A** | 0.5 | 0.05 | 0.5926 | -1.23 | 0.540/12.94 | 0.545 |
| | | 1.50 | 0.5897 | -1.71 | 0.622/13.15 | 0.632 |
| | 1.0 | 0.05 | 0.5929 | -1.19 | 0.645/15.54 | 0.650 |
| | | 1.50 | 0.5969 | -0.510 | 0.747/15.74 | 0.748 |
| | 2.0 | 0.05 | 0.5977 | -0.386 | 0.892/20.12 | 0.893 |
| | | 1.50 | 0.5969 | -0.512 | 0.833/19.79 | 0.834 |
| **E1** | 0.5 | 0.05 | 0.5881 | -1.99 | 0.968/20.72 | 0.983 |
| | | 1.50 | 0.5783 | -3.62 | 0.921/21.40 | 0.969 |
| | 1.0 | 0.05 | 0.6051 | 0.853 | 1.14/27.33 | 1.15 |
| | | 1.50 | 0.5888 | -1.87 | 1.16/27.34 | 1.17 |
| | 2.0 | 0.05 | 0.5830 | -2.84 | 1.77/36.11 | 1.80 |
| | | 1.50 | 0.5789 | -3.51 | 1.64/35.46 | 1.68 |
| **E2** | 0.5 | 0.05 | 0.5863 | -2.28 | 0.964/22.58 | 0.983 |
| | | 1.50 | 0.5819 | -3.02 | 1.01/23.19 | 1.04 |
| | 1.0 | 0.05 | 0.5980 | -0.329 | 1.20/29.35 | 1.20 |
| | | 1.50 | 0.5987 | -0.220 | 1.17/30.13 | 1.17 |
| | 2.0 | 0.05 | 0.5929 | -1.18 | 1.80/40.21 | 1.81 |
| | | 1.50 | 0.5877 | -2.05 | 1.74/40.26 | 1.75 |

**Table 6.6:** Mean estimate $\bar{\sigma}$, mean relative bias $\bar{b}$ ,empirical variance $S^2_{\text{emp}}$ and MSE of $\sigma$ (From the second-stage analysis).

| ***R*** | $\sigma$ | $Q$ | $\bar{\sigma}$ | $\bar{b}$ | $S^2_{\text{emp}}(\times 1e-2)$ | MSE$(\times 1e-2)$ |
|---|---|---|---|---|---|---|
| **A** | 0.05 | 0.5 | 0.065 | 0.293 | 0.232 | 0.253 |
| | | 1.0 | 0.069 | 0.373 | 0.266 | 0.301 |
| | | 2.0 | 0.074 | 0.476 | 0.462 | 0.518 |
| | 1.50 | 0.5 | 2.039 | 0.360 | 178.8 | 207.9 |
| | | 1.0 | 2.043 | 0.362 | 251.5 | 280.9 |
| | | 2.0 | 2.184 | 0.456 | 382.9 | 429.7 |
| **E1** | 0.05 | 0.5 | 0.082 | 0.650 | 1.806 | 1.911 |
| | | 1.0 | 0.073 | 0.454 | 0.651 | 0.702 |
| | | 2.0 | 0.125 | 1.495 | 5.959 | 6.518 |
| | 1.50 | 0.5 | 2.484 | 0.656 | 425.5 | 522.2 |
| | | 1.0 | 2.638 | 0.759 | 961.9 | 1091 |
| | | 2.0 | 3.696 | 1.464 | 3082 | 3564 |
| **E2** | 0.05 | 0.5 | 0.083 | 0.653 | 1.018 | 1.125 |
| | | 1.0 | 0.082 | 0.640 | 0.836 | 0.938 |
| | | 2.0 | 0.114 | 1.281 | 3.252 | 3.662 |
| | 1.50 | 0.5 | 2.582 | 0.722 | 629.1 | 746.3 |
| | | 1.0 | 2.431 | 0.621 | 683.8 | 770.5 |
| | | 2.0 | 3.592 | 1.395 | 2933 | 3271 |

# Chapter 7

# A frequentist approach for the analysis of $^{18}$O-labeled mass spectra using a heteroscedastic random-effect Markov-chain-based model

MS data can be subject to technical and/or biological variability. *Technical variability* is related to the between-spectra variability of intensity measurements even for the same sample. *Biological variability* is related to the variability of measurements for different biological samples.

It is worth noting that, the estimation of different sources of variability in the context of mass spectrometry data, has never been addressed in the methods proposed for the analysis of the $^{18}$O-labeled mass spectrometry experiments. In this chapter, we present a heteroscedastic regression model with random effects, in the frequentist framework, as a method to account for the between-spectra variability.

## 7.1 Model formulation

The model formulation is similar to the one defined in (6.1)–(6.2) and (5.7)–(5.8), modified by including a spectrum-specific relative abundance parameter $Q_i$ for the $i$th spectrum. More specifically, consider a peptide, which has $l \geq 5$ isotopic variants (including the monoisotopic variant). The observed peak intensity $y_{ij}$ of the $j$th observed peak, where $j = 1, 2, \ldots$, in the $i$th mass spectrum is defined as

$$\begin{aligned} y_{ij} &= \mu_{ij} + \varepsilon_{ij}, && (7.1)\\ \text{where} \quad \varepsilon_{ij} &\sim N(0, \sigma^2 \mu_{ij}^{2\theta}), && (7.2) \end{aligned}$$

and $\varepsilon_{ij}$'s are independent. The mean intensity $\mu_{ij}$ of the $j$th peak in the $i$th spectrum is expressed as follows:

$$\mu_{ij} \equiv \mathrm{E}(y_{ij}) = \begin{cases} H_i R_j + Q_i H_i \sum_{k=0}^{min(4,j-1)} P_k R_{j-k} & \text{if } 1 \leq j \leq l, \\ Q_i H_i \sum_{k=j-l}^{4} P_k R_{j-k} & \text{if } l+1 \leq j \leq l+4. \end{cases} \qquad (7.3)$$

We assume that $H_i$ and $Q_i$ are random, i.e.,

$$\begin{pmatrix} H_i \\ Q_i \end{pmatrix} \sim N\left( \begin{pmatrix} H \\ Q \end{pmatrix}, \begin{pmatrix} \sigma_H^2 & \sigma_{HQ} \\ \sigma_{HQ} & \sigma_Q^2 \end{pmatrix} \right). \qquad (7.4)$$

The parameters of interest are $Q$, $\sigma_Q^2$ and $\sigma_H^2$. Parameters $Q$ and $\sigma_Q^2$ capture, respectively, the (mean) relative abundance of the peptide in the two peptide samples and biological variability across different spectra. Parameter $\sigma_H^2$ is related to the technical variability of intensity measurements across the different spectra. It is worth noting that for technical replicates of mass spectra, there is no need to account for the variability of the relative abundance. In such case, $Q_i$, $Q$ and $\sigma_Q^2$ can be replaced by a single fixed-effect $Q$ and (7.4) degenerates to a univariate normal distribution for $H_i$.

## 7.2 Estimation and inference

### 7.2.1 Estimation approach for the homoscedastic regression model

As an initial step, the model was implemented with homoscedastic residual variance. The optimization approach for the homoscedastic random effects model is based on maximizing the marginal likelihood function by 'eliminating' the random parameters $H_i$ and/or $Q_i$. This can be done by integrating out these random effects over their assumed distributions.

More specifically, assuming that $n$ joint spectra are available, define $\boldsymbol{H} = (H_1, \ldots, H_n)$ and $\boldsymbol{Q} = (Q_1, \ldots, Q_n)$. For the homoscedastic model, with the mean structure specified by (7.3)–(7.4), the joint likelihood is given by

$$L_{\text{Marginal}}(\boldsymbol{\beta}, \sigma^2) = \int\int L_{\text{ML}}(\boldsymbol{\beta}, \sigma^2) F(\boldsymbol{H}, \boldsymbol{Q}) d(\boldsymbol{H}, \boldsymbol{Q}), \tag{7.5}$$

with the corresponding log-likelihood of $L_{\text{ML}}(\boldsymbol{\beta}, \sigma^2)$, omitting the constant terms, given as

$$l_{\text{ML}}(\boldsymbol{\beta}, \sigma^2) = -\sum_{i=1}^{n}\sum_{j=1}^{m} \log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} \left[y_{ij} - \mu_{ij}(\boldsymbol{\beta})\right]^2. \tag{7.6}$$

In (7.5), $F(\boldsymbol{H}, \boldsymbol{Q})$ denotes the product of the density of the assumed joint distribution for the random effects $H_i$ and $Q_i$, specified by (7.4). Vector $\boldsymbol{\beta}$ contains all the parameters used to model the mean value $\mu_{ij}$, i.e., $\boldsymbol{\beta} = (H_1, \ldots, H_n, Q_1, \ldots, Q_n, \lambda, R_1, \ldots, R_{m-4})$ .

As the integral, shown in (7.5), has no closed form, a solution is to use numerical approximation of the integral. For this purpose, an adaptive quadrature function is used. The advantage of using an adaptive quadrature function is that it adapts the integral such that more weight is assigned to the parameter space where the distribution of the random effects is concentrated (Molenberghs and Verbeke 2005). Consequently, the adaptive quadrature function leads to more accurate approximation of the integral.

The corresponding REML-likelihood becomes

$$L_{\text{REML}}(\boldsymbol{\beta}, \sigma^2) = \left|\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{X}_i\right|^{1/2} \times \left|\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i\right|^{-1/2} \times L_{\text{Marginal}}(\boldsymbol{\beta}, \sigma^2), \tag{7.7}$$

where $\boldsymbol{X}_i$ is a design matrix for the $i$th spectrum, involving the shift probabilities $\boldsymbol{P}$ (defined in Secction 5.2.2). Thus, matrix $\boldsymbol{X}_i$ contains only one unknown parameter ($\lambda$). (For linear mixed models when no unknown parameter is involved in the design matrix $\boldsymbol{X}_i$, the term $\left|\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{X}_i\right|$ becomes a constant and can then be ignored from the REML-likelihood function.) Matrix $\boldsymbol{W}_i$ is the inverse of the covariance matrix $\boldsymbol{V}_i$, where $\boldsymbol{V}_i$ is the variance-covariance matrix for the intensity values of the $i$th spectrum. Matrix $\boldsymbol{V}_i$ contains the sum of the residual variance $\sigma^2$ and variance of the random effects on the main diagonal and only the variance of the random effects on the off-diagonal.

### 7.2.2 Estimation approach for the heteroscedastic regression model

Similar to equation (7.5), the joint likelihood for the heteroscedastic model (7.1)–(7.4) is given by

$$L_{\text{Marginal}}(\boldsymbol{\beta}, \theta, \sigma^2) = \int\int L_{\text{ML}}(\boldsymbol{\beta}, \theta, \sigma^2) F(\boldsymbol{H}, \boldsymbol{Q}) d(\boldsymbol{H}, \boldsymbol{Q}), \tag{7.8}$$

where the corresponding log-likelihood of $L_{\text{ML}}(\boldsymbol{\beta}, \theta, \sigma^2)$, omitting the constant terms, is in the same form as equation (6.3). The only difference is that the vector $\boldsymbol{\beta}$ now contains spectrum-specific relative abundance parameters, i.e., $(Q_1, \ldots, Q_n)$.

To maximize the marginal likelihood shown in (7.8), an adaptive quadrature function can be considered. The maximization of the marginal likelihood, shown in (7.8), is numerically complex and is very sensitive to the choice of initial values. An alternative is to perform a two-stage analysis, described by Davidian and Giltinan (1995). The first stage of the analysis entails a PL-GLS algorithm for the estimation of the mean-dependent variance function, by treating all the effects as fixed ones. The second stage produces the estimates of the random effect(s) based on the individual estimates obtained from the first stage.

More specifically, for the first stage, the PL-GLS algorithm is the same as described in Section 6.2, but with the spectrum-specific relative abundance parameter $Q_i$ included.

At the second stage, the spectrum-specific parameter estimates $\widehat{H}_i$ and $\widehat{Q}_i$, estimated from the first stage, are treated as "data" to obtain the estimates for $H$, $\sigma_H^2$, $Q$ and $\sigma_Q^2$.

More specifically, the following model is used:

$$(\widehat{\boldsymbol{\eta}}_{\boldsymbol{i}}) \sim N\left(\boldsymbol{\eta}, \boldsymbol{D} + \boldsymbol{C}_{\boldsymbol{i}}\right). \tag{7.9}$$

where $\widehat{\boldsymbol{\eta}}_{\boldsymbol{i}} = \begin{pmatrix} \widehat{H}_i \\ \widehat{Q}_i \end{pmatrix}$, $\boldsymbol{\eta} = \begin{pmatrix} H \\ Q \end{pmatrix}$ and $\boldsymbol{D} = \begin{pmatrix} \sigma_H^2 & \sigma_{HQ} \\ \sigma_{HQ} & \sigma_Q^2 \end{pmatrix}$. Matrix $\boldsymbol{C}_i$ is the variance-covariance matrix for $\widehat{H}_i$ and $\widehat{Q}_i$ obtained from the first stage. The inclusion of $\boldsymbol{C}_i$ accounts for the variability of the estimates of $\widehat{H}_i$ and $\widehat{Q}_i$ around their true values and thus can be deemed as the uncertainty of the within-subject (spectra) variability for small-sample applications. We now wish to estimate $\boldsymbol{\eta}$ and $\boldsymbol{D}$ as the mean and variance-covariance matrix of the random effects. A straightforward way for the estimation is to maximize the log-likelihood, resulting from (7.9). Ignoring constant terms, the log-likelihood is

$$l(\boldsymbol{\eta}, \boldsymbol{D}) = -\sum_{i=1}^{n} \left[\log|\boldsymbol{D} + \boldsymbol{C}_i| + (\widehat{\boldsymbol{\eta}}_{\boldsymbol{i}} - \boldsymbol{\eta})^T (\boldsymbol{D} + \boldsymbol{C}_i)^{-1} (\widehat{\boldsymbol{\eta}}_{\boldsymbol{i}} - \boldsymbol{\eta})\right]. \tag{7.10}$$

Maximizing the log-likelihood function in (7.10) can be approached in two ways: by direct maximization of the function shown in (7.10), or by using the EM algorithm.

The EM algorithm is performed in three steps:

(i) Set $k = 0$ and obtain starting values as

$$\widehat{\boldsymbol{\eta}}_{(0)} = n^{-1}\sum_{i=1}^{n}\widehat{\boldsymbol{\eta}}_{\boldsymbol{i}}, \quad \widehat{\boldsymbol{D}}_{(0)} = (n-1)^{-1}\sum_{i=1}^{n} \left(\widehat{\boldsymbol{\eta}}_{\boldsymbol{i}} - \widehat{\boldsymbol{\eta}}_{(0)}\right)\left(\widehat{\boldsymbol{\eta}}_{\boldsymbol{i}} - \widehat{\boldsymbol{\eta}}_{(0)}\right)^T.$$

(ii) "E-step": set $k = k + 1$ and produce current empirical Bayes estimates of the $\boldsymbol{\eta}_{\boldsymbol{i}}$, $i = 1, \ldots, n$, given by

$$\tilde{\boldsymbol{\eta}}_{i,(k+1)} = \left(\widehat{\boldsymbol{D}}_k^{-1} + \boldsymbol{C}_i^{-1}\right)^{-1} \left(\boldsymbol{C}_i^{-1}\widehat{\boldsymbol{\eta}}_i + \widehat{\boldsymbol{D}}_k^{-1}\widehat{\boldsymbol{\eta}}_{(k)}\right).$$

(iii) "M-step": obtain updated estimates as

$$\widehat{\boldsymbol{\eta}}_{(k+1)} = n^{-1}\sum_{i=1}^{n}\tilde{\boldsymbol{\eta}}_{i,(k+1)},$$

$$\widehat{\boldsymbol{D}}_{(k+1)} = n^{-1}\sum_{i=1}^{n} \left(\widehat{\boldsymbol{D}}_{(k)}^{-1} + \boldsymbol{C}_i^{-1}\right)^{-1} + n^{-1}\sum_{i=1}^{n} \left(\tilde{\boldsymbol{\eta}}_{i,(k+1)} - \widehat{\boldsymbol{\eta}}_{(k+1)}\right)\left(\tilde{\boldsymbol{\eta}}_{i,(k+1)} - \widehat{\boldsymbol{\eta}}_{(k+1)}\right)^T.$$

(iv) Iterate between steps (ii)–(iii) until convergence.

For spectra with technical replicates, the vectors and matrices degenerate to the corresponding scalers and vectors for the random effect of $\boldsymbol{H}$.

Irrespectively of the estimation approaches used, standard errors of the estimates can be obtained from the inverse of negative Hessian of the marginal loglikelihood (7.8), computed at the estimated values of the parameters.

## 7.3 Results

In this section, we present results of the application of the model to both real-life and simulated data.

### 7.3.1 Application to bovine cytochrome C data set

In this section, we present an application of the model to the controlled experiment of the enzymatic labeling of bovine cytochrome C peptides (see Section 3.1).

**Random *H***

In this section, the results of the model by including only the random effect for $H$ (not for $Q$) are presented. This is reasonable, as the data set contains technical replicates with same biological (labeled and unlabeled) samples for different spectra, which should give the same value of $Q$. Practically, the estimation approaches were implemented by using Matlab 2009a. In particular, function $fminunc$ for unconstrained optimization problems was used. As an initial step, the homoscedastic model, with mean structure defined by (7.3)–(7.4), was fitted. In this respect, the model was estimated by maximizing the marginal likelihood via a numerical approximation of the integral using function *quadl* with adaptive Lobatto quadrature. As in Chapters 5 and 6, the proportions of water impurities of the heavy-oxygen water were assumed to be equal to $p_{16} = 2\%$ and $p_{17} = 1\%$. The true values of isotopic ratios $R_i$ were calculated from the atomic composition of the peptides by using the convolution method, developed by Rockwood (1995). As the duration of the experiment is not known, we estimated products $\lambda\tau$ instead of $\lambda$.

Tables 7.3 and 7.4 present results of the analysis of the three peptides for the spectra under the homoscedastic residual variance assumption, for which the intended value of relative abundance $Q$ was equal to 1/3 and 3/1, respectively. The results are mostly consistent with those shown in Chapters 5 and 6. For each peptide, a considerable amount of between-spectra variability of the intensity measurements, represented by the parameter $\sigma_H^2$, is worth noting. This, on the one hand, demonstrates the advantage of using the $^{18}$O-labeling strategy, since the variability can be removed from the comparison of the peptide abundance in the unlabeled and labeled samples. Moreover, it shows that a random effects model accounting for the between-spectra variability is useful.

The second analysis assumes heteroscedasticity for the residual variance (7.2). The analysis is based on the marginal log-likelihood (7.8), by using the point estimates obtained from the two-stage analysis, with the optimization approach presented in Section 7.2.2, as the initial values. The differences of the point estimates obtained from the analysis based on the marginal log-likelihood (7.8) and those of the two-stage analysis with direct log-likelihood (7.10) for the random effects and the EM algorithm were mostly at the magnitude of $10^{-7}$ to $10^{-11}$ and can therefore be ignored. Thus, we present only the results obtained from the marginal log-likelihood.

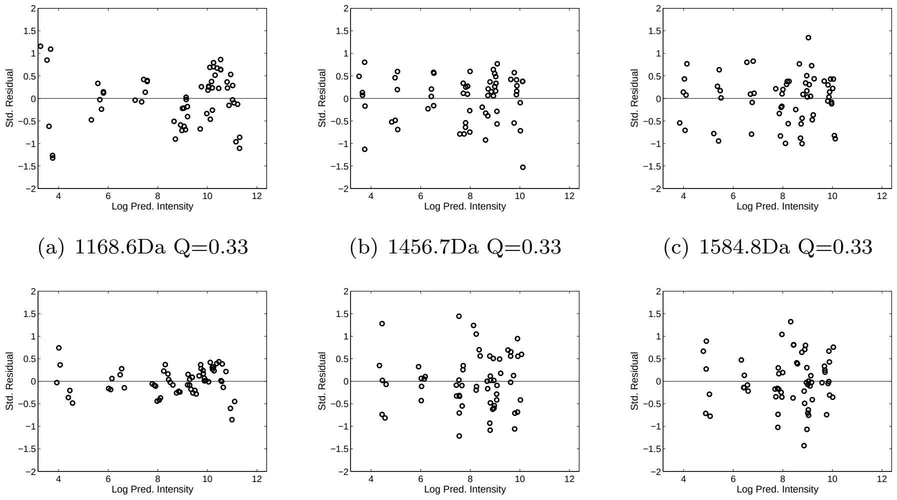
The results, shown in Tables 7.5 and 7.6 are, in general, consistent with the results of the homoscedastic model, shown in Tables 7.3 and 7.4.

The point estimates of $Q$ are in agreement, especially for the two peptides with masses 1456.7 and 1584.8 Da, while the standard errors for the heteroscedastic model are smaller for $Q = 1/3$. This indicates that the use of a (mean-)variance function results in a precision gain for the parameters in the mean structure.

It is also important to note that for peptide with mass 1168.6 Da, the point estimates of the isotopic ratios for the heteroscedastic model are much closer to the true values, although they are still more biased than for the other two peptides. Taking into account the standard errors of these parameter estimates, the heteroscedastic model shows less departure from the true values. Thus, the model provides better estimates of these isotopic distribution parameters.

Interestingly, standard errors of $Q$ in Tables 7.5 and 7.6, as compared with the ones shown in Tables 6.1 and 6.2 (with fixed effects), are slightly smaller. This indicates the consideration of between-spectra variability may lead to more precise parameter estimates.

To check the goodness of fit of the model, scatter plots of the standardized (rescaled with the corresponding power-of-the-mean variance function) conditional residuals versus the logarithm of the predicted intensity values are presented in Figure 7.1. The symmetry of the clouds of the residuals around the horizontal line at zero indicates the adequacy of the model with respect to its mean structure. No systematic trend of the spread of the residuals along the abscissa indicates the constant nature of the standardized residuals and thus implies the appropriateness of the specified power-of-the-mean variance function.



(a) 1168.6Da Q=0.33

(b) 1456.7Da Q=0.33

(c) 1584.8Da Q=0.33

(d) 1168.6Da Q=3

(e) 1456.7Da Q=3

(f) 1584.8Da Q=3

**Figure 7.1:** Scatter plots of the standardized residuals versus the logarithm of predicted intensity values.

### Random $\boldsymbol{H}$ and $\boldsymbol{Q}$

The data set was re-analyzed by assuming the heteroscedastic model with both random $\boldsymbol{H}$ and $\boldsymbol{Q}$. Two types of estimation approaches were considered for this purpose: the analysis based on the marginal log-likelihood (7.8) and the two-stage analysis with maximization of direct log-likelihood (7.10) for the random effects. For the analysis based on the marginal log-likelihood (7.8), the estimates obtained from the two-stage analysis were used as the initial values. As the difference of the results obtained from the two approaches was at the magnitude of $10^{-7}$ to $10^{-11}$, their results can be treated as identical. It should be noted that, as the magnitude of the estimates for $\sigma_H^2$ and $\sigma_Q^2$ differ drastically, numerical problems may arise when working with the EM algorithm for the two-stage analysis. Alternatively, one could consider the rescaling of the response variable, by, e.g., multiplying a factor of $10^{-4}$. For practical implementation, function *dblquad* in Matlab 2009a was used for the double integration (with adaptive Lobatto quadrature) of the two random effects $H_i$ and $Q_i$, needed for the calculation of the marginal log-likelihood function in (7.8).

Tables 7.7 and 7.8 show the results of fitting the model with both $H_i$ and $Q_i$ as random effects. The results of the fixed effects are quite similar to those shown in Tables 7.5 and 7.6. It can also be seen that the estimates of $\sigma_Q^2$ are quite small. This is understandable, since the spectra are technical replicates for the same biological samples. It is worth noting that the estimates of the covariance parameter $\sigma_{HQ}$ for the two random effects are non-significant. Taking this into account, we refitted the model by assuming $\sigma_{HQ} = 0$. The results are presented in Tables 7.9 and 7.10.

Comparing with Tables 7.7 and 7.8, the fixed-effect estimates are identical to those in Tables 7.9 and 7.10. This is not surprising, since the fixed-effect estimates were obtained in the first stage of the analysis, in which the same estimation approach of PL-GLS was performed for both models. Moreover, the estimates of the random-effect distribution for the two models are also very similar.

### 7.3.2 Application with biological replicates: a simulation study

For the application with biological replicates, we present a simulation study. As has been observed from the real-life analyses of the bovine cytothrome C data set, the covariance $\sigma_{HQ}$ can be treated as zero, i.e., we can assume that $H_i$ and $Q_i$ are independent. The parameters in this simulation study, except of $\sigma_Q$, were chosen based on values obtained for the peptide with mass 1584.8 Da from the data of bovine cytochrome C peptides (Tables 7.9 and 7.10). The observed intensity values in the generated data sets were a sum of the mean intensity $\mu_{ij}$ and a random error term $\varepsilon_{ij}$ as in equations (7.1)-(7.2), and (7.3)-(7.4), truncated to be zero if negative. To avoid numerical problems related to zero intensity values for the least abundant peaks, $\sigma_Q$ was chosen as a compromise between the full representation of between-biological-sample variability (to be large enough) and the occurrence of numerical problems (to be small enough). In the simulations, two settings with two different relative abundances were considered, each with 100 data sets. For each data set, 6 biological replicates of mass spectra were assumed to be available. The two settings were:

$$\begin{array}{lllll} \text{Setting 1:} & Q = 1/3, & \sigma_Q = 0.05, & H = 24000, & \sigma_H = 2100 \\ \text{Setting 2:} & Q = 3, & \sigma_Q = 0.5, & H = 8000, & \sigma_H = 880 \end{array}$$

The other parameters were chosen as follows: $\sigma = 0.40$, $\theta = 0.60$, $\lambda\tau = 8.4$, and $M = 1584.76$ Da.

We chose the isotopic ratios to be the ratios from the average isotopic distribution estimated at mass $M = 1584.76$ Da by a Poisson approximation, as proposed by Breen *et al.* (2000). Again, estimates of the two-stage analysis with direct log-likelihood (7.10) for the random effects were used as the initial values for the analysis based on the marginal log-likelihood (7.8) maximization.

**Table 7.1:** Simulation results of the two settings – Mean estimate (M.Est.), mean relative bias (M.R.B.), empirical standard error $S_{\text{emp}}$ and model-based standard error $S_{\text{mb}}$, using two-stage analysis with direct log-likelihood.

| Parameter | Setting 1 | | | Setting 2 | | |
|---|---|---|---|---|---|---|
| | M.Est. | M.R.B. | $S_{\text{emp}}/S_{\text{mb}}$ | M.Est. | M.R.B. | $S_{\text{emp}}/S_{\text{mb}}$ |
| $R_2$ | 0.9114 | 0.0010 | 0.0044/0.0070 | 0.9125 | 0.0022 | 0.0051/0.0092 |
| $R_3$ | 0.4154 | 0.0023 | 0.0024/0.0032 | 0.4153 | 0.0020 | 0.0029/0.0035 |
| $R_4$ | 0.1262 | 0.0037 | 0.0015/0.0037 | 0.1257 | -0.0009 | 0.0010/0.0020 |
| $R_5$ | 0.0287 | 0.0022 | 0.0006/0.0010 | 0.0287 | 0.0032 | 0.0004/0.0006 |
| $R_6$ | 0.0052 | -0.0050 | 0.0002/0.0005 | 0.0052 | 0.0044 | 0.0001/0.0001 |
| $\mu_H$ | 24314.8 | 0.0131 | 1015/1226 | 8076.5 | 0.0096 | 319.2/351.8 |
| $\sigma_H$ | 2194.7 | 0.0451 | 481.2/581.0 | 920.6 | 0.0461 | 182.3/237.7 |
| $\mu_Q$ | 0.3376 | 0.0129 | 0.0701/0.0964 | 3.0261 | 0.0087 | 0.3024/0.3413 |
| $\sigma_Q$ | 0.0479 | -0.0417 | 0.0083/0.0092 | 0.4712 | -0.0576 | 0.0986/0.1249 |
| $\sigma$ | 0.3690 | -0.0755 | 0.0698/0.0805 | 0.3683 | -0.0792 | 0.0481/0.0716 |
| $\theta$ | 0.6321 | 0.0536 | 0.1212/0.1357 | 0.6402 | 0.0670 | 0.0929/0.0895 |
| $\lambda\tau$ | 8.7275 | 0.0390 | 0.6215/0.7826 | 8.4035 | 0.0004 | 0.0588/0.0813 |
| $llk$ | -421.85 | – | 62.22 | -437.47 | – | 32.98 |

Tables 7.1 and 7.2 show the results of the model fitted by using the two-stage analysis (with direct log-likelihood for the random effects) and the analysis based on the marginal log-likelihood (7.8), respectively. It is clear that the results are nearly identical except for the estimates of $\lambda\tau$, which show a slight difference for the two estimation approaches.

For both estimation approaches, the point estimates of the isotopic ratios $R_j$, $\lambda\tau$, as well as $Q$ are very close to the true values and with negligible biases. Regarding the parameters that reflect the between-spectra and between-biological-sample variability, i.e., $\sigma_H$ and $\sigma_Q$, the estimates are close to the true values. The estimates of $\theta$ and $\sigma$ also correspond to the true values. The simulation shows that all of the parameters are estimated with a negligible bias. Thus, the simulation indicates good performance of the model, when fitted to mass spectra data with biological replicates, simulated under the correct model specification.

From Tables 7.1 and 7.2, the average model (marginal) log-likelihood values, de-

**Table 7.2:** Simulation results of the two settings – Mean estimate (M.Est.), mean relative bias (M.R.B.), empirical standard error $S_{\mathrm{emp}}$ and model-based standard error $S_{\mathrm{mb}}$, using the analysis based on the marginal log-likelihood (7.8).

| Parameter | Setting 1 | | | Setting 2 | | |
|---|---|---|---|---|---|---|
| | M.Est. | M.R.B. | $S_{\mathrm{emp}}/S_{\mathrm{mb}}$ | M.Est. | M.R.B. | $S_{\mathrm{emp}}/S_{\mathrm{mb}}$ |
| $R_2$ | 0.9114 | 0.0010 | 0.0044/0.0070 | 0.9125 | 0.0022 | 0.0051/0.0092 |
| $R_3$ | 0.4154 | 0.0023 | 0.0024/0.0032 | 0.4153 | 0.0020 | 0.0029/0.0035 |
| $R_4$ | 0.1262 | 0.0037 | 0.0015/0.0037 | 0.1257 | -0.0009 | 0.0010/0.0020 |
| $R_5$ | 0.0287 | 0.0022 | 0.0006/0.0010 | 0.0287 | 0.0032 | 0.0004/0.0006 |
| $R_6$ | 0.0052 | -0.0050 | 0.0002/0.0005 | 0.0052 | 0.0044 | 0.0001/0.0001 |
| $\mu_H$ | 24314.8 | 0.0131 | 1015/1359 | 8076.5 | 0.0096 | 319.2/351.8 |
| $\sigma_H$ | 2194.7 | 0.0451 | 481.2/565.2 | 920.6 | 0.0461 | 182.3/237.7 |
| $\mu_Q$ | 0.3376 | 0.0129 | 0.0701/0.0942 | 3.0261 | 0.0087 | 0.3024/0.3443 |
| $\sigma_Q$ | 0.0479 | -0.0417 | 0.0083/0.0092 | 0.4712 | -0.0576 | 0.0986/0.1249 |
| $\sigma$ | 0.3690 | -0.0755 | 0.0698/0.0805 | 0.3683 | -0.0792 | 0.0481/0.0716 |
| $\theta$ | 0.6321 | 0.0536 | 0.1212/0.1359 | 0.6402 | 0.0670 | 0.0929/0.0895 |
| $\lambda\tau$ | 8.7293 | 0.0392 | 0.7101/0.7324 | 8.3720 | -0.0033 | 0.2186/0.3600 |
| $llk$ | -419.48 | – | 61.84 | -435.69 | – | 32.39 |

noted as $llk$, for the analysis based on the marginal log-likelihood (7.8) are slightly larger than those for the two-stage analysis with direct log-likelihood for the random effects. It should be noted that for the two-stage analysis, the model log-likelihood values for around 5 out of the 100 data sets, are slightly smaller than for the analysis based on the marginal log-likelihood (7.8), while all the rest 95 data sets show identical values for the two approaches. Despite the slight improvement in the marginal log-likelihood values for the analysis based on the marginal log-likelihood (7.8), the two approaches can still be treated as equivalent since the parameter estimates are equally precisely estimated.

## 7.4 Discussion

We have presented an extension of the model, introduced in Chapter 6. The model, presented in this chapter, allows for the estimation of technical and/or biological variability in the context of $^{18}$O-labeled MS data, the topic of which has not been addressed thus far. In the model, we considered the heteroscedastic nature of residual variance.

The results of the application to real-life technical replicates, in general, are consistent with the true parameter values, except for the peptide with mass 1168.6 Da. It is possible that the bias may be caused by some experimental factors unknown to us. In the simulation study, which intended to show an application to the biological replicates, the parameters were well estimated with negligible bias. For both applications, feasibility of incorporating the between-spectra variability via the inclusion of one or more random effects is ascertained. The comparison of our method with the one proposed in Chapter 6 indicates that adjusting for extra sources of variability gives more precise estimates of the parameters of interest.

Numerical complexity of the developed method is tolerable. On average, fitting the model with heteroscedasticity for each peptide, as presented in Tables 7.5 and 7.6, took about 25.6s on a HP8530p laptop using Matlab 2009a under Windows Vista®. Thus, the method can be applied to a high-throughput environment, based upon the automation of the selection of peak-clusters for fitting the model.

Several further extensions of the proposed methodology are still possible. For instance, a Bayesian formulation of the model, which would allow for the use of prior information about, e.g., the isotopic distribution of the peptide, can be proposed. This type of extension will be reported in the next chapter.

Table 7.3: Results of the analysis of the data for $Q = 1/3$ for the homoscedastic model with random $\boldsymbol{H}$ (Est.: estimate; SE: standard error.)

| | 1584.8 Da | | | 1456.7 Da | | | 1168.6 Da | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | TRUE | Est. | SE | TRUE | Est. | SE | TRUE | Est. | SE |
| $H$ | – | 22121.0 | 876.1 | – | 22820.8 | 801.4 | – | 64585.3 | 3956.8 |
| $\sigma_H^2$ | – | 4598290.9 | 3975 | – | 3896216.2 | 9958 | – | 96806404.4 | 318637 |
| $Q$ | 0.3333 | 0.3382 | 0.0030 | 0.3333 | 0.3417 | 0.0084 | 0.3333 | 0.5543 | 0.0147 |
| $\lambda\tau$ | – | 7.1627 | 0.000003 | – | 7.0348 | 0.0002 | – | 4.7177 | 0.0006 |
| $\sigma$ | – | 72.1850 | 0.0221 | – | 135.9390 | 0.0899 | – | 1342.1489 | 0.9025 |
| $R_2$ | 0.8703 | 0.8608 | 0.0017 | 0.7933 | 0.7892 | 0.0014 | 0.6645 | 0.8249 | 0.0006 |
| $R_3$ | 0.4223 | 0.3980 | 0.0015 | 0.3567 | 0.3277 | 0.0025 | 0.2454 | 0.2880 | 0.0076 |
| $R_4$ | 0.1478 | 0.1233 | 0.0013 | 0.1166 | 0.0880 | 0.0024 | 0.0653 | 0.0249 | 0.0083 |
| $R_5$ | 0.0413 | 0.0357 | 0.0029 | 0.0306 | 0.0259 | 0.0077 | 0.0139 | 0.0610 | 0.0132 |
| $R_6$ | 0.0097 | 0.0067 | 0.0025 | 0.0068 | 0.0024 | 0.0067 | 0.0025 | 0.0000 | 0.0002 |

Table 7.4: Results of the analysis of the data for $Q = 3/1$ for the homoscedastic model with random $\boldsymbol{H}$ (Est.: estimate; SE: standard error.)

| | 1584.8 Da | | | 1456.7 Da | | | 1168.6 Da | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | TRUE | Est. | SE | TRUE | Est. | SE | TRUE | Est. | SE |
| $H$ | – | 8528.9 | 361.3 | – | 8878.8 | 351.3 | – | 23630.3 | 2249.9 |
| $\sigma_H^2$ | – | 783295.8 | 805.7 | – | 732142.5 | 128.6 | – | 30237814.1 | 4710 |
| $Q$ | (2.4) | 2.4122 | 0.0063 | (2.4) | 2.3895 | 0.0102 | (2.4) | 2.0074 | 0.0146 |
| $\lambda\tau$ | – | 9.3945 | 0.0001 | – | 11.7708 | 0.0001 | – | 19.9795 | 0.000001 |
| $\sigma$ | – | 81.2602 | 0.0305 | – | 109.7680 | 0.0102 | – | 1086.6977 | 2.8333 |
| $R_2$ | 0.8703 | 0.8611 | 0.0022 | 0.7933 | 0.7741 | 0.0028 | 0.6645 | 0.7541 | 0.0098 |
| $R_3$ | 0.4223 | 0.4184 | 0.0018 | 0.3567 | 0.3349 | 0.0022 | 0.2454 | 0.3069 | 0.0087 |
| $R_4$ | 0.1478 | 0.1350 | 0.0016 | 0.1166 | 0.0967 | 0.0021 | 0.0653 | 0.0739 | 0.0087 |
| $R_5$ | 0.0413 | 0.0338 | 0.0018 | 0.0306 | 0.0213 | 0.0022 | 0.0139 | 0.0067 | 0.0261 |
| $R_6$ | 0.0097 | 0.0075 | 0.0019 | 0.0068 | 0.0046 | 0.0022 | 0.0025 | 0.0093 | 0.0151 |

**Table 7.5:** Results of the analysis of the data for $Q = 1/3$ for the heteroscedastic model with random $\boldsymbol{H}$ (Est.: estimate; SE: standard error.)

| | 1584.8 Da | | | 1456.7 Da | | | 1168.6 Da | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | TRUE | Est. | SE | TRUE | Est. | SE | TRUE | Est. | SE |
| $H$ | – | 22268.0 | 818.6 | – | 23025.1 | 759.0 | – | 69292.6 | 4523.2 |
| $\sigma_H^2$ | – | 3999670.8 | 5697.8 | – | 3362302.8 | 2210.7 | – | 104782380.6 | 462811.1 |
| $Q$ | 0.3333 | 0.3368 | 0.0016 | 0.3333 | 0.3384 | 0.0026 | 0.3333 | 0.5130 | 0.0131 |
| $\lambda\tau$ | – | 7.3254 | 0.0000001 | – | 7.3947 | 0.000002 | – | 5.8627 | 0.00002 |
| $\sigma$ | – | 0.3732 | 0.00002 | – | 0.3454 | 0.00002 | – | 0.5461 | 0.0002 |
| $\theta$ | – | 0.6250 | 0.0133 | – | 0.6901 | 0.0134 | – | 0.8333 | 0.0113 |
| $R_2$ | 0.8703 | 0.8568 | 0.0039 | 0.7933 | 0.7806 | 0.0061 | 0.6645 | 0.7556 | 0.0223 |
| $R_3$ | 0.4223 | 0.3976 | 0.0022 | 0.3567 | 0.3273 | 0.0031 | 0.2454 | 0.2812 | 0.0085 |
| $R_4$ | 0.1478 | 0.1244 | 0.0010 | 0.1166 | 0.0904 | 0.0014 | 0.0653 | 0.0556 | 0.0036 |
| $R_5$ | 0.0413 | 0.0331 | 0.0007 | 0.0306 | 0.0211 | 0.0007 | 0.0139 | 0.0095 | 0.0004 |
| $R_6$ | 0.0097 | 0.0084 | 0.0003 | 0.0068 | 0.0057 | 0.0003 | 0.0025 | 0.0012 | 0.0002 |

**Table 7.6:** Results of the analysis of the data for $Q = 3/1$ for the heteroscedastic model with random $\boldsymbol{H}$ (Est.: estimate; SE: standard error.)

| | 1584.8 Da | | | 1456.7 Da | | | 1168.6 Da | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | TRUE | Est. | SE | TRUE | Est. | SE | TRUE | Est. | SE |
| $H$ | – | 8588.6 | 326.8 | – | 8924.8 | 320.0 | – | 23542.7 | 2245.1 |
| $\sigma_H^2$ | – | 633672.2 | 563.2 | – | 600231.3 | 543.5 | – | 28422220.7 | 29380.5 |
| $Q$ | (2.4) | 2.3992 | 0.0114 | (2.4) | 2.3785 | 0.0145 | (2.4) | 2.1206 | 0.0468 |
| $\lambda\tau$ | – | 9.6154 | 0.000004 | – | 12.7681 | 0.000003 | – | 20.0000 | 0.0001 |
| $\sigma$ | – | 0.4121 | 0.00002 | – | 0.4705 | 0.00002 | – | 0.1139 | 0.00002 |
| $\theta$ | – | 0.6250 | 0.0126 | – | 0.6372 | 0.0126 | – | 0.9375 | 0.0125 |
| $R_2$ | 0.8703 | 0.8607 | 0.0043 | 0.7933 | 0.7739 | 0.0049 | 0.6645 | 0.7185 | 0.0177 |
| $R_3$ | 0.4223 | 0.4204 | 0.0026 | 0.3567 | 0.3374 | 0.0028 | 0.2454 | 0.2900 | 0.0083 |
| $R_4$ | 0.1478 | 0.1355 | 0.0014 | 0.1166 | 0.0975 | 0.0013 | 0.0653 | 0.0698 | 0.0048 |
| $R_5$ | 0.0413 | 0.0338 | 0.0005 | 0.0306 | 0.0213 | 0.0005 | 0.0139 | 0.0121 | 0.0005 |
| $R_6$ | 0.0097 | 0.0073 | 0.0002 | 0.0068 | 0.0044 | 0.0002 | 0.0025 | 0.0015 | 0.0001 |

**Table 7.7:** Results of the analysis of the data for $Q = 1/3$ for the heteroscedastic model with random $\boldsymbol{H}$ and $\boldsymbol{Q}$ (Est.: estimate; SE: standard error.)

| | 1584.8 Da | | | 1456.7 Da | | | 1168.6 Da | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | TRUE | Est. | SE | TRUE | Est. | SE | TRUE | Est. | SE |
| $H$ | – | 22321.7 | 1148 | – | 23096.2 | 1014 | – | 70541.7 | 5317 |
| $\sigma_H^2$ | – | 6199325.9 | 3549 | – | 5055473.6 | 8484 | – | 126561704.9 | 58170 |
| $Q$ | 0.3333 | 0.3345 | 0.0091 | 0.3333 | 0.3296 | 0.0120 | 0.3333 | 0.5065 | 0.0440 |
| $\sigma_Q^2$ | – | 0.000215 | 0.0000001 | – | 0.000486 | 0.00000004 | – | 0.000026 | 0.00000001 |
| $\sigma_{HQ}$ | – | -24.56 | 20.68 | – | -15.36 | 50.14 | – | 15.25 | 57.04 |
| $\lambda\tau$ | – | 7.6091 | 0.000003 | – | 7.8237 | 0.000001 | – | 5.9543 | 0.000001 |
| $\sigma$ | – | 0.4801 | 0.00002 | – | 0.4466 | 0.00004 | – | 0.5642 | 0.0000003 |
| $\theta$ | – | 0.7744 | 0.0500 | – | 0.8302 | 0.0474 | – | 0.9333 | 0.0318 |
| $R_2$ | 0.8703 | 0.8554 | 0.0267 | 0.7933 | 0.7785 | 0.0277 | 0.6645 | 0.7473 | 0.0675 |
| $R_3$ | 0.4223 | 0.3983 | 0.0135 | 0.3567 | 0.3279 | 0.0134 | 0.2454 | 0.2777 | 0.0313 |
| $R_4$ | 0.1478 | 0.1254 | 0.0048 | 0.1166 | 0.0917 | 0.0046 | 0.0653 | 0.0563 | 0.0080 |
| $R_5$ | 0.0413 | 0.0328 | 0.0022 | 0.0306 | 0.0209 | 0.0018 | 0.0139 | 0.0095 | 0.0017 |
| $R_6$ | 0.0097 | 0.0084 | 0.0007 | 0.0068 | 0.0057 | 0.0006 | 0.0025 | 0.0012 | 0.0002 |

**Table 7.8:** Results of the analysis of the data for $Q = 3/1$ for the heteroscedastic model with random $\boldsymbol{H}$ and $\boldsymbol{Q}$ (Est.: estimate; SE: standard error.)

| | 1584.8 Da | | | 1456.7 Da | | | 1168.6 Da | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | TRUE | Est. | SE | TRUE | Est. | SE | TRUE | Est. | SE |
| $H$ | – | 8600.1 | 426.7 | – | 8923.7 | 338.7 | – | 23939.1 | 3723 |
| $\sigma_H^2$ | – | 746673.7 | 1322 | – | 1328470.7 | 4283 | – | 46105776.1 | 35205 |
| $Q$ | (2.4) | 2.3940 | 0.0623 | (2.4) | 2.3828 | 0.0716 | (2.4) | 2.1343 | 0.2508 |
| $\sigma_Q^2$ | – | 0.004235 | 0.000001 | – | 0.000657 | 0.0000003 | – | 0.005497 | 0.0000002 |
| $\sigma_{HQ}$ | – | -1.6509 | 112.4 | – | -50.58 | 39.13 | – | -488.2 | 312.1 |
| $\lambda\tau$ | – | 9.6850 | 0.00003 | – | 12.7096 | 0.00002 | – | 19.9047 | 0.00003 |
| $\sigma$ | – | 0.3805 | 0.00003 | – | 0.4370 | 0.00002 | – | 0.2279 | 0.00002 |
| $\theta$ | – | 0.7771 | 0.0445 | – | 0.7771 | 0.0184 | – | 1.0375 | 0.0275 |
| $R_2$ | 0.8703 | 0.8602 | 0.0203 | 0.7933 | 0.7740 | 0.0183 | 0.6645 | 0.7138 | 0.0938 |
| $R_3$ | 0.4223 | 0.4212 | 0.0113 | 0.3567 | 0.3380 | 0.0094 | 0.2454 | 0.2885 | 0.0432 |
| $R_4$ | 0.1478 | 0.1357 | 0.0042 | 0.1166 | 0.0976 | 0.0034 | 0.0653 | 0.0695 | 0.0118 |
| $R_5$ | 0.0413 | 0.0338 | 0.0014 | 0.0306 | 0.0214 | 0.0011 | 0.0139 | 0.0120 | 0.0021 |
| $R_6$ | 0.0097 | 0.0073 | 0.0004 | 0.0068 | 0.0044 | 0.0003 | 0.0025 | 0.0015 | 0.0003 |

**Table 7.9:** Results of the analysis of the data for $Q = 1/3$ for the heteroscedastic model with random $\boldsymbol{H}$ and $\boldsymbol{Q}$ assuming $\sigma_{HQ} = 0$ (Est.: estimate; SE: standard error.)

| | 1584.8 Da | | | 1456.7 Da | | | 1168.6 Da | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | TRUE | Est. | SE | TRUE | Est. | SE | TRUE | Est. | SE |
| $H$ | – | 22321.7 | 1192 | – | 23096.2 | 1085 | – | 70541.7 | 5265 |
| $\sigma_H^2$ | – | 6307482.2 | 6834 | – | 5193374.5 | 574.8 | – | 126463528.3 | 144346 |
| $Q$ | 0.3333 | 0.3328 | 0.0089 | 0.3333 | 0.3284 | 0.0125 | 0.3333 | 0.5057 | 0.0443 |
| $\sigma_Q^2$ | – | 0.000216 | 0.0000001 | – | 0.000487 | 0.0000004 | – | 0.000026 | 0.0000002 |
| $\lambda\tau$ | – | 7.6091 | 0.000004 | – | 7.8237 | 0.00001 | – | 5.9543 | 0.00001 |
| $\sigma$ | – | 0.4801 | 0.00001 | – | 0.4466 | 0.0002 | – | 0.5642 | 0.00001 |
| $\theta$ | – | 0.7744 | 0.0474 | – | 0.8302 | 0.0484 | – | 0.9333 | 0.0315 |
| $R_2$ | 0.8703 | 0.8554 | 0.0250 | 0.7933 | 0.7785 | 0.0278 | 0.6645 | 0.7473 | 0.0676 |
| $R_3$ | 0.4223 | 0.3983 | 0.0128 | 0.3567 | 0.3279 | 0.0134 | 0.2454 | 0.2777 | 0.0314 |
| $R_4$ | 0.1478 | 0.1254 | 0.0047 | 0.1166 | 0.0917 | 0.0046 | 0.0653 | 0.0563 | 0.0080 |
| $R_5$ | 0.0413 | 0.0328 | 0.0022 | 0.0306 | 0.0209 | 0.0018 | 0.0139 | 0.0095 | 0.0017 |
| $R_6$ | 0.0097 | 0.0084 | 0.0007 | 0.0068 | 0.0057 | 0.0006 | 0.0025 | 0.0012 | 0.0002 |

**Table 7.10:** Results of the analysis of the data for $Q = 3/1$ for the heteroscedastic model with random $\boldsymbol{H}$ and $\boldsymbol{Q}$ assuming $\sigma_{HQ} = 0$ (Est.: estimate; SE: standard error.)

| | 1584.8 Da | | | 1456.7 Da | | | 1168.6 Da | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | TRUE | Est. | SE | TRUE | Est. | SE | TRUE | Est. | SE |
| $H$ | – | 8600.1 | 414.3 | – | 8923.8 | 440.1 | – | 23939.1 | 3004 |
| $\sigma_H^2$ | – | 746673.7 | 1322.4 | – | 924743.1 | 898.5 | – | 46105776.1 | 10082 |
| $Q$ | (2.4) | 2.3940 | 0.0613 | (2.4) | 2.3803 | 0.0635 | (2.4) | 2.1343 | 0.2456 |
| $\sigma_Q^2$ | – | 0.004235 | 0.000001 | – | 0.005253 | 0.00003 | – | 0.005497 | 0.000002 |
| $\lambda\tau$ | – | 9.6850 | 0.000002 | – | 12.7096 | 0.000002 | – | 19.9047 | 0.00001 |
| $\sigma$ | – | 0.3805 | 0.00003 | – | 0.4371 | 0.0001 | – | 0.2279 | 0.00001 |
| $\theta$ | – | 0.7771 | 0.0398 | – | 0.7771 | 0.0356 | – | 1.0375 | 0.0445 |
| $R_2$ | 0.8703 | 0.8602 | 0.0194 | 0.7933 | 0.7740 | 0.0180 | 0.6645 | 0.7138 | 0.0894 |
| $R_3$ | 0.4223 | 0.4212 | 0.0109 | 0.3567 | 0.3380 | 0.0094 | 0.2454 | 0.2885 | 0.0414 |
| $R_4$ | 0.1478 | 0.1357 | 0.0041 | 0.1166 | 0.0976 | 0.0035 | 0.0653 | 0.0695 | 0.0113 |
| $R_5$ | 0.0413 | 0.0338 | 0.0014 | 0.0306 | 0.0214 | 0.0011 | 0.0139 | 0.0120 | 0.0020 |
| $R_6$ | 0.0097 | 0.0073 | 0.0004 | 0.0068 | 0.0044 | 0.0003 | 0.0025 | 0.0015 | 0.0002 |

# Chapter 8

# A Bayesian approach for the analysis of $^{18}$O-labeled mass spectra using a heteroscedastic fixed-effects Markov-chain-based model

In this chapter, we formulate the heteroscedastic model, presented in Chapter 6, within the Bayesian framework. Using the Bayesian approach allows for the incorporation of prior information that could be helpful to analyze the data. In particular, such information exists for the isotopic distribution (see Section 3.2.2). Moreover, it allows inclusion of random effects that can be used to capture the biological variability of the peptide abundance. We investigate the operational characteristics of the model by applying it to real-life MS data set (the bovine cytochrome C data) and by conducting a simulation study.

## 8.1 Model formulation

### 8.1.1 The likelihood

The likelihood of the moel is the same as in equations (6.1), (6.2) and (5.7)–(5.8).

### 8.1.2 Prior and posterior distributions

Non-informative normal priors were defined for the logarithm of all the parameters except for $\lambda$, the common residual variance $\sigma$, and the power parameter $\theta$. For $\theta$, a non-informative normal prior distribution was considered. For $\lambda$, a non-informative normal prior was defined for it with a box-cox transformation: $\lambda' = \log\left(\frac{\lambda}{20/\tau-\lambda}\right)$. More specifically, the following priors were used for these parameters:

$$\begin{aligned}
\log(H_i) &\sim N\left(0, \frac{1}{\tau_1}\right), && (8.1)\\
\log(Q) &\sim N\left(0, \frac{1}{\tau_2}\right), && (8.2)\\
\log(R_j) &\sim N\left(0, \frac{1}{\tau_3}\right), && (8.3)\\
\sigma^{-2} &\sim \Gamma(\alpha, \beta), && (8.4)\\
\theta &\sim N\left(0, \frac{1}{\tau_4}\right), && (8.5)\\
\lambda' &\sim N\left(0, \frac{1}{\tau_5}\right), && (8.6)
\end{aligned}$$

where $\alpha$, $\beta$, and $\tau_1, \ldots, \tau_5$ are positive constants close to zero. Since the variance function of the model, shown in (6.1), is dependent on the mean structure parameters, there are no closed form posterior distributions for the parameters. As a result, the posterior distributions need to be evaluated by numerical (sampling) methods, e.g., via Metropolis-Hasting algorithm with acception-rejection rules.

## 8.2 Practical implementation

In the following sections, the implementation of the Markov-chain-based model, using WinBUGS and JAGS, will be introduced.

### 8.2.1 WinBUGS through WBDiff

Since the implementation of the model entails the estimation of Markov Chain transition probabilities through matrix exponential, the Bayesian model cannot be implemented directly through WinBUGS. However, WBDiff, which namely is a WinBUGS Differential Interface, makes such application in WinBUGS feasible. WBDiff is built for WinBUGS to do differential equations, and hence can handle matrix exponential as well. The software can be downloaded from `http://www.winbugs-development.org.uk/wbdiff.html` with user manual available from Lunn (2004).

The transition matrix $\boldsymbol{T}$ is of the form (5.6) with $p_{16}$, $p_{17}$, and $p_{18}$ being the known proportions of oxygen isotopes that exist in the heavy-oxygen-water due to water impurities. The transition matrix with element $T_{ij}$ can be interpreted as the probability of moving from state $i$ to state $j$ after the next exchange (reaction).

To implement the matrix exponential shown in equation (5.7) via differential equations, let $\boldsymbol{\pi}(t)$ denote the vector of the transition rate, given time $t$, such that $\boldsymbol{S}(\tau) = \boldsymbol{S}_0 e^{-\lambda\tau} e^{\boldsymbol{T}\lambda\tau} = \frac{\partial \boldsymbol{\pi}(t)}{\partial t} e^{-\lambda\tau}$. The vector $\boldsymbol{\pi}(t) = (\pi_1(t), \pi_2(t), \ldots, \pi_6(t))$ contains transition rates related to the six states of oxygen combinations for the carboxyl-terminus, expressed in (2.10). Then the matrix exponential can be written down as the differential equations shown as follows:

$$\begin{cases} \frac{\partial \pi_1(t)}{\partial t} & = \pi_1(t)T_{11}\lambda\tau + \pi_2(t)T_{21}\lambda\tau + \pi_3(t)T_{31}\lambda\tau \\ \frac{\partial \pi_2(t)}{\partial t} & = \pi_1(t)T_{12}\lambda\tau + \pi_2(t)T_{22}\lambda\tau + \pi_3(t)T_{32}\lambda\tau + \pi_4(t)T_{42}\lambda\tau + \pi_5(t)T_{52}\lambda\tau \\ \frac{\partial \pi_3(t)}{\partial t} & = \pi_1(t)T_{13}\lambda\tau + \pi_2(t)T_{23}\lambda\tau + \pi_3(t)T_{33}\lambda\tau + \pi_5(t)T_{53}\lambda\tau + \pi_6(t)T_{63}\lambda\tau \\ \frac{\partial \pi_4(t)}{\partial t} & = \pi_2(t)T_{24}\lambda\tau + \pi_4(t)T_{44}\lambda\tau + \pi_5(t)T_{54}\lambda\tau \\ \frac{\partial \pi_5(t)}{\partial t} & = \pi_2(t)T_{25}\lambda\tau + \pi_3(t)T_{35}\lambda\tau + \pi_4(t)T_{45}\lambda\tau + \pi_5(t)T_{55}\lambda\tau + \pi_6(t)T_{65}\lambda\tau \\ \frac{\partial \pi_6(t)}{\partial t} & = \pi_3(t)T_{36}\lambda\tau + \pi_5(t)T_{56}\lambda\tau + \pi_6(t)T_{66}\lambda\tau \end{cases} \tag{8.7}$$

The differential equations in (8.7) can be implemented in WinBUGS using the WBDiff as an interface. An example of the WinBUGS code for the Bayesian model is shown in Appendix C.

### 8.2.2 JAGS-Just Another Gibbs Sampler

Alternatively, the matrix exponential can be implemented in JAGS. In JAGS 1.0.3, matrix exponential is defined as an internal function `mexp` by loading the `msm` module (User manual available at Plummer 2009).

## 8.3 Results

In this section, we present results of an application of the model to the controlled experiment of the six replicated mass spectra of bovine cytochrome C peptides. We also show results of a simulation study, undertaken to check the statistical properties of the implemented method.

### 8.3.1 Bovine cytochrome C data sets

The model was applied to three fragments (peptides at 1168.6Da, $1456,7$Da and 1584.8Da respectively) of the replicated mass spectra of bovine cytochrome C (see Section 3.1).

The model application to the data was analyzed using WinBUGS 1.4 through WBDiff. As an initial step, a homoscedastic model, by using priors shown in (8.1)–(8.6), was fitted to the data. Later, a heteroscedastic model with power of the mean variance function, defined by (6.1)–(6.2) and (5.7)–(5.8), was applied to the same data sets, by using (8.5) as the prior for the power parameter $\theta$. Again, as the true reaction time $\tau$ was unknown, the product of $\lambda\tau$ was estimated. Tables 8.2 to 8.7 show the statistical results of the homoscedastic and heteroscedastic models.

Several patterns can be observed in these tables. For each of the peptides, the estimated values of $\theta$ of the heteroscedastic model, from the two experiments with relative abundances $3/1$ and $1/3$, are not significantly different from each other, by taking into account the precision expressed as the 95% credible interval. This suggests the chosen functional form of dependence of residual variance on the intensity was appropriate.

It is also worth noting that, for the peptides with masses 1584.8 Da and 1456.7 Da, the point estimates for $Q$, and for the isotopic ratios $\boldsymbol{R}$, shown in Tables 8.2 to 8.5, for both models, are very close to the true values.

When $Q = 3/1$, its value is consistently estimated around 2.4, especially for the two peptides with masses $1456,7$ and 1584.8Da. This agrees with the estimates in the frequentist approach presented in the Chapters 5 and 6. Moreover, for the two peptides, the estimates of $\lambda\tau$, for the two experiments with different relative abundance values, are also very similar.

For the peptide with mass 1168.6 Da, the estimates for $Q$ and $\boldsymbol{R}$, shown in Tables 8.6 and 8.7, differ considerably from the true values. The estimates of the heteroscedastic models are slightly closer to the true values, implying a possible improvement when a proper variance function is used.

### 8.3.2 A simulation study

The simulation was performed based on the assumption of power-of-the-mean variance function for the residuals. The settings were the same as in the simulation study for the frequentist approach, presented in Section 6.4. Both heteroscedastic and homoscedastic models were applied to the simulated data. The intention of the simulation was to check the statistical properties of the proposed model under the correct model assumptions.

The simulation was done in R 2.8.0 with `R2jags` package linking the software JAGS 1.0.3 to R, since it was more convenient to read data and compile the models through R with a large amount of data sets. From the results of the simulation, we observed that the biases and MSE of the isotopic ratio ($\boldsymbol{R}$) estimates were negligible. Thus, we only discuss the results of the parameters of $Q$, $\lambda$, $\theta$ and $\sigma$. Tables B.1 to B.12 (Appendix B) show the results of different parameter estimates in the simulation. The summary statistics shown in the tables are based on the medians of the corresponding parameters. Figures 8.2 to 8.6 give a graphical representation of the summary statistics shown in Tables B.1 to B.9.

The results of the estimation of parameters $\lambda$ and $Q$, shown in Figures 8.2 to 8.5, show that the MSE is smaller for the heteroscedastic model (denoted in figures by *VAR*) than the homoscedastic model (denoted by *CON*). This indicates that a correct specification of residual variance gives more precise parameter estimates.

Figures 8.4 and 8.5 also show that when $\lambda$ increases, $Q$ is better estimated as its MSE becomes smaller. This finding agrees with the results shown in Section 6.4, for the frequentist model. It indicates that, given a fixed reaction time $\tau$, when labeling is more complete, $Q$ can be better estimated.

The results of $\lambda$, when its true value is equal to 0.10 and the residual variance $\sigma$ is equal to 1.50, are not shown in the figures because the estimates were seriously biased and very imprecise. This is related to the convergence problem, as explained in Section 6.4.

### Convergence

The convergence of the continuation procedure for the parameters were checked by Geweke statistics (Geweke 1992). From the simulation, it was found that, when $\lambda = 0.01$ and $\sigma = 1.50$, there was problem for the posterior distribution of $\lambda$ to converge. Table 8.1 shows the percentage of the converged data sets for parameter $\lambda$ for the settings with $\lambda = 0.01$ and $\sigma = 1.50$.

**Table 8.1:** Percentage of converged data sets per setting for $\lambda$.

| $\sigma$ | $\lambda$ | $Q$ | Breen | | E1 | | E2 | |
|---|---|---|---|---|---|---|---|---|
| | | | CON | VAR | CON | VAR | CON | VAR |
| 1.50 | 0.10 | 0.5 | 1.00% | 0.00% | 0.80% | 0.00% | 1.00% | 0.00% |
| | | 1.0 | 0.60% | 0.20% | 0.20% | 0.00% | 1.20% | 0.20% |
| | | 2.0 | 0.20% | 0.60% | 1.00% | 0.20% | 1.00% | 0.60% |

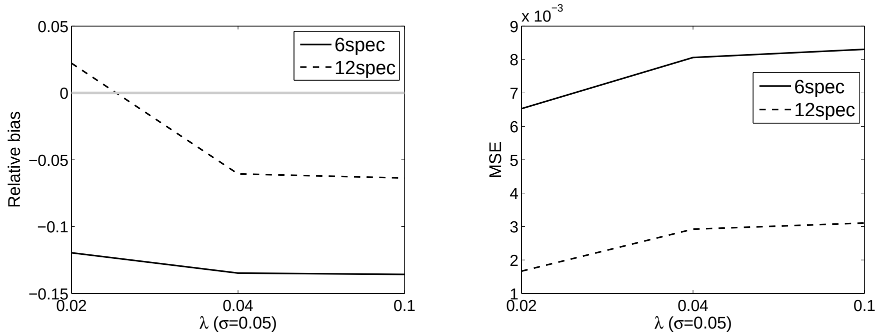
The reason for the non-convergence has been explained in Section 6.4.2.

In the Bayesian framework, however, the non-convergence of one parameter doesn't hinder the estimation of other parameters, since they are sampled from their conditional posterior distributions. One may argue about the validity of their joint distribution if some of them don't converge. Alternatively, one can think of performing a two-stage analysis, as described in Section 6.4.2, to resolve the problem.

### Sample size for the estimation of $\theta$

Figure 8.6 shows that, when $\sigma = 1.50$, the bias for $\theta$ is negligible. However, this is not the case when $\sigma = 0.05$. In such case, $\theta$ was severely underestimated. The underestimation of $\theta$ may be due to the lack of information available in the data. A solution to increase the information content of the data is to increase the number of available spectra.

The simulation for the settings with **A** ratios was repeated by increasing the number of spectra from 6 to 12. Figure 8.1 shows a comparison of the bias and



(a) $\bar{b}$

(b) MSE

**Figure 8.1:** Comparison of the estimation of $\theta$ from data sets with 6 and 12 spectra for seetings with Breen ratios ($\sigma = 0.05$).

MSE for the estimates of $\theta$, obtained from the data sets containing 6 and 12 spectra. Clearly, the underestimation is diminished and the MSE becomes much smaller when number of spectra is increased.

## 8.4 Discussion

In this chapter, we have implemented the heteroscedastic model, introduced in Chapter 6, in the Bayesian framework. Using the Bayesian approach allows for the incorporation of prior information that could be helpful to analyze the data. In particular, such information exists for the isotopic distribution. We assessed the performances of the model via both a real-life data application and a simulation study.

The application of the models to the bovine cytochrome C data sets, in general, produced unbiased estimation, except for the peptide with mass 1168.6Da. For this peptide, the results were biased both for the frequentist and Bayesian approaches. The bias might be caused by some experimental factors unknown to us.

The simulation study addresses the importance of a correct specification of the residual variance. The study shows that a correct specification for the form of the residual variance leads to a precision gain. It is worth noting that for the power parameter $\theta$ in the variance function to be well estimated, there should be enough information in the data. Such information content can be increased by providing more replications of spectra.

The results of the analysis using the Bayesian model agrees, in general, with the ones implemented in the frequentist framework (presented in Chapter 6). An advantage of the Bayesian approach over the frequentist approach, despite the possibility of incorporating prior information, is that the estimation approach is relatively easier and straightforward. Recall that in the frequentist framework, when the model incorporates both a mean-dependent variance function and random effects, a one-stage analysis is computationally complex and very sensitive to the choice of initial values.

The results of the application of our method to data from a real-life, controlled experiment and to a simulation study confirm feasibility and satisfactory performance of the proposed modeling approach. The model is flexible for further extensions, e.g., considering the estimation of different sources of variability, in the context of MS data, by including random effeccts. In the next chapter, we present an extended model by including random effects of $H$ and $Q$, to account for the between-spectra and biological variability, in the Bayesian framework.

## Statistical results of the case study:

**Table 8.2:** Results of the analysis of the data for $Q = 3/1$ at 1584.8Da (HOMOSC.: homoscedastic model; HETEROSC.: heteroscedastic model).

| Parameter | TRUE | HOMOSC. | | | HETEROSC. | | |
|---|---|---|---|---|---|---|---|
| | | Mean | Median | 95% c.i. | Mean | Median | 95% c.i. |
| $H_1$ | – | 8400.3 | 8402.0 | (8296.5, 8498.4) | 8432.9 | 8433.1 | (8293.7, 8583.9) |
| $H_2$ | – | 8258.8 | 8259.1 | (8158.4, 8358.1) | 8270.9 | 8272.0 | (8125.1, 8414.6) |
| $H_3$ | – | 7441.6 | 7442.1 | (7343.9, 7535.3) | 7538.9 | 7538.8 | (7403.5, 7684.9) |
| $H_4$ | – | 9868.8 | 9871.2 | (9561.4, 9973.7) | 9895.0 | 9893.4 | (9734.2, 10060.0) |
| $H_5$ | – | 9527.3 | 9528.8 | (9417.0, 9630.5) | 9618.7 | 9617.3 | (9456.7, 9796.9) |
| $H_6$ | – | 8420.6 | 8422.2 | (8309.6, 8520.1) | 8504.8 | 8502.3 | (8358.1, 8658.1) |
| $Q$ | *2.4* | 2.3857 | 2.3855 | (2.3594, 2.4127) | 2.3581 | 2.3583 | (2.3221, 2.3899) |
| $\lambda\tau$ | – | 8.2320 | 8.2080 | (7.8000, 8.8320) | 8.6040 | 8.6040 | (8.2080, 9.0360) |
| $\sigma$ | – | 123.2132 | 121.3509 | (100.3990, 151.1763) | 1.1490 | 0.3602 | (0.1278, 3.7930) |
| $\theta$ | – | – | – | – | 0.6517 | 0.6780 | (0.3975, 0.8025) |
| $R_2$ | 0.8703 | 0.8635 | 0.8634 | (0.8570, 0.8705) | 0.8617 | 0.8617 | (0.8500, 0.8732) |
| $R_3$ | 0.4223 | 0.4255 | 0.4255 | (0.4201, 0.4307) | 0.4302 | 0.4300 | (0.4229, 0.4386) |
| $R_4$ | 0.1478 | 0.1374 | 0.1375 | (0.1324, 0.1425) | 0.1388 | 0.1389 | (0.1353, 0.1423) |
| $R_5$ | 0.0413 | 0.0362 | 0.0364 | (0.0299, 0.0412) | 0.0367 | 0.0367 | (0.0352, 0.0383) |
| $R_6$ | 0.0098 | 0.0085 | 0.0083 | (0.0055, 0.0133) | 0.0097 | 0.0097 | (0.0090, 0.0104) |

**Table 8.3:** Results of the analysis of the data for $Q = 1/3$ at 1584.8Da (HOMOSC.: homoscedastic model; HETEROSC.: heteroscedastic model).

| Parameter | TRUE | HOMOSC. | | | HETEROSC. | | |
|---|---|---|---|---|---|---|---|
| | | Mean | Median | 95% c.i. | Mean | Median | 95% c.i. |
| $H_1$ | – | 23001.6 | 22999.0 | (22876.0, 23126.9) | 23021.5 | 23017.9 | (22775.6, 23271.2) |
| $H_2$ | – | 22418.0 | 22418.2 | (22295.0, 22547.9) | 22431.3 | 22431.6 | (22192.4, 22673.1) |
| $H_3$ | – | 21347.4 | 21345.5 | (21225.4, 21467.2) | 21397.8 | 21398.4 | (21166.0, 21630.3) |
| $H_4$ | – | 23832.4 | 23833.2 | (23705.5, 23953.6) | 23987.7 | 23989.7 | (23728.8, 24256.8) |
| $H_5$ | – | 18534.7 | 18534.9 | (18413.7, 18655.5) | 18525.9 | 18528.3 | (18318.6, 18722.6) |
| $H_6$ | – | 24598.7 | 24597.1 | (24476.4, 24724.7) | 24780.1 | 24779.2 | (24527.3, 25047.5) |
| $Q$ | 0.3333 | 0.3270 | 0.3272 | (0.3179, 0.3373) | 0.3260 | 0.3260 | (0.3213, 0.3308) |
| $\lambda\tau$ | – | 7.3680 | 7.1880 | (6.0240, 9.0720) | 7.2720 | 7.2360 | (6.6240, 8.1360) |
| $\sigma$ | – | 83.803 | 82.5033 | (68.6678, 103.0714) | 0.5256 | 0.3599 | (0.1404, 1.1882) |
| $\theta$ | – | – | – | – | 0.6273 | 0.6284 | (0.4841, 0.7486) |
| $R_2$ | 0.8703 | 0.8616 | 0.8617 | (0.8576, 0.8655) | 0.8572 | 0.8572 | (0.8490, 0.8652) |
| $R_3$ | 0.4223 | 0.4122 | 0.4122 | (0.4067, 0.4171) | 0.4102 | 0.4103 | (0.4053, 0.4148) |
| $R_4$ | 0.1478 | 0.1315 | 0.1315 | (0.1228, 0.1384) | 0.1315 | 0.1315 | (0.1284, 0.1346) |
| $R_5$ | 0.0413 | 0.0416 | 0.0415 | (0.0346, 0.0486) | 0.0387 | 0.0387 | (0.0371, 0.0406) |
| $R_6$ | 0.0098 | 0.0104 | 0.0106 | (0.0067, 0.0156) | 0.0130 | 0.0130 | (0.0121, 0.0138) |

**Table 8.4:** Results of the analysis of the data for $Q = 3/1$ at 1456.7Da (HOMOSC.: homoscedastic model; HETEROSC.: heteroscedastic model).

| Parameter | TRUE | HOMOSC. | | | HETEROSC. | | |
|---|---|---|---|---|---|---|---|
| | | Mean | Median | 95% c.i. | Mean | Median | 95% c.i. |
| $H_1$ | – | 8773.5 | 8772.2 | (8663.3, 8885.2) | 8720.8 | 8720.3 | (8568.3, 8870.5) |
| $H_2$ | – | 8860.9 | 8860.7 | (8749.1, 8971.8) | 8820.2 | 8819.8 | (8670.9, 8973.7) |
| $H_3$ | – | 7784.9 | 7784.4 | (7682.6, 7889.6) | 7847.6 | 7846.5 | (7710.6, 7993.0) |
| $H_4$ | – | 10205.1 | 10203.9 | (10082.1, 10329.7) | 10232.0 | 10232.0 | (10064.8, 10407.4) |
| $H_5$ | – | 9763.0 | 9762.9 | (9647.9, 9878.8) | 9813.9 | 9812.8 | (9645.8, 9989.7) |
| $H_6$ | – | 8644.6 | 8643.9 | (8536.4, 8757.3) | 8745.1 | 8742.4 | (8596.3, 8909.4) |
| $Q$ | *2.4* | 2.3777 | 2.3775 | (2.3517, 2.4035) | 2.3623 | 2.3633 | (2.3270, 2.3939) |
| $\lambda\tau$ | – | 10.9440 | 10.8120 | (9.4320, 12.7200) | 11.6400 | 11.5080 | (10.2960, 13.2000) |
| $\sigma$ | – | 133.2802 | 131.1575 | (108.6815, 164.6927) | 0.5938 | 0.4258 | (0.1665, 1.2280) |
| $\theta$ | – | – | – | – | 0.6513 | 0.6553 | (0.5289, 0.7731) |
| $R_2$ | 0.7933 | 0.7758 | 0.7759 | (0.7690, 0.7826) | 0.7762 | 0.7763 | (0.7651, 0.7868) |
| $R_3$ | 0.3567 | 0.3396 | 0.3396 | (0.3342, 0.3449) | 0.3430 | 0.3430 | (0.3368, 0.3493) |
| $R_4$ | 0.1166 | 0.1008 | 0.1008 | (0.0955, 0.1060) | 0.1016 | 0.1015 | (0.0987, 0.1045) |
| $R_5$ | 0.0306 | 0.0233 | 0.0237 | (0.0148, 0.0286) | 0.0245 | 0.0245 | (0.0235, 0.0256) |
| $R_6$ | 0.0068 | 0.0057 | 0.0057 | (0.0004, 0.0089) | 0.0071 | 0.0070 | (0.0066, 0.0076) |

**Table 8.5:** Results of the analysis of the data for $Q = 1/3$ at 1456.7Da (HOMOSC.: homoscedastic model; HETEROSC.: heteroscedastic model).

| Parameter | TRUE | HOMOSC. | | | HETEROSC. | | |
|---|---|---|---|---|---|---|---|
| | | Mean | Median | 95% c.i. | Mean | Median | 95% c.i. |
| $H_1$ | – | 24749.4 | 24750.1 | (24499.6, 25001.2) | 24867.0 | 24852.0 | (24360.1, 25457.3) |
| $H_2$ | – | 22374.6 | 22376.7 | (22126.6, 22618.9) | 22883.5 | 22871.3 | (22334.2, 23476.7) |
| $H_3$ | – | 22235.7 | 22238.0 | (21988.8, 22473.4) | 22674.8 | 22667.0 | (22179.8, 23214.7) |
| $H_4$ | – | 24557.7 | 24558.6 | (24307.3, 24797.0) | 24832.9 | 24828.4 | (24266.6, 25408.4) |
| $H_5$ | – | 19651.4 | 19650.0 | (19412.1, 19891.5) | 19713.6 | 19716.8 | (19247.8, 20178.4) |
| $H_6$ | – | 24424.7 | 24427.3 | (24174.9, 24671.2) | 25919.5 | 25179.1 | (24635.0, 25806.4) |
| $Q$ | 0.3333 | 0.3598 | 0.3601 | (0.3507, 0.3676) | 0.3294 | 0.3297 | (0.3208, 0.3370) |
| $\lambda\tau$ | – | 16.3800 | 16.6560 | (14.4600, 17.5080) | 8.7480 | 8.8080 | (7.3800, 9.8640) |
| $\sigma$ | – | 161.1173 | 158.8647 | (132.1106, 197.2109) | 0.2629 | 0.2108 | (0.1007, 0.5082) |
| $\theta$ | – | – | – | – | 0.7639 | 0.7667 | (0.6607, 0.8627) |
| $R_2$ | 0.7933 | 0.7905 | 0.7905 | (0.7836, 0.7977) | 0.7769 | 0.7770 | (0.7611, 0.7926) |
| $R_3$ | 0.3567 | 0.3485 | 0.3485 | (0.3425, 0.3543) | 0.3415 | 0.3416 | (0.3335, 0.3490) |
| $R_4$ | 0.1166 | 0.1061 | 0.1061 | (0.1007, 0.1116) | 0.1030 | 0.1030 | (0.1002, 0.1059) |
| $R_5$ | 0.0306 | 0.0328 | 0.0327 | (0.0227, 0.0407) | 0.0275 | 0.0275 | (0.0261, 0.0292) |
| $R_6$ | 0.0068 | 0.0044 | 0.0043 | (0.0026, 0.0065) | 0.0112 | 0.0112 | (0.0104, 0.0119) |

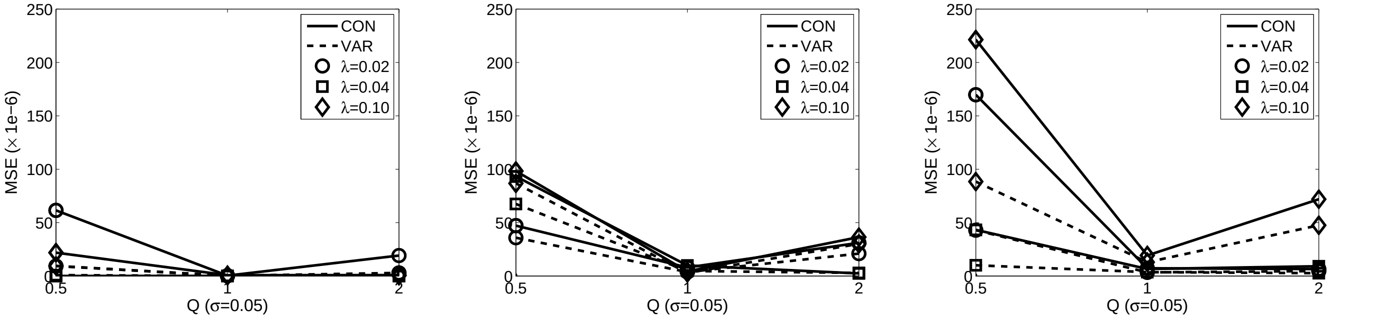
**Table 8.6:** Results of the analysis of the data for $Q = 3/1$ at 1168.6Da (HOMOSC.: homoscedastic model; HETEROSC.: heteroscedastic model).

| Parameter | TRUE | HOMOSC. | | | HETEROSC. | | |
|---|---|---|---|---|---|---|---|
| | | Mean | Median | 95% c.i. | Mean | Median | 95% c.i. |
| $H_1$ | – | 32402.1 | 32407.5 | (31148.7, 33694.7) | 32833.4 | 32834.2 | (31229.7, 34649.2) |
| $H_2$ | – | 26655.6 | 26663.6 | (25562.9, 27733.2) | 27550.7 | 27521.9 | (26285.9, 28929.0) |
| $H_3$ | – | 27570.4 | 27594.5 | (26363.5, 28674.2) | 29538.4 | 29534.8 | (28096.3, 30975.1) |
| $H_4$ | – | 18437.2 | 18435.5 | (17529.7, 19372.0) | 17180.0 | 17165.4 | (16359.0, 18112.2) |
| $H_5$ | – | 20856.3 | 20861.4 | (19869.9, 21815.7) | 19779.5 | 19769.0 | (18828.0, 20866.3) |
| $H_6$ | – | 19322.5 | 19329.4 | (18336.0, 20289.8) | 18826.6 | 18818.1 | (17894.7, 19784.6) |
| $Q$ | *2.4* | 2.0330 | 2.0326 | (1.9674, 2.1034) | 2.1346 | 2.1353 | (2.0408, 2.2293) |
| $\lambda\tau$ | – | 15.5040 | 15.6480 | (13.6920, 16.5600) | 17.0800 | 17.1240 | (16.2840, 17.7240) |
| $\sigma$ | – | 1204.438 | 1186.492 | (989.9005, 1466.329) | 0.0956 | 0.0842 | (0.0491, 0.1634) |
| $\theta$ | – | – | – | – | 0.9577 | 0.9593 | (0.8873, 1.0216) |
| $R_2$ | 0.6645 | 0.7568 | 0.7568 | (0.7342, 0.7798) | 0.7137 | 0.7141 | (0.6786, 0.7478) |
| $R_3$ | 0.2454 | 0.3096 | 0.3095 | (0.2917, 0.3281) | 0.2900 | 0.2899 | (0.2734, 0.3062) |
| $R_4$ | 0.0653 | 0.0800 | 0.0801 | (0.0636, 0.0946) | 0.0723 | 0.0722 | (0.0676, 0.0768) |
| $R_5$ | 0.0139 | 0.0036 | 0.0035 | (0.0023, 0.0051) | 0.0144 | 0.0144 | (0.0133, 0.0154) |
| $R_6$ | 0.0025 | 0.0053 | 0.0052 | (0.0027, 0.0080) | 0.0037 | 0.0037 | (0.0034, 0.0040) |

**Table 8.7:** Results of the analysis of the data for $Q = 1/3$ at 1168.6Da (HOMOSC.: homoscedastic model; HETEROSC.: heteroscedastic model).

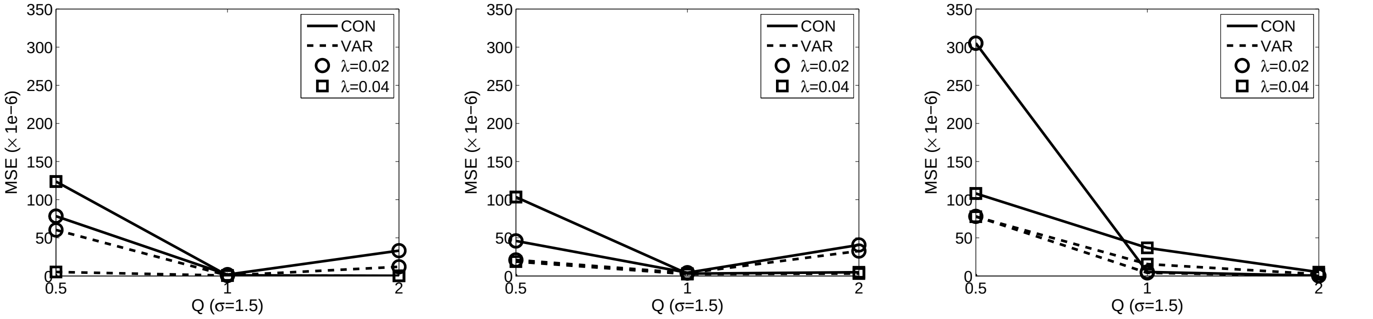
| Parameter | TRUE | HOMOSC. | | | HETEROSC. | | |
|---|---|---|---|---|---|---|---|
| | | Mean | Median | 95% c.i. | Mean | Median | 95% c.i. |
| $H_1$ | – | 75245.8 | 75242.3 | (72985.3, 77564.3) | 87319.1 | 87179.7 | (79783.1, 96097.2) |
| $H_2$ | – | 73669.0 | 73646.9 | (71581.8, 75948.2) | 88076.3 | 87918.9 | (81255.8, 96498.6) |
| $H_3$ | – | 63588.3 | 63585.8 | (61422.7, 65726.7) | 75376.6 | 75167.1 | (69101.6, 82509.8) |
| $H_4$ | – | 71619.6 | 71601.5 | (69427.8, 73940.7) | 77272.2 | 77067.4 | (71406.1, 84527.4) |
| $H_5$ | – | 48695.4 | 48687.6 | (46695.1, 50786.3) | 50596.9 | 50476.6 | (46784.7, 55115.4) |
| $H_6$ | – | 61968.9 | 61988.6 | (59748.0, 64080.2) | 66931.3 | 66838.9 | (61500.5, 72929.8) |
| $Q$ | 0.3333 | 0.5266 | 0.5268 | (0.5008, 0.5512) | 0.4565 | 0.4568 | (0.4268, 0.4861) |
| $\lambda\tau$ | – | 3.9720 | 3.9720 | (3.4560, 4.5840) | 5.6400 | 5.8200 | (5.2440, 6.5760) |
| $\sigma$ | – | 17551.493 | 1527.288 | (1267.013, 1900.911) | 0.0888 | 0.0781 | (0.0452, 0.1582) |
| $\theta$ | – | – | – | – | 1.0086 | 1.0098 | (0.9382, 1.0717) |
| $R_2$ | 0.6645 | 0.8243 | 0.8243 | (0.7992, 0.8478) | 0.7372 | 0.7367 | (0.6653, 0.7986) |
| $R_3$ | 0.2454 | 0.3336 | 0.3334 | (0.3146, 0.3543) | 0.2950 | 0.2955 | (0.2689, 0.3185) |
| $R_4$ | 0.0653 | 0.0305 | 0.0303 | (0.0163, 0.0481) | 0.0627 | 0.0628 | (0.0562, 0.0685) |
| $R_5$ | 0.0139 | 0.0525 | 0.0540 | (0.0302, 0.0697) | 0.0137 | 0.0137 | (0.0123, 0.0150) |
| $R_6$ | 0.0025 | 0.0026 | 0.0025 | (0.0012, 0.0042) | 0.0049 | 0.0049 | (0.0043, 0.0054) |

## Graphical representation of the simulation results:



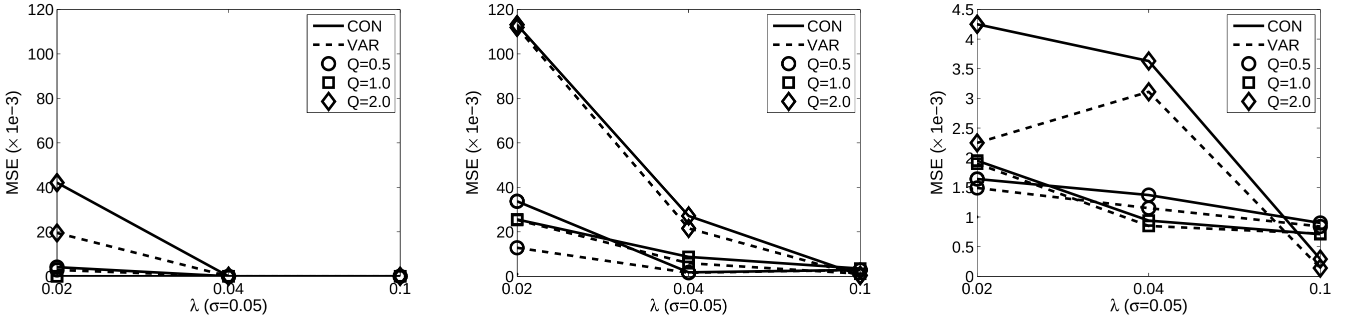
(a) **A** (b) **E1** (c) **E2**

**Figure 8.2:** Graphical representation of the MSE of $\lambda$ for settings with $\sigma = 0.05$.



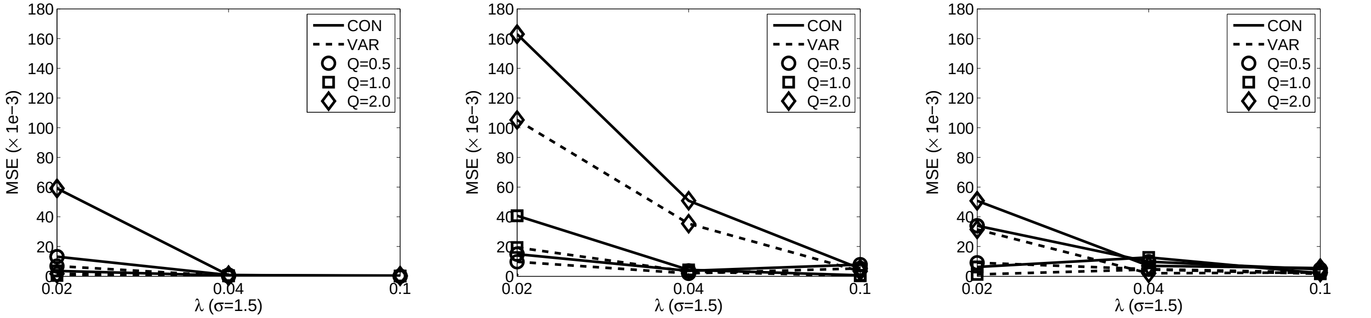
(a) **A** (b) **E1** (c) **E2**

**Figure 8.3:** Graphical representation of the MSE of $\lambda$ for settings with $\sigma = 1.50$ (excluding $\lambda = 0.10$).



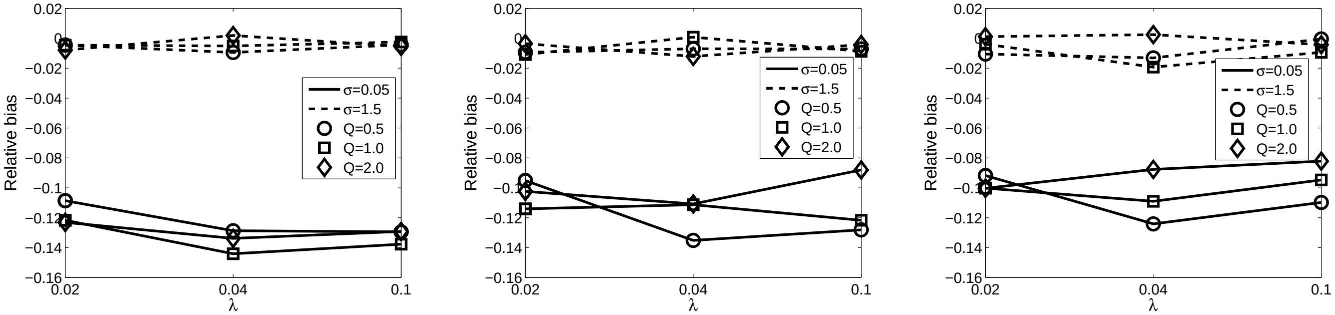
(a) **A** (b) **E1** (c) **E2**

**Figure 8.4:** Graphical representation of the MSE of $Q$ for settings with $\sigma = 0.05$.



(a) **A** (b) **E1** (c) **E2**

**Figure 8.5:** Graphical representation of the MSE of $Q$ for settings with $\sigma = 1.50$.



(a) Breen (b) E1 (c) E2

**Figure 8.6:** Graphical representation of the mean relative bias of $\theta$.

# Chapter 9

# A Bayesian approach for the analysis of $^{18}$O-labeled mass spectra using a heteroscedastic random-effect Markov-chain-based model

In this chapter, we extend the Bayesian model, defined in Chapter 8, by incorporating the random effect(s) to account for the between-spectra technical/biological variability. To investigate the performance of the model, we apply it to both the bovine cytochrome C data set and to data from a simulation study.

## 9.1 Model formulation

### 9.1.1 The likelihood

Based on the model presented in Chapter 8, to account for the between-spectra (biological) variability, we need to define spectrum-specific relative abundance parameter, i.e., $Q_i$ for the $i$th spectrum.

For the model implementation, we assume no covariance between random effects $H_i$ and $Q_i$, i.e., $\sigma_{HQ} = 0$, as a priori such an association would not be expected (as can be observed from Chapter 7). The resulting likelihood is then shown in equations (7.1)–(7.2) and (7.3), by assuming that$H_i$ and $Q_i$ are random, i.e.,

$$\begin{aligned} H_i &\sim N\left(H, \sigma_H^2\right), &(9.1)\\ Q_i &\sim N\left(Q, \sigma_Q^2\right). &(9.2)\end{aligned}$$

### 9.1.2 Prior and posterior distributions

The priors for the parameters were defined in the same way as explained in Section 8.1. For $\sigma_H^2$ and $\sigma_Q^2$, non-informative conjugate (inverse-gamma) priors were used. For $H$ and $Q$, non-informative normal priors were defined for the logarithm of them. More specifically, the following priors were used:

$$\begin{aligned} \sigma_H^{-2} &\sim \Gamma\left(\alpha_1, \beta_1\right), &(9.3)\\ \sigma_Q^{-2} &\sim \Gamma\left(\alpha_2, \beta_2\right), &(9.4)\\ H &\sim N\left(0, \frac{1}{\tau_6}\right), &(9.5)\\ Q &\sim N\left(0, \frac{1}{\tau_7}\right), &(9.6)\end{aligned}$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2, \tau_6$, and $\tau_7$ are positive constants close to zero. Same as in Chapter 8, there is no analytical solution for the posterior distributions of the parameters and they need to be evaluated via some numerical (sampling) methods.

The analysis was done using WinBUGS 1.4 through WBDiff (the interface of differential equations).

## 9.2 Results

In this section, we present results of the application of the model to the six replicated mass spectra of bovine cytochrome C peptides. We also show results of a simulation study, undertaken to check the statistical properties of the proposed model.

### 9.2.1 Bovine cytochrome C data sets

The model was applied to three peptides (at 1168.6 Da, 1456.7 Da and 1584.8 Da) of the replicated mass spectra of bovine cytochrome C (see Section 3.1).

The model was fitted to the data by using WinBUGS 1.4 through WBDiff. Tables 9.2 to 9.7 show the statistical results of the random-effect models.

Several common characteristics can be observed in these tables. First, for each peptide, a considerable amount of between-spectra variability of abundance, represented by $\sigma_H^2$, can be observed. Moreover, compared with $\sigma_H^2$, the magnitude of between-spectra variability for the relative abundance, denoted as $\sigma_Q^2$, is negligible (for model denoted as "Random $\boldsymbol{H}$ and $\boldsymbol{Q}$"). This is because, for each of the peptide, the two samples from the six spectra are basically the same. Thus, in principle they should produce exactly the same value of the relative abundance parameter.

The results are, in general, similar to the ones presented in Chapter 8 for the fixed effect model. However, there are several additional patterns that can be observed. First, it is interesting to note that the important parameters, $\lambda\tau$ and isotopic ratios $\boldsymbol{R}$ are better estimated than in the fixed-effect model. For $\boldsymbol{R}$, accounting for its precision by the 95% credible intervals, the estimates are closer to the true values than in the fixed-effect model, even for the peptide with mass 1168.6 Da. For $\lambda\tau$, the estimates of the two random-effect models are stable, i.e., for each peptide with the same relative abundance, the $\lambda\tau$ estimates are more alike and show, in general, narrower 95% credible intervals than those in the fixed-effect model, presented in the Chapter 8.

### 9.2.2 A simulation study reflecting biological variability

In this section, we present a simulation study of the setting with biological variability, by applying the Bayesian model with random $H_i$ and $Q_i$. The settings were the same as presented in Section 7.3.2.

The simulation was preformed using *R2WinBUGS*, the interface to call the application of WinBUGS 1.4 via R, with the discrete-time Markov-chain-based model implemented through WBDiff.

The results of the simulation study are presented in Table 9.1. The point estimates of the isotopic ratios, $\lambda\tau$, as well as mean relative abundance $Q$ are very close to the true values. Regarding the parameters that reflect the technical and biological variability, i.e., $\sigma_H$ and $\sigma_Q$, they are estimated with negligible bias. The estimates of power-of-the-mean variance function parameters $\theta$ and $\sigma$ also show agreement with their true values. The simulation shows satisfactory performance of the model, implemented within the Bayesian framework, under the correct model specification.

**Table 9.1:** Simulation results of the two settings – Mean estimate (M.Est.), mean relative bias (M.R.B.), empirical standard error $S_{\text{emp}}$ and model-based standard error $S_{\text{mb}}$.

| Parameter | Setting 1 | | | Setting 2 | | |
|---|---|---|---|---|---|---|
| | M.Est. | M.R.B. | $S_{\text{emp}}/S_{\text{mb}}$ | M.Est. | M.R.B | $S_{\text{emp}}/S_{\text{mb}}$ |
| $R_2$ | 0.9110 | 0.0006 | 0.0034/0.0035 | 0.9095 | -0.0010 | 0.0037/0.0060 |
| $R_3$ | 0.4186 | 0.0100 | 0.0020/0.0020 | 0.4277 | 0.0320 | 0.0024/0.0037 |
| $R_4$ | 0.1261 | 0.0022 | 0.0013/0.0013 | 0.1258 | 0.0002 | 0.0009/0.0013 |
| $R_5$ | 0.0288 | 0.0059 | 0.0005/0.0006 | 0.0290 | 0.0114 | 0.0004/0.0006 |
| $R_6$ | 0.0052 | 0.0029 | 0.0002/0.0002 | 0.0053 | 0.0157 | 0.0001/0.0002 |
| $\mu_H$ | 24030.5 | 0.0013 | 933.7/997.2 | 8080.1 | 0.0100 | 306.7/441.7 |
| $\sigma_H$ | 2024.6 | -0.0359 | 671.5/726.5 | 853.8 | -0.0298 | 290.1/353.7 |
| $\mu_Q$ | 0.3460 | 0.0381 | 0.0984/0.1160 | 2.8887 | -0.0371 | 0.2498/0.2686 |
| $\sigma_Q$ | 0.0477 | -0.0450 | 0.0129/0.0175 | 0.4762 | -0.0475 | 0.1723/0.1690 |
| $\sigma$ | 0.4533 | 0.1332 | 0.1828/0.2496 | 0.3563 | -0.1093 | 0.1169/0.1813 |
| $\theta$ | 0.5955 | -0.0076 | 0.0554/0.0567 | 0.6405 | 0.0675 | 0.0480/0.0763 |
| $\lambda_T$ | 7.9933 | -0.0484 | 0.3317/0.4205 | 7.8722 | -0.0628 | 0.0633/0.0761 |

## 9.3 Discussion

In this chapter, we have introduced a Bayesian model, with random effects, for the estimation of technical and biological variability.

The application of the models to the bovine cytochrome C data set, in general, gives unbiased estimation. The estimates of the most important parameters are improved with smaller bias and better precision, as compared to the fixed-effect models. Thus, the incorporation of various sources of variability, in the context of MS data, by the inclusion of one or more random effects is useful.

The results of the application of our method to data from a real-life, controlled experiment and to a simulation study confirm feasibility and satisfactory performance of the proposed modeling approach. The computational speed of the model is tolerable can be feasibly implemented in a high-throughput environment. On average, fitting the model, in the Bayesian framework, for each peptide, took about 3 minutes on a HP8530p laptop under Windows Vista®. The model can be extended to, e.g., the shape representation of the peaks. In the next chapter, we will discuss the implementation and applications of such a model.

## Statistical results of the case study:

**Table 9.2:** Results of the analysis of the data for $Q = 3/1$ at 1584.8Da.

| Parameter | TRUE | Random $\boldsymbol{H}$ | | | Random $\boldsymbol{H}$ and $\boldsymbol{Q}$ | | |
|---|---|---|---|---|---|---|---|
| | | Mean | Median | 95% c.i. | Mean | Median | 95% c.i. |
| $\mu_H$ | – | 8559 | 8543 | (7675, 9524) | 8618 | 8623 | (7676, 9455) |
| $\sigma^2_H$ | – | 1208000 | 826700 | (281700, 4502000) | 1154000 | 787800 | (261100, 4305000) |
| $Q$ | *2.4* | 2.3760 | 2.3770 | (2.3440, 2.4070) | 2.3660 | 2.3650 | (2.2870, 2.4430) |
| $\sigma^2_Q$ | – | – | – | – | 0.0097 | 0.0069 | (0.0023, 0.0348) |
| $\lambda\tau$ | – | 8.9520 | 8.9400 | (8.5440, 9.4560) | 8.9520 | 8.9400 | (8.5200, 9.4800) |
| $\sigma$ | – | 0.2575 | 0.2031 | (0.0729, 0.7070) | 0.5984 | 0.4202 | (0.1118, 1.4142) |
| $\theta$ | – | 0.7493 | 0.7548 | (0.6101, 0.8801) | 0.6586 | 0.6532 | (0.5089, 0.8212) |
| $R_2$ | 0.8703 | 0.8648 | 0.8648 | (0.8531, 0.8769) | 0.8606 | 0.8607 | (0.8483, 0.8728) |
| $R_3$ | 0.4223 | 0.4309 | 0.4308 | (0.4225, 0.4394) | 0.4277 | 0.4275 | (0.4207, 0.4353) |
| $R_4$ | 0.1478 | 0.1366 | 0.1366 | (0.1327, 0.1404) | 0.1359 | 0.1359 | (0.1328, 0.1395) |
| $R_5$ | 0.0413 | 0.0340 | 0.0340 | (0.0327, 0.0354) | 0.0339 | 0.0339 | (0.0326, 0.0353) |
| $R_6$ | 0.0098 | 0.0074 | 0.0074 | (0.0070, 0.0078) | 0.0074 | 0.0074 | (0.0069, 0.0079) |

**Table 9.3:** Results of the analysis of the data for $Q = 1/3$ at 1584.8Da.

| Parameter | TRUE | Random $\boldsymbol{H}$ | | | Random $\boldsymbol{H}$ and $\boldsymbol{Q}$ | | |
|---|---|---|---|---|---|---|---|
| | | Mean | Median | 95% c.i. | Mean | Median | 95% c.i. |
| $H$ | – | 22290 | 22300 | (19970, 24520) | 22220 | 22220 | (20020, 24460) |
| $\sigma^2_H$ | – | 7128000 | 4959000 | (1674000, 25470000) | 7498000 | 5353000 | (1797000, 26630000) |
| $Q$ | 0.3333 | 0.3314 | 0.3314 | (0.3271, 0.3360) | 0.3169 | 0.3173 | (0.2855, 0.3522) |
| $\sigma^2_Q$ | – | – | – | – | 0.0056 | 0.0042 | (0.0015, 0.0170) |
| $\lambda\tau$ | – | 6.5640 | 6.5520 | (6.0960, 7.0680) | 6.6324 | 6.6132 | (6.2184, 7.1712) |
| $\sigma$ | – | 0.6359 | 0.5341 | (0.2134, 1.8740) | 0.5557 | 0.4368 | (0.1813, 1.1916) |
| $\theta$ | – | 0.5947 | 0.5957 | (0.4492, 0.7124) | 0.6056 | 0.6028 | (0.4813, 0.7215) |
| $R_2$ | 0.8703 | 0.8596 | 0.8595 | (0.8527, 0.8662) | 0.8576 | 0.8576 | (0.8497, 0.8652) |
| $R_3$ | 0.4223 | 0.4053 | 0.4053 | (0.4009, 0.4096) | 0.4046 | 0.4047 | (0.4001, 0.4088) |
| $R_4$ | 0.1478 | 0.1245 | 0.1245 | (0.1212, 0.1278) | 0.1247 | 0.1247 | (0.1219, 0.1275) |
| $R_5$ | 0.0413 | 0.0336 | 0.0335 | (0.0317, 0.0355) | 0.0334 | 0.0334 | (0.0319, 0.0352) |
| $R_6$ | 0.0098 | 0.0085 | 0.0085 | (0.0076, 0.0092) | 0.0085 | 0.0085 | (0.0078, 0.0092) |

**Table 9.4:** Results of the analysis of the data for $Q = 3/1$ at 1456.7Da.

| Parameter | TRUE | Random $\boldsymbol{H}$ | | | Random $\boldsymbol{H}$ and $\boldsymbol{Q}$ | | |
|---|---|---|---|---|---|---|---|
| | | Mean | Median | 95% c.i. | Mean | Median | 95% c.i. |
| $H$ | – | 8855 | 8895 | (7380, 9795) | 8873 | 8877 | (7472, 9787) |
| $\sigma_H^2$ | – | 1591000 | 1032000 | (351000, 5946000) | 1307000 | 891000 | (303500, 4687000) |
| $Q$ | *2.4* | 2.3940 | 2.3940 | (2.3640, 2.4230) | 2.3705 | 2.3705 | (2.3062, 2.4536) |
| $\sigma_Q^2$ | – | – | – | – | 0.0076 | 0.0055 | (0.0018, 0.0272) |
| $\lambda\tau$ | – | 12.6960 | 12.0320 | (9.7440, 13.1840) | 11.2320 | 11.1600 | (10.2480, 12.6720) |
| $\sigma$ | – | 0.4318 | 0.3786 | (0.1885, 1.0330) | 0.6001 | 0.4426 | (0.1756, 1.3149) |
| $\theta$ | – | 0.6720 | 0.6758 | (0.5560, 0.7625) | 0.6431 | 0.6444 | (0.5182, 0.7595) |
| $R_2$ | 0.7933 | 0.7763 | 0.7762 | (0.7661, 0.7865) | 0.7774 | 0.7775 | (0.7684, 0.7865) |
| $R_3$ | 0.3567 | 0.3405 | 0.3405 | (0.3349, 0.3462) | 0.3432 | 0.3432 | (0.3380, 0.3486) |
| $R_4$ | 0.1166 | 0.0984 | 0.0984 | (0.0959, 0.1011) | 0.1014 | 0.1014 | (0.0988, 0.1039) |
| $R_5$ | 0.0306 | 0.0214 | 0.0214 | (0.0203, 0.0223) | 0.0245 | 0.0245 | (0.0234, 0.0255) |
| $R_6$ | 0.0068 | 0.0044 | 0.0044 | (0.0041, 0.0048) | 0.0070 | 0.0070 | (0.0066, 0.0075) |

**Table 9.5:** Results of the analysis of the data for $Q = 1/3$ at 1456.7Da.

| Parameter | TRUE | Random $\boldsymbol{H}$ | | | Random $\boldsymbol{H}$ and $\boldsymbol{Q}$ | | |
|---|---|---|---|---|---|---|---|
| | | Mean | Median | 95% c.i. | Mean | Median | 95% c.i. |
| $H$ | – | 22750 | 22810 | (20460, 24730) | 22980 | 23000 | (20780, 25010) |
| $\sigma_H^2$ | – | 6779000 | 4476000 | (1537000, 25910000) | 6455000 | 4506000 | (1503000, 23250000) |
| $Q$ | 0.3333 | 0.3319 | 0.3320 | (0.3239, 0.3397) | 0.3337 | 0.3308 | (0.3034, 0.3698) |
| $\sigma_Q^2$ | – | – | – | – | 0.0055 | 0.0041 | (0.0015, 0.0177) |
| $\lambda\tau$ | – | 6.8700 | 6.8382 | (6.2292, 7.7088) | 6.9060 | 6.8532 | (6.1824, 7.8924) |
| $\sigma$ | – | 0.3464 | 0.3085 | (0.1481, 0.7885) | 0.4263 | 0.3606 | (0.1613, 0.8165) |
| $\theta$ | – | 0.7338 | 0.7362 | (0.6204, 0.8303) | 0.6967 | 0.6944 | (0.5961, 0.8021) |
| $R_2$ | 0.7933 | 0.7857 | 0.7857 | (0.7723, 0.7987) | 0.7807 | 0.7809 | (0.7659, 0.7953) |
| $R_3$ | 0.3567 | 0.3359 | 0.3358 | (0.3284, 0.3434) | 0.3337 | 0.3336 | (0.3269, 0.3409) |
| $R_4$ | 0.1166 | 0.0915 | 0.0916 | (0.0877, 0.0954) | 0.0909 | 0.0909 | (0.0872, 0.0949) |
| $R_5$ | 0.0306 | 0.0213 | 0.0213 | (0.0197, 0.0233) | 0.0213 | 0.0212 | (0.0199, 0.0228) |
| $R_6$ | 0.0068 | 0.0058 | 0.0058 | (0.0053, 0.0064) | 0.0058 | 0.0058 | (0.0052, 0.0064) |

**Table 9.6:** Results of the analysis of the data for $Q = 3/1$ at 1168.6Da.

| Parameter | TRUE | Random $\boldsymbol{H}$ | | | Random $\boldsymbol{H}$ and $\boldsymbol{Q}$ | | |
|---|---|---|---|---|---|---|---|
| | | Mean | Median | 95% c.i. | Mean | Median | 95% c.i. |
| $H$ | – | 21910 | 22490 | (13770, 27780) | 22020 | 22000 | (14500, 28600) |
| $\sigma_H^2$ | – | 91880000 | 60750000 | (19490000, 350500000) | 62480000 | 45340000 | (15640000, 209200000) |
| $Q$ | *2.4* | 2.0900 | 2.0910 | (2.0120, 2.1680) | 2.1359 | 2.1364 | (2.0307, 2.2426) |
| $\sigma_Q^2$ | – | – | – | – | 0.0102 | 0.0073 | (0.0023, 0.0361) |
| $\lambda\tau$ | – | 17.9520 | 18.3960 | (16.2480, 19.2600) | 18.8160 | 19.2720 | (16.0440, 19.8720) |
| $\sigma$ | – | 0.1494 | 0.1369 | (0.0799, 0.2930) | 0.1314 | 0.0999 | (0.0541, 0.2940) |
| $\theta$ | – | 0.9155 | 0.9188 | (0.8379, 0.9776) | 0.9310 | 0.9362 | (0.8233, 1.0044) |
| $R_2$ | 0.6645 | 0.6962 | 0.6959 | (0.6692, 0.7238) | 0.6974 | 0.6970 | (0.6618, 0.7223) |
| $R_3$ | 0.2454 | 0.2810 | 0.2808 | (0.2684, 0.2949) | 0.2816 | 0.2813 | (0.2689, 0.2947) |
| $R_4$ | 0.0653 | 0.0682 | 0.0682 | (0.0640, 0.0727) | 0.0686 | 0.0686 | (0.0647, 0.0737) |
| $R_5$ | 0.0139 | 0.0118 | 0.0118 | (0.0110, 0.0127) | 0.0124 | 0.0124 | (0.0115, 0.0134) |
| $R_6$ | 0.0025 | 0.0015 | 0.0015 | (0.0013, 0.0016) | 0.0017 | 0.0017 | (0.0015, 0.0020) |

**Table 9.7:** Results of the analysis of the data for $Q = 1/3$ at 1168.6Da.

| Parameter | TRUE | Random $\boldsymbol{H}$ | | | Random $\boldsymbol{H}$ and $\boldsymbol{Q}$ | | |
|---|---|---|---|---|---|---|---|
| | | Mean | Median | 95% c.i. | Mean | Median | 95% c.i. |
| $H$ | – | 66370 | 66080 | (57630, 77740) | 69250 | 69230 | (57270, 81200) |
| $\sigma_H^2$ | – | 138800000 | 88830000 | (29870000, 499800000) | 193000000 | 128700000 | (39500000, 736800000) |
| $Q$ | 0.3333 | 0.5112 | 0.5117 | (0.4815, 0.5416) | 0.4872 | 0.4852 | (0.4087, 0.5703) |
| $\sigma_Q^2$ | – | – | – | – | 0.0093 | 0.0065 | (0.0021, 0.0332) |
| $\lambda\tau$ | – | 5.1468 | 5.1492 | (4.3824, 5.8620) | 5.3568 | 5.3220 | (4.6548, 6.2460) |
| $\sigma$ | – | 1.1080 | 0.9935 | (0.5099, 2.2170) | 1.3561 | 1.1476 | (0.5704, 2.5538) |
| $\theta$ | – | 0.7785 | 0.7813 | (0.6939, 0.8574) | 0.7448 | 0.7461 | (0.6583, 0.8279) |
| $R_2$ | 0.6645 | 0.8021 | 0.8004 | (0.7561, 0.8578) | 0.7625 | 0.7617 | (0.7167, 0.8133) |
| $R_3$ | 0.2454 | 0.3203 | 0.3202 | (0.2949, 0.3466) | 0.3049 | 0.3048 | (0.2798, 0.3303) |
| $R_4$ | 0.0653 | 0.0564 | 0.0563 | (0.0485, 0.0648) | 0.0556 | 0.0554 | (0.0482, 0.0643) |
| $R_5$ | 0.0139 | 0.0097 | 0.0096 | (0.0080, 0.0123) | 0.0095 | 0.0095 | (0.0074, 0.0119) |
| $R_6$ | 0.0025 | 0.0013 | 0.0013 | (0.0010, 0.0017) | 0.0014 | 0.0014 | (0.0009, 0.0020) |

# Chapter 10

# A model for the analysis of $^{18}$O-labeled mass spectra with shape representation of the data

In this chapter, we introduce an extended model, which takes the shape of the peak envelopes into account. A straightforward advantage of defining a shape model is that it can be directly applied to the original data set, for which all the information content is preserved. The resulting parameter estimates are expected to be more precise.

## 10.1 Model formulation

In this section, we first describe the model formulation for both the mean and variance structure. The corresponding estimation approach will be introduced afterwards.

### 10.1.1 Mean structure

The mean structure of the model is intrinsically the same as shown in (6.2), but with a multiplicative factor of a shape function. For the shape function, we consider an asymmetric Laplace distribution function, to account for the right-skewness nature of the peak envelopes (as can be observed from Figure 2.9). To model the observed peak intensities, assuming $l$ isotopic variants are observed for each peptide, we define that the mean intensity $\mu_{ij}$ of the observed one $y_{ij}$, which is the $j$th observed intensity from the $i$th spectrum, to be

$\mu_{ij} \equiv \mathrm{E}(y_{ij}) =$

$$\begin{cases} H_iR_m\psi(x_{ij};M_1+(m-1)S,\sigma_s,\kappa) & \\ \quad +QH_i\psi(x_{ij};M_2+(m-5)S,\sigma_s,\kappa)\sum_{k=0}^{min(4,m-1)}P_kR_{m-k} & \text{if } 1\le m\le l \\ QH_i\psi(x_{ij};M_2+(m-5)S,\sigma_s,\kappa)\sum_{k=m-l}^{4}P_kR_{m-k} & \text{if } l+1\le m\le l+4 \end{cases}, \tag{10.1}$$

where $x_{ij}$ is the $j$th mass coordinate in the $i$th spectrum; $m$ is the index for the observed peak; $M_1$ and $M_2$ are, respectively, the monoisotopic masses of the unlabeled and labeled peptide samples. Parameter $S$ reflects the average difference in molecular weight of the isotopes and is usually very close to one. Thus, in (10.1), $S$ is defined to be the difference in mass locations between two neighboring isotopic peaks of the same peptide, and is assumed to be constant over all the isotopic peaks for both peptide samples. Parameters $\sigma_s$ and $\kappa$ are, respectively, the spread and skewness parameters for the asymmetric Laplace distribution function $\psi(x_{ij};M+(m-1)S,\sigma_s,\kappa)$, where $\psi(x_{ij};M+(m-1)S,\sigma_s,\kappa)=$

$$\begin{cases} F(x_{ij}|M+(m-1)S,\sigma_s,\kappa)-F(x_{ij-1}|M+(m-1)S,\sigma_s,\kappa) & \text{if } j\ge 2, \\ F(x_{ij}|M+(m-1)S,\sigma_s,\kappa) & \text{if } j=1, \end{cases} \tag{10.2}$$

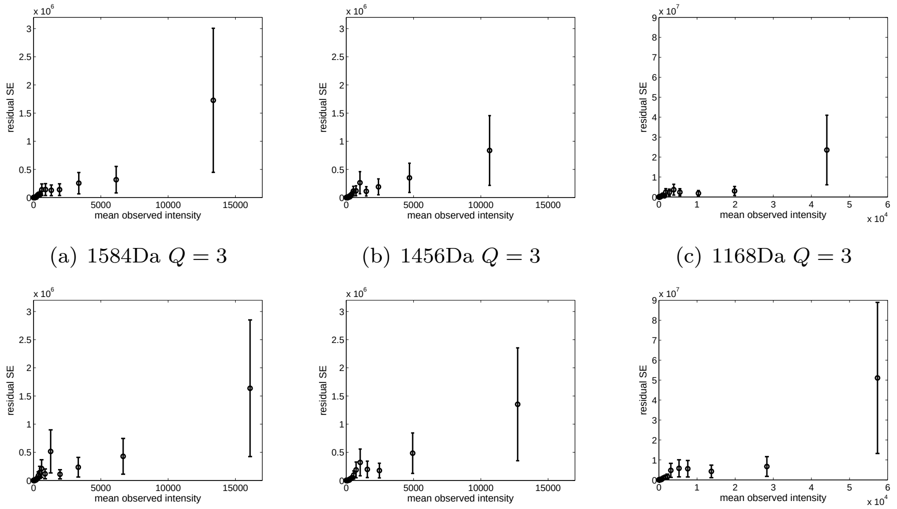
with $F(x_{ij}|M+(m-1)S,\sigma_s,\kappa)$ the *cdf* function of asymmetric Laplace calculated at $x_{ij}$ with mean $M+(m-1)S$ and standard deviation $\sigma_s$, i.e., $F(x_{ij}|M+(m-1)S,\sigma_s,\kappa)=$

$$\begin{cases} \frac{\kappa^2}{1+\kappa^2}\exp\left[-\frac{\sqrt{2}}{\sigma_s\kappa}|x_{ij}-(M+(m-1)S)\,|\right] & \text{if } x_{ij}<M+(m-1)S, \\ 1-\frac{1}{1+\kappa^2}\exp\left[-\frac{\sqrt{2}\kappa}{\sigma_s}|x_{ij}-(M+(m-1)S)\,|\right] & \text{if } x_{ij}\ge M+(m-1)S. \end{cases}$$

Similar to the models for the stick representation, terms $H_iR_m\psi(x_{ij};M_1+(m-1)S,\sigma_s,\kappa)$ and $QH_i\psi(x_{ij};M_2+(m-)S,\sigma_s,\kappa)P_kR_{m-k}$ in (10.1) denote the contributions to the mean values of the $j$th observed intensity from the $m$th isotopic variant from Samples I and II respectively.

### 10.1.2 Data exploration for a suitable variance function

For the pre-processing algorithm, we used the strategy proposed by Valkenborg *et al.* (2009). For the baseline correction, however, using the algorithm, implemented by Valkenborg *et al.* (2009), turned out to produce numerical problems due to many zero intensity values after baseline correction. This caused problems in evaluating a suitable variance function and in estimating the chosen variance function. Alternatively, we corrected the baseline by subtracting the minimum intensity value of the $i$th spectrum across the $m + 4$ observed peaks for a certain peptide samples. This is empirically valid since the baseline in around 10 Da mass range is fairly constant.



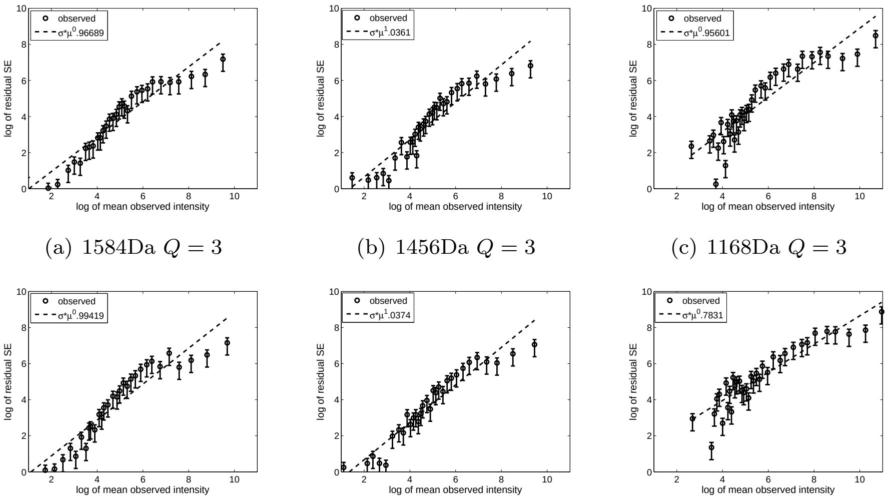
(a) 1584Da $Q = 3$ (b) 1456Da $Q = 3$ (c) 1168Da $Q = 3$

(d) 1584Da $Q = 1/3$ (e) 1456Da $Q = 1/3$ (f) 1168Da $Q = 1/3$

**Figure 10.1:** Scatter plots of grouped residual standard error with 95% confidence interval (on $y$) versus mean (observed) intensity (on $x$).

In order to find a suitable variance function for the residual variance of the model, a homoscedastic model with a mean structure defined in (10.1) was fitted, using least squares, to the bovine cytochrome C data set (see Section 3.1). The residuals were first ordered according to their corresponding observed intensity values and then grouped. The variances and the mean (observed) intensities were calculated based on the grouped residuals.

Figure 10.1 shows the scatter plots of the residual standard errors, together with



(a) 1584Da $Q = 3$ (b) 1456Da $Q = 3$ (c) 1168Da $Q = 3$

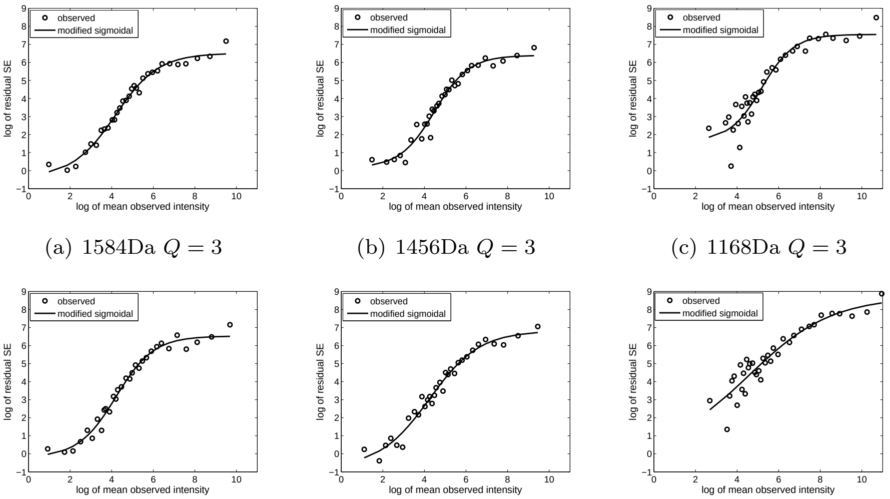
(d) 1584Da $Q = 1/3$ (e) 1456Da $Q = 1/3$ (f) 1168Da $Q = 1/3$

**Figure 10.2:** Scatter plots of the logarithmic residual standard error, with 95% confidence interval, versus the logarithmic mean (observed) intensity, to check the appropriateness of the power-of-the-mean variance function.

the corresponding 95% confidence intervals, versus the mean (observed) intensities. It is apparent from Figure 10.1 that the fit is far from linear and the scatter of the observations mostly concentrate around low intensity values. Moreover, the widths of the 95% confidence intervals are far from constant, but increase drastically along the intensity value scale. Thus, a logarithmic transformation was considered for exploring a suitable variance function. The scatter plots were then re-drawn using the logarithmic transformation for both scales. The result is shown in Figure 10.2. The 95% confidence intervals are now fairly constant. However, the scatters show a sigmoidal shape.

Figure 10.2 indicates that for the model of the shape representation, instead of applying a power-of-the-mean variance function as defined in Chapter 6, a sigmoidal variance function should be considered. However, a conventional sigmoidal function is limited in both the scale and the shape. Thus, we considered the modified sigmoidal function:

$$V = d + \frac{a}{1 + \exp\left[-b(U - c)\right]}, \tag{10.3}$$



(a) 1584Da $Q = 3$ (b) 1456Da $Q = 3$ (c) 1168Da $Q = 3$

(d) 1584Da $Q = 1/3$ (e) 1456Da $Q = 1/3$ (f) 1168Da $Q = 1/3$

**Figure 10.3:** Scatter plots of the logarithm of residual standard error versus the logarithm of mean (observed) intensity and the fitted sigmoidal model (10.3).

where $a > 0$, $b > 0$, and $V$ and $U$ are, respectively, the logarithm of the residual standard error and the logarithm of the observed intensity $y_{ij}$. The modified sigmoidal function, defined in (10.3), was fitted to the residuals, obtained from the homoscedastic model, using least squares approach. Figure 10.3 indicates good fit of the proposed modified sigmoidal function for all the six data sets.

The flexibility of the model is reflected by the different aspects of the model shape being controlled by different parameters. More specifically, parameter $a$ influences the elongation of the sigmoidal shape along the $y$ axis, while parameter $b$ controls the stretch of the sigmoidal shape along the $x$ axis and the direction of the sigmoid (if $b > 0$, the function is monotonely increasing; if $b < 0$, it's monotonely decreasing). Parameters $c$ and $d$ adjust the shift of the sigmoid on the $x$ and $y$ axes, respectively. Another advantage of using the modified sigmoidal function as a variance function is that the additive factor $d$ on the log-scale becomes multiplicative on the original scale. This means the $\exp(d)$ can be viewed as a common baseline residual variance $\sigma^2$, which can be profiled out when, for instance, a PL-GLS algorithm is considered. Therefore, in the PL step to estimate the variance function parameters, only three parameters $a$, $b$ and $c$ need to be estimated. The details of PL-GLS algorithm for the model will be discussed in Section 10.2.

Thus, the resulting model takes the following form:

$$y_{ij} \quad = \mu_{ij} + \varepsilon_{ij}, \tag{10.4}$$

where $\varepsilon_{ij} \sim N\left(0, \sigma^2 g^2\left(\mu_{ij}; a, b, c\right)\right)$, with

$$g\left(\mu_{ij}; a, b, c\right) \quad = \quad \exp\left(\frac{a}{1+\exp\left\{-b\left[\log\left(\mu_{ij}\right)-c\right]\right\}}\right), \; a > 0, \; b > 0, \tag{10.5}$$

and that $\varepsilon_{ij}$'s are independent. The mean intensity $\mu_{ij}$ is defined in equation (10.1).

## 10.2 Estimation and inference

Assume that we have got $n$ joint spectra, each with $M$ observed mass (or intensity) coordinates. The model, specified by (10.1)–(10.5), can be fitted to observed data by using various methods (Carroll and Ruppert 1988, Davidian and Giltinan 1995). The starting point for them is the log-likelihood, given by
$l_{\text{ML}}(\boldsymbol{\beta}, a, b, c, \sigma^2) =$

$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{M}\log\left\{\sigma^2 g^2\left(\mu_{ij}(\boldsymbol{\beta}); a, b, c\right)\right\} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{M}\left\{\frac{y_{ij}-\mu_{ij}(\boldsymbol{\beta})}{g\left(\mu_{ij}(\boldsymbol{\beta}); a, b, c\right)}\right\}^2, \tag{10.6}$$

where $\boldsymbol{\beta} = (H_1, \ldots, H_n, Q, \lambda, R_1, \ldots, R_l, M_1, M_2, S, \sigma_s, \kappa)$ is a parameter vector that includes all the parameters used to model the mean value, as specified in (10.1).

Maximum-likelihood (ML) estimates of $\boldsymbol{\beta}$, $a$, $b$, $c$, and $\sigma^2$ can be obtained by simultaneously maximizing log-likelihood function (10.6) with respect to these parameters. In general, however, this is a numerically complex task, which requires finding an optimum in a multidimensional parameter space. This task can be simplified by observing that, if we assume $\boldsymbol{\beta}$, $a$, $b$, and $c$ are *known*, the estimator for $\sigma^2$ is given by

$$\widehat{\sigma}^2_{\text{ML}} = \frac{1}{nM}\sum_{i=1}^{n}\sum_{j=1}^{M}\left\{\frac{y_{ij}-\mu_{ij}(\boldsymbol{\beta})}{g\left(\mu_{ij}(\boldsymbol{\beta}); a, b, c\right)}\right\}^2. \tag{10.7}$$

By plugging expression (10.7) in (10.6) and omitting constant terms, we obtain the following *log-profile-likelihood* function, which depends only on $a$, $b$, $c$, and $\boldsymbol{\beta}$ (not on

$\sigma^2$):

$$l^*_{\text{ML}}(\boldsymbol{\beta}, a, b, c) =$$

$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{M}\log\left\{g^2\left(\mu_{ij}(\boldsymbol{\beta}); a, b, c\right)\right\} - \frac{nM}{2}\log\left[\sum_{i=1}^{n}\sum_{j=1}^{M}\left\{\frac{y_{ij}-\mu_{ij}(\boldsymbol{\beta})}{g\left(\mu_{ij}(\boldsymbol{\beta}); a, b, c\right)}\right\}^2\right] =$$

$$-\frac{nM}{2}\log\left[\left\{\prod_{i=1}^{n}\prod_{j=1}^{M}g\left(\mu_{ij}(\boldsymbol{\beta}); a, b, c\right)\right\}^{\frac{2}{nM}}\right] - \frac{nM}{2}\log\left[\sum_{i=1}^{n}\sum_{j=1}^{M}\left\{\frac{y_{ij}-\mu_{ij}(\boldsymbol{\beta})}{g\left(\mu_{ij}(\boldsymbol{\beta}); a, b, c\right)}\right\}^2\right]. \tag{10.8}$$

Maximizing (10.8) with respect to $a$, $b$, $c$, and $\boldsymbol{\beta}$ allows obtaining estimates for these parameters. The estimates can then be used to compute the ML-estimate of $\sigma^2$ from (10.7). However, it is well known that the ML-estimator is biased downwards. Thus, especially when the number of spectra is small, it is better to replace it by the following REML-estimator:

$$\widehat{\sigma}^2_{\text{REML}} = \frac{1}{(nM-p)}\sum_{i=1}^{n}\sum_{j=1}^{M}\left\{\frac{y_{ij}-\mu_{ij}(\boldsymbol{\beta})}{g\left(\mu_{ij}(\boldsymbol{\beta}); a, b, c\right)}\right\}^2, \tag{10.9}$$

where $p$ denotes the number parameters in the model.

The use of log-profile-likelihood (10.8) still requires a simultaneous maximization of the function over $\boldsymbol{\beta}$, $a$, $b$ and $c$. Moreover, the use of log-likelihood (10.6) or of log-profile-likelihood (10.8) assumes that the data fulfill all the assumptions of the model, defined in (10.1)–(10.4) and (10.5).

An alternative estimation approach is to use a PL-GLS algorithm (Davidian and Giltinan 1995), which is more robust to mis-specifications of the model and simpler numerically (see Section 6.2). In the case of the modified sigmoidal variance, as specified in (10.5), the approach can be deduced in a similar way as presented in Section 6.2. Namely, log-profile-likelihood (10.8) can be expressed as

$$l^*_{\text{ML}}(\boldsymbol{\beta}, a, b, c) \quad = \quad -\frac{nM}{2}\log\left[\left\{\widetilde{g}\left(\mu_{ij}(\boldsymbol{\beta}); a, b, c\right)\right\}^2\sum_{i=1}^{n}\sum_{j=1}^{M}\left\{\frac{y_{ij}-\mu_{ij}(\boldsymbol{\beta})}{g\left(\mu_{ij}(\boldsymbol{\beta}); a, b, c\right)}\right\}^2\right], \tag{10.10}$$

where $\widetilde{g}\left(\mu_{ij}(\boldsymbol{\beta}); a, b, c\right) = \left\{\prod_{i=1}^{n}\prod_{j=1}^{M}g\left(\mu_{ij}(\boldsymbol{\beta}); a, b, c\right)\right\}^{\frac{1}{nM}}$. It follows that maximization of (10.10) is equivalent to minimization of

$$l^{**}_{\mathrm{ML}}(\boldsymbol{\beta},a,b,c) =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{M}\left[\{y_{ij}-\mu_{ij}(\boldsymbol{\beta})\}\left\{\frac{\widetilde{g}\left(\mu_{ij}(\boldsymbol{\beta});a,b,c\right)}{g\left(\mu_{ij}(\boldsymbol{\beta});a,b,c\right)}\right\}\right]^2 \equiv \sum_{i=1}^{n}\sum_{j=1}^{M}\{f_{ij}(\boldsymbol{\beta},a,b,c)\}^2. \tag{10.11}$$

Thus, minimization of (10.11), either over $a$, $b$ and $c$ (while keeping $\boldsymbol{\beta}$ fixed) or over $(\boldsymbol{\beta},a,b,c)$, can be viewed as an ordinary least squares (OLS) problem for a linear model with all data equal to 0 and $f_{ij}(\boldsymbol{\beta},a,b,c)$ as the fitted mean structure. It can be also viewed as a weighted least squares (WLS) problem for estimating $\boldsymbol{\beta}$, with weight

$$w_{ij}(\boldsymbol{\beta},a,b,c) = \frac{\widetilde{g}\left(\mu_{ij}(\boldsymbol{\beta});a,b,c\right)}{g\left(\mu_{ij}(\boldsymbol{\beta});a,b,c\right)}.$$

As a result, the following algorithm can be used to estimate $\boldsymbol{\beta}$, $a$, $b$, $c$, and $\sigma^2$:

1. Set $k=0$. Use an initial estimate $\widehat{\boldsymbol{\beta}}^{(0)}$ of $\boldsymbol{\beta}$.

2. Set $k=k+1$.

3. While keeping $\widehat{\boldsymbol{\beta}}^{(k-1)}$ fixed, compute estimates $\widehat{a}^{(k)}$, $\widehat{b}^{(k)}$ and $\widehat{c}^{(k)}$ of $a$, $b$ and $c$, respectively, from (10.11), by using OLS.

4. Compute weights $w_{ij}^{(k)}(\widehat{\boldsymbol{\beta}}^{(k-1)},\widehat{a}^{(k)},\widehat{b}^{(k)},\widehat{c}^{(k)})$. While keeping the weights fixed, obtain estimate $\widehat{\boldsymbol{\beta}}^{(k)}$ of $\boldsymbol{\beta}$ by using WLS.

5. Iterate between steps 2–4 until convergence.

6. Use the obtained estimates of $\boldsymbol{\beta}$ and $a$, $b$, $c$ to compute an estimate of $\sigma^2$ from (10.7) or (10.9).

Irrespectively of the estimation approach used, standard errors of the estimates of $\boldsymbol{\beta}$, $a$, $b$, $c$, and $\sigma^2$ can be obtained from the inverse of the negative Hessian of log-likelihood (10.6), computed at the estimated values of the parameters.

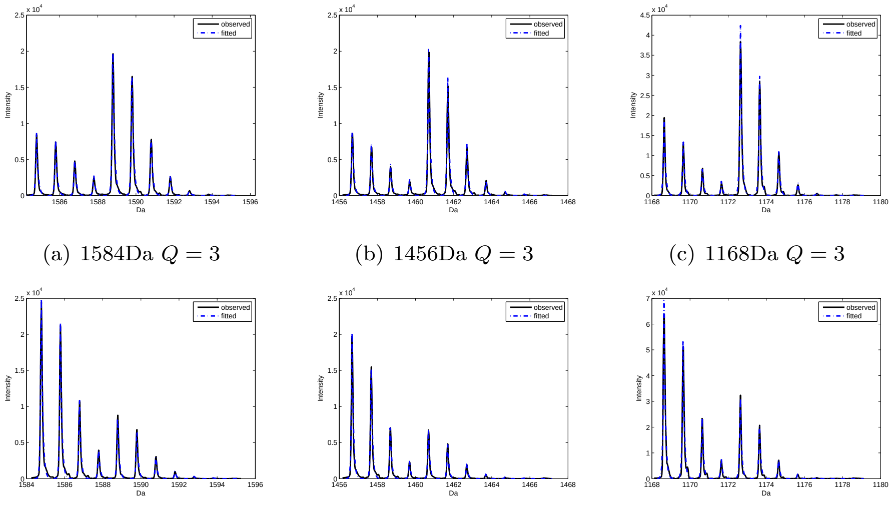
### 10.2.1 Practical implementation

For the practical implementation, we considered an unconstrained estimation approach, as explained in Section 5.2.4. More specifically, for the parameters that need to be constrained positively, a logarithmic transformation was considered. For $\lambda$, as in Chapters 5 to 9, we considered the Box-Cox transformation: $\lambda = \lambda_0\exp(\lambda')/\{\exp(\lambda')+1\}$.

It was observed that, to avoid numerical problems, parameters $M_1$, $M_2$, and $S$ need to be constrained within a certain range. For the monoisotopic masses of the two peptide samples, $M_1$ and $M_2$, they could be constrained within 1 Da range, at a particular peak observed from a mass spectrum. For this purpose, an inverse-logit transformation was used. To be more specific, suppose $M_1$ and $M_2$ are observed to vary within intervals of $[M_{1_0}, M_{1_0} + 1]$ and $[M_{2_0}, M_{2_0} + 1]$ respectively, the transformation then takes the form: $M_1 = M_{1_0} + \frac{\exp(M_1')}{1+\exp(M_1')}$ and $M_2 = M_{2_0} + \frac{\exp(M_2')}{1+\exp(M_2')}$. With this respect, the transformed scales of $M_1'$ and $M_2'$ can take any real values.

The parameter $S$, which is the difference in mass locations between two neighboring isotopic peaks, usually varies within a small range around 1.0015 Da. The value 1.0015 was obtained as an average difference in the isotopic masses. Thus, $S$ can be constrained in the interval of $[1.0000, 1.0050]$. To this aim, a re-scaled inverse-logit transformation was used, yielding $S = 1 + 0.0050\frac{\exp(S')}{1+\exp(S')}$, where the transformed scale $S'$ is allowed to take any real values.

## 10.3 Application to bovine cytochrome C data



(a) 1584Da $Q = 3$

(b) 1456Da $Q = 3$

(c) 1168Da $Q = 3$

(d) 1584Da $Q = 1/3$

(e) 1456Da $Q = 1/3$

(f) 1168Da $Q = 1/3$

**Figure 10.4:** Observed versus fitted (with predicted mean intensity) spectrum.

We apply the model to the controlled experiment of the enzymatic labeling of bovine cytochrome C peptides. The estimation approaches were implemented by using Matlab 2009a with function *fminunc* for unconstrained optimization problems. It should be noted that, unlike the models for the stick representation, due to the inclusion of the shape function, the gradient functions of the model parameters can no longer be expressed analytically. For this reason, a medium-scale search using the Gauss-Newton algorithm (instead of Newton-Raphson algorithm) was performed. The proportions of water impurities of the heavy-oxygen water were still assumed to be equal to $p_{16} = 2\%$ and $p_{17} = 1\%$. As stated earlier, to avoid numerical problem, the baseline correction algorithm proposed by Valkenborg *et al.* (2009) was not applied. Instead, the baseline was assumed to be constant for each peptide samples within each spectrum. Thus, the baseline correction was done by subtracting the minimum intensity value of a specific spectrum from the observed intensity values of the spectrum.

Tables 10.2 and 10.3 present, respectively, the PL-GLS estimates of the model for the three analyzed peptides for the controlled experiment with intended relative abundances $Q = 1/3$ and $Q = 3/1$.

Apart from the patterns of the parameter estimates observed also in Section 6.3, several improvements over the stick-representation models are worth mentioning. First, for the two peptides with masses 1456.7 Da and 1584.8 Da, the $\lambda\tau$ estimates, for the two experiments with relative abundances $Q = 1/3$ and $Q = 3/1$, are closer to each other. This is expected as the oxygen incorporation rate should be relatively constant for the same peptide. Furthermore, the 95% confidence intervals for the relative abundance $Q$ and the isotopic ratios $\boldsymbol{R}$, in general, contain the true values, especially for the two peptides with masses 1456.7 Da and 1584.8 Da. This is an important improvement over the stick-representation models. Recall that, in the stick-representation models (see Section 6.3), the standard errors were much smaller, yielding the confidence intervals slightly deviating away from the true values.

Figure 10.4 shows the comparison of the observed versus the fitted spectrum for one of the six spectra. The fitted intensities were obtained by the predicted mean intensity values of the model. The figure shows that the asymmetric Laplace function provides a good approximation for the shape of the peak envelopes. The fitted spectra are quite compatible with the observed ones, especially for the peptides with masses 1456.7 Da and 1584.8 Da.

## 10.4 A simulation study

In this section, we present results of a simulation study, undertaken to check the statistical properties of the proposed model.

### 10.4.1 Simulation settings

In the simulation study, the proportions of heavy-oxygen water impurities were assumed to be equal to $p_{16} = 2\%$ and $p_{17} = 1\%$. For this simulation study, we considered as in the bovine cytochrome C data sets, six technical replicates. Possible variability due to, e.g., laser fluctuations and inefficient crystallization, was simulated by using six different reference intensities (obtained as a rough average from the bovine cytochrome C data set), namely, $H_1 = 129500, H_2 = 126800, H_3 = 117500, H_4 = 135600, H_5 = 105500$ and $H_6 = 136000$.

The data sets were generated with combinations of settings for different parameters shown as below:

$$\begin{array}{rl}
Q: & \{0.5 \quad 2\} \\
\lambda\tau: & \{4.8 \quad 9.6\} \\
\boldsymbol{R}: & \{\mathbf{E1} \quad \mathbf{E2}\} \\
S: & 1.0015 \\
\sigma_s: & 0.0800 \\
M_1: & 2001.053 \\
M_2: & \{M_1 + 4S + 0\sigma_s \quad M_1 + 4S + 1.5\sigma_s \quad M_1 + 4S + 3\sigma_s\}
\end{array}$$

For the settings of the isotopic ratios $\boldsymbol{R}$, similar to Section 6.4, **E1** and **E2** are the isotopic distributions with the second isotopic peak being the least and most abundant among all the peptides around 2001 Da from the NCBI data. The other parameters were chosen as: $\kappa = 0.7500$, $a = 5.0000$, $b = 0.5000$, $c = 4.5000$ and $\sigma = 15.00$. The combinations of these parameters lead to 24 settings in total. Table 10.1 shows the numbering of the 24 settings with different combinations of the parameters.

For each of the settings, 200 replicated data sets, with random noise, were generated. In particular, the variance of the random noise of the replicated data sets was assumed to follow the modified sigmoidal function (10.5). The randomly generated noise, based on the assumption of (10.5), was then added to the mean intensity

**Table 10.1:** Numbering of the 24 simulation settings.

| | | $Q = 0.5$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | E1 | | | E2 | | |
| | | $M_2 : 0\sigma_s$ | $M_2 : 1.5\sigma_s$ | $M_2 : 3\sigma_s$ | $M_2 : 0\sigma_s$ | $M_2 : 1.5\sigma_s$ | $M_2 : 3\sigma_s$ |
| $\lambda\tau$ | 4.8 | set1 | set2 | set3 | set4 | set5 | set6 |
| | 9.6 | set7 | set8 | set9 | set10 | set11 | set12 |
| | | $Q = 2$ | | | | | |
| | | E1 | | | E2 | | |
| | | $M_2 : 0\sigma_s$ | $M_2 : 1.5\sigma_s$ | $M_2 : 3\sigma_s$ | $M_2 : 0\sigma_s$ | $M_2 : 1.5\sigma_s$ | $M_2 : 3\sigma_s$ |
| $\lambda\tau$ | 4.8 | set13 | set14 | set15 | set16 | set17 | set18 |
| | 9.6 | set19 | set20 | set21 | set22 | set23 | set24 |

values, and truncated to be zero if negative. Based on the truncated intensity values, a baseline was added to them. The baseline was generated by using a Gaussian density function, representing a slight 'bumping' shape as represented by the 'true baseline' in Figure 10.5. To correct for the baseline, as done also in the analysis of bovine cytochrome C data, a minimum intensity value observed for that spectrum was subtracted from the 'observed' intensity values of the spectrum.

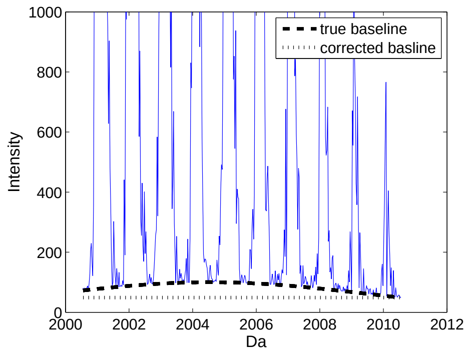


**Figure 10.5:** Graphical demonstration of baseline correction.

One of the purposes of the simulation was to check the possible improvement of the shape-representation model over the model for the stick representation, by retaining the full content of the information from the mass spectra data. To this aim, the same simulated data sets were also analyzed by the model of the stick representation, after extracting the information from the original data for the stick representation (see Section 4.2).

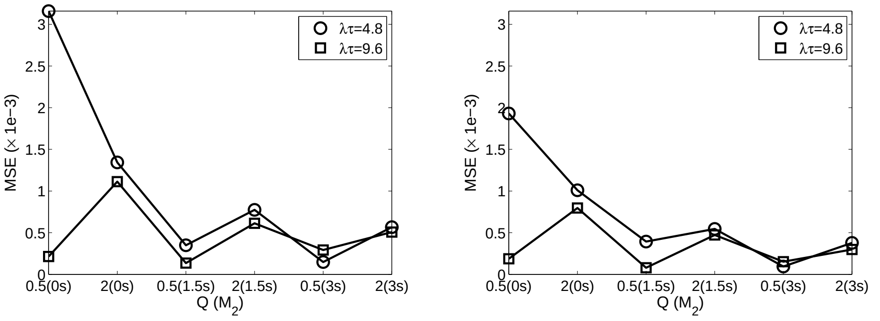
A graphical representation of the 24 settings of the simulated data sets is shown in Figures 10.9 and 10.10. The corresponding stick representation of the 24 settings is presented in Figures 10.11 and 10.12.

To apply the model for the stick representation (introduced in Chapter 6), the validity of the assumption of the power-of-the-mean variance function for the stick representation needs to be checked. Figures 10.13 and 10.14 show the scatter plots of the (grouped) logarithm of standard errors of the random noise versus the logarithm of the mean intensity values. The random noise corresponds to those intensities that represent the observed peaks, i.e., the maximum intensity of each observed peak, in the stick representation. In general, the scatter plots can be viewed as linear, which corroborates the validity of the assumption of the variance function, namely the power-of-the-mean variance function. Thus, for the analysis of the stick representation, such variance function was applied.

### 10.4.2 Results of the simulation study

Tables D.1 to D.7 (Appendix D) present a comparison of the simulation results of the models for the shape and stick representations. A common pattern, observed from these tables, when $M_2 = 4S + 1.5\sigma_s$ and $M_2 = 4S + 3\sigma_s$, is that the shape-representation model gives less biased and more precise estimates, as indicated by the mean relative bias $\bar{b}$ and the empirical standard error $S_{emp}$, respectively. For the settings with $M_2 = 4S + 1.5\sigma_s$ and $M_2 = 4S + 3\sigma_s$, i.e., when at a specific observed peak, the 'tilt' of the centroid of the peaks of the two peptide samples is larger (equal to $4S + 1.5\sigma_s$ and $4S + 3\sigma_s$) than for the settings with $M_2 = 4S + 0\sigma_s$, $\bar{b}$ and $S_{emp}$ of the shape model are considerably smaller than those of the stick-representation model. This is because, when the 'tilt' is large, the summary statistics used for the stick representation, takes only the maximum intensity of a certain peak. The summary statistics would then only reflect the information of the more abundant peptide sample, while losing most of the information content of the other, less abundant one. Thus, it illustrates the advantage of using the shape model.

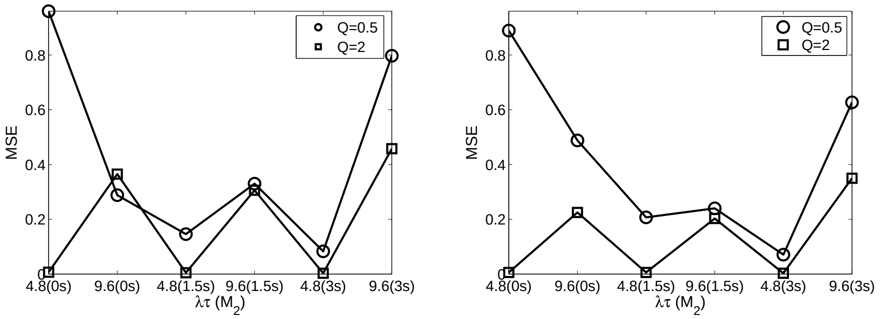
Figures 10.6 and 10.7 give a graphical representation of the MSE of $Q$ and $\lambda\tau$ of the shape model for the various settings. From Figure 10.6, it can be seen that, when $\lambda\tau$ is larger (9.6), the MSE for the relative abundance $Q$, in generally, is smaller. Note that this was also observed in the simulation for the stick representation (presented in Chapter 6). It indicates that the more complete the labeling, the better $Q$ can



(a) E1

(b) E2

**Figure 10.6:** MSE of $Q$ for various settings of the shape model.



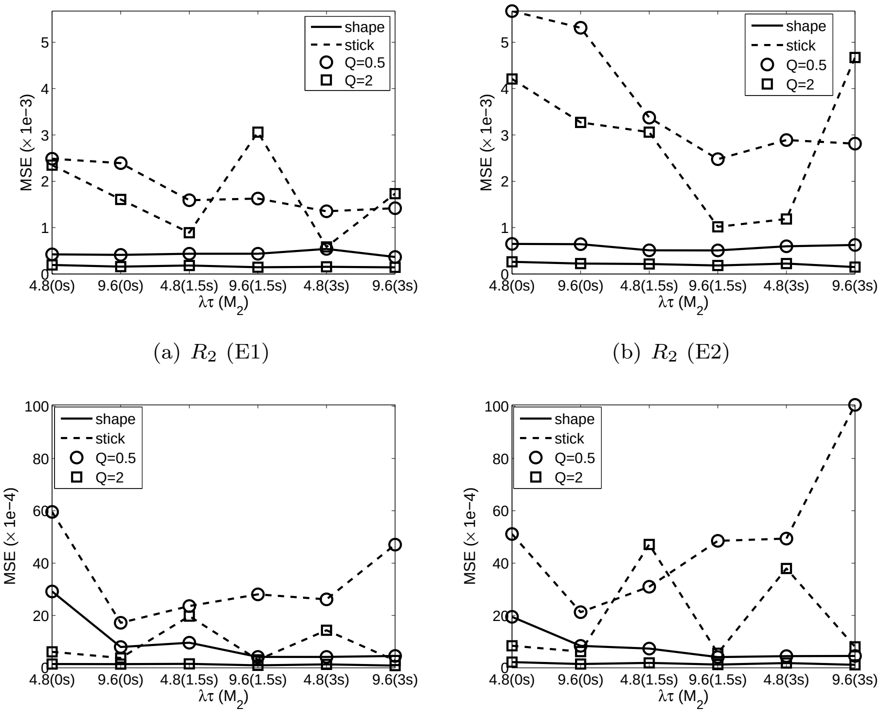
(a) E1

(b) E2

**Figure 10.7:** MSE of $\lambda\tau$ for various settings of the shape model.

be estimated. In should be noted that, again, similar to the simulation for the stick representation (presented in Chapter 6), the MSE of the product of oxygen incorporation rate and reaction duration $\lambda\tau$, shown in Figure 10.7, is smaller when the labeled peptide sample is more abundant, i.e., when $Q = 2$. This is reasonable since the information about the oxygen exchange mainly comes from the labeled peptide sample.

Figure 10.8 presents graphically the MSE for both the shape and stick (representation) models for isotopic ratios $R_2$ and $R_3$. These figures indicate that the model for the stick representation consistently show larger MSE than the shape-representation model. Moreover, for the shape model, the MSE of the ratios is smaller when the



(a) $R_2$ (E1)

(b) $R_2$ (E2)

(c) $R_3$ (E1)

(d) $R_3$ (E2)

**Figure 10.8:** *MSE* of $R_2$ and $R_3$ for settings of E1 and E2 (dashed lines: model for the stick representation; solid lines: model for the shape representation).

second (labeled) peptide sample is more abundant. This may be due to the fact that, for most of the isotopic peaks of the labeled peptides, they are not overlapped with those of the unlabeled ones. Thus, they give more information for these ratio estimates. This implies that a more abundant labeled peptide sample provides more information content for estimation of the isotopic ratios. As the results of the other isotopic ratios ($R_4$–$R_6$) show similar patterns, they are not presented graphically.

Tables D.8 to D.12 (Appendix D) show the results of $M_1$, $M_2$, $\kappa$, $\sigma_s$, and $S$ for the shape model. They are, in general, estimated with a negligible bias. It is worth mentioning, however, that the estimates of the spread parameter of the asymmetric Laplace function $\sigma_s$ (presented in Table D.11) show consistently an upward bias. This can be explained by observing in Figure 10.5 that the (constant) baseline correction does not fully correct for the true baseline, especially for the first to the seventh observed peaks. These peaks, are usually the most abundant ones. The heavy tails induced by the baseline, result in the parameter estimation of the shape function trying to adapt to these tails, and provide, consequently, an upward bias for the estimate of the spread parameter.

The results of the estimates of the parameters for the variance function are presented in Tables D.13 to D.16 (Appendix D). These results are mostly estimated with negligible bias.

The simulation confirms the satisfactory performance of the shape-representation model, and verifies its advantages than the model for the stick representation.

## 10.5 Discussion and conclusion

Most of the existing methods (Mirgorodskaya *et al.* 2000, Rao *et al.* 2005, López-Ferrer *et al.* 2006, Eckel-Passow *et al.* 2006, Ramos-Fernández *et al.* 2007) based the analyses on the stick representation, which is an important limitation. Since working with the stick representation implies an information reduction, especially when the data reflect worse quality, e.g., when the peaks of the two peptide samples are 'tilted' instead of exhibiting a complete overlap. As a result, the information reduction imposes biased estimates of the parameters of interest. As an alternative, Ramos-Fernández *et al.* (2007) proposed an analyzing approach, based on the shape representation of the MS data, by approximating the asymmetry nature of the peak envelopes using a mixture of two Gaussian functions. The use of a mixture of Gaussian functions can not only cause numerical complexity, but may lead to the non-identifiability issue, when a peak envelope can be equally well approximated by several mixture patterns. Furthermore, in their approach, the masses and isotopic distributions were assumed to be known, with the assumption of known peptide atomic sequences. This assumption is often unrealistic, as in reality the sequences or peptide chemical compositions are rarely known.

To address the limitations of these methods, in this chapter, we have presented a model for the enzymatically $^{18}$O-labeled MS for the shape representation. In the model, we used the asymmetric Laplace distribution function to approximate the peak shape envelopes, avoiding the possible non-identifiability issue. Moreover, our modeling approach allows for the estimation of both peptide masses and their isotopic distributions, releasing the assumption of known chemical compositions. In particular, the model is based on the one for the stick representation (presented in Chapter 6), by accounting for the (mean-dependent) heteroscedastic nature of the residual variance for the MS data. We implemented the model by a double iteration of PL-GLS algorithm, which is more robust than the likelihood-maximization approach when distributional assumptions are violated.

The results of the application to the real-life data were, in general, consistent with the true parameter values for two of three analyzed peptides. The estimates for the parameters of interest, in general, showed agreement with the estimates obtained for the stick representation, and their 95% confidence intervals mostly contained the true values.

In the simulation study, the relative abundance parameter was estimated with better precision when the labeling was more complete. On the other hand, the parameter, related to the labeling step, i.e., the product of oxygen incorporation rate and reaction duration, and the isotopic ratio parameters, showed better estimation when the labeled peptide sample was more abundant. This is because the labeled peptide gives the most information for the estimation of these parameters. The simulation illustrates the improved performances of the shape model compared with the model for the stick representation, when the peaks of the two peptide samples are 'tilted' instead of exhibiting a complete overlap.

The computational speed of the model, presented in this chapter, was estimated to be roughly 4 minutes for each peptide, on a HP8530p laptop using Matlab 2009a under Windows Vista®.

## Statistical results of the case study:

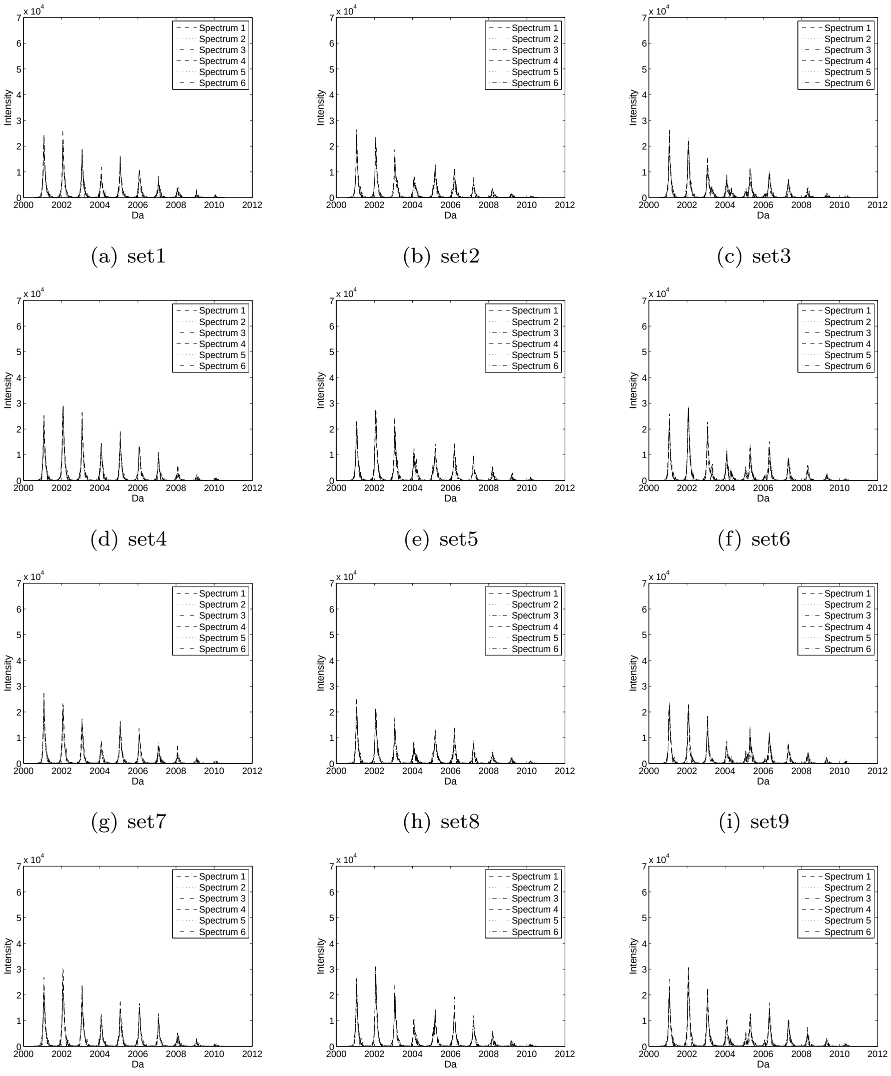
**Table 10.2:** Results of the analysis of the data for $Q = 1/3$.

| | 1584.8 Da | | | 1456.7 Da | | | 1168.6 Da | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | TRUE | Est. | SE | TRUE | Est. | SE | TRUE | Est. | SE |
| $H_1$ | – | 140584.1 | 2479.8 | – | 143020.7 | 2301.9 | – | 558465.4 | 5582.5 |
| $H_2$ | – | 138828.9 | 2469.3 | – | 138115.3 | 2276.2 | – | 566407.7 | 5585.9 |
| $H_3$ | – | 124616.7 | 2421.3 | – | 130093.7 | 2285.3 | – | 467854.7 | 5424.2 |
| $H_4$ | – | 148885.3 | 2534.9 | – | 147210.0 | 2297.8 | – | 511043.2 | 5353.1 |
| $H_5$ | – | 111185.8 | 2407.8 | – | 111198.5 | 2257.9 | – | 335006.8 | 5052.5 |
| $H_6$ | – | 148003.1 | 2520.5 | – | 147296.5 | 2294.8 | – | 441897.0 | 5249.5 |
| $Q$ | 0.3333 | 0.3107 | 0.0552 | 0.3333 | 0.3072 | 0.0114 | 0.3333 | 0.4342 | 0.0270 |
| $M_1$ | 1584.76 | 1584.7560 | 0.0008 | 1456.66 | 1456.6681 | 0.0006 | 1168.61 | 1168.6211 | 0.0005 |
| $M_2$ | 1588.77 | 1588.7634 | 0.0025 | 1460.67 | 1460.6774 | 0.0016 | 1172.62 | 1172.6281 | 0.0004 |
| $\lambda\tau$ | – | 8.2058 | 0.1647 | – | 7.0719 | 1.1761 | – | 5.8806 | 0.5285 |
| $\kappa$ | – | 0.7076 | 0.0070 | – | 0.8575 | 0.0093 | – | 0.8024 | 0.0104 |
| $\sigma_s$ | – | 0.0882 | 0.0014 | – | 0.0829 | 0.0007 | – | 0.0714 | 0.0009 |
| $S$ | 1.0015 | 1.0033 | 0.0007 | 1.0015 | 1.0040 | 0.0007 | 1.0015 | 1.0040 | 0.0008 |
| $\sigma$ | – | 26.8757 | 4.3311 | – | 54.5165 | 1.5328 | – | 85.8688 | 0.3973 |
| $a$ | – | 3.9700 | 0.2311 | – | 3.1869 | 0.0340 | – | 3.6037 | 0.0669 |
| $b$ | – | 0.6286 | 0.0676 | – | 0.8811 | 0.0217 | – | 1.0674 | 0.0287 |
| $c$ | – | 5.0525 | 0.00002 | – | 5.0631 | 0.000001 | – | 5.6036 | 0.000002 |
| $R_2$ | 0.8703 | 0.8606 | 0.0208 | 0.7933 | 0.7883 | 0.0113 | 0.6645 | 0.7231 | 0.0030 |
| $R_3$ | 0.4223 | 0.4077 | 0.0198 | 0.3567 | 0.3294 | 0.0096 | 0.2454 | 0.2663 | 0.0174 |
| $R_4$ | 0.1478 | 0.1356 | 0.0351 | 0.1166 | 0.0921 | 0.0097 | 0.0653 | 0.0502 | 0.0111 |
| $R_5$ | 0.0413 | 0.0464 | 0.0115 | 0.0306 | 0.0304 | 0.0140 | 0.0139 | 0.0151 | 0.0204 |
| $R_6$ | 0.0097 | 0.0169 | 0.0396 | 0.0068 | 0.0070 | 0.0067 | 0.0025 | 0.0100 | 0.0049 |

**Table 10.3:** Results of the analysis of the data for $Q = 3/1$.

| | 1584.8 Da | | | 1456.7 Da | | | 1168.6 Da | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | TRUE | Est. | SE | TRUE | Est. | SE | TRUE | Est. | SE |
| $H_1$ | – | 48776.2 | 1342.9 | – | 50752.0 | 1348.0 | – | 209822.6 | 2384.9 |
| $H_2$ | – | 48405.6 | 1339.4 | – | 50943.4 | 1350.2 | – | 173986.7 | 2150.3 |
| $H_3$ | – | 44943.0 | 1283.8 | – | 46536.1 | 1293.1 | – | 182871.1 | 2208.1 |
| $H_4$ | – | 56546.1 | 1458.9 | – | 58554.0 | 1456.1 | – | 112792.2 | 1809.0 |
| $H_5$ | – | 58853.4 | 1483.0 | – | 58430.2 | 1446.3 | – | 132629.1 | 1902.5 |
| $H_6$ | – | 52252.3 | 1373.9 | – | 54103.6 | 1382.6 | – | 125139.8 | 1858.4 |
| $Q$ | *2.4* | 2.4800 | 0.0555 | *2.4* | 2.4502 | 0.0526 | *2.4* | 2.1564 | 0.0215 |
| $M_1$ | 1584.76 | 1584.7479 | 0.0017 | 1456.66 | 1456.6577 | 0.0016 | 1168.61 | 1168.6190 | 0.0007 |
| $M_2$ | 1588.77 | 1588.7549 | 0.0009 | 1460.67 | 1460.6651 | 0.0010 | 1172.62 | 1172.6262 | 0.0004 |
| $\lambda\tau$ | – | 8.4015 | 0.4462 | – | 9.1833 | 0.5065 | – | 20.0000 | 0.0000 |
| $\kappa$ | – | 0.6567 | 0.0070 | – | 0.7370 | 0.0091 | – | 0.7912 | 0.0064 |
| $\sigma_s$ | – | 0.0869 | 0.0008 | – | 0.0867 | 0.0008 | – | 0.0829 | 0.0005 |
| $S$ | 1.0015 | 1.0029 | 0.0008 | 1.0015 | 1.0029 | 0.0008 | 1.0015 | 1.0027 | 0.0004 |
| $\sigma$ | – | 23.6448 | 0.5910 | – | 33.1379 | 0.6935 | – | 46.1972 | 0.6986 |
| $a$ | – | 4.1196 | 0.0184 | – | 3.7354 | 0.0161 | – | 3.6805 | 0.0162 |
| $b$ | – | 0.5546 | 0.0114 | – | 0.5990 | 0.0135 | – | 1.3530 | 0.0386 |
| $c$ | – | 5.0278 | 0.00002 | – | 5.0401 | 0.000001 | – | 5.5129 | 0.0000005 |
| $R_2$ | 0.8703 | 0.8673 | 0.0148 | 0.7933 | 0.7863 | 0.0132 | 0.6645 | 0.7272 | 0.0065 |
| $R_3$ | 0.4223 | 0.4138 | 0.0120 | 0.3567 | 0.3384 | 0.0108 | 0.2454 | 0.2862 | 0.0053 |
| $R_4$ | 0.1478 | 0.1374 | 0.0100 | 0.1166 | 0.1022 | 0.0090 | 0.0653 | 0.0878 | 0.0042 |
| $R_5$ | 0.0413 | 0.0367 | 0.0047 | 0.0306 | 0.0273 | 0.0020 | 0.0139 | 0.0193 | 0.0009 |
| $R_6$ | 0.0097 | 0.0111 | 0.0019 | 0.0068 | 0.0095 | 0.0022 | 0.0025 | 0.0075 | 0.0014 |

## Graphical representation of the simulation study:



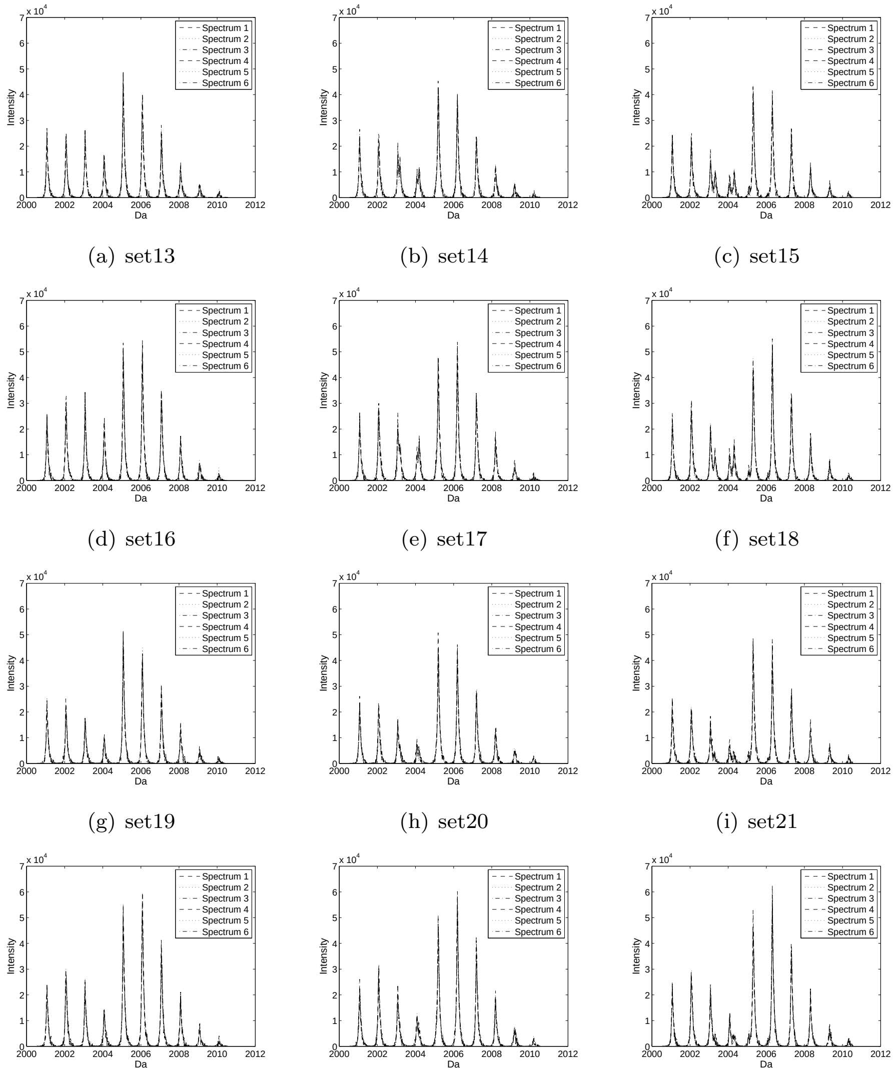
(a) set1 (b) set2 (c) set3

(d) set4 (e) set5 (f) set6

(g) set7 (h) set8 (i) set9

(j) set10 (k) set11 (l) set12

**Figure 10.9:** Graphical representation for the simulation settings of the shape model (set1∼12).



(a) set13

(b) set14

(c) set15

(d) set16

(e) set17

(f) set18

(g) set19

(h) set20

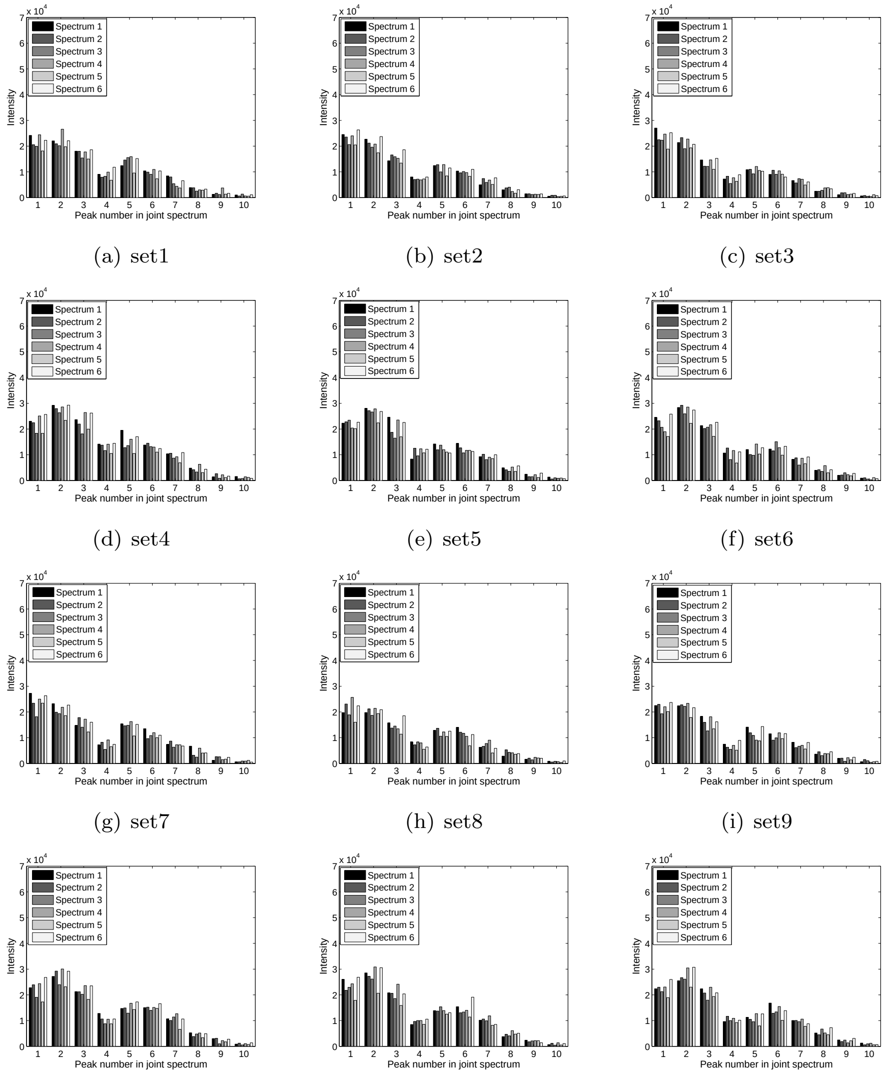
(i) set21

(j) set22

(k) set23

(l) set24

**Figure 10.10:** Graphical representation for the simulation settings of the shape model (set13$\sim$24).



(a) set1

(b) set2

(c) set3

(d) set4

(e) set5

(f) set6

(g) set7

(h) set8

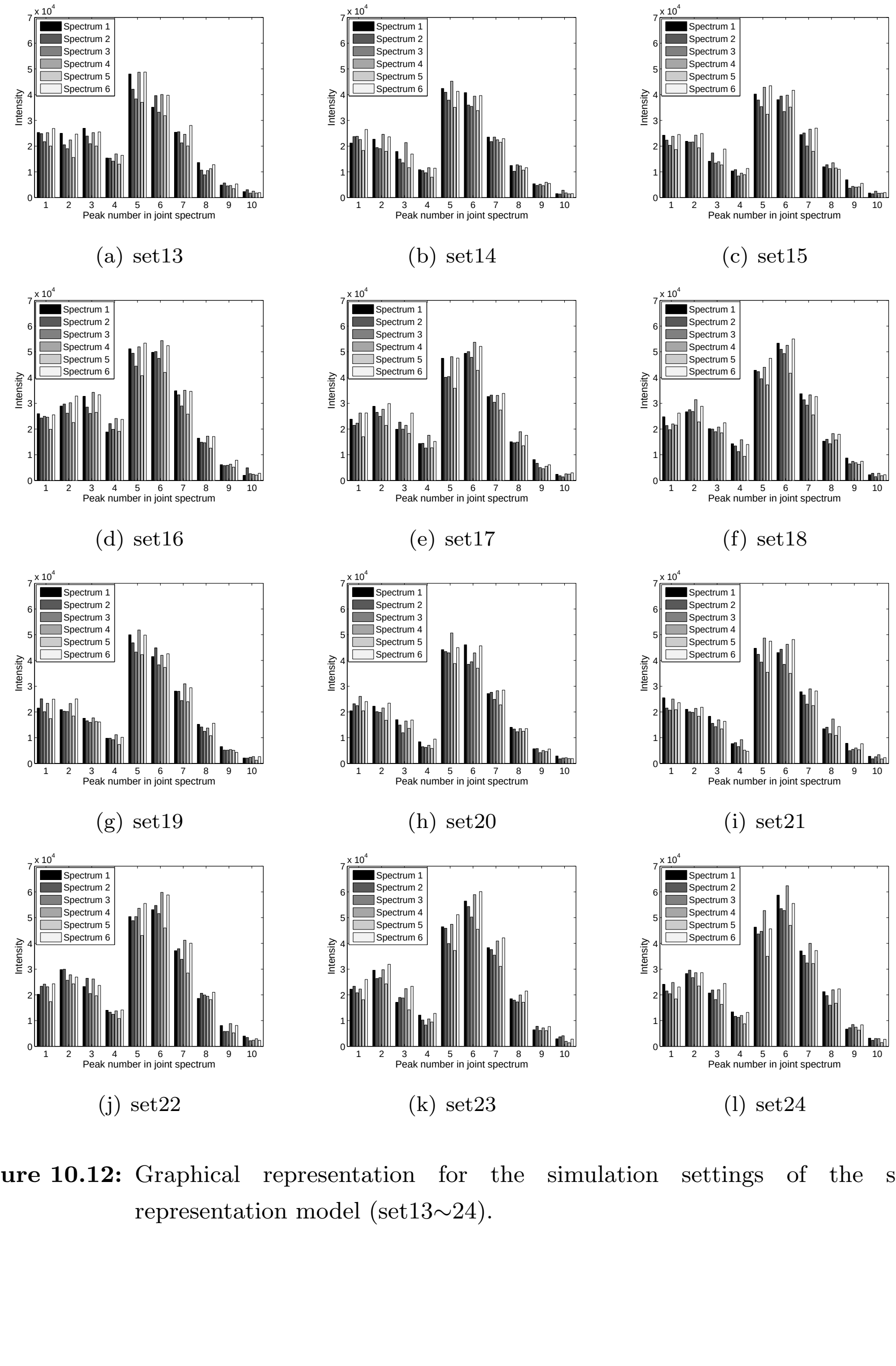
(i) set9

(j) set10

(k) set11

(l) set12

**Figure 10.11:** Graphical representation for the simulation settings of the stick-representation model (set1∼12).



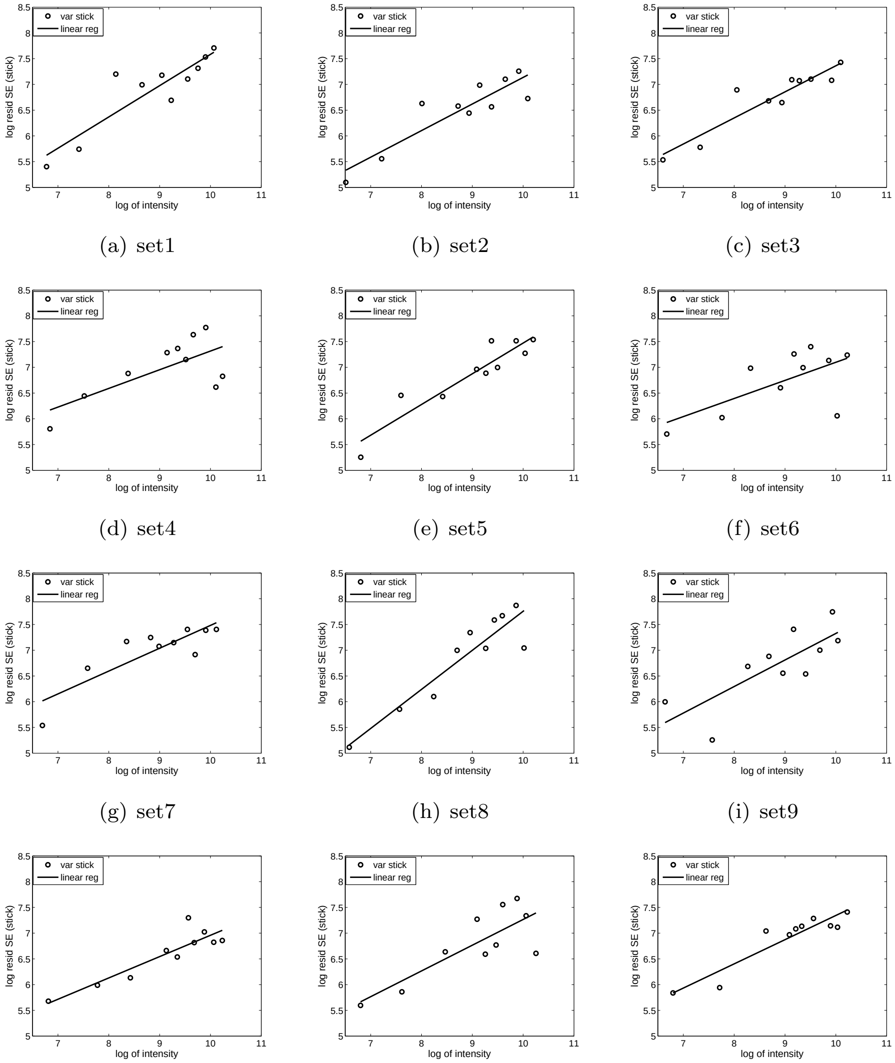
(a) set13 (b) set14 (c) set15

(d) set16 (e) set17 (f) set18

(g) set19 (h) set20 (i) set21

(j) set22 (k) set23 (l) set24

**Figure 10.12:** Graphical representation for the simulation settings of the stick-representation model (set13∼24).



(a) set1

(b) set2

(c) set3

(d) set4

(e) set5

(f) set6

(g) set7

(h) set8

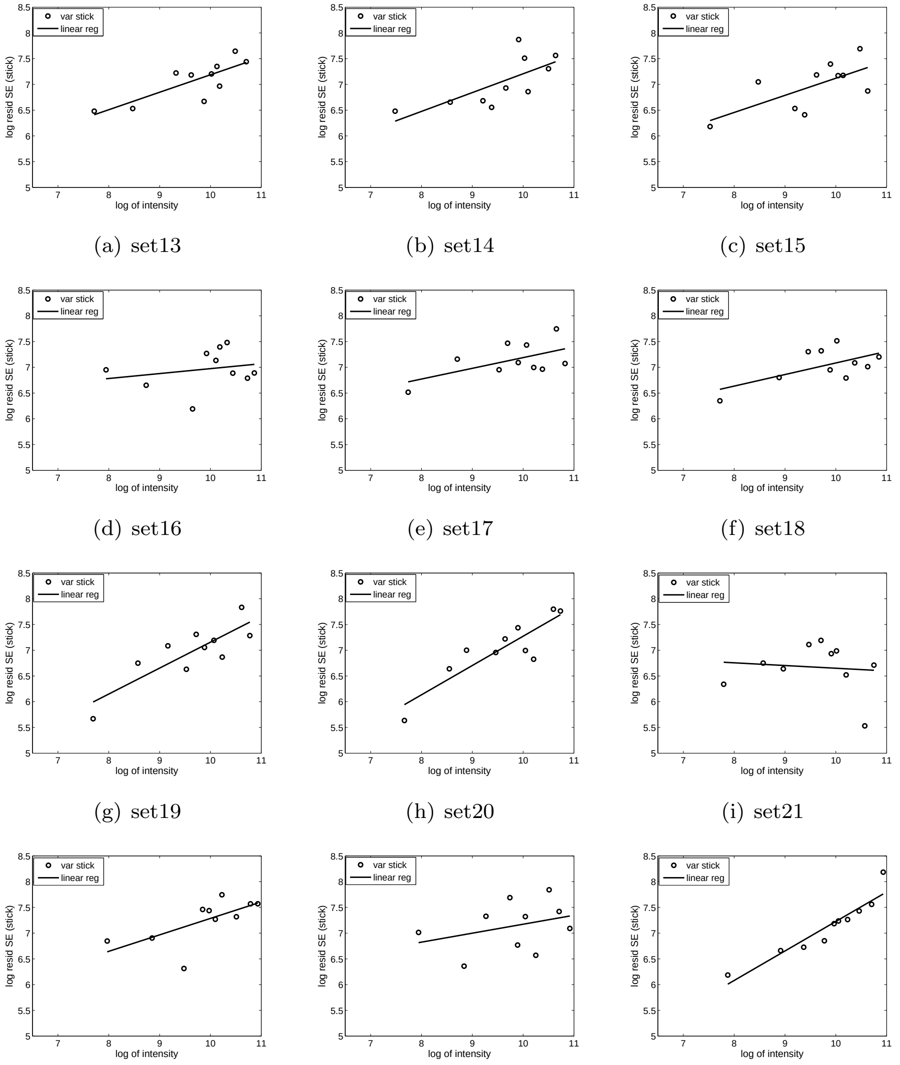
(i) set9

(j) set10

(k) set11

(l) set12

**Figure 10.13:** EDA for the power-of-the-mean residual variance function (set1∼12).



(a) set13 (b) set14 (c) set15

(d) set16 (e) set17 (f) set18

(g) set19 (h) set20 (i) set21

(j) set22 (k) set23 (l) set24

**Figure 10.14:** EDA for the power-of-the-mean residual variance function (set13∼24).

# Part III

# Analysis of mass-spectrometry data with overlapping peptides

# Chapter 11

# Introduction to the problem of overlapping peptides

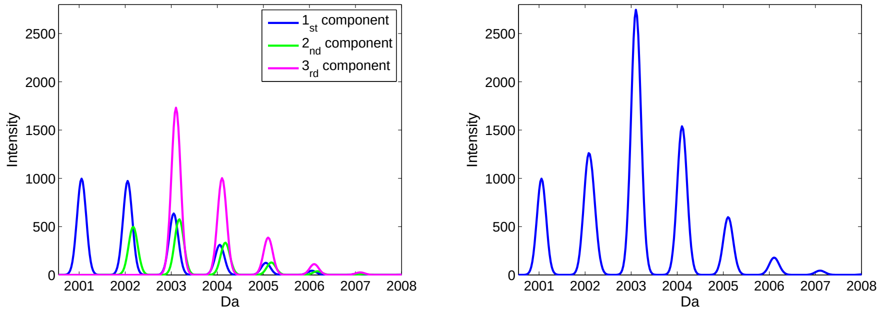
In this chapter, we introduce the problem of the quantification of overlapping peptides in a high-resolution MALDI-TOF-MS. We also briefly review the existing methods aimed at solving the problem. Some basic notations for our modeling approach are also defined.

As stated in Section 2.3, a peptide produces a series of peaks, called isotopic peaks, seperated by roughly multiples of one Da. Their relative heights are related to the probabilities of the isotopic distribution of the peptide.

A 'cluster' of peaks observed in a mass spectrum can be produced by more than one peptide. This happens if two peptides differ in mass by only a few units. Such peptides are called overlapping peptides. Clearly, the identification of the relative abundances, as well as of the exact masses of the overlapping peptides, is of interest.

## 11.1 Problem description

Figure 11.1 illustrates a possible scenario for the case of no measurement noise. It shows, in panel (a), isotopic peaks for three overlapping peptides. The resulting observed *joint spectrum* is presented in panel (b), with a 'cluster' of superimposed peptide peaks. Our key interest is to quantify the true underlying peptides, as displayed in Figure 11.1a. The quantification means a proper assessment of: 1) the number



(a) Component spectrum

(b) Observed spectrum

**Figure 11.1:** The observed spectrum and its corresponding true underlying peptide components.

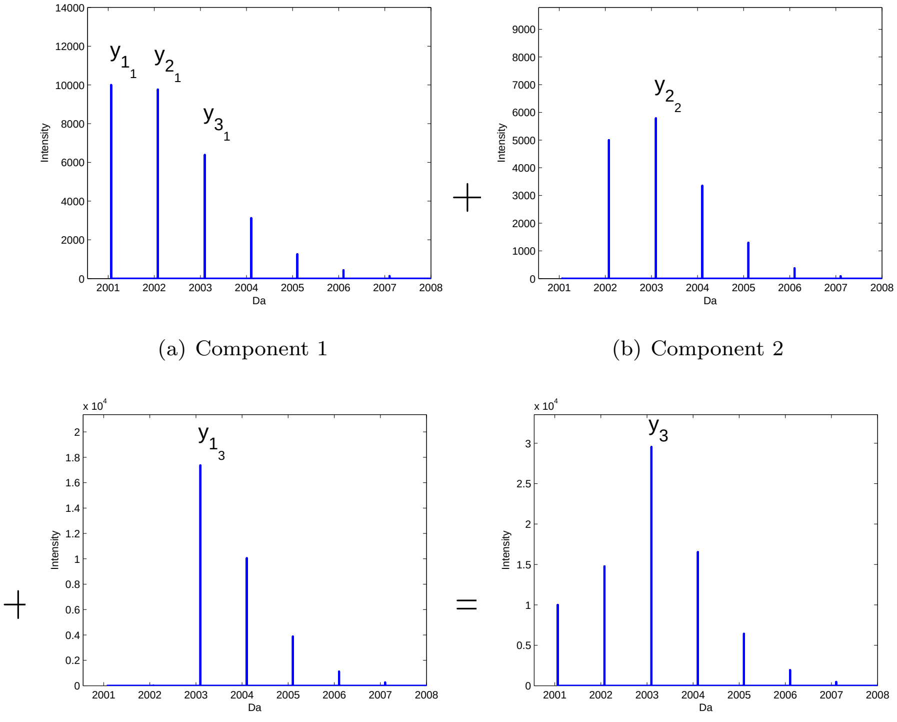
of overlapping peptides (components), 2) the *monoisotopic masses* of the peptides, i.e., the masses of the isotopic variants that contain the most abundant isotopes of chemical elements constructing the peptides, and 3) the corresponding abundances of the peptides.

To this aim, we consider both stick and shape representations of the data.

Modeling the shape representation allows one to consider all the measurements in a mass spectrum, but requires the use of an appropriate function describing the peak-shape (see Chapter 10). As an alternative, the stick representation (definition in Section 2.5) uses only the summary statistic of the measurements and thus yields a simpler model.

### 11.1.1 Stick representation of a mass spectrum with overlapping peptides

One way to obtain the data of stick representation is to sum all the intensities belonging to one observed peak, and to use this summed intensity as a representative measure of that peak. A graphical example of an observed spectrum in the stick representation is shown in Figure 11.2. Intrinsically, the observed peak intensity is a sum of intensities of isotopic peaks from different peptides that contribute to this certain peak. Let $y_{l_d}$ denote the abundance of the $l$th isotopic peak of the $d$th peptide. For the example shown in Figure 11.2, the intensity of the third stick (bin) $y_3$ is a sum of intensities of the third isotopic peak of the first peptide, $y_{3_1}$, of the second isotopic peak of the second peptide, $y_{2_2}$, and of the monoisotopic peak of the third peptide, $y_{1_3}$, i.e., $y_3 = y_{3_1} + y_{2_2} + y_{1_3}$.



(a) Component 1

(b) Component 2

(c) Component 3

(d) Joint spectrum

**Figure 11.2:** Graphical explanation of model definition for the stick representation.

To work with the stick representation, several assumptions have to be made:

1. There are approximately the same number of data points within each bin (corresponding to each observed peak). This assumption holds as the number of data points per peak is locally constant (within around 20 Da range);

2. There is an exact overlap with respect to the modes of the peaks from different peptides. This is because for the stick representation, only one data point would represent each peak within a bin. Thus, a slight shift of the two peptide peaks within the same bin would not be detectable;

3. The difference in the monoisotopic masses of overlapping peptides is integers of 1 Da. This is because the peak numbers, rather than the true monoisotopic mass coordinates, will be used for the modeling of the stick representation;

4. The monoisotopic masses of two overlapping peptides do not appear at the same observed peak. In other words, the difference of monoisotopic masses of two overlapping peptides is at least 1 Da, because an exact overlap of two peptides are not distinguishable.

### 11.1.2 Shape representation of a mass spectrum with overlapping peptides

The stick representation may not work equivalently well as the shape representation, due to information reduction. Alternatively, one may want to consider the use of an appropriate shape function for the shape representation.

For this purpose, a suitable function depicting the shape of the peak envelope is needed.

By referring to Figure 11.1, let us assume that we have a measurement $y_i^*$ at mass coordinate $x_i$ and that this measurement belongs to the third observed peak. The intensity at the mass coordinate $x_i$ is a sum of intensity measurements of all the isotopic peaks of the three peptides that contribute to the mass coordinate, i.e., $y_i^* = y_{3_1}^* + y_{2_2}^* + y_{1_3}^*$. $y_{l_d}^*$ is the peak intensity of $l$th isotopic peak of the $d$th peptide at mass location $x_i$ in the shape representation, calculated from: $y_{l_d}^* = y_{l_d}\psi(x_i|\mu_{l_d}, \sigma_s)$. $\psi(x_i|\mu_{l_d}, \sigma_s)$ is a shape function, approximating the shape of the peak envelopes, with $\sum_i \psi(x_i|\mu_{l_d}, \sigma_s) = 1$; $\mu_{l_d}$ and $\sigma_s$ are the mean and spread for the shape function.

## 11.2 Existing methods

Several existing methods to tackle the problem of overlapping peptides in the MALDI-TOF MS data have been proposed. In this section, we discuss these methods and mention the limitations while using the methods.

Breen *et al.* (2000) suggested to model the isotopic distribution by a Poisson approximation (details explained in Section 2.3), which can also be used to identify overlapping peptides. The method is based on the stick representation. This method often fails due to the lack of information about the mass location of these peptides and due to the discrepancy between the true isotopic distribution and the Poisson-approximated one. Such discrepancy will not only result in the bias of the isotopic distribution itself, but also of the quantification of the overlapping peptides which is of particular interest, e.g., the relative abundance(s) of these peptides.

Schulz-Trieglaff *et al.* (2007) and Lange *et al.* (2006) developed a peak-picking algorithm by means of a wavelet function, combined with a greedy search to identify the overlapping peptides. This method has three limitations: 1) the multi-stage analysis involved for the greedy search, would pose a difficulty in, e.g., estimating precision of the estimates, obtained at different stages, 2) often no unique solution can be found for the wavelet functions to fit to the peptide profiles, and 3) greedy search is often problematic in that it can either include noise peaks as peptide peaks or discard peptide peaks, depending on the fit to the wavelet functions. These limitations acting together can lead to non-identification or mis-identification of the overlapping peptides.

As an alternative, we propose a Bayesian modeling approach to address the problem of overlapping peptides in the MS data. Chapter 12 and Chapter 13 describe, respectively, models based on the stick and shape representations. In Chapter 14, we present an improved method, based on the shape representation, using a Bayesian model averaging approach.

# Chapter 12

# A Bayesian model for the stick representation of a mass spectrum with overlapping peptides

## 12.1 Introduction

Due to the limitations of the existing methods for the quantification of overlapping peptides, mentioned in Section 11.2, as an alternative, we propose a Bayesian modeling approach to address the problem. The advantage of the Bayesian methodology over frequentist approaches is that prior information can be incorporated in the model. Such information can be obtained from MS data available in public databases like, e.g., the NCBI data (refer to Section 3.1).

In this chapter, we formulate the Bayesian model for the stick representation of a mass spectrum with overlapping peptides. We first explain the priors used for the parameters and the resulting posterior distributions. Then, an application to the bovine cytochrome C data set is presented, followed by a simulation study, to assess the statistical performances of the proposed the model.

## 12.2 Model formulation for the stick representation

We assume that the number of the overlapping peptides, $D$ is known. In the stick representation, the height of $i$th peak, $y_i$, is assumed to be normally distributed with mean $E(y_i)$ and a constant variance $\sigma^2$. Thus,

$$y_i \sim N(E(y_i), \sigma^2),\ i = 1, \ldots, N, \tag{12.1}$$

with

$$\begin{aligned} E(y_i) &= f(\boldsymbol{H}, \boldsymbol{R}, \boldsymbol{M}) \\ &\equiv \sum_{d=1}^{D} \sum_{l=1}^{L} E(y_{l_d}) \\ &= \sum_{d=1}^{D} \sum_{l=1}^{L} H_d R_{l_d} I(M_d + l - 1 - i = 0) \end{aligned} \tag{12.2}$$

where $N$ is the number of sticks (peaks) in a mass-spectrum that are modeled and $L$ is the number of isotopic variants of the peptides. The mean structure shown in equation (12.2) is based on the formulation for the stick representation explained in Section 11.1. In equation (12.2), $H_d$ is the abundance of the $d$th overlapping peptide ($d = 1, 2, \ldots, D$) and $\boldsymbol{H} = (H_1, \ldots, H_D)$. The monoisotopic-mass-index of the $d$th peptide in the stick representation is denoted by $M_d$ with $M_1 = 1$ and $\boldsymbol{M} = (M_1, M_2, \ldots, M_D)$. Note that we order the peptides according to their increasing monoisotopic masses, i.e., $M_d < M_{d+1}$. Parameter $R_{l_d}$ is the $l$th common reference isotopic ratio for the $d$th peptide and $\boldsymbol{R} = (R_{1_1}, R_{2_1}, \ldots, R_{L_1}, R_{1_2}, R_{2_2}, \ldots, R_{L_2}, \ldots, R_{1_D}, R_{2_D}, ..., R_{L_D})$ is a vector containing the isotopic ratios for all peptides. The indicator function $I(A)$ is equal to one if expression $A$ is true and zero otherwise.

## 12.3 Prior distributions

We formulate a Bayesian model for the observed MS data by assuming prior distributions for parameters of the model, defined in (12.1)–(12.2). As there is usually no prior information available for $H_d$ and $\sigma^2$, we specify non-informative priors for these parameters. Note that, in the stick representation, we do not distinguish between monoisotopic masses that fall in the same bin. Hence, for the mass-indices $M_d$, we also specify a non-informative prior. This allows to fully estimate these parameters from the data. The prior distributions are specified as follows:

$$\begin{aligned} H_d &\sim N\left(0, \frac{1}{\tau}\right) \text{ with } \tau \sim \Gamma(\alpha^*, \beta^*), && (12.3) \\ \sigma^{-2} &\sim \Gamma(\alpha, \beta), && (12.4) \\ (M_2, \ldots, M_D) &\sim \text{Multinomial}\{1, (\pi_2, \ldots, \pi_{N-1})\}, && (12.5) \end{aligned}$$

where $\pi_2 = \ldots = \pi_{N-1} = \frac{1}{N-1}$ and $\alpha$, $\beta$, $\alpha^*$, and $\beta^*$ are positive constants close to zero. To avoid redundancy, $\boldsymbol{M}$ is ordered such that $M_d < M_{d+1}, d = 1, ..., D-1$.

For the isotopic ratios $\boldsymbol{R}$, we consider informative priors, as explained in the next section.

### 12.3.1 Informative prior for the isotopic ratios $\boldsymbol{R}$

#### Priors obtained from the polynomial models

The NCBI data can be used to extract information about possible forms of the isotopic distribution of peptides. We initially fitted a multivariate polynomial model to the ratios simultaneously, to obtain the model-based mean and covariance structure. This, however, turned out to be practically infeasible due to the heterogeneity of covariance structure for different isotopic ratios across different mass ranges. Alternatively, a semi-model-based and semi-empirical approach was performed. This was done by fitting a univariate polynomial model (3.1) to each of the isotopic ratios. In this way, the model-based mean structure could be obtained. The variances were calculated empirically from the residuals of these polynomial models.

The resulting model coefficient estimates of the polynomial models for isotopic ratios $l = 2$ to 8 are shown in Table 12.1. They allow to infer the form of the isotopic distribution of a peptide with monoisotopic mass $m$. The variances of the model residuals are shown in Table 12.2.

The prior for each of the isotopic ratios can then be defined as a normal prior with mean and variance shown in Tables 12.1 and 12.2. More explicitly, the prior distribution is on the log-scale of $C_l$: $N(\mu_{C_l^*}, \sigma^2_{C_l^*})$.

As the priors for the logarithm of different consecutive ratios are assumed to be independent, the prior for the logarithms of $R_l$ ($l = 2, \ldots, 8$) is $N(\sum_{i=2}^{l} \mu_{C_i^*}, \sum_{i=2}^{l} \sigma^2_{C_i^*})$. As a result, the prior for $R_l$ is log-normal with the mean $\sum_{i=2}^{l} \mu_{C_i^*}$ and variance $\sum_{i=2}^{l} \sigma^2_{C_i^*}$. Figure 12.1 shows an overlay of the observed isotopic ratios $R_l$ versus their transformed

**Table 12.1:** The polynomial model coefficient estimates.

| | $C_2^*$ | $C_3^*$ | $C_4^*$ | $C_5^*$ | $C_6^*$ | $C_7^*$ | $C_8^*$ |
|---|---|---|---|---|---|---|---|
| $\beta_0$ | -2.5835 | -2.6283 | -2.9429 | -3.1161 | -3.2939 | -3.4508 | -3.6021 |
| $\beta_1$ | 3.2954 | 2.3416 | 2.4265 | 2.3733 | 2.4299 | 2.4994 | 2.5967 |
| $\beta_2$ | -1.7098 | -1.0856 | -1.2003 | -1.1854 | -1.2464 | -1.3110 | -1.3932 |
| $\beta_3$ | 0.4594 | 0.2772 | 0.3197 | 0.3176 | 0.3386 | 0.3600 | 0.3865 |
| $\beta_4$ | -0.0466 | -0.0274 | -0.0324 | -0.0323 | -0.0347 | -0.0372 | -0.0401 |
| $\sigma^2$ | 0.0035 | 0.0008 | 0.0006 | 0.0010 | 0.0012 | 0.0016 | 0.0019 |

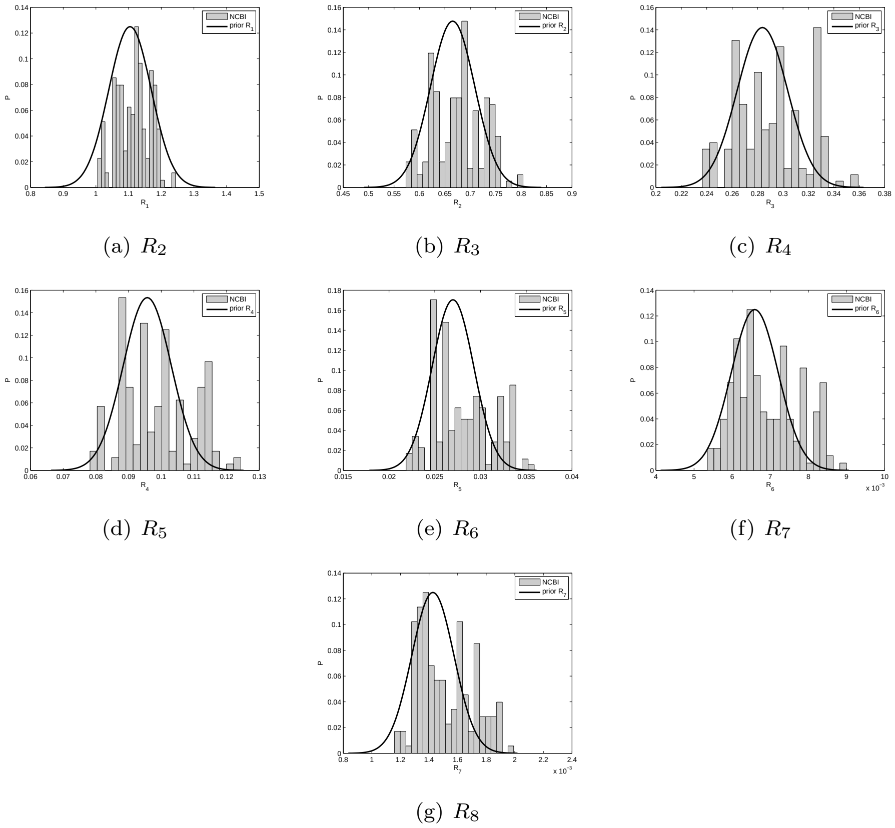
**Table 12.2:** The variances of the model residuals $\sigma^2_{C_l^*}$.

| $\sigma^2_{C_2^*}$ | $\sigma^2_{C_3^*}$ | $\sigma^2_{C_4^*}$ | $\sigma^2_{C_5^*}$ | $\sigma^2_{C_6^*}$ | $\sigma^2_{C_7^*}$ | $\sigma^2_{C_8^*}$ |
|---|---|---|---|---|---|---|
| 0.003478 | 0.000817 | 0.000609 | 0.000998 | 0.001233 | 0.001598 | 0.001956 |

prior distributions at around 2001 Da. This figure indicates the priors correctly capture the possible ranges for the values of the ratios.

### Reparameterization as a virtual constraint of ratio estimates

Using the priors for the logarithm of consecutive ratios $C_l^*$, described in Section 12.3.1, the priors of the common-reference ratios are linked by the corresponding consecutive ratios. More specifically, as $R_j = C_1C_2 \dots C_j$, the increase of $C_l$ will result in an increase in the common-reference ratios $R_l$–$R_L$. For instance, consider a situation when the second peptide is much more abundant than the first and it starts to overlap from the third observed peak (as presented in Figure 12.2a). The model could either capture the true scenario (represented by Figure 12.2b) by recognizing the third (much more abundant) peak as an advent of (the monoisotopic peak of) an overlapping peptide; or it could treat it as the second isotopic peak, which is much more abundant than the monoisotopic one (depicted in Figure 12.2c). The latter case could happen when $C_{2_2}$ is estimated much larger and far beyond the reasonable range of the ratio estimates, resulting in the over-estimation of common-reference ratios $R_{2_2}$ to $R_{L_2}$. In this way, it still yields equivalently good fit to the true scenario, leading to non-identifiability issue. Alternatively, to circumvent the problem, the ratios can be reparameterized. Let $h_l$ denote the abundance of the $l$th isotopic peak, such that



(a) $R_2$

(b) $R_3$

(c) $R_4$

(d) $R_5$

(e) $R_6$

(f) $R_7$

(g) $R_8$

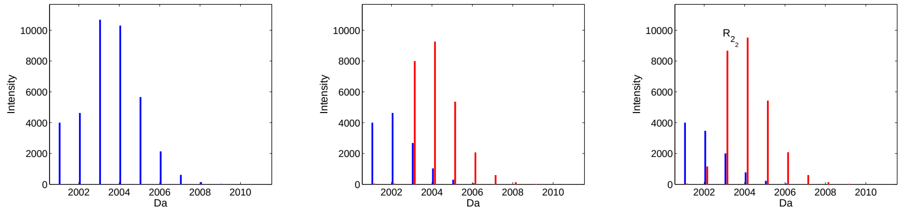
**Figure 12.1:** Prior density plots for the reference ratios $R_l$ versus the NCBI data.

$\sum\limits_{i=1}^{L} h_l = 1$. We then have $R_l = h_l/h_1$, and:

$$R_2 = \sum_{l=2}^{L} R_l - \sum_{l=3}^{L} R_l = \frac{1-h_1}{h_1} - \sum_{l=3}^{L} R_l. \qquad (12.6)$$

For isotopic ratios $R_l$ ($l = 3, \ldots, L$), we still use the priors defined in Section 12.3.1. For $R_2$, instead of putting a prior on it, we use the equality relationship in (12.6) and define a prior for $(1-h_1)/h_1$. The reparameterization becomes a natural constraint for the common-reference ratios. This is because the increase of the other ratios, as shown in equation (12.6), would result in the shrinkage of $R_2$ (given $h_1$).

The prior for $(1-h_1)/h_1$ was obtained by fitting a model with monoisotopic



(a) Observed spectrum (b) True scenario (c) False scenario

**Figure 12.2:** Graphical representation of non-identifiability with the use of prior $C_l^*$: $N(\mu_{C_l^*}, \sigma^2_{C_l^*})$. Panel (a): the observed spectrum as a mixture of two peptides; panel (b): the underlying true scenario; panel (c): the wrong scenario due to the over-estimation of $R_{2_2}$–$R_{L_2}$.

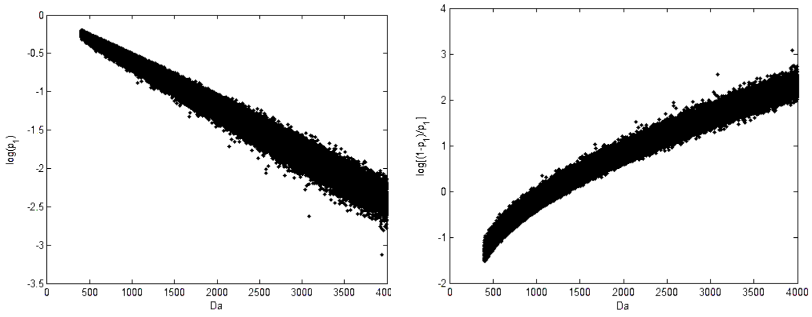
mass $m$ as a covariate to the isotopic distributions of the NCBI data set. Figure 12.3 shows the scatter plot of the log and log odds scales of $h_1$ (from the NCBI data set) versus the monoitotopic mass $m$. It is clear that a linear regression on the log scale of $h_1$, i.e., $\log(h_1) = \alpha + \beta m + \varepsilon$, is a good choice. However, the increase of the 'bandwidth' of the scatter along $m$ implies that the variance of $\varepsilon$ cannot be a constant. On the other hand, Figure 12.3(b) shows that the variance of the log odds scale of $h_1$ is fairly constant. Thus, an alternative can be to fit a model to the log odds of $h_1$. Transforming the linear regression on the log scale to log odds scale, the model function becomes: $y = \log\left(\frac{1-h_1}{h_1}\right) = \log\left[1 - \exp\left(\alpha + \beta m\right)\right] - (\alpha + \beta m) + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2_{h_1})$.

Figure 12.4 indicates that the predicted mean follows the pattern of the observed log odds of $h_1$. Strictly speaking, normality assumption is not perfectly satisfied as indicated by the QQ-plot shown in Figure 12.5. Given the large number of observations, normality could be hardly perfectly met. We assume roughly, that the residuals follow a normal distribution. Consequently, the prior for the log odds scale is defined as:

$$\log\left(\frac{1-h_1}{h_1}\right) \sim N(\mu_{h_1}, \sigma^2_{h_1}), \tag{12.7}$$

where $\mu_{h_1} = \log\left[1 - \exp\left(\alpha + \beta m\right)\right] - (\alpha + \beta m)$. The prior for the odds of $h_1$ is then log-normal, i.e., $\frac{1-h_1}{h_1} \sim \text{Log-normal}(\mu_{h_1}, \sigma^2_{h_1})$.

The resulting prior for the common-reference (isotopic) ratio $R_{l_d}$ is log-normal,



(a) $\log(h_1)$

(b) $\log\left(\frac{1-h_1}{h_1}\right)$

**Figure 12.3:** Scatter plots of functions of $h_1$ versus monoisotopic mass.

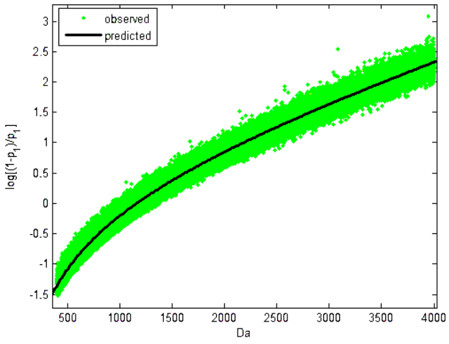


**Figure 12.4:** Fitted mean versus observed log odds of $h_1$.

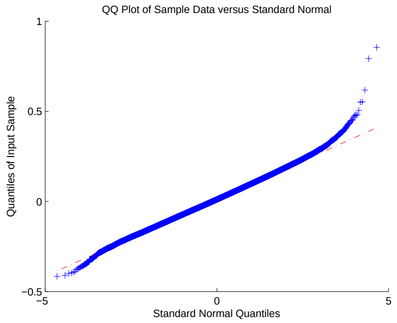


**Figure 12.5:** QQ plot for the model residuals on the log odds of $h_1$.

i.e.,

$$R_{l_d} \sim \text{Log-normal}\left(\sum_{i=1}^{l}\mu_i, \sum_{i=1}^{l}\sigma_i^2\right), \text{ where } l = 3, \ldots, L, \tag{12.8}$$

and $R_{2_d} = \frac{1-h_1}{h_1} - \sum_{l=3}^{L} R_{l_d}$ with the prior:

$$\frac{1-h_1}{h_1} \sim \text{Log-normal}(\mu_{h_1}, \sigma_{h_1}^2). \tag{12.9}$$

## 12.4 Conditional posterior distributions

The normal prior distribution for $H_d$, defined in (12.3), is a conjugate prior given the normal likelihood, which results from (12.1)–(12.2). Hence, the conditional posterior distribution of $H_d$ is also a normal distribution, which can be computed analytically. More specifically:

$$p(H_d|\boldsymbol{y}, \boldsymbol{M}, \sigma^2) \propto N(\mu_{H_d}, \sigma_{H_d}^2), \tag{12.10}$$

with $\sigma_{H_d}^2 = \left[\tau + \sigma^{-2}\left(\sum_{l=1}^{L} R_l^2\right)\right]^{-1}$ and

$$\begin{aligned}\mu_{H_d} &= \sigma_{H_d}^2 \left\{0 + \sigma^{-2}\left[\sum_{i=1}^{N} y_i \left(\sum_{l=1}^{L} R_l I(M_d + l - 1 - i)\right)\right]\right\} \\ &= \frac{\sum_{i=1}^{N} y_i \left[\sum_{l=1}^{m} R_l I(M_d + l - 1 - i = 0)\right]}{\tau\sigma^2 + \sum_{l=1}^{m} R_l^2}\end{aligned}$$

The prior distribution for $\sigma^{-2}$, defined in (12.4), given the normal likelihood, is also a conjugate prior, and leads to a Gamma conditional posterior distribution:

$$p(\sigma^{-2}|\boldsymbol{y}, \boldsymbol{H}, \boldsymbol{M}) \propto \Gamma\left(\alpha + \frac{N}{2}, \beta + \frac{\sum_{i=1}^{N}(y_i - E(y_i))^2}{2}\right) \tag{12.11}$$

The conditional posterior distribution for $M_d$ is multinomial with the probabilities of belonging to each category updated based on the prior, i.e.,

$$p(M_d = j|\boldsymbol{y}, \boldsymbol{H}, \sigma^2) \propto \text{Multinomial}(1, \boldsymbol{p}) \tag{12.12}$$

$$
\begin{aligned}
\text{with } p_j &= \frac{\pi_j L(M_d = j|\boldsymbol{y}, \boldsymbol{H}, \sigma^2)}{\sum\limits_{s=2}^{N} \pi_s L(M_d = s|\boldsymbol{y}, \boldsymbol{H}, \sigma^2)}, j = 2, ..., N.\\
&= \frac{\exp\left[-\frac{\sum\limits_{i=1}^{N}(y_i - E(y_i))^2}{2\sigma^2}\right]}{\sum\limits_{s}\exp\left[-\frac{\sum\limits_{i=1}^{N}(y_i - E(y_i))^2}{2\sigma^2}\right]}\\
&= \frac{\exp\left[-\frac{\sum\limits_{i=1}^{N}\left(y_i - \sum\limits_{d=1}^{D}\sum\limits_{l=1}^{L} H_d R_l I(j + l - 1 - i = 0)\right)^2}{2\sigma^2}\right]}{\sum\limits_{s=2}^{N}\exp\left[-\frac{\sum\limits_{i=1}^{N}\left(y_i - \sum\limits_{d=1}^{D}\sum\limits_{l=1}^{L} H_d R_l \cdot I(s + l - 1 - i = 0)\right)^2}{2\sigma^2}\right]}
\end{aligned}
$$

There is no analytical solution for the conditional posterior distributions of $C_{l_d}$ and $R_{l_d}$. These distributions therefore need to be evaluated by numerical (sampling) methods, e.g., a Metropolis-Hasting algorithm with acception-rejection rules.

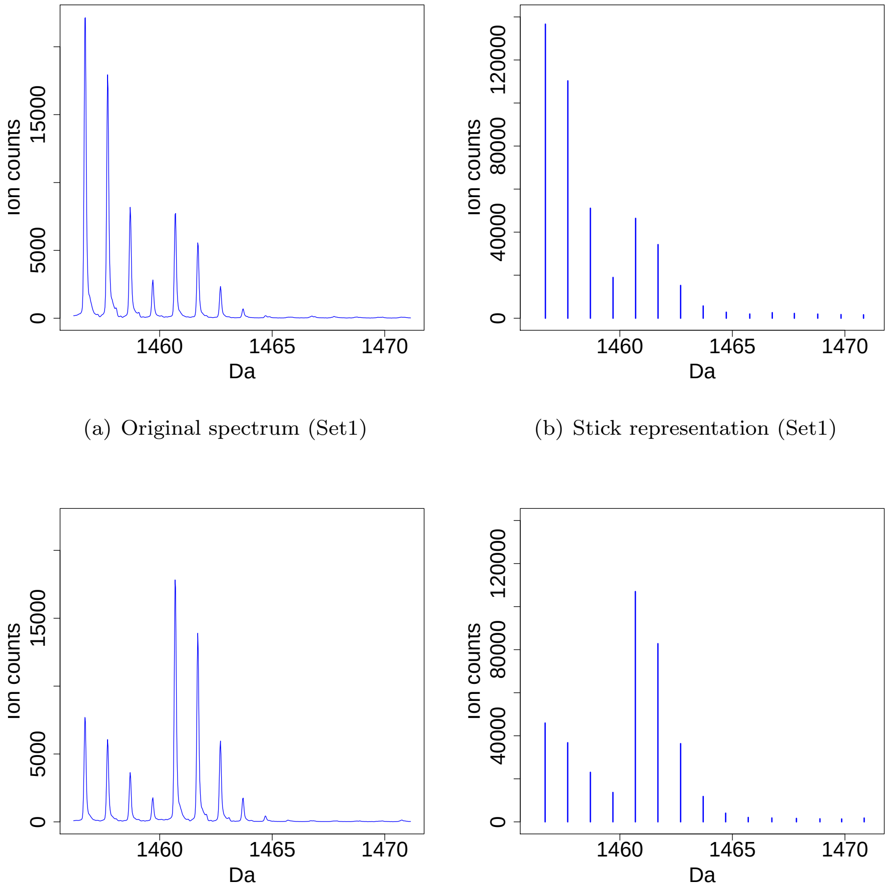
## 12.5 Application to the data

To investigate the performance of the developed model (12.1)–(12.2), we fitted it to real-life data. The model was fitted by using the *R* package *R2WinBUGS*, built in *R* to automatically call the *WinBUGS1.4* software, which allows to fit Bayesian models. To implement the model in *WinBUGS1.4*, the priors, specified in Section 12.3, were considered.

### 12.5.1 Bovine cytochrome C mass spectra

The model was applied to a data set of replicated joint mass spectra obtained for peptides of bovine cytochrome C (for details refer to Section 3.2).

In the $^{18}$O labeling strategy, the labeled peptide ideally receives two $^{18}O$-atoms at its carboxyl terminus, which leads to a four Da mass shift of the corresponding peptide peaks when analyzed by a mass spectrometer (see Section 2.6.1). Thus, each



(a) Original spectrum (Set1)

(b) Stick representation (Set1)

(c) Original spectrum (Set2)

(d) Stick representation (Set2)

**Figure 12.6:** Graphical representation of the first $(H_2/H_1 = 1/3)$ and the second $(H_2/H_1 = 3/1)$ data sets.

spectrum can be treated as containing pairs $(D = 2)$ of overlapping peptides with a four Da difference in the monoisotopic masses ($M_2^* = M_1^* + 4$, with $M_1^*$ and $M_2^*$ denoting the monoisotopic masses of the two peptides).

For the analysis purposes, we chose two peptides with monoisotopic masses of 1456.66 Da and 1584.76 Da. For each peptide, we considered six spectra for each of two different relative abundances (1/3 or 3/1) of the $^{16}$O and $^{18}$O labeled peptides.

This results in the following four settings, each with 6 spectra:

Setting 1: $M_1^* = 1456.66248$ $H_2/H_1 = 1/3$ $M_2 = 5$
Setting 2: $M_1^* = 1456.66248$ $H_2/H_1 = 3/1$ $M_2 = 5$
Setting 3: $M_1^* = 1584.75744$ $H_2/H_1 = 1/3$ $M_2 = 5$
Setting 4: $M_1^* = 1584.75744$ $H_2/H_1 = 3/1$ $M_2 = 5$

An example of the original spectra and of the corresponding stick representations is shown in Figure 12.6.

### 12.5.2 Results of the model fit

The parameters of main interest are:

- the estimated index of the monoisotopic mass location of the second peptide, $M_2$, for model (12.1)–(12.2);
- the relative abundance $H_2/H_1$.

Note that, usually, instead of the relative abundance $H_2/H_1$, individual abundances $H_1$ and $H_2$ of the overlapping peptides would be of interest. However, in the analyzed experiment, only their ratio $H_2/H_1$ was controlled. Thus, it is of interest to verify whether the proposed model estimates correctly the relative abundance.

In this respect, as has been observed in Chapter 5, the experiment for settings 2 and 4 was not well conducted. The achieved value of relative abundance $H_2/H_1$ was about 2.4, not 3. This value was therefore assumed as the true relative abundance.

Since the atomic compositions of all the peptides in the data set are known, the true isotopic distributions can be computed based on the atomic compositions using a Fourier transform, as proposed by Rockwood (1995). The computed isotopic distributions, transformed to isotopic ratios, are then used as the true values for these ratios.

#### Fit for the individual spectrum

Tables 12.4 to 12.15 show the point estimates (medians of the posterior distributions) and the 95% credible intervals for the parameters of model (12.1)–(12.2) for the stick representation, based on 20,000 samples (with 10,000 burn-in samples) from the posterior distribution. The tables contain results for each technical replicate.

Several patterns can be observed from these tables. First, for all of the spectra, the monoisotopic mass index for the second peptide $M_2$ is correctly estimated. Second, the point estimates of the relative abundance $H_2/H_1$ for all the cases are underestimated and the 95% credible intervals are fairly wide. For settings 2 and 4, the 95% credible interval covers the true value, which is in general not the case for the other two settings.

It is also worth noting that the point estimates of the isotopic ratios mostly show an upward bias, especially for the first peptide, i.e., for parameters $R_{l_1}$. However, the 95% credible intervals of these ratio estimates for both peptides cover the true values.

### Simultaneous fit to six technical replications of spectra

For each setting, there are 6 technical replicates of the mass spectra. It is intuitively reasonable to fit a model simultaneously to the six data sets by appropriately incorporating the between-spectra variability. The variability can then be studied. With this respect, the model was modified by allowing all the spectra to share the same parameters except of the reference abundance, the residual standard deviation, and the relative abundance parameters. More specifically, the reference abundance, i.e., the abundance of the first peptide, was defined to be spectrum specific and denoted as $H_{1_j}$ for the $j$th spectrum. Similarly, the residual standard deviation is denoted as $\sigma_j$ for the $j$th spectrum. The model is reparameterized to incorporate a random effect parameter $Q_j$, being the relative abundance between the two peptides for the $j$th spectrum. For this relative abundance effect, in order to estimate the between-spectra variability, $Q_j$ is assumed to be a random variable. As such, the abundance of the second peptide $H_{2_j}$ for the $j$th spectrum is no longer a parameter, but a product of $H_{1_j}$ and $Q_j$, i.e., $H_{2_j} = H_{1_j}Q_j$. More specifically, based on the model defined in (12.1)-(12.2), the intensity of the $i$th peak in the $j$th spectrum is:

$$y_{ij} \sim N(E(y_{ij}), \sigma_j^2), \; i = 1, \ldots, N, \tag{12.13}$$

with

$$\begin{aligned} E(y_{ij}) &= f(\boldsymbol{H}, \boldsymbol{R}, \boldsymbol{M}, \boldsymbol{Q}) \\ &= \sum_{l=1}^{L} H_{1_j} R_{l_1} I(M_1 + l - 1 - i = 0) + \sum_{l=1}^{L} H_{2_j} R_{l_2} I(M_2 + l - 1 - i = 0) \\ &= \sum_{l=1}^{L} H_{1_j} R_{l_1} I(l - i = 0) + \sum_{l=1}^{L} H_{1_j} Q_j R_{l_2} I(M_2 + l - 1 - i = 0), \end{aligned} \tag{12.14}$$

where $Q_j$ is random:

$$Q_j \sim N(\overline{Q}, \sigma_Q^2).$$

Non-informative (conjugate) priors are used for $\overline{Q}$ and $\sigma_Q^2$, i.e.,

$$\begin{aligned} \overline{Q} &\sim N\left(0, \frac{1}{\tau_Q}\right), \\ \sigma_Q^{-2} &\sim \Gamma\left(\alpha_Q, \beta_Q\right), \end{aligned}$$

where $\tau_Q, \alpha_Q$ and $\beta_Q$ are positive constants close to zero. The results of fitting the model (12.13)–(12.14), based on 20,000 samples from the posterior distributions, are presented in Tables 12.16 and 12.17. The monoisotopic mass index for the second peptide $M_2$ is again correctly estimated for all the settings. There is still slight under-estimation for the mean relative abundance parameter $\overline{Q}$, but it is closer to the true value when the between-spectra variability is incorporated. The estimates of between-spectra variability of the relative abundance, captured by $\sigma_Q^2$, are very small. This is reasonable since, for technical replicates, samples for different spectra are the same and thus their relative abundance is expected to be the same as well.

It is also interesting to note that the isotopic ratios for the first peptide, for all of the four settings, are over-estimated. On the other hand, the ratios for the second peptide are estimated very close to the true values, regarding their point estimates. The 95% credible intervals of these ratio estimates for this peptide contain the true values as well. This is perhaps due to the incorporation of the between-spectra variability for the second peptide via the relative abundance parameter, which allows for the correction of the bias for estimating the other parameters that are related to this peptide.

The over-estimation of the isotopic ratios of the first peptide may be caused by the fact that in the experiment, in which $^{18}$O-labeling is used, a part of peptide molecules from a labeled sample (the second peptide) do not get a complete label (see Section 2.6.1). These incompletely labeled molecules additionally overlap with the molecules from the unlabeled sample (the first peptide). Because of this effect, the intensity measurements in the observed joint spectrum may not reflect the true isotopic distributions of the overlapping peptides. Rather, the peaks of the first peptide appear to be more abundant than they actually are and thus causes an upward bias. The under-estimation of the relative abundance may also be due to this effect.

## 12.6 A simulation study

In this section, we present a simulation study that was performed to investigate the performance of the proposed model (12.1)–(12.2) for the stick representation of a mass spectrum with overlapping peptides.

### 12.6.1 Simulation settings

In the simulation, only two overlapping peptides were considered. For the isotopic distribution, we chose three sets of ratios: an average one (denoted by **A**) by a Poisson approximation (see Section 2.3); the one with extremely small ratios (denoted by **E1**); and the one with extremely large ratios (denoted by **E2**) within $20001 \pm 0.5$ Da mass range. Details of the three sets of isotopic ratios are given in Section 6.4 (see Figure 6.3).

These three sets of isotopic distributions lead to 9 combinations for the isotopic ratios of the two peptides. They are respectively: **AA**, **AE1**, **AE2**, **E1A**, **E2A**, **E1E1**, **E2E2**, **E1E2** and **E2E1**.

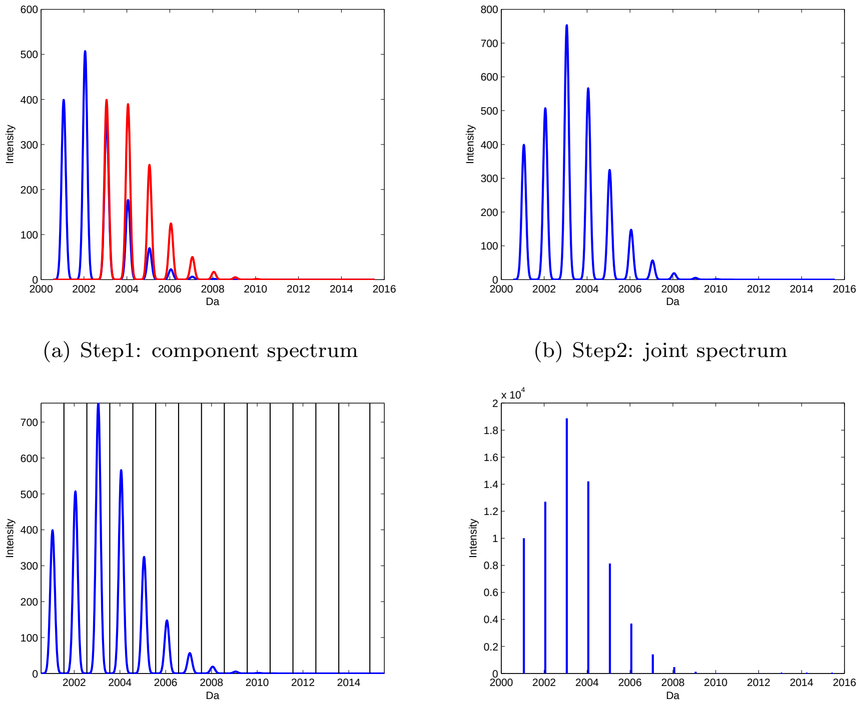
For each of the isotopic ratio combinations, three sets of $M_2$ values and 5 sets of $H_2 : H_1$ values were considered, This gives 15 settings with respect to the combinations of these parameters. The coding of these settings, which will be used in figures in the following section, is shown in Table 12.3.

**Table 12.3:** The coding of the simulation settings for the stick representation model.

| $M_2$ | $H_2 : H_1$ | | | | |
|---|---|---|---|---|---|
| | 1:1 | 0.5:1 | 0.2:1 | 1:0.5 | 1:0.2 |
| 2 | 21 | 22 | 23 | 24 | 25 |
| 3 | 31 | 32 | 33 | 34 | 35 |
| 5 | 51 | 52 | 53 | 54 | 55 |

For each of the settings shown in Table 12.3, the true values of $H_1$ and $H_2$ were obtained by multiplying $10^4$. For instance, for the setting with $H_2 : H_1 = 0.5 : 1$, $H_2$ is equal to 5000 and $H_1$ is equal to $10^4$. For all of the settings, $\sigma$ was chosen to be 10. For demonstration purposes, we generated 10 data sets per setting.

In the simulation study, one of the aims was to investigate the influence of the information reduction imposed by the stick representation. For this purpose, we generated the data sets based on the shape representation (see Figure 12.7(a)–(b)). Normal density function was used to approximate the peak shape. The peaks were then binned and summed within each bin (see Figure 12.7(c)). Afterwards, random normal noise with $\sigma = 10$ was added to each summed peak and was truncated to zero if the sum was negative (see Figure 12.7(d)). Figure 12.7 illustrates the four steps of data generation.



(a) Step1: component spectrum

(b) Step2: joint spectrum

(c) Step3: binned spectrum

(d) Step4: sticks

**Figure 12.7:** Illustration of the four steps of data generation (no measurement error).

In order to investigate the influence of either ignoring or considering the variability of the isotopic distributions, we fit two models to the simulated data sets. First, the model, which accounts for the variability of the isotopic distribution, obtained from (12.1)–(12.2) was considered. Second, a model obtained by replacing isotopic ratios $R_{l_d}$ by fixed values, obtained from an average computed values according to the method proposed by Breen *et al.* (2000), was used.

## 12.6.2 Simulation results

Figures 12.8 and 12.9 show the mean relative bias of the isotopic ratio ($R_{l_1}$ and $R_{l_2}$) estimates for the model with variable isotopic ratios. Severe bias of the isotopic ratios can be observed in Figures 12.8 and 12.9, especially for the smaller ratios (represented by larger bullet points in the figures). This may be due to the lack of information in the data for the ratios to be correctly estimated. Figures 12.10 and 12.11 indicates that these ratio estimates are closer to their prior means, which confirms the hypothesis about the insufficient amount of information in the data.

Figures 12.12 and 12.13 show the mean relative bias of $H_1$ and $H_2$ for the two models. For the estimates of $H_1$, it can be observed that they were slightly better estimated for the model with variable isotopic ratios, except for the settings of **AA**, **AE1** and **AE2**. As these are the settings when the first peptide has an average isotopic distribution, the model with fixed isotopic ratios correctly specify the isotopic distribution of the peptide. Figure 12.13 shows that, in general, the two models exhibit similar bias for $H_2$. In other words, the model with variable isotopic ratios does not offer much improvement. This may be due to the bias of the isotopic ratio estimates towards their prior means in the model.

As shown in Figure 12.14, the mean relative bias for $M_2$ is consistently zero except for a few settings when the second peptide is 5 times less abundant than the first one. Again, the model with variable isotopic ratios performs slightly better in identifying the monoisotopic mass index for the second peptide in that there appears bias only when overlap starts from the second observed peak (Figure 12.14(d)(f)(h)). These are the settings for which the second peptide is much less abundant than the first, and thus it is more difficult to be detected.

Figure 12.15 indicates that, for all the settings, $\sigma$ is severely over-estimated for both models, except for the settings of **AA** when using the model with fixed isotopic ratios. As it has been mentioned, these are the settings when the correct specification of the isotopic ratios is used for the model. For the model with variable isotopic ratios, as it has been seen from Figures 12.10 and 12.11, the isotopic ratio estimates are biased towards the prior means, thus even for the setting of **AA**, $\sigma$ was over-estimated.

## 12.7 Discussion

In this chapter, we have presented a model for the stick representation of a mass spectrum (and multiple spectra) with overlapping peptides. We applied the model to the bovine cytochrome C data set and investigated its statistical performance via a simulation study. To work with the stick representation, one needs to accept the assumptions for modeling the representation (see Section 11.1).

In the real-life data application, the second peptide was correctly detected regarding its monoisotopic mass index. The isotopic ratios of the first peptide were well estimated. On the other hand, the ratio estimates for the second peptide were in general over-estimated with an upward bias, while the relative abundance was in general under-estimated. This may be due to the incomplete labeling, given that the data were obtained from of $^{18}$O-labeling experiment. The incomplete labeling causes part of the peptide molecules of the labeled (second) peptide to have multiple shifts from 0 to 4 Da and get overlapped with the first peptide. Consequently, the first peptide, observed in a spectrum, becomes more abundant while the second peptide becomes less abundant (See Part II of the dissertation for more details). This not only distorts the isotopic distributions (especially of the first peptide), but their relative abundance.

The simulation study indicates that the model with variable isotopic ratios, in general, performs slightly better than the one with fixed ratios. The latter model works well only when it correctly specifies the isotopic distributions for the data, which in reality is not likely to happen. In the model with variable isotopic ratios, the ratio estimates were closer to their prior means instead of their true values. This indicates an insufficient amount of information imposed by the information reduction for the stick representation. A possible way to improve this is to use the shape representation of the mass spectrum. This will be considered in the next chapter.

**Table 12.4:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with $H_2/H_1 = 1/3$ for **Spectrum 1**.

| Parameter | Set 1 | | | | Set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8108 | 0.8106 | (0.78, 0.8428) | 0.8703 | 0.8812 | 0.8808 | (0.8591, 0.9064) |
| $R_{3_1}$ | 0.3567 | 0.3738 | 0.374 | (0.3539, 0.393) | 0.4223 | 0.4446 | 0.4448 | (0.4276, 0.4608) |
| $R_{4_1}$ | 0.1166 | 0.1240 | 0.124 | (0.1159, 0.1322) | 0.1478 | 0.1588 | 0.1588 | (0.1494, 0.1679) |
| $R_{5_1}$ | 0.0306 | 0.0325 | 0.0324 | (0.0295, 0.0356) | 0.0413 | 0.0445 | 0.0445 | (0.0409, 0.0482) |
| $R_{2_2}$ | 0.7933 | 0.8028 | 0.8029 | (0.7367, 0.8679) | 0.8703 | 0.8571 | 0.8559 | (0.8032, 0.9151) |
| $R_{3_2}$ | 0.3567 | 0.3651 | 0.3644 | (0.3312, 0.4013) | 0.4223 | 0.4191 | 0.4186 | (0.3861, 0.454) |
| $R_{4_2}$ | 0.1166 | 0.1200 | 0.1199 | (0.1075, 0.1324) | 0.1478 | 0.1479 | 0.1477 | (0.1346, 0.1621) |
| $R_{5_2}$ | 0.0306 | 0.0315 | 0.0315 | (0.0277, 0.0352) | 0.0413 | 0.0416 | 0.0415 | (0.0372, 0.0461) |
| $M_2$ | 5 | 5 | 5 | (5, 5) | 5 | 5 | 5 | (5, 5) |
| $\sigma$ | – | 1777.839 | 1709 | (1177.975, 2764.025) | – | 1214.990 | 1166 | (779.4975, 1940.025) |
| $H_2/H_1$ | 1/3 | 0.3063 | 0.3060 | (0.2834, 0.3310) | 1/3 | 0.3093 | 0.3093 | (0.2910, 0.3269) |

**Table 12.5:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with $H_2/H_1 = 3/1$ for **Spectrum 1**.

| Parameter | Set 2 | | | | Set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8282 | 0.8266 | (0.746, 0.9166) | 0.8703 | 0.9083 | 0.9087 | (0.8127, 1.006) |
| $R_{3_1}$ | 0.3567 | 0.3819 | 0.3806 | (0.3392, 0.4306) | 0.4223 | 0.4514 | 0.4512 | (0.3971, 0.5055) |
| $R_{4_1}$ | 0.1166 | 0.1254 | 0.1249 | (0.1104, 0.1434) | 0.1478 | 0.1595 | 0.1593 | (0.1394, 0.1807) |
| $R_{5_1}$ | 0.0306 | 0.0329 | 0.0329 | (0.0284, 0.0378) | 0.0413 | 0.0446 | 0.0445 | (0.0384, 0.0515) |
| $R_{2_2}$ | 0.7933 | 0.7931 | 0.7924 | (0.7336, 0.8553) | 0.8703 | 0.8755 | 0.8747 | (0.8135, 0.9418) |
| $R_{3_2}$ | 0.3567 | 0.3599 | 0.3598 | (0.3295, 0.3916) | 0.4223 | 0.4274 | 0.4268 | (0.3927, 0.4656) |
| $R_{4_2}$ | 0.1166 | 0.1183 | 0.1182 | (0.1072, 0.1297) | 0.1478 | 0.1506 | 0.1504 | (0.137, 0.1656) |
| $R_{5_2}$ | 0.0306 | 0.0310 | 0.0309 | (0.0278, 0.0345) | 0.0413 | 0.0423 | 0.0422 | (0.0376, 0.0474) |
| $M_2$ | 5 | 5 | 5 | (5, 5) | 5 | 5 | 5 | (5, 5) |
| $\sigma$ | – | 3231.511 | 3116 | (2160, 4983.025) | – | 3293.170 | 3172 | (2172, 5133) |
| $H_2/H_1$ | *2.4* | 2.2314 | 2.2274 | (1.9586, 2.5248) | *2.4* | 2.2189 | 2.2142 | (1.9435, 2.5200) |

**Table 12.6:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with $H_2/H_1 = 1/3$ for **Spectrum 2**.

| Parameter | Set 1 | | | | Set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8052 | 0.8049 | (0.7788, 0.8327) | 0.8703 | 0.8763 | 0.876 | (0.8577, 0.8956) |
| $R_{3_1}$ | 0.3567 | 0.3730 | 0.3732 | (0.3553, 0.3895) | 0.4223 | 0.4387 | 0.4388 | (0.4246, 0.4524) |
| $R_{4_1}$ | 0.1166 | 0.1239 | 0.1239 | (0.1162, 0.1315) | 0.1478 | 0.1568 | 0.1569 | (0.1484, 0.165) |
| $R_{5_1}$ | 0.0306 | 0.0324 | 0.0324 | (0.0296, 0.0354) | 0.0413 | 0.0439 | 0.0439 | (0.0406, 0.0474) |
| $R_{2_2}$ | 0.7933 | 0.7933 | 0.7932 | (0.7338, 0.8531) | 0.8703 | 0.8752 | 0.8745 | (0.8236, 0.9289) |
| $R_{3_2}$ | 0.3567 | 0.3605 | 0.3598 | (0.3303, 0.3942) | 0.4223 | 0.4291 | 0.4288 | (0.3981, 0.4614) |
| $R_{4_2}$ | 0.1166 | 0.1185 | 0.1184 | (0.1072, 0.1305) | 0.1478 | 0.1516 | 0.1513 | (0.1387, 0.165) |
| $R_{5_2}$ | 0.0306 | 0.0311 | 0.0311 | (0.0276, 0.0347) | 0.0413 | 0.0426 | 0.0426 | (0.0382, 0.0470) |
| $M_2$ | 5 | 5 | 5 | (5, 5) | 5 | 5 | 5 | (5, 5) |
| $\sigma$ | – | 1170.793 | 1125 | (771.1975, 1854) | – | 799.3267 | 768 | (514.495, 1273.025) |
| $H_2/H_1$ | 1/3 | 0.3058 | 0.3057 | (0.2854, 0.3267) | 1/3 | 0.3011 | 0.3011 | (0.2857, 0.3162) |

**Table 12.7:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with $H_2/H_1 = 3/1$ for **Spectrum 2**.

| Parameter | Set 2 | | | | Set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8314 | 0.8292 | (0.7491, 0.9198) | 0.8703 | 0.9087 | 0.9092 | (0.8104, 1.009) |
| $R_{3_1}$ | 0.3567 | 0.3834 | 0.3821 | (0.3398, 0.4321) | 0.4223 | 0.4514 | 0.4515 | (0.3954, 0.5063) |
| $R_{4_1}$ | 0.1166 | 0.1260 | 0.1255 | (0.1107, 0.1436) | 0.1478 | 0.1595 | 0.1594 | (0.1390, 0.1809) |
| $R_{5_1}$ | 0.0306 | 0.0330 | 0.033 | (0.0285, 0.0379) | 0.0413 | 0.0446 | 0.0445 | (0.0382, 0.0515) |
| $R_{2_2}$ | 0.7933 | 0.8015 | 0.8008 | (0.7416, 0.8637) | 0.8703 | 0.8811 | 0.8805 | (0.8185, 0.9479) |
| $R_{3_2}$ | 0.3567 | 0.3643 | 0.3643 | (0.3337, 0.3962) | 0.4223 | 0.4305 | 0.43 | (0.3955, 0.4693) |
| $R_{4_2}$ | 0.1166 | 0.1197 | 0.1197 | (0.1085, 0.1312) | 0.1478 | 0.1517 | 0.1515 | (0.138, 0.1668) |
| $R_{5_2}$ | 0.0306 | 0.0314 | 0.0313 | (0.0282, 0.0350) | 0.0413 | 0.0426 | 0.0425 | (0.0378, 0.0477) |
| $M_2$ | 5 | 5 | 5 | (5, 5) | 5 | 5 | 5 | (5, 5) |
| $\sigma$ | – | 3763.184 | 3629 | (2512, 5789.025) | – | 3917.537 | 3775 | (2587, 6088.05) |
| $H_2/H_1$ | *2.4* | 2.2048 | 2.2007 | (1.9390, 2.4901) | *2.4* | 2.2111 | 2.2061 | (1.9325, 2.5127) |

**Table 12.8:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with $H_2/H_1 = 1/3$ for **Spectrum 3**.

| Parameter | Set 1 | | | | Set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8049 | 0.8047 | (0.7765, 0.8359) | 0.8703 | 0.8770 | 0.8767 | (0.8558, 0.9002) |
| $R_{3_1}$ | 0.3567 | 0.3713 | 0.3714 | (0.3522, 0.3898) | 0.4223 | 0.4383 | 0.4384 | (0.4224, 0.454) |
| $R_{4_1}$ | 0.1166 | 0.1233 | 0.1233 | (0.1153, 0.1314) | 0.1478 | 0.1570 | 0.1571 | (0.1478, 0.1658) |
| $R_{5_1}$ | 0.0306 | 0.0323 | 0.0323 | (0.0294, 0.0354) | 0.0413 | 0.0440 | 0.0440 | (0.0405, 0.0477) |
| $R_{2_2}$ | 0.7933 | 0.8014 | 0.8011 | (0.7368, 0.8652) | 0.8703 | 0.8639 | 0.8633 | (0.8104, 0.92) |
| $R_{3_2}$ | 0.3567 | 0.3645 | 0.3639 | (0.3322, 0.4002) | 0.4223 | 0.4234 | 0.4231 | (0.3911, 0.4575) |
| $R_{4_2}$ | 0.1166 | 0.1198 | 0.1197 | (0.1077, 0.1322) | 0.1478 | 0.1495 | 0.1493 | (0.1361, 0.1634) |
| $R_{5_2}$ | 0.0306 | 0.0315 | 0.0315 | (0.0278, 0.0352) | 0.0413 | 0.0420 | 0.042 | (0.0376, 0.0465) |
| $M_2$ | 5 | 5 | 5 | (5, 5) | 5 | 5 | 5 | (5, 5) |
| $\sigma$ | – | 1748.928 | 1681 | (1158, 2733) | – | 1170.213 | 1122 | (752.9975, 1865) |
| $H_2/H_1$ | 1/3 | 0.3037 | 0.3035 | (0.2809, 0.3275) | 1/3 | 0.3023 | 0.3023 | (0.2850, 0.3192) |

**Table 12.9:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with $H_2/H_1 = 3/1$ for **Spectrum 3**.

| Parameter | Set 2 | | | | Set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8302 | 0.8285 | (0.7477, 0.9185) | 0.8703 | 0.9079 | 0.9079 | (0.8095, 1.006) |
| $R_{3_1}$ | 0.3567 | 0.3829 | 0.3816 | (0.3399, 0.432) | 0.4223 | 0.4512 | 0.4516 | (0.3957, 0.5052) |
| $R_{4_1}$ | 0.1166 | 0.1258 | 0.1252 | (0.1108, 0.1436) | 0.1478 | 0.1594 | 0.1593 | (0.1389, 0.1807) |
| $R_{5_1}$ | 0.0306 | 0.0330 | 0.0329 | (0.0285, 0.0379) | 0.0413 | 0.0446 | 0.0445 | (0.0382, 0.0515) |
| $R_{2_2}$ | 0.7933 | 0.7964 | 0.7957 | (0.7380, 0.8586) | 0.8703 | 0.8748 | 0.8736 | (0.8131, 0.9425) |
| $R_{3_2}$ | 0.3567 | 0.3611 | 0.361 | (0.331, 0.3924) | 0.4223 | 0.4274 | 0.4268 | (0.3923, 0.4663) |
| $R_{4_2}$ | 0.1166 | 0.1186 | 0.1186 | (0.1077, 0.1302) | 0.1478 | 0.1506 | 0.1503 | (0.137, 0.1655) |
| $R_{5_2}$ | 0.0306 | 0.0311 | 0.0310 | (0.0279, 0.0346) | 0.0413 | 0.0422 | 0.0421 | (0.0375, 0.0473) |
| $M_2$ | 5 | 5 | 5 | (5, 5) | 5 | 5 | 5 | (5, 5) |
| $\sigma$ | – | 3432.690 | 3310 | (2290, 5298.025) | – | 3542.675 | 3414 | (2334.95, 5516) |
| $H_2/H_1$ | *2.4* | 2.2564 | 2.2560 | (1.9836, 2.5500) | *2.4* | 2.2441 | 2.2397 | (1.9618, 2.5488) |

**Table 12.10:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with $H_2/H_1 = 1/3$ for **Spectrum 4**.

| Parameter | Set 1 | | | | Set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8096 | 0.8094 | (0.7821, 0.8388) | 0.8703 | 0.8766 | 0.8763 | (0.8569, 0.8975) |
| $R_{3_1}$ | 0.3567 | 0.3756 | 0.3757 | (0.3570, 0.3929) | 0.4223 | 0.4412 | 0.4414 | (0.4261, 0.4554) |
| $R_{4_1}$ | 0.1166 | 0.1247 | 0.1247 | (0.1168, 0.1327) | 0.1478 | 0.1574 | 0.1575 | (0.1487, 0.1658) |
| $R_{5_1}$ | 0.0306 | 0.0327 | 0.0326 | (0.0298, 0.0357) | 0.0413 | 0.0441 | 0.0441 | (0.0407, 0.0476) |
| $R_{2_2}$ | 0.7933 | 0.7956 | 0.7954 | (0.7352, 0.8578) | 0.8703 | 0.8703 | 0.8695 | (0.8182, 0.9650) |
| $R_{3_2}$ | 0.3567 | 0.3620 | 0.3613 | (0.3303, 0.3964) | 0.4223 | 0.4251 | 0.4247 | (0.3941, 0.4587) |
| $R_{4_2}$ | 0.1166 | 0.1190 | 0.119 | (0.1072, 0.1310) | 0.1478 | 0.1501 | 0.1499 | (0.1373, 0.1639) |
| $R_{5_2}$ | 0.0306 | 0.0313 | 0.0313 | (0.0277, 0.0349) | 0.0413 | 0.0422 | 0.0421 | (0.0379, 0.0466) |
| $M_2$ | 5 | 5 | 5 | (5, 5) | 5 | 5 | 5 | (5, 5) |
| $\sigma$ | – | 1438.052 | 1381 | (948.7975, 2266) | – | 942.0071 | 903.75 | (608.4975, 1501.025) |
| $H_2/H_1$ | 1/3 | 0.3062 | 0.3061 | (0.2848, 0.3286) | 1/3 | 0.3053 | 0.3054 | (0.2890, 0.3212) |

**Table 12.11:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with $H_2/H_1 = 3/1$ for **Spectrum 4**.

| Parameter | Set 2 | | | | Set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8300 | 0.8286 | (0.7486, 0.9165) | 0.8703 | 0.9068 | 0.9072 | (0.8105, 1.007) |
| $R_{3_1}$ | 0.3567 | 0.3827 | 0.3814 | (0.3404, 0.4305) | 0.4223 | 0.4507 | 0.4506 | (0.3959, 0.5055) |
| $R_{4_1}$ | 0.1166 | 0.1257 | 0.1253 | (0.1108, 0.1431) | 0.1478 | 0.1592 | 0.1591 | (0.1389, 0.1804) |
| $R_{5_1}$ | 0.0306 | 0.0330 | 0.0329 | (0.0285, 0.0378) | 0.0413 | 0.0446 | 0.0444 | (0.0383, 0.0515) |
| $R_{2_2}$ | 0.7933 | 0.7976 | 0.7968 | (0.7384, 0.8593) | 0.8703 | 0.8694 | 0.8684 | (0.8089, 0.9355) |
| $R_{3_2}$ | 0.3567 | 0.3614 | 0.3613 | (0.3316, 0.3929) | 0.4223 | 0.4246 | 0.4241 | (0.3902, 0.4632) |
| $R_{4_2}$ | 0.1166 | 0.1187 | 0.1187 | (0.1078, 0.1302) | 0.1478 | 0.1496 | 0.1493 | (0.1363, 0.1643) |
| $R_{5_2}$ | 0.0306 | 0.0311 | 0.0310 | (0.0279, 0.0347) | 0.0413 | 0.0419 | 0.0418 | (0.0373, 0.0470) |
| $M_2$ | 5 | 5 | 5 | (5, 5) | 5 | 5 | 5 | (5, 5) |
| $\sigma$ | – | 3406.034 | 3282.5 | (2275.975, 5249.1) | – | 3469.54 | 3343 | (2287, 5420) |
| $H_2/H_1$ | *2.4* | 2.2437 | 2.2394 | (1.9760, 2.5303) | *2.4* | 2.2519 | 2.2478 | (1.9716, 2.5556) |

**Table 12.12:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with $H_2/H_1 = 1/3$ for **Spectrum 5**.

| Parameter | Set 1 | | | | Set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8082 | 0.808 | (0.7804, 0.8379) | 0.8703 | 0.8912 | 0.8911 | (0.8697, 0.9134) |
| $R_{3_1}$ | 0.3567 | 0.3745 | 0.3746 | (0.3556, 0.3925) | 0.4223 | 0.4411 | 0.441 | (0.4255, 0.4567) |
| $R_{4_1}$ | 0.1166 | 0.1244 | 0.1243 | (0.1164, 0.1324) | 0.1478 | 0.1579 | 0.1579 | (0.1491, 0.1666) |
| $R_{5_1}$ | 0.0306 | 0.0326 | 0.0325 | (0.0297, 0.0357) | 0.0413 | 0.0443 | 0.0442 | (0.0408, 0.0479) |
| $R_{2_2}$ | 0.7933 | 0.8005 | 0.8004 | (0.7385, 0.8630) | 0.8703 | 0.8545 | 0.8533 | (0.802, 0.912) |
| $R_{3_2}$ | 0.3567 | 0.3636 | 0.3631 | (0.3313, 0.3993) | 0.4223 | 0.4189 | 0.4186 | (0.3872, 0.4522) |
| $R_{4_2}$ | 0.1166 | 0.1196 | 0.1195 | (0.1074, 0.132) | 0.1478 | 0.1479 | 0.1477 | (0.1351, 0.1617) |
| $R_{5_2}$ | 0.0306 | 0.0314 | 0.0314 | (0.0278, 0.0351) | 0.0413 | 0.0416 | 0.0415 | (0.0373, 0.0459) |
| $M_2$ | 5 | 5 | 5 | (5, 5) | 5 | 5 | 5 | (5, 5) |
| $\sigma$ | – | 1726.21 | 1658 | (1140, 2709) | – | 1218.135 | 1170 | (787.4, 1928) |
| $H_2/H_1$ | 1/3 | 0.3024 | 0.3022 | (0.2803, 0.3252) | 1/3 | 0.3081 | 0.3082 | (0.2906, 0.3248) |

**Table 12.13:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with $H_2/H_1 = 3/1$ for **Spectrum 5**.

| Parameter | Set 2 | | | | Set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8275 | 0.8259 | (0.7466, 0.9148) | 0.8703 | 0.9083 | 0.909 | (0.8104, 1.009) |
| $R_{3_1}$ | 0.3567 | 0.3818 | 0.3807 | (0.3394, 0.4294) | 0.4223 | 0.4515 | 0.4517 | (0.3965, 0.5063) |
| $R_{4_1}$ | 0.1166 | 0.1254 | 0.1249 | (0.1104, 0.1429) | 0.1478 | 0.1595 | 0.1594 | (0.139, 0.1809) |
| $R_{5_1}$ | 0.0306 | 0.0329 | 0.0329 | (0.0284, 0.0377) | 0.0413 | 0.0446 | 0.0445 | (0.0383, 0.0515) |
| $R_{2_2}$ | 0.7933 | 0.7900 | 0.7891 | (0.733, 0.8508) | 0.8703 | 0.8701 | 0.8692 | (0.8088, 0.9368) |
| $R_{3_2}$ | 0.3567 | 0.3582 | 0.358 | (0.3292, 0.3895) | 0.4223 | 0.4250 | 0.4245 | (0.3906, 0.4639) |
| $R_{4_2}$ | 0.1166 | 0.1177 | 0.1176 | (0.107, 0.129) | 0.1478 | 0.1497 | 0.1494 | (0.1362, 0.1646) |
| $R_{5_2}$ | 0.0306 | 0.0308 | 0.0308 | (0.0277, 0.0343) | 0.0413 | 0.0420 | 0.0419 | (0.0373, 0.0471) |
| $M_2$ | 5 | 5 | 5 | (5, 5) | 5 | 5 | 5 | (5, 5) |
| $\sigma$ | – | 3752.857 | 3615.5 | (2501, 5810.025) | – | 3972.488 | 3827 | (2608, 6205.075) |
| $H_2/H_1$ | *2.4* | 2.2461 | 2.2423 | (1.9842, 2.5249) | *2.4* | 2.2479 | 2.2443 | (1.9696, 2.5474) |

**Table 12.14:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with $H_2/H_1 = 1/3$ for **Spectrum 6**.

| Parameter | Set 1 | | | | Set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8178 | 0.8175 | (0.7902, 0.8469) | 0.8703 | 0.8822 | 0.8819 | (0.8619, 0.9042) |
| $R_{3_1}$ | 0.3567 | 0.3789 | 0.3791 | (0.3598, 0.3962) | 0.4223 | 0.4441 | 0.4443 | (0.4286, 0.459) |
| $R_{4_1}$ | 0.1166 | 0.1262 | 0.1261 | (0.1181, 0.1342) | 0.1478 | 0.1594 | 0.1594 | (0.1504, 0.1681) |
| $R_{5_1}$ | 0.0306 | 0.0330 | 0.033 | (0.0301, 0.0361) | 0.0413 | 0.0446 | 0.0447 | (0.0411, 0.0483) |
| $R_{2_2}$ | 0.7933 | 0.7981 | 0.7981 | (0.7375, 0.8593) | 0.8703 | 0.8706 | 0.8698 | (0.8183, 0.9267) |
| $R_{3_2}$ | 0.3567 | 0.3626 | 0.3621 | (0.3311, 0.3968) | 0.4223 | 0.4260 | 0.4257 | (0.3943, 0.4594) |
| $R_{4_2}$ | 0.1166 | 0.1192 | 0.1191 | (0.1075, 0.1311) | 0.1478 | 0.1504 | 0.1502 | (0.1372, 0.1641) |
| $R_{5_2}$ | 0.0306 | 0.0313 | 0.0313 | (0.0277, 0.0349) | 0.0413 | 0.0422 | 0.0422 | (0.0378, 0.0467) |
| $M_2$ | 5 | 5 | 5 | (5, 5) | 5 | 5 | 5 | (5, 5) |
| $\sigma$ | – | 1653.986 | 1588 | (1086, 2596.025) | – | 1155.196 | 1108 | (736.7975, 1857) |
| $H_2/H_1$ | 1/3 | 0.3115 | 0.3113 | (0.2897, 0.3341) | 1/3 | 0.3060 | 0.3060 | (0.2894, 0.3224) |

**Table 12.15:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with $H_2/H_1 = 3/1$ for **Spectrum 6**.

| Parameter | Set 2 | | | | Set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8305 | 0.8287 | (0.749, 0.9184) | 0.8703 | 0.9067 | 0.9074 | (0.8083, 1.006) |
| $R_{3_1}$ | 0.3567 | 0.3831 | 0.3819 | (0.3406, 0.4316) | 0.4223 | 0.4507 | 0.451 | (0.395, 0.5048) |
| $R_{4_1}$ | 0.1166 | 0.1259 | 0.1254 | (0.1109, 0.1436) | 0.1478 | 0.1592 | 0.1592 | (0.1387, 0.1807) |
| $R_{5_1}$ | 0.0306 | 0.0330 | 0.0330 | (0.0286, 0.0379) | 0.0413 | 0.0446 | 0.0445 | (0.0382, 0.0514) |
| $R_{2_2}$ | 0.7933 | 0.7993 | 0.7985 | (0.7395, 0.8624) | 0.8703 | 0.8696 | 0.8688 | (0.8092, 0.9343) |
| $R_{3_2}$ | 0.3567 | 0.3626 | 0.3624 | (0.3323, 0.3947) | 0.4223 | 0.4249 | 0.4244 | (0.3904, 0.463) |
| $R_{4_2}$ | 0.1166 | 0.1191 | 0.119 | (0.1081, 0.1307) | 0.1478 | 0.1497 | 0.1494 | (0.1363, 0.1644) |
| $R_{5_2}$ | 0.0306 | 0.0312 | 0.0312 | (0.028, 0.0348) | 0.0413 | 0.0420 | 0.0419 | (0.0373, 0.047) |
| $M_2$ | 5 | 5 | 5 | (5, 5) | 5 | 5 | 5 | (5, 5) |
| $\sigma$ | – | 3992.293 | 3850 | (2660.975, 6149.05) | – | 4285.128 | 4129 | (2825, 6690) |
| $H_2/H_1$ | *2.4* | 2.1914 | 2.1872 | (1.9293, 2.4726) | *2.4* | 2.2471 | 2.2426 | (1.9652, 2.5520) |

## Statistical results of the model fitted simultaneously to multiple spectra of the case study:

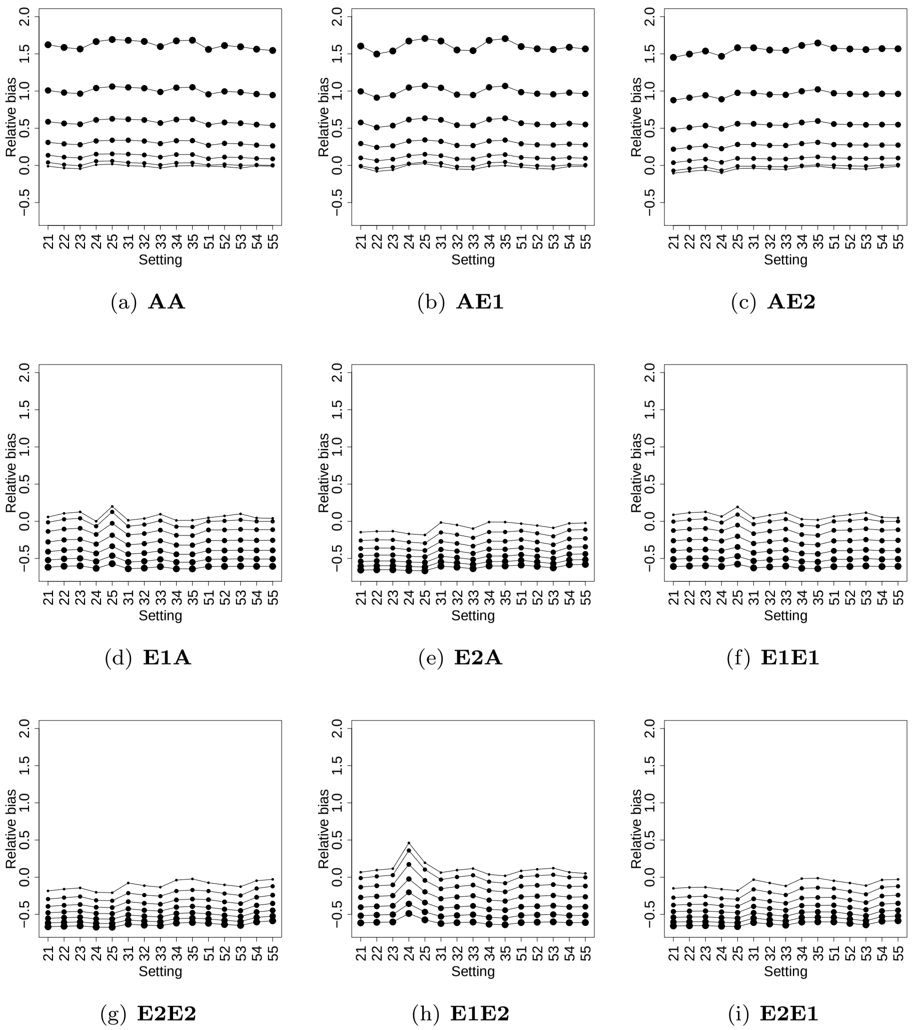
**Table 12.16:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with $\overline{Q} = 1/3$.

| Par. | Set 1 | | | | Set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8154 | 0.8155 | (0.8055, 0.8259) | 0.8703 | 0.8800 | 0.8799 | (0.8741, 0.8861) |
| $R_{3_1}$ | 0.3567 | 0.3818 | 0.3817 | (0.3739, 0.3893) | 0.4223 | 0.4459 | 0.4459 | (0.4405, 0.4511) |
| $R_{4_1}$ | 0.1166 | 0.1323 | 0.1322 | (0.1266, 0.1382) | 0.1478 | 0.1682 | 0.1683 | (0.1634, 0.1724) |
| $R_{5_1}$ | 0.0306 | 0.0347 | 0.0347 | (0.0323, 0.0374) | 0.0413 | 0.0472 | 0.0472 | (0.0441, 0.0503) |
| $R_{2_2}$ | 0.7933 | 0.7932 | 0.7932 | (0.7614, 0.8252) | 0.8703 | 0.8721 | 0.8720 | (0.8511, 0.8943) |
| $R_{3_2}$ | 0.3567 | 0.3600 | 0.3603 | (0.3416, 0.3781) | 0.4223 | 0.4229 | 0.4228 | (0.4089, 0.4378) |
| $R_{4_2}$ | 0.1166 | 0.1197 | 0.1197 | (0.1115, 0.1277) | 0.1478 | 0.1515 | 0.1516 | (0.1435, 0.1595) |
| $R_{5_2}$ | 0.0306 | 0.0317 | 0.0316 | (0.0288, 0.0347) | 0.0413 | 0.0432 | 0.0431 | (0.0399, 0.0468) |
| $M_2$ | 5 | 5 | 5 | (5, 5) | 5 | 5 | 5 | (5, 5) |
| $\sigma_1$ | – | 1434 | 1391 | (974.9, 2128) | – | 888.2 | 859.2 | (602.5, 1335) |
| $\sigma_2$ | – | 1435 | 1388 | (979.9, 2150) | – | 889.8 | 860.6 | (603.5, 1350) |
| $\sigma_3$ | – | 1434 | 1389 | (979.7, 2171) | – | 890.3 | 861.5 | (607.8, 1346) |
| $\sigma_4$ | – | 1444 | 1389 | (978.1, 2204) | – | 893.2 | 861.4 | (601.4, 1360) |
| $\sigma_5$ | – | 1440 | 1392 | (977.2, 2172) | – | 893.1 | 858.7 | (606.4, 1362) |
| $\sigma_6$ | – | 1441 | 1396 | (983.8, 2183) | – | 884.7 | 857.1 | (603.7, 1336) |
| $\overline{Q}$ | 1/3 | 0.3111 | 0.3110 | (0.2871, 0.3356) | 1/3 | 0.3044 | 0.3044 | (0.2818, 0.3276) |
| $\sigma_Q^2$ | – | 0.0007759 | 0.0005297 | (0.0001748, 0.002804) | – | 0.0007298 | 0.0004981 | (0.000164, 0.002509) |

**Table 12.17:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with $\overline{Q} = 3/1$.

| Par. | Set 2 | | | | Set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8571 | 0.8579 | (0.8112, 0.8997) | 0.8703 | 0.9339 | 0.9348 | (0.8810, 0.9848) |
| $R_{3_1}$ | 0.3567 | 0.4271 | 0.4272 | (0.3966, 0.4575) | 0.4223 | 0.5143 | 0.5136 | (0.4762, 0.5576) |
| $R_{4_1}$ | 0.1166 | 0.1443 | 0.1444 | (0.1305, 0.1568) | 0.1478 | 0.1878 | 0.1875 | (0.1705 0.2070) |
| $R_{5_1}$ | 0.0306 | 0.0377 | 0.0377 | (0.0339, 0.0416) | 0.0413 | 0.0525 | 0.0525 | (0.0469, 0.0587) |
| $R_{2_2}$ | 0.7933 | 0.7910 | 0.7910 | (0.7682, 0.8137) | 0.8703 | 0.8572 | 0.8574 | (0.8340, 0.8803) |
| $R_{3_2}$ | 0.3567 | 0.3536 | 0.3532 | (0.3391, 0.3697) | 0.4223 | 0.4161 | 0.4160 | (0.4002, 0.4327) |
| $R_{4_2}$ | 0.1166 | 0.1159 | 0.1158 | (0.1086, 0.1238) | 0.1478 | 0.1461 | 0.1461 | (0.1379, 0.1545) |
| $R_{5_2}$ | 0.0306 | 0.0304 | 0.0304 | (0.0277, 0.0332) | 0.0413 | 0.0411 | 0.0411 | (0.0375, 0.0444) |
| $M_2$ | 5 | 5 | 5 | (5, 5) | 5 | 5 | 5 | (5, 5) |
| $\sigma_1$ | – | 3257 | 3157 | (2231, 4885) | – | 3251 | 3149 | (2200, 4921) |
| $\sigma_2$ | – | 3272 | 3175 | (2211, 4949) | – | 3248 | 3135 | (2207, 5022) |
| $\sigma_3$ | – | 3259 | 3169 | (2229, 4857) | – | 3256 | 3155 | (2210, 4894) |
| $\sigma_4$ | – | 3256 | 3151 | (2214, 4870) | – | 3250 | 3142 | (2193, 4842) |
| $\sigma_5$ | – | 3245 | 3144 | (2217, 4800) | – | 3248 | 3139 | (2200, 4898) |
| $\sigma_6$ | – | 3241 | 3142 | (2228, 4835) | – | 3259 | 3159 | (2193, 4885) |
| $\overline{Q}$ | *2.4* | 2.2440 | 2.2420 | (2.1370, 2.3560) | *2.4* | 2.3110 | 2.3110 | (2.1990, 2.4310) |
| $\sigma_Q^2$ | – | 0.004559 | 0.002372 | (0.000401, 0.02242) | – | 0.004581 | 0.0024 | (0.0004319, 0.02182) |

## Graphical representation of the simulation results:



(a) **AA**

(b) **AE1**

(c) **AE2**

(d) **E1A**

(e) **E2A**

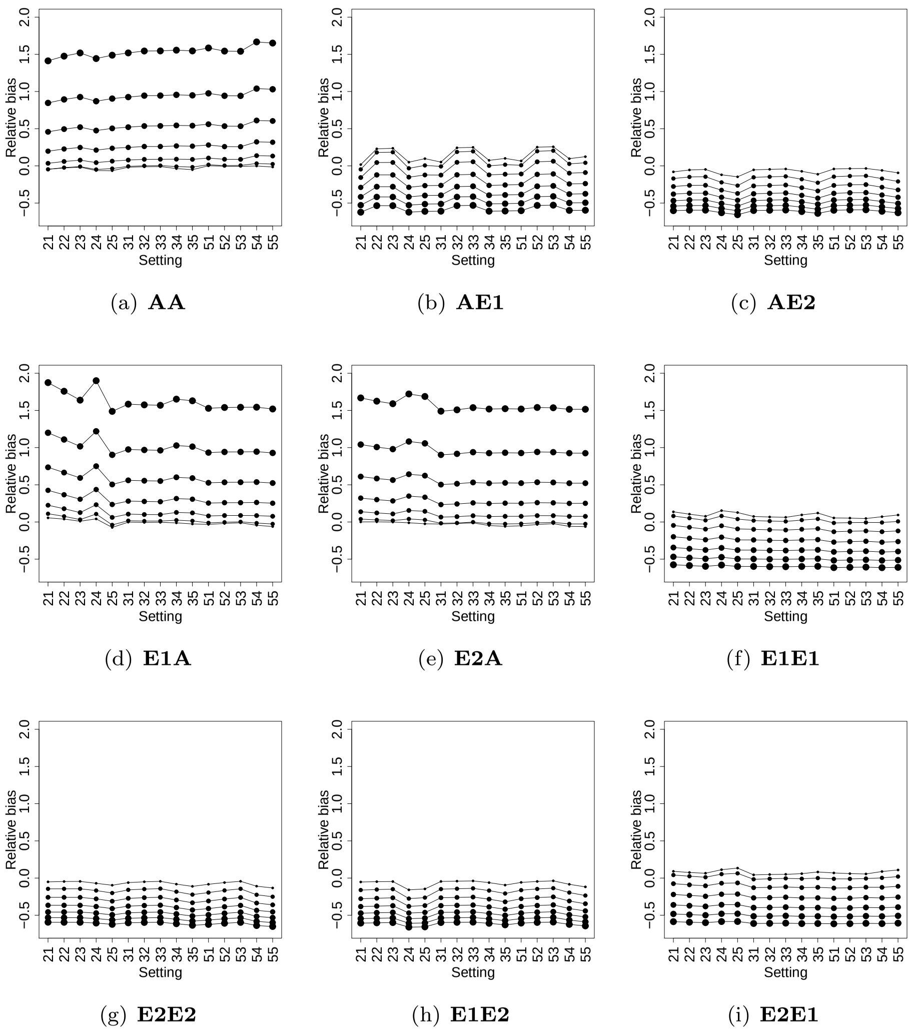
(f) **E1E1**

(g) **E2E2**

(h) **E1E2**

(i) **E2E1**

**Figure 12.8:** Mean relative bias for $R_{l_1}$ for model with variable isotopic ratios ($R_{2_1}$ – $R_{8_1}$: bullet points from small to large).



(a) **AA**

(b) **AE1**

(c) **AE2**

(d) **E1A**

(e) **E2A**

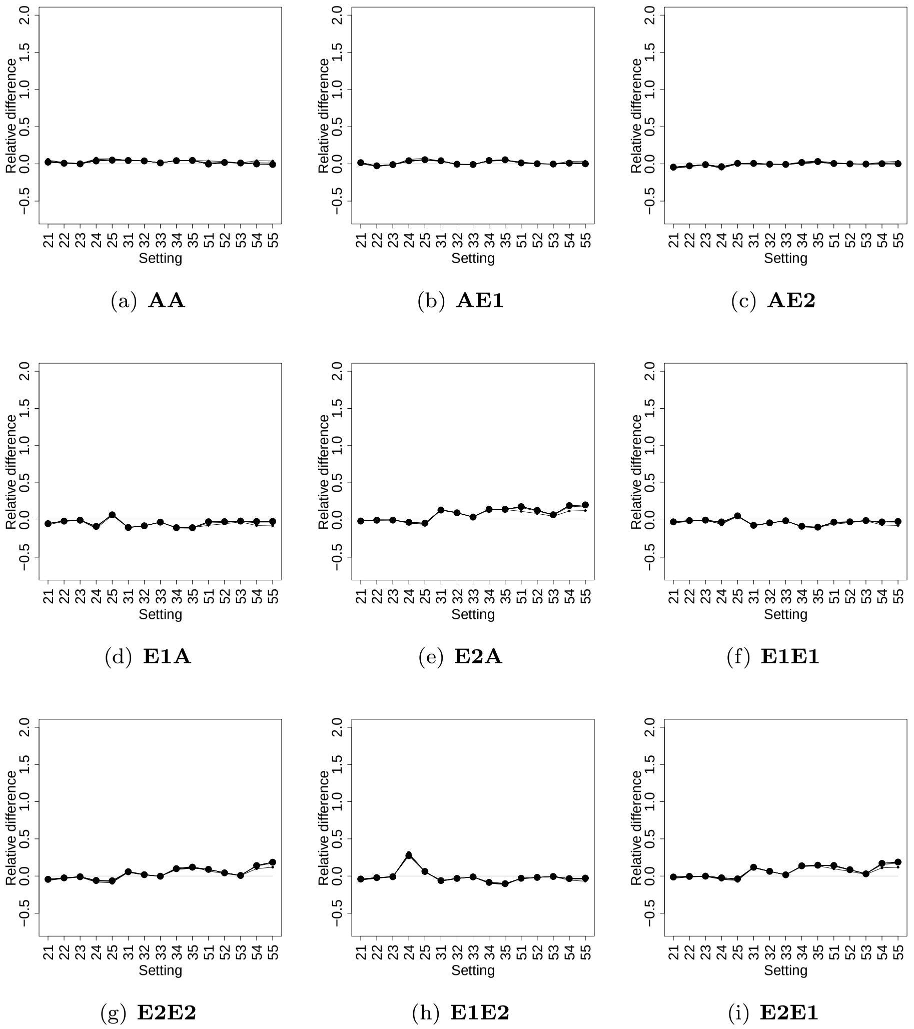
(f) **E1E1**

(g) **E2E2**

(h) **E1E2**

(i) **E2E1**

**Figure 12.9:** Mean relative bias for $R_{l_2}$ for model with variable isotopic ratios ($R_{2_2}$ – $R_{8_2}$: bullet points from small to large).



(a) **AA**

(b) **AE1**

(c) **AE2**

(d) **E1A**

(e) **E2A**

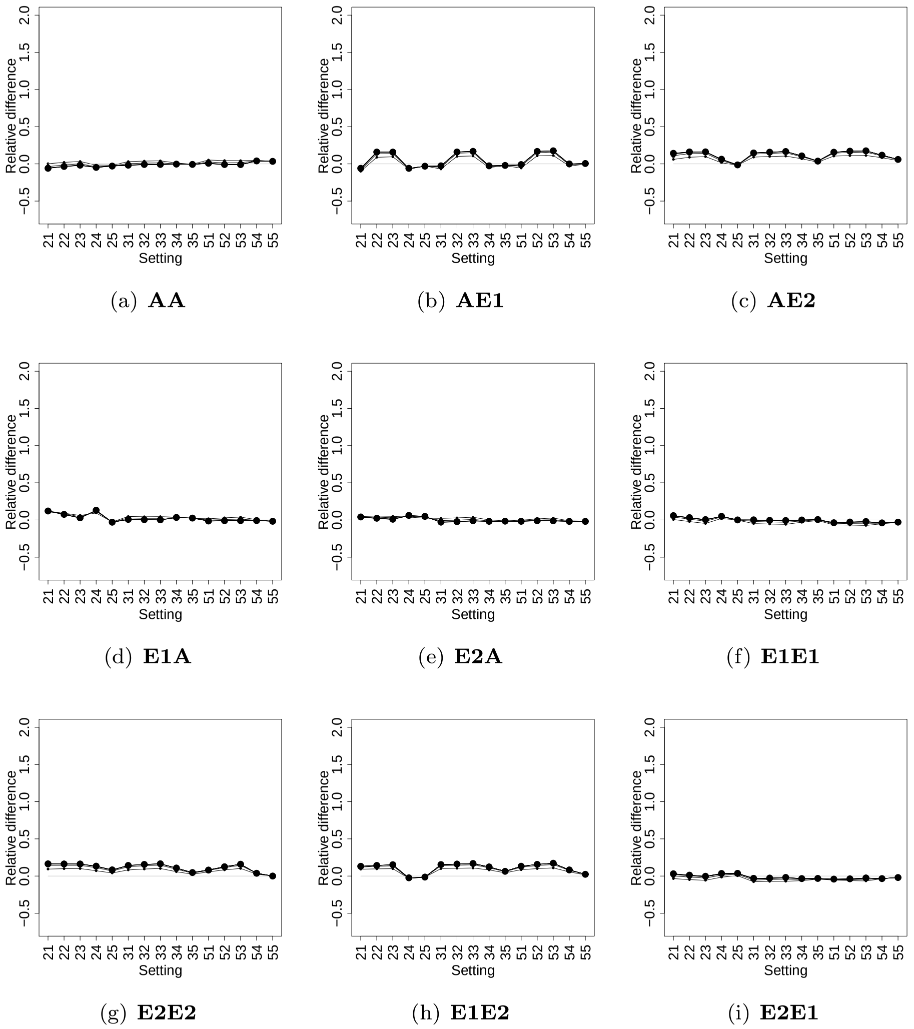
(f) **E1E1**

(g) **E2E2**

(h) **E1E2**

(i) **E2E1**

**Figure 12.10:** Mean relative difference for $R_{l_1}$ w.r.t. its prior mean for model with variable isotopic ratios ($R_{2_1}$ – $R_{8_1}$: bullet points from small to large).

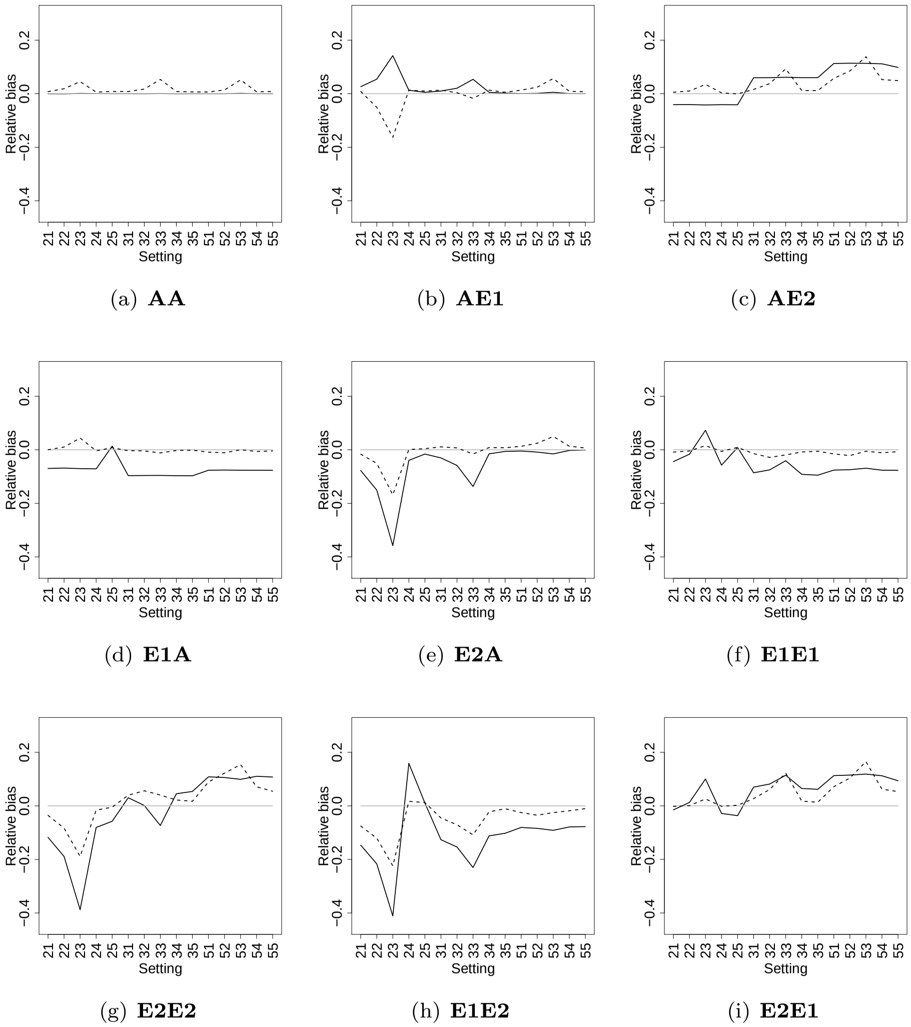


(a) **AA** (b) **AE1** (c) **AE2**

(d) **E1A** (e) **E2A** (f) **E1E1**

(g) **E2E2** (h) **E1E2** (i) **E2E1**

**Figure 12.11:** Mean relative difference for $R_{l_2}$ w.r.t. its prior mean for model with variable isotopic ratios ($R_{2_2}$ – $R_{8_2}$: bullet points from small to large).



(a) **AA**

(b) **AE1**

(c) **AE2**

(d) **E1A**

(e) **E2A**

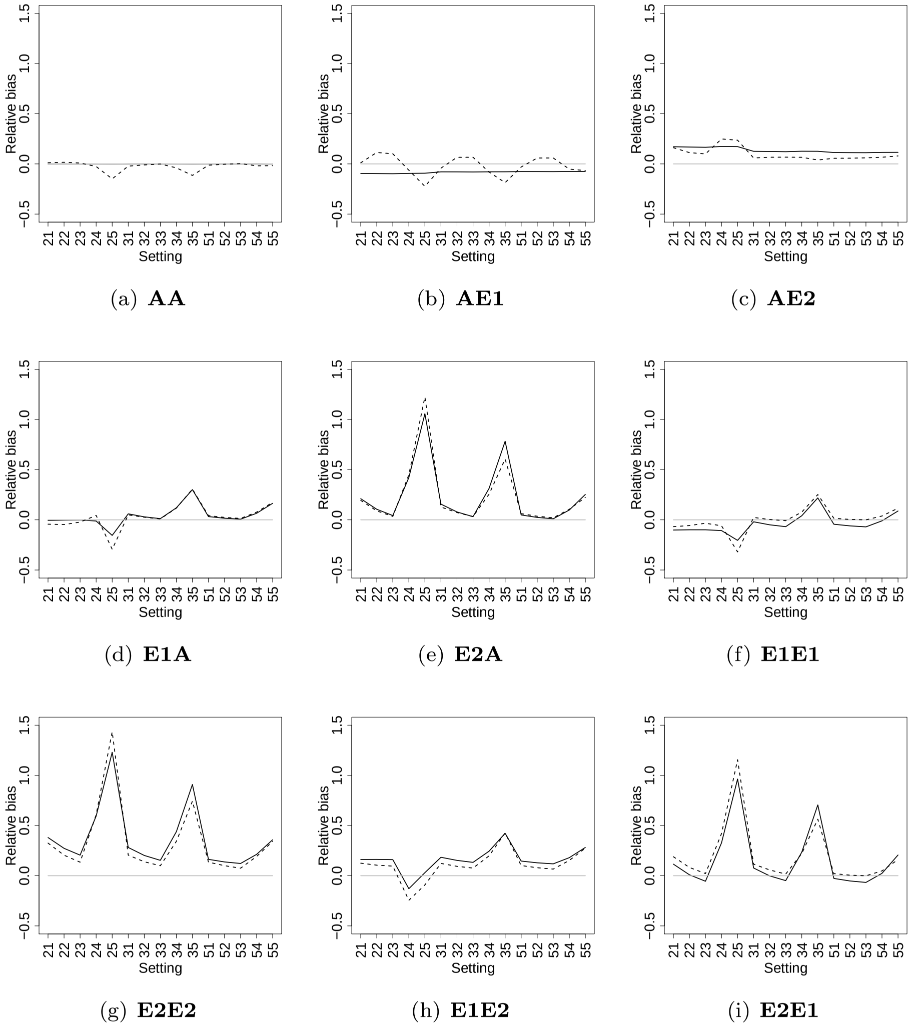
(f) **E1E1**

(g) **E2E2**

(h) **E1E2**

(i) **E2E1**

**Figure 12.12:** Mean relative bias for $H_1$ for models with fixed and variable isotopic ratios (solid lines: model with fixed isotopic ratios; dashed lines: model with variable isotopic ratios).



(a) **AA**

(b) **AE1**

(c) **AE2**

(d) **E1A**

(e) **E2A**

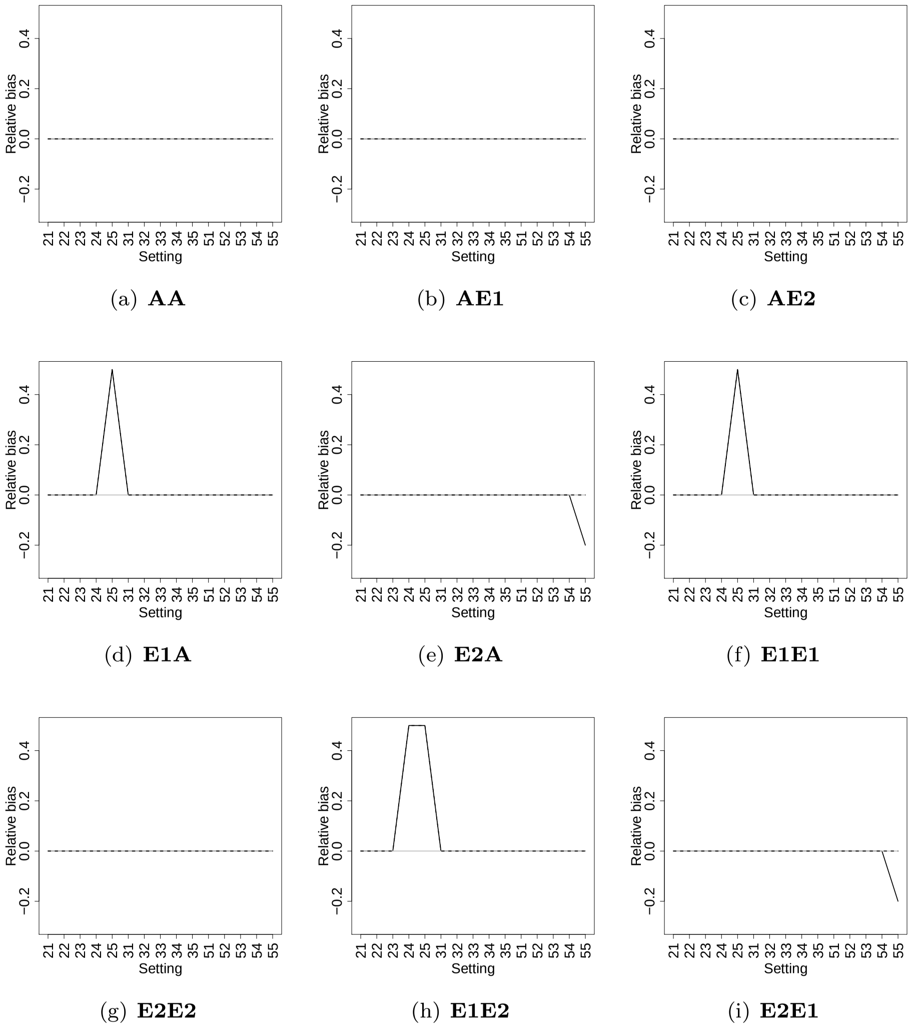
(f) **E1E1**

(g) **E2E2**

(h) **E1E2**

(i) **E2E1**

**Figure 12.13:** Mean relative bias for $H_2$ for models with fixed and variable isotopic ratios (solid lines: model with fixed isotopic ratios; dashed lines: model with variable isotopic ratios).

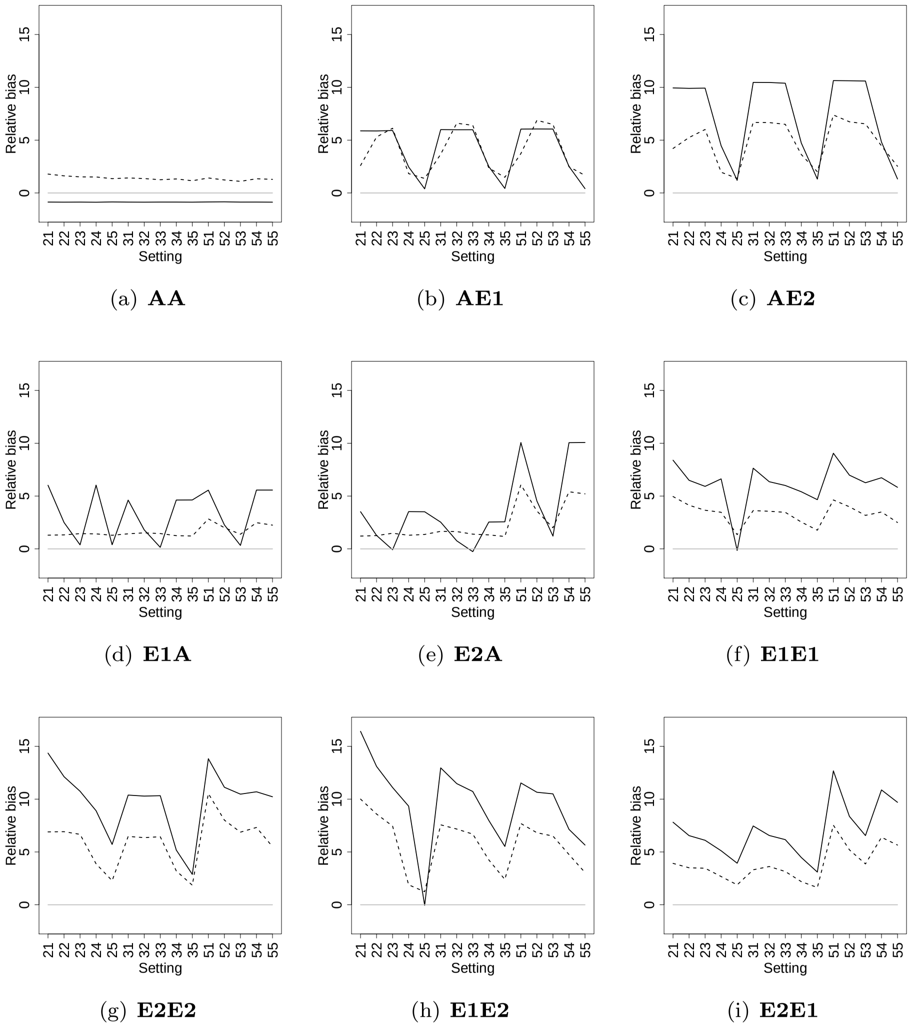


(a) **AA** (b) **AE1** (c) **AE2**

(d) **E1A** (e) **E2A** (f) **E1E1**

(g) **E2E2** (h) **E1E2** (i) **E2E1**

**Figure 12.14:** Mean relative bias for $M_2$ for models with fixed and variable isotopic ratios (solid lines: model with fixed isotopic ratios; dashed lines: model with variable isotopic ratios).



(a) **AA** (b) **AE1** (c) **AE2**

(d) **E1A** (e) **E2A** (f) **E1E1**

(g) **E2E2** (h) **E1E2** (i) **E2E1**

**Figure 12.15:** Mean relative bias for $\sigma$ for models with fixed and variable isotopic ratios (solid lines: model with fixed isotopic ratios; dashed lines: model with variable isotopic ratios).

# Chapter 13

# A Bayesian model for the shape representation of a mass spectrum with overlapping peptides

## 13.1 Introduction

The sum of the peak intensities, used in the stick representation of the data (see Section 11.1), is a rough approximation of the area under the peak-curve, which corresponds to the abundance of a particular isotopic variant. The approximation requires the assumption that the number of data points per bin is approximately the same. In general, this assumption is plausible, because the number of data points per 1 Da is fairly constant within a 20 Da mass-range. Nevertheless, the stick representation of data implies two potential limitations:

1. non-identifiability of the overlapping peptides when the difference of the monoisotopic masses is less than 1 Da;
2. inability to estimate the exact monoisotopic masses of the overlapping peptides.

Moreover, as it has been observed in Chapter 12, the information reduction results in biased estimates of isotopic ratios. To address these limitations, in this chapter, we consider a model for the peak-shape representation of a spctrum. By taking into account the peak shapes, all measurements in a mass spectrum are used and the limitations of the stick representation are hopefully avoided. In this section, we describe a Bayesian model for the shape representation.

## 13.2 Model formulation for the shape representation

Again, we assume that the number of the overlapping peptides, $D$, is known. Essentially, the model formulation is similar to the one presented in Chapter 12 for the stick representation of the MS data. The main modification is to replace the indicator function $I(M_d + l - 1 - i = 0)$ in equation (12.2) with a suitable peak-shape function $\psi(x_i; \mu_{l_d}, \sigma_s)$. Thus, for the observed intensities $y_i^*$ $(i = 1, \ldots, N^*)$ we assume the following model:

$$y_i^* \sim N(E(y_i^*), \sigma^2) \tag{13.1}$$

with

$$\begin{aligned} E(y_i^*) &= f(\boldsymbol{H}, \boldsymbol{R}, \boldsymbol{M}^*, \sigma_s, S) \\ &= \sum_{d=1}^{D} \sum_{l=1}^{L} H_d R_{l_d} \psi(x_i; M_d^* + (l-1)S, \sigma_s) \end{aligned} \tag{13.2}$$

where $x_i$ is the mass coordinate corresponding to intensity $y_i^*$, $\boldsymbol{M}^* = (M_1^*, M_1^*, ..., M_D^*)$ is a vector of monoisotopic masses of the $D$ overlapping peptides, with $M_d^* < M_{d+1}^*$, $S$ is the difference in mass locations between two neighboring isotopic peaks of the same peptide, and assumed to be constant over all the isotopic peaks for all the overlapping peptides. In (13.2), $\psi(x; \mu, \sigma_s^2)$ is a function of a chosen distribution, defined for the shape of the peaks. With this respect, either *cdf*(cumulative distribution function) or *pdf*(probability distribution function) can be used. To approximate the (underlying) continuous mass coordinate, we chose to use the *cdf*, which is also a more accurate approximation of area under the curve, especially when the dispersion parameter, $\sigma_s$, takes very small values. For a normal distribution function, the area under the curve between two neighboring mass coordinates is:

$\psi(x_i; M_d^* + (l-1)S, \sigma_s^2) =$

$$\begin{cases} \Phi(x_i|M_d^* + (l-1)S, \sigma_s^2) - \Phi(x_{i-1}|M_d^* + (l-1)S, \sigma_s^2) & \text{if } i \geq 2, \\ \Phi(x_i|M_d^* + (l-1)S, \sigma_s^2) - \Phi(0|M_d^* + (l-1)S, \sigma_s^2) & \text{if } i = 1 \end{cases}, \tag{13.3}$$

with $\Phi(x_i|M_d^* + (l-1)S, \sigma_s^2)$ corresponding to the normal *cdf* function calculated at $x_i$ with mean $M_d^* + (l-1)S$ and variance $\sigma_s^2$.

Peaks in MS data often exhibit a right-skewed shape. Thus, an alternative is to approximate the shape by a function that accounts for an asymmetric shape. Asymmetric Laplace function can serve for this purpose. In this case, an extra shape parameter – the skewness parameter $\kappa$ should be included and the shape function becomes:

$\psi(x_i; M_d^* + (l-1)S, \sigma_s, \kappa) =$

$$\begin{cases} F(x_i|M_d^* + (l-1)S, \sigma_s, \kappa) - F(x_{i-1}|M_d^* + (l-1)S, \sigma_s, \kappa) & \text{if } i \geq 2, \\ F(x_i|M_d^* + (l-1)S, \sigma_s, \kappa) - F(0|M_d^* + (l-1)S, \sigma_s, \kappa) & \text{if } i = 1, \end{cases} \tag{13.4}$$

with $F(x_i|M_d^* + (l-1)S, \sigma_s, \kappa)$ denoting the *cdf* function of an asymmetric Laplace distribution calculated at $x_j$ with mean $M_d^* + (l-1)S$ and standard deviation $\sigma_s$, i.e., $F(x_i|M_d^* + (l-1)S, \sigma_s, \kappa) =$

$$\begin{cases} \frac{\kappa^2}{1+\kappa^2} \exp\left[-\frac{\sqrt{2}}{\sigma_s \kappa} |x_i - (M_d^* + (l-1)S)\,|\right] & \text{if } x_i < M_d^* + (l-1)S, \\ 1 - \frac{1}{1+\kappa^2} \exp\left[-\frac{\sqrt{2}\kappa}{\sigma_s} |x_i - (M_d^* + (l-1)S)\,|\right] & \text{if } x_i \geq M_d^* + (l-1)S. \end{cases}$$

## 13.3 Prior distributions

For $H_d$, $\sigma^2$, $\sigma_s$, $S$, and $\kappa$, in model (13.1)–(13.2), we use the following non-informative priors:

$$\begin{aligned} H_d &\sim N\left(0, \frac{1}{\tau}\right) I(H_d \geq 0), & (13.5) \\ \sigma^{-2} &\sim \Gamma(\alpha, \beta), & (13.6) \\ \sigma_s &\sim N(0, 10^6) I(0 \leq \sigma_s \leq 0.5), & (13.7) \\ S &\sim N\left(1, \frac{1}{\tau_s}\right) \text{ with } \tau_s \sim \Gamma\left(\alpha^{**}, \beta^{**}\right) I\left(\tau_s \geq 1600\right), & (13.8) \\ \kappa &\sim U(0.01, 0.99), & (13.9) \end{aligned}$$

where $\tau$, $\alpha$, $\beta$, $\alpha^*$, $\beta^*$ $\alpha^{**}$ and $\beta^{**}$ are positive constants close to zero. To avoid numerical problems, $H_d$ is constrained to be non-negative and the hyper-prior for $H_d$ (see equation (12.3)) is replaced by a constant $\tau$. The peak-width parameter $\sigma_s$ is constrained to be positive and not larger than 0.5, because peaks observed in a spectrum, see right panel of Figure 2.7, are clearly separated from each other, with the width of a peak not larger than 1 Da. Parameter $S$ is the average difference in mass difference of two neighboring isotopic peaks of a peptide and is usually very close to one. This is reflected in the prior by setting a lower bound to the precision parameter $\tau_s$. Finally, the skewness parameter $\kappa$ is assumed to be a priori smaller than one, since the observed peak shapes are always (at least in the MALDI-TOF data) skewed to the right.

The informative prior for the isotopic ratios $R_{l_d}$ is the same as defined in (12.8)–(12.9) (details see Section 12.3.1).

As mentioned in Section 3.1, some prior information for the estimation of monoisotopic masses $\boldsymbol{M^*}$ is available from public databases, e.g., the NCBI data. This is discussed in the following section.

### 13.3.1 Informative prior for the monoisotopic masses $\boldsymbol{M^*}$

Figure 3.1 shows that monoisotopic masses appear in 'clusters' of a "bell" shape, which indicates that a suitable choice for the prior distribution of $\boldsymbol{M}^*$ may be a mixture of normals. More concretely, the prior for the monoisotopic mass of the $d$th peptide is defined as:

$$M_d^* \sim \sum_{g=1}^{G} \pi N(\eta_g, \sigma_m^2), \tag{13.10}$$

where $\pi = 1/G$, and $G$ is the number of normal components. For example, from Figure 3.1, it follows that, if the monoisotopic mass is likely to appear in any of the 5 clusters, we would choose $G = 5$, i.e., define 5 normal components for the mixture. To avoid adding any subjective information, at which cluster the monoisotopic mass $M_d^*$ is likely to occur, we define the probability of the mixture normal components to be equal, i.e., $\pi = 1/G$. The means $\eta_g$ and the variance $\sigma_m^2$, are estimated from the NCBI data. Variance $\sigma_m^2$ is assumed to be constant for all the different components and is chosen to be equal to the maximum value of the variances of these components from the NCBI data.

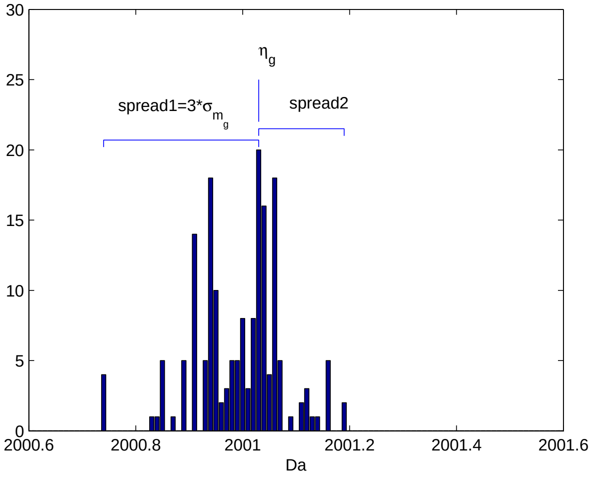


**Figure 13.1:** Graphical demonstration of the estimation of mean $\eta_g$ and standard deviation $\sigma_{m_g}$ for the prior normal density of $M_d^*$.

To estimate the parameters, taking a mass window of 2000 ± 2.5 Da as an example (Figure 3.1), each 'cluster' of the histogram is treated as a component of the normal mixture. Because of the assymetric shape of these clusters, we define the normal means $\eta_g$ to be the mode (instead of the mean) of each histogram (see Figure 13.1). By assuming that the observed peptides in a histogram represent 99.9% of all the peptides around the particular cluster of masses, the standard deviation ($\sigma_{m_g}$) of the normal density can then be defined as one third of the observed spread around the median $\eta_g$ of that histogram. As the histograms are assymetric, the spread to the left (*spread*1) and the right (*spread*2) of the mode are usually unequal. We define the spread of each histogram to be the maximum of the two. As an example shown in Figure 13.1, the standard deviation $\sigma_{m_g}$ is defined to be one third of the left spread *spread*1. Finally, the common variance is then taken as $\sigma_m^2 = \max(\sigma_{m_g}^2)$.

### Practical implementation

Practically, in the Bayesian framework, sampling of $M_d^*$ from the normal mixture was carried out in four steps:

1. estimate the posterior probability of the normal mixture for the $g$th component,

denoted by $\pi_g^*$;

2. conditional on the vector of posterior probabilities $\boldsymbol{\pi}^*$ for these mixture components, sample a component out of the $G$ normal mixture components, by means of an indicator variable, denoted as $\imath$ ($\imath = 1, \cdots, G$);

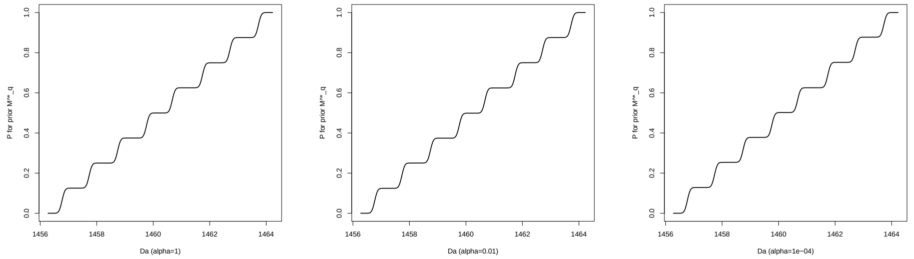
3. conditional on $\imath$ (a certain component being selected), sample the error, denoted by $\varepsilon_{M_d^*}$, from the normal prior: $N(0, \sigma_m^2)$;

4. consequently, $M_d^* = \eta_\imath + \varepsilon_{M_d^*}$.

For the $1^{st}$ step, a non-informative Dirichlet prior was used. Zhu and Lu (2004) suggested that for probability settings, a non-informative prior is not a *flat* prior (when the parameter of the prior of Dirichlet $\alpha_g = 1$, $g = 1, \ldots, G$), but a prior that has probability mass close to zero everywhere except the boundaries. This means that the Dirichlet prior should have all $\alpha_g$ equal to a positive value, but as small as possible. Such a non-informative prior is similar in the spirit to Jeffrey's prior. The intuition behind this theory is that, by specifying each of the $\alpha_g$ equal to 1, one subjectively adds one observation to each of the categories (normal components). Thus, information content is falsely increased by these pseudo-observations. As a result, the posterior distribution would also be very flat even if data indicate clear signal about $M_d^*$.

Figure 13.2 shows an example of the cumulative probability for the prior of $M_d^*$ with the three different choices of $\alpha_g$. They look quite similar and all of them seem to be quite non-informative as the cumulative probability function increases steadily across the possible mass range. However, huge difference in the posterior distributions can be caused by the different choices of $\alpha_g$ values. Figure 13.3 demonstrates the resulting posterior cumulative probability with the three choices of $\alpha_g$. It indicates that when the data contain clear information about the location of $M_d^*$, smaller values of $\alpha_g$ lead to more probability mass around the correct mass location whereas $\alpha_g = 1$ leads to a very flat posterior distribution. Therefore, to estimate $M_d^*$ properly, one needs to be cautious about the choice of $\alpha_g$, i.e., to choose $\alpha_g$ as small as possible.

The $2^{nd}$ step can be realized by drawing a sample of $\imath$ from a multinomial distribution with probabilities $\boldsymbol{\pi}^*$, obtained from the Dirichlet distribution. One may argue that the $3^{rd}$ and $4^{th}$ steps can be combined into one step, because they are equivalent to sampling directly from $N(\eta_\imath, \sigma_m^2)$. In practice however, this single step requires



(a) $\alpha_g = 1$ (b) $\alpha_g = 0.01$ (c) $\alpha_g = 0.0001$

**Figure 13.2:** Graphical representation of the cumulative probability of the prior for $M_d^*$.

specifying an initial value for the mass $M_d^*$. Such initial value will force the MCMC samples to stick with the normal component, at which the initial value is specified.

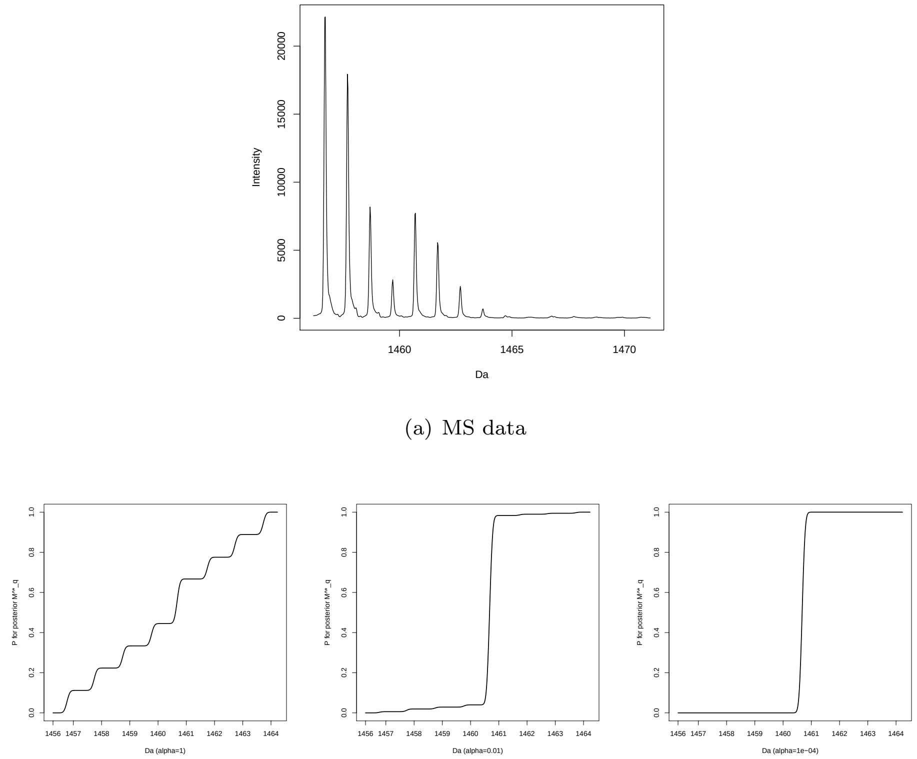
### The analytical solution of the prior mean and variance

The mean and variance of the prior normal mixture, given the Dirichlet prior for $\boldsymbol{\pi}$, can be computed analytically:

$$E\left[\sum_{g=1}^{G}\pi_g\left(\eta_g+\varepsilon_g\right)\right]=\frac{1}{G}\sum_{g=1}^{G}\eta_g. \tag{13.11}$$

$$\begin{aligned}
&\operatorname{Var}\left[\sum_{g=1}^{G}\pi_g\left(\eta_g+\varepsilon_g\right)\right]\\
=&E\left\{\operatorname{Var}\left[\sum_{g=1}^{G}\pi_g\left(\eta_g+\varepsilon_g\right)|\eta_g,\sigma_m^2\right]\right\}+\operatorname{Var}\left\{E\left[\sum_{g=1}^{G}\pi_g\left(\eta_g+\varepsilon_g\right)|\eta_g,\sigma_m^2\right]\right\}\\
=&E\left[\sum_{g=1}^{G}\pi_g^2\sigma_m^2\right]+\operatorname{Var}\left[\sum_{g=1}^{G}\pi_g\eta_g\right]\\
=&\sum_{g=1}^{G}\sigma_m^2E\left(\pi_g^2\right)+\sum_{g=1}^{G}\operatorname{Var}\left(\pi_g\eta_g\right)+2\sum_{i}\sum_{j}\operatorname{Cov}\left(\pi_i\eta_i,\pi_j\eta_j\right)
\end{aligned} \tag{13.12}$$

When $\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_G)$ with $\alpha_1, \ldots, \alpha_G = \alpha$, each of $\pi_g$ marginally



(a) MS data

(b) $\alpha_g = 1$

(c) $\alpha_g = 0.01$

(d) $\alpha_g = 0.0001$

**Figure 13.3:** Graphical representation of the cumulative probability of the posterior for $M_d^*$ by applying the priors of Figure 13.2 to data shown in (a).

follows a Beta distribution, i.e., $\pi_g \sim \text{Beta}\left(\alpha, (G-1)\alpha\right)$. Therefore, we have:

$$E\left(\pi_g^2\right) = \frac{\alpha\left(\alpha+1\right)}{G\alpha\left(G\alpha+1\right)},$$

$$\text{Var}\left(\pi_g\right) = \frac{\alpha\left(G\alpha-\alpha\right)}{G^2\alpha^2\left(G\alpha+1\right)} = \frac{G-1}{G^2\left(G\alpha+1\right)},$$

and

$$\text{Cov}\left(\pi_i, \pi_j\right) = \frac{-\alpha^2}{G^2\alpha^2\left(G\alpha+1\right)} = -\frac{1}{G^2\left(G\alpha+1\right)}.$$

Equation (13.12) can thus be written as:

$$
\begin{aligned}
&Var\left[\sum_{g=1}^{G}\pi_g\left(\eta_g+\varepsilon_g\right)\right]\\
&= \frac{(\alpha+1)\sum\limits_{g=1}^{G}\sigma_m^2}{G(G\alpha+1)} + \frac{G-1}{G^2(G\alpha+1)}\sum_{g=1}^{G}\eta_g^2 - 2\frac{1}{G^2(G\alpha+1)}\sum_{i=1}^{G-1}\sum_{j=i+1}^{G}\eta_i\eta_j\\
&= \frac{(\alpha+1)\sigma_m^2}{G\alpha+1} + \frac{G-1}{G^2(G\alpha+1)}\sum_{g=1}^{G}\eta_g^2 - 2\frac{1}{G^2(G\alpha+1)}\sum_{i=1}^{G-1}\sum_{j=i+1}^{G}\eta_i\eta_j. \qquad (13.13)
\end{aligned}
$$

## 13.4 Conditional posterior distributions

The conditional posterior distributions of $H_d$ and $\sigma^2$ can be derived similarly as in Section 12.4. Their analytical solutions are:

$$p(H_d|\boldsymbol{y},\boldsymbol{M}^*,\sigma^2,\sigma_s,S) \propto N(\mu_{H_d},\sigma_{H_d}^2)I(H_d \geq 0) \qquad (13.14)$$

with

$$\mu_{H_d} = \frac{\sum\limits_{i=1}^{N^*} y_i^*\left(\sum\limits_{l=1}^{L} R_l\cdot\psi(x_i-(M_d^*+(l-1)S),\sigma_s)\right)}{\tau\sigma^2+\sum\limits_{i=1}^{N^*}\left\{\sum\limits_{l=1}^{L}\left[R_l\cdot\psi(x_i-(M_d^*+(l-1)S),\sigma_s)\right]\right\}^2}$$

and

$$\sigma_{H_d}^2 = \frac{1}{\left\{\tau+\sigma^{-2}\left[\sum\limits_{i=1}^{N^*}\left(\sum\limits_{l=1}^{L}R_l\cdot\psi(x_i-(M_d^*+(l-1)S),\sigma_s)\right)^2\right]\right\}},$$

$$p(\sigma^{-2}|\boldsymbol{y}^*,\boldsymbol{H},\boldsymbol{M}^*,\sigma_s,S) \propto \Gamma\left(\alpha+\frac{N^*}{2},\beta+\frac{\sum\limits_{i=1}^{N^*}[y_i^*-E(y_i^*)]^2}{2}\right) \qquad (13.15)$$

Clearly, the posterior for $M_d^*$ should also be a mixture of normal, taking the form of $\sum\limits_{g=1}^{G}\pi_g^* N(\eta_g^*,\sigma_m^{2*})$, with $\pi_g$ equal to the probability of the $g^{th}$ normal component, obtained from the Dirichlet distribution. However, no analytical solution exists for the conditional posterior distribution of $M_d^*$, as well as for parameters $\sigma_s$, $\kappa$, $S$, and $C_{l_d}^*$. These distributions therefore need to be evaluated by numerical (sampling) methods, e.g., a Metropolis-Hasting algorithm with acception-rejection rules.

## 13.5 Application to the data

To investigate the performance of the developed model (13.1)–(13.2), we fitted it to real-life data. The model was fitted by using the *R* package *R2WinBUGS*, built in *R* to automatically call the *WinBUGS1.4* software, which allows to fit Bayesian models. To implement the model in *WinBUGS1.4*, the priors, specified in Section 13.3, were considered.

The model was applied to the same peptides of the data set with replicated joint mass spectra of bovine cytochrome C from LC Packings, as described in Chapter 12. More specifically, we considered four settings of the data, each with 6 spectra:

| | | | |
|---|---|---|---|
| Setting 1: | $M_1^* = 1456.66$ | $M_2^* = 1460.67$ | $H_2/H_1 = 1/3$ |
| Setting 2: | $M_1^* = 1456.66$ | $M_2^* = 1460.67$ | $H_2/H_1 = 3/1$ |
| Setting 3: | $M_1^* = 1584.76$ | $M_2^* = 1588.77$ | $H_2/H_1 = 1/3$ |
| Setting 4: | $M_1^* = 1584.76$ | $M_2^* = 1588.77$ | $H_2/H_1 = 3/1$ |

An example of the original spectra is shown in Figure 12.6.

### 13.5.1 Results of the model fit

The parameters of main interest are:

- the monoisotopic masses of the two peptides, $\boldsymbol{M^*}$;
- the relative abundance $H_2/H_1$.

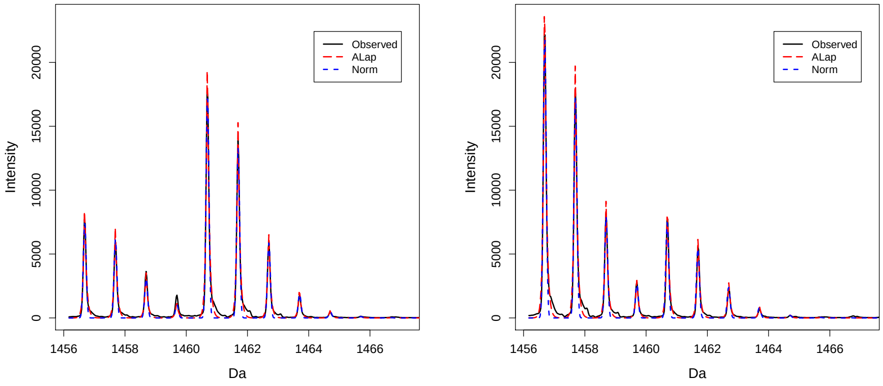
**Fit for the individual spectrum**

Tables 13.3 to 13.14 show the point estimates (means and medians of the posterior distributions) and the 95% credible intervals for the parameters of model (13.1)–(13.2) for the shape representation with normal *cdf* as a function for the peak shape for the four settings, with 6 spectra, based on 20,000 samples (with 10,000 burn-in samples) from the posterior distribution. Tables 13.15 to 13.26 show the results of the model with an asymmetric Laplace distribution as the shape function for the same spectra, based on the same number of samples of the posterior distribution.

Several patterns can be observed from the tables. First, for all of the data sets, the monoisotopic mass of the second peptide $M_2^*$ is estimated at the correct, $5^{th}$ peak, indicating the appropriateness of the model specification and the prior distribution for the parameter. Second, the point estimates of the relative abundance $H_2/H_1$ for all the cases are underestimated, whereas the 95% credible intervals are much narrower than the ones shown in Section 12.5. For settings 2 and 4, the 95% credible interval covers the true value, which is in general not the case for the other two settings.

It is also worth noting that the point estimates of the isotopic ratios are much closer to the true values with their 95% credible intervals much narrower than those for the stick model, presented in Section 12.5.

Figure 13.4 shows the fitted spectra of the models with the normal and asymmetric Laplace distribution (shape-)functions versus the observed spectra for the peptide with monoisotopic mass 1456.66 Da. It can be seen that asymmetric Laplace distribution function provides a better fit to the peak envelopes than the normal-density function. This is ascertained by the fact that the residual variances obtained from the model with the asymmetric Laplace shape function are consistently smaller than the ones with normal-density function (refer to Tables 13.3 to 13.26).



(a) $H_2/H_1 = 3/1$ (b) $H_2/H_1 = 1/3$

**Figure 13.4:** The fitted spectra using normal-density and asymmetric Laplace as the shape function versus the observed spectra for peptide with monoisotopic mass $M_1^* = 1456.66$Da for the $1^{st}$ replicate (of the mass spectra).

### Simultaneous fit to six technical replications of spectra

Similar to the model for the stick representation, presented in Chapter 12, a model incorporating random effects for the relative abundance parameter, denoted as $Q_j$ for the $j$th spectrum, can be fitted simultaneously to the six technical replicates of the spectra. More specifically, based on the model defined in (13.1)-(13.2), the intensity of the $i$th coordinate in the $j$ spectrum is:

$$y^*_{ij} \sim N(E(y^*_{ij}), \sigma^2_j), \; i = 1, \ldots, N^*, \tag{13.16}$$

with

$$\begin{aligned} E(y^*_{ij}) &= f(\boldsymbol{H}, \boldsymbol{R}, \boldsymbol{M^*}, \boldsymbol{Q}, \sigma_s, S) \\ &= \sum_{l=1}^{L} H_{1_j} R_{l_1} \psi(x_{ij}; M^*_1 + (l-1)S, \sigma_s) + \sum_{l=1}^{L} H_{2_j} R_{l_2} \psi(x_{ij}; M^*_2 + (l-1)S, \sigma_s) \\ &= \sum_{l=1}^{L} H_{1_j} R_{l_1} \psi(x_{ij}; M^*_1 + (l-1)S, \sigma_s) + \sum_{l=1}^{L} H_{1_j} Q_j R_{l_2} \psi(x_{ij}; M^*_s + (l-1)S, \sigma_s), \end{aligned} \tag{13.17}$$

where $Q_j$ is random:

$$Q_j \sim N(\overline{Q}, \sigma^2_Q).$$

The results of fitting the model with normal and asymmetric Laplace distribution functions are presented, respectively, in Tables 13.27–13.28, Tables 13.29–13.30. The monoisotopic mass of the second peptide $M^*_2$ is again estimated correctly at the $5^{th}$ peak, for all the settings. There is still slight under-estimation for the mean relative abundance parameter $\overline{Q}$, but the estimates are closer to the true value. The estimates of between-spectra variability of the relative abundance, captured by $\sigma^2_Q$, are again very small due to the fact that the samples from different spectra are the same biological samples.

The isotopic ratio estimates are closer to the true values with the 95% credible intervals much narrower than in the model for the stick representation (see Section 12.5). This is because the shape representation retains the whole information content from the original data. This can be viewed as an important improvement over the stick-representation model.

## 13.6 A simulation study

For illustration purpose and simplicity, the simulation study was based on the model with a normal distribution function. In the simulation study, eight settings, each with 100 simulated data sets, were considered. These settings are shown in Table 13.1.

The other parameters were chosen as follows:

**Table 13.1:** The eight settings for the simulation study.

| | set1 | set2 | set3 | set4 | set5 | set6 | set7 | set8 |
|---|---|---|---|---|---|---|---|---|
| $M_1^*$ | 2000.90 | 2000.90 | 2000.90 | 2000.90 | 2000.90 | 2000.90 | 2000.90 | 2000.90 |
| $M_2^*$ | 2000.94 | 2000.94 | 2001.94 | 2001.94 | 2004.94 | 2004.94 | 2006.94 | 2006.94 |
| $H_2/H_1$ | 0.2 | 5 | 0.2 | 5 | 0.2 | 5 | 0.2 | 5 |
| $\boldsymbol{R}$ | **E1E2** | **E1E2** | **AA** | **AA** | **AA** | **AA** | **E1E2** | **E1E2** |

$$M_1^* = 2000.90, \quad H_1 = 10000, \quad \sigma = 10, \quad \sigma_s = 0.08, \quad S = 1.0015.$$

Figure 13.5 shows the graphical representation of the eight settings. It can be seen that in settings 4, 6, 7 and 8, the separation of the two peptides is discernible. However, for the other four settings, the separation is less clear, which suggests that it may be more difficult to correctly quantify the two peptides.

Table 13.2 shows that, for all of the 100 data sets for settings 6, 7 and 8, $\pi_g^* = 1$ only for the peak that truly bears the monoisotopic mass of the second peptide and 0 elsewhere. This means the conditional posterior distribution for $M_2^*$ is a normal at the correct peak for $M_2^*$, instead of a mixture of normal distributions. However, for setting 4, even if the separation is clearly seen in the data, in 5% of the simulated data sets, the model chose the wrong peak as the monoisotopic peak of the second peptide. For the other more difficult settings, i.e., settings 1–3, and 5, the model chose different peaks as the monoisotopic peak of the second peptide, for the 100 replicated data sets. This results in the posterior distribution of $M_2^*$ to be a mixture of normal components, with probability of each component corresponding to the mean value of the probability of $\pi_g^*$ shown in Table 13.2.

Tables 13.31 and 13.32 show the summary statistics for the eight settings. For settings 6–8, the parameters were, in general, well estimated. However, even for these settings, the model-based standard errors, $\sigma_{mb}$, were slightly under-estimated as compared to the empirical ones $\sigma_{emp}$. On the other hand, for the other five settings, the point estimates, especially for the monoisotopic mass $M_2^*$ and for the isotopic ratios $R_{l_2}$ of the second peptide, are seriously biased. Moreover, the model based standard errors $\sigma_{mb}$ are severely under-estimated, as compared with the empirical ones, especially for the two parameters of interest, i.e., $M_2^*$ and $H_2/H_1$. The under-estimation of $\sigma_{mb}$ and the serious bias of the point estimates lead to 95% credible intervals incapable of incorporating the true values of the parameters.

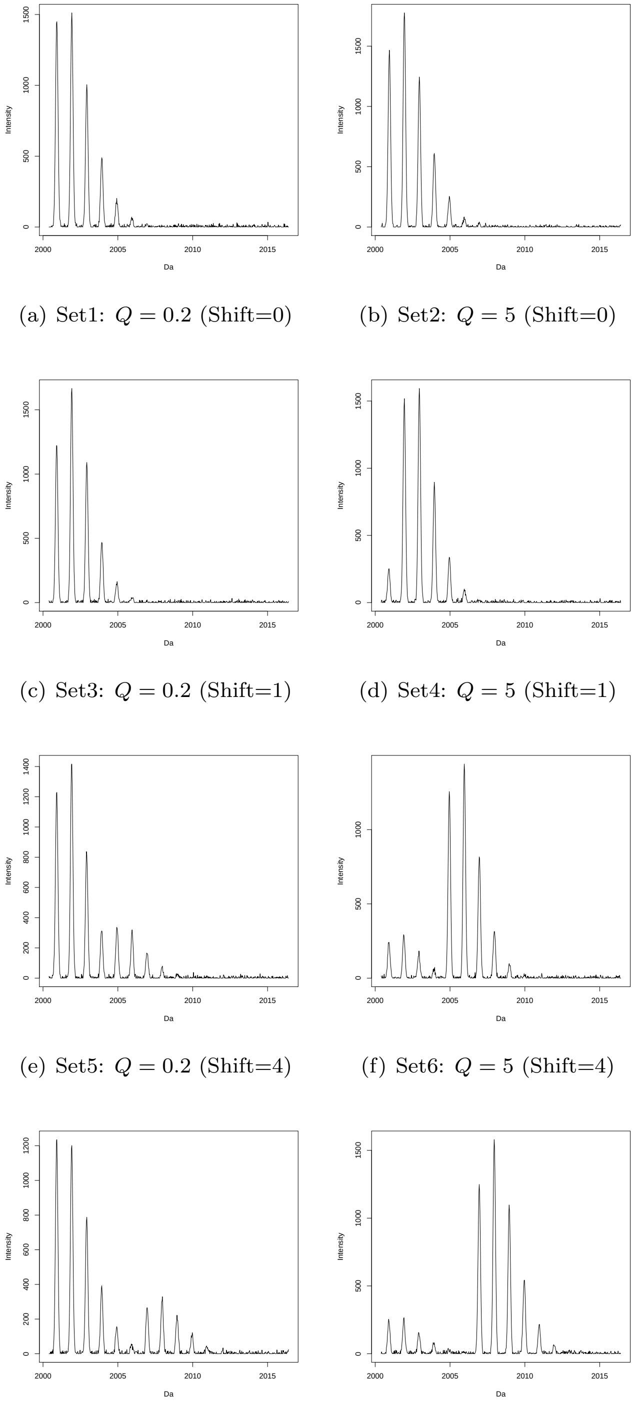
Note that, among all the parameters, the most important parameter is the monoisotopic mass of the overlapping peptide $M_2^*$, since the estimation of the remaining parameters is conditional on its estimate. This indicates that when $M_2^*$ is incorrectly estimated, different estimates for the other parameters will be obtained. To estimate $M_2^*$ correctly, it is crucial that the right component of the normal mixture is picked up.

Figure 13.6 shows the trace plots of indicator variable $\imath$, after the convergence, for the chosen normal component for the estimation of $M_2^*$ from the 100 data sets for each setting (the true values of $\imath$ are plotted as the thick grey lines). What is expected for the samples of $\imath$ is that for settings showing clear separation of the overlapping peptides (settings 4, 6, 7, and 8), $\imath$ should converge to the correct value, implying that the posterior of $M_2^*$ should be a normal, instead of a mixture of normal, at the correct peak location. For the settings where the separation is less clear, $\imath$ is allowed to vary across different values. This implies that the posterior of $M_2^*$ for these settings should be a mixture of normal distributions, incorporating the normal component, at which the true value of $M_2^*$ locates. In order for the model-based standard errors $\sigma_{mb}$ to be compatible with the empirical ones, the variability of $\imath$ within a single data set should be more or less equal to that between the 100 data sets. This is, nevertheless, not observed from Figure 13.6.

From Figure 13.6, it is clear for the settings 6-8 that the samples converged to the true value of $\imath$. This is not the case for setting 4, even though it is also a setting that shows a clear peptide separation. For the other four settings, the traces of the 100 data sets are mostly parallel to each other and $\imath$, for most of these data sets, was not correctly sampled (with the correct values of $\imath$ indicated by the thick grey lines). This means that, for most of the 100 data sets, for the four difficult settings, the posterior of $M_2^*$ is a normal (not a normal mixture) distribution that does not contain the true value of $M_2^*$. On the other hand, for a few of the data sets, the samples of $\imath$ vary across different values, indicating that the posterior for these data sets is a normal mixture. However, it can be seen from Figure 13.6b (for setting 2) that, between the sample number of 900 to 1000, the thick grey line does not overlay with samples of $\imath$ for any of the 100 data sets. This means that, for this setting, none of the posterior distributions of $M_2^*$ contain the true value.

**Table 13.2:** Mean probabilities of the mixture of normal for the mass of the second peptide $\pi_g^*$ (the ones in bold should be estimated with the largest probability).

| | mean probabilities of multinomial $\bar{\pi}_g^*$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | set1 | set2 | set3 | set4 | set5 | set6 | set7 | set8 |
| Shift=0 | **0.1400** | **0.0101** | 0.0714 | 0 | 0.0200 | 0 | 0 | 0 |
| Shift=1 | 0.5800 | 0.9192 | **0.7653** | **0.9500** | 0.0100 | 0 | 0 | 0 |
| Shift=2 | 0.0500 | 0.0303 | 0.0510 | 0.0500 | 0.0300 | 0 | 0 | 0 |
| Shift=3 | 0.0300 | 0.0303 | 0.0306 | 0 | 0.1200 | 0 | 0 | 0 |
| Shift=4 | 0.0700 | 0.0101 | 0.0204 | 0 | **0.5300** | **1** | 0 | 0 |
| Shift=5 | 0.1200 | 0.0000 | 0.0000 | 0 | 0.2700 | 0 | 0 | 0 |
| Shift=6 | 0.0100 | 0.0000 | 0.0102 | 0 | 0.0200 | 0 | **1** | **1** |
| Shift=7 | 0.0000 | 0.0000 | 0.0510 | 0 | 0.0000 | 0 | 0 | 0 |



(a) Set1: $Q = 0.2$ (Shift=0)

(b) Set2: $Q = 5$ (Shift=0)

(c) Set3: $Q = 0.2$ (Shift=1)

(d) Set4: $Q = 5$ (Shift=1)

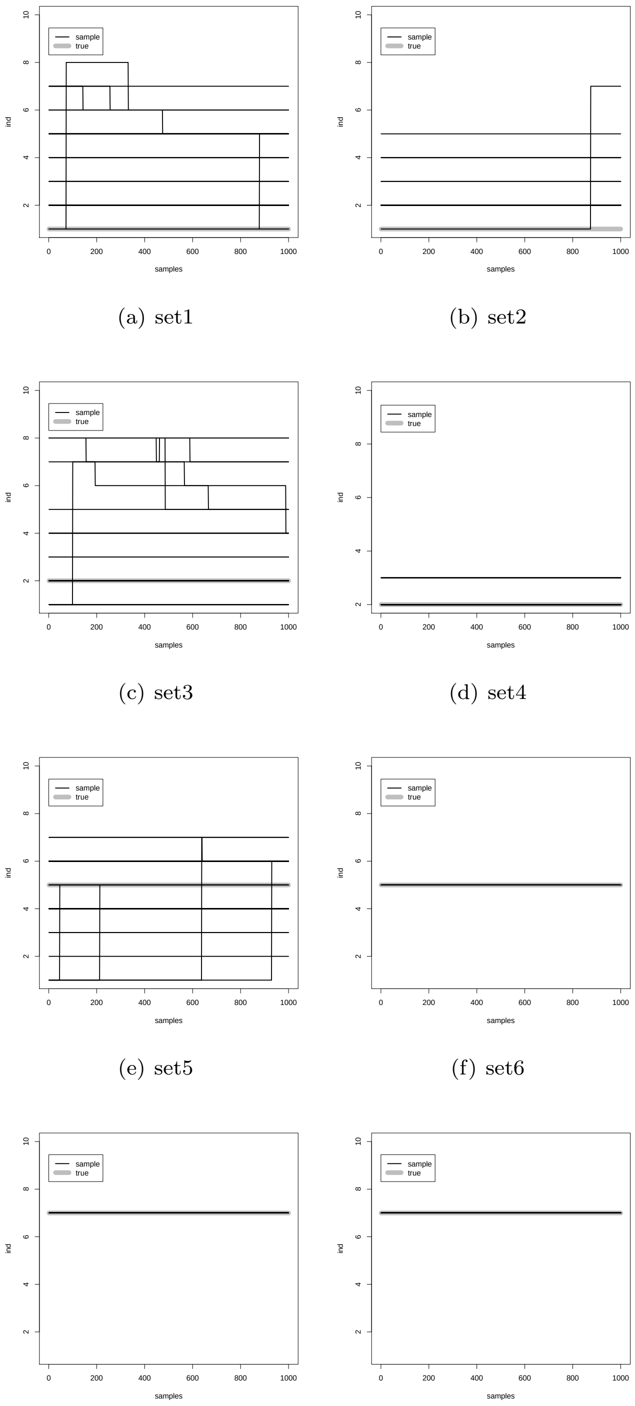
(e) Set5: $Q = 0.2$ (Shift=4)

(f) Set6: $Q = 5$ (Shift=4)

(g) Set7: $Q = 0.2$ (Shift=6)

(h) Set8: $Q = 5$ (Shift=6)

**Figure 13.5:** Graphical representation of the 8 settings of simulated data sets.



(a) set1

(b) set2

(c) set3

(d) set4

(e) set5

(f) set6

(g) set7

(h) set8

**Figure 13.6:** Samples of the indicator variable $\imath$ from different data sets as compared with the true values.

## 13.7 Discussion

In this chapter, we have presented a Bayesian model for the shape representation of a mass spectrum with overlapping peptides. We applied the model to the bovine cytochrome C data set and investigated its statistical performance via a simulation study.

The real-life data from the bovine cytochrome C data set show a clear separation of peaks for the overlapping peptides. In this case, the second peptide was correctly detected by the model and the monoisotopic mass of the peptide wa correctly estimated. The isotopic ratios, as well as the relative abundance of the two peptides, were much better estimated than in the stick-representation model. This is because the shape representation retains all the information content of the data.

The simulation study, although it showed good estimation for some of the settings with a clear separation of the peptides, produced biased results for most settings. Even for settings, for which the separation was obvious, for some of the data sets, the monoisotopic mass estimates were still biased due to the posterior distribution of the parameter converging to a wrong component of the normal mixture. For the more difficult settings, in most cases, the parameter estimates were severely biased. Moreover, for all of the settings, for the important parameters, the 95% credible intervals were very narrow and failed to contain the true parameter values. This indicates that the model under-estimates uncertainties when estimating the parameters.

The problem implies that a more elaborate method, which can, on one hand, produce correct estimates for the settings with good separation, and on the other hand, correctly estimate the model uncertainty as an indication for the difficult settings, should be implemented. In the next chapter, we will present an alternative method that can be used to this aim.

## Individual fit to the six spectra:

**Table 13.3:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with normal-density shape-function for **Spectrum 1** ($H_2/H_1 = 1/3$).

| Parameter | Data set 1 | | | | Data set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.7913 | 0.7913 | (0.774, 0.8088) | 0.8703 | 0.8618 | 0.8618 | (0.8427, 0.8814) |
| $R_{3_1}$ | 0.3567 | 0.3599 | 0.3598 | (0.3477, 0.3725) | 0.4223 | 0.4326 | 0.4326 | (0.4184, 0.4466) |
| $R_{4_1}$ | 0.1166 | 0.1189 | 0.1188 | (0.1124, 0.1255) | 0.1478 | 0.1537 | 0.1537 | (0.1458, 0.1618) |
| $R_{5_1}$ | 0.0306 | 0.0312 | 0.0312 | (0.0287, 0.0339) | 0.0413 | 0.0432 | 0.0431 | (0.0398, 0.0467) |
| $M_1^*$ | 1456.66 | 1456.679 | 1456.679 | (1456.678, 1456.68) | 1584.76 | 1584.777 | 1584.777 | (1584.776, 1584.778) |
| $R_{2_2}$ | 0.7933 | 0.7752 | 0.7750 | (0.7286, 0.8233) | 0.8703 | 0.8372 | 0.8369 | (0.7854, 0.8904) |
| $R_{3_2}$ | 0.3567 | 0.3481 | 0.3478 | (0.3227, 0.3747) | 0.4223 | 0.4049 | 0.4046 | (0.375, 0.4361) |
| $R_{4_2}$ | 0.1166 | 0.1141 | 0.114 | (0.1044, 0.1244) | 0.1478 | 0.1423 | 0.1421 | (0.1303, 0.1552) |
| $R_{5_2}$ | 0.0306 | 0.0299 | 0.0299 | (0.0268, 0.0333) | 0.0413 | 0.0399 | 0.0398 | (0.0358, 0.0443) |
| $M_2^*$ | 1460.67 | 1460.687 | 1460.687 | (1460.685, 1460.690) | 1588.77 | 1588.785 | 1588.785 | (1588.782, 1588.788) |
| $\sigma$ | – | 309.92 | 309.7 | (293.7975, 327.3) | – | 335.4336 | 335.2 | (317.5, 354.3) |
| $\sigma_s$ | – | 0.0470 | 0.0470 | (0.0463, 0.0478) | – | 0.0499 | 0.0499 | (0.0491, 0.0507) |
| $S$ | – | 1.0026 | 1.003 | (1.002, 1.004) | – | 1.0027 | 1.003 | (1.002, 1.004) |
| $H_2/H_1$ | 1/3 | 0.3086 | 0.3086 | (0.2941, 0.3230) | 1/3 | 0.3084 | 0.3084 | (0.2931, 0.3240) |

**Table 13.4:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with normal-density shape-function for **Spectrum 1** ($H_2/H_1 = 3/1$).

| Parameter | Data set 2 | | | | Data set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8213 | 0.821 | (0.7829, 0.8599) | 0.8703 | 0.9170 | 0.9168 | (0.8732, 0.9627) |
| $R_{3_1}$ | 0.3567 | 0.4104 | 0.4104 | (0.3851, 0.4365) | 0.4223 | 0.5030 | 0.5029 | (0.4722, 0.5346) |
| $R_{4_1}$ | 0.1166 | 0.1383 | 0.1383 | (0.1276, 0.1494) | 0.1478 | 0.1832 | 0.1831 | (0.1691, 0.1977) |
| $R_{5_1}$ | 0.0306 | 0.0363 | 0.0363 | (0.0328, 0.0401) | 0.0413 | 0.0514 | 0.0513 | (0.0465, 0.0567) |
| $M_1^*$ | 1456.66 | 1456.674 | 1456.674 | (1456.672, 1456.676) | 1584.76 | 1584.772 | 1584.772 | (1584.770, 1584.774) |
| $R_{2_2}$ | 0.7933 | 0.7689 | 0.7689 | (0.7504, 0.7875) | 0.8703 | 0.8554 | 0.8553 | (0.8342, 0.8771) |
| $R_{3_2}$ | 0.3567 | 0.3366 | 0.3366 | (0.3241, 0.3493) | 0.4223 | 0.4082 | 0.4082 | (0.394, 0.4231) |
| $R_{4_2}$ | 0.1166 | 0.1093 | 0.1093 | (0.1032, 0.1157) | 0.1478 | 0.1427 | 0.1426 | (0.135, 0.1506) |
| $R_{5_2}$ | 0.0306 | 0.0287 | 0.0286 | (0.0263, 0.0311) | 0.0413 | 0.0400 | 0.0400 | (0.0368, 0.0433) |
| $M_2^*$ | 1460.67 | 1460.682 | 1460.682 | (1460.681, 1460.683) | 1588.77 | 1588.78 | 1588.78 | (1588.779, 1588.781) |
| $\sigma$ | – | 255.4553 | 255.3 | (241.8, 269.5) | – | 275.1429 | 275 | (260.1, 290.9) |
| $\sigma_s$ | – | 0.0464 | 0.0464 | (0.0456, 0.0472) | – | 0.0499 | 0.0499 | (0.0490, 0.0507) |
| $S$ | – | 1.0026 | 1.003 | (1.001, 1.004) | – | 1.0026 | 1.003 | (1.001, 1.004) |
| $H_2/H_1$ | 2.4 | 2.2978 | 2.2971 | (2.2152, 2.3834) | 2.4 | 2.2667 | 2.2658 | (2.1767, 2.3609) |

**Table 13.5:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with normal-density shape-function for **Spectrum 2** ($H_2/H_1 = 1/3$).

| Parameter | Data set 1 | | | | Data set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.7875 | 0.7874 | (0.7701, 0.8049) | 0.8703 | 0.8584 | 0.8583 | (0.8395, 0.8778) |
| $R_{3_1}$ | 0.3567 | 0.3607 | 0.3606 | (0.3483, 0.3734) | 0.4223 | 0.4265 | 0.4265 | (0.4127, 0.4405) |
| $R_{4_1}$ | 0.1166 | 0.1189 | 0.1189 | (0.1127, 0.1256) | 0.1478 | 0.1513 | 0.1513 | (0.1435, 0.1593) |
| $R_{5_1}$ | 0.0306 | 0.0312 | 0.0312 | (0.0288, 0.0339) | 0.0413 | 0.0425 | 0.0424 | (0.0392, 0.0460) |
| $M_1^*$ | 1456.66 | 1456.669 | 1456.669 | (1456.668, 1456.670) | 1584.76 | 1584.766 | 1584.766 | (1584.765, 1584.767) |
| $R_{2_2}$ | 0.7933 | 0.7698 | 0.7695 | (0.7234, 0.8182) | 0.8703 | 0.8565 | 0.8559 | (0.8038, 0.9121) |
| $R_{3_2}$ | 0.3567 | 0.3451 | 0.3448 | (0.3202, 0.3723) | 0.4223 | 0.4159 | 0.4156 | (0.385, 0.4485) |
| $R_{4_2}$ | 0.1166 | 0.1132 | 0.113 | (0.1035, 0.1237) | 0.1478 | 0.1461 | 0.1459 | (0.1333, 0.1596) |
| $R_{5_2}$ | 0.0306 | 0.0297 | 0.0296 | (0.0266, 0.0331) | 0.0413 | 0.0410 | 0.0409 | (0.0367, 0.0456) |
| $M_2^*$ | 1460.67 | 1460.677 | 1460.677 | (1460.675, 1460.679) | 1588.77 | 1588.774 | 1588.774 | (1588.772, 1588.777) |
| $\sigma$ | – | 253.3394 | 253.2 | (240.1, 267.4) | – | 267.5314 | 267.4 | (253.4, 282.8) |
| $\sigma_s$ | – | 0.0435 | 0.0435 | (0.0428, 0.0443) | – | 0.0463 | 0.0463 | (0.0455, 0.0471) |
| $S$ | – | 1.0026 | 1.003 | (1.002, 1.003) | – | 1.0028 | 1.003 | (1.002, 1.004) |
| $H_2/H_1$ | 1/3 | 0.3074 | 0.3073 | (0.2932, 0.3218) | 1/3 | 0.3022 | 0.3022 | (0.2868, 0.3176) |

**Table 13.6:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with normal-density shape-function for **Spectrum 2** ($H_2/H_1 = 3/1$).

| Parameter | Data set 2 | | | | Data set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8301 | 0.83 | (0.7909, 0.8703) | 0.8703 | 0.9244 | 0.9239 | (0.8792, 0.9706) |
| $R_{3_1}$ | 0.3567 | 0.4132 | 0.413 | (0.3873, 0.4395) | 0.4223 | 0.5027 | 0.5025 | (0.4709, 0.5351) |
| $R_{4_1}$ | 0.1166 | 0.1391 | 0.139 | (0.1283, 0.1502) | 0.1478 | 0.1827 | 0.1826 | (0.1684, 0.1974) |
| $R_{5_1}$ | 0.0306 | 0.0365 | 0.0365 | (0.0330, 0.0404) | 0.0413 | 0.0512 | 0.0512 | (0.0463, 0.0565) |
| $M_1^*$ | 1456.66 | 1456.682 | 1456.682 | (1456.680, 1456.684) | 1584.76 | 1584.780 | 1584.780 | (1584.778, 1584.782) |
| $R_{2_2}$ | 0.7933 | 0.7797 | 0.7797 | (0.7611, 0.7988) | 0.8703 | 0.8577 | 0.8576 | (0.8358, 0.88) |
| $R_{3_2}$ | 0.3567 | 0.3433 | 0.3432 | (0.3303, 0.3563) | 0.4223 | 0.4111 | 0.411 | (0.3961, 0.4269) |
| $R_{4_2}$ | 0.1166 | 0.1117 | 0.1116 | (0.1056, 0.1182) | 0.1478 | 0.1436 | 0.1435 | (0.1359, 0.1518) |
| $R_{5_2}$ | 0.0306 | 0.0293 | 0.0293 | (0.0270, 0.0318) | 0.0413 | 0.0402 | 0.0402 | (0.0370, 0.0436) |
| $M_2^*$ | 1460.67 | 1460.689 | 1460.689 | (1460.688, 1460.690) | 1588.77 | 1588.788 | 1588.788 | (1588.786, 1588.789) |
| $\sigma$ | – | 296.4185 | 296.2 | (280.8, 313.3) | – | 326.6484 | 326.4 | (309, 345.8) |
| $\sigma_s$ | – | 0.0496 | 0.0496 | (0.0488, 0.0505) | – | 0.0533 | 0.0533 | (0.0524, 0.0543) |
| $S$ | – | 1.0022 | 1.002 | (1.001, 1.003) | – | 1.0024 | 1.002 | (1.001, 1.004) |
| $H_2/H_1$ | 2.4 | 2.2822 | 2.2820 | (2.1993, 2.3694) | 2.4 | 2.2715 | 2.2709 | (2.1795, 2.3676) |

**Table 13.7:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with normal-density shape-function for **Spectrum 3** ($H_2/H_1 = 1/3$).

| Parameter | Data set 1 | | | | Data set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.7849 | 0.7847 | (0.7681, 0.8022) | 0.8703 | 0.8601 | 0.8602 | (0.8405, 0.8793) |
| $R_{3_1}$ | 0.3567 | 0.3566 | 0.3566 | (0.3443, 0.369) | 0.4223 | 0.4264 | 0.4264 | (0.4127, 0.4406) |
| $R_{4_1}$ | 0.1166 | 0.1178 | 0.1178 | (0.1116, 0.1241) | 0.1478 | 0.1519 | 0.1518 | (0.1442, 0.1599) |
| $R_{5_1}$ | 0.0306 | 0.0310 | 0.0309 | (0.0285, 0.0336) | 0.0413 | 0.0427 | 0.0426 | (0.0394, 0.0462) |
| $M_1^*$ | 1456.66 | 1456.674 | 1456.674 | (1456.673, 1456.675) | 1584.76 | 1584.772 | 1584.772 | (1584.771, 1584.773) |
| $R_{2_2}$ | 0.7933 | 0.7774 | 0.7773 | (0.7305, 0.8257) | 0.8703 | 0.8437 | 0.8434 | (0.7936, 0.8972) |
| $R_{3_2}$ | 0.3567 | 0.3492 | 0.3491 | (0.3235, 0.3759) | 0.4223 | 0.4086 | 0.4082 | (0.3793, 0.4401) |
| $R_{4_2}$ | 0.1166 | 0.1145 | 0.1145 | (0.1047, 0.1249) | 0.1478 | 0.1435 | 0.1432 | (0.1313, 0.1565) |
| $R_{5_2}$ | 0.0306 | 0.0300 | 0.0300 | (0.0269, 0.0334) | 0.0413 | 0.0402 | 0.0402 | (0.0361, 0.0447) |
| $M_2^*$ | 1460.67 | 1460.682 | 1460.682 | (1460.680, 1460.684) | 1588.77 | 1588.780 | 1588.78 | (1588.777, 1588.782) |
| $\sigma$ | – | 323.908 | 323.7 | (306.8, 342) | – | 330.9926 | 330.8 | (313.2, 350.1) |
| $\sigma_s$ | – | 0.0442 | 0.0442 | (0.0435, 0.0449) | – | 0.0470 | 0.0470 | (0.0462, 0.0477) |
| $S$ | – | 1.0027 | 1.003 | (1.002, 1.004) | – | 1.0025 | 1.002 | (1.002, 1.003) |
| $H_2/H_1$ | 1/3 | 0.3052 | 0.3052 | (0.2909, 0.3195) | 1/3 | 0.3045 | 0.3044 | (0.2896, 0.3196) |

**Table 13.8:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with normal-density shape-function for **Spectrum 3** ($H_2/H_1 = 3/1$).

| Parameter | Data set 2 | | | | Data set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8320 | 0.8318 | (0.7932, 0.8718) | 0.8703 | 0.8249 | 0.8247 | (0.8011, 0.8511) |
| $R_{3_1}$ | 0.3567 | 0.4139 | 0.4138 | (0.3888, 0.4401) | 0.4223 | 0.4519 | 0.4518 | (0.4389, 0.4662) |
| $R_{4_1}$ | 0.1166 | 0.1393 | 0.1392 | (0.1288, 0.1503) | 0.1478 | 0.1748 | 0.1747 | (0.1697, 0.1803) |
| $R_{5_1}$ | 0.0306 | 0.0366 | 0.0365 | (0.0330, 0.0404) | 0.0413 | 0.0446 | 0.0446 | (0.0431, 0.0464) |
| $M_1^*$ | 1456.66 | 1456.664 | 1456.664 | (1456.663, 1456.666) | 1584.76 | 1584.7612 | 1584.7612 | (1584.7612, 1584.7613) |
| $R_{2_2}$ | 0.7933 | 0.7732 | 0.7732 | (0.7552, 0.7913) | 0.8703 | 0.8259 | 0.8259 | (0.8107, 0.8395) |
| $R_{3_2}$ | 0.3567 | 0.3355 | 0.3355 | (0.323, 0.348) | 0.4223 | 0.3960 | 0.3960 | (0.3884, 0.4031) |
| $R_{4_2}$ | 0.1166 | 0.1087 | 0.1087 | (0.1029, 0.1151) | 0.1478 | 0.1310 | 0.1310 | (0.1281, 0.1341) |
| $R_{5_2}$ | 0.0306 | 0.0285 | 0.0285 | (0.0262, 0.0309) | 0.0413 | 0.0350 | 0.0350 | (0.0342, 0.0361) |
| $M_2^*$ | 1460.67 | 1460.672 | 1460.672 | (1460.671, 1460.673) | 1588.77 | 1588.7689 | 1588.7689 | (1588.7689, 1588.7690) |
| $\sigma$ | – | 276.8692 | 276.7 | (262.1, 292.4) | – | 311.9810 | 311.8920 | (294.2483, 330.9076) |
| $\sigma_s$ | – | 0.0447 | 0.0447 | (0.044, 0.0455) | – | 0.0486 | 0.0485 | (0.0484, 0.0491) |
| $S$ | – | 1.0028 | 1.003 | (1.002, 1.004) | – | 1.0025 | 1.0025 | (1.0025, 1.0025) |
| $H_2/H_1$ | 2.4 | 2.3261 | 2.3251 | (2.2433, 2.4116) | 2.4 | 2.1173 | 2.1171 | (2.0503, 2.1852) |

**Table 13.9:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with normal-density shape-function for **Spectrum 4** ($H_2/H_1 = 1/3$).

| Parameter | Data set 1 | | | | Data set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.7895 | 0.7896 | (0.7724, 0.8065) | 0.8703 | 0.8611 | 0.861 | (0.8418, 0.8806) |
| $R_{3_1}$ | 0.3567 | 0.3629 | 0.3629 | (0.3507, 0.3754) | 0.4223 | 0.4299 | 0.4299 | (0.4156, 0.444) |
| $R_{4_1}$ | 0.1166 | 0.1198 | 0.1198 | (0.1134, 0.1265) | 0.1478 | 0.1522 | 0.1522 | (0.1443, 0.1602) |
| $R_{5_1}$ | 0.0306 | 0.0315 | 0.0315 | (0.0290, 0.0341) | 0.0413 | 0.0427 | 0.0427 | (0.0394, 0.0463) |
| $M_1^*$ | 1456.66 | 1456.676 | 1456.676 | (1456.675, 1456.676) | 1584.76 | 1584.773 | 1584.773 | (1584.772, 1584.774) |
| $R_{2_2}$ | 0.7933 | 0.7665 | 0.7663 | (0.7207, 0.8139) | 0.8703 | 0.8583 | 0.8576 | (0.8071, 0.9127) |
| $R_{3_2}$ | 0.3567 | 0.3446 | 0.3445 | (0.3194, 0.3708) | 0.4223 | 0.4142 | 0.4139 | (0.3843, 0.4454) |
| $R_{4_2}$ | 0.1166 | 0.1129 | 0.1128 | (0.1033, 0.1232) | 0.1478 | 0.1455 | 0.1454 | (0.1331, 0.1586) |
| $R_{5_2}$ | 0.0306 | 0.0296 | 0.0296 | (0.0266, 0.0330) | 0.0413 | 0.0408 | 0.0407 | (0.0366, 0.0452) |
| $M_2^*$ | 1460.67 | 1460.684 | 1460.684 | (1460.682, 1460.686) | 1588.77 | 1588.781 | 1588.781 | (1588.778, 1588.784) |
| $\sigma$ | – | 287.6014 | 287.4 | (272.9, 303.5) | – | 299.6430 | 299.5 | (283.3, 316.7025) |
| $\sigma_s$ | – | 0.0430 | 0.0430 | (0.0424, 0.0437) | – | 0.0456 | 0.0457 | (0.0449, 0.0464) |
| $S$ | – | 1.0024 | 1.002 | (1.001, 1.003) | – | 1.0026 | 1.002 | (1.002, 1.004) |
| $H_2/H_1$ | 1/3 | 0.3103 | 0.3103 | (0.2962, 0.3245) | 1/3 | 0.3063 | 0.3063 | (0.2911, 0.3218) |

**Table 13.10:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with normal-density shape-function for **Spectrum 4** ($H_2/H_1 = 3/1$).

| Parameter | Data set 2 | | | | Data set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8256 | 0.8254 | (0.7868, 0.8653) | 0.8703 | 0.9185 | 0.918 | (0.8748, 0.9642) |
| $R_{3_1}$ | 0.3567 | 0.4099 | 0.4098 | (0.3847, 0.4356) | 0.4223 | 0.5026 | 0.5025 | (0.4716, 0.5345) |
| $R_{4_1}$ | 0.1166 | 0.1384 | 0.1383 | (0.128, 0.1496) | 0.1478 | 0.1832 | 0.1831 | (0.1693, 0.1981) |
| $R_{5_1}$ | 0.0306 | 0.0363 | 0.0363 | (0.0329, 0.0400) | 0.0413 | 0.0514 | 0.0513 | (0.0465, 0.0568) |
| $M_1^*$ | 1456.66 | 1456.667 | 1456.668 | (1456.666, 1456.669) | 1584.76 | 1584.764 | 1584.764 | (1584.762, 1584.766) |
| $R_{2_2}$ | 0.7933 | 0.7736 | 0.7735 | (0.7555, 0.7927) | 0.8703 | 0.8414 | 0.8413 | (0.8204, 0.8625) |
| $R_{3_2}$ | 0.3567 | 0.3365 | 0.3364 | (0.324, 0.3494) | 0.4223 | 0.4043 | 0.4042 | (0.3898, 0.4194) |
| $R_{4_2}$ | 0.1166 | 0.1090 | 0.1089 | (0.1028, 0.1153) | 0.1478 | 0.1409 | 0.1409 | (0.1333, 0.1488) |
| $R_{5_2}$ | 0.0306 | 0.0285 | 0.0285 | (0.0262, 0.0310) | 0.0413 | 0.394 | 0.0394 | (0.0363, 0.0428) |
| $M_2^*$ | 1460.67 | 1460.675 | 1460.675 | (1460.674, 1460.676) | 1588.77 | 1588.772 | 1588.772 | (1588.771, 1588.773) |
| $\sigma$ | – | 284.5025 | 284.3 | (269.5, 300.6) | – | 300.0844 | 299.9 | (284.2, 317.2) |
| $\sigma_s$ | – | 0.0447 | 0.0447 | (0.0440, 0.0455) | – | 0.0472 | 0.0472 | (0.0464, 0.0481) |
| $S$ | – | 1.0028 | 1.003 | (1.002, 1.004) | – | 1.0026 | 1.003 | (1.001, 1.004) |
| $H_2/H_1$ | 2.4 | 2.3083 | 2.3081 | (2.2246, 2.3952) | 2.4 | 2.3122 | 2.3116 | (2.2210, 2.4075) |

**Table 13.11:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with normal-density shape-function for **Spectrum 5** ($H_2/H_1 = 1/3$).

| Parameter | Data set 1 | | | | Data set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.7897 | 0.7896 | (0.7727, 0.8073) | 0.8703 | 0.8747 | 0.8747 | (0.8554, 0.8943) |
| $R_{3_1}$ | 0.3567 | 0.3623 | 0.3623 | (0.3502, 0.3749) | 0.4223 | 0.4310 | 0.4309 | (0.4171, 0.4449) |
| $R_{4_1}$ | 0.1166 | 0.1196 | 0.1195 | (0.1132, 0.1261) | 0.1478 | 0.1530 | 0.153 | (0.1453, 0.1609) |
| $R_{5_1}$ | 0.0306 | 0.0314 | 0.0314 | (0.0289, 0.0340) | 0.0413 | 0.0430 | 0.0430 | (0.0397, 0.0466) |
| $M_1^*$ | 1456.66 | 1456.671 | 1456.670 | (1456.670, 1456.671) | 1584.76 | 1584.768 | 1584.768 | (1584.767, 1584.769) |
| $R_{2_2}$ | 0.7933 | 0.7795 | 0.7794 | (0.7329, 0.8275) | 0.8703 | 0.8358 | 0.8352 | (0.7851, 0.889) |
| $R_{3_2}$ | 0.3567 | 0.3501 | 0.3499 | (0.3245, 0.3769) | 0.4223 | 0.4053 | 0.4051 | (0.3757, 0.4366) |
| $R_{4_2}$ | 0.1166 | 0.1148 | 0.1147 | (0.105, 0.1252) | 0.1478 | 0.1424 | 0.1423 | (0.1304, 0.1553) |
| $R_{5_2}$ | 0.0306 | 0.0301 | 0.0301 | (0.027, 0.0336) | 0.0413 | 0.0399 | 0.0399 | (0.0358, 0.0444) |
| $M_2^*$ | 1460.67 | 1460.678 | 1460.678 | (1460.676, 1460.681) | 1588.77 | 1588.776 | 1588.776 | (1588.773, 1588.778) |
| $\sigma$ | – | 325.5337 | 325.3 | (308.9, 343.5) | – | 348.0792 | 347.9 | (329.7, 367.9) |
| $\sigma_s$ | – | 0.0462 | 0.0462 | (0.0455, 0.0470) | – | 0.0490 | 0.0490 | (0.0482, 0.0498) |
| $S$ | – | 1.0027 | 1.003 | (1.002, 1.004) | – | 1.0026 | 1.003 | (1.002, 1.004) |
| $H_2/H_1$ | 1/3 | 0.3035 | 0.3036 | (0.2895, 0.3176) | 1/3 | 0.3099 | 0.3098 | (0.2949, 0.3252) |

**Table 13.12:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with normal-density shape-function for **Spectrum 5** ($H_2/H_1 = 3/1$).

| Parameter | Data set 2 | | | | Data set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8189 | 0.8188 | (0.7813, 0.8581) | 0.8703 | 0.8197 | 0.8203 | (0.7902, 0.8496) |
| $R_{3_1}$ | 0.3567 | 0.4083 | 0.4081 | (0.3835, 0.4341) | 0.4223 | 0.4481 | 0.4475 | (0.4334, 0.4628) |
| $R_{4_1}$ | 0.1166 | 0.1376 | 0.1375 | (0.1271, 0.1488) | 0.1478 | 0.1735 | 0.1733 | (0.1679, 0.1792) |
| $R_{5_1}$ | 0.0306 | 0.0361 | 0.0361 | (0.0326, 0.0400) | 0.0413 | 0.0439 | 0.0438 | (0.0418, 0.0466) |
| $M_1^*$ | 1456.66 | 1456.663 | 1456.664 | (1456.662, 1456.665) | 1584.76 | 1584.7607 | 1584.7607 | (1584.7507, 1584.7608) |
| $R_{2_2}$ | 0.7933 | 0.7617 | 0.7617 | (0.7438, 0.7801) | 0.8703 | 0.8207 | 0.8212 | (0.8012, 0.8330) |
| $R_{3_2}$ | 0.3567 | 0.3361 | 0.336 | (0.3237, 0.3488) | 0.4223 | 0.3948 | 0.3948 | (0.3880, 0.4027) |
| $R_{4_2}$ | 0.1166 | 0.1089 | 0.1089 | (0.1028, 0.1153) | 0.1478 | 0.1308 | 0.1308 | (0.1279, 0.1345) |
| $R_{5_2}$ | 0.0306 | 0.0285 | 0.0285 | (0.0263, 0.0310) | 0.0413 | 0.0346 | 0.0346 | (0.0338, 0.0358) |
| $M_2^*$ | 1460.67 | 1460.671 | 1460.671 | (1460.670, 1460.672) | 1588.77 | 1588.7684 | 1588.7684 | (1588.7684, 1588.7684) |
| $\sigma$ | – | 324.0355 | 323.8 | (307.2, 342.4) | – | 366.5905 | 366.5007 | (345.9104, 388.7287) |
| $\sigma_s$ | – | 0.0438 | 0.0438 | (0.0430, 0.0445) | – | 0.0478 | 0.0478 | (0.0472, 0.0482) |
| $S$ | – | 1.0026 | 1.003 | (1.002, 1.004) | – | 1.0026 | 1.0026 | (1.0026, 1.0026) |
| $H_2/H_1$ | 2.4 | 2.2996 | 2.2989 | (2.2188, 2.3851) | 2.4 | 2.1100 | 2.1099 | (2.0406, 2.1795) |

**Table 13.13:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with normal-density shape-function for **Spectrum 6** ($H_2/H_1 = 1/3$).

| Parameter | Data set 1 | | | | Data set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8005 | 0.8004 | (0.7829, 0.8182) | 0.8703 | 0.8646 | 0.8645 | (0.845, 0.8846) |
| $R_{3_1}$ | 0.3567 | 0.3665 | 0.3664 | (0.3537, 0.3795) | 0.4223 | 0.4326 | 0.4325 | (0.4182, 0.447) |
| $R_{4_1}$ | 0.1166 | 0.1214 | 0.1213 | (0.1148, 0.128) | 0.1478 | 0.1540 | 0.154 | (0.1462, 0.1620) |
| $R_{5_1}$ | 0.0306 | 0.0319 | 0.0319 | (0.0294, 0.0346) | 0.0413 | 0.0432 | 0.0432 | (0.0399, 0.0468) |
| $M_1^*$ | 1456.66 | 1456.672 | 1456.672 | (1456.671, 1456.672) | 1584.76 | 1584.769 | 1584.769 | (1584.768, 1584.770) |
| $R_{2_2}$ | 0.7933 | 0.7727 | 0.7725 | (0.7265, 0.8193) | 0.8703 | 0.8565 | 0.8561 | (0.8045, 0.9112) |
| $R_{3_2}$ | 0.3567 | 0.3463 | 0.346 | (0.3212, 0.3724) | 0.4223 | 0.4147 | 0.4145 | (0.3845, 0.4466) |
| $R_{4_2}$ | 0.1166 | 0.1136 | 0.1135 | (0.1039, 0.1237) | 0.1478 | 0.1457 | 0.1455 | (0.1333, 0.1591) |
| $R_{5_2}$ | 0.0306 | 0.0298 | 0.0298 | (0.0267, 0.0331) | 0.0413 | 0.0408 | 0.0408 | (0.0366, 0.0455) |
| $M_2^*$ | 1460.67 | 1460.679 | 1460.679 | (1460.677, 1460.681) | 1588.77 | 1588.776 | 1588.776 | (1588.774, 1588.779) |
| $\sigma$ | – | 330.4222 | 330.2 | (313.2, 348.9025) | – | 353.0758 | 352.9 | (334.2, 373.1) |
| $\sigma_s$ | – | 0.0449 | 0.0449 | (0.0442, 0.0457) | – | 0.0476 | 0.0476 | (0.0468, 0.0484) |
| $S$ | – | 1.0024 | 1.002 | (1.001, 1.003) | – | 1.0024 | 1.002 | (1.001, 1.003) |
| $H_2/H_1$ | 1/3 | 0.3158 | 0.3158 | (0.3013, 0.3303) | 1/3 | 0.3078 | 0.3078 | (0.2923, 0.3234) |

**Table 13.14:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with normal-density shape-function for **Spectrum 6** ($H_2/H_1 = 3/1$).

| Parameter | Data set 2 | | | | Data set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8240 | 0.8237 | (0.7867, 0.8634) | 0.8703 | 0.9175 | 0.9173 | (0.8074, 0.962) |
| $R_{3_1}$ | 0.3567 | 0.4120 | 0.4117 | (0.3869, 0.4381) | 0.4223 | 0.5058 | 0.5055 | (0.4756, 0.5375) |
| $R_{4_1}$ | 0.1166 | 0.1391 | 0.139 | (0.1285, 0.1503) | 0.1478 | 0.1847 | 0.1846 | (0.1708, 0.1993) |
| $R_{5_1}$ | 0.0306 | 0.0365 | 0.0365 | (0.0330, 0.0403) | 0.0413 | 0.0518 | 0.0517 | (0.0469, 0.0572) |
| $M_1^*$ | 1456.66 | 1456.67 | 1456.67 | (1456.668, 1456.672) | 1584.76 | 1584.768 | 1584.768 | (1584.766, 1584.770) |
| $R_{2_2}$ | 0.7933 | 0.7757 | 0.7756 | (0.758, 0.7943) | 0.8703 | 0.8414 | 0.8414 | (0.8212, 0.8622) |
| $R_{3_2}$ | 0.3567 | 0.3412 | 0.3411 | (0.3286, 0.354) | 0.4223 | 0.4055 | 0.4054 | (0.3907, 0.4201) |
| $R_{4_2}$ | 0.1166 | 0.1107 | 0.1106 | (0.1046, 0.117) | 0.1478 | 0.1416 | 0.1416 | (0.1341, 0.1496) |
| $R_{5_2}$ | 0.0306 | 0.0290 | 0.0290 | (0.0267, 0.0314) | 0.0413 | 0.0396 | 0.0396 | (0.0364, 0.0429) |
| $M_2^*$ | 1460.67 | 1460.677 | 1460.677 | (1460.676, 1460.678) | 1588.77 | 1588.775 | 1588.775 | (1588.774, 1588.777) |
| $\sigma$ | – | 317.6234 | 317.4 | (300.9975, 335.5) | – | 350.4413 | 350.2 | (331.7, 370.6) |
| $\sigma_s$ | – | 0.0473 | 0.0473 | (0.0465, 0.0481) | – | 0.0505 | 0.0505 | (0.0497, 0.0514) |
| $S$ | – | 1.0025 | 1.002 | (1.001, 1.004) | – | 1.0024 | 1.002 | (1.001, 1.004) |
| $H_2/H_1$ | 2.4 | 2.2511 | 2.2505 | (2.1709, 2.3349) | 2.4 | 2.3075 | 2.3069 | (2.2181, 2.4018) |

**Table 13.15:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with asymmetric Laplace shape-function for **Spectrum 1** ($H_2/H_1 = 1/3$).

| Parameter | Data set 1 | | | | Data set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.7907 | 0.7906 | (0.7754, 0.8066) | 0.8703 | 0.8619 | 0.8619 | (0.8456, 0.8785) |
| $R_{3_1}$ | 0.3567 | 0.3595 | 0.3595 | (0.3481, 0.3711) | 0.4223 | 0.4337 | 0.4337 | (0.4215, 0.446) |
| $R_{4_1}$ | 0.1166 | 0.1190 | 0.119 | (0.1128, 0.1253) | 0.1478 | 0.1545 | 0.1545 | (0.1473, 0.1618) |
| $R_{5_1}$ | 0.0306 | 0.0313 | 0.0313 | (0.0288, 0.0339) | 0.0413 | 0.0434 | 0.0434 | (0.0401, 0.0469) |
| $M_1^*$ | 1456.66 | 1456.674 | 1456.674 | (1456.673, 1456.675) | 1584.76 | 1584.764 | 1584.764 | (1584.763, 1584.765) |
| $R_{2_2}$ | 0.7933 | 0.7722 | 0.7718 | (0.7299, 0.8165) | 0.8703 | 0.8344 | 0.834 | (0.7902, 0.8803) |
| $R_{3_2}$ | 0.3567 | 0.3462 | 0.3462 | (0.3222, 0.3708) | 0.4223 | 0.4023 | 0.402 | (0.3756, 0.4299) |
| $R_{4_2}$ | 0.1166 | 0.1135 | 0.1135 | (0.1043, 0.1233) | 0.1478 | 0.1413 | 0.1411 | (0.1301, 0.1532) |
| $R_{5_2}$ | 0.0306 | 0.0298 | 0.0298 | (0.0269, 0.0330) | 0.0413 | 0.0396 | 0.0396 | (0.0357, 0.0439) |
| $M_2^*$ | 1460.67 | 1460.682 | 1460.682 | (1456.679, 1460.684) | 1588.77 | 1588.773 | 1588.773 | (1588.771, 1588.775) |
| $\sigma$ | – | 278.3557 | 278.2 | (263.5, 294.1) | – | 272.6676 | 272.5 | (258, 288.4) |
| $\sigma_s$ | – | 0.0742 | 0.0742 | (0.0728, 0.0756) | – | 0.0767 | 0.0767 | (0.0753, 0.0782) |
| $\kappa$ | – | 0.8660 | 0.866 | (0.8454, 0.8869) | – | 0.7593 | 0.7591 | (0.7426, 0.7767) |
| $S$ | – | 1.0022 | 1.002 | (1.001, 1.003) | – | 1.0028 | 1.003 | (1.002, 1.004) |
| $H_2/H_1$ | 1/3 | 0.3090 | 0.3090 | (0.2963, 0.3221) | 1/3 | 0.3093 | 0.3093 | (0.2962, 0.3223) |

**Table 13.16:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with asymmetric Laplace shape-function for **Spectrum 1** ($H_2/H_1 = 3/1$).

| Parameter | Data set 2 | | | | Data set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8209 | 0.8208 | (0.7848, 0.8578) | 0.8703 | 0.9127 | 0.9128 | (0.8755, 0.9512) |
| $R_{3_1}$ | 0.3567 | 0.4154 | 0.4153 | (0.3906, 0.4406) | 0.4223 | 0.5192 | 0.5191 | (0.4918, 0.5472) |
| $R_{4_1}$ | 0.1166 | 0.1407 | 0.1407 | (0.1301, 0.1516) | 0.1478 | 0.1923 | 0.1922 | (0.1793, 0.2061) |
| $R_{5_1}$ | 0.0306 | 0.0370 | 0.0369 | (0.0334, 0.0407) | 0.0413 | 0.0539 | 0.0539 | (0.0491, 0.0592) |
| $M_1^*$ | 1456.66 | 1456.668 | 1456.668 | (1456.666, 1456.671) | 1584.76 | 1584.762 | 1584.762 | (1584.760, 1584.763) |
| $R_{2_2}$ | 0.7933 | 0.7685 | 0.7686 | (0.7515, 0.7854) | 0.8703 | 0.8566 | 0.8565 | (0.8393, 0.8741) |
| $R_{3_2}$ | 0.3567 | 0.3363 | 0.3363 | (0.3247, 0.3484) | 0.4223 | 0.4071 | 0.4071 | (0.3948, 0.4198) |
| $R_{4_2}$ | 0.1166 | 0.1092 | 0.1091 | (0.1035, 0.1151) | 0.1478 | 0.1419 | 0.1419 | (0.135, 0.1491) |
| $R_{5_2}$ | 0.0306 | 0.0286 | 0.0286 | (0.0264, 0.031) | 0.0413 | 0.0398 | 0.0397 | (0.0367, 0.0430) |
| $M_2^*$ | 1460.67 | 1460.676 | 1460.676 | (1460.675, 1460.678) | 1588.77 | 1588.771 | 1588.771 | (1588.769, 1588.772) |
| $\sigma$ | – | 236.7468 | 236.6 | (224, 250.5) | – | 223.3746 | 223.3 | (211, 236.8025) |
| $\sigma_s$ | – | 0.0734 | 0.0734 | (0.0719, 0.0748) | – | 0.0770 | 0.0770 | (0.0756, 0.0784) |
| $\kappa$ | – | 0.8622 | 0.8622 | (0.8378, 0.8871) | – | 0.8108 | 0.8109 | (0.7948, 0.8268) |
| $S$ | – | 1.0018 | 1.002 | (1.001, 1.003) | – | 1.0029 | 1.003 | (1.002, 1.004) |
| $H_2/H_1$ | 2.4 | 2.3019 | 2.3012 | (2.2252, 2.3836) | 2.4 | 2.2714 | 2.2707 | (2.1987, 2.3475) |

**Table 13.17:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with asymmetric Laplace shape-function for **Spectrum 2** ($H_2/H_1 = 1/3$).

| Parameter | Data set 1 | | | | Data set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.7896 | 0.7896 | (0.7738, 0.8053) | 0.8703 | 0.8584 | 0.8583 | (0.8431, 0.874) |
| $R_{3_1}$ | 0.3467 | 0.3622 | 0.3622 | (0.3509, 0.3735) | 0.4223 | 0.4274 | 0.4274 | (0.4157, 0.4393) |
| $R_{4_1}$ | 0.1166 | 0.1194 | 0.1194 | (0.1133, 0.1255) | 0.1478 | 0.1520 | 0.152 | (0.1449, 0.1591) |
| $R_{5_1}$ | 0.0306 | 0.0314 | 0.0314 | (0.0289, 0.0340) | 0.0413 | 0.0427 | 0.0427 | (0.0395, 0.0460) |
| $M_1^*$ | 1456.66 | 1456.658 | 1456.658 | (1456.657, 1456.659) | 1584.76 | 1584.758 | 1584.758 | (1584.757, 1584.759) |
| $R_{2_2}$ | 0.7933 | 0.7710 | 0.7706 | (0.7288, 0.8155) | 0.8703 | 0.8547 | 0.8544 | (0.8108, 0.9007) |
| $R_{3_2}$ | 0.3567 | 0.3448 | 0.3446 | (0.3207, 0.3699) | 0.4223 | 0.4137 | 0.4135 | (0.3865, 0.4417) |
| $R_{4_2}$ | 0.1166 | 0.1131 | 0.1129 | (0.1038, 0.123) | 0.1478 | 0.1452 | 0.1451 | (0.1338, 0.1575) |
| $R_{5_2}$ | 0.0306 | 0.0297 | 0.0296 | (0.0267, 0.0330) | 0.0413 | 0.0407 | 0.0407 | (0.0367, 0.0451) |
| $M_2^*$ | 1460.67 | 1460.666 | 1460.666 | (1456.664, 1460.668) | 1588.77 | 1588.767 | 1588.767 | (1588.765, 1588.769) |
| $\sigma$ | – | 223.0668 | 222.9 | (211.2, 235.9) | – | 213.4271 | 213.3 | (201.9, 225.8) |
| $\sigma_s$ | – | 0.0674 | 0.0674 | (0.0661, 0.0687) | – | 0.0714 | 0.0714 | (0.0702, 0.0726) |
| $\kappa$ | – | 0.7853 | 0.7848 | (0.7674, 0.8053) | – | 0.8273 | 0.8273 | (0.8113, 0.8427) |
| $S$ | – | 1.0035 | 1.004 | (1.003, 1.004) | – | 1.0030 | 1.003 | (1.002, 1.004) |
| $H_2/H_1$ | 1/3 | 0.3071 | 0.3071 | (0.2946, 0.3199) | 1/3 | 0.3019 | 0.3020 | (0.2896, 0.3144) |

**Table 13.18:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with asymmetric Laplace shape-function for **Spectrum 2** ($H_2/H_1 = 3/1$).

| Parameter | Data set 2 | | | | Data set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8288 | 0.8285 | (0.7944, 0.8644) | 0.8703 | 0.9209 | 0.9204 | (0.8844, 0.9599) |
| $R_{3_1}$ | 0.3567 | 0.4215 | 0.4214 | (0.3977, 0.4456) | 0.4223 | 0.5211 | 0.5207 | (0.4936, 0.5494) |
| $R_{4_1}$ | 0.1166 | 0.1432 | 0.1431 | (0.1329, 0.1542) | 0.1478 | 0.1932 | 0.1931 | (0.1797, 0.2072) |
| $R_{5_1}$ | 0.0306 | 0.0376 | 0.0375 | (0.0341, 0.0414) | 0.0413 | 0.0542 | 0.0542 | (0.0491, 0.0595) |
| $M_1^*$ | 1456.66 | 1456.674 | 1456.674 | (1456.673, 1456.676) | 1584.76 | 1584.765 | 1584.765 | (1584.763, 1584.767) |
| $R_{2_2}$ | 0.7933 | 0.7798 | 0.7798 | (0.7636, 0.7964) | 0.8703 | 0.8591 | 0.859 | (0.8418, 0.8765) |
| $R_{3_2}$ | 0.3567 | 0.3420 | 0.342 | (0.3305, 0.3537) | 0.4223 | 0.4107 | 0.4107 | (0.3982, 0.4233) |
| $R_{4_2}$ | 0.1166 | 0.1112 | 0.1112 | (0.1054, 0.1171) | 0.1478 | 0.1430 | 0.143 | (0.136, 0.1503) |
| $R_{5_2}$ | 0.0306 | 0.0291 | 0.0291 | (0.0269, 0.0315) | 0.0413 | 0.0400 | 0.0400 | (0.0370, 0.0433) |
| $M_2^*$ | 1460.67 | 1460.682 | 1460.682 | (1460.681, 1460.683) | 1588.77 | 1588.773 | 1588.773 | (1588.772, 1588.775) |
| $\sigma$ | – | 255.6178 | 255.5 | (241.7, 270.4) | – | 256.3566 | 256.2 | (242.1, 271.4) |
| $\sigma_s$ | – | 0.0778 | 0.0778 | (0.0764, 0.0792) | – | 0.0799 | 0.0798 | (0.0766, 0.0832) |
| $\kappa$ | – | 0.8433 | 0.8433 | (0.8234, 0.8621) | – | 0.7602 | 0.76 | (0.744, 0.7770) |
| $S$ | – | 1.0027 | 1.003 | (1.002, 1.004) | – | 1.0029 | 1.003 | (1.002, 1.004) |
| $H_2/H_1$ | 2.4 | 2.2836 | 2.2827 | (2.2117, 2.3598) | 2.4 | 2.2788 | 2.2782 | (2.2052, 2.3558) |

**Table 13.19:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with asymmetric Laplace shape-function for **Spectrum 3** ($H_2/H_1 = 1/3$).

| Parameter | Data set 1 | | | | Data set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.7870 | 0.7869 | (0.7715, 0.8027) | 0.8703 | 0.8602 | 0.8602 | (0.8442, 0.8762) |
| $R_{3_1}$ | 0.3567 | 0.3578 | 0.3578 | (0.3463, 0.3693) | 0.4223 | 0.4269 | 0.4268 | (0.415, 0.439) |
| $R_{4_1}$ | 0.1166 | 0.1183 | 0.1183 | (0.1121, 0.1246) | 0.1478 | 0.1526 | 0.1526 | (0.1453, 0.1601) |
| $R_{5_1}$ | 0.0306 | 0.0311 | 0.0311 | (0.0287, 0.0336) | 0.0413 | 0.0429 | 0.0429 | (0.0396, 0.0463) |
| $M_1^*$ | 1456.66 | 1456.664 | 1456.664 | (1456.663, 1456.665) | 1584.76 | 1584.765 | 1584.765 | (1584.764, 1584.766) |
| $R_{2_2}$ | 0.7933 | 0.7773 | 0.777 | (0.7346, 0.8219) | 0.8703 | 0.8410 | 0.8407 | (0.7952, 0.8885) |
| $R_{3_2}$ | 0.3567 | 0.3486 | 0.3484 | (0.3246, 0.3736) | 0.4223 | 0.4060 | 0.4059 | (0.3788, 0.4341) |
| $R_{4_2}$ | 0.1166 | 0.1143 | 0.1142 | (0.105, 0.1243) | 0.1478 | 0.1425 | 0.1425 | (0.131, 0.1545) |
| $R_{5_2}$ | 0.0306 | 0.0300 | 0.0300 | (0.0270, 0.0333) | 0.0413 | 0.0400 | 0.0399 | (0.0360, 0.0442) |
| $M_2^*$ | 1460.67 | 1460.672 | 1460.672 | (1456.670, 1460.674) | 1588.77 | 1588.774 | 1588.774 | (1588.771, 1588.776) |
| $\sigma$ | – | 290.6453 | 290.4 | (275.1, 307.1) | – | 280.2653 | 280 | (265.2, 296.5) |
| $\sigma_s$ | – | 0.0689 | 0.0689 | (0.0676, 0.0702) | – | 0.0730 | 0.0730 | (0.0717, 0.0743) |
| $\kappa$ | – | 0.8024 | 0.8023 | (0.7831, 0.8213) | – | 0.8482 | 0.8482 | (0.8317, 0.8648) |
| $S$ | – | 1.0036 | 1.004 | (1.003, 1.005) | – | 1.0024 | 1.002 | (1.002, 1.003) |
| $H_2/H_1$ | 1/3 | 0.3055 | 0.3055 | (0.2925, 0.3185) | 1/3 | 0.3042 | 0.3042 | (0.2912, 0.3174) |

**Table 13.20:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with asymmetric Laplace shape-function for **Spectrum 3** ($H_2/H_1 = 3/1$).

| Parameter | Data set 2 | | | | Data set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8315 | 0.8315 | (0.7961, 0.8681) | 0.8703 | 0.9173 | 0.9168 | (0.8768, 0.9598) |
| $R_{3_1}$ | 0.3567 | 0.4188 | 0.4185 | (0.3947, 0.4436) | 0.4223 | 0.5134 | 0.5132 | (0.4842, 0.5439) |
| $R_{4_1}$ | 0.1166 | 0.1417 | 0.1416 | (0.1315, 0.1529) | 0.1478 | 0.1881 | 0.1879 | (0.1742, 0.2024) |
| $R_{5_1}$ | 0.0306 | 0.0372 | 0.0372 | (0.0337, 0.0410) | 0.0413 | 0.0527 | 0.0527 | (0.0477, 0.0582) |
| $M_1^*$ | 1456.66 | 1456.657 | 1456.657 | (1456.655, 1456.658) | 1584.76 | 1584.751 | 1584.751 | (1584.748, 1584.754) |
| $R_{2_2}$ | 0.7933 | 0.7741 | 0.774 | (0.7575, 0.7904) | 0.8703 | 0.8476 | 0.8476 | (0.8289, 0.867) |
| $R_{3_2}$ | 0.3567 | 0.3347 | 0.3347 | (0.3233, 0.3465) | 0.4223 | 0.4078 | 0.4078 | (0.3942, 0.4218) |
| $R_{4_2}$ | 0.1166 | 0.1084 | 0.1083 | (0.1026, 0.1143) | 0.1478 | 0.1416 | 0.1416 | (0.1345, 0.1492) |
| $R_{5_2}$ | 0.0306 | 0.0284 | 0.0284 | (0.0262, 0.0308) | 0.0413 | 0.0396 | 0.0396 | (0.0366, 0.0429) |
| $M_2^*$ | 1460.67 | 1460.664 | 1460.664 | (1460.663, 1460.665) | 1588.77 | 1588.757 | 1588.757 | (1588.755, 1588.760) |
| $\sigma$ | – | 253.6929 | 253.6 | (240.1, 268.1) | – | 273.5024 | 273.3 | (258.4, 289.6025) |
| $\sigma_s$ | – | 0.0698 | 0.0698 | (0.0685, 0.0712) | – | 0.0740 | 0.0740 | (0.0725, 0.0756) |
| $\kappa$ | – | 0.8384 | 0.8384 | (0.8192, 0.8574) | – | 0.7843 | 0.7837 | (0.7572, 0.8155) |
| $S$ | – | 1.0041 | 1.004 | (1.003, 1.005) | – | 1.0026 | 1.003 | (1.002, 1.004) |
| $H_2/H_1$ | 2.4 | 2.3286 | 2.3279 | (2.2530, 2.4077) | 2.4 | 2.3109 | 2.3102 | (2.2279, 2.3981) |

**Table 13.21:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with asymmetric Laplace shape-function for **Spectrum 4** ($H_2/H_1 = 1/3$).

| Parameter | Data set 1 | | | | Data set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.7891 | 0.7889 | (0.7741, 0.8044) | 0.8703 | 0.8611 | 0.861 | (0.8441, 0.8788) |
| $R_{3_1}$ | 0.3567 | 0.3625 | 0.3625 | (0.3513, 0.3736) | 0.4223 | 0.4300 | 0.4299 | (0.4171, 0.4427) |
| $R_{4_1}$ | 0.1166 | 0.1198 | 0.1197 | (0.1138, 0.126) | 0.1478 | 0.1523 | 0.1523 | (0.1449, 0.1599) |
| $R_{5_1}$ | 0.0306 | 0.0315 | 0.0314 | (0.0290, 0.0341) | 0.0413 | 0.0428 | 0.0428 | (0.0395, 0.0462) |
| $M_1^*$ | 1456.66 | 1456.670 | 1456.670 | (1456.669, 1456.671) | 1584.76 | 1584.761 | 1584.761 | (1584.760, 1584.763) |
| $R_{2_2}$ | 0.7933 | 0.7635 | 0.7635 | (0.7224, 0.8074) | 0.8703 | 0.8575 | 0.8574 | (0.8086, 0.907) |
| $R_{3_2}$ | 0.3567 | 0.3427 | 0.3423 | (0.3197, 0.3674) | 0.4223 | 0.4127 | 0.4126 | (0.3838, 0.4424) |
| $R_{4_2}$ | 0.1166 | 0.1124 | 0.1122 | (0.1032, 0.1221) | 0.1478 | 0.1450 | 0.1449 | (0.1332, 0.1573) |
| $R_{5_2}$ | 0.0306 | 0.0295 | 0.0295 | (0.0266, 0.0323) | 0.0413 | 0.0407 | 0.0406 | (0.0366, 0.0450) |
| $M_2^*$ | 1460.67 | 1460.678 | 1460.678 | (1456.676, 1460.680) | 1588.77 | 1588.770 | 1588.770 | (1588.767, 1588.772) |
| $\sigma$ | – | 257.251 | 257.1 | (243.8, 271.5) | – | 267.2495 | 267.1 | (252.8, 282.6) |
| $\sigma_s$ | – | 0.0671 | 0.0671 | (0.0659, 0.0683) | – | 0.0711 | 0.0711 | (0.0698, 0.0726) |
| $\kappa$ | – | 0.8675 | 0.8676 | (0.8495, 0.8851) | – | 0.7756 | 0.7755 | (0.7508, 0.8016) |
| $S$ | – | 1.0030 | 1.003 | (1.002, 1.004) | – | 1.0028 | 1.003 | (1.002, 1.004) |
| $H_2/H_1$ | 1/3 | 0.3106 | 0.3107 | (0.2978, 0.3232) | 1/3 | 0.3069 | 0.3068 | (0.2933, 0.3212) |

**Table 13.22:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with asymmetric Laplace shape-function for **Spectrum 4** ($H_2/H_1 = 3/1$).

| Parameter | Data set 2 | | | | Data set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8256 | 0.8255 | (0.7902, 0.8617) | 0.8703 | 0.9175 | 0.9174 | (0.8773, 0.9594) |
| $R_{3_1}$ | 0.3567 | 0.4165 | 0.4163 | (0.3922, 0.4417) | 0.4223 | 0.5118 | 0.5116 | (0.4823, 0.5417) |
| $R_{4_1}$ | 0.1166 | 0.1416 | 0.1415 | (0.1309, 0.153) | 0.1478 | 0.1882 | 0.188 | (0.1745, 0.2024) |
| $R_{5_1}$ | 0.0306 | 0.0372 | 0.0372 | (0.0337, 0.0410) | 0.0413 | 0.0528 | 0.0527 | (0.0479, 0.0581) |
| $M_1^*$ | 1456.66 | 1456.658 | 1456.658 | (1456.656, 1456.659) | 1584.76 | 1584.759 | 1584.759 | (1584.757, 1584.761) |
| $R_{2_2}$ | 0.7933 | 0.7756 | 0.7756 | (0.7593, 0.7922) | 0.8703 | 0.8411 | 0.841 | (0.8224, 0.86) |
| $R_{3_2}$ | 0.3567 | 0.3360 | 0.3359 | (0.3243, 0.3479) | 0.4223 | 0.4036 | 0.4036 | (0.3902, 0.4175) |
| $R_{4_2}$ | 0.1166 | 0.1086 | 0.1085 | (0.1028, 0.1144) | 0.1478 | 0.1405 | 0.1404 | (0.1333, 0.1478) |
| $R_{5_2}$ | 0.0306 | 0.0284 | 0.0284 | (0.0262, 0.0308) | 0.0413 | 0.0393 | 0.0393 | (0.0362, 0.0425) |
| $M_2^*$ | 1460.67 | 1460.665 | 1460.665 | (1460.664, 1460.666) | 1588.77 | 1588.766 | 1588.766 | (1588.764, 1588.768) |
| $\sigma$ | – | 256.0149 | 255.8 | (242.3, 270.7) | – | 267.6753 | 267.5 | (252.7, 283.5) |
| $\sigma_s$ | – | 0.0696 | 0.0696 | (0.0682, 0.0710) | – | 0.0744 | 0.0744 | (0.0729, 0.0758) |
| $\kappa$ | – | 0.8023 | 0.8022 | (0.7847, 0.8212) | – | 0.8542 | 0.8543 | (0.826, 0.8816) |
| $S$ | – | 1.0039 | 1.004 | (1.003, 1.005) | – | 1.0024 | 1.002 | (1.001, 1.004) |
| $H_2/H_1$ | 2.4 | 2.3096 | 2.3090 | (2.2343, 2.3869) | 2.4 | 2.3121 | 2.3116 | (2.2302, 2.3983) |

**Table 13.23:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with asymmetric Laplace shape-function for **Spectrum 5** ($H_2/H_1 = 1/3$).

| Parameter | Data set 1 | | | | Data set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.7899 | 0.7899 | (0.775, 0.8050) | 0.8703 | 0.8756 | 0.8755 | (0.8586, 0.8927) |
| $R_{3_1}$ | 0.3567 | 0.3625 | 0.3625 | (0.3513, 0.3736) | 0.4223 | 0.4312 | 0.4312 | (0.4188, 0.4438) |
| $R_{4_1}$ | 0.1166 | 0.1197 | 0.1197 | (0.1137, 0.1259) | 0.1478 | 0.1535 | 0.1535 | (0.1461, 0.161) |
| $R_{5_1}$ | 0.0306 | 0.0315 | 0.0314 | (0.0291, 0.0341) | 0.0413 | 0.0432 | 0.0431 | (0.0399, 0.0466) |
| $M_1^*$ | 1456.66 | 1456.662 | 1456.662 | (1456.661, 1456.663) | 1584.76 | 1584.762 | 1584.762 | (1584.761, 1584.764) |
| $R_{2_2}$ | 0.7933 | 0.7776 | 0.7776 | (0.7358, 0.821) | 0.8703 | 0.8341 | 0.8335 | (0.7884, 0.8825) |
| $R_{3_2}$ | 0.3567 | 0.3486 | 0.3483 | (0.3252, 0.3729) | 0.4223 | 0.4037 | 0.4033 | (0.3767, 0.4325) |
| $R_{4_2}$ | 0.1166 | 0.1143 | 0.1142 | (0.1051, 0.1241) | 0.1478 | 0.1418 | 0.1417 | (0.1306, 0.1539) |
| $R_{5_2}$ | 0.0306 | 0.0300 | 0.0300 | (0.0270, 0.0333) | 0.0413 | 0.0398 | 0.0397 | (0.0358, 0.0441) |
| $M_2^*$ | 1460.67 | 1460.669 | 1460.669 | (1456.667, 1460.671) | 1588.77 | 1588.769 | 1588.769 | (1588.766, 1588.772) |
| $\sigma$ | – | 284.1292 | 284 | (269.2, 300.2) | – | 302.2285 | 302 | (285.8, 319.7) |
| $\sigma_s$ | – | 0.0720 | 0.072 | (0.0707, 0.0733) | – | 0.0767 | 0.0767 | (0.0753, 0.0780) |
| $\kappa$ | – | 0.8238 | 0.8238 | (0.8077, 0.8404) | – | 0.8575 | 0.8573 | (0.837, 0.8781) |
| $S$ | – | 1.0038 | 1.004 | (1.003, 1.005) | – | 1.0027 | 1.003 | (1.002, 1.004) |
| $H_2/H_1$ | 1/3 | 0.3037 | 0.3037 | (0.2913, 0.3160) | 1/3 | 0.3096 | 0.3096 | (0.2960, 0.3232) |

**Table 13.24:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with asymmetric Laplace shape-function for **Spectrum 5** ($H_2/H_1 = 3/1$).

| Parameter | Data set 2 | | | | Data set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8174 | 0.817 | (0.7825, 0.8532) | 0.8703 | 0.9117 | 0.9117 | (0.8722, 0.9529) |
| $R_{3_1}$ | 0.3567 | 0.4147 | 0.4146 | (0.3912, 0.4394) | 0.4223 | 0.5096 | 0.5095 | (0.4805, 0.5394) |
| $R_{4_1}$ | 0.1166 | 0.1408 | 0.1407 | (0.1305, 0.1516) | 0.1478 | 0.1873 | 0.1872 | (0.1734, 0.2017) |
| $R_{5_1}$ | 0.0306 | 0.0369 | 0.0369 | (0.0335, 0.0407) | 0.0413 | 0.0525 | 0.0525 | (0.0476, 0.0579) |
| $M_1^*$ | 1456.66 | 1456.656 | 1456.656 | (1456.655, 1456.658) | 1584.76 | 1584.749 | 1584.749 | (1584.747, 1584.752) |
| $R_{2_2}$ | 0.7933 | 0.7616 | 0.7617 | (0.7454, 0.778) | 0.8703 | 0.8419 | 0.8418 | (0.8231, 0.8607) |
| $R_{3_2}$ | 0.3567 | 0.3348 | 0.3347 | (0.3235, 0.3463) | 0.4223 | 0.4057 | 0.4057 | (0.3921, 0.4194) |
| $R_{4_2}$ | 0.1166 | 0.1082 | 0.1082 | (0.1025, 0.1141) | 0.1478 | 0.1409 | 0.1409 | (0.1338, 0.1485) |
| $R_{5_2}$ | 0.0306 | 0.0283 | 0.0283 | (0.0261, 0.0307) | 0.0413 | 0.0394 | 0.0394 | (0.0364, 0.0426) |
| $M_2^*$ | 1460.67 | 1460.664 | 1460.663 | (1460.663, 1460.664) | 1588.77 | 1588.756 | 1588.756 | (1588.754, 1588.758) |
| $\sigma$ | – | 287.7251 | 287.6 | (272.2, 304.3) | – | 310.075 | 309.8 | (292.9, 328.4) |
| $\sigma_s$ | – | 0.0681 | 0.0681 | (0.0669, 0.0694) | – | 0.0722 | 0.0722 | (0.0707, 0.0737) |
| $\kappa$ | – | 0.8366 | 0.8363 | (0.8191, 0.8554) | – | 0.7688 | 0.7688 | (0.7453, 0.7923) |
| $S$ | – | 1.0037 | 1.004 | (1.003, 1.005) | – | 1.0029 | 1.003 | (1.002, 1.004) |
| $H_2/H_1$ | 2.4 | 2.3020 | 2.3018 | (2.2292, 2.3781) | 2.4 | 2.3084 | 2.3081 | (2.2267, 2.3930) |

**Table 13.25:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with asymmetric Laplace shape-function for **Spectrum 6** ($H_2/H_1 = 1/3$).

| Parameter | Data set 1 | | | | Data set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.7992 | 0.7993 | (0.7836, 0.8149) | 0.8703 | 0.8649 | 0.8649 | (0.8482, 0.8817) |
| $R_{3_1}$ | 0.3567 | 0.3661 | 0.3661 | (0.3543, 0.3779) | 0.4223 | 0.4333 | 0.4334 | (0.4208, 0.4458) |
| $R_{4_1}$ | 0.1166 | 0.1214 | 0.1214 | (0.1152, 0.1278) | 0.1478 | 0.1548 | 0.1547 | (0.1475, 0.1623) |
| $R_{5_1}$ | 0.0306 | 0.0319 | 0.0319 | (0.0294, 0.0346) | 0.0413 | 0.0435 | 0.0434 | (0.0402, 0.0470) |
| $M_1^*$ | 1456.66 | 1456.667 | 1456.667 | (1456.666, 1456.668) | 1584.76 | 1584.756 | 1584.756 | (1584.755, 1584.757) |
| $R_{2_2}$ | 0.7933 | 0.7711 | 0.7708 | (0.7296, 0.8148) | 0.8703 | 0.8555 | 0.8552 | (0.8091, 0.9039) |
| $R_{3_2}$ | 0.3567 | 0.3451 | 0.3449 | (0.3212, 0.3701) | 0.4223 | 0.4132 | 0.4129 | (0.3855, 0.4425) |
| $R_{4_2}$ | 0.1166 | 0.1132 | 0.1131 | (0.1039, 0.123) | 0.1478 | 0.1451 | 0.145 | (0.1334, 0.1576) |
| $R_{5_2}$ | 0.0306 | 0.0297 | 0.0297 | (0.0267, 0.0330) | 0.0413 | 0.0407 | 0.0407 | (0.0366, 0.0450) |
| $M_2^*$ | 1460.67 | 1460.674 | 1460.674 | (1456.672, 1460.676) | 1588.77 | 1588.765 | 1588.765 | (1588.762, 1588.766) |
| $\sigma$ | – | 301.1815 | 301 | (285.3, 318.2) | – | 299.9874 | 299.8 | (284, 317.5) |
| $\sigma_s$ | – | 0.0708 | 0.0708 | (0.0694, 0.0721) | – | 0.0732 | 0.0732 | (0.0718, 0.0747) |
| $\kappa$ | – | 0.8768 | 0.8765 | (0.8558, 0.8988) | – | 0.7618 | 0.7615 | (0.7448, 0.7805) |
| $S$ | – | 1.0022 | 1.002 | (1.001, 1.003) | – | 1.0028 | 1.003 | (1.002, 1.003) |
| $H_2/H_1$ | 1/3 | 0.3158 | 0.3157 | (0.3027, 0.3289) | 1/3 | 0.3084 | 0.3084 | (0.2951, 0.3221) |

**Table 13.26:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with asymmetric Laplace shape-function for **Spectrum 6** ($H_2/H_1 = 3/1$).

| Parameter | Data set 2 | | | | Data set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8248 | 0.8247 | (0.7886, 0.8611) | 0.8703 | 0.9159 | 0.9155 | (0.8766, 0.9569) |
| $R_{3_1}$ | 0.3567 | 0.4148 | 0.4146 | (0.3935, 0.4424) | 0.4223 | 0.5140 | 0.5138 | (0.4851, 0.5438) |
| $R_{4_1}$ | 0.1166 | 0.1418 | 0.1417 | (0.131, 0.1527) | 0.1478 | 0.1892 | 0.1891 | (0.1756, 0.2034) |
| $R_{5_1}$ | 0.0306 | 0.0372 | 0.0372 | (0.0337, 0.0411) | 0.0413 | 0.0531 | 0.0530 | (0.0481, 0.0584) |
| $M_1^*$ | 1456.66 | 1456.659 | 1456.659 | (1456.657, 1456.661) | 1584.76 | 1584.76 | 1584.76 | (1584.758, 1584.762) |
| $R_{2_2}$ | 0.7933 | 0.7780 | 0.7779 | (0.7609, 0.7951) | 0.8703 | 0.8425 | 0.8425 | (0.8238, 0.8609) |
| $R_{3_2}$ | 0.3567 | 0.3415 | 0.3415 | (0.3295, 0.3536) | 0.4223 | 0.4054 | 0.4053 | (0.3921, 0.419) |
| $R_{4_2}$ | 0.1166 | 0.1106 | 0.1106 | (0.1047, 0.1168) | 0.1478 | 0.1414 | 0.1414 | (0.1341, 0.1487) |
| $R_{5_2}$ | 0.0306 | 0.0290 | 0.0290 | (0.0267, 0.0314) | 0.0413 | 0.0395 | 0.0395 | (0.0365, 0.0427) |
| $M_2^*$ | 1460.67 | 1460.666 | 1460.666 | (1460.665, 1460.668) | 1588.77 | 1588.768 | 1588.768 | (1588.766, 1588.770) |
| $\sigma$ | – | 292.8212 | 292.7 | (277.1, 309.2) | – | 316.6342 | 316.5 | (299, 335.6) |
| $\sigma_s$ | – | 0.0722 | 0.0722 | (0.0697, 0.0749) | – | 0.0791 | 0.0791 | (0.0776, 0.0807) |
| $\kappa$ | – | 0.7960 | 0.7957 | (0.7779, 0.8155) | – | 0.8415 | 0.8416 | (0.8163, 0.8643) |
| $S$ | – | 1.0036 | 1.004 | (1.003, 1.005) | – | 1.0023 | 1.002 | (1.001, 1.003) |
| $H_2/H_1$ | 2.4 | 2.2516 | 2.2515 | (2.1764, 2.3294) | 2.4 | 2.3039 | 2.3032 | (2.2228, 2.3885) |

## Simultaneous fit to the six spectra:

**Table 13.27:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with normal distribution function.

| Par. | Data set 1 | | | | Data set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.7901 | 0.7901 | (0.7808, 0.7992) | 0.8703 | 0.8624 | 0.8623 | (0.8527, 0.8722) |
| $R_{3_1}$ | 0.3567 | 0.3616 | 0.3616 | (0.3543, 0.3687) | 0.4223 | 0.4316 | 0.4316 | (0.4239, 0.4394) |
| $R_{4_1}$ | 0.1166 | 0.1203 | 0.1203 | (0.1155, 0.1252) | 0.1478 | 0.1551 | 0.1551 | (0.1495, 0.1607) |
| $R_{5_1}$ | 0.0306 | 0.0316 | 0.0316 | (0.0293, 0.034) | 0.0413 | 0.0436 | 0.0436 | (0.0405, 0.0469) |
| $M_1^*$ | 1456.66 | 1456.673 | 1456.673 | (1456.673, 1456.674) | 1584.76 | 1584.771 | 1584.771 | (1584.771, 1584.772) |
| $R_{2_2}$ | 0.7933 | 0.7652 | 0.7652 | (0.7382, 0.7923) | 0.8703 | 0.8410 | 0.841 | (0.8113, 0.8711) |
| $R_{3_2}$ | 0.3567 | 0.3371 | 0.3371 | (0.3202, 0.3546) | 0.4223 | 0.3999 | 0.3997 | (0.3811, 0.4197) |
| $R_{4_2}$ | 0.1166 | 0.1102 | 0.1101 | (0.1028, 0.118) | 0.1478 | 0.1400 | 0.1399 | (0.131, 0.1492) |
| $R_{5_2}$ | 0.0306 | 0.0289 | 0.0289 | (0.0264, 0.0317) | 0.0413 | 0.0393 | 0.0392 | (0.0358, 0.0429) |
| $M_2^*$ | 1460.67 | 1460.681 | 1460.681 | (1460.68, 1460.683) | 1588.77 | 1588.779 | 1588.779 | (1588.778, 1588.780) |
| $\sigma_1$ | – | 361.4743 | 361.3 | (340.4, 383.8) | – | 351.1390 | 351 | (331.6, 372.2) |
| $\sigma_2$ | – | 291.8526 | 291.7 | (274.8, 309.7) | – | 468.4105 | 467.9 | (442.1, 497.2) |
| $\sigma_3$ | – | 392.5628 | 392.3 | (369.4, 417.2) | – | 483.824 | 483.4 | (455.5, 513.6) |
| $\sigma_4$ | – | 419.5152 | 419.1 | (394, 446.9) | – | 312.5077 | 312.2 | (295.7, 330.4) |
| $\sigma_5$ | – | 471.3549 | 471.2 | (442.6, 501.2025) | – | 427.0039 | 426.7 | (402.1, 453.5) |
| $\sigma_6$ | – | 385.0705 | 384.8 | (363.1, 409.3) | – | 377.7365 | 377.4 | (356.9, 400.1) |
| $\sigma_s$ | – | 0.0454 | 0.0454 | (0.0450, 0.0458) | – | 0.0482 | 0.0482 | (0.0478, 0.0486) |
| $S$ | – | 1.0026 | 1.002 | (1.002, 1.003) | – | 1.0026 | 1.003 | (1.002, 1.003) |
| $\overline{Q}$ | 1/3 | 0.3104 | 0.3104 | (0.2869, 0.334) | 1/3 | 0.3081 | 0.3082 | (0.2841, 0.3324) |
| $\sigma_Q^2$ | – | 0.00076 | 0.00053 | (0.00018, 0.00271) | – | 0.00077 | 0.00053 | (0.00018, 0.00285) |

**Table 13.28:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with normal distribution function.

| Par. | Data set 2 | | | | Data set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8073 | 0.8071 | (0.7876, 0.8275) | 0.8703 | 0.8932 | 0.8932 | (0.8716, 0.9156) |
| $R_{3_1}$ | 0.3567 | 0.4498 | 0.4497 | (0.4344, 0.4657) | 0.4223 | 0.5563 | 0.5563 | (0.5385, 0.5741) |
| $R_{4_1}$ | 0.1166 | 0.1629 | 0.1628 | (0.1539, 0.1721) | 0.1478 | 0.2233 | 0.2233 | (0.2118, 0.235) |
| $R_{5_1}$ | 0.0306 | 0.0428 | 0.0427 | (0.0393, 0.0464) | 0.0413 | 0.0626 | 0.0626 | (0.0578, 0.0677) |
| $M_1^*$ | 1456.66 | 1456.668 | 1456.668 | (1456.668, 1456.669) | 1584.76 | 1584.768 | 1584.768 | (1584.767, 1584.769) |
| $R_{2_2}$ | 0.7933 | 0.7747 | 0.7746 | (0.7657, 0.7839) | 0.8703 | 0.8492 | 0.8493 | (0.839, 0.8593) |
| $R_{3_2}$ | 0.3567 | 0.3342 | 0.3342 | (0.3272, 0.3413) | 0.4223 | 0.4063 | 0.4063 | (0.3986, 0.4139) |
| $R_{4_2}$ | 0.1166 | 0.1060 | 0.106 | (0.1017, 0.1104) | 0.1478 | 0.1393 | 0.1392 | (0.1341, 0.1446) |
| $R_{5_2}$ | 0.0306 | 0.0277 | 0.0277 | (0.0258, 0.0298) | 0.0413 | 0.0388 | 0.0388 | (0.0362, 0.0415) |
| $M_2^*$ | 1460.67 | 1460.676 | 1460.676 | (1460.675, 1460.676) | 1588.77 | 1588.776 | 1588.776 | (1588.776, 1588.777) |
| $\sigma_1$ | – | 322.6952 | 322.5 | (304.2, 342.3025) | – | 393.7457 | 393.5 | (371.3, 418.0025) |
| $\sigma_2$ | – | 544.3902 | 543.8 | (513.4, 577.5) | – | 357.8858 | 357.7 | (337.8, 379.3) |
| $\sigma_3$ | – | 316.5448 | 316.3 | (299, 335.4) | – | 295.3017 | 295.2 | (279.3, 312.4) |
| $\sigma_4$ | – | 285.8252 | 285.5 | (270.9, 302.3) | – | 327.2925 | 327 | (309, 347) |
| $\sigma_5$ | – | 398.0664 | 397.8 | (375.1, 422.2) | – | 355.3285 | 355.2 | (335.8, 375.9) |
| $\sigma_6$ | – | 317.6863 | 317.5 | (300.8, 335.9) | – | 424.1894 | 423.9 | (400, 449.9) |
| $\sigma_s$ | – | 0.0461 | 0.0461 | (0.0458, 0.0465) | – | 0.0491 | 0.0491 | (0.0488, 0.0496) |
| $S$ | – | 1.0027 | 1.003 | (1.002, 1.003) | – | 1.0025 | 1.003 | (1.002, 1.003) |
| $\overline{Q}$ | 2.4 | 2.2915 | 2.292 | (2.234, 2.347) | 2.4 | 2.2933 | 2.294 | (2.238, 2.349) |
| $\sigma_Q^2$ | – | 0.00219 | 0.00134 | (0.00032, 0.00934) | – | 0.00190 | 0.00118 | (0.00030, 0.00805) |

**Table 13.29:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with asymmetric Laplace distribution function.

| Par. | Data set 1 | | | | Data set 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.7913 | 0.7913 | (0.7827, 0.7997) | 0.8703 | 0.8627 | 0.8626 | (0.8541, 0.8717) |
| $R_{3_1}$ | 0.3567 | 0.3629 | 0.3629 | (0.3562, 0.3695) | 0.4223 | 0.4322 | 0.4322 | (0.4253, 0.4392) |
| $R_{4_1}$ | 0.1166 | 0.1209 | 0.1209 | (0.1162, 0.1256) | 0.1478 | 0.1558 | 0.1557 | (0.1505, 0.161) |
| $R_{5_1}$ | 0.0306 | 0.0318 | 0.0317 | (0.0295, 0.0342) | 0.0413 | 0.0438 | 0.0437 | (0.0408, 0.0469) |
| $M_1^*$ | 1456.66 | 1456.663 | 1456.663 | (1456.661, 1456.665) | 1584.76 | 1584.757 | 1584.757 | (1584.756, 1584.758) |
| $R_{2_2}$ | 0.7933 | 0.7681 | 0.768 | (0.7426, 0.7941) | 0.8703 | 0.8427 | 0.8425 | (0.8155, 0.8702) |
| $R_{3_2}$ | 0.3567 | 0.3384 | 0.3383 | (0.3222, 0.3552) | 0.4223 | 0.3993 | 0.3992 | (0.3817, 0.4177) |
| $R_{4_2}$ | 0.1166 | 0.1107 | 0.1106 | (0.1036, 0.1181) | 0.1478 | 0.1397 | 0.1396 | (0.1313, 0.1485) |
| $R_{5_2}$ | 0.0306 | 0.0291 | 0.0291 | (0.0265, 0.0318) | 0.0413 | 0.0392 | 0.0392 | (0.0360, 0.0428) |
| $M_2^*$ | 1460.67 | 1460.671 | 1460.671 | (1460.669, 1460.673) | 1588.77 | 1588.766 | 1588.766 | (1588.765, 1588.767) |
| $\sigma_1$ | – | 367.4027 | 367.4 | (342.1, 393.1) | – | 284.841 | 284.6 | (269.2, 301.7) |
| $\sigma_2$ | – | 245.9896 | 245.6 | (230, 263.7) | – | 405.1527 | 404.8 | (381.5, 430.8) |
| $\sigma_3$ | – | 333.6486 | 333 | (311.7, 358.5) | – | 520.4787 | 520.4 | (490.7, 552.4) |
| $\sigma_4$ | – | 463.3965 | 463.5 | (430.1975, 496.1) | – | 279.4288 | 279.1 | (264.1975, 296.1) |
| $\sigma_5$ | – | 408.194 | 407.3 | (379.3, 440.8) | – | 434.491 | 434.1 | (409.6, 461.6025) |
| $\sigma_6$ | – | 406.6493 | 406.5 | (377.9, 436.2025) | – | 310.2171 | 310.1 | (293, 328.5) |
| $\sigma_s$ | – | 0.0712 | 0.0712 | (0.0705, 0.0720) | – | 0.0745 | 0.0745 | (0.0737, 0.0752) |
| $\kappa$ | – | 0.8045 | 0.8039 | (0.7889, 0.8233) | – | 0.7591 | 0.759 | (0.7479, 0.7705) |
| $S$ | – | 1.0027 | 1.003 | (1.002, 1.003) | – | 1.0029 | 1.003 | (1.002, 1.003) |
| $\overline{Q}$ | 1/3 | 0.3098 | 0.3098 | (0.2865, 0.3333) | 1/3 | 0.3085 | 0.3085 | (0.2855, 0.3322) |
| $\sigma_Q^2$ | – | 0.00076 | 0.00052 | (0.00018, 0.00273) | – | 0.00076 | 0.00052 | (0.00018, 0.00269) |

**Table 13.30:** Means, medians and 95% credible intervals based on the samples from posterior distributions for the parameters of the model with asymmetric Laplace distribution function.

| Par. | Data set 2 | | | | Data set 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | True | Mean | Median | 95% c.i. | True | Mean | Median | 95% c.i. |
| $R_{2_1}$ | 0.7933 | 0.8064 | 0.8065 | (0.7882, 0.8242) | 0.8703 | 0.8912 | 0.8911 | (0.8712, 0.9112) |
| $R_{3_1}$ | 0.3567 | 0.4559 | 0.4559 | (0.4413, 0.4706) | 0.4223 | 0.5627 | 0.5626 | (0.5468, 0.579) |
| $R_{4_1}$ | 0.1166 | 0.1687 | 0.1686 | (0.16, 0.1776) | 0.1478 | 0.2320 | 0.2321 | (0.2207, 0.2434) |
| $R_{5_1}$ | 0.0306 | 0.0444 | 0.0444 | (0.0409, 0.0482) | 0.0413 | 0.0652 | 0.0652 | (0.0602, 0.0705) |
| $M_1^*$ | 1456.66 | 1456.658 | 1456.658 | (1456.657, 1456.659) | 1584.76 | 1584.756 | 1584.756 | (1584.755, 1584.757) |
| $R_{2_2}$ | 0.7933 | 0.7768 | 0.7767 | (0.7686, 0.7853) | 0.8703 | 0.8516 | 0.8516 | (0.8424, 0.8607) |
| $R_{3_2}$ | 0.3567 | 0.3346 | 0.3346 | (0.3282, 0.3411) | 0.4223 | 0.4077 | 0.4077 | (0.4007, 0.4147) |
| $R_{4_2}$ | 0.1166 | 0.1057 | 0.1057 | (0.1016, 0.1099) | 0.1478 | 0.1394 | 0.1394 | (0.1345, 0.1442) |
| $R_{5_2}$ | 0.0306 | 0.0276 | 0.0276 | (0.0257, 0.0296) | 0.0413 | 0.0388 | 0.0388 | (0.0363, 0.0414) |
| $M_2^*$ | 1460.67 | 1460.665 | 1460.665 | (1460.664, 1460.666) | 1588.77 | 1588.763 | 1588.763 | (1588.762, 1588.763) |
| $\sigma_1$ | – | 318.8121 | 318.6 | (300.8, 337.9) | – | 348.2114 | 347.9 | (327.0975, 370.4) |
| $\sigma_2$ | – | 532.7193 | 532.5 | (503.1, 564.2) | – | 284.6756 | 284.4 | (267.9975, 302.6025) |
| $\sigma_3$ | – | 279.207 | 279 | (263.7, 295.7) | – | 266.8882 | 266.7 | (252.6975, 282.5) |
| $\sigma_4$ | – | 250.2610 | 250.2 | (236.9, 264.4) | – | 315.3316 | 315.1 | (297.0975, 334.5) |
| $\sigma_5$ | – | 344.3234 | 344 | (325.1, 365.2) | – | 306.5559 | 306.4 | (290, 324.3) |
| $\sigma_6$ | – | 288.2081 | 288 | (272.8, 304.6) | – | 411.1908 | 411 | (387.4, 436.5) |
| $\sigma_s$ | – | 0.0716 | 0.0716 | (0.0710, 0.0723) | – | 0.0764 | 0.0764 | (0.0757, 0.0772) |
| $\kappa$ | – | 0.8007 | 0.8006 | (0.7919, 0.8101) | – | 0.7674 | 0.7673 | (0.7571, 0.7781) |
| $S$ | – | 1.0039 | 1.004 | (1.003, 1.004) | – | 1.0029 | 1.003 | (1.002, 1.003) |
| $\overline{Q}$ | 2.4 | 2.2910 | 2.291 | (2.239, 2.344) | 2.4 | 2.2972 | 2.297 | (2.247, 2.348) |
| $\sigma_Q^2$ | – | 0.00210 | 0.00135 | (0.00033, 0.00829) | – | 0.001660 | 0.001088 | (0.000285, 0.006464) |

## Simulation results of the eight settings:

**Table 13.31:** Summary statistics (mean estimates $\bar{\hat{\theta}}$, empirical standard errors $\sigma_{emp}$ and model based standard errors $\sigma_{mb}$).

| Parameter | set1 | | | set2 | | | set3 | | | set4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ |
| $R_{2_1}$ | 0.9763 | 0.9644 | 0.0490/0.0084 | 0.9763 | 1.0145 | 0.0631/0.0117 | 1.1577 | 1.2543 | 0.0558/0.0102 | 1.1577 | 1.5896 | 0.9267/0.0854 |
| $R_{3_1}$ | 0.6385 | 0.6172 | 0.0491/0.0089 | 0.6385 | 0.6331 | 0.0572/0.0167 | 0.6702 | 0.7768 | 0.0644/0.0122 | 0.6702 | 0.8053 | 0.2652/0.0531 |
| $R_{4_1}$ | 0.3125 | 0.2913 | 0.0315/0.0061 | 0.3125 | 0.2796 | 0.0282/0.0098 | 0.2586 | 0.3207 | 0.0338/0.0077 | 0.2586 | 0.2772 | 0.0069/0.0179 |
| $R_{5_1}$ | 0.1277 | 0.1133 | 0.0141/0.0037 | 0.1277 | 0.1018 | 0.0086/0.0050 | 0.0749 | 0.0967 | 0.0125/0.0038 | 0.0749 | 0.0929 | 0.0022/0.0070 |
| $M_1^*$ | 2000.90 | 2000.906 | 0.0008/0.0002 | 2000.90 | 2000.934 | 0.0004/0.0002 | 2000.90 | 2000.900 | 0.0004/0.0003 | 2000.90 | 2000.901 | 0.0064/0.0013 |
| $R_{2_2}$ | 1.2708 | 1.1007 | 0.1401/0.1230 | 1.2708 | 1.0360 | 0.1193/0.0873 | 1.1577 | 1.2180 | 0.0984/0.0955 | 1.1577 | 1.1727 | 0.1041/0.0169 |
| $R_{3_2}$ | 0.8872 | 0.6620 | 0.0173/0.0397 | 0.8872 | 0.6556 | 0.0190/0.0348 | 0.6702 | 0.6678 | 0.0214/0.0363 | 0.6702 | 0.6771 | 0.0873/0.0127 |
| $R_{4_2}$ | 0.4431 | 0.2882 | 0.0061/0.0197 | 0.4431 | 0.2905 | 0.0077/0.0179 | 0.2586 | 0.2776 | 0.0089/0.0180 | 0.2586 | 0.2602 | 0.0337/0.0055 |
| $R_{5_2}$ | 0.1750 | 0.0984 | 0.0024/0.0076 | 0.1750 | 0.1017 | 0.0033/0.0072 | 0.0749 | 0.0923 | 0.0026/0.0068 | 0.0749 | 0.0787 | 0.0062/0.0027 |
| $M_2^*$ | 2000.94 | 2002.645 | 1.6026/0.0527 | 2000.94 | 2002.055 | 0.4913/0.0226 | 2001.94 | 2002.488 | 1.6254/0.0593 | 2001.94 | 2001.992 | 0.2177/0.0009 |
| $\sigma$ | 10 | 7.6873 | 0.3352/0.2180 | 10 | 7.6568 | 0.3052/0.2158 | 10 | 7.6193 | 0.3223/0.2490 | 10 | 7.8147 | 0.9477/0.2278 |
| $\sigma_s$ | 0.08 | 0.0813 | 0.0004/0.0002 | 0.08 | 0.0811 | 0.0002/0.0001 | 0.08 | 0.0803 | 0.0005/0.0002 | 0.08 | 0.0798 | 0.0004/0.0002 |
| $S$ | 1.0015 | 1.0014 | 0.0011/0.0004 | 1.0015 | 1.0012 | 0.0008/0.0004 | 1.0015 | 1.0036 | 0.0018/0.0004 | 1.0015 | 1.0009 | 0.0013/0.0005 |
| $H_2/H_1$ | 0.2 | 0.0703 | 0.0424/0.0089 | 5 | 0.2146 | 0.0496/0.0117 | 0.2 | 0.1041 | 0.0522/0.0102 | 5 | 4.7323 | 0.2842/0.1145 |

**Table 13.32:** Summary statistics (mean estimates $\bar{\hat{\theta}}$, empirical standard errors $\sigma_{emp}$ and model based standard errors $\sigma_{mb}$).

| Parameter | set5 | | | set6 | | | set7 | | | set8 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ |
| $R_{2_1}$ | 1.1577 | 1.1542 | 0.0244/0.0054 | 1.1577 | 1.1637 | 0.0250/0.0199 | 0.9763 | 0.9759 | 0.0048/0.0039 | 0.9763 | 0.9790 | 0.0236/0.0190 |
| $R_{3_1}$ | 0.6702 | 0.6568 | 0.0634/0.0044 | 0.6702 | 0.6720 | 0.0172/0.0149 | 0.6385 | 0.6388 | 0.0031/0.0033 | 0.6385 | 0.6427 | 0.0178/0.0152 |
| $R_{4_1}$ | 0.2586 | 0.2401 | 0.0484/0.0038 | 0.2586 | 0.2682 | 0.0103/0.0109 | 0.3125 | 0.3120 | 0.0035/0.0029 | 0.3125 | 0.3052 | 0.0126/0.0120 |
| $R_{5_1}$ | 0.0749 | 0.1420 | 0.0773/0.0057 | 0.0749 | 0.0962 | 0.0010/0.0074 | 0.1277 | 0.1241 | 0.0035/0.0027 | 0.1277 | 0.1036 | 0.0049/0.0071 |
| $M_1^*$ | 2000.90 | 2000.899 | 0.0008/0.0003 | 2000.90 | 2000.900 | 0.00011/0.0009 | 2000.90 | 2000.900 | 0.0003/0.0003 | 2000.90 | 2000.900 | 0.0011/0.0010 |
| $R_{2_2}$ | 1.1577 | 1.0016 | 0.2770/0.0406 | 1.1577 | 1.1612 | 0.0049/0.0044 | 1.2708 | 1.2057 | 0.0231/0.0200 | 1.2708 | 1.2660 | 0.0054/0.0045 |
| $R_{3_2}$ | 0.6702 | 0.6695 | 0.2421/0.0239 | 0.6702 | 0.6722 | 0.0039/0.0033 | 0.8872 | 0.8266 | 0.0163/0.0162 | 0.8872 | 0.8834 | 0.0042/0.0037 |
| $R_{4_2}$ | 0.2586 | 0.3103 | 0.1317/0.0145 | 0.2586 | 0.2598 | 0.0032/0.0027 | 0.4431 | 0.3866 | 0.0158/0.0128 | 0.4431 | 0.4398 | 0.0034/0.0030 |
| $R_{5_2}$ | 0.0749 | 0.1020 | 0.0309/0.0072 | 0.0749 | 0.0788 | 0.0027/0.0024 | 0.1750 | 0.1163 | 0.0056/0.0080 | 0.1750 | 0.1702 | 0.0033/0.0028 |
| $M_2^*$ | 2004.94 | 2004.960 | 1.0072/0.0630 | 2004.94 | 2004.940 | 0.0004/0.0003 | 2006.94 | 2006.940 | 0.0010/0.0009 | 2006.94 | 2006.940 | 0.0003/0.0002 |
| $\sigma$ | 10 | 9.8833 | 3.7120/0.5880 | 10 | 7.8690 | 0.3281/0.2232 | 10 | 8.2762 | 0.3302/0.2410 | 10 | 8.2775 | 0.3150/0.2433 |
| $\sigma_s$ | 0.08 | 0.0800 | 0.0003/0.0002 | 0.08 | 0.0800 | 0.0002/0.0002 | 0.08 | 0.0800 | 0.0003/0.0002 | 0.08 | 0.0801 | 0.0002/0.0002 |
| $S$ | 1.0015 | 1.0024 | 0.0015/0.0003 | 1.0015 | 1.0014 | 0.0005/0.0002 | 1.0015 | 1.0015 | 0.0005/0.0002 | 1.0015 | 1.0016 | 0.0005/0.0002 |
| $H_2/H_1$ | 0.2 | 0.1855 | 0.0431/0.0076 | 5 | 4.9980 | 0.0745/0.0672 | 0.2 | 0.2120 | 0.0030/0.0027 | 5 | 5.0181 | 0.0844/0.0687 |

# Chapter 14

# A Bayesian model averaging approach for the shape representation of a mass spectrum with overlapping peptides

## 14.1 Introduction

As it has been observed in Chapter 13, the Bayesian (mixture) model performs well when MS data show a clear separation of the overlapping peptides, i.e., when the separation is clearly seen in the data and when the mass difference of the overlapping peptides is at least around 4 Da. However, when the separation is less apparent or the mass difference is smaller, the Bayesian (mixture) model does not perform well. In particular, it provides biased estimates of the monoisotopic mass of the overlapping peptide. To tackle the problem, in this chapter, we introduce the Bayesian model averaging approach.

## 14.2 Model implementation for Bayesian model averaging

In the Bayesian model averaging approach, the model formulation is given by equations (13.1)-(13.2). The difference lies in the prior distribution for the monoisotopic mass $M_d^*$. Instead of specifying a mixture of $G$ normal distributions, defined in equation (13.10), $G$ separate models are fitted, each with a normal prior $N(\eta_g, \sigma_m^2)$, where $G$ is the number of (normal) components, which can possibly contain the true value of the monoisotopic mass of the overlapping peptide. Taking Figure 3.1 as an example, suppose that, the monoisotopic mass $M_2^*$ is likely to appear around the mass range of 1997.5–2002.5 Da, then there shows five clusters (normal components) that could contain the true value of $M_2^*$. Thus, $G$ is equal to five for this case. The resulting parameter estimates will be a weighted sum of the $G$ candidate models. This means the point estimate of a parameter $\theta$ can be treated as the weighted average of the model-specific estimate $\widehat{\theta}_g$:

$$\hat{\theta} = \sum_{g=1}^{G} w_g \widehat{\theta}_g, \tag{14.1}$$

where $w_g$ is the weight of the $g$th model. As suggested by Burham and Anderson (2002), $w_g$ can be computed based on the Information Criteria. Based on the DIC (Deviance Information Criterion) of each model, $w_g$ can be computed as follows:

$$w_g = \frac{\exp\left(-\frac{1}{2}\Delta DIC_g\right)}{\left[\sum\limits_{g=1}^{G} \exp\left(-\frac{1}{2}\Delta DIC_g\right)\right]}, \tag{14.2}$$

where $\Delta DIC_g = DIC_g - \underset{g}{min}(DIC_g)$. The standard error can be computed as suggested by Burham and Anderson (2002) and Eicher *et al.* (2009):

$$\hat{\sigma}(\theta) = \sum_{g=1}^{G} w_g \sqrt{\widehat{\sigma}_g(\theta)^2 + \left(\widehat{\theta}_g - \widehat{\theta}\right)^2}, \tag{14.3}$$

where $\widehat{\theta}_g$ and $\widehat{\sigma}_g(\theta)$ are, respectively, the point estimate and the standard error for parameter $\theta$ in the $g$th candidate model.

In the next section, we show the performance of Bayesian model averaging approach applied to a simulation study as well as to the bovine cytochrome C data set.

## 14.3 Results

### 14.3.1 A simulation study

For illustration purposes and simplicity, the simulation was based on the model with normal distribution (shape-)function (see Section 13.6). We considered 30 settings, accounting for various mass differences of the overlapping peptides, with both clear and unclear peptide separation. The details of the settings are shown in Table 14.1. Let $shift$ be the integer of the mass difference of the two overlapping peptides, and $tilt$ be the mass difference after the decimal point, with $tilt < 1$. As a result, the mass difference of the two overlapping peptides is equal to $M_2 - M_1 = shift + tilt$. Or, in other words, $M^2 = M_1 + shift + tilt$. It may be difficult to quantify two overlapping peptides when the mass difference between two peptides is too small, i.e., either $shift$ or $tilt$ is very small. Thus, it is of interest to investigate different settings with combinations of the two parameters.

The other parameters were chosen based on real-life data:

$M_1^* = 2000.90, \quad H_1 = 10000, \quad \sigma = 10, \quad \sigma_s = 0.08, \quad S = 1.0015.$

For each of the settings, 100 simulated data sets with a random error were generated, based on the model defined in equations (13.1)–(13.2). Figures 14.1 to 14.3 show the graphical representation of the 30 settings. It can be seen that settings 1–3, 5–7, 9–16, 18–19 and 21 are difficult ones, for which the location of the second, overlapping peptide is not immediately obvious. These settings are the ones with either much less abundant second, overlapping peptide (e.g., setting 21), or with very small mass difference (e.g., setting 2).

Model (13.1)-(13.2) was fitted by using the *R* package *R2WinBUGS*, built in *R* to automatically call the *WinBUGS1.4* software, which allows to fit Bayesian models. The DIC values were obtained as an automatic output from the *R2WinBUGS* package.

Tables E.1 to E.4 show the average weights of the eight candidate models for the 30 settings. Tables E.5 to E.12 show the summary statistics, i.e., the average point estimates of the 100 data sets (denoted as $\bar{\hat{\theta}}$), the mean model-based standard error of the 100 data sets (denoted as $\sigma_{mb}$) and the empirical standard error (denoted as

**Table 14.1:** The combinations of parameters used for the 30 settings of the simulation study.

| | set1 | set2 | set3 | set4 | set5 | set6 | set7 | set8 |
|---|---|---|---|---|---|---|---|---|
| *shift* | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| *tilt* | 0.04 | 0.04 | 0.04 | 0.04 | 0.16 | 0.16 | 0.16 | 0.16 |
| $H_2/H_1$ | 0.2 | 5 | 0.2 | 5 | 0.5 | 2 | 0.5 | 2 |
| Isotopoic Ratios | **E1E2** | **E1E2** | **AA** | **AA** | **E2A** | **E2A** | **AE1** | **AE1** |

| | set9 | set10 | set11 | set12 | set13 | set14 | set15 | set16 |
|---|---|---|---|---|---|---|---|---|
| *shift* | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| *tilt* | 0.24 | 0.24 | 0.24 | 0.16 | 0.16 | 0.16 | 0.04 | 0.04 |
| $H_2/H_1$ | 1 | 1 | 2 | 0.2 | 5 | 0.2 | 0.5 | 0.5 |
| Isotopoic Ratios | **E2E1** | **E1E1** | **E2A** | **E1E2** | **E1E2** | **AA** | **E2A** | **AE1** |

| | set17 | set18 | set19 | set20 | set21 | set22 | set23 | set24 |
|---|---|---|---|---|---|---|---|---|
| *shift* | 1 | 0 | 0 | 1 | 4 | 4 | 6 | 6 |
| *tilt* | 0.04 | 0.04 | 0.16 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |
| $H_2/H_1$ | 2 | 1 | 1 | 1 | 0.2 | 5 | 0.2 | 5 |
| Isotopoic Ratios | **AE1** | **E2E1** | **E2E1** | **E1E1** | **AA** | **AA** | **E1E2** | **E1E2** |

| | set25 | set26 | set27 | set28 | set29 | set30 | | |
|---|---|---|---|---|---|---|---|---|
| *shift* | 4 | 4 | 6 | 6 | 4 | 6 | | |
| *tilt* | 0.16 | 0.16 | 0.16 | 0.16 | 0.24 | 0.24 | | |
| $H_2/H_1$ | 0.5 | 2 | 0.5 | 2 | 1 | 1 | | |
| Isotopoic Ratios | **AE1** | **AE1** | **E2A** | **E2A** | **E2E2** | **E2E1** | | |

$\sigma_{emp}$), computed from the point estimates of the 100 simulated data sets.

The graphical representation of the summary statistics are shown in Figures 14.4 to 14.17. The point estimates of the mass for the second (overlapping) peptide $M_2^*$, shown in Figure 14.4, correctly represent the true mass of the peptide, except only for settings 1–3, 6, 15 and 18. For these settings, the 95% credible intervals, computed based on the model averaging, are very wide. Thus, most of them still contain the true values of $M_2^*$. This is clearly an improvement over the previous approach, presented in Chapter 13. The wide credible intervals are an indication of settings, for which the separation of the overlapping peptides is not easily discernible. For these same (difficult) settings, the 95% credible intervals of the relative abundance $H_2/H_1$, shown in Figure 14.5, contain zero and thus can be viewed as another indication that the second, overlapping peptide is difficult to be found. For the remaining settings, even for some of those, for which the presence of the second peptide is not clear from the data, the Bayesian model averaging approach is able to estimate the monoisotopic masses of the two overlapping peptides and to correctly quantify their relative abundance. A slight bias for the estimation of $H_2/H_1$ is only observed for setting 21.

Figures 14.6 to 14.9 show the estimation of the isotopic ratios for the first peptide. For the settings, for which the separation of the overlapping peptides can be detected by the model averaging approach, the point estimates of the isotopic ratios, especially for $R_{2_1}$, $R_{3_1}$ and $R_{4_1}$, show negligible bias with their 95% credible intervals including, in general, the true values. For $R_{5_1}$, however, the estimation is comparatively worse, as can be observed from Figure 14.9. This is because $R_{5_1}$ is much less abundant than the other ratios, making it more difficult to estimate. Moreover, the corresponding isotopic peak of $R_{5_1}$, in most of the settings, gets overlapped with the (more abundant) isotopic peaks of the second peptide. This makes $R_{5_1}$ even more difficult to estimate. Similar patterns can be observed from Figures 14.10 – 14.13.

Figures 14.14, 14.16 and 14.17 show, in general, unbiased estimates for parameters $M_1^*$, $\sigma_s$, and $S$, respectively. The estimates of residual standard deviation parameter $\sigma$, shown in Figure 14.15 are, in general, also well estimated with only a slight underestimation.

Figures 14.18 – 14.22 present the fit of the predicted spectra (using the average point estimates obtained from Tables E.5 to E.12 in Appendix E) versus the observed spectra. These figures indicate, for most of the settings that, the fitted spectra correspond to the observed spectra, except only for settings 6 and 15. For these settings, since the point estimates of $M_2^*$ appear in the middle of two observed peaks, the resulting isotopic peaks of the second (overlapping) peptides, appear around the 'valleys' in between the two neighboring observed peaks.

As can be seen from Table 14.1, settings 1 to 20 are the settings, for which the monoisotopic mass difference of the two overlapping peptides is at most around one Da, i.e., $shift = 0$ or 1. These settings can be viewed as the more difficult due to the relatively small difference in the monoisotopic mass, i.e., $M_2^* - M_1^*$. Table 14.2 gives a summary of whether or not the model is able to produce correct estimates for these settings with combinations of $shift$, *tilt*, and $H_2/H_1$, based on the simulation study. Note that the poor estimates are produced when the mass difference $M_2^* - M_1^*$ or the relative abundance $H_2/H_1$ is too small. For the mass difference, it can either happen when $shift$ or *tilt* is very small. When $shift$ is small, e.g., when $shift = 0$, the isotopic peaks of the two peptides, appearing at the same observed peaks, exhibit (almost) complete overlap, and thus it becomes difficult for the model to detect a peak envelope, as a result of a mixture of two isotopic peak envelopes.

In particular, Table 14.2 indicates that, in general, when $M_2^* - M_1^* \geq 0.16$, the model produces the correct parameter estimates. The special case happens when

**Table 14.2:** Correctness of model estimates for settings with various combinations of $shift$, $tilt$ and $H_2/H_1$. (+: correct estimation; −: wrong estimation.)

| | 1 | | 2 | | | | 5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $R_H(=H_2/H_1)$ | $R_H$ | $R_H$ | $1/R_H$ | $R_H$ | $1/R_H$ | $R_H$ | $1/R_H$ | $R_H$ | $1/R_H$ | $R_H$ |
| $shift$ | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| $tilt = 0.04$ | − | + | − | − | + | + | − | − | − | + |
| $tilt = 0.16$ | + | + | + | − | + | + | + | + | + | + |
| $tilt = 0.24$ | + | + | + | + | + | + | + | + | + | + |

$M_2^* - M_1^* = 0.16$ and $H_2/H_1 = 2$. In such case, the mixture of the isotopic peak envelopes of the two peptides is difficult to be discerned, because the isotopic peaks of the two peptides show a complete overlap.

### 14.3.2 Application to bovine cytochrome C mass spectra

To investigate the performance of the model averaging approach when applied to real-life data, we fitted the model to the bovine cytochrome C data set. The model was applied to the same peptides of the data set with replicated joint mass spectra of bovine cytochrome C from LC Packings, as described in Chapter 13. The fit the model, we used the *R* package *R2WinBUGS*, built in *R* to automatically call the *WinBUGS1.4* software.

#### Fit for the individual spectrum

As the results of the six replicated spectra exhibit similar pattern, we only present one of them. These results are shown in Tables 14.3 and 14.4. The results are quite comparable with the ones shown in Section 13.5.1. In particular, the point estimates of the relative abundance $H_2/H_1$ for all the cases are slightly underestimated. The isotopic ratios are well-estimated, and, in general, their 95% credible intervals contain the true values.

#### Simultaneous fit to six technical replications of spectra

Similar to the model presented in Chapter 13, a model, specified in equations (13.16)–(13.17), incorporating random effects for the relative abundance parameter, denoted as $Q_j$ with $Q_j \sim N(\overline{Q}, \sigma_Q^2)$, can be fitted simultaneously to the six technical replicates of the spectra.

**Table 14.3:** Means, and standard errors based on model averaging for the parameters of the model with asymmetric Laplace shape-function.

| Parameter | Data set 1 | | | Data set 3 | | |
|---|---|---|---|---|---|---|
| | True | Mean | S.E. | True | Mean | S.E. |
| $R_{2_1}$ | 0.7933 | 0.7907 | 0.008395 | 0.8703 | 0.8598 | 0.008567 |
| $R_{3_1}$ | 0.3567 | 0.3593 | 0.006669 | 0.4223 | 0.4347 | 0.006836 |
| $R_{4_1}$ | 0.1166 | 0.1229 | 0.005162 | 0.1478 | 0.1575 | 0.005743 |
| $R_{5_1}$ | 0.0306 | 0.0328 | 0.002562 | 0.0413 | 0.0441 | 0.003436 |
| $M_1^*$ | 1456.66 | 1456.674 | 0.0006 | 1584.76 | 1584.764 | 0.0006 |
| $R_{2_2}$ | 0.7933 | 0.7842 | 0.02669 | 0.8703 | 0.8487 | 0.02809 |
| $R_{3_2}$ | 0.3567 | 0.3508 | 0.01589 | 0.4223 | 0.4072 | 0.01712 |
| $R_{4_2}$ | 0.1166 | 0.1202 | 0.007787 | 0.1478 | 0.1493 | 0.009163 |
| $R_{5_2}$ | 0.0306 | 0.0322 | 0.002467 | 0.0413 | 0.0430 | 0.003287 |
| $M_2^*$ | 1460.67 | 1460.682 | 0.0011 | 1588.77 | 1588.773 | 0.0011 |
| $\sigma$ | – | 278.7780 | 7.7737 | – | 273.1138 | 7.8657 |
| $\sigma_s$ | – | 0.0742 | 0.0007 | – | 0.0767 | 0.0007 |
| $\kappa$ | – | 0.8662 | 0.01141 | – | 0.7591 | 0.008672 |
| $S$ | – | 1.0022 | 0.0006 | – | 1.0028 | 0.0005 |
| $H_2/H_1$ | 1/3 | 0.3059 | 0.007047 | 1/3 | 0.3064 | 0.007222 |

**Table 14.4:** Means, and standard errors based on model averaging for the parameters of the model with asymmetric Laplace shape-function.

| Parameter | Data set 2 | | | Data set 4 | | |
|---|---|---|---|---|---|---|
| | True | Mean | S.E. | True | Mean | S.E. |
| $R_{2_1}$ | 0.7933 | 0.7615 | 0.01930 | 0.8703 | 0.8419 | 0.01850 |
| $R_{3_1}$ | 0.3567 | 0.4357 | 0.01472 | 0.4223 | 0.5305 | 0.01507 |
| $R_{4_1}$ | 0.1166 | 0.1536 | 0.009825 | 0.1478 | 0.1973 | 0.01213 |
| $R_{5_1}$ | 0.0306 | 0.0324 | 0.002528 | 0.0413 | 0.0431 | 0.003362 |
| $M_1^*$ | 1456.66 | 1456.668 | 0.0012 | 1584.76 | 1584.762 | 0.0007 |
| $R_{2_2}$ | 0.7933 | 0.7717 | 0.008791 | 0.8703 | 0.8571 | 0.008971 |
| $R_{3_2}$ | 0.3567 | 0.3362 | 0.006908 | 0.4223 | 0.4057 | 0.007161 |
| $R_{4_2}$ | 0.1166 | 0.1104 | 0.005162 | 0.1478 | 0.1417 | 0.005569 |
| $R_{5_2}$ | 0.0306 | 0.0315 | 0.002299 | 0.0413 | 0.0419 | 0.002860 |
| $M_2^*$ | 1460.67 | 1460.676 | 0.0009 | 1588.77 | 1588.771 | 0.0005 |
| $\sigma$ | – | 232.6181 | 6.6199 | – | 217.1335 | 6.5086 |
| $\sigma_s$ | – | 0.0734 | 0.0007 | – | 0.0770 | 0.0007 |
| $\kappa$ | – | 0.8637 | 0.01402 | – | 0.8095 | 0.007834 |
| $S$ | – | 1.0018 | 0.0006 | – | 1.0029 | 0.0004 |
| $H_2/H_1$ | 2.4 | 2.2519 | 0.03791 | 2.4 | 2.2181 | 0.03462 |

The results with the asymmetric Laplace distribution function are presented in Tables 14.5 and 14.6.

In general, the results are consistent with those shown in Section 13.5.1. More particularly, there is still a slight under-estimation of the mean relative abundance parameter $\overline{Q}$, but with 95% credible intervals containing the true values. The estimates of between-spectra variability of the relative abundance, captured by $\sigma_Q^2$, are again very small since for these technical replicates, the values of $\boldsymbol{Q}$ are expected to be the same.

The isotopic ratios are well-estimated with their 95% credible intervals, in general, covering the true values. The monoisotopic masses of the two peptides, ignoring their rounding errors, are correctly estimated.

## 14.4 Concluding remarks

In this chapter, we have presented the model for the shape representation of a mass spectrum with overlapping peptides, fitted by using the Bayesian model averaging. We have compared its performance with the Bayesian (mixture) model, presented in Chapter 13, via the application to the bovine cytochrome C data set and a simulation study. The approach has two advantages, as compared with the Bayesian (mixture) modeling approach:

- It produces unbiased estimates for all settings that show clear (visually discernible) or unclear separation of the overlapping peptides in the MS data;
- The model uncertainty, measured by the 95% credible intervals of the parameters, gives an indication of the difficulty to quantify the second, overlapping peptide.

The results of the application to the real-life data show, in general, unbiased estimates of the parameters and are consistent with the ones presented in Chapter 13. This means that the Bayesian (mixture) modeling approach, introduced in Chapter 13, produce equally good estimates as the Bayesian model averaging approach, presented in this chapter. This is because the settings for the real-life data exhibit apparent separation for the two overlapping peptides.

In the simulation study, when applying Bayeisan model averaging, we observed, in general, unbiased estimates for the settings with either clear or unclear separation for the overlapping peptides in the simulated MS data. Moreover, for the settings, for which the quantification of the second, overlapping peptide was difficult, the 95% credible intervals of the parameter estimates were wider and still contained the true values. This indicates that the width of the 95% credible intervals, by using the Bayesian model averaging approach, gives an indication of whether or not the separation of the peptides can be made.

The separability of the overlapping peptides depends on the mass difference of the peptides. When Bayesian model averaging approach is applied, it produces unbiased estimates for the parameters related to the overlapping peptides, when the monoisotoipc mass difference is at least 0.16 Da, which is roughly a half of the width of an isotopic peak, observed in a MALDI-TOF mass spectrum. This indicates that the two overlapping peptides can be correctly quantified by using the Bayesian model averaging approach when the mass difference of the two peptides is at least a half of the width of an isotopic peak. A smaller mass difference, i.e., less than a half of the isotopic peak width, would suggest a complete overlap of the peptides and would make the quantification infeasible. The computational speed for each of the models, presented in this chapter, was estimated to be approximately 20 minutes, on a HP8530p laptop under Windows Vista®. Bearing in mind that the models are highly complex ones, being implemented in the Bayesian framework, the numerical complexity can be treated as tolerable to be implemented in a high-throughput environment.

It should be noted that the validity of this approach is based on a proper preprocessing procedure (for details of pre-processing, refer to Valkenborg *et al.* 2009). More specifically, it assumes that a cluster of peptide peaks is correctly found after noise filtering. This implies that, if a part of the isotopic peaks of a cluster is treated as noise-generated and are discarded, the Bayesian model averaging approach would yield biased estimation.

The approach implemented so far, only handles situations when the number of overlapping peptides is known. Future research is needed for approaches allowing estimation of the number of overlapping peptides.

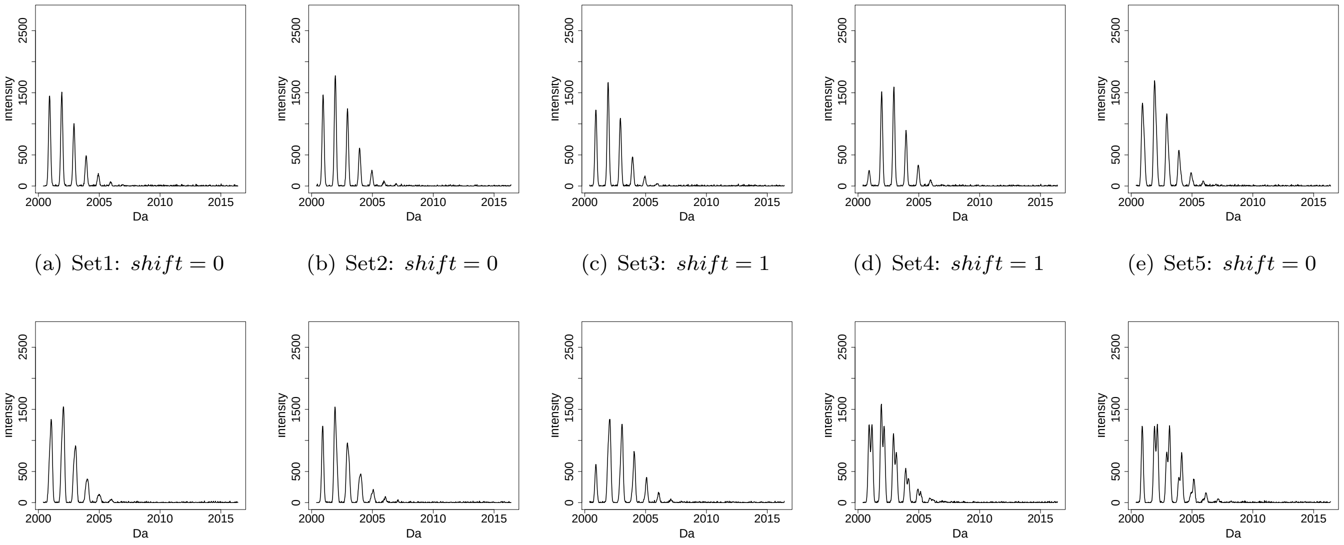
## Statistical results of the simultaneous fit to the MS of the case study:

**Table 14.5:** Means, and standard errors based on model averaging for the parameters of the model with asymmetric Laplace shape-function.

| Parameter | Data set 1 | | | Data set 3 | | |
|---|---|---|---|---|---|---|
| | True | Mean | S.E. | True | Mean | S.E. |
| $R_{2_1}$ | 0.7933 | 0.7911 | 0.004388 | 0.8703 | 0.8621 | 0.004478 |
| $R_{3_1}$ | 0.3567 | 0.3627 | 0.003612 | 0.4223 | 0.4319 | 0.003639 |
| $R_{4_1}$ | 0.1166 | 0.1230 | 0.003217 | 0.1478 | 0.1580 | 0.003313 |
| $R_{5_1}$ | 0.0306 | 0.0331 | 0.002544 | 0.0413 | 0.0446 | 0.003352 |
| $M_1^*$ | 1456.66 | 1456.663 | 0.0008 | 1584.76 | 1584.757 | 0.0004 |
| $R_{2_2}$ | 0.7933 | 0.7787 | 0.01541 | 0.8703 | 0.8543 | 0.01730 |
| $R_{3_2}$ | 0.3567 | 0.3383 | 0.01052 | 0.4223 | 0.3998 | 0.01127 |
| $R_{4_2}$ | 0.1166 | 0.1052 | 0.006465 | 0.1478 | 0.1433 | 0.007471 |
| $R_{5_2}$ | 0.0306 | 0.0322 | 0.002418 | 0.0413 | 0.0428 | 0.003174 |
| $M_2^*$ | 1460.67 | 1460.671 | 0.0010 | 1588.77 | 1588.766 | 0.0005 |
| $\sigma_1$ | – | 364.3375 | 12.7121 | – | 284.6342 | 8.4533 |
| $\sigma_2$ | – | 248.0763 | 8.4008 | – | 404.3642 | 12.8051 |
| $\sigma_3$ | – | 336.6604 | 11.7703 | – | 521.7210 | 15.8608 |
| $\sigma_4$ | – | 458.8311 | 15.9636 | – | 279.7208 | 8.0997 |
| $\sigma_5$ | – | 412.7333 | 15.2735 | – | 435.6440 | 13.2885 |
| $\sigma_6$ | – | 402.8385 | 14.2977 | – | 309.4983 | 9.2123 |
| $\sigma_s$ | – | 0.0713 | 0.0004 | – | 0.0744 | 0.0004 |
| $\kappa$ | – | 0.8083 | 0.007911 | – | 0.7578 | 0.005169 |
| $S$ | – | 1.0026 | 0.0005 | – | 1.0029 | 0.0003 |
| $\overline{Q}$ | 1/3 | 0.3088 | 0.2521 | 1/3 | 0.3072 | 0.2489 |
| $\sigma_Q^2$ | – | 0.00060 | 0.0008202 | – | 0.00057 | 0.0006084 |

**Table 14.6:** Means, and standard errors based on model averaging for the parameters of the model with asymmetric Laplace shape-function.

| Parameter | Data set 2 | | | Data set 4 | | |
|---|---|---|---|---|---|---|
| | True | Mean | S.E. | True | Mean | S.E. |
| $R_{2_1}$ | 0.7933 | 0.7791 | 0.009461 | 0.8703 | 0.8911 | 0.01042 |
| $R_{3_1}$ | 0.3567 | 0.4545 | 0.007774 | 0.4223 | 0.5628 | 0.008514 |
| $R_{4_1}$ | 0.1166 | 0.1041 | 0.007140 | 0.1478 | 0.2321 | 0.005739 |
| $R_{5_1}$ | 0.0306 | 0.0327 | 0.002528 | 0.0413 | 0.0652 | 0.002620 |
| $M_1^*$ | 1456.66 | 1456.658 | 0.0005 | 1584.76 | 1584.756 | 0.0006 |
| $R_{2_2}$ | 0.7933 | 0.7755 | 0.004216 | 0.8703 | 0.8516 | 0.004638 |
| $R_{3_2}$ | 0.3567 | 0.3343 | 0.003387 | 0.4223 | 0.4077 | 0.003552 |
| $R_{4_2}$ | 0.1166 | 0.1027 | 0.002959 | 0.1478 | 0.1394 | 0.002487 |
| $R_{5_2}$ | 0.0306 | 0.0299 | 0.001830 | 0.0413 | 0.0389 | 0.001313 |
| $M_2^*$ | 1460.66 | 1460.665 | 0.0003 | 1588.76 | 1588.763 | 0.0004 |
| $\sigma_1$ | – | 316.8024 | 9.5611 | – | 348.1936 | 11.1470 |
| $\sigma_2$ | – | 531.6636 | 15.7337 | – | 284.7879 | 8.7807 |
| $\sigma_3$ | – | 277.2690 | 8.1782 | – | 266.7438 | 7.6838 |
| $\sigma_4$ | – | 246.6829 | 6.9257 | – | 315.2750 | 9.5970 |
| $\sigma_5$ | – | 341.6068 | 10.3204 | – | 306.4608 | 8.7765 |
| $\sigma_6$ | – | 284.7614 | 8.0729 | – | 410.9400 | 12.5120 |
| $\sigma_s$ | – | 0.0716 | 0.0003 | – | 0.0764 | 0.0004 |
| $\kappa$ | – | 0.8016 | 0.004752 | – | 0.7675 | 0.005519 |
| $S$ | – | 1.0039 | 0.0003 | – | 1.0029 | 0.0003 |
| $\overline{Q}$ | 2.4 | 2.2832 | 0.3226 | 2.4 | 2.2979 | 0.2693 |
| $\sigma_Q^2$ | – | 0.00219 | 0.006632 | – | 0.00169 | 0.006297 |



(a) Set1: $shift = 0$

(b) Set2: $shift = 0$

(c) Set3: $shift = 1$

(d) Set4: $shift = 1$

(e) Set5: $shift = 0$

(f) Set6: $shift = 0$

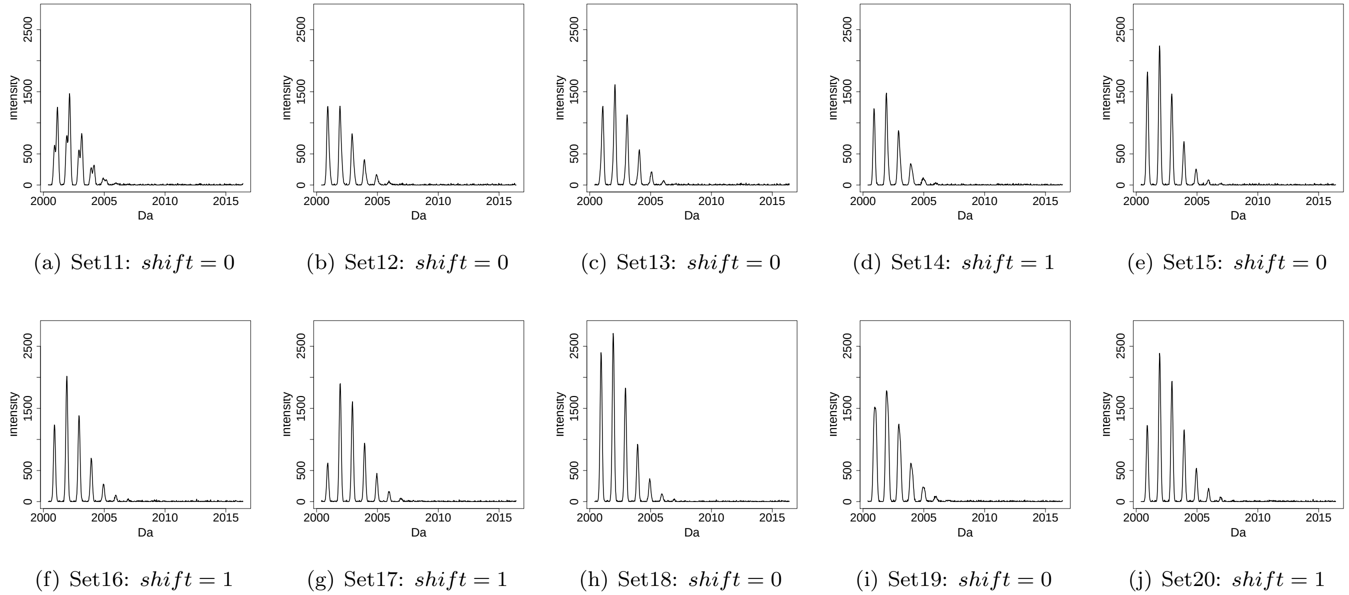
(g) Set7: $shift = 1$

(h) Set8: $shift = 1$

(i) Set9: $shift = 0$

(j) Set10: $shift = 1$

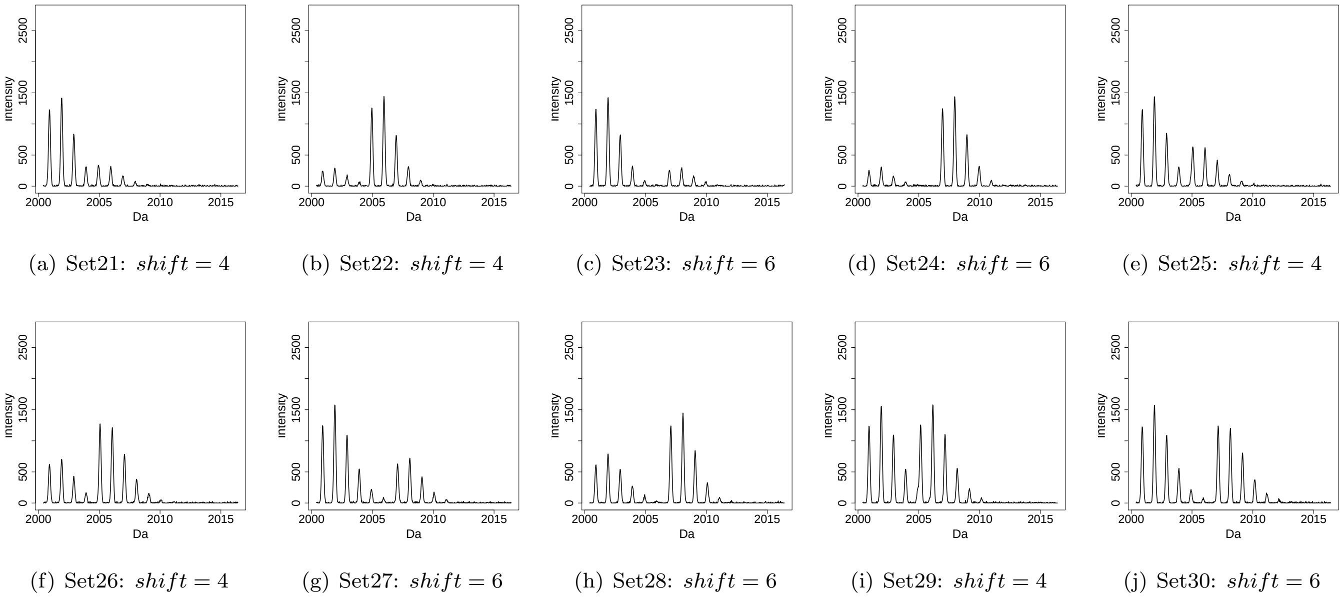
**Figure 14.1:** Graphical representation of settings $1 \sim 10$ of simulated data sets.



(a) Set11: $shift = 0$ (b) Set12: $shift = 0$ (c) Set13: $shift = 0$ (d) Set14: $shift = 1$ (e) Set15: $shift = 0$

(f) Set16: $shift = 1$ (g) Set17: $shift = 1$ (h) Set18: $shift = 0$ (i) Set19: $shift = 0$ (j) Set20: $shift = 1$

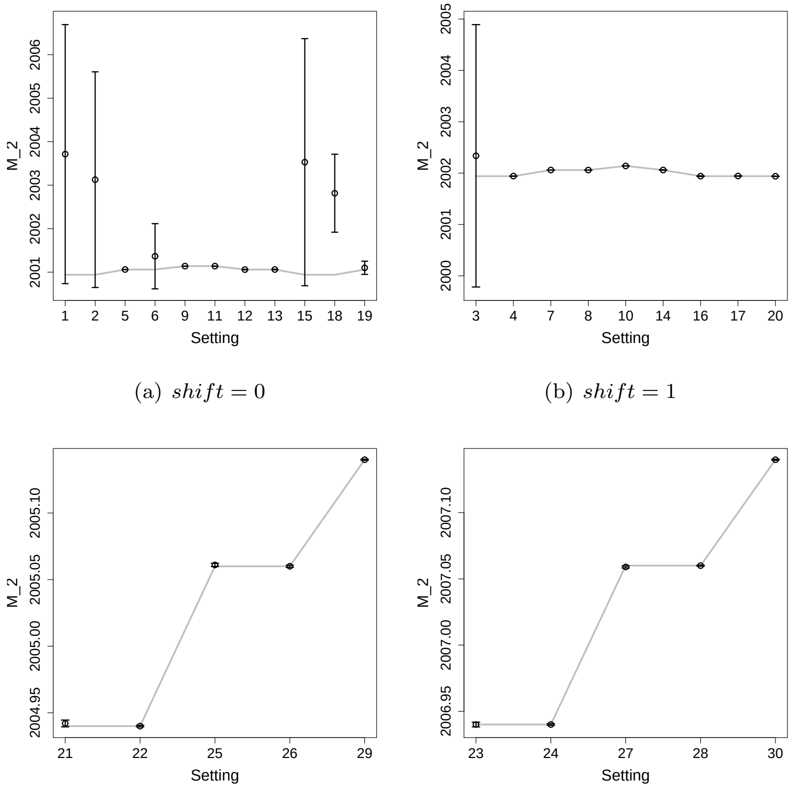
**Figure 14.2:** Graphical representation of settings $11 \sim 20$ of simulated data sets.



(a) Set21: $shift = 4$
(b) Set22: $shift = 4$
(c) Set23: $shift = 6$
(d) Set24: $shift = 6$
(e) Set25: $shift = 4$

(f) Set26: $shift = 4$
(g) Set27: $shift = 6$
(h) Set28: $shift = 6$
(i) Set29: $shift = 4$
(j) Set30: $shift = 6$

**Figure 14.3:** Graphical representation of settings $21 \sim 30$ of simulated data sets.



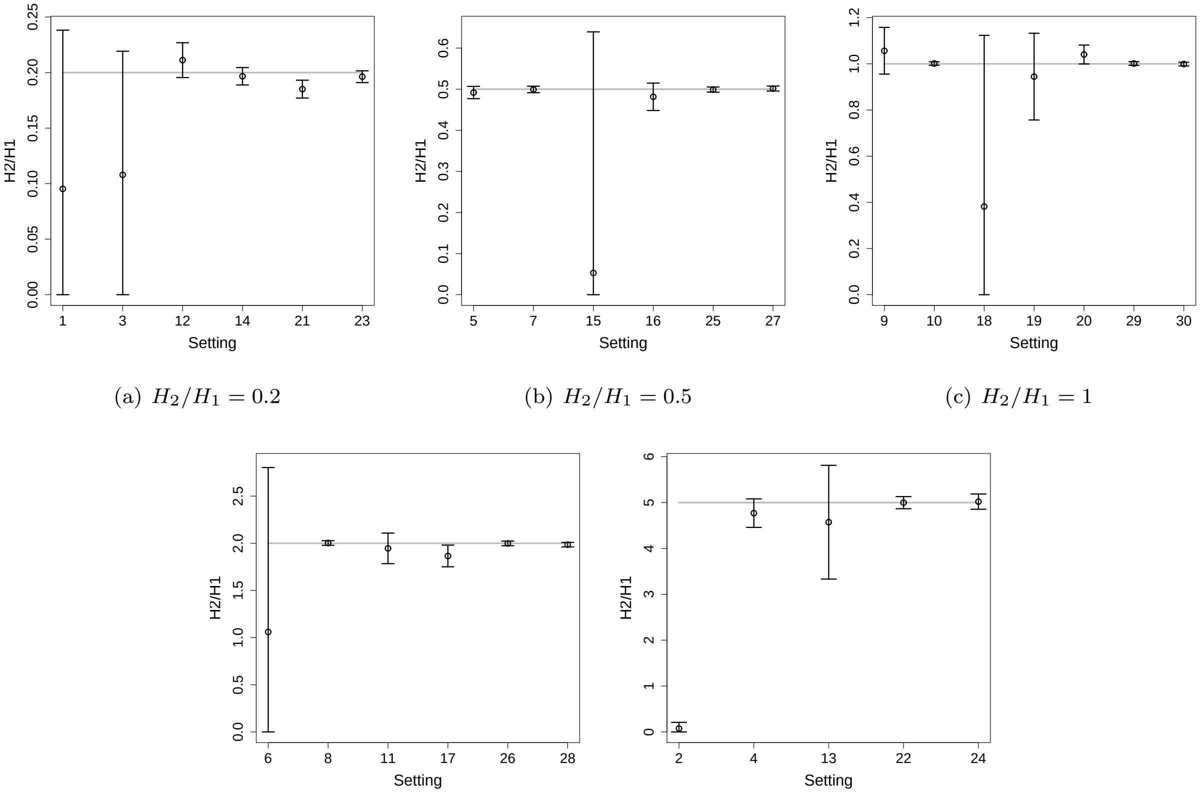
(a) $shift = 0$

(b) $shift = 1$

(c) $shift = 4$

(d) $shift = 6$

**Figure 14.4:** Graphical representation of the average point estimates with the 95% credible intervals for $M_2^*$ (True values indicated by the grey line).



(a) $H_2/H_1 = 0.2$

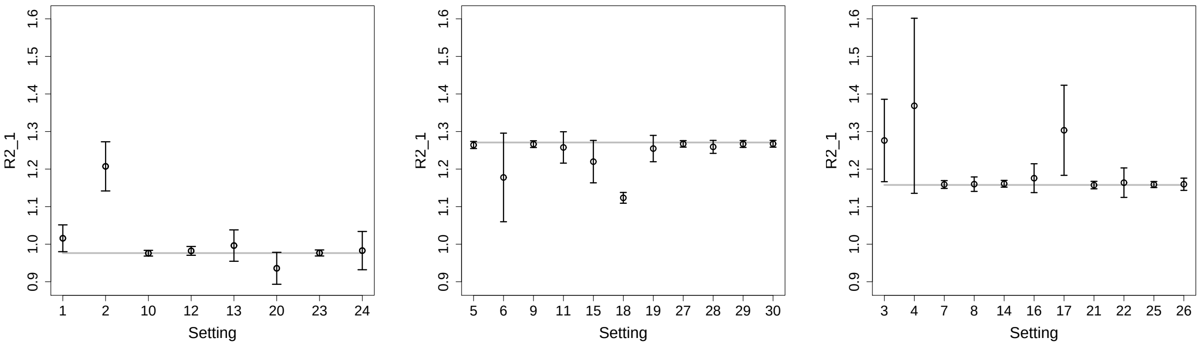
(b) $H_2/H_1 = 0.5$

(c) $H_2/H_1 = 1$

(d) $H_2/H_1 = 2$

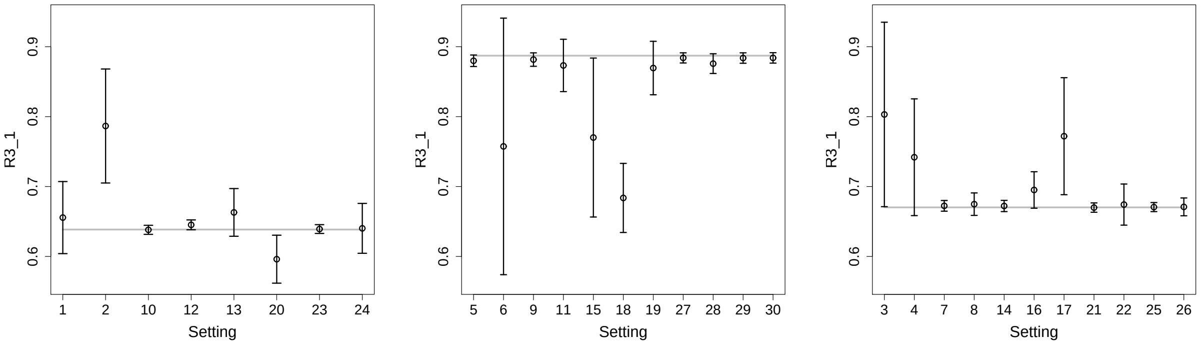
(e) $H_2/H_1 = 5$

**Figure 14.5:** Graphical representation of the average point estimates with the 95% credible intervals for $H_2/H_1$ (True value indicated by the horizontal grey line).



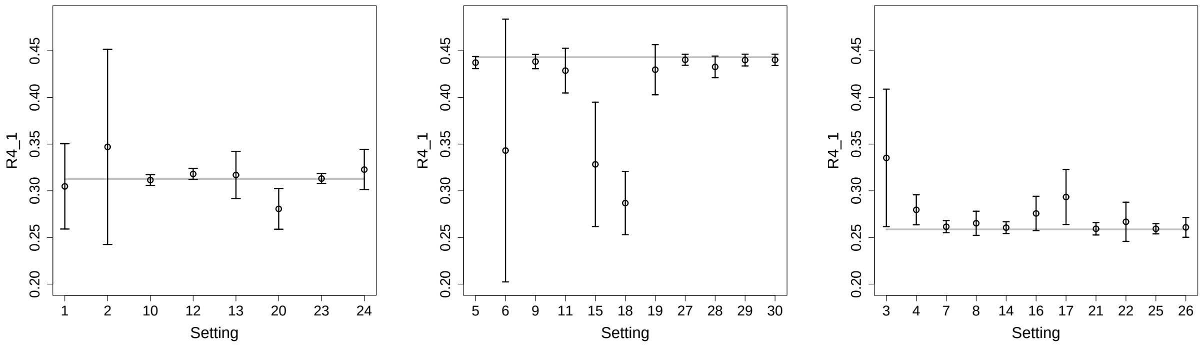
(a) E1 (b) E2 (c) A

**Figure 14.6:** Graphical representation of the average point estimates with the 95% credible intervals for $R_{2_1}$ (True value indicated by the horizontal grey line).



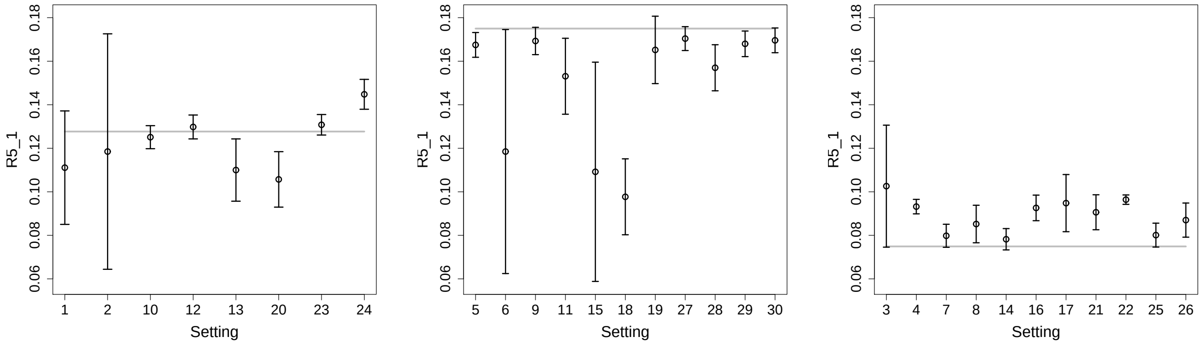
(a) E1 (b) E2 (c) A

**Figure 14.7:** Graphical representation of the average point estimates with the 95% credible intervals for $R_{3_1}$ (True value indicated by the horizontal grey line).



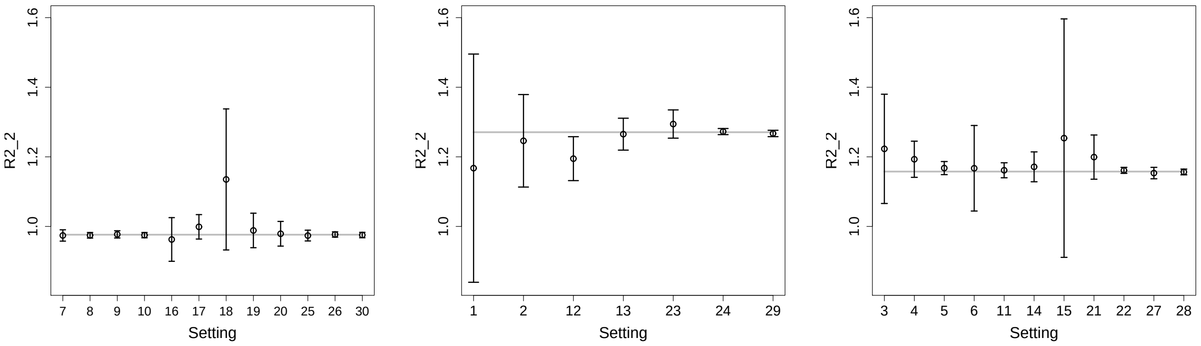
(a) E1 (b) E2 (c) A

**Figure 14.8:** Graphical representation of the average point estimates with the 95% credible intervals for $R_{4_1}$ (True value indicated by the horizontal grey line).



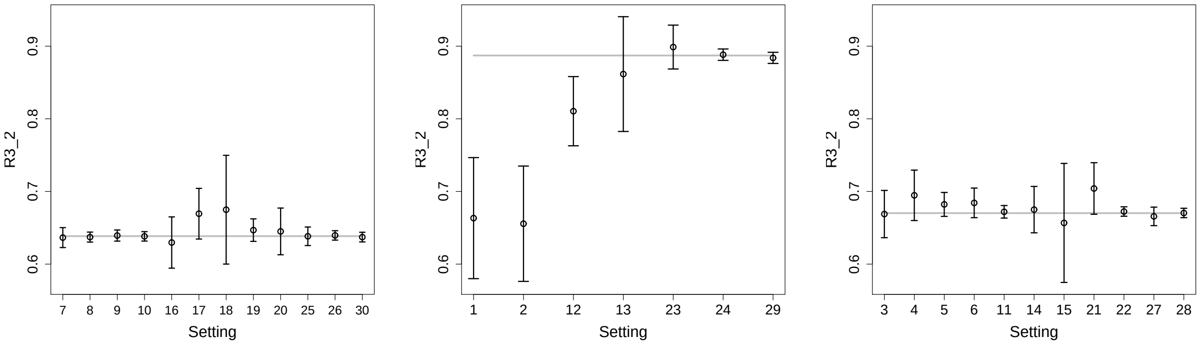
(a) E1 (b) E2 (c) A

**Figure 14.9:** Graphical representation of the average point estimates with the 95% credible intervals for $R_{5_1}$ (True value indicated by the horizontal grey line).



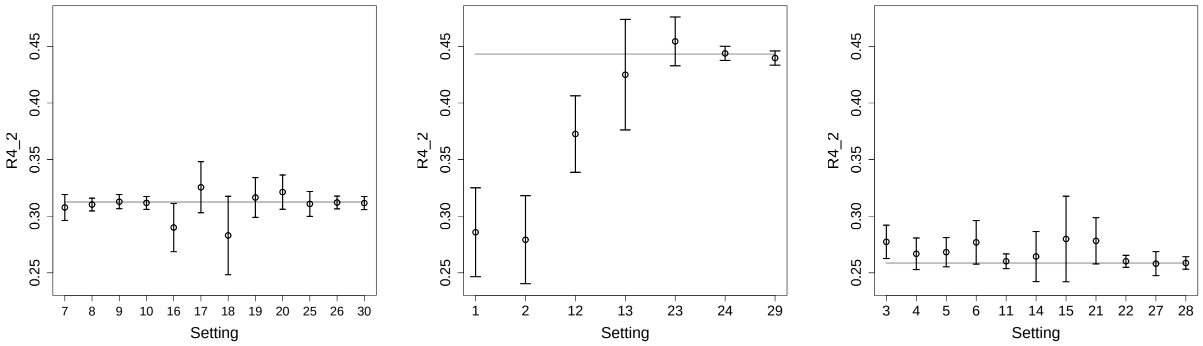
(a) E1 (b) E2 (c) A

**Figure 14.10:** Graphical representation of the average point estimates with the 95% credible intervals for $R_{2_2}$ (True value indicated by the horizontal grey line).



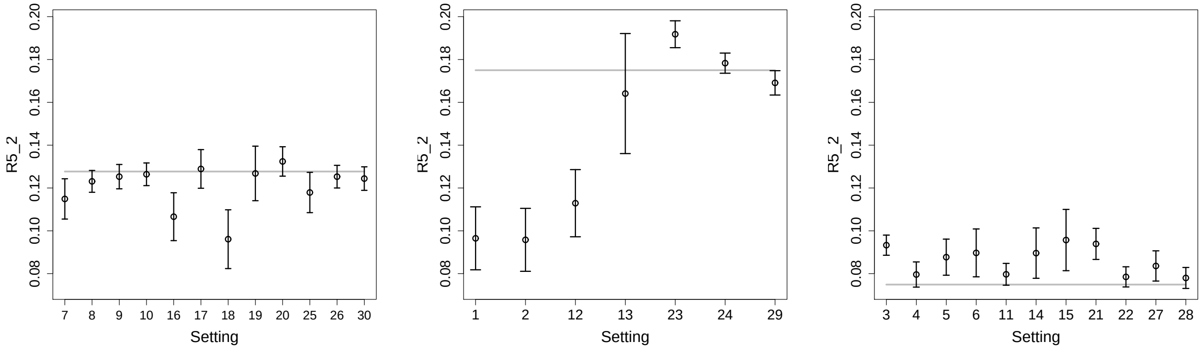
(a) E1 (b) E2 (c) A

**Figure 14.11:** Graphical representation of the average point estimates with the 95% credible intervals for $R_{3_2}$ (True value indicated by the horizontal grey line).



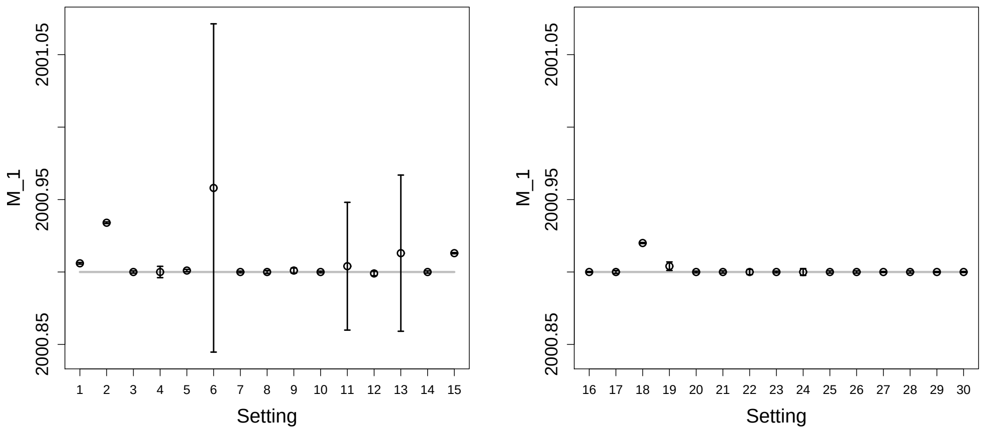
(a) E1 (b) E2 (c) A

**Figure 14.12:** Graphical representation of the average point estimates with the 95% credible intervals for $R_{4_2}$ (True value indicated by the horizontal grey line).



(a) E1 (b) E2 (c) A

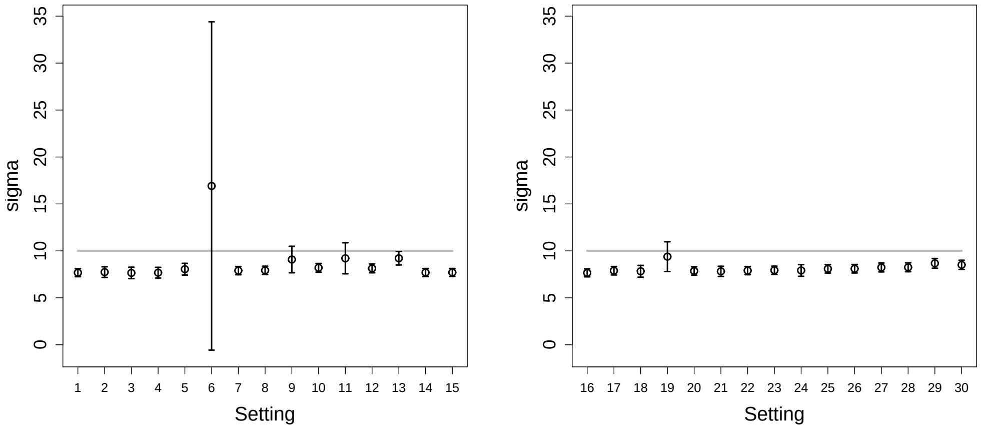
**Figure 14.13:** Graphical representation of the average point estimates with the 95% credible intervals for $R_{5_2}$ (True value indicated by the horizontal grey line).



(a) Set1–15

(b) Set16–30

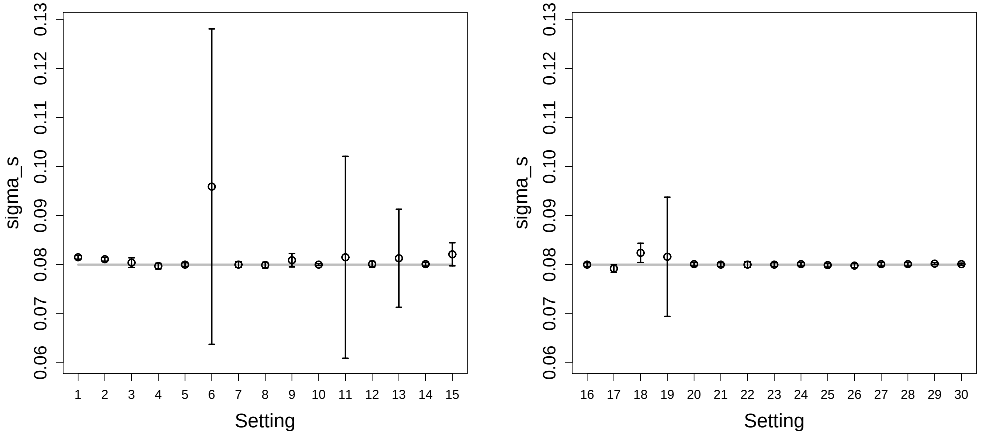
**Figure 14.14:** Graphical representation of the average point estimates with the 95% credible intervals for $M_1^*$ (True value indicated by the horizontal grey line).



(a) Set1–15

(b) Set16–30

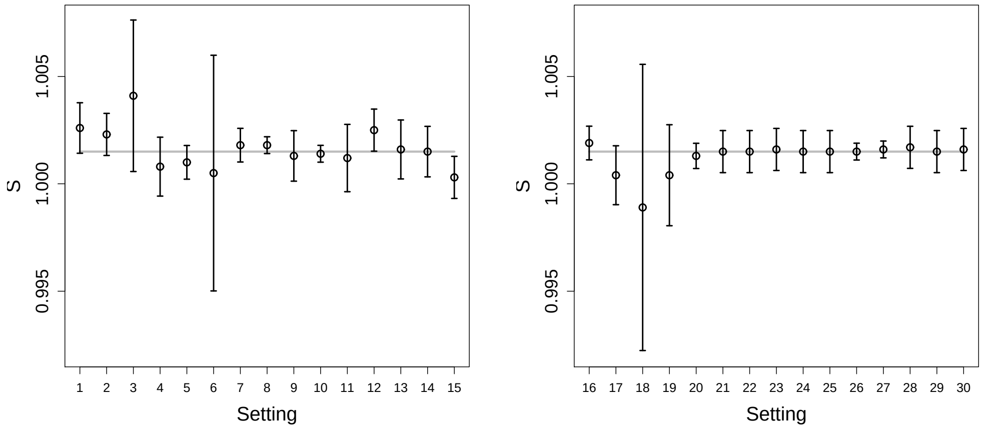
**Figure 14.15:** Graphical representation of the average point estimates with the 95% credible intervals for $\sigma$ (True value indicated by the horizontal grey line).



(a) Set1–15

(b) Set16–30

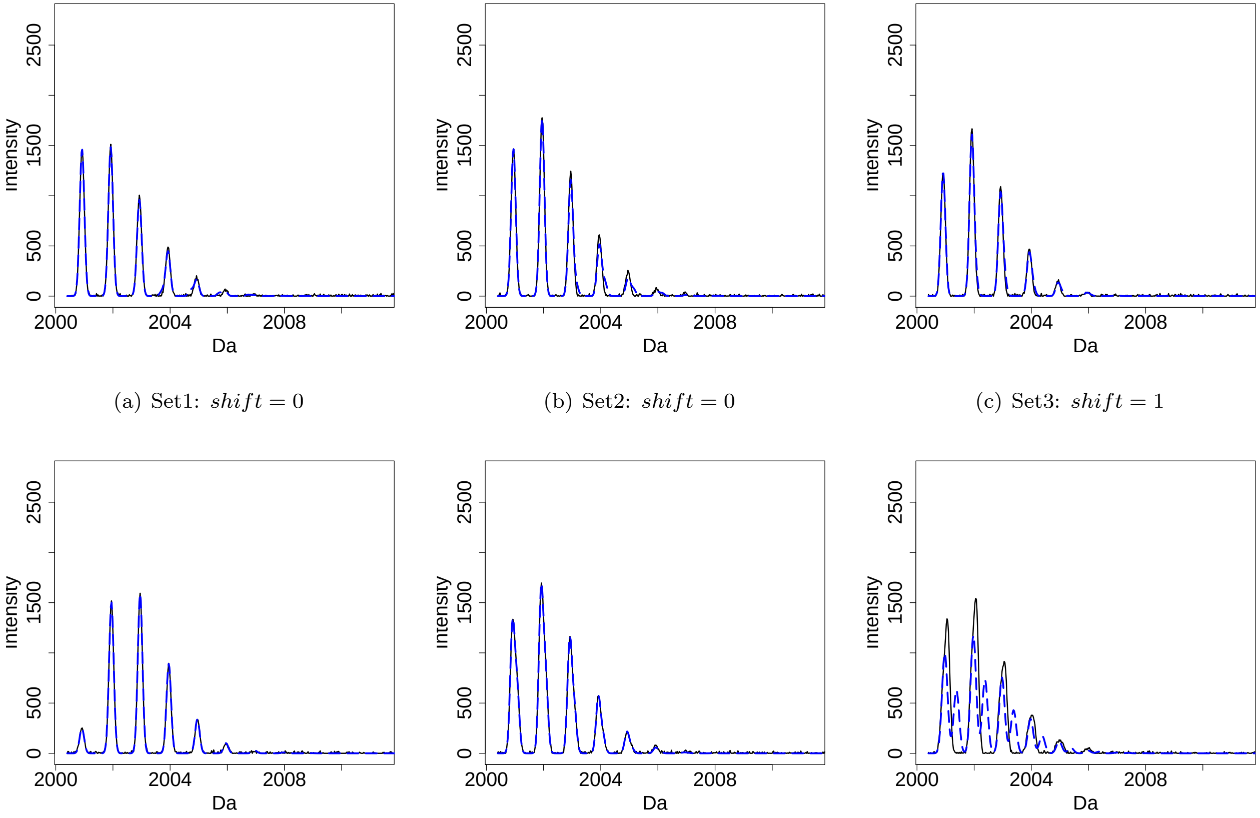
**Figure 14.16:** Graphical representation of the average point estimates with the 95% credible intervals for $\sigma_s$ (True value indicated by the horizontal grey line).



(a) Set1–15

(b) Set16–30

**Figure 14.17:** Graphical representation of the average point estimates with the 95% credible intervals for $S$ (True value indicated by the horizontal grey line).



(a) Set1: $shift = 0$

(b) Set2: $shift = 0$

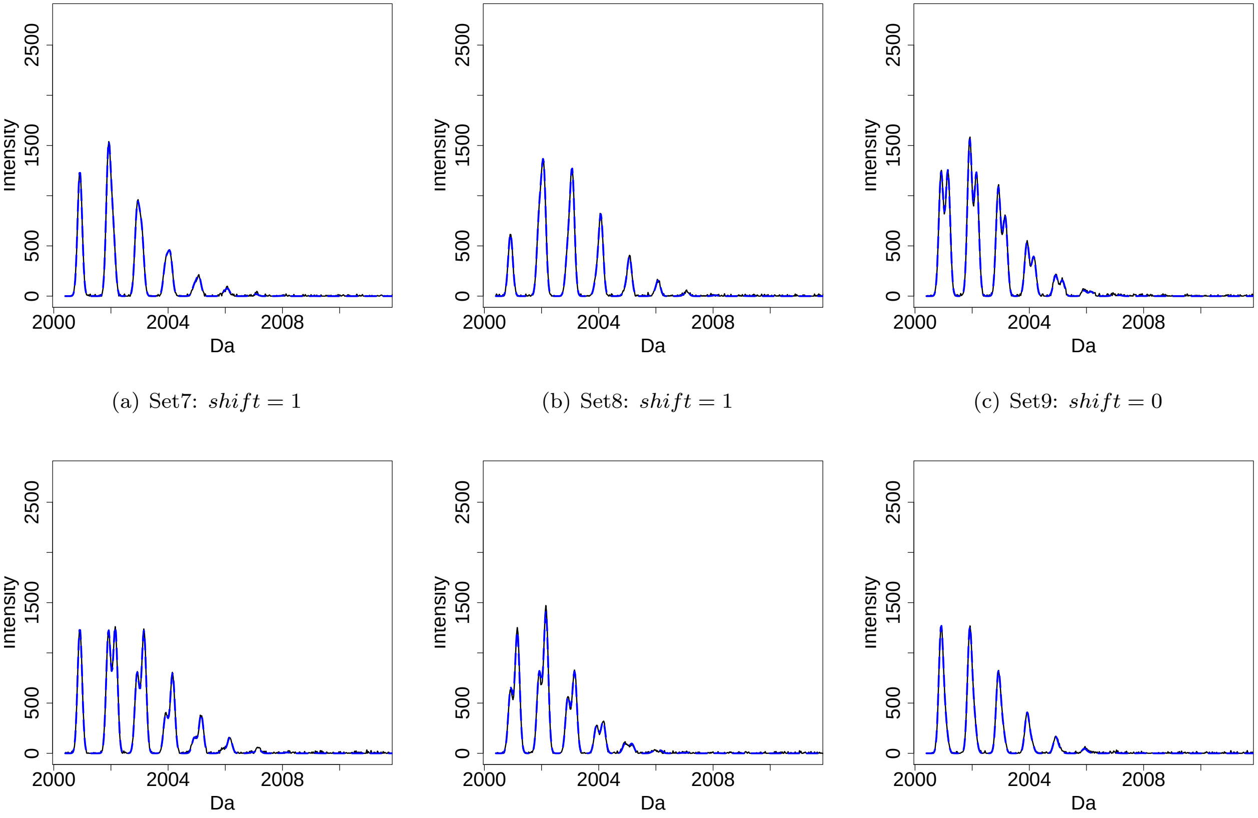
(c) Set3: $shift = 1$

(d) Set4: $shift = 1$

(e) Set5: $shift = 0$

(f) Set6: $shift = 0$

**Figure 14.18:** Observed (black solid line) versus predicted (blue dashed line) spectra (predicted intensity values calculated based on the point estimates of settings $1 \sim 6$.



(a) Set7: $shift = 1$

(b) Set8: $shift = 1$

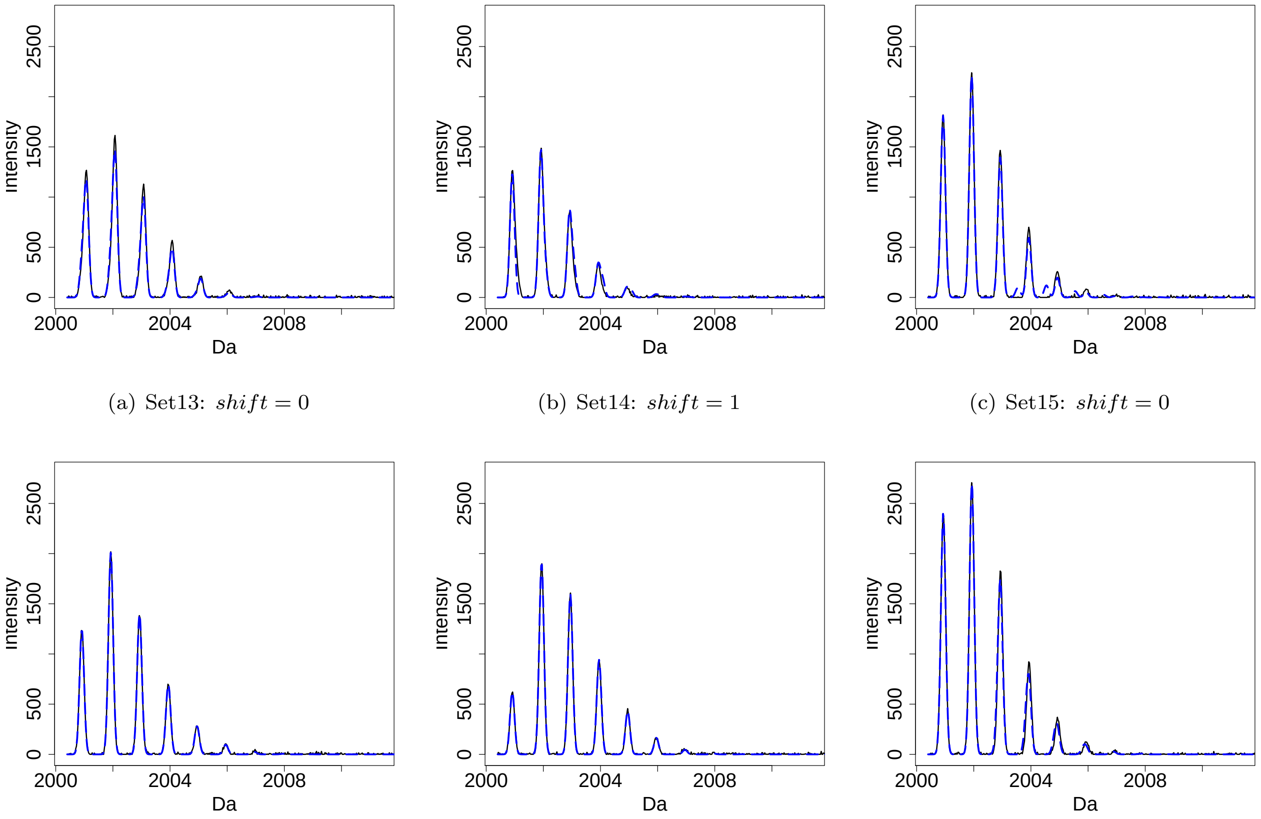
(c) Set9: $shift = 0$

(d) Set10: $shift = 1$

(e) Set11: $shift = 0$

(f) Set12: $shift = 0$

**Figure 14.19:** Observed (black solid line) versus predicted (blue dashed line) spectra (predicted intensity values calculated based on the point estimates of settings $7 \sim 12$.



(a) Set13: $shift = 0$

(b) Set14: $shift = 1$

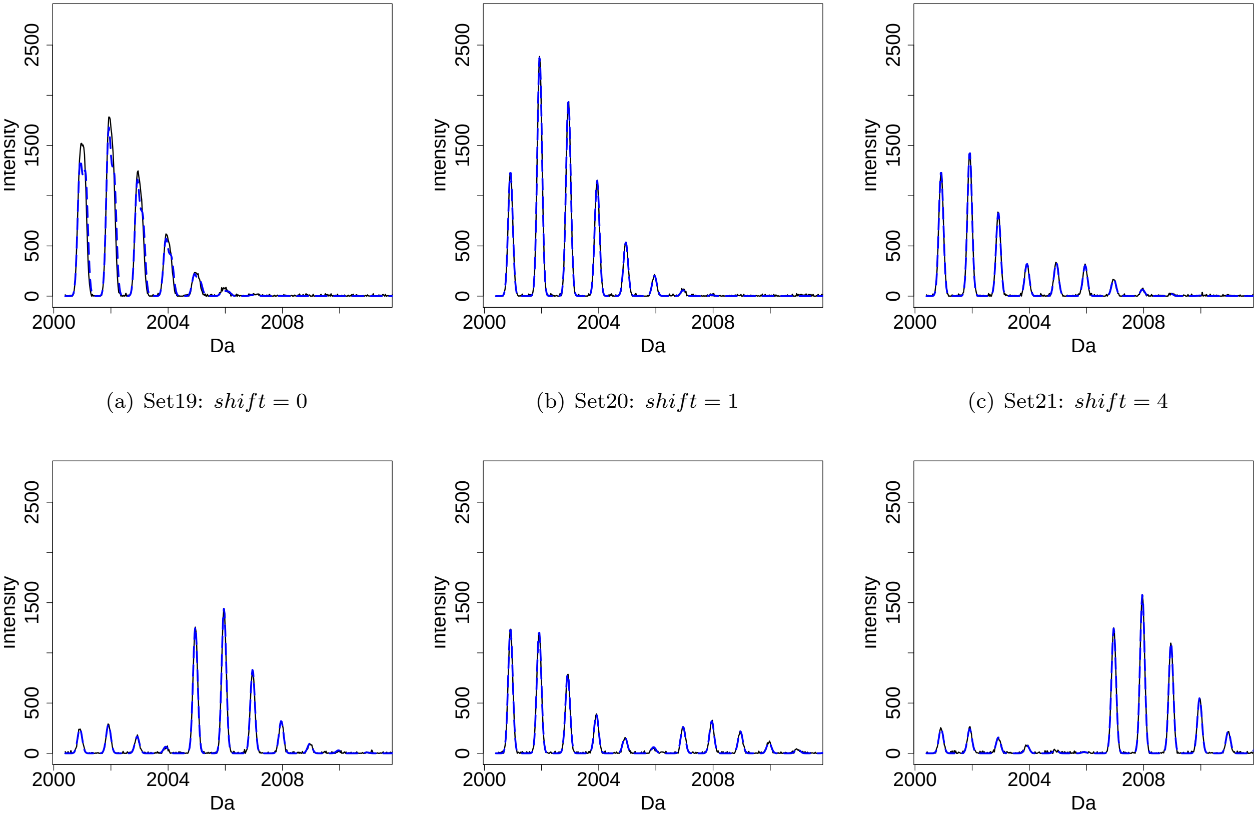
(c) Set15: $shift = 0$

(d) Set16: $shift = 1$

(e) Set17: $shift = 1$

(f) Set18: $shift = 0$

**Figure 14.20:** Observed (black solid line) versus predicted (blue dashed line) spectra (predicted intensity values calculated based on the point estimates of settings $13 \sim 18$.



(a) Set19: $shift = 0$

(b) Set20: $shift = 1$

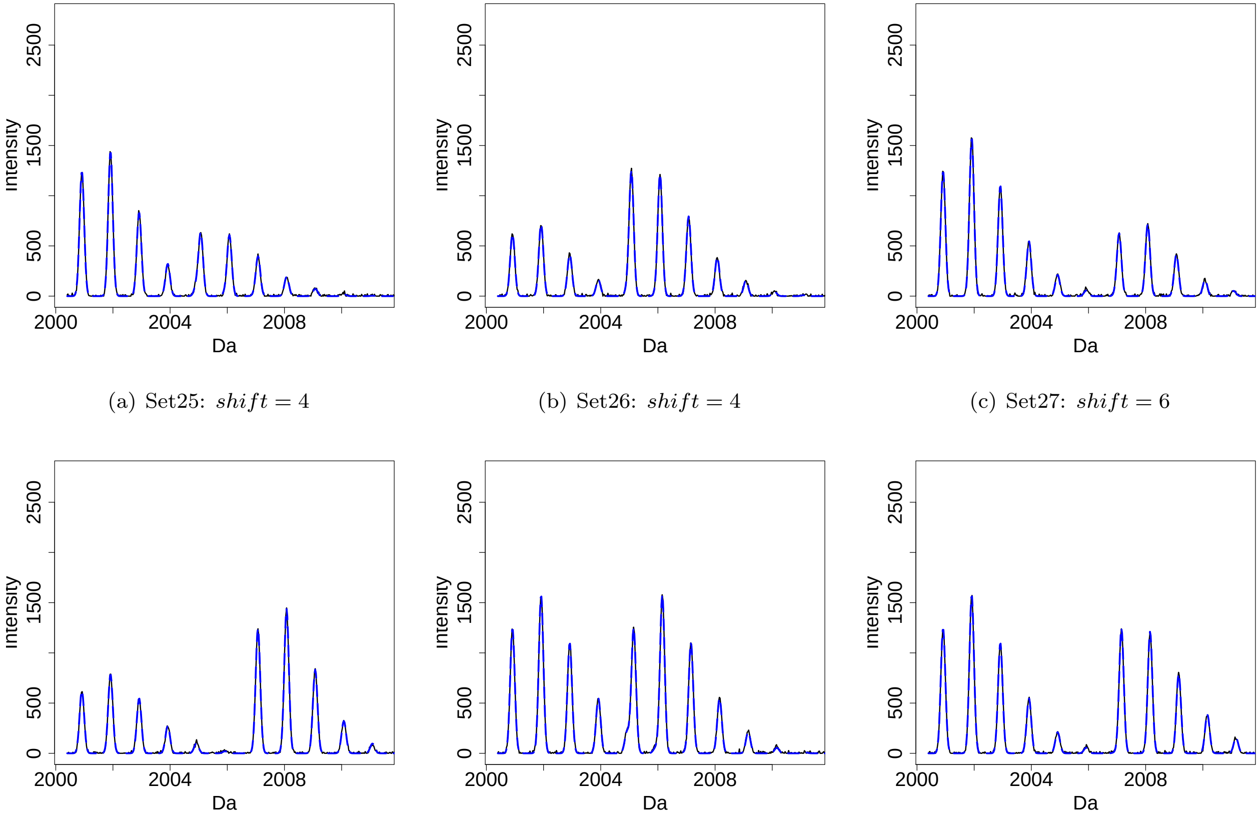
(c) Set21: $shift = 4$

(d) Set22: $shift = 4$

(e) Set23: $shift = 6$

(f) Set24: $shift = 6$

**Figure 14.21:** Observed (black solid line) versus predicted (blue dashed line) spectra (predicted intensity values calculated based on the point estimates of settings $19 \sim 24$.



(a) Set25: $shift = 4$

(b) Set26: $shift = 4$

(c) Set27: $shift = 6$

(d) Set28: $shift = 6$

(e) Set29: $shift = 4$

(f) Set30: $shift = 6$

**Figure 14.22:** Observed (black solid line) versus predicted (blue dashed line) spectra (predicted intensity values calculated based on the point estimates of settings $25 \sim 30$.

# Chapter 15

# Concluding remarks and directions for future research

Proteomics has gained a growing amount of attention as it plays an irreplaceable role in the pharmaceutical and biological applications. This dissertation is devoted to the statistical modeling approaches for the application of proteomics based mass spectrometry data. Some case studies have been addressed, considering both labeled and label-free experiments.

## Modeling of enzymatic $^{18}$O-labeled mass spectra

A part of the dissertation focused on the analysis of enzymatic $^{18}$O-labeled mass spectra data. We considered models for both the stick and shape representations of the spectrum.

As compared with the existing methods (Mirgorodskaya *et al.* 2000, Rao *et al.* 2005, López-Ferrer *et al.* 2006, Eckel-Passow *et al.* 2006, Ramos-Fernández *et al.* 2007), our method does not require an additional experimental step. In addition, our method provides extensions in several ways:

- It accounts for the possible presence of all the oxygen isotopes in the heavy-oxygen water.

- It incorporates the estimation of the isotopic distribution parameters, avoiding bias introduced by using a fixed average distribution.

- It is a unified modeling framework, in which all parameters of interest are simultaneously estimated from the data. It can easily accommodate different parameterizations, and provide necessary estimates of precision.
- It incorporates the heteroscedastic nature by using a mean-dependent variance function.
- It allows for the estimation of biological and/or technical variability of the mass spectra, by including random effects.
- It can be implemented in both the frequentist and Bayesian framework.

The analysis of our modeling approach justified the advantage of the shape model than the models for the stick representation since the former retained the full information of the MS data and turned out to produce more precise parameter estimates. Furthermore, the model, which takes into account the heteroscedastic nature of the MS data gave more precise parameter estimates, than the homoscedastic model.

The relative abundance of the labeled and unlabeled peptide samples, in general, was more precisely estimated when the labeling was more complete. The isotopic distribution parameters and the oxygen incorporation rate parameter of the labeling step were better estimated when the labeled sample was more abundant. This is because the labeled peptide sample provides more information for the estimation of these two sets of parameters.

In the simulations, we encountered a non-estimability issue for the parameter of oxygen incorporation rate when the labeling was more complete. The problem can be solved by performing a two-stage analysis.

For future research, one can think of the inclusion of informative priors for the Bayesian model. This should yield more precision gain, in terms of the parameter estimation. In particular, the prior information exists for the isotopic distributions (see Part III of the dissertation). Moreover, the shape-representation model can be formulated in the Bayesian framework, by including random effects to capture the technical and/or biological variability of the MS data. In such case, the Bayesian model is expected to be numerically simpler than the frequentist model, as for the latter, a one-stage analysis for such a complicated model would likely to be infeasible. On the other hand, the implementation of an automated processing of an MS experiment can also be conducted, as the developed modeling approach is a fast tool, capable of handling the high-throughput MS data.

## The quantification of overlapping peptides in MALDI-TOF mass spectra

Another part of the dissertation deals with the analysis of the overlapping peptides in MALDI-TOF mass spectra. In the analysis, we assumed unknown chemical compositions (masses) of the peptides. We considered both stick and shape representations of the spectrum.

For the stick representation, the estimation, especially for the isotopic ratios, was biased, due to the use of summary statistics of the data. Moreover, some strong assumptions have to be made for one to feasibly work with the stick representation.

The modeling approach based on the shape representation avoids the aforementioned limitations. A Bayesian (mixture) model, however, produced biased estimates and an under-estimation of parameter uncertainty. The Bayesian model averaging approach becomes a solution to tackle the problem. Provided the mass difference of the overlapping peptides is at least equal to a half of the width of an isotopic peak (around 0.16 Da), the Bayesian model averaging turned out to be an effective method to quantify the overlapping peptides, when masses of these peptides were unknown.

Although the proposed method focuses on the application to singly-charged MALDI-TOF mass spectrum, it can be modified to apply also for the multiply-charged mass spectrum by a modification of the expression for the mean structure of the model and the prior distributions for the corresponding parameters. The proposed modeling approach, assuming unknown masses of the overlapping peptides, can be modified for the application to the cases when the masses are known. In such a case, the masses of the peptides in the model can be fixed with known values and the model simplifies.

The validity of the approach is conditional on the pre-processing algorithm. To be more specific, we assume that the peptide peak features were correctly found. Based on the pre-processing algorithm proposed by Valkenborg *et al.* (2009), this means that all the peptide peak features have to be abundant enough to be distinguished from the noise.

Some further extensions can be considered. For instance, the estimation of the number of overlapping peptides can be considered by adding an additional step of the model selection procedure. In particular, the number of overlapping peptides can be estimated based on the model selection criteria for models with different number of overlapping peptides. In addition, the feature (peptide peaks) finding algorithm can include an additional step, by comparing the average isotopic distribution with the ones estimated from a certain ‘cluster’ of peaks. Strong agreement of the estimated ones with the average isotopic distribution can be viewed as a measure to distinguish ‘clusters’ of features from noise.

# Bibliography

Antoniak, C.E. (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, **2(6)**, 1152–1174.

Bates, D.M. and Watts, D.G. (1988) *Nonlinear regression analysis and its applications.* New York: John Wiley.

Belle, A., Tanay, A., Bitincka, L., Shamir, R. and K.OShea E. (2006). Quantification of protein half-lives in the budding yeast proteome. *Proceedings of the National Academy of Sciences*, **103(35)**, 1300413009.

Blei, D.M. and Jordan, M.I. (2006) Variational inference for Dirichlet process mixtures. *International Society for Bayesian Analysis*, **1**, 121–144.

Bordes, L., Chauveau, D. and Vandekerkhove, P. (2006) An EM algorithm for a semiparametric mixture model. *http://hal.archives-ouvertes.fr/docs/00/05/77/50/PDF/Chauveau_EMSP.pdf.*

Box, G.E.P. and Tiao, G.C. (1992) *Bayesian inference in statistical analysis.* New York: Jonh Wiley.

Breen, E.J., Hopwood, F.G., Williams, K.L. and Wilkins, M.R. (2000) Automatic Poisson peak harvesting for high throughput protein identification. *Electrophoresis*, **21**, 2243–2251.

Burnham, K.P. and Anderson D.R. (1998) *Model selection and inference: a practical information-theoretic approach.* New York: Springer-Verlag.

Burnham, K.P. and Anderson, D.R. (2002) *Model selection and multimodel inference: a practical information-theoretic approach.* New York: Springer.

Carlin, B.P. and Louis, T.A. (1998) *Bayes and empirical Bayes methods for data anlsysis.* London: Chapman & Hall/CRC.

Carlin, B.P. and Louis, T.A. (2009) *Bayesian methods for data analysis.* London: Chapman & Hall/CRC.

Carroll, R.J. and Ruppert, D. (1998) *Transformation and weighting in regression.* London: Chapman & Hall/CRC Press.

Clayton, D. and Rasbash, J. (1999) Estimation in large crossed random-effect models by data augmentation. *Journal of Royal Statistical Society*, **162(3)**, 425–436.

Congdon, P. (2003) *Applied Bayesian Modeling.* New York: John Wiley & Sons, Ltd.

Coombes, K., Tsavachidis, S., Morris, J., Baggerly, K., Hung, M. and Kuerer, H. (2005) Improved peak detection and quantification of mass spectrometry data acquired from seldi denoising spectra with the undecimated discreta wavelet transform. *Proteomics*, **5(16)**, 4107-4117.

Davidian, M. and Giltinan, D.M. (1995) *Nonlinear models for repeated measurement data.* London: Chapman & Hall/CRC Press.

Davidian, M. (2009) Lecture notes: nonlinear models for univariate and multivariate response. *Department of Statistics, North Caroline State University.*

Dhingraa, V., Gupta, M. and Fu, Z.F. (2005) New frontiers in proteomics research: a perspective. *International Journal of Pharmaceutics*, **299(1–2)**, 1-18.

Diebolt, J. and Robert, C.P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Royal Statistical Society*, **56(2)**, 363–375.

Dijkstra, M., Roelofsen, H., Vonk, R.J. and Jansen, R.C. (2006) Peak quantification in surface-enhanced laser desorption/ionization by using mixture models. *Proteomics*, **6**, 5106–5116.

Eckel-Passow, J.E., Oberg, A.L., Therneau, T.M., Mason, C.J., Mahoney, D.W., Johnson, K.L., Olson, J.E. and Bergen H.R. 3rd. (2006) Regression analysis for comparing protein samples with $^{16}$O/ O18 stable-isotope labelled mass-spectrometry. *Bioinformatics*, **22**, 2739–2745.

Eicher, T., Lenkoski, A. and Raftery, A.E. (2009) Bayesian model averaging and endogeneity under model uncertainty: an application to development determinants. *Working papers from University of Washington, Department of Economics*, UWEC-2009-19.

Escobar, M.D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *Journal of America Statistical Association*, **90(430)**, 577-588.

Faes, C., Aerts, M., Geys, H. and Molenberghs, G. (2007) Model averaging using fractional polynomials to estimate a safe level of exposure. *Risk Analysis*, **27(1)**, 111–123.

Fenyö, D. and Beavis, R.C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical Chemistry*, **75(4)**, 768–774.

Frühwirth-Schnatter, S. (1992) Data augmentation and dynamic linear models. *http://statmath.wu-wien.ac.at/*.

Gay, S., Binz, P.A., C., Hochstrasser, D.F. and Appel, R.D. (1999) Modeling peptide mass fingerprinting data using the atomic composition of peptides. *Electrophoresis*, **20**, 3527–3534.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004) *Bayesian data analysis.* London: Chapman & Hall/CRC.

Gevaert, K., Goethals, M., Martens, L., Van Damme, J., Staes, A., Thomas, G.R. and Vandekerckhove, J. (2003) Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted n-terminal peptides. *Nature Biotechnology*, **21**, 566–569.

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to calculating posterior moments. *Bayesian statistics, 4th. eds.* Oxford: Oxford University Press.

Givens, G.H., Smith, D.D. and Tweedie, R.L. (1997) Publication bias in Meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science*, **12(4)**, 221–250.

Green, P.J. (1995) Reversible jump Markov Chain Monte Carlo computation and Bayeisan model determination. *Biometrika*, **82(4)**, 711–732.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H. and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, **17**, 994–999.

Hartley, H.O. (1961) The modified Gauss-Newton method for the fitting of non-linear regression functions by least squares. *Technometrics*, **3**, 269–280.

Hastings, C.A., Norton, S.M. and Roy, S. (2002) New algorithm for processing and peak detection in liquid chromatography/mass spectrometry data. *Rapid Comunications in Mass Spectrometry*, **16**, 462–467.

Hazen, G.B. (1992) Stochastic trees: a new technique for temporal medical decision modeling. *Medical Decision Making*, **12**, 163-178.

Hazen, G.B. (1993) Factored stochastic trees: a tool for solving complex temporal medical decision models. *Medical Decision Making*, **13**, 227-236.

Hazen, G.B. and Pellissier, J.M. (1996) Recursive utility for stochastic trees. *Operations Research*, **44**, 788-809.

Hazen, G.B., Pellissier, J.M. and Sounderpandian, J. (1998) Stochastic-tree models in medical decision making. *Interfaces*, **28**, 64-80.

Hazen, G.B. (2000) Preference factoring for stochastic trees. *Management Science*, **46**, 389-403.

Higham, N.J. (2005) The scaling and squaring method for the matrix exponential revisited. *Society for Industrial and Applied Mathematics*, **26(4)**, 1179–1193.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999) Bayesian model averaging: a tutorial. *Statistical Science*, **14(4)**, 382-417.

Imai, K. and Van Dyk, D.A. (2005) A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*, **124**, 311-334.

Ishwaran, H. and James, L.F. (2001) Gibbs sampling methods for stick-breaking priors. *Journal of American Statistical Association*, **96(453)**, 161–173.

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J. and Thiébaut, R. (2007) Robustness of the linear mixed model to misspecified error distribution. *Computational statistics and data analysis*, **51(10)**, 5142–5154.

Jurisica, I. and Wigle, D. (2006) *Knowledge discovery in proteomics.* Boca Raton: CRC Press.

Kempka, M., Sjödahl, J., Björk, A. and Roeraade, J. (2004) Improved method for peak picking in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, **18**, 1208–1212.

Lange, E., Gröpl, C., Reinert, K., Kohlbacher, O. and Hildebrandt, A. (2006) High-accuracy peak picking of proteomics data using wavelet techniques. *Pacific Symposium on Biocomputing*, **11**, 243–254.

Lewis, J.K., Wei, J. and Siuzdak, G. (2000) Matrix-assisted laser desorption/ionization mass spectrometry in peptide and protein analysis. *Encyclopedia of Analytical Chemistry*, **R.A. Meyers (Ed.)**, 5880-5894.

Lin, T.I., Lee, J.C. and Ho, H.J. (2006) On fast supervised learning for normal mixuture models with missing information. *Journal of the Pattern Recognition Society*, **39**, 1177–1187.

Listgarten, J. and Emili, A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Molecular & Cellular Proteomics*, **4**, 419–434.

Liu, J.S. and Wu, Y.N. (1999) Parameter expansion for data augmentation. *Journal of American Statistical Association*, **94(448)**, 1264–1274.

López-Ferrer, D., Ramos-Fernández, A., Martinez-Bartolomé, S., Garca-Ruiz, P. and Vázquez, J. (2006) Quantitative proteomics using $^{16}$O/$^{18}$O labelling and linear ion trap mass spectrometry. *Proteomics*, **6**, S4–S11.

Lunn, D. (2004) WinBUGS Differential Interface – worked examples. *Imperial College of Medicine, London, UK.*

Marco, V.B.D. and Bombi, G.G. (2001) Mathematical functions for the representation of chromatographic peaks. *Journal of Chormatography A*, **931**, 1-30.

Mathews, C. K., van Holde, K. E. and Ahern, K. G. (1999) *Biochemistry (3rd edn.)*. Addison Wesley Longman: San Francisco.

McLachlan, G.J. and Peel, D. (2000) *Finite mixture models.* New York: John Wiley.

McLachlan, G.J. and Basford, K.E. (1988) *Mixture models: inference and applications to clustering.* New York: Marcel Dekker, Inc.

Mirgorodskaya, O.A., Kozmin, Y.P., Titov, M.I., Körner, R., Sönksen, C.P. and Roepstorff, P. (2000) Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using $^{18}$O-labelled internal standards. *Rapid Comunications in Mass Spectrometry*, **14**, 1226–1232.

Miyagi, M. and Rao, K.C. (2007) Proteolitic $^{18}$O-labeling strategies for quantitative proteomics. *Mass Spectrometry Reviews*, **26**, 121–136.

Molenberghs, G. and Verbeke, G. (2000) *Linear mixed models for longitudinal data.* New York: Springer.

Molenberghs, G. and Verbeke, G.(2005) *Models for discrete longitudinal data.* New York: Springer.

Morris, J.S., Brown, P.J., Baggerly, K.A. and Coombes, K.R. (2006) Analysis of mass spectrometry data using Bayesian wavelet-based functional mixed models. *Working paper, http://www.bepress.com/mdandersonbiostat/paper22.*

Namata, H., Aerts, M., Faes, C. and Teunis, P. (2008) Model averaging in microbial rish assessment using fractional polynomials. *Risk Analysis*, **2008**, *28(4)*, 891–905.

Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W. (1996) *Applied LInear Statistical Models (Fourth Edition).* Boston, Mass: McGraw-Hill.

Plummer, M. (2004) JAGS version 1.0.3 manual. *http://www-ice.iarc.fr/∼martyn/software/jags/jags_user_manual.pdf.*

Qian, W.J., Monroe, M.E., Liu, T., Jacobs, J.M., Anderson, G.A., Shen, Y.F., Moore, R.J., Anderson, D.J., Zhang, R., Calvano, S.E., Lowry, S.F., Xiao, W.Z., Moldawer, L.L., Davis, R.W., Tompkins, R.G., Camp, D.G. and Smith, F.D. (2005) Quantitative proteome analysis of human plasma following in Vivo lipopolysaccharide administration using $^{16}$O/$^{18}$O labeling and the accurate mass and time tag approach. *Molecular & Cellular Proteomics*, **4**, 700–709.

Qu, P.P. and Qu, Y.S. (2000) A Bayesian approach to finite mixture models in bioassy via data augmentation and Gibbs sampling and its application to insecticide resistance. *Biometrics*, **56**, 1249–1255.

Ramos-Fernández, A., López-Ferrer, D. and Vázquez, J. (2007) Improved method for differential expression proteomics using trypsin-catalyzed $^{18}$O-labelling with a correction for labeling efficiency. *Molecular & Cellular Proteomics*, **6**, 1274–1286.

Rao, K., Carruth, R. and Miyagi, M. (2005) Proteolic $^{18}$O-labelling by peptidyl-lys metalloendopeptidase for comparative proteomics. *Journal of Proteome Research*, **4**, 507–514.

Richardson, S. and Green, P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components. *Journal of Royal Statistical Society: Series B (Methdological)*, **59(4)**, 731–792.

Rockwood, A.L. (1995) Relationship of fourier transforms to isotope distribution calculations. *Rapid Communications in Mass Spectrometry*, **9**, 103–105.

Rogers, S., Girolami, M., Kolch, W., Waters, K.M., Liu, T., Thrall, B. and Wiley, H.S. (2008) Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics*, **24**, 2894-2900.

Sandra, K., Verleysen, K., Labeur, L., D'Hondt, F., Thomas, G., Kas, K., Gevaert, K., Vandekerckhove, J. and Sandra, P. (2007) Combination of cofradic and high temperature-extended column lengh conventional liquid chromatography: A very efficient way to tackle complex protein samples, such as serum. *Journal of Separation Science*, **30**, 658–668.

Schafer, J.L. (1997) *Analysis of incomplete multivariate data.* London: Chapman & Hall.

Schulz-Trieglaff, O., Hussong, R., Gröpl, C., Hildebrandt, A. and Reinert, K. (2007) A fast and accurate algorithm for the quantification of peptides from mass spectrometry Data. Research in Computational Molecular Biology, **LNBI 4453**, 473–487.

Searle, S.R. (1982) *Matrix algebra useful for statistics.* New York: John Wiley.

Seber, G.A.F. and Wild, C.J. (1989) *Nonlinear regression.* New York: John Wiley.

Sekhar Rao, K.C., Palamalai, V., Dunlevy, J.R. and Miyagi, M. (2005) Peptide-lys metalloendopeptidase-catalyzed $^{18}$O labeling for comparative proteomics. *Molecular & Cellular Proteomics*, **4**, 1550–1557.

Senko, M., Beu, S. and McLafferty, F. (1995) Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distribution. *Journal of the American Society for Mass Spectrometry*, **6**, 229–233.

Staes, A., Demol, H., Van Damme, J., Martens, L., Vandekerckhove, J. and Gevaert, K. (2004) Global differential non-gel proteomics by quantitative and stable labeling of tryptic peptides with oxygen-18. *Journal of Proteome Research*, **3**, 786–791.

Stephens, M. (2000) Bayeisan analysis of mixture models with an unknown number of components - an alternative to resersible jump methods. *The Annals of Statistics*, **28(1)**, 40–74.

Storms, F., Van Der Heijden, R., Tjaden, U. and Van Der Greef, J. (2006) Considerations for proteolytic labeling-optimization of $^{18}$O incorporation and prohibiiton of back-exchange. *Rapid Communications in Mass Spectrometry*, **20**, 3491–3497.

Stoyanova, R., Kuesel, A.C. and Brown, T.R. (1995) Application of principle-component analysis for NMR spectral quantitation. *Journal of Magnetic Resonance, Series A*, **115**, 265–269.

Tanner, M.A. and Wong, W.H. (1987) The calculation of posterior distributions by data augmentation. *Journal of American Statistical Association*, **82(398)**, 528–540.

Titterington, D.M., Smith A.F.M. and Makov, U.E. (1985) *Statistical analysis of finite mixture distributions.* New York: John Wiley.

Valentine, S.J., Kulchania, M., Srebalus Barnes, C.A. and Clemmer, D.E. (2001) Multidimensional separations of complex peptide mixtures: a combined high-performance liquid chromatography/ion mobility/time-of-flight mass spectrometry approach. *International Journal of Mass Spectrometry*, **212(1–3)**, 97–109.

Valkenborg, D., Assam, P., Thomas, G., Krols, L., Kas, K. and Burzykowski, T. (2007) Using a Poisson approximation to predict the isotopic distribution of sulphur-containing peptides in a peptide-centric proteomic approach. *Rapid Communications in Mass Spectrometry*, **21**, 3387–3391.

Valkenborg, D. (2008) Ph.D. dissertation: *Statistical methods for the analysis of high-resolution mass spectrometry data.* I-BIOSTAT, Hasselt University, Belgium.

Valkenborg, D., Jansen, I. and Burzykowski, T. (2008) A model-based method for the prediction of the isotopic distribution of peptides. *Journal of the American Society for Mass Spectrometry*, **19**, 703–712.

Valkenborg, D., Van Sanden, S., Lin, D., Kasim, A., Zhu, Q., Haldermans, P., Jansen, I., Shkedy, Z., and Burzykowski, T. (2008) A Cross-Validation Study to Select a Classification Procedure for Clinical Diagnosis Based on Proteomic Mass Spectrometry. *Statistical Applications in Genetics and Molecular Biology: Competition on Clinical Mass Spectrometry based Proteomics Diagnostics. Volume 7, Issue 2, Article 12.*

Valkenborg, D., Thomas, G., Krols, L., Kas, K. and Burzykowski, T. (2009) A strategy for the prior processing of high-resolution mass spectral data obtained from high-dimensional combined fractional diagnoal chromatography. *Journal of Mass Spectrometry*, **44(4)**, 516–529.

Van Dyk, D.A. and Meng, X.L. (2001) The art of data augmentation. *Journal of Computational and Graphical Statistics*, **10(1)**, 1–50.

Van Sanden, S., Lin, D. and Burzykowski, T. (2007) Performance of classification methods in a microarray setting: a simulation study. *Technical Report With Implemented R Functions.*

Wang, Y.K., Quinn, D.F., Ma, Z.X. and Fu, E.W. (2002) Inverse labeling– mass spectrometry for the rapid identification of differentially expressed protein markers/targets. *Journal of Chormatography B*, **782**, 291-306.

Wang, Y.G. and Lin, X. (2005) Effects of variance-function misspecification in analysis of longitudinal data. *Biometrics*, **61**, 413-421.

Wang, Y., Zhou, X.B., Wang, H.H., Li, K., Yao, L.X. and Wong, S.T.C. (2008) Reversible jump MCMC approach for peak identification for stroke SELDI mass spectrometry using mixture model. *Bioinformatics*, **24**, 407-413.

Wei, G.C.G. and Tanner, M.A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of American Statistical Association*, **85(411)**, 699–704.

Welton, N.J. and Ades A.E. (2005) Estimation of Markov chain transition probabilities and rates from fully and partially observed data: uncertanity propagation, evidence synthesis, and model calibration. *Medical Decision Making*, **25**, 633-645.

Yasui, Y., McLerran, D., Adam, B.L., Winget, M., Thornquist, M. and Feng, Z. (2003) An automated peak identification/calibration procedure for high-dimension protein measures from mass spectrometers. *Journal of Biomedicine and Biotechnology*, **4**, 242-248.

Yergey J.A. (1983) A general approach to calculating isotopic distributions for mass spectrometry. *International Journal of Mass Spectrometry and Ion Physics*, **52**, 337–349.

Zhang, N.S., Wei, B.C. and Lin J.G. (2005) Generalized nonlinear models and variance function estimation. *Computational Statistics and Data Analysis*, **48**, 549-570.

Zhu, M. and Lu, Y. (2004) The Counter-intuitive non-informative prior for the Bernoulli family. *Journal of Statistics Education*, **Volume 12**, Number 2.

Zhu, Q., Valkenborg, D. and Burzykowski, T. (2010) A Markov-chain-based heteroscedastic regression model for the analysis of high-resolution enzymatically $^{18}$O-labeled mass spectra. *Journal of Proteome Research,*, **9(5)**, 2669-2677.

## Appendix A

# Simulation results of the rrequentist heteroscedastic model for enzymatically $^{18}$O-labeled mass spectra

**Table A.1:** Mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$ and mean model based variance $S^2_{\text{mb}}$ for $\lambda$ for settings with **A** ratios (at 2001.05Da).

| $\lambda$ | $Q$ | $\sigma$ | $\bar{b}$ $(\times 1e-3)$ | | | $S^2_{\text{emp}}/S^2_{\text{mb}}$ $(\times 1e-7)$ | | | MSE $(\times 1e-7)$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LS | PLK | PL-GLS | LS | PLK | PL-GLS | LS | PLK | PL-GLS |
| 0.02 | 0.5 | 0.05 | -0.04 | -0.02 | -0.02 | 0.004/0.008 | 0.003/0.005 | 0.003/0.005 | 0.004 | 0.003 | 0.003 |
| | | 1.50 | 0.10 | -0.05 | -0.32 | 4.31/7.10 | 2.79/3.89 | 2.78/7.61 | 4.31 | 2.79 | 2.78 |
| | 1.0 | 0.05 | -0.006 | -0.02 | -0.02 | 0.002/0.002 | 0.001/0.001 | 0.001/0.002 | 0.002 | 0.001 | 0.001 |
| | | 1.50 | 1.05 | 1.43 | 1.32 | 1.56 /1.41 | 0.95/1.21 | 0.95/1.93 | 1.57 | 0.95 | 0.95 |
| | 2.0 | 0.05 | 0.006 | -0.001 | -0.001 | 0.0005/0.0005 | 0.0005/0.011 | 0.0005/0.0007 | 0.0005 | 0.0005 | 0.0005 |
| | | 1.50 | -0.50 | -0.40 | -0.41 | 0.45/0.31 | 0.42/0.45 | 0.41/0.82 | 0.45 | 0.42 | 0.41 |
| 0.04 | 0.5 | 0.05 | 0.03 | 0.04 | 0.04 | 0.03/0.03 | 0.02/0.03 | 0.02/0.09 | 0.03 | 0.02 | 0.02 |
| | | 1.50 | 3.50 | 2.43 | 2.39 | 24.28/28.89 | 19.47/24.65 | 19.48/30.08 | 24.47 | 19.56 | 19.57 |
| | 1.0 | 0.05 | -0.08 | -0.06 | -0.06 | 0.007/0.004 | 0.006/0.008 | 0.006/0.009 | 0.007 | 0.006 | 0.006 |
| | | 1.50 | 0.74 | 0.98 | 0.97 | 5.86/3.76 | 4.71/5.56 | 4.71/6.39 | 5.87 | 4.72 | 4.72 |
| | 2.0 | 0.05 | 0.006 | 0.0004 | 0.0004 | 0.0018/0.0011 | 0.0016/0.0028 | 0.0016/0.0025 | 0.0018 | 0.0016 | 0.0016 |
| | | 1.50 | -0.12 | -0.01 | -0.003 | 1.60/0.96 | 1.45/2.80 | 1.45/1.40 | 1.60 | 1.45 | 1.45 |
| 0.10 | 0.5 | 0.05 | 0.59 | 0.66 | 0.73 | 29.90/33.04 | 25.80/33.92 | 25.84/13.87 | 29.93 | 25.84 | 25.90 |
| | | 1.50 | -39.38 | -1.90 | -77.68 | 6970/14520 | 5407/6050 | 4081/5100 | 7125 | 5407 | 4684 |
| | 1.0 | 0.05 | 0.35 | 0.35 | 0.35 | 6.17/4.29 | 5.00/8.26 | 5.00/18.12 | 6.18 | 5.01 | 5.01 |
| | | 1.50 | -8.50 | 23.94 | -12.41 | 3142/26020 | 2170/3689 | 2441/6700 | 3149 | 2227 | 2457 |
| | 2.0 | 0.05 | 0.06 | 0.06 | 0.06 | 1.630/1.13 | 1.24/4.14 | 1.24/5.36 | 1.63 | 1.24 | 1.24 |
| | | 1.50 | 24.29 | 16.56 | 9.13 | 2184/10350 | 1731/2403 | 1326/3200 | 2243 | 1758 | 1334 |

**Table A.2:** Mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$ and mean model based variance $S^2_{\text{mb}}$ for $\lambda$ for settings with **E1** ratios.

| $\lambda$ | $Q$ | $\sigma$ | $\bar{b}$ $(\times 1e-3)$ | | | $S^2_{\text{emp}}/S^2_{\text{mb}}$ $(\times 1e-7)$ | | | MSE $(\times 1e-7)$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LS | PLK | PL-GLS | LS | PLK | PL-GLS | LS | PLK | PL-GLS |
| 0.02 | 0.5 | 0.05 | -0.03 | -0.01 | -0.006 | 0.007/0.012 | 0.006/0.013 | 0.006/0.004 | 0.007 | 0.006 | 0.006 |
| | | 1.50 | 5.37 | 4.11 | 3.66 | 6.99/11.39 | 6.28/9.25 | 6.34/8.98 | 7.11 | 6.35 | 6.39 |
| | 1.0 | 0.05 | -0.007 | -0.014 | -0.015 | 0.0024/0.0021 | 0.0017/0.0038 | 0.0016/0.0053 | 0.0024 | 0.0017 | 0.0017 |
| | | 1.50 | 0.34 | 1.19 | 1.05 | 2.12/1.88 | 1.48/3.23 | 1.48/5.17 | 2.12 | 1.48 | 1.48 |
| | 2.0 | 0.05 | -0.03 | -0.03 | -0.03 | 0.0007/0.0005 | 0.0006/0.0012 | 0.0006/0.0006 | 0.0007 | 0.0006 | 0.0006 |
| | | 1.50 | 0.22 | 0.67 | 0.58 | 0.58/0.42 | 0.52/0.57 | 0.52/0.71 | 0.58 | 0.52 | 0.52 |
| 0.04 | 0.5 | 0.05 | 0.02 | 0.01 | 0.01 | 0.038/0.044 | 0.037/0.064 | 0.037/0.046 | 0.039 | 0.037 | 0.037 |
| | | 1.50 | 4.14 | 2.96 | 2.94 | 37.08/42.86 | 33.89/68.74 | 34.07/56.00 | 37.36 | 34.03 | 34.21 |
| | 1.0 | 0.05 | 0.01 | 0.0007 | 0.0008 | 0.010/0.005 | 0.009/0.013 | 0.009/0.014 | 0.010 | 0.009 | 0.009 |
| | | 1.50 | -0.52 | -0.25 | -0.24 | 8.38/4.83 | 7.38/16.05 | 7.41/22.00 | 8.39 | 7.38 | 7.41 |
| | 2.0 | 0.05 | -0.008 | -0.009 | -0.009 | 0.0023/0.001 | 0.0022/0.0012 | 0.0022/0.008 | 0.0023 | 0.0022 | 0.0022 |
| | | 1.50 | -0.04 | 0.02 | 0.02 | 2.22/1.18 | 2.00/5.42 | 2.01/6.60 | 2.22 | 2.00 | 2.01 |
| 0.10 | 0.5 | 0.05 | 1.19 | 1.27 | 1.39 | 42.06/45.68 | 37.69/40.10 | 37.84/50.13 | 42.20 | 37.85 | 38.04 |
| | | 1.50 | -50.02 | -9.09 | -95.58 | 7964/14060 | 6924/7382 | 5367/6900 | 8215 | 6932 | 6281 |
| | 1.0 | 0.05 | 1.14 | 1.08 | 1.08 | 12.03/5.40 | 10.85/24.73 | 10.86/61.56 | 12.16 | 10.97 | 10.97 |
| | | 1.50 | -31.28 | 1.86 | -33.14 | 3558/30260 | 3591/4405 | 2945/4580 | 3656 | 3591 | 3055 |
| | 2.0 | 0.05 | -0.19 | -0.22 | -0.22 | 2.36/1.38 | 2.08/2.06 | 2.08/3.04 | 2.36 | 2.08 | 2.08 |
| | | 1.50 | 34.18 | 34.36 | 21.98 | 2599/16280 | 2328/3957 | 1754/1600 | 2716 | 2446 | 1803 |

**Table A.3:** Mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$ and mean model based variance $S^2_{\text{mb}}$ for $\lambda$ for settings with **E2** ratios.

| $\lambda$ | $Q$ | $\sigma$ | $\bar{b}$ $(\times 1e-3)$ | | | $S^2_{\text{emp}}/S^2_{\text{mb}}$ $(\times 1e-7)$ | | | MSE $(\times 1e-7)$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LS | PLK | PL-GLS | LS | PLK | PL-GLS | LS | PLK | PL-GLS |
| 0.02 | 0.5 | 0.05 | 0.07 | 0.05 | 0.06 | 0.005/0.010 | 0.004/0.010 | 0.004/0.004 | 0.005 | 0.004 | 0.004 |
| | | 1.50 | 0.74 | 0.09 | -0.27 | 4.28/9.45 | 3.93/9.10 | 3.95/3.93 | 4.28 | 3.93 | 3.95 |
| | 1.0 | 0.05 | -0.03 | -0.03 | -0.03 | 0.002/0.002 | 0.001/0.003 | 0.001/0.005 | 0.002 | 0.001 | 0.001 |
| | | 1.50 | 1.04 | 0.97 | -0.58 | 1.87/1.37 | 1.34/2.30 | 1.58/4.56 | 1.88 | 1.35 | 1.58 |
| | 2.0 | 0.05 | 0.01 | 0.0002 | 0.0001 | 0.00062/0.0004 | 0.00058/0.0012 | 0.00058/0.0007 | 0.00062 | 0.00058 | 0.00058 |
| | | 1.50 | -0.44 | -0.06 | -0.08 | 0.55/0.37 | 0.52/0.63 | 0.52/0.71 | 0.55 | 0.52 | 0.52 |
| 0.04 | 0.5 | 0.05 | 0.05 | 0.05 | 0.05 | 0.031/0.038 | 0.030/0.065 | 0.030/0.028 | 0.031 | 0.030 | 0.030 |
| | | 1.50 | 2.57 | 2.33 | 2.33 | 27.58/35.91 | 26.12/33.78 | 26.12/29.00 | 27.68 | 26.21 | 26.21 |
| | 1.0 | 0.05 | -0.03 | -0.04 | -0.04 | 0.009/0.005 | 0.007/0.002 | 0.007/0.004 | 0.009 | 0.007 | 0.007 |
| | | 1.50 | 0.03 | -0.24 | -0.27 | 7.28/4.25 | 6.49/9.70 | 6.50/8.34 | 7.28 | 6.49 | 6.50 |
| | 2.0 | 0.05 | -0.004 | -0.003 | -0.003 | 0.0022/0.0013 | 0.0021/0.0020 | 0.0021/0.0048 | 0.0022 | 0.0021 | 0.0021 |
| | | 1.50 | 0.27 | 0.50 | 0.51 | 2.30/1.12 | 2.17/3.09 | 2.17/4.33 | 2.30 | 2.18 | 2.18 |
| 0.10 | 0.5 | 0.05 | -0.11 | -0.14 | -0.03 | 32.61/38.09 | 31.34/34.41 | 31.45/42.40 | 32.61 | 31.34 | 31.45 |
| | | 1.50 | -40.12 | -5.40 | -10.36 | 7343/11060 | 3338/3644 | 3909/4780 | 7504 | 3341 | 3920 |
| | 1.0 | 0.05 | 0.24 | 0.33 | 0.33 | 9.15/4.85 | 8.61/7.57 | 8.61/8.40 | 9.15 | 8.62 | 8.62 |
| | | 1.50 | -32.05 | -2.55 | -42.97 | 3404/31050 | 2617/3199 | 2540/3650 | 3506 | 2618 | 2725 |
| | 2.0 | 0.05 | -0.10 | 0.002 | 0.002 | 2.10/1.33 | 1.90/2.22 | 1.90/4.12 | 2.10 | 1.90 | 1.90 |
| | | 1.50 | 18.86 | 23.60 | 13.23 | 2132/9681 | 1286/2245 | 1774/3220 | 2168 | 1341 | 1791 |

**Table A.4:** Mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$ and mean model based variance $S^2_{\text{mb}}$ for $Q$ for settings with **A** ratios (at 2001.05Da).

| $Q$ | $\lambda$ | $\sigma$ | $\bar{b}$ ($\times 1e-5$) | | | $S^2_{\text{emp}}/S^2_{\text{mb}}$ ($\times 1e-6$) | | | MSE ($\times 1e-6$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LS | PLK | PL-GLS | LS | PLK | PL-GLS | LS | PLK | PL-GLS |
| 0.5 | 0.02 | 0.05 | 3.19 | 4.81 | 4.81 | 0.368/1.25 | 0.332/0.390 | 0.332/0.362 | 0.368 | 0.333 | 0.332 |
| | | 1.50 | 79.13 | 155.8 | 152.4 | 303.2/1152 | 271.9/363.0 | 272.1/333.8 | 303.4 | 272.5 | 272.7 |
| | 0.04 | 0.05 | -3.83 | -3.32 | -3.34 | 0.139/0.262 | 0.138/0.177 | 0.138/0.149 | 0.139 | 0.138 | 0.138 |
| | | 1.50 | 33.09 | 87.73 | 77.44 | 130.6/233.7 | 127.3/142.6 | 127.4/133.1 | 130.6 | 127.5 | 127.6 |
| | 0.10 | 0.05 | 0.67 | 0.02 | -0.22 | 0.118/0.154 | 0.118/0.138 | 0.118/0.115 | 0.118 | 0.118 | 0.118 |
| | | 1.50 | 755.3 | 683.2 | 843.4 | 68.25/125.2 | 64.86/78.46 | 60.05/110.5 | 82.51 | 76.53 | 77.83 |
| 1.0 | 0.02 | 0.05 | -3.26 | -2.52 | -2.51 | 0.724/0.623 | 0.571/0.655 | 0.571/0.666 | 0.725 | 0.572 | 0.572 |
| | | 1.50 | -11.50 | 12.14 | 14.22 | 701.7/563.6 | 602.9/586.3 | 602.7/587.2 | 701.7 | 602.9 | 602.8 |
| | 0.04 | 0.05 | 3.08 | 2.53 | 2.53 | 0.285/0.159 | 0.268/0.274 | 0.268/0.274 | 0.286 | 0.268 | 0.269 |
| | | 1.50 | 26.35 | 37.82 | 36.68 | 258.3/134.3 | 245.2/246.7 | 245.7/247.1 | 258.4 | 245.3 | 245.8 |
| | 0.10 | 0.05 | -1.19 | -0.96 | -0.97 | 0.224/0.098 | 0.215/0.216 | 0.215/0.216 | 0.224 | 0.215 | 0.215 |
| | | 1.50 | 316.8 | 232.8 | 310.6 | 146.6/87.81 | 146.8/186.4 | 141.1/204.7 | 156.6 | 152.2 | 150.7 |
| 2.0 | 0.02 | 0.05 | -1.00 | 0.41 | 0.41 | 1.66/0.917 | 1.53/1.41 | 1.53/1.62 | 1.67 | 1.53 | 1.53 |
| | | 1.50 | 160.8 | 128.3 | 129.0 | 1717/832.9 | 1587/1644 | 1582/1470 | 1727 | 1593 | 1589 |
| | 0.04 | 0.05 | -2.76 | -2.06 | -2.06 | 0.609/0.269 | 0.593/0.742 | 0.593/0.694 | 0.612 | 0.595 | 0.595 |
| | | 1.50 | -37.55 | -61.46 | -60.15 | 534.2/237.8 | 523.8/640.9 | 523.1/607.2 | 534.7 | 525.3 | 524.5 |
| | 0.1 | 0.05 | 0.30 | 0.42 | 0.42 | 0.567/0.204 | 0.514/0.817 | 0.514/0.534 | 0.567 | 0.514 | 0.514 |
| | | 1.50 | 88.14 | 70.52 | 83.48 | 407.0/180.5 | 399.3/631.5 | 400.4/488.0 | 410.2 | 401.3 | 403.2 |

**Table A.5:** Mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$ and mean model based variance $S^2_{\text{mb}}$ for $Q$ for settings with **E1** ratios.

| $Q$ | $\lambda$ | $\sigma$ | $\bar{b}$ ($\times 1e-5$) | | | $S^2_{\text{emp}}/S^2_{\text{mb}}$ ($\times 1e-6$) | | | MSE ($\times 1e-6$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LS | PLK | PL-GLS | LS | PLK | PL-GLS | LS | PLK | PL-GLS |
| 0.5 | 0.02 | 0.05 | -8.34 | -0.89 | -1.22 | 0.696/2.36 | 0.616/0.873 | 0.618/0.820 | 0.698 | 0.616 | 0.618 |
| | | 1.50 | -597.8 | -290.7 | -277.7 | 688.5/2092 | 625.5/794.0 | 631.4/700.3 | 697.4 | 627.7 | 633.3 |
| | 0.04 | 0.05 | -3.22 | -3.28 | -3.29 | 0.257/0.377 | 0.245/0.357 | 0.245/0.228 | 0.257 | 0.246 | 0.245 |
| | | 1.50 | -63.50 | 38.82 | 28.02 | 198.8/343.3 | 194.0/266.6 | 194.5/208.6 | 198.9 | 194.0 | 194.5 |
| | 0.10 | 0.05 | -3.66 | -2.97 | -3.38 | 0.168/0.204 | 0.163/0.189 | 0.163/0.166 | 0.169 | 0.163 | 0.163 |
| | | 1.50 | 969.0 | 907.6 | 1165 | 104.9/165.4 | 98.51/134.1 | 96.48/139.6 | 128.3 | 119.1 | 130.4 |
| 1.0 | 0.02 | 0.05 | -0.17 | 0.84 | 0.80 | 1.21/0.898 | 1.09/1.05 | 1.09/1.17 | 1.21 | 1.09 | 1.09 |
| | | 1.50 | -80.44 | -79.98 | -68.62 | 1099/820.3 | 1005/951.1 | 1010/1015 | 1099 | 1005 | 1011 |
| | 0.04 | 0.05 | 0.64 | 0.79 | 0.79 | 0.370/0.192 | 0.347/0.373 | 0.347/0.373 | 0.370 | 0.347 | 0.347 |
| | | 1.50 | 108.9 | 118.3 | 114.9 | 314.2/178.9 | 306.1/346.5 | 307.6/347.5 | 315.4 | 307.5 | 308.9 |
| | 0.10 | 0.05 | -3.59 | -3.41 | -3.42 | 0.344/0.122 | 0.332/0.284 | 0.332/0.284 | 0.345 | 0.334 | 0.334 |
| | | 1.50 | 557.9 | 419.8 | 491.5 | 202.9/209.8 | 203.8/244.5 | 204.3/260.1 | 234.0 | 221.4 | 228.4 |
| 2.0 | 0.02 | 0.05 | 1.38 | 1.82 | 1.82 | 2.43/1.22 | 2.23/4.05 | 2.23/2.30 | 2.43 | 2.23 | 2.23 |
| | | 1.50 | 141.5 | 63.89 | 68.76 | 2279/1093 | 2013/3031 | 2008/2057 | 2287 | 2015 | 2010 |
| | 0.04 | 0.05 | 0.09 | 0.05 | 0.04 | 0.820/0.327 | 0.796/0.903 | 0.796/0.844 | 0.820 | 0.796 | 0.796 |
| | | 1.50 | 13.03 | -5.03 | -1.68 | 700.2/293.8 | 675.7/712.0 | 677.1/761.0 | 700.2 | 675.7 | 677.1 |
| | 0.1 | 0.05 | -0.15 | 0.13 | 0.13 | 0.781/0.251 | 0.752/0.822 | 0.752/0.658 | 0.781 | 0.752 | 0.752 |
| | | 1.50 | 156.8 | 113.3 | 155.1 | 582.8/220.3 | 566.3/636.8 | 541.1/607.1 | 592.6 | 571.4 | 550.7 |

**Table A.6:** Mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$ and mean model based variance $S^2_{\text{mb}}$ for $Q$ for settings with **E2** ratios.

| $Q$ | $\lambda$ | $\sigma$ | $\bar{b}$ ($\times 1e-5$) | | | $S^2_{\text{emp}}/S^2_{\text{mb}}$ ($\times 1e-6$) | | | MSE ($\times 1e-6$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LS | PLK | PL-GLS | LS | PLK | PL-GLS | LS | PLK | PL-GLS |
| 0.5 | 0.02 | 0.05 | 0.81 | -5.21 | -5.27 | 0.556/1.73 | 0.453/0.431 | 0.454/0.669 | 0.556 | 0.454 | 0.455 |
| | | 1.50 | -175.2 | 26.55 | 29.16 | 561.0/1548 | 464.2/598.6 | 464.7/580.3 | 561.8 | 464.2 | 464.7 |
| | 0.04 | 0.05 | -7.00 | -6.84 | -6.86 | 0.184/0.297 | 0.177/0.279 | 0.177/0.184 | 0.186 | 0.178 | 0.178 |
| | | 1.50 | 5.68 | 33.48 | 21.81 | 168.9/266.9 | 160.6/179.9 | 160.7/166.2 | 168.9 | 160.6 | 160.7 |
| | 0.10 | 0.05 | 0.26 | 0.09 | -0.27 | 0.137/0.171 | 0.133/0.223 | 0.133/0.133 | 0.137 | 0.133 | 0.133 |
| | | 1.50 | 836.6 | 777.2 | 1042 | 81.22/129.3 | 75.30/41.55 | 72.09/183.4 | 98.72 | 90.40 | 89.22 |
| 1.0 | 0.02 | 0.05 | 5.27 | 4.79 | 4.78 | 0.884/0.809 | 0.838/0.874 | 0.838/0.957 | 0.887 | 0.840 | 0.841 |
| | | 1.50 | 64.26 | 106.3 | 102.2 | 883.0/734.9 | 833.1/909.3 | 816.1/878.2 | 883.4 | 834.2 | 817.1 |
| | 0.04 | 0.05 | 0.95 | 1.31 | 1.31 | 0.314/0.171 | 0.302/0.314 | 0.302/0.314 | 0.314 | 0.302 | 0.302 |
| | | 1.50 | -8.85 | 5.87 | 7.39 | 265.2/154.3 | 253.7/284.1 | 254.9/284.3 | 265.2 | 253.7 | 255.0 |
| | 0.10 | 0.05 | 0.66 | 0.42 | 0.41 | 0.257/0.108 | 0.257/0.235 | 0.257/0.235 | 0.257 | 0.257 | 0.257 |
| | | 1.50 | 472.5 | 416.0 | 484.9 | 152.8/96.69 | 146.0/203.1 | 145.3/207.2 | 175.1 | 163.3 | 166.8 |
| 2.0 | 0.02 | 0.05 | -2.11 | 0.31 | 0.32 | 1.97/1.06 | 1.91/4.02 | 1.91/2.08 | 1.97 | 1.91 | 1.91 |
| | | 1.50 | 120.4 | 78.75 | 83.60 | 2052/952.2 | 1928/1118 | 1927/1849 | 2058 | 1931 | 1930 |
| | 0.04 | 0.05 | -2.02 | -1.99 | -1.99 | 0.756/0.303 | 0.740/0.836 | 0.740/0.739 | 0.758 | 0.741 | 0.741 |
| | | 1.50 | 37.40 | 20.35 | 23.22 | 647.4/268.8 | 644.7/711.8 | 644.5/656.5 | 648.0 | 644.9 | 644.7 |
| | 0.1 | 0.05 | 1.38 | 0.97 | 0.97 | 0.522/0.230 | 0.487/0.518 | 0.487/0.559 | 0.522 | 0.488 | 0.488 |
| | | 1.50 | 184.6 | 176.8 | 188.0 | 419.0/205.9 | 391.3/402.3 | 394.8/505.1 | 432.7 | 403.8 | 408.9 |

**Table A.7:** Mean estimate $\bar{\theta}$, mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$ and mean model based variance $S^2_{\text{mb}}$ for $\theta$ for settings with **A** ratios (at 2001.05Da).

| $Q$ | $\lambda$ | $\sigma$ | $\bar{\theta}$ | | $\bar{b}(\times 1e-2)$ | | $S^2_{\text{emp}}/S^2_{\text{mb}}$ $(\times 1e-2)$ | | MSE$(\times 1e-2)$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | PLK | PL-GLS | PLK | PL-GLS | PLK | PL-GLS | PLK | PL-GLS |
| 0.5 | 0.02 | 0.05 | 0.5849 | 0.5850 | -2.51 | -2.50 | 0.405/0.610 | 0.404/0.601 | 0.428 | 0.426 |
| | | 1.50 | 0.5833 | 0.5840 | -2.79 | -2.67 | 0.453/0.544 | 0.449/0.523 | 0.481 | 0.474 |
| | 0.04 | 0.05 | 0.5858 | 0.5858 | -2.37 | -2.36 | 0.486/0.746 | 0.486/0.743 | 0.507 | 0.506 |
| | | 1.50 | 0.5888 | 0.5893 | -1.86 | -1.78 | 0.544/0.674 | 0.541/0.657 | 0.556 | 0.552 |
| | 0.10 | 0.05 | 0.5952 | 0.5966 | -0.793 | -0.571 | 0.617/0.880 | 0.652/0.867 | 0.620 | 0.653 |
| | | 1.50 | 0.5940 | 0.5939 | -0.993 | -1.02 | 0.514/0.685 | 0.521/0.695 | 0.518 | 0.524 |
| 1.0 | 0.02 | 0.05 | 0.5926 | 0.5939 | -1.23 | -1.02 | 0.536/0.690 | 0.577/0.675 | 0.541 | 0.581 |
| | | 1.50 | 0.5869 | 0.5873 | -2.18 | -2.11 | 0.487/0.633 | 0.485/0.622 | 0.504 | 0.501 |
| | 0.04 | 0.05 | 0.5936 | 0.5937 | -1.06 | -1.06 | 0.656/0.985 | 0.655/0.984 | 0.660 | 0.660 |
| | | 1.50 | 0.6040 | 0.6043 | 0.674 | 0.721 | 0.600/0.695 | 0.600/0.690 | 0.601 | 0.602 |
| | 0.10 | 0.05 | 0.5967 | 0.5967 | -0.545 | -0.542 | 0.702/1.14 | 0.702/0.913 | 0.703 | 0.703 |
| | | 1.50 | 0.5999 | 0.5978 | -0.015 | -0.364 | 0.736/1.13 | 0.731/1.15 | 0.736 | 0.732 |
| 2.0 | 0.02 | 0.05 | 0.6008 | 0.6008 | 0.139 | 0.140 | 0.530/0.742 | 0.530/0.742 | 0.530 | 0.530 |
| | | 1.50 | 0.5942 | 0.5943 | -0.970 | -0.942 | 0.485/0.652 | 0.485/0.649 | 0.488 | 0.488 |
| | 0.04 | 0.05 | 0.6015 | 0.6015 | 0.242 | 0.252 | 0.807/1.33 | 0.808/1.13 | 0.808 | 0.808 |
| | | 1.50 | 0.5897 | 0.5899 | -1.72 | -1.68 | 0.755/1.50 | 0.755/1.49 | 0.765 | 0.766 |
| | 0.1 | 0.05 | 0.5895 | 0.5896 | -1.75 | -1.74 | 0.881/0.972 | 0.879/0.935 | 0.892 | 0.890 |
| | | 1.50 | 0.5874 | 0.5880 | -2.11 | -2.01 | 0.849/2.14 | 0.849/2.12 | 0.865 | 0.859 |

**Table A.8:** Mean estimate $\bar{\theta}$, mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$ and mean model based variance $S^2_{\text{mb}}$ for $\theta$ for settings with **E1** ratios.

| $Q$ | $\lambda$ | $\sigma$ | $\bar{\theta}$ | | $\bar{b}(\times 1e-2)$ | | $S^2_{\text{emp}}/S^2_{\text{mb}}$ $(\times 1e-2)$ | | MSE$(\times 1e-2)$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | PLK | PL-GLS | PLK | PL-GLS | PLK | PL-GLS | PLK | PL-GLS |
| 0.5 | 0.02 | 0.05 | 0.5794 | 0.6023 | -3.43 | 0.388 | 0.598/1.31 | 1.95/3.27 | 0.640 | 1.95 |
| | | 1.50 | 0.5719 | 0.5900 | -4.69 | -1.66 | 0.564/0.975 | 1.60/1.93 | 0.643 | 1.61 |
| | 0.04 | 0.05 | 0.5932 | 0.5932 | -1.14 | -1.13 | 0.750/1.57 | 0.750/1.57 | 0.754 | 0.754 |
| | | 1.50 | 0.5807 | 0.5816 | -3.22 | -3.06 | 0.787/1.06 | 0.775/1.07 | 0.824 | 0.808 |
| | 0.10 | 0.05 | 0.5868 | 0.5869 | -2.20 | -2.18 | 0.943/3.17 | 0.943/3.17 | 0.961 | 0.960 |
| | | 1.50 | 0.5796 | 0.5799 | -3.41 | -3.35 | 0.980/3.20 | 0.958/3.06 | 1.02 | 0.999 |
| 1.0 | 0.02 | 0.05 | 0.5871 | 0.5984 | -2.15 | -0.264 | 0.726/1.72 | 1.46/1.71 | 0.742 | 1.47 |
| | | 1.50 | 0.5813 | 0.5907 | -3.12 | -1.56 | 0.744/1.76 | 1.27/1.67 | 0.779 | 1.28 |
| | 0.04 | 0.05 | 0.5930 | 0.5931 | -1.16 | -1.16 | 1.15/3.81 | 1.15/3.80 | 1.15 | 1.15 |
| | | 1.50 | 0.6032 | 0.6036 | 0.538 | 0.595 | 1.09/2.36 | 1.09/2.35 | 1.09 | 1.09 |
| | 0.10 | 0.05 | 0.5898 | 0.5898 | -1.70 | -1.69 | 1.09/3.86 | 1.09/3.86 | 1.10 | 1.10 |
| | | 1.50 | 0.6006 | 0.5998 | 0.103 | -0.032 | 1.32/3.78 | 1.30/3.79 | 1.32 | 1.30 |
| 2.0 | 0.02 | 0.05 | 0.6007 | 0.6008 | 0.121 | 0.128 | 0.928/1.84 | 0.928/1.84 | 0.928 | 0.928 |
| | | 1.50 | 0.5940 | 0.5955 | -1.01 | -0.752 | 0.943/1.52 | 0.869/2.30 | 0.947 | 0.871 |
| | 0.04 | 0.05 | 0.5915 | 0.5918 | -1.42 | -1.37 | 1.42/8.55 | 1.42/8.46 | 1.43 | 1.42 |
| | | 1.50 | 0.5956 | 0.5959 | -0.734 | -0.680 | 1.46/7.26 | 1.46/7.19 | 1.46 | 1.46 |
| | 0.1 | 0.05 | 0.5887 | 0.5888 | -1.89 | -1.87 | 1.79/1.91 | 1.79/1.91 | 1.81 | 1.81 |
| | | 1.50 | 0.5838 | 0.5847 | -2.70 | -2.55 | 1.68/1.08 | 1.70/1.56 | 1.70 | 1.72 |

**Table A.9:** Mean estimate $\bar{\theta}$, mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$ and mean model based variance $S^2_{\text{mb}}$ for $\theta$ for settings with **E2** ratios.

| $Q$ | $\lambda$ | $\sigma$ | $\bar{\theta}$ | | $\bar{b}(\times 1e-2)$ | | $S^2_{\text{emp}}/S^2_{\text{mb}}$ $(\times 1e-2)$ | | MSE$(\times 1e-2)$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | PLK | PL-GLS | PLK | PL-GLS | PLK | PL-GLS | PLK | PL-GLS |
| 0.5 | 0.02 | 0.05 | 0.5758 | 0.6137 | -4.03 | 2.28 | 0.594/1.32 | 3.01/2.24 | 0.653 | 3.03 |
| | | 1.50 | 0.5717 | 0.6075 | -4.72 | 1.25 | 0.639/1.34 | 2.93/1.28 | 0.719 | 2.94 |
| | 0.04 | 0.05 | 0.5834 | 0.5835 | -2.77 | -2.75 | 0.757/1.96 | 0.757/1.95 | 0.785 | 0.785 |
| | | 1.50 | 0.5811 | 0.5818 | -3.15 | -3.04 | 0.797/2.24 | 0.794/2.16 | 0.833 | 0.828 |
| | 0.10 | 0.05 | 0.5853 | 0.5854 | -2.45 | -2.43 | 0.940/2.66 | 0.940/2.66 | 0.962 | 0.961 |
| | | 1.50 | 0.5797 | 0.5821 | -3.38 | -2.98 | 1.07/4.70 | 1.11/4.81 | 1.11 | 1.14 |
| 1.0 | 0.02 | 0.05 | 0.5937 | 0.6063 | -1.05 | 1.05 | 0.806/1.61 | 1.55/1.58 | 0.810 | 1.55 |
| | | 1.50 | 0.5915 | 0.5978 | -1.41 | -0.371 | 0.824/1.51 | 1.17/1.48 | 0.831 | 1.17 |
| | 0.04 | 0.05 | 0.5972 | 0.5972 | -0.475 | -0.467 | 1.21/4.13 | 1.21/4.13 | 1.21 | 1.21 |
| | | 1.50 | 0.5948 | 0.5951 | -0.868 | -0.810 | 1.11/3.70 | 1.11/3.66 | 1.12 | 1.12 |
| | 0.10 | 0.05 | 0.5866 | 0.5967 | -0.559 | -0.553 | 1.24/5.25 | 1.24/5.25 | 1.24 | 1.24 |
| | | 1.50 | 0.5987 | 0.5989 | -0.222 | -0.185 | 1.36/4.25 | 1.30/4.13 | 1.36 | 1.30 |
| 2.0 | 0.02 | 0.05 | 0.6011 | 0.6012 | 0.186 | 0.194 | 0.946/2.19 | 0.946/2.19 | 0.946 | 0.946 |
| | | 1.50 | 0.5957 | 0.5960 | -0.710 | -0.670 | 1.04/2.95 | 1.04/2.92 | 1.04 | 1.04 |
| | 0.04 | 0.05 | 0.6085 | 0.6087 | 1.42 | 1.45 | 1.65/3.28 | 1.65/3.17 | 1.65 | 1.65 |
| | | 1.50 | 0.5982 | 0.5986 | -0.292 | -0.238 | 1.60/2.44 | 1.60/2.34 | 1.60 | 1.60 |
| | 0.1 | 0.05 | 0.5884 | 0.5886 | -1.94 | -1.90 | 1.71/2.60 | 1.70/2.58 | 1.72 | 1.72 |
| | | 1.50 | 0.5841 | 0.5868 | -2.65 | -2.20 | 1.86/2.14 | 1.85/2.64 | 1.88 | 1.86 |

**Table A.10:** Mean estimate $\bar{\sigma}$, mean relative bias $\bar{b}$ ,empirical variance $S^2_{\text{emp}}$ and mean model-based variance $S^2_{\text{mb}}$ for $\sigma$ for settings with **A** ratios (at 2001.05Da).

| $\sigma$ | $Q$ | $\lambda$ | $\bar{\sigma}$ | | | $\bar{b}$ | | $S^2_{\text{emp}}(\times 1e-2)$ | | MSE$(\times 1e-2)$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LS | PLK | PL-GLS | PLK | PL-GLS | PLK | PL-GLS | PLK | PL-GLS |
| 0.05 | 0.5 | 0.02 | 11.72 | 0.065 | 0.065 | 0.300 | 0.298 | 0.158 | 0.155 | 0.180 | 0.177 |
| | | 0.04 | 11.89 | 0.067 | 0.067 | 0.343 | 0.342 | 0.161 | 0.161 | 0.191 | 0.190 |
| | | 0.10 | 11.98 | 0.064 | 0.064 | 0.287 | 0.279 | 0.219 | 0.216 | 0.240 | 0.236 |
| | 1.0 | 0.02 | 14.29 | 0.065 | 0.065 | 0.306 | 0.297 | 0.187 | 0.184 | 0.210 | 0.206 |
| | | 0.04 | 14.50 | 0.067 | 0.067 | 0.344 | 0.344 | 0.241 | 0.241 | 0.270 | 0.270 |
| | | 0.10 | 14.47 | 0.066 | 0.066 | 0.328 | 0.328 | 0.238 | 0.238 | 0.265 | 0.265 |
| | 2.0 | 0.02 | 18.62 | 0.062 | 0.062 | 0.237 | 0.237 | 0.215 | 0.215 | 0.229 | 0.229 |
| | | 0.04 | 18.61 | 0.067 | 0.067 | 0.346 | 0.345 | 0.301 | 0.301 | 0.331 | 0.331 |
| | | 0.10 | 18.23 | 0.080 | 0.080 | 0.604 | 0.601 | 0.770 | 0.766 | 0.861 | 0.856 |
| 1.50 | 0.5 | 0.02 | 355.0 | 2.014 | 1.997 | 0.342 | 0.332 | 128.6 | 122.3 | 155.0 | 147.1 |
| | | 0.04 | 355.9 | 1.986 | 1.975 | 0.324 | 0.317 | 159.2 | 154.5 | 182.8 | 177.1 |
| | | 0.10 | 360.3 | 1.906 | 1.911 | 0.271 | 0.274 | 138.5 | 142.5 | 155.0 | 159.4 |
| | 1.0 | 0.02 | 429.1 | 2.032 | 2.023 | 0.355 | 0.349 | 155.8 | 153.3 | 184.2 | 180.7 |
| | | 0.04 | 434.4 | 1.809 | 1.805 | 0.206 | 0.203 | 157.8 | 157.0 | 167.3 | 166.3 |
| | | 0.10 | 435.7 | 2.003 | 2.034 | 0.335 | 0.356 | 306.3 | 308.7 | 331.6 | 337.2 |
| | 2.0 | 0.02 | 558.3 | 1.932 | 1.928 | 0.288 | 0.286 | 157.5 | 156.6 | 176.1 | 175.0 |
| | | 0.04 | 554.0 | 2.247 | 2.243 | 0.498 | 0.495 | 332.4 | 330.7 | 338.2 | 385.8 |
| | | 0.10 | 543.7 | 2.378 | 2.364 | 0.585 | 0.576 | 451.9 | 448.8 | 529.0 | 523.4 |

**Table A.11:** Mean estimate $\bar{\sigma}$, mean relative bias $\bar{b}$ ,empirical variance $S^2_{\text{emp}}$ and mean model-based variance $S^2_{\text{mb}}$ for $\sigma$ for settings with **E1** ratios.

| $\sigma$ | $Q$ | $\lambda$ | $\bar{\sigma}$ | | | $\bar{b}$ | | $S^2_{\text{emp}}(\times 1e-2)$ | | MSE$(\times 1e-2)$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LS | PLK | PL-GLS | PLK | PL-GLS | PLK | PL-GLS | PLK | PL-GLS |
| 0.05 | 0.5 | 0.02 | 11.75 | 0.074 | 0.072 | 0.481 | 0.431 | 0.357 | 0.372 | 0.414 | 0.418 |
| | | 0.04 | 11.76 | 0.069 | 0.069 | 0.379 | 0.378 | 0.348 | 0.347 | 0.384 | 0.383 |
| | | 0.10 | 11.82 | 0.079 | 0.078 | 0.571 | 0.570 | 0.800 | 0.717 | 0.785 | 0.798 |
| | 1.0 | 0.02 | 14.07 | 0.075 | 0.074 | 0.492 | 0.473 | 0.425 | 0.432 | 0.485 | 0.488 |
| | | 0.04 | 14.20 | 0.083 | 0.083 | 0.651 | 0.650 | 0.862 | 0.861 | 0.968 | 0.967 |
| | | 0.10 | 14.22 | 0.083 | 0.083 | 0.670 | 0.669 | 0.703 | 0.702 | 0.815 | 0.814 |
| | 2.0 | 0.02 | 18.39 | 0.072 | 0.072 | 0.438 | 0.438 | 0.441 | 0.440 | 0.489 | 0.488 |
| | | 0.04 | 18.11 | 0.097 | 0.096 | 0.931 | 0.925 | 1.685 | 1.669 | 1.901 | 1.883 |
| | | 0.10 | 17.96 | 0.116 | 0.116 | 1.324 | 1.321 | 3.374 | 3.366 | 3.813 | 3.802 |
| 1.50 | 0.5 | 0.02 | 347.4 | 2.277 | 2.199 | 0.518 | 0.466 | 191.5 | 200.0 | 251.8 | 248.9 |
| | | 0.04 | 353.7 | 2.348 | 2.314 | 0.565 | 0.543 | 357.6 | 328.6 | 429.5 | 394.8 |
| | | 0.10 | 353.8 | 2.569 | 2.530 | 0.713 | 0.687 | 682.0 | 658.9 | 796.3 | 765.0 |
| | 1.0 | 0.02 | 422.6 | 2.365 | 2.317 | 0.576 | 0.546 | 378.6 | 371.2 | 453.4 | 438.2 |
| | | 0.04 | 432.6 | 2.196 | 2.190 | 0.464 | 0.460 | 489.3 | 487.4 | 537.8 | 535.0 |
| | | 0.10 | 429.0 | 2.447 | 2.458 | 0.632 | 0.639 | 700.4 | 704.2 | 790.2 | 796.1 |
| | 2.0 | 0.02 | 548.9 | 2.292 | 2.288 | 0.528 | 0.526 | 595.1 | 590.3 | 657.8 | 652.4 |
| | | 0.04 | 543.2 | 2.819 | 2.812 | 0.879 | 0.874 | 1541 | 1530 | 1715 | 1702 |
| | | 0.10 | 535.8 | 3.369 | 3.350 | 1.246 | 1.233 | 1817 | 1798 | 2166 | 2141 |

**Table A.12:** Mean estimate $\bar{\sigma}$, mean relative bias $\bar{b}$ ,empirical variance $S^2_{\text{emp}}$ and mean model-based variance $S^2_{\text{mb}}$ for $\sigma$ for settings with **E2** ratios.

| $\sigma$ | $Q$ | $\lambda$ | $\bar{\sigma}$ | | | $\bar{b}$ | | $S^2_{\text{emp}}(\times 1e-2)$ | | MSE$(\times 1e-2)$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LS | PLK | PL-GLS | PLK | PL-GLS | PLK | PL-GLS | PLK | PL-GLS |
| 0.05 | 0.5 | 0.02 | 13.32 | 0.076 | 0.073 | 0.528 | 0.457 | 0.293 | 0.311 | 0.363 | 0.364 |
| | | 0.04 | 13.38 | 0.077 | 0.077 | 0.533 | 0.532 | 0.379 | 0.379 | 0.450 | 0.449 |
| | | 0.10 | 13.36 | 0.080 | 0.080 | 0.593 | 0.591 | 0.465 | 0.464 | 0.553 | 0.552 |
| | 1.0 | 0.02 | 16.05 | 0.072 | 0.071 | 0.447 | 0.419 | 0.344 | 0.349 | 0.394 | 0.392 |
| | | 0.04 | 16.19 | 0.082 | 0.082 | 0.633 | 0.632 | 0.840 | 0.840 | 0.941 | 0.940 |
| | | 0.10 | 16.21 | 0.085 | 0.085 | 0.700 | 0.699 | 0.977 | 0.977 | 1.099 | 1.099 |
| | 2.0 | 0.02 | 20.80 | 0.073 | 0.073 | 0.456 | 0.455 | 0.461 | 0.460 | 0.513 | 0.512 |
| | | 0.04 | 20.74 | 0.093 | 0.092 | 0.852 | 0.848 | 1.818 | 1.797 | 2.000 | 1.977 |
| | | 0.10 | 20.45 | 0.114 | 0.113 | 1.270 | 1.264 | 3.675 | 3.651 | 4.079 | 4.050 |
| 1.50 | 0.5 | 0.02 | 395.1 | 2.410 | 2.298 | 0.607 | 0.532 | 294.8 | 312.0 | 377.7 | 375.7 |
| | | 0.04 | 400.4 | 2.391 | 2.374 | 0.594 | 0.583 | 406.3 | 397.8 | 485.7 | 474.2 |
| | | 0.10 | 399.7 | 2.749 | 2.721 | 0.833 | 0.814 | 1045 | 1066 | 1201 | 1215 |
| | 1.0 | 0.02 | 482.5 | 2.201 | 2.172 | 0.467 | 0.448 | 317.5 | 316.6 | 366.6 | 361.9 |
| | | 0.04 | 487.2 | 2.474 | 2.465 | 0.649 | 0.643 | 774.0 | 766.5 | 868.8 | 859.5 |
| | | 0.10 | 487.9 | 2.497 | 2.455 | 0.665 | 0.637 | 746.7 | 728.1 | 846.2 | 819.2 |
| | 2.0 | 0.02 | 620.3 | 2.442 | 2.435 | 0.628 | 0.624 | 689.6 | 681.5 | 778.2 | 769.0 |
| | | 0.04 | 617.8 | 3.068 | 3.060 | 1.045 | 1.040 | 2324 | 2309 | 2569 | 2552 |
| | | 0.10 | 614.8 | 3.870 | 3.752 | 1.580 | 1.501 | 3168 | 3037 | 3729 | 3544 |

# Appendix B

# Simulation results of the Bayesian heteroscedastic model for enzymatically $^{18}$O-labeled mass spectra

**Table B.1:** Mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$, mean model based variance $S^2_{\text{mb}}$ and MSE of $\lambda$ for settings with **A** ratios.

| $\lambda$ | $Q$ | $\sigma$ | $\bar{b}$ | | $S^2_{\text{emp}}/S^2_{\text{mb}}$ | | MSE | |
|---|---|---|---|---|---|---|---|---|
| | | | Con | Var | Con | Var | Con | Var |
| 0.02 | 0.5 | 0.05 | 0.01102 | -0.01086 | 6.928e-5/0.0001932 | 5.987e-6/1.170e-7 | 6.933e-5 | 6.034e-6 |
| | | 1.50 | 0.02378 | -0.01843 | 6.694e-5/0.0001895 | 4.338e-6/8.778e-6 | 6.717e-5 | 4.474e-6 |
| | 1.0 | 0.05 | -0.01365 | -0.02196 | 4.281e-5/1.090e-5 | 6.840e-6/9.628e-6 | 4.288e-5 | 7.033e-6 |
| | | 1.50 | 0.03856 | -0.02566 | 4.257e-5/0.0001100 | 4.412e-6/2.313e-7 | 4.317e-5 | 4.676e-6 |
| | 2.0 | 0.05 | -0.05310 | -0.01770 | 1.495e-5/5.703e-8 | 1.516e-5/1.987e-5 | 1.608e-5 | 1.529e-5 |
| | | 1.50 | -0.03495 | -0.02088 | 1.361e-5/1.058e-7 | 1.153e-5/2.406e-5 | 1.410e-5 | 1.171e-5 |
| 0.04 | 0.5 | 0.05 | -0.01314 | -0.02356 | 5.701e-5/1.123e-6 | 3.696e-5/1.122e-7 | 5.728e-5 | 3.785e-5 |
| | | 1.50 | -0.03228 | -0.02097 | 1.627e-5/5.024e-6 | 6.259e-6/1.570e-6 | 1.794e-5 | 6.963e-6 |
| | 1.0 | 0.05 | -6.644e-5 | -0.001068 | 5.773e-8/8.579e-10 | 3.377e-8/9.786e-8 | 5.774e-8 | 3.560e-8 |
| | | 1.50 | 0.002477 | -0.02142 | 4.898e-7/7.665e-7 | 5.990e-8/4.233e-7 | 4.996e-7 | 7.940e-7 |
| | 2.0 | 0.05 | 0.003000 | -0.007378 | 7.180e-6/1.430e-9 | 2.702e-6/1.617e-9 | 7.195e-6 | 2.789e-6 |
| | | 1.50 | 0.001814 | -0.02010 | 1.476e-6/1.877e-7 | 6.220e-7/2.838e-6 | 1.481e-6 | 1.268e-6 |
| 0.10 | 0.5 | 0.05 | 0.005850 | -0.02761 | 5.138e-5/2.212e-5 | 1.903e-5/1.900e-6 | 5.172e-5 | 2.665e-5 |
| | | 1.50 | 0.6516 | 0.5218 | 0.0001307/4.469e-5 | 0.0001930/2.672e-5 | 0.004377 | 0.002916 |
| | 1.0 | 0.05 | 0.001081 | -0.008568 | 5.256e-5/8.798e-7 | 1.768e-5/4.179e-7 | 5.257e-5 | 1.842e-5 |
| | | 1.50 | 0.6621 | 0.6255 | 1.252e-5/1.880e-5 | 0.0002624/3.200e-5 | 0.004396 | 0.004175 |
| | 2.0 | 0.05 | -0.001708 | -0.003913 | 1.908e-5/2.180e-7 | 8.500e-6/1.886e-7 | 1.911e-5 | 8.653e-6 |
| | | 1.50 | 0.6510 | 0.5684 | 0.0001336/3.442e-5 | 0.001148/8.666e-5 | 0.004372 | 0.004379 |

**Table B.2:** Mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$, mean model based variance $S^2_{\text{mb}}$ and MSE of $\lambda$ for settings with **E1** ratios.

| $\lambda$ | $Q$ | $\sigma$ | $\bar{b}$ | | $S^2_{\text{emp}}/S^2_{\text{mb}}$ | | MSE | |
|---|---|---|---|---|---|---|---|---|
| | | | Con | Var | Con | Var | Con | Var |
| 0.02 | 0.5 | 0.05 | 0.008361 | 0.09784 | 4.717e-5/0.0001363 | 3.207e-5/2.077e-6 | 4.719e-5 | 3.589e-5 |
| | | 1.50 | 0.02628 | 0.03859 | 4.563e-5/0.0001960 | 2.045e-5/0.0001788 | 4.591e-5 | 2.104e-5 |
| | 1.0 | 0.05 | 0.002003 | -0.04111 | 8.027e-6/5.023e-6 | 3.029e-6/1.581e-7 | 8.029e-6 | 3.705e-6 |
| | | 1.50 | 0.02856 | 0.03824 | 3.952e-6/3.538e-7 | 3.200e-6/3.800e-5 | 4.278e-6 | 3.785e-6 |
| | 2.0 | 0.05 | -0.1147 | -0.1746 | 2.585e-5/9.456e-8 | 8.842e-6/1.812e-5 | 3.111e-5 | 2.103e-5 |
| | | 1.50 | -0.08814 | -0.1305 | 3.753e-5/2.083e-7 | 2.590e-5/2.367e-5 | 4.063e-5 | 3.271e-5 |
| 0.04 | 0.5 | 0.05 | -0.06842 | -0.06629 | 8.689e-5/3.766e-6 | 6.045e-5/8.709e-6 | 9.328e-5 | 6.748e-5 |
| | | 1.50 | 0.01001 | -0.05291 | 0.0004334/9.493e-5 | 1.433e-5/3.065e-6 | 0.0004335 | 1.881e-5 |
| | 1.0 | 0.05 | 1.236e-5 | -0.04028 | 9.872e-6/1.078e-9 | 2.131e-6/3.624e-7 | 9.872e-6 | 4.727e-6 |
| | | 1.50 | -0.0008734 | -0.01924 | 3.002e-6/1.028e-6 | 1.827e-6/8.128e-7 | 3.004e-6 | 2.419e-6 |
| | 2.0 | 0.05 | -0.001758 | -0.02649 | 2.450e-6/9.505e-10 | 1.794e-6/2.837e-8 | 2.455e-6 | 2.917e-6 |
| | | 1.50 | -0.002947 | -0.01772 | 4.873e-6/2.498e-7 | 2.699e-6/2.153e-7 | 4.887e-6 | 3.202e-6 |
| 0.10 | 0.5 | 0.05 | 0.007315 | -0.03744 | 0.0002927/4.172e-5 | 7.277e-5/9.234e-6 | 0.0002933 | 8.679e-5 |
| | | 1.50 | 0.6057 | 0.2833 | 0.0008743/3.212e-5 | 0.0003987/1.607e-5 | 0.004542 | 0.001201 |
| | 1.0 | 0.05 | 0.002218 | -0.01597 | 4.244e-6/1.157e-6 | 3.368e-7/8.756e-7 | 4.293e-6 | 2.887e-6 |
| | | 1.50 | 0.6564 | 0.5056 | 6.578e-5/2.787e-5 | 0.0001991/2.621e-5 | 0.004374 | 0.002755 |
| | 2.0 | 0.05 | -0.003675 | -0.01844 | 3.631e-5/2.929e-7 | 2.670e-5/1.603e-6 | 3.644e-5 | 3.011e-5 |
| | | 1.50 | 0.6553 | 0.5573 | 0.0001062/2.999e-5 | 0.0001723/5.285e-5 | 0.004400 | 0.003278 |

**Table B.3:** Mean relative bias $\bar{b}$, empirical variance $S^2_{\mathrm{emp}}$, mean model based variance $S^2_{\mathrm{mb}}$ and MSE of $\lambda$ for settings with **E2** ratios.

| $\lambda$ | $Q$ | $\sigma$ | $\bar{b}$ | | $S^2_{\mathrm{emp}}/S^2_{\mathrm{mb}}$ | | MSE | |
|---|---|---|---|---|---|---|---|---|
| | | | Con | Var | Con | Var | Con | Var |
| 0.02 | 0.5 | 0.05 | 0.05282 | 0.06272 | 0.0001687/6.584e-5 | 4.138e-5/9.275e-6 | 0.0001698 | 4.296e-5 |
| | | 1.50 | 0.01329 | 0.06826 | 0.0003051/0.0001087 | 7.624e-5/0.0001395 | 0.0003052 | 7.811e-5 |
| | 1.0 | 0.05 | -0.0006426 | 0.01339 | 7.281e-6/2.604e-8 | 3.358e-6/6.037e-6 | 7.281e-6 | 3.430e-6 |
| | | 1.50 | 0.01575 | 0.009722 | 5.173e-6/2.953e-7 | 4.014e-6/4.364e-6 | 5.272e-6 | 4.052e-6 |
| | 2.0 | 0.05 | 0.001823 | -0.01249 | 6.528e-6/1.006e-8 | 4.597e-6/1.580e-6 | 6.529e-6 | 4.659e-6 |
| | | 1.50 | 0.002581 | -0.007212 | 3.260e-7/1.486e-7 | 1.613e-7/6.385e-7 | 3.287e-7 | 1.821e-7 |
| 0.04 | 0.5 | 0.05 | -0.03583 | 0.01056 | 4.122e-5/6.196e-7 | 9.863e-6/6.905e-6 | 4.327e-5 | 1.004e-5 |
| | | 1.50 | -0.03065 | -0.02072 | 0.0001067/4.950e-5 | 7.695e-5/1.038e-5 | 0.0001082 | 7.764e-5 |
| | 1.0 | 0.05 | -0.004993 | -0.04036 | 6.574e-6/9.196e-9 | 1.042e-6/9.380e-7 | 6.614e-6 | 3.648e-6 |
| | | 1.50 | 0.003517 | -0.02591 | 3.690e-5/9.214e-7 | 1.454e-5/7.749e-7 | 3.692e-5 | 1.561e-5 |
| | 2.0 | 0.05 | -2.141e-6 | -0.04033 | 9.194e-6/2.627e-10 | 7.100e-8/6.998e-8 | 9.194e-6 | 2.673e-6 |
| | | 1.50 | -0.002837 | -0.03150 | 5.037e-6/2.359e-7 | 7.314e-7/3.128e-7 | 5.049e-6 | 2.319e-6 |
| 0.10 | 0.5 | 0.05 | -0.01176 | -0.02978 | 0.0002200/2.894e-5 | 7.954e-5/7.074e-6 | 0.0002214 | 8.841e-5 |
| | | 1.50 | 0.6510 | 0.4134 | 0.0001777/3.412e-5 | 0.0002969/1.962e-5 | 0.004416 | 0.002006 |
| | 1.0 | 0.05 | -0.0007639 | -0.02263 | 1.929e-5/1.036e-6 | 7.773e-6/1.175e-5 | 1.929e-5 | 1.290e-5 |
| | | 1.50 | 0.6587 | 0.5424 | 5.697e-5/2.975e-5 | 0.0001603/2.995e-5 | 0.004395 | 0.003102 |
| | 2.0 | 0.05 | -0.007416 | -0.009750 | 7.137e-5/2.808e-7 | 4.656e-5/4.527e-6 | 7.192e-5 | 4.751e-5 |
| | | 1.50 | 0.6380 | 0.5325 | 0.0003702/2.850e-5 | 0.0003276/4.718e-5 | 0.004441 | 0.003163 |

**Table B.4:** Mean relative bias $\bar{b}$, empirical variance $S^2_{\mathrm{emp}}$, mean model based variance $S^2_{\mathrm{mb}}$ and MSE of $Q$ for settings with **A** ratios.

| $Q$ | $\lambda$ | $\sigma$ | $\bar{b}$ | | $S^2_{\mathrm{emp}}/S^2_{\mathrm{mb}}$ | | MSE | |
|---|---|---|---|---|---|---|---|---|
| | | | Con | Var | Con | Var | Con | Var |
| | 0.02 | 0.05 | -0.006669 | 0.001626 | 0.04087/0.001640 | 0.002484/0.003961 | 0.04088 | 0.002485 |
| | | 1.50 | -0.02846 | 0.01797 | 0.03533/0.003750 | 0.007419/0.01148 | 0.03554 | 0.007500 |
| 0.5 | 0.04 | 0.05 | 0.02615 | 0.004610 | 0.02427/0.005389 | 0.001837/1.2912e-7 | 0.02444 | 0.001843 |
| | | 1.50 | 0.06262 | -0.003446 | 0.005060/0.002466 | 0.001590/0.0003709 | 0.006041 | 0.001593 |
| | 0.10 | 0.05 | 0.01499 | 0.005010 | 0.02255/0.002971 | 0.001477/9.419e-8 | 0.02260 | 0.001484 |
| | | 1.50 | 0.02058 | -0.002415 | 0.002043/0.001890 | 0.001941/4.146e-5 | 0.002149 | 0.001943 |
| | 0.02 | 0.05 | -0.001535 | 0.003136 | 0.005158/0.005075 | 0.003310/0.002516 | 0.005160 | 0.003320 |
| | | 1.50 | 0.02483 | 0.003714 | 0.05064/0.01357 | 0.004351/0.007700 | 0.05126 | 0.004365 |
| 1.0 | 0.04 | 0.05 | 4.916e-5 | 0.002165 | 0.002949/2.974e-7 | 0.001700/0.0001333 | 0.002949 | 0.001705 |
| | | 1.50 | 0.01296 | 0.0005067 | 0.003495/0.0002484 | 0.003866/0.0002329 | 0.003663 | 0.003866 |
| | 0.10 | 0.05 | -1.787e-5 | -0.00009779 | 0.002395/1.999e-7 | 0.0007845/0.0002531 | 0.002395 | 0.0007846 |
| | | 1.50 | 0.001473 | -0.007731 | 0.0008563/8.431e-5 | 0.0006313/0.0001234 | 0.0008585 | 0.0006911 |
| | 0.02 | 0.05 | 0.007708 | 0.002294 | 0.05282/0.03407 | 0.007400/0.001279 | 0.05306 | 0.007421 |
| | | 1.50 | 0.003887 | -0.0005702 | 0.04991/0.03883 | 0.01440/0.001382 | 0.04997 | 0.01440 |
| 2.0 | 0.04 | 0.05 | -0.001026 | -0.0004800 | 0.001999/5.299e-7 | 0.0008603/6.518e-7 | 0.002003 | 0.0008612 |
| | | 1.50 | 0.001794 | 0.0002757 | 0.005573/0.0004825 | 0.001874/0.0006046 | 0.005576 | 0.001874 |
| | 0.1 | 0.05 | 0.01739 | 0.0006205 | 0.006044/0.008720 | 0.0003334/6.341e-7 | 0.006047 | 0.0003349 |
| | | 1.50 | 0.01214 | -0.01004 | 0.05210/0.0002484 | 0.009141/0.0005767 | 0.05210 | 0.009544 |

**Table B.5:** Mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$, mean model based variance $S^2_{\text{mb}}$ and MSE of $Q$ for settings with **E1** ratios.

| $Q$ | $\lambda$ | $\sigma$ | $\bar{b}$ | | $S^2_{\text{emp}}/S^2_{\text{mb}}$ | | MSE | |
|---|---|---|---|---|---|---|---|---|
| | | | Con | Var | Con | Var | Con | Var |
| 0.5 | 0.02 | 0.05 | -0.01612 | -0.05900 | 0.03366/0.001934 | 0.01196/0.004823 | 0.03372 | 0.01283 |
| | | 1.50 | 0.08438 | -0.02350 | 0.01312/0.007108 | 0.009602/0.02141 | 0.01490 | 0.009740 |
| | 0.04 | 0.05 | 0.005713 | -0.01006 | 0.001771/0.02095 | 0.001555/0.001529 | 0.001779 | 0.001580 |
| | | 1.50 | 0.05076 | -0.004173 | 0.002974/0.01851 | 0.001890/0.04947 | 0.003618 | 0.001895 |
| | 0.10 | 0.05 | -0.0002105 | -0.003559 | 0.002896/0.004156 | 0.002861/0.04720 | 0.002896 | 0.002864 |
| | | 1.50 | 0.005923 | -0.008106 | 0.007818/0.003998 | 0.005273/0.05314 | 0.007826 | 0.005289 |
| 1.0 | 0.02 | 0.05 | -0.0002159 | 0.01068 | 0.02557/0.0001284 | 0.02524/0.01176 | 0.02557 | 0.02535 |
| | | 1.50 | 0.02020 | -0.009803 | 0.04032/0.001473 | 0.01929/0.02904 | 0.04073 | 0.01938 |
| | 0.04 | 0.05 | 9.504e-6 | 0.004157 | 0.008751/3.980e-7 | 0.005881/0.004613 | 0.008751 | 0.005898 |
| | | 1.50 | 0.008948 | 0.003450 | 0.004212/0.001298 | 0.003159/0.01151 | 0.004292 | 0.003128 |
| | 0.10 | 0.05 | -5.905e-5 | -0.001498 | 0.003476/2.554e-7 | 0.001167/0.01286 | 0.003476 | 0.001169 |
| | | 1.50 | -0.005651 | -0.007184 | 0.0006191/0.0001144 | 0.0002382/0.01007 | 0.0006510 | 0.0002898 |
| 2.0 | 0.02 | 0.05 | 0.04800 | 0.008642 | 0.1038/0.04995 | 0.1116/0.03090 | 0.1131 | 0.1119 |
| | | 1.50 | 0.07142 | 0.02060 | 0.1426/0.09900 | 0.1035/0.02886 | 0.1630 | 0.1052 |
| | 0.04 | 0.05 | 0.003682 | 0.01916 | 0.02708/0.005067 | 0.02007/0.006456 | 0.02714 | 0.02154 |
| | | 1.50 | 0.003737 | 0.01255 | 0.05079/0.009815 | 0.03480/0.007416 | 0.05085 | 0.03543 |
| | 0.1 | 0.05 | -0.0003838 | 0.002424 | 0.0005451/5.205e-7 | 0.0002579/0.006637 | 0.0005457 | 0.0002814 |
| | | 1.50 | -0.006008 | 0.002590 | 0.004974/0.0002872 | 0.003333/0.004556 | 0.005118 | 0.003360 |

**Table B.6:** Mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$, mean model based variance $S^2_{\text{mb}}$ and MSE of $Q$ for settings with **E2** ratios.

| $Q$ | $\lambda$ | $\sigma$ | $\bar{b}$ | | $S^2_{\text{emp}}/S^2_{\text{mb}}$ | | MSE | |
|---|---|---|---|---|---|---|---|---|
| | | | Con | Var | Con | Var | Con | Var |
| 0.5 | 0.02 | 0.05 | -0.007620 | -0.008933 | 0.001621/0.003790 | 0.001474/0.003077 | 0.001636 | 0.001494 |
| | | 1.50 | 0.1646 | -0.02575 | 0.02723/0.003644 | 0.008896/0.02712 | 0.03400 | 0.009062 |
| | 0.04 | 0.05 | 0.09485 | -0.006884 | 0.04625/0.007329 | 0.002903/0.01132 | 0.04850 | 0.002915 |
| | | 1.50 | 0.03843 | 0.009131 | 0.009413/0.008726 | 0.004773/0.02881 | 0.009782 | 0.004794 |
| | 0.10 | 0.05 | 0.006087 | 0.01369 | 0.009739/0.004929 | 0.004061/0.02841 | 0.009748 | 0.004108 |
| | | 1.50 | 0.001765 | -0.006969 | 0.004649/0.0003728 | 0.002504/0.02893 | 0.004649 | 0.002516 |
| 1.0 | 0.02 | 0.05 | 0.002051 | -0.02269 | 0.001950/0.0007679 | 0.001382/0.005984 | 0.001954 | 0.001897 |
| | | 1.50 | 0.02186 | -0.003470 | 0.005768/0.001530 | 0.001218/0.002702 | 0.006246 | 0.001230 |
| | 0.04 | 0.05 | 0.0001579 | 0.004473 | 0.005983/0.001927 | 0.003889/4.110e-6 | 0.005983 | 0.003909 |
| | | 1.50 | 0.009293 | -0.003566 | 0.01261/0.002065 | 0.004525/0.001781 | 0.01270 | 0.004538 |
| | 0.10 | 0.05 | 0.0002179 | -0.001814 | 0.002644/2.254e-7 | 0.002606/0.004978 | 0.002644 | 0.002609 |
| | | 1.50 | -0.004950 | -0.007241 | 0.001716/0.0009614 | 0.001576/0.001242 | 0.001962 | 0.001628 |
| 2.0 | 0.02 | 0.05 | -0.001484 | -0.01000 | 0.004243/2.928e-6 | 0.001845/0.003336 | 0.004252 | 0.002245 |
| | | 1.50 | 0.002719 | -0.004412 | 0.05077/0.008328 | 0.03144/0.001617 | 0.05080 | 0.03152 |
| | 0.04 | 0.05 | -2.154e-5 | 0.001227 | 0.007630/6.276e-7 | 0.003191/1.353e-6 | 0.007630 | 0.003197 |
| | | 1.50 | 0.0009192 | 0.003189 | 0.007019/0.0005592 | 0.001927/0.002221 | 0.007022 | 0.001968 |
| | 0.1 | 0.05 | -1.441e-5 | 0.003296 | 0.0002897/4.727e-7 | 0.0001007/0.001467 | 0.0002897 | 0.0001442 |
| | | 1.50 | -0.006304 | -0.0001904 | 0.005299/0.0002538 | 0.002519/0.0003318 | 0.005458 | 0.002519 |

**Table B.7:** Mean estimate $\bar{\theta}$, mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$, mean model based variance $S^2_{\text{mb}}$ and MSE of $\theta$ for settings with **A** ratios.

| $Q$ | $\lambda$ | $\sigma$ | $\bar{\theta}$ | $\bar{b}$ | $S^2_{\text{emp}}/S^2_{\text{mb}}$ | MSE |
|---|---|---|---|---|---|---|
| 0.5 | 0.02 | 0.05 | 0.5443 | -0.09282 | 0.001394/0.008363 | 0.004495 |
| | | 1.50 | 0.5900 | -0.01666 | 0.003234/0.004807 | 0.003334 |
| | 0.04 | 0.05 | 0.5306 | -0.1157 | 0.001452/0.004971 | 0.006268 |
| | | 1.50 | 0.5941 | -0.009791 | 0.003586/0.004546 | 0.003621 |
| | 0.10 | 0.05 | 0.5320 | -0.1133 | 0.001723/0.003403 | 0.006345 |
| | | 1.50 | 0.6012 | 0.001943 | 0.003162/0.004371 | 0.003164 |
| 1.0 | 0.02 | 0.05 | 0.5310 | -0.1150 | 0.001272/0.009123 | 0.006032 |
| | | 1.50 | 0.5911 | -0.01480 | 0.003093/0.006348 | 0.003172 |
| | 0.04 | 0.05 | 0.5205 | -0.1325 | 0.001199/0.003844 | 0.007519 |
| | | 1.50 | 0.5976 | -0.003921 | 0.003599/0.005869 | 0.003604 |
| | 0.10 | 0.05 | 0.5185 | -0.1359 | 0.001323/0.002619 | 0.007973 |
| | | 1.50 | 0.6028 | 0.004635 | 0.003774/0.004115 | 0.003782 |
| 2.0 | 0.02 | 0.05 | 0.5282 | -0.1196 | 0.001378/0.01057 | 0.006531 |
| | | 1.50 | 0.5919 | -0.01356 | 0.002901/0.005814 | 0.002967 |
| | 0.04 | 0.05 | 0.5191 | -0.1348 | 0.001517/0.002791 | 0.008060 |
| | | 1.50 | 0.5935 | -0.01080 | 0.003885/0.004956 | 0.003927 |
| | 0.1 | 0.05 | 0.5185 | -0.1358 | 0.001664/0.002706 | 0.008304 |
| | | 1.50 | 0.5937 | -0.01058 | 0.003515/0.005086 | 0.003555 |

**Table B.8:** Mean estimate $\bar{\theta}$, mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$, mean model based variance $S^2_{\text{mb}}$ and MSE of $\theta$ for settings with **E1** ratios.

| $Q$ | $\lambda$ | $\sigma$ | $\bar{\theta}$ | $\bar{b}$ | $S^2_{\text{emp}}/S^2_{\text{mb}}$ | MSE |
|---|---|---|---|---|---|---|
| 0.5 | 0.02 | 0.05 | 0.5429 | -0.09513 | 0.002811/0.02105 | 0.006069 |
| | | 1.50 | 0.5944 | -0.009414 | 0.003504/0.007470 | 0.003536 |
| | 0.04 | 0.05 | 0.5189 | -0.1352 | 0.002666/0.004905 | 0.009246 |
| | | 1.50 | 0.5958 | -0.006976 | 0.004130/0.009950 | 0.004148 |
| | 0.10 | 0.05 | 0.5232 | -0.1281 | 0.003282/0.009913 | 0.009185 |
| | | 1.50 | 0.5957 | -0.007135 | 0.004904/0.01220 | 0.004923 |
| 1.0 | 0.02 | 0.05 | 0.5315 | -0.1141 | 0.003016/0.03233 | 0.007704 |
| | | 1.50 | 0.5935 | -0.01082 | 0.003917/0.01081 | 0.003959 |
| | 0.04 | 0.05 | 0.5332 | -0.1113 | 0.003674/0.01245 | 0.008132 |
| | | 1.50 | 0.6004 | 0.0007440 | 0.004471/0.01179 | 0.004471 |
| | 0.10 | 0.05 | 0.5270 | -0.1217 | 0.004018/0.006371 | 0.009346 |
| | | 1.50 | 0.5948 | -0.008726 | 0.004917/0.01039 | 0.004944 |
| 2.0 | 0.02 | 0.05 | 0.5386 | -0.1024 | 0.003795/0.05511 | 0.007569 |
| | | 1.50 | 0.5978 | -0.003601 | 0.003812/0.01310 | 0.003817 |
| | 0.04 | 0.05 | 0.5335 | -0.1108 | 0.004195/0.007222 | 0.008618 |
| | | 1.50 | 0.5927 | -0.01214 | 0.004650/0.009916 | 0.004703 |
| | 0.1 | 0.05 | 0.5472 | -0.08805 | 0.005022/0.007040 | 0.007813 |
| | | 1.50 | 0.5975 | -0.004210 | 0.005574/0.01044 | 0.005580 |

**Table B.9:** Mean estimate $\bar{\theta}$, mean relative bias $\bar{b}$, empirical variance $S^2_{\text{emp}}$, mean model based variance $S^2_{\text{mb}}$ and MSE of $\theta$ for settings with **E2** ratios.

| $Q$ | $\lambda$ | $\sigma$ | $\bar{\theta}$ | $\bar{b}$ | $S^2_{\text{emp}}/S^2_{\text{mb}}$ | MSE |
|---|---|---|---|---|---|---|
| 0.5 | 0.02 | 0.05 | 0.5449 | -0.09183 | 0.003093/0.02576 | 0.006129 |
| | | 1.50 | 0.5937 | -0.01053 | 0.003319/0.009010 | 0.003359 |
| | 0.04 | 0.05 | 0.5255 | -0.1241 | 0.003359/0.006160 | 0.008903 |
| | | 1.50 | 0.5921 | -0.01314 | 0.004390/0.01320 | 0.004452 |
| | 0.10 | 0.05 | 0.5341 | -0.1098 | 0.004161/0.006776 | 0.008503 |
| | | 1.50 | 0.5997 | -0.0004683 | 0.004677/0.009608 | 0.004677 |
| 1.0 | 0.02 | 0.05 | 0.5399 | -0.1002 | 0.003773/0.02281 | 0.007386 |
| | | 1.50 | 0.5976 | -0.003956 | 0.004046/0.01136 | 0.004051 |
| | 0.04 | 0.05 | 0.5346 | -0.1090 | 0.004041/0.008035 | 0.008322 |
| | | 1.50 | 0.5884 | -0.01929 | 0.005437/0.01190 | 0.005571 |
| | 0.10 | 0.05 | 0.5432 | -0.09473 | 0.004839/0.006792 | 0.008070 |
| | | 1.50 | 0.5943 | -0.009427 | 0.004763/0.009679 | 0.004795 |
| 2.0 | 0.02 | 0.05 | 0.5393 | -0.1003 | 0.004328/0.004724 | 0.007949 |
| | | 1.50 | 0.6007 | 0.001087 | 0.004191/0.008806 | 0.004191 |
| | 0.04 | 0.05 | 0.5474 | -0.08767 | 0.004753/0.007907 | 0.007520 |
| | | 1.50 | 0.6015 | 0.002433 | 0.004927/0.01030 | 0.004929 |
| | 0.1 | 0.05 | 0.5507 | -0.08219 | 0.004392/0.006314 | 0.006824 |
| | | 1.50 | 0.5974 | -0.004291 | 0.005005/0.01044 | 0.005012 |

**Table B.10:** Mean estimate $\bar{\sigma}$, mean relative bias $\bar{b}$ ,empirical variance $S^2_{\mathrm{emp}}$, model-based variance $S^2_{\mathrm{mb}}$, mean model-based variance $S^2_{\mathrm{mb}}$ and MSE of $\sigma$ for settings with **A** ratios.

| $\sigma$ | $Q$ | $\lambda$ | $\bar{\sigma}$ | | $\bar{b}$ | $S^2_{\mathrm{emp}}/S^2_{\mathrm{mb}}$ | MSE |
|---|---|---|---|---|---|---|---|
| | | | Con | Var | Var | | |
| 0.05 | 0.5 | 0.02 | 18.5894 | 0.09819 | 0.9639 | 0.0007254/0.0004653 | 0.003048 |
| | | 0.04 | 19.0040 | 0.1152 | 1.3042 | 0.0009614/0.0007845 | 0.005214 |
| | | 0.10 | 14.2154 | 0.09035 | 0.8069 | 0.0004353/0.0008671 | 0.002063 |
| | 1.0 | 0.02 | 15.8446 | 0.1141 | 1.2824 | 0.0009159/0.0009105 | 0.005027 |
| | | 0.04 | 12.7127 | 0.1290 | 1.5792 | 0.001292/0.001826 | 0.007527 |
| | | 0.10 | 12.6556 | 0.08517 | 0.7034 | 0.0005435/0.001777 | 0.001781 |
| | 2.0 | 0.02 | 16.3881 | 0.1272 | 1.5434 | 0.001194/0.001553 | 0.007149 |
| | | 0.04 | 16.2937 | 0.1455 | 1.9102 | 0.001710/0.004273 | 0.01083 |
| | | 0.10 | 15.9068 | 0.1433 | 1.8652 | 0.001947/0.004377 | 0.01064 |
| 1.50 | 0.5 | 0.02 | 335.7952 | 1.9898 | 0.3265 | 0.4861/2.1879 | 0.7260 |
| | | 0.04 | 323.4883 | 1.9743 | 0.3162 | 0.5041/2.7104 | 0.7290 |
| | | 0.10 | 322.3756 | 1.9743 | 0.3162 | 0.6331/2.7103 | 0.8581 |
| | 1.0 | 0.02 | 401.7372 | 2.0558 | 0.3705 | 0.5397/2.1775 | 0.8486 |
| | | 0.04 | 392.5880 | 2.2346 | 0.4897 | 1.3709/6.8713 | 1.9105 |
| | | 0.10 | 392.5763 | 1.9847 | 0.3231 | 0.6040/2.6555 | 0.8390 |
| | 2.0 | 0.02 | 521.690 | 2.2841 | 0.5227 | 1.2440/8.2922 | 1.8587 |
| | | 0.04 | 493.2701 | 2.0687 | 0.3791 | 0.6139/2.1772 | 0.9374 |
| | | 0.10 | 482.0026 | 2.8335 | 0.8890 | 2.3688/9.6139 | 4.1469 |

**Table B.11:** Mean estimate $\bar{\sigma}$, mean relative bias $\bar{b}$ ,empirical variance $S^2_{\text{emp}}$, model-based variance $S^2_{\text{mb}}$, mean model-based variance $S^2_{\text{mb}}$ and MSE of $\sigma$ for settings with **E1** ratios.

| $\sigma$ | $Q$ | $\lambda$ | $\bar{\sigma}$ | | $\bar{b}$ | $S^2_{\text{emp}}/S^2_{\text{mb}}$ | MSE |
|---|---|---|---|---|---|---|---|
| | | | Con | Var | Var | | |
| 0.05 | 0.5 | 0.02 | 15.2336 | 0.1173 | 1.3469 | 0.002149/0.002048 | 0.006684 |
| | | 0.04 | 12.1991 | 0.1392 | 1.7838 | 0.002912/0.006908 | 0.01087 |
| | | 0.10 | 12.1509 | 0.1341 | 1.6823 | 0.003627/0.007744 | 0.01070 |
| | 1.0 | 0.02 | 15.0681 | 0.1340 | 1.6807 | 0.003329/0.006147 | 0.01039 |
| | | 0.04 | 14.5067 | 0.1156 | 1.3116 | 0.003982/0.009409 | 0.008283 |
| | | 0.10 | 14.5330 | 0.1203 | 1.4061 | 0.004313/0.009689 | 0.009256 |
| | 2.0 | 0.02 | 18.7745 | 0.1120 | 1.2395 | 0.003494/0.006812 | 0.007335 |
| | | 0.04 | 18.5059 | 0.1039 | 1.0789 | 0.003635/0.008005 | 0.006545 |
| | | 0.10 | 18.3506 | 0.1037 | 1.0733 | 0.003514/0.008314 | 0.006394 |
| 1.50 | 0.5 | 0.02 | 379.2682 | 2.0831 | 0.3887 | 0.5155/1.7074 | 0.8554 |
| | | 0.04 | 401.1458 | 2.5882 | 0.7255 | 2.1841/5.7799 | 3.3682 |
| | | 0.10 | 386.8660 | 2.6634 | 0.7756 | 2.3842/5.9524 | 3.7377 |
| | 1.0 | 0.02 | 452.3595 | 2.6594 | 0.7729 | 2.0070/6.7894 | 3.3512 |
| | | 0.04 | 448.8193 | 2.7686 | 0.8457 | 2.5571/5.4196 | 4.1664 |
| | | 0.10 | 438.0552 | 2.7417 | 0.8278 | 2.6940/5.5326 | 4.2358 |
| | 2.0 | 0.02 | 1953.030 | 2.7265 | 0.8177 | 2.4175/4.3031 | 3.9218 |
| | | 0.04 | 578.5213 | 2.8660 | 0.9107 | 2.8325/4.6992 | 4.6984 |
| | | 0.10 | 551.8279 | 2.7443 | 0.8295 | 2.7014/4.1197 | 4.2496 |

**Table B.12:** Mean estimate $\bar{\sigma}$, mean relative bias $\bar{b}$ ,empirical variance $S^2_{\text{emp}}$, model-based variance $S^2_{\text{mb}}$, mean model-based variance $S^2_{\text{mb}}$ and MSE of $\sigma$ for settings with **E2** ratios.

| $\sigma$ | $Q$ | $\lambda$ | $\bar{\sigma}$ | | $\bar{b}$ | $S^2_{\text{emp}}/S^2_{\text{mb}}$ | MSE |
|---|---|---|---|---|---|---|---|
| | | | Con | Var | Var | | |
| 0.05 | 0.5 | 0.02 | 19.9978 | 0.1226 | 1.4514 | 0.002613/0.003438 | 0.007879 |
| | | 0.04 | 13.6390 | 0.1327 | 1.6534 | 0.003439/0.009084 | 0.01027 |
| | | 0.10 | 13.6609 | 0.1279 | 1.5590 | 0.003925/0.008920 | 0.01000 |
| | 1.0 | 0.02 | 16.4048 | 0.1308 | 1.6155 | 0.003516/0.008056 | 0.01004 |
| | | 0.04 | 16.5354 | 0.1090 | 1.1797 | 0.003777/0.009524 | 0.007256 |
| | | 0.10 | 16.5585 | 0.1032 | 1.0640 | 0.003516/0.009037 | 0.006347 |
| | 2.0 | 0.02 | 21.2499 | 0.1211 | 1.4225 | 0.003212/0.008211 | 0.008270 |
| | | 0.04 | 21.1953 | 0.09346 | 0.8691 | 0.002788/0.006929 | 0.004677 |
| | | 0.10 | 20.8847 | 0.08657 | 0.7315 | 0.002027/0.007547 | 0.003365 |
| 1.50 | 0.5 | 0.02 | 430.8158 | 2.5387 | 0.6924 | 2.0333/6.2857 | 3.1121 |
| | | 0.04 | 426.3120 | 2.7563 | 0.8375 | 2.5173/6.1427 | 4.0954 |
| | | 0.10 | 410.6182 | 2.7572 | 0.8382 | 2.7048/6.8906 | 4.2855 |
| | 1.0 | 0.02 | 511.5862 | 2.6370 | 0.7580 | 2.6987/7.2538 | 3.9916 |
| | | 0.04 | 501.4285 | 2.7189 | 0.8126 | 2.5706/4.1126 | 4.0563 |
| | | 0.10 | 499.2004 | 2.1012 | 0.4008 | 0.5633/0.8372 | 0.9247 |
| | 2.0 | 0.02 | 643.1848 | 2.3306 | 0.5537 | 2.9703/5.1681 | 3.6602 |
| | | 0.04 | 631.4164 | 2.8110 | 0.8741 | 3.1187/2.7966 | 4.8374 |
| | | 0.10 | 632.6614 | 2.7446 | 0.8297 | 2.7175/3.8970 | 4.2665 |

# Appendix C

# WinBUGS code for the Bayesian model of $^{18}$O-labeled mass spectra

The following lines give the WinBUGS code implemented for the model with mean-power-variance function:

```
model {
###Matrix exponential implemented by differential
equations:
    solution[1:n.grid, 1:dim] <- ode(init[1:dim], grid[1:n.grid], D(pi[1:dim], t), origin, tol)

    D(pi[c_11], t) <- p16*lambda*pi[c_11]+p16/2*lambda*pi[c_12]+p16/2*lambda*pi[c_13]
    D(pi[c_12], t) <- (p16+p17)/2*lambda*pi[c_12]+p17*lambda*pi[c_11]+p17/2*lambda*pi[c_13]+p16*lambda*pi[c_14]+p16/2*lambda*pi[c_15]
    D(pi[c_13], t) <- (p16+p18)/2*lambda*pi[c_13]+p18*lambda*pi[c_11]+p18/2*lambda*pi[c_12]+p16/2*lambda*pi[c_15]+p16*lambda*pi[c_16]
    D(pi[c_14], t) <- p17*lambda*pi[c_14]+p17/2*lambda*pi[c_12]+p17/2*lambda*pi[c_15]
    D(pi[c_15], t) <- (p17+p18)/2*lambda*pi[c_15]+p18/2*lambda*pi[c_12]+p17/2*lambda*pi[c_13]+p18*lambda*pi[c_14]+p17*lambda*pi[c_16]
    D(pi[c_16], t) <- p18*lambda*pi[c_16]+p18/2*lambda*pi[c_13]+p18/2*lambda*pi[c_15]

    vec[1]<-exp(-lambda*grid[n.grid])*solution[n.grid,1]
    vec[2]<-exp(-lambda*grid[n.grid])*solution[n.grid,2]
    vec[3]<-exp(-lambda*grid[n.grid])*(solution[n.grid,3]+solution[n.grid,4])
    vec[4]<-exp(-lambda*grid[n.grid])*solution[n.grid,5]
    vec[5]<-exp(-lambda*grid[n.grid])*solution[n.grid,6]

###Model mean structure and variance function:
    for(i in 1:n.rep){
        log.H[i]~dnorm(0.0,1.0E-6)  #prior for log(H[i]) with \tau_1=1.0E-6.
        H[i]<-exp(log.H[i])
        mean[i,1]<-H[i]+H[i]*Q*vec[1]
        mean[i,2]<-H[i]*R[1]+H[i]*Q*vec[2]+H[i]*Q*R[1]*vec[1]
        mean[i,3]<-H[i]*R[2]+H[i]*Q*vec[3]+H[i]*Q*R[1]*vec[2]+H[i]*Q*R[2]*vec[1]
        mean[i,4]<-H[i]*R[3]+H[i]*Q*vec[4]+H[i]*Q*R[1]*vec[3]+H[i]*Q*R[2]*vec[2]+H[i]*Q*R[3]*vec[1]
        mean[i,5]<-H[i]*R[4]+H[i]*Q*vec[5]+H[i]*Q*R[1]*vec[4]+H[i]*Q*R[2]*vec[3]+H[i]*Q*R[3]*vec[2]+H[i]*Q*R[4]*vec[1]
        mean[i,6]<-H[i]*R[5]+H[i]*Q*R[1]*vec[5]+H[i]*Q*R[2]*vec[4]+H[i]*Q*R[3]*vec[3]+H[i]*Q*R[4]*vec[2]+H[i]*Q*R[5]*vec[1]
        mean[i,7]<-H[i]*Q*R[2]*vec[5]+H[i]*Q*R[3]*vec[4]+H[i]*Q*R[4]*vec[3]+H[i]*Q*R[5]*vec[2]
        mean[i,8]<-H[i]*Q*R[3]*vec[5]+H[i]*Q*R[4]*vec[4]+H[i]*Q*R[5]*vec[3]
        mean[i,9]<-H[i]*Q*R[4]*vec[5]+H[i]*Q*R[5]*vec[4]
        mean[i,10]<-H[i]*Q*R[5]*vec[5]
        for(j in 1:n.obs){
              sigmafun[i,j]<-1/tau*pow(mean[i,j],2*theta)  #variance function \sigma^2*mu[i,j]^{2\theta}
              taufun[i,j]<-1/sigmafun[i,j]
              y[i,j]~dnorm(mean[i,j],taufun[i,j])
        }
    }

#   Initial conditions:
    init[1] <- 0.167; init[2] <- 0.167; init[3] <- 0.167; init[4] <- 0.167; init[5] <- 0.167; init[6] <- 0.167;
    #   Prior distributions:
    tau~dgamma(0.001,0.001)  #prior for 1/sigma^2 with \alpha and \beta=0.001
    sigma<-1/sqrt(tau)
    lambda.prime~dnorm(0.0,1.0E-6) #prior for \lambda' with \tau_5=1.0E-6
    lambda<-20/120*exp(lambda.prime)/(exp(lambda.prime)+1)  #back-transform for lambda
    log.Q~dnorm(0.0,1.0E-6) #prior for log(Q) with \tau_2=1.0E-6
    Q<-exp(log.Q)
    for(j in 1:5){
    log.R[j]~dnorm(0.0,1.0E-6) #prior for log(R[j]) with \tau_3=1.0E-6
    R[j]<-exp(log.R[j])
    }
    theta~dnorm(0.0,1.0E-6) #prior for \theta with \tau_4=1.0E-6
}
```

# Appendix D

# Simulation results of the shape model for enzymatically $^{18}$O-labeled mass spectra

**Table D.1:** Mean relative bias $\bar{b}$, empirical standard error $S_{emp}$, model based standard error $S_{mb}$ and mean squared error $MSE$ for both the stick and shape models for $Q$.

| $M_2$ | $Q$ | $\lambda\tau$ | $R$ | $\bar{b}(\times 1e-2)$ | | $S_{emp}/S_{mb}(\times 1e-2)$ | | $MSE(\times 1e-3)$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Shape | Stick | Shape | Stick | Shape | Stick |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | -10.38 | -14.94 | 2.156/2.279 | 3.099/3.713 | 3.159 | 6.541 |
| | | | E2 | -7.761 | -13.67 | 2.061/2.216 | 3.162/3.406 | 1.931 | 5.669 |
| | | 9.6 | E1 | -1.867 | -2.547 | 1.132/1.413 | 2.260/2.476 | 0.215 | 0.673 |
| | | | E2 | -1.647 | -2.142 | 1.090/1.218 | 1.855/2.200 | 0.187 | 0.459 |
| | 2 | 4.8 | E1 | -0.311 | -0.345 | 3.612/3.664 | 6.317/7.079 | 1.343 | 4.039 |
| | | | E2 | -0.588 | -0.231 | 2.951/3.174 | 5.797/6.041 | 1.009 | 3.382 |
| | | 9.6 | E1 | -0.052 | 0.091 | 3.335/3.975 | 5.680/5.904 | 1.113 | 3.230 |
| | | | E2 | -0.183 | 0.428 | 2.797/2.876 | 4.812/5.117 | 0.796 | 2.389 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | -2.913 | -16.55 | 1.169/1.215 | 1.875/2.276 | 0.349 | 7.200 |
| | | | E2 | -3.349 | -16.61 | 1.065/1.159 | 1.590/2.006 | 0.394 | 7.152 |
| | | 9.6 | E1 | 1.098 | -5.372 | 1.024/1.357 | 2.154/2.623 | 0.135 | 1.185 |
| | | | E2 | 0.159 | -5.772 | 0.891/1.213 | 1.894/2.432 | 0.080 | 1.191 |
| | 2 | 4.8 | E1 | -0.065 | -9.800 | 2.779/2.787 | 5.456/5.969 | 0.774 | 41.40 |
| | | | E2 | -0.350 | -8.59 | 2.229/2.258 | 4.960/4.965 | 0.546 | 32.01 |
| | | 9.6 | E1 | 0.020 | -0.263 | 2.474/2.493 | 4.228/5.456 | 0.613 | 1.815 |
| | | | E2 | -0.021 | 0.512 | 2.175/2.261 | 3.412/4.956 | 0.473 | 1.269 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | 1.274 | -17.49 | 1.036/1.516 | 2.062/2.404 | 0.148 | 8.076 |
| | | | E2 | 0.631 | -17.19 | 0.914/1.351 | 1.617/2.212 | 0.094 | 7.652 |
| | | 9.6 | E1 | 2.830 | -6.758 | 0.960/1.301 | 2.509/2.957 | 0.292 | 1.772 |
| | | | E2 | 1.867 | -6.341 | 0.807/1.019 | 1.789/2.817 | 0.152 | 1.325 |
| | 2 | 4.8 | E1 | -0.130 | -12.38 | 2.367/2.411 | 4.618/5.377 | 0.567 | 63.42 |
| | | | E2 | -0.072 | -10.84 | 1.939/2.000 | 3.910/4.453 | 0.378 | 48.53 |
| | | 9.6 | E1 | 0.056 | -0.735 | 2.252/2.410 | 4.217/6.031 | 0.508 | 1.994 |
| | | | E2 | -0.089 | -0.272 | 1.719/1.981 | 3.843/5.390 | 0.299 | 1.507 |

**Table D.2:** Mean relative bias $\bar{b}$, empirical standard error $S_{emp}$, model based standard error $S_{mb}$ and mean squared error $MSE$ for both the stick and shape models for $\lambda\tau$.

| $M_2$ | $Q$ | $\lambda\tau$ | $R$ | $\bar{b}(\times 1e-2)$ | | $S_{emp}/S_{mb}(\times 1e-2)$ | | $MSE$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Shape | Stick | Shape | Stick | Shape | Stick |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | 18.68 | 52.93 | 39.51/39.83 | 226.5/355.5 | 0.960 | 11.58 |
| | | | E2 | 18.19 | 46.59 | 35.55/37.27 | 195.9/241.2 | 0.889 | 8.839 |
| | | 9.6 | E1 | 0.758 | 5.172 | 53.21/53.88 | 272.5/215.6 | 0.288 | 7.672 |
| | | | E2 | 1.985 | 14.91 | 67.24/73.56 | 370.8/557.2 | 0.488 | 15.79 |
| | 2 | 4.8 | E1 | -0.756 | -0.643 | 7.610/8.794 | 14.04/14.92 | 0.007 | 0.021 |
| | | | E2 | -0.452 | 0.326 | 7.298/7.605 | 14.24/14.76 | 0.006 | 0.021 |
| | | 9.6 | E1 | -4.816 | -0.324 | 38.88/54.20 | 167.7/272.9 | 0.365 | 2.813 |
| | | | E2 | -2.819 | 0.047 | 38.95/55.23 | 143.3/194.9 | 0.225 | 2.054 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | 4.587 | 316.7 | 31.28/40.33 | 0.00007/0.005 | 0.146 | 231.0 |
| | | | E2 | 6.751 | 316.7 | 32.00/33.06 | 0.00008/0.006 | 0.207 | 231.0 |
| | | 9.6 | E1 | 1.472 | 108.3 | 55.63/67.21 | 0.0001/0.005 | 0.330 | 108.2 |
| | | | E2 | 1.002 | 108.3 | 48.08/68.54 | 0.0001/0.005 | 0.240 | 108.2 |
| | 2 | 4.8 | E1 | 0.725 | 75.22 | 6.115/7.358 | 90.05/100.0 | 0.005 | 13.85 |
| | | | E2 | 1.069 | 77.54 | 5.685/6.646 | 87.07/92.38 | 0.006 | 14.61 |
| | | 9.6 | E1 | -3.427 | 108.3 | 44.47/44.95 | 0.00007/0.005 | 0.306 | 108.2 |
| | | | E2 | -1.806 | 108.3 | 41.58/43.18 | 0.0001/0.006 | 0.203 | 108.2 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | 5.003 | 316.7 | 15.78/59.25 | 0.00004/0.004 | 0.083 | 231.0 |
| | | | E2 | 4.675 | 316.7 | 14.21/62.75 | 0.00008/0.005 | 0.071 | 231.0 |
| | | 9.6 | E1 | -7.310 | 108.3 | 55.15/56.07 | 0.0001/0.005 | 0.797 | 108.2 |
| | | | E2 | -4.986 | 108.3 | 63.07/79.66 | 0.0002/0.007 | 0.627 | 108.2 |
| | 2 | 4.8 | E1 | -0.392 | 130.7 | 5.581/6.850 | 205.2/567.3 | 0.003 | 43.58 |
| | | | E2 | -0.246 | 129.9 | 5.358/5.931 | 219.8/432.8 | 0.003 | 43.69 |
| | | 9.6 | E1 | -6.251 | 108.3 | 31.29/38.15 | 0.00006/0.004 | 0.458 | 108.2 |
| | | | E2 | -5.320 | 108.3 | 29.85/36.51 | 0.00008/0.005 | 0.350 | 108.2 |

**Table D.3:** Mean relative bias $\bar{b}$, empirical standard error $S_{emp}$, model based standard error $S_{mb}$ and mean squared error $MSE$ for both the stick and shape models for $R_2$.

| $M_2$ | $Q$ | $\lambda\tau$ | $R$ | $\bar{b}(\times 1e-3)$ | | $S_{emp}/S_{mb}(\times 1e-2)$ | | $MSE(\times 1e-3)$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Shape | Stick | Shape | Stick | Shape | Stick |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | 1.803 | -36.89 | 2.049/2.267 | 3.448/4.082 | 0.423 | 2.486 |
| | | | E2 | 2.430 | -48.70 | 2.527/2.588 | 4.293/5.008 | 0.648 | 5.673 |
| | | 9.6 | E1 | 3.732 | -35.56 | 1.994/2.395 | 3.443/4.000 | 0.411 | 2.391 |
| | | | E2 | 3.470 | -48.83 | 2.497/2.565 | 3.823/4.740 | 0.643 | 5.312 |
| | 2 | 4.8 | E1 | 1.520 | -42.58 | 1.384/1.467 | 2.490/2.641 | 0.194 | 2.348 |
| | | | E2 | 1.941 | -46.41 | 1.603/1.674 | 2.700/3.070 | 0.263 | 4.207 |
| | | 9.6 | E1 | 2.684 | -35.43 | 1.230/1.316 | 2.028/2.342 | 0.158 | 1.608 |
| | | | E2 | 1.586 | -40.50 | 1.485/1.562 | 2.492/2.735 | 0.225 | 3.270 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | 3.590 | -19.16 | 2.059/2.253 | 3.523/4.665 | 0.436 | 1.591 |
| | | | E2 | 0.677 | -30.48 | 2.257/2.439 | 4.329/6.036 | 0.510 | 3.374 |
| | | 9.6 | E1 | 2.685 | -13.29 | 2.074/2.346 | 3.818/5.076 | 0.437 | 1.626 |
| | | | E2 | -0.431 | -15.46 | 2.256/2.516 | 4.573/6.729 | 0.509 | 2.477 |
| | 2 | 4.8 | E1 | 1.010 | -20.73 | 1.343/1.742 | 2.195/2.608 | 0.182 | 0.891 |
| | | | E2 | -0.917 | -38.55 | 1.457/1.469 | 2.569/2.867 | 0.214 | 3.059 |
| | | 9.6 | E1 | 1.053 | 4.062 | 1.198/1.121 | 2.306/2.786 | 0.145 | 0.548 |
| | | | E2 | 2.036 | 12.59 | 1.324/1.497 | 2.759/3.588 | 0.182 | 1.017 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | 8.092 | -10.85 | 1.993/2.187 | 3.521/5.243 | 0.541 | 1.352 |
| | | | E2 | 5.240 | -16.40 | 2.353/2.366 | 4.954/7.236 | 0.598 | 2.889 |
| | | 9.6 | E1 | 6.374 | 2.507 | 1.811/2.153 | 3.758/6.119 | 0.367 | 1.418 |
| | | | E2 | 6.954 | 14.24 | 2.340/2.403 | 4.983/8.632 | 0.626 | 2.811 |
| | 2 | 4.8 | E1 | 4.173 | 0.831 | 1.177/1.226 | 2.416/2.693 | 0.155 | 0.584 |
| | | | E2 | 0.943 | -15.06 | 1.494/1.438 | 2.862/2.948 | 0.225 | 1.185 |
| | | 9.6 | E1 | 3.218 | 36.58 | 1.151/1.120 | 2.139/3.213 | 0.143 | 1.733 |
| | | | E2 | 3.251 | 47.55 | 1.149/1.296 | 3.192/4.268 | 0.149 | 4.670 |

**Table D.4:** Mean relative bias $\bar{b}$, empirical standard error $S_{emp}$, model based standard error $S_{mb}$ and mean squared error $MSE$ for both the stick and shape models for $R_3$.

| $M_2$ | $Q$ | $\lambda\tau$ | $R$ | $\bar{b}(\times 1e-2)$ | | $S_{emp}/S_{mb}(\times 1e-2)$ | | $MSE(\times 1e-4)$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Shape | Stick | Shape | Stick | Shape | Stick |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | 7.823 | 10.76 | 2.049/2.416 | 3.512/3.838 | 29.15 | 59.57 |
| | | | E2 | 4.229 | 6.596 | 2.324/2.639 | 4.107/4.571 | 19.48 | 51.11 |
| | | 9.6 | E1 | 3.680 | 4.803 | 1.549/2.001 | 2.791/3.214 | 7.921 | 17.19 |
| | | | E2 | 2.215 | 2.828 | 2.126/2.188 | 3.866/3.912 | 8.383 | 21.24 |
| | 2 | 4.8 | E1 | 0.457 | -1.085 | 1.162/1.248 | 2.353/2.478 | 1.436 | 6.018 |
| | | | E2 | 0.472 | -1.281 | 1.405/1.406 | 2.668/2.850 | 2.151 | 8.410 |
| | | 9.6 | E1 | 0.219 | -1.308 | 1.010/1.077 | 1.745/1.906 | 1.391 | 3.742 |
| | | | E2 | 0.274 | -1.251 | 1.167/1.981 | 2.223/2.221 | 1.421 | 6.173 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | 3.917 | 5.951 | 1.820/1.924 | 3.024/3.695 | 9.568 | 23.58 |
| | | | E2 | 2.046 | 4.513 | 2.007/2.066 | 3.867/5.011 | 7.324 | 30.98 |
| | | 9.6 | E1 | 2.283 | 6.558 | 1.416/1.987 | 3.244/4.082 | 4.131 | 28.06 |
| | | | E2 | 1.111 | 6.303 | 1.766/2.095 | 4.151/5.680 | 4.089 | 48.50 |
| | 2 | 4.8 | E1 | 0.214 | -6.346 | 0.122/1.588 | 1.820/2.055 | 1.497 | 19.73 |
| | | | E2 | 0.089 | -7.331 | 1.366/1.405 | 2.200/2.258 | 1.872 | 47.14 |
| | | 9.6 | E1 | 0.404 | -0.941 | 0.926/0.951 | 1.635/2.144 | 0.924 | 3.034 |
| | | | E2 | 0.191 | -0.062 | 1.095/1.175 | 2.358/2.788 | 1.228 | 5.561 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | 1.957 | 6.370 | 1.609/1.700 | 3.100/4.261 | 4.153 | 26.15 |
| | | | E2 | 1.188 | 5.908 | 1.821/1.958 | 4.679/6.126 | 4.429 | 49.37 |
| | | 9.6 | E1 | 1.806 | 9.054 | 1.454/1.835 | 3.691/4.892 | 3.443 | 47.05 |
| | | | E2 | 1.342 | 10.08 | 1.761/1.994 | 4.539/7.132 | 4.521 | 100.5 |
| | 2 | 4.8 | E1 | 0.550 | -5.211 | 1.076/1.134 | 1.805/2.041 | 1.280 | 14.33 |
| | | | E2 | 0.065 | -6.517 | 1.344/1.312 | 2.120/2.245 | 1.809 | 37.93 |
| | | 9.6 | E1 | 0.356 | -0.104 | 0.880/0.963 | 1.661/2.419 | 0.826 | 2.763 |
| | | | E2 | 0.367 | 1.706 | 1.003/1.092 | 2.391/3.189 | 1.113 | 8.009 |

**Table D.5:** Mean relative bias $\bar{b}$, empirical standard error $S_{emp}$, model based standard error $S_{mb}$ and mean squared error $MSE$ for both the stick and shape models for $R_4$.

| $M_2$ | $Q$ | $\lambda\tau$ | $R$ | $\bar{b}(\times 1e-2)$ | | $S_{emp}/S_{mb}(\times 1e-2)$ | | $MSE(\times 1e-4)$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Shape | Stick | Shape | Stick | Shape | Stick |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | 17.90 | 20.91 | 1.554/1.613 | 2.537/2.750 | 33.71 | 49.12 |
| | | | E2 | 11.63 | 13.78 | 2.025/2.154 | 3.477/3.323 | 30.64 | 49.36 |
| | | 9.6 | E1 | 9.944 | 9.369 | 1.221/1.692 | 2.048/2.141 | 11.15 | 12.77 |
| | | | E2 | 6.104 | 4.429 | 1.502/1.668 | 2.556/2.552 | 9.569 | 10.38 |
| | 2 | 4.8 | E1 | 1.712 | -0.262 | 0.826/1.025 | 1.577/2.503 | 0.968 | 2.493 |
| | | | E2 | 1.274 | -1.247 | 1.026/1.069 | 1.869/1.948 | 1.371 | 3.800 |
| | | 9.6 | E1 | 1.504 | 0.485 | 0.719/0.822 | 1.344/2.284 | 0.738 | 1.829 |
| | | | E2 | 0.767 | -1.098 | 0.828/0.876 | 1.576/2.479 | 0.801 | 2.722 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | 10.76 | 8.331 | 1.237/1.616 | 1.881/2.574 | 12.83 | 10.31 |
| | | | E2 | 6.405 | 4.394 | 1.271/1.626 | 2.4483.245 | 9.669 | 9.785 |
| | | 9.6 | E1 | 8.339 | 7.998 | 1.044/1.700 | 2.195/2.708 | 7.879 | 11.07 |
| | | | E2 | 4.849 | 5.599 | 1.271/1.651 | 2.762/3.689 | 6.232 | 13.78 |
| | 2 | 4.8 | E1 | 1.216 | 3.125 | 0.813/0.936 | 1.526/2.484 | 0.805 | 3.283 |
| | | | E2 | 0.329 | -1.785 | 0.930/0.965 | 1.649/2.597 | 0.886 | 3.344 |
| | | 9.6 | E1 | 1.532 | -2.322 | 0.728/0.808 | 1.285/1.457 | 0.759 | 2.178 |
| | | | E2 | 0.935 | -2.001 | 0.838/0.843 | 1.630/1.877 | 0.874 | 3.443 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | 5.285 | 7.798 | 1.178/1.454 | 2.334/2.839 | 4.115 | 11.39 |
| | | | E2 | 2.958 | 4.657 | 1.320/1.540 | 3.153/3.938 | 3.461 | 14.20 |
| | | 9.6 | E1 | 6.142 | 11.26 | 1.108/1.584 | 2.287/3.248 | 4.912 | 17.60 |
| | | | E2 | 3.383 | 9.425 | 1.300/1.599 | 3.175/4.588 | 3.936 | 27.52 |
| | 2 | 4.8 | E1 | 1.446 | 4.524 | 0.714/0.919 | 1.438/1.465 | 0.715 | 4.066 |
| | | | E2 | 0.805 | -1.608 | 0.845/0.970 | 1.478/1.573 | 0.841 | 2.691 |
| | | 9.6 | E1 | 1.565 | -0.807 | 0.652/0.819 | 1.302/1.624 | 0.664 | 1.760 |
| | | | E2 | 0.863 | 0.750 | 7.987/8.592 | 1.773/2.176 | 0.784 | 3.256 |

**Table D.6:** Mean relative bias $\bar{b}$, empirical standard error $S_{emp}$, model based standard error $S_{mb}$ and mean squared error $MSE$ for both the stick and shape models for $R_5$.

| $M_2$ | $Q$ | $\lambda\tau$ | $R$ | $\bar{b}(\times 1e-2)$ | | $S_{emp}/S_{mb}(\times 1e-2)$ | | $MSE(\times 1e-4)$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Shape | Stick | Shape | Stick | Shape | Stick |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | 20.66 | 45.41 | 1.259/1.775 | 2.323/2.405 | 8.337 | 38.01 |
| | | | E2 | 12.35 | 35.06 | 1.506/1.956 | 2.673/2.683 | 6.938 | 44.78 |
| | | 9.6 | E1 | 13.20 | 33.35 | 1.188/1.891 | 2.023/2.264 | 4.168 | 21.6 |
| | | | E2 | 8.593 | 22.00 | 1.267/1.628 | 2.579/3.491 | 3.868 | 21.47 |
| | 2 | 4.8 | E1 | 6.111 | 8.974 | 0.667/1.143 | 1.079/1.300 | 1.035 | 2.438 |
| | | | E2 | 4.480 | 3.263 | 0.768/1.226 | 1.268/2.200 | 1.205 | 1.933 |
| | | 9.6 | E1 | 4.375 | 6.013 | 0.566/0.926 | 0.931/0.948 | 0.623 | 1.438 |
| | | | E2 | 2.826 | 2.743 | 0.695/0.970 | 1.149/2.099 | 0.727 | 1.551 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | 13.35 | 15.65 | 0.822/1.56 | 2.146/3.106 | 3.494 | 8.482 |
| | | | E2 | 8.827 | 6.551 | 0.998/1.788 | 2.200/4.065 | 3.383 | 6.156 |
| | | 9.6 | E1 | 12.21 | 18.65 | 0.797/1.509 | 2.194/3.208 | 2.993 | 10.32 |
| | | | E2 | 7.042 | 8.016 | 0.873/2.107 | 3.182/4.724 | 2.282 | 12.09 |
| | 2 | 4.8 | E1 | 0.753 | 4.763 | 0.520/0.976 | 1.028/1.039 | 0.279 | 1.415 |
| | | | E2 | -0.115 | -1.967 | 0.659/0.942 | 1.186/1.172 | 0.434 | 1.526 |
| | | 9.6 | E1 | 0.365 | 6.681 | 0.503/0.833 | 0.980/1.236 | 0.255 | 1.667 |
| | | | E2 | 0.784 | 3.876 | 0.700/0.822 | 1.188/1.537 | 0.509 | 1.872 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | 12.87 | 14.53 | 0.710/1.668 | 2.536/3.990 | 3.125 | 9.771 |
| | | | E2 | 8.353 | 2.883 | 0.862/1.604 | 3.273/5.452 | 2.879 | 10.96 |
| | | 9.6 | E1 | 12.43 | 21.15 | 0.655/1.481 | 2.795/4.020 | 2.873 | 14.88 |
| | | | E2 | 8.507 | 13.67 | 0.814/1.806 | 3.797/5.744 | 2.880 | 20.14 |
| | 2 | 4.8 | E1 | 3.975 | 4.669 | 0.538/0.907 | 0.990/1.062 | 0.539 | 1.324 |
| | | | E2 | 2.562 | -2.192 | 0.636/0.896 | 1.271/1.815 | 0.606 | 1.763 |
| | | 9.6 | E1 | 3.416 | 9.313 | 0.444/0.791 | 0.923/1.351 | 0.382 | 2.224 |
| | | | E2 | 2.746 | 5.065 | 0.549/0.804 | 1.300/1.797 | 0.532 | 2.477 |

**Table D.7:** Mean relative bias $\bar{b}$, empirical standard error $S_{emp}$, model based standard error $S_{mb}$ and mean squared error $MSE$ for both the stick and shape models for $R_6$.

| $M_2$ | $Q$ | $\lambda\tau$ | $R$ | $\bar{b}$ | | $S_{emp}/S_{mb}(\times 1e-2)$ | | $MSE(\times 1e-4)$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Shape | Stick | Shape | Stick | Shape | Stick |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | 0.266 | 0.930 | 0.605/1.522 | 1.301/1.529 | 1.691 | 17.87 |
| | | | E2 | 0.193 | 0.805 | 0.772/3.595 | 1.566/1.622 | 1.835 | 23.97 |
| | | 9.6 | E1 | 0.207 | 0.796 | 0.574/1.242 | 1.166/1.475 | 1.313 | 13.20 |
| | | | E2 | 0.166 | 0.673 | 0.638/2.091 | 1.212/1.651 | 1.322 | 16.49 |
| | 2 | 4.8 | E1 | 0.127 | 0.352 | 0.368/2.835 | 0.711/0.737 | 0.437 | 2.826 |
| | | | E2 | 0.104 | 0.278 | 0.401/0.794 | 0.785/0.850 | 0.519 | 3.174 |
| | | 9.6 | E1 | 0.103 | 0.304 | 0.316/0.794 | 0.605/0.621 | 0.299 | 2.099 |
| | | | E2 | 0.078 | 0.232 | 0.378/1.270 | 0.707/0.712 | 0.346 | 2.285 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | 0.211 | 0.755 | 0.417/2.356 | 1.144/1.836 | 1.003 | 11.97 |
| | | | E2 | 0.152 | 0.674 | 0.521/2.982 | 1.399/2.069 | 1.039 | 17.01 |
| | | 9.6 | E1 | 0.197 | 0.823 | 0.464/1.686 | 1.200/1.749 | 0.937 | 14.11 |
| | | | E2 | 0.128 | 0.743 | 0.504/2.958 | 1.413/2.170 | 0.800 | 20.32 |
| | 2 | 4.8 | E1 | 0.045 | 0.273 | 0.305/1.145 | 0.623/0.664 | 0.131 | 1.800 |
| | | | E2 | 0.023 | 0.182 | 0.382/0.949 | 0.740/0.769 | 0.163 | 1.651 |
| | | 9.6 | E1 | 0.041 | 0.353 | 0.308/0.941 | 0.649/0.920 | 0.126 | 2.752 |
| | | | E2 | 0.020 | 0.283 | 0.358/0.828 | 0.692/1.183 | 0.141 | 3.142 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | 0.202 | 0.816 | 0.374/2.245 | 1.310/1.916 | 0.905 | 14.17 |
| | | | E2 | 0.167 | 0.689 | 0.478/2.655 | 1.670/2.376 | 1.157 | 18.52 |
| | | 9.6 | E1 | 0.185 | 0.923 | 0.401/2.574 | 1.294/1.918 | 0.801 | 17.61 |
| | | | E2 | 0.146 | 0.826 | 0.464/3.374 | 1.511/2.072 | 0.924 | 24.91 |
| | 2 | 4.8 | E1 | 0.077 | 0.305 | 0.302/0.951 | 0.590/0.700 | 0.203 | 2.091 |
| | | | E2 | 0.054 | 0.215 | 0.373/0.884 | 0.663/0.806 | 0.235 | 1.975 |
| | | 9.6 | E1 | 0.055 | 0.398 | 0.291/0.829 | 0.663/1.021 | 0.142 | 3.403 |
| | | | E2 | 0.043 | 0.348 | 0.361/0.790 | 0.901/1.690 | 0.192 | 4.833 |

**Table D.8:** Mean estiamtes $\bar{\mu}$, mean relative bias $\bar{b}$, empirical standard error $S_{emp}$ and model based standard error $S_{mb}$ for both the stick and shape models for $M_1$.

| $M_2$ | $Q$ | True | $\lambda\tau$ | $R$ | $\bar{\mu}$ | $\bar{b}$ $(\times 1e-7)$ | $S_{emp}/S_{mb}$ $(\times 1e-4)$ |
|---|---|---|---|---|---|---|---|
| | | | | | Shape | Shape | Shape |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | 2001.053 | 2001.0534 | 2.012 | 8.959/9.970 |
| | | | E2 | 2001.053 | 2001.0533 | 1.371 | 7.829/9.579 |
| | | 9.6 | E1 | 2001.053 | 2001.0533 | 1.691 | 9.117/11.33 |
| | | | E2 | 2001.053 | 2001.0533 | 1.569 | 8.420/8.976 |
| | 2 | 4.8 | E1 | 2001.053 | 2001.0533 | 1.507 | 7.316/8.434 |
| | | | E2 | 2001.053 | 2001.0531 | 0.733 | 6.051/6.893 |
| | | 9.6 | E1 | 2001.053 | 2001.0532 | 0.929 | 7.776/8.586 |
| | | | E2 | 2001.053 | 2001.0531 | 0.728 | 6.447/7.147 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | 2001.053 | 2001.0532 | 1.184 | 9.461/11.59 |
| | | | E2 | 2001.053 | 2001.0532 | 1.051 | 8.295/12.56 |
| | | 9.6 | E1 | 2001.053 | 2001.0535 | 2.486 | 9.226/12.44 |
| | | | E2 | 2001.053 | 2001.0533 | 1.392 | 8.516/19.20 |
| | 2 | 4.8 | E1 | 2001.053 | 2001.0534 | 1.808 | 7.486/9.296 |
| | | | E2 | 2001.053 | 2001.0533 | 1.549 | 6.909/7.740 |
| | | 9.6 | E1 | 2001.053 | 2001.0532 | 1.156 | 7.995/8.954 |
| | | | E2 | 2001.053 | 2001.0532 | 1.012 | 6.515/7.569 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | 2001.053 | 2001.0538 | 3.750 | 8.632/8.930 |
| | | | E2 | 2001.053 | 2001.0537 | 3.553 | 7.679/10.52 |
| | | 9.6 | E1 | 2001.053 | 2001.0539 | 4.412 | 9.128/10.14 |
| | | | E2 | 2001.053 | 2001.0538 | 3.838 | 8.465/11.02 |
| | 2 | 4.8 | E1 | 2001.053 | 2001.0531 | 0.667 | 7.713/11.21 |
| | | | E2 | 2001.053 | 2001.0531 | 0.446 | 6.848/8.571 |
| | | 9.6 | E1 | 2001.053 | 2001.0532 | 0.846 | 8.258/10.39 |
| | | | E2 | 2001.053 | 2001.0533 | 1.301 | 6.692/8.512 |

**Table D.9:** Mean estiamtes $\bar{\mu}$, mean relative bias $\bar{b}$, empirical standard error $S_{emp}$ and model based standard error $S_{mb}$ for both the stick and shape models for $M_2$.

| $M_2$ | $Q$ | True | $\lambda\tau$ | $R$ | $\bar{\mu}$ | $\bar{b}(\times 1e-7)$ | $S_{emp}/S_{mb}(\times 1e-4)$ |
|---|---|---|---|---|---|---|---|
| | | | | | Shape | Shape | Shape |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | 2005.059 | 2005.0594 | 1.787e | 14.43/19.46 |
| | | | E2 | 2005.059 | 2005.0592 | 0.754 | 13.58/16.74 |
| | | 9.6 | E1 | 2005.059 | 2005.0590 | -0.032 | 13.81/18.77 |
| | | | E2 | 2005.059 | 2005.0593 | 1.447 | 13.59/15.30 |
| | 2 | 4.8 | E1 | 2005.059 | 2005.0592 | 0.924 | 6.361/7.553 |
| | | | E2 | 2005.059 | 2005.0592 | 0.850 | 5.235/6.237 |
| | | 9.6 | E1 | 2005.059 | 2005.0592 | 1.130 | 5.854/6.986 |
| | | | E2 | 2005.059 | 2005.0591 | 0.411 | 5.497/5.865 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | 2005.179 | 2005.1784 | -3.104 | 14.64/18.78 |
| | | | E2 | 2005.179 | 2005.1785 | -2.349 | 12.51/15.30 |
| | | 9.6 | E1 | 2005.179 | 2005.1789 | -0.509 | 12.84/19.52 |
| | | | E2 | 2005.179 | 2005.1790 | -0.131 | 11.15/18.26 |
| | 2 | 4.8 | E1 | 2005.179 | 2005.1792 | 1.209 | 5.359/6.688 |
| | | | E2 | 2005.179 | 2005.1791 | 0.698 | 4.773/5.676 |
| | | 9.6 | E1 | 2005.179 | 2005.1793 | 1.363 | 5.666/6.582 |
| | | | E2 | 2005.179 | 2005.1792 | 1.008 | 4.705/5.859 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | 2005.299 | 2005.3007 | 8.532 | 12.54/14.96 |
| | | | E2 | 2005.299 | 2005.3005 | 7.305 | 10.06/14.25 |
| | | 9.6 | E1 | 2005.299 | 2005.3009 | 9.599 | 13.46/15.95 |
| | | | E2 | 2005.299 | 2005.3008 | 8.943 | 12.05/13.59 |
| | 2 | 4.8 | E1 | 2005.299 | 2005.2995 | 2.562 | 5.367/7.525 |
| | | | E2 | 2005.299 | 2005.2995 | 2.403 | 4.614/7.048 |
| | | 9.6 | E1 | 2005.299 | 2005.2997 | 3.577 | 5.532/7.667 |
| | | | E2 | 2005.299 | 2005.2997 | 3.266 | 4.667/7.407 |

**Table D.10:** Mean estiamtes $\bar{\mu}$, mean relative bias $\bar{b}$, empirical standard error $S_{emp}$ and model based standard error $S_{mb}$ for both the stick and shape models for $\kappa$.

| $M_2$ | $Q$ | True | $\lambda\tau$ | $R$ | $\bar{\mu}$ | $\bar{b}(\times 1e-3)$ | $S_{emp}/S_{mb}(\times 1e-3)$ |
|---|---|---|---|---|---|---|---|
| | | | | | Shape | Shape | Shape |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | 0.75 | 0.7550 | 6.699 | 5.797/9.803 |
| | | | E2 | 0.75 | 0.7538 | 5.005 | 5.870/9.774 |
| | | 9.6 | E1 | 0.75 | 0.7554 | 7.212 | 6.746/10.77 |
| | | | E2 | 0.75 | 0.7546 | 6.075 | 6.103/8.859 |
| | 2 | 4.8 | E1 | 0.75 | 0.7535 | 4.607 | 4.707/9.738 |
| | | | E2 | 0.75 | 0.7528 | 3.716 | 4.382/6.884 |
| | | 9.6 | E1 | 0.75 | 0.7528 | 3.739 | 5.105/8.467 |
| | | | E2 | 0.75 | 0.7513 | 1.697 | 4.500/6.515 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | 0.75 | 0.7511 | 1.503 | 7.902/11.10 |
| | | | E2 | 0.75 | 0.7497 | -0.390 | 6.667/13.67 |
| | | 9.6 | E1 | 0.75 | 0.7553 | 7.055 | 6.885/10.18 |
| | | | E2 | 0.75 | 0.7537 | 4.892 | 5.959/21.45 |
| | 2 | 4.8 | E1 | 0.75 | 0.7550 | 6.725 | 5.120/9.105 |
| | | | E2 | 0.75 | 0.7530 | 4.016 | 4.848/7.862 |
| | | 9.6 | E1 | 0.75 | 0.7548 | 6.381 | 5.629/8.028 |
| | | | E2 | 0.75 | 0.7531 | 4.074 | 4.435/7.257 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | 0.75 | 0.7493 | -0.904 | 7.636/14.85 |
| | | | E2 | 0.75 | 0.7490 | -1.315 | 6.176/19.26 |
| | | 9.6 | E1 | 0.75 | 0.7557 | 7.664 | 7.339/13.01 |
| | | | E2 | 0.75 | 0.7542 | 5.590 | 7.339/18.26 |
| | 2 | 4.8 | E1 | 0.75 | 0.7519 | 2.517 | 5.683/8.901 |
| | | | E2 | 0.75 | 0.7512 | 1.580 | 4.208/7.472 |
| | | 9.6 | E1 | 0.75 | 0.7526 | 3.472 | 5.934/8.105 |
| | | | E2 | 0.75 | 0.7524 | 3.236 | 4.914/6.983 |

**Table D.11:** Mean estiamtes $\bar{\mu}$, mean relative bias $\bar{b}$, empirical standard error $S_{emp}$ and model based standard error $S_{mb}$ for both the stick and shape models for $\sigma_s$.

| $M_2$ | $Q$ | True | $\lambda\tau$ | $R$ | $\bar{\mu}$ | $\bar{b}$ | $S_{emp}/S_{mb}$ ($\times 1e-3$) |
|---|---|---|---|---|---|---|---|
| | | | | | Shape | Shape | Shape |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | 0.08 | 0.0889 | 0.111 | 1.253/1.569 |
| | | | E2 | 0.08 | 0.0870 | 0.088 | 1.014/1.248 |
| | | 9.6 | E1 | 0.08 | 0.0891 | 0.113 | 0.979/1.647 |
| | | | E2 | 0.08 | 0.0872 | 0.090 | 1.032/1.065 |
| | 2 | 4.8 | E1 | 0.08 | 0.0841 | 0.051 | 0.723/1.729 |
| | | | E2 | 0.08 | 0.0831 | 0.039 | 0.609/2.643 |
| | | 9.6 | E1 | 0.08 | 0.0839 | 0.049 | 0.698/1.333 |
| | | | E2 | 0.08 | 0.0829 | 0.037 | 0.596/2.630 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | 0.08 | 0.0899 | 0.123 | 1.244/1.245 |
| | | | E2 | 0.08 | 0.0878 | 0.098 | 1.085/1.224 |
| | | 9.6 | E1 | 0.08 | 0.0895 | 0.119 | 1.182/1.394 |
| | | | E2 | 0.08 | 0.0875 | 0.093 | 1.121/1.402 |
| | 2 | 4.8 | E1 | 0.08 | 0.0844 | 0.055 | 0.814/1.426 |
| | | | E2 | 0.08 | 0.0834 | 0.043 | 0.631/3.568 |
| | | 9.6 | E1 | 0.08 | 0.0840 | 0.050 | 0.779/1.642 |
| | | | E2 | 0.08 | 0.0831 | 0.038 | 0.588/2.687 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | 0.08 | 0.0885 | 0.106 | 1.195/1.337 |
| | | | E2 | 0.08 | 0.0868 | 0.085 | 0.991/1.167 |
| | | 9.6 | E1 | 0.08 | 0.0884 | 0.105 | 1.283/1.392 |
| | | | E2 | 0.08 | 0.0868 | 0.084 | 0.960/1.039 |
| | 2 | 4.8 | E1 | 0.08 | 0.0837 | 0.047 | 0.817/2.702 |
| | | | E2 | 0.08 | 0.0828 | 0.035 | 0.588/2.187 |
| | | 9.6 | E1 | 0.08 | 0.0836 | 0.045 | 0.767/2.482 |
| | | | E2 | 0.08 | 0.0827 | 0.034 | 0.674/2.020 |

**Table D.12:** Mean estiamtes $\bar{\mu}$, mean relative bias $\bar{b}$, empirical standard error $S_{emp}$ and model based standard error $S_{mb}$ for both the stick and shape models for $S$.

| $M_2$ | $Q$ | True | $\lambda\tau$ | $R$ | $\bar{\mu}$ | $\bar{b}(\times 1e-5)$ | $S_{emp}/S_{mb}(\times 1e-4)$ |
|---|---|---|---|---|---|---|---|
| | | | | | Shape | Shape | Shape |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | 1.0015 | 1.0014 | -5.306 | 5.160/8.717 |
| | | | E2 | 1.0015 | 1.0015 | -3.025 | 4.838/5/119 |
| | | 9.6 | E1 | 1.0015 | 1.0015 | 0.541 | 5.473/6.943 |
| | | | E2 | 1.0015 | 1.0015 | -3.005 | 4.695/5.922 |
| | 2 | 4.8 | E1 | 1.0015 | 1.0015 | -1.200 | 3.485/6.706 |
| | | | E2 | 1.0015 | 1.0015 | -2.473 | 3.232/4.970 |
| | | 9.6 | E1 | 1.0015 | 1.0015 | -2.376 | 3.340/6.615 |
| | | | E2 | 1.0015 | 1.0015 | -3.850 | 2.894/4.731 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | 1.0015 | 1.0019 | 42.42 | 5.341/6.971 |
| | | | E2 | 1.0015 | 1.0018 | 26.64 | 4.953/5.137 |
| | | 9.6 | E1 | 1.0015 | 1.0016 | 13.78 | 5.445/5.793 |
| | | | E2 | 1.0015 | 1.0015 | 4.505 | 4.598/4.785 |
| | 2 | 4.8 | E1 | 1.0015 | 1.0015 | -3.395 | 2.982/3.999 |
| | | | E2 | 1.0015 | 1.0015 | -3.439 | 2.479/3.097 |
| | | 9.6 | E1 | 1.0015 | 1.0014 | -5.295 | 3.065/4.196 |
| | | | E2 | 1.0015 | 1.0014 | -7.566 | 2.555/2.605 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | 1.0015 | 1.0006 | -86.22 | 4.370/13.83 |
| | | | E2 | 1.0015 | 1.0008 | -71.16 | 3.946/13.23 |
| | | 9.6 | E1 | 1.0015 | 1.0007 | -82.27 | 4.840/5.396 |
| | | | E2 | 1.0015 | 1.0008 | -66.24 | 3.914/6.649 |
| | 2 | 4.8 | E1 | 1.0015 | 1.0010 | -46.60 | 2.836/4.355 |
| | | | E2 | 1.0015 | 1.0012 | -33.38 | 2.509/5.072 |
| | | 9.6 | E1 | 1.0015 | 1.0010 | -47.82 | 3.012/5.425 |
| | | | E2 | 1.0015 | 1.0011 | -37.40 | 2.536/3.208 |

**Table D.13:** Mean relative bias $\bar{b}$, empirical standard error $S_{emp}$, model based standard error $S_{mb}$ and mean squared error $MSE$ for both the stick and shape models for $\sigma$.

| $M_2$ | $Q$ | $\lambda\tau$ | $R$ | $\bar{b}$ | $S_{emp}/S_{mb}$ | $MSE$ |
|---|---|---|---|---|---|---|
| | | | | Shape | Shape | Shape |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | -0.221 | 1.757/1.960 | 14.09 |
| | | | E2 | -0.049 | 1.587/1.690 | 3.068 |
| | | 9.6 | E1 | -0.258 | 1.851/2.128 | 18.38 |
| | | | E2 | -0.091 | 1.691/2.116 | 4.713 |
| | 2 | 4.8 | E1 | 0.355 | 1.830/2.248 | 31.65 |
| | | | E2 | 0.530 | 1.867/2.242 | 66.74 |
| | | 9.6 | E1 | 0.478 | 2.193/2.766 | 56.27 |
| | | | E2 | 0.662 | 2.042/2.497 | 0.010 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | -0.256 | 1.744/1.972 | 17.75 |
| | | | E2 | -0.092 | 1.580/1.898 | 4.382 |
| | | 9.6 | E1 | -0.275 | 1.904/2.474 | 20.69 |
| | | | E2 | -0.088 | 1.663/1.851 | 4.442 |
| | 2 | 4.8 | E1 | 0.301 | 1.954/2.149 | 24.54 |
| | | | E2 | 0.529 | 1.800/1.898 | 66.29 |
| | | 9.6 | E1 | 0.439 | 1.945/2.146 | 47.13 |
| | | | E2 | 0.699 | 2.188/2.399 | 0.011 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | -0.125 | 1.611/2.537 | 6.094 |
| | | | E2 | 0.009 | 1.617/2.584 | 2.631 |
| | | 9.6 | E1 | -0.196 | 1.851/2.510 | 12.08 |
| | | | E2 | -0.039 | 1.607/2.571 | 2.928 |
| | 2 | 4.8 | E1 | 0.352 | 1.748/1.948 | 30.92 |
| | | | E2 | 0.575 | 1.994/2.437 | 78.32 |
| | | 9.6 | E1 | 0.397 | 1.948/1.956 | 39.21 |
| | | | E2 | 0.659 | 2.475/2.475 | 0.010 |

**Table D.14:** Mean relative bias $\bar{b}$, empirical standard error $S_{emp}$, model based standard error $S_{mb}$ and mean squared error $MSE$ for both the stick and shape models for $a$.

| $M_2$ | $Q$ | $\lambda\tau$ | $R$ | $\bar{b}(\times 1e-2)$ | $S_{emp}/S_{mb}$ | $MSE(\times 1e-2)$ |
|---|---|---|---|---|---|---|
| | | | | Shape | Shape | Shape |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | 5.145 | 0.284/0.320 | 14.68 |
| | | | E2 | 0.369 | 0.209/0.323 | 4.388 |
| | | 9.6 | E1 | 6.567 | 0.288/0.343 | 19.06 |
| | | | E2 | 1.812 | 0.231/0.320 | 6.173 |
| | 2 | 4.8 | E1 | -5.870 | 0.166/0.311 | 11.38 |
| | | | E2 | -8.427 | 0.154/0.339 | 20.11 |
| | | 9.6 | E1 | -7.983 | 0.185/0.305 | 19.36 |
| | | | E2 | -10.06 | 0.151/0.330 | 27.56 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | 6.224 | 0.305/0.325 | 18.98 |
| | | | E2 | 1.400 | 0.211/0.325 | 4.953 |
| | | 9.6 | E1 | 8.061 | 0.352/0.325 | 28.66 |
| | | | E2 | 1.779 | 0.226/0.333 | 5.882 |
| | 2 | 4.8 | E1 | -5.296 | 0.185/0.314 | 10.43 |
| | | | E2 | -8.737 | 0.141/0.342 | 21.08 |
| | | 9.6 | E1 | -7.338 | 0.168/0.305 | 16.28 |
| | | | E2 | -10.67 | 0.159/0.343 | 30.97 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | 0.562 | 0.219/0.321 | 4.882 |
| | | | E2 | -1.939 | 0.202/0.322 | 5.002 |
| | | 9.6 | E1 | 4.535 | 0.294/0.323 | 13.81 |
| | | | E2 | 0.235 | 0.208/0.323 | 4.337 |
| | 2 | 4.8 | E1 | -6.568 | 0.162/0.338 | 13.41 |
| | | | E2 | -9.545 | 0.149/0.355 | 24.98 |
| | | 9.6 | E1 | -6.390 | 0.177/0.315 | 13.35 |
| | | | E2 | -10.01 | 0.170/0.332 | 27.91 |

**Table D.15:** Mean relative bias $\bar{b}$, empirical standard error $S_{emp}$, model based standard error $S_{mb}$ and mean squared error $MSE$ for both the stick and shape models for $b$.

| $M_2$ | $Q$ | $\lambda\tau$ | $R$ | $\bar{b}$ | $S_{emp}/S_{mb}(\times 1e-2)$ | $MSE(\times 1e-2)$ |
|---|---|---|---|---|---|---|
| | | | | Shape | Shape | Shape |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | 0.175 | 4.482/6.525 | 0.964 |
| | | | E2 | 0.221 | 4.056/5.107 | 1.390 |
| | | 9.6 | E1 | 0.160 | 6.007/9.352 | 1.001 |
| | | | E2 | 0.206 | 4.004/4.756 | 1.225 |
| | 2 | 4.8 | E1 | 0.250 | 3.534/4.850 | 1.689 |
| | | | E2 | 0.254 | 3.595/6.761 | 1.740 |
| | | 9.6 | E1 | 0.263 | 3.962/4.859 | 1.887 |
| | | | E2 | 0.263 | 3.314/3.712 | 1.841 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | 0.193 | 4.456/5.700 | 1.132 |
| | | | E2 | 0.234 | 3.773/4.399 | 1.507 |
| | | 9.6 | E1 | 0.144 | 4.850/5.097 | 0.755 |
| | | | E2 | 0.202 | 3.854/4.106 | 1.168 |
| | 2 | 4.8 | E1 | 0.255 | 3.964/4.950 | 1.876 |
| | | | E2 | 0.277 | 3.460/3.788 | 2.034 |
| | | 9.6 | E1 | 0.248 | 3.446/4.812 | 1.657 |
| | | | E2 | 0.274 | 3.745/4.977 | 2.017 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | 0.269 | 3.878/4.125 | 1.963 |
| | | | E2 | 0.283 | 4.094/5.072 | 2.166 |
| | | 9.6 | E1 | 0.182 | 4.463/5.043 | 1.026 |
| | | | E2 | 0.227 | 3.720/4.021 | 1.430 |
| | 2 | 4.8 | E1 | 0.269 | 3.627/4.666 | 1.938 |
| | | | E2 | 0.282 | 3.494/7.479 | 2.106 |
| | | 9.6 | E1 | 0.236 | 3.589/4.401 | 1.522 |
| | | | E2 | 0.265 | 4.000/9.927 | 1.916 |

**Table D.16:** Mean relative bias $\bar{b}$, empirical standard error $S_{emp}$, model based standard error $S_{mb}$ and mean squared error $MSE$ for both the stick and shape models for $c$.

| $M_2$ | $Q$ | $\lambda\tau$ | $R$ | $\bar{b}$ | $S_{emp}/S_{mb}$ $(\times 1e-2)$ | $MSE$ |
|---|---|---|---|---|---|---|
| | | | | Shape | Shape | Shape |
| $0\sigma_s$ | 0.5 | 4.8 | E1 | 0.129 | 2.164/3.971 | 0.338 |
| | | | E2 | 0.148 | 1.899/2.465 | 0.444 |
| | | 9.6 | E1 | 0.128 | 8.066/12.61 | 0.336 |
| | | | E2 | 0.147 | 1.786/2.319 | 0.436 |
| | 2 | 4.8 | E1 | 0.200 | 1.395/4.406 | 0.813 |
| | | | E2 | 0.211 | 1.658/3.412 | 0.903 |
| | | 9.6 | E1 | 0.206 | 1.342/1.678 | 0.861 |
| | | | E2 | 0.223 | 1.411/2.403 | 1.006 |
| $1.5\sigma_s$ | 0.5 | 4.8 | E1 | 0.129 | 2.037/2.316 | 0.335 |
| | | | E2 | 0.147 | 1.878/1.895 | 0.439 |
| | | 9.6 | E1 | 0.129 | 2.191/2.555 | 0.336 |
| | | | E2 | 0.147 | 1.948/1.993 | 0.435 |
| | 2 | 4.8 | E1 | 0.193 | 1.388/1.737 | 0.756 |
| | | | E2 | 0.211 | 1.350/1.593 | 0.901 |
| | | 9.6 | E1 | 0.200 | 1.324/1.669 | 0.807 |
| | | | E2 | 0.223 | 1.360/3.370 | 1.007 |
| $3\sigma_s$ | 0.5 | 4.8 | E1 | 0.128 | 2.133/2.517 | 0.332 |
| | | | E2 | 0.147 | 1.891/2.598 | 0.436 |
| | | 9.6 | E1 | 0.129 | 2.333/3.426 | 0.336 |
| | | | E2 | 0.147 | 1.822/6.048 | 0.438 |
| | 2 | 4.8 | E1 | 0.189 | 1.426/2.079 | 0.724 |
| | | | E2 | 0.209 | 3.742/4.741 | 0.882 |
| | | 9.6 | E1 | 0.197 | 1.232/2.052 | 0.786 |
| | | | E2 | 0.219 | 4.515/8.938 | 0.977 |

# Appendix E

# Simulation results for the Bayesian model averaging approach to quantify the overlapping peptides

Tables E.1 to E.4 show the average weights of the 100 simulated data sets for each of the 8 models.

**Table E.1:** Weights of the 8 models in settings 1–8 (the ones in bold should receive the largest weight across the 8 models).

| | ave. weight $\bar{\pi}_i$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | set1 | set2 | set3 | set4 | set5 | set6 | set7 | set8 |
| Shift=0 | **0.1354** | **0.1818** | 0.2411 | 0.0000 | **1** | **0.6198** | 0 | 0 |
| Shift=1 | 0.0771 | 0.0703 | **0.5780** | **1.0000** | 0 | 0.3802 | **1** | **1** |
| Shift=2 | 0.2482 | 0.4120 | 0.0475 | 0.0000 | 0 | 0.0000 | 0 | 0 |
| Shift=3 | 0.2310 | 0.1668 | 0.0251 | 0 | 0 | 0.0000 | 0 | 0 |
| Shift=4 | 0.1468 | 0.1078 | 0.0128 | 0 | 0 | 0.0000 | 0 | 0 |
| Shift=5 | 0.0706 | 0.0218 | 0.0232 | 0 | 0 | 0.0000 | 0 | 0 |
| Shift=6 | 0.0507 | 0.0223 | 0.0206 | 0 | 0 | 0.0000 | 0 | 0 |
| Shift=7 | 0.0402 | 0.0171 | 0.0499 | 0 | 0 | 0.0000 | 0 | 0 |

**Table E.2:** Weights of the 8 models in settings 9–16 (the ones in bold should receive the largest weight across the 8 models).

| | ave. weight $\bar{\pi}_i$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | set9 | set10 | set11 | set12 | set13 | set14 | set15 | set16 |
| Shift=0 | **1** | 0 | **1** | **1** | **1** | 0 | **0.1003** | 0 |
| Shift=1 | 0 | **1** | 0 | 0 | 0 | **1** | 0.0895 | **1** |
| Shift=2 | 0 | 0 | 0 | 0 | 0 | 0 | 0.4313 | 0 |
| Shift=3 | 0 | 0 | 0 | 0 | 0 | 0 | 0.1521 | 0 |
| Shift=4 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0897 | 0 |
| Shift=5 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0344 | 0 |
| Shift=6 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0420 | 0 |
| Shift=7 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0607 | 0 |

**Table E.3:** Weights of the 8 models in the 6 settings (the ones in bold should receive the largest weight across the 8 models).

| | ave. weight $\bar{\pi}_i$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | set17 | set18 | set19 | set20 | set21 | set22 | set23 | set24 |
| Shift=0 | 0 | **0.0595** | **0.9800** | 0 | 0 | 0 | 0.0000 | 0 |
| Shift=1 | **1** | 0.0076 | 0.0000 | **1** | 0 | 0 | 0 | 0 |
| Shift=2 | 0 | 0.9306 | 0.0198 | 0 | 0 | 0 | 0 | 0 |
| Shift=3 | 0 | 0.0022 | 0.0002 | 0 | 0 | 0 | 0 | 0 |
| Shift=4 | 0 | 0.0001 | 0.0000 | 0 | **1** | **1.0000** | 0 | 0 |
| Shift=5 | 0 | 0.0000 | 0.0000 | 0 | 0 | 0.0000 | 0.0000 | 0 |
| Shift=6 | 0 | 0.0000 | 0.0000 | 0 | 0 | 0 | **1.0000** | **1** |
| Shift=7 | 0 | 0.0000 | 0.0000 | 0 | 0 | 0 | 0.0000 | 0 |

**Table E.4:** Weights of the 8 models in the 8 settings (the ones in bold should receive the largest weight across the 8 models).

| | ave. weight $\bar{\pi}_i$ | | | | | |
|---|---|---|---|---|---|---|
| | set25 | set26 | set27 | set28 | set29 | set30 |
| Shift=0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Shift=1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Shift=2 | 0 | 0 | 0 | 0 | 0 | 0 |
| Shift=3 | 0 | 0 | 0 | 0 | 0 | 0 |
| Shift=4 | **1** | **1** | 0 | 0 | **1** | 0 |
| Shift=5 | 0 | 0 | 0 | 0 | 0 | 0 |
| Shift=6 | 0 | 0 | **1** | **1** | 0 | **1** |
| Shift=7 | 0 | 0 | 0 | 0 | 0 | 0 |

Tables E.5 and E.12 show the summary statistics of model averaging for the 8 settings.

**Table E.5:** Summary statistics of model averaging (mean estimates $\bar{\hat{\theta}}$, empirical standard errors $\sigma_{emp}$) and model based standard errors $\sigma_{mb}$.

| Parameter | set1 | | | set2 | | | set3 | | | set4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ |
| $R_{2_1}$ | 0.9763 | 1.0157 | 0.0182/0.0200 | 0.9763 | 1.2072 | 0.0334/0.0338 | 1.1577 | 1.2761 | 0.0560/0.0625 | 1.1577 | 1.3685 | 0.1189/0.1346 |
| $R_{3_1}$ | 0.6385 | 0.6556 | 0.0263/0.0281 | 0.6385 | 0.7866 | 0.0416/0.0517 | 0.6702 | 0.8031 | 0.0673/0.0704 | 0.6702 | 0.7419 | 0.0426/0.0468 |
| $R_{4_1}$ | 0.3125 | 0.3047 | 0.0233/0.0292 | 0.3125 | 0.3470 | 0.0533/0.0653 | 0.2586 | 0.3352 | 0.0376/0.0403 | 0.2586 | 0.2796 | 0.0082/0.0182 |
| $R_{5_1}$ | 0.1277 | 0.1111 | 0.0133/0.0169 | 0.1277 | 0.1185 | 0.0276/0.0295 | 0.0749 | 0.1026 | 0.0143/0.0153 | 0.0749 | 0.0932 | 0.0017/0.0070 |
| $M_1^*$ | 2000.90 | 2000.906 | 0.0003/0.0004 | 2000.90 | 2000.934 | 0.0003/0.0004 | 2000.90 | 2000.900 | 0.0004/0.0004 | 2000.90 | 2000.900 | 0.0020/0.0023 |
| $R_{2_2}$ | 1.2708 | 1.1675 | 0.1673/0.1835 | 1.2708 | 1.2461 | 0.0678/0.0934 | 1.1577 | 1.2230 | 0.0801/0.1196 | 1.1577 | 1.1931 | 0.0265/0.0282 |
| $R_{3_2}$ | 0.8872 | 0.6633 | 0.0425/0.0433 | 0.8872 | 0.6556 | 0.0405/0.0425 | 0.6702 | 0.6689 | 0.0166/0.0384 | 0.6702 | 0.6947 | 0.0177/0.0199 |
| $R_{4_2}$ | 0.4431 | 0.2858 | 0.0200/0.0204 | 0.4431 | 0.2792 | 0.0198/0.0192 | 0.2586 | 0.2774 | 0.0075/0.0184 | 0.2586 | 0.2668 | 0.0071/0.0084 |
| $R_{5_2}$ | 0.1750 | 0.0965 | 0.0075/0.0075 | 0.1750 | 0.0958 | 0.0075/0.0077 | 0.0749 | 0.0933 | 0.0024/0.0070 | 0.0749 | 0.0796 | 0.0030/0.0039 |
| $M_2^*$ | 2000.94 | 2003.714 | 1.5196/1.6031 | 2000.94 | 2003.127 | 1.2650/1.3378 | 2001.94 | 2002.337 | 1.3033/1.4417 | 2001.94 | 2001.942 | 0.0012/0.0014 |
| $\sigma$ | 10 | 7.6750 | 0.2184/0.3275 | 10 | 7.7370 | 0.2882/0.3422 | 10 | 7.6530 | 0.3122/0.3482 | 10 | 7.6789 | 0.2915/0.3413 |
| $\sigma_s$ | 0.08 | 0.0815 | 0.0002/0.0003 | 0.08 | 0.0811 | 0.0002/0.0002 | 0.08 | 0.0804 | 0.0005/0.0006 | 0.08 | 0.0797 | 0.0003/0.0004 |
| $S$ | 1.0015 | 1.0026 | 0.0006/0.0007 | 1.0015 | 1.0023 | 0.0005/0.0007 | 1.0015 | 1.0041 | 0.0018/0.0020 | 1.0015 | 1.0008 | 0.0007/0.0007 |
| $H_2/H_1$ | 0.2 | 0.0953 | 0.0729/0.0738 | 5 | 0.0739 | 0.0683/0.0786 | 0.2 | 0.1079 | 0.0568/0.0620 | 5 | 4.7687 | 0.1586/0.1790 |

**Table E.6:** Summary statistics of model averaging (mean estimates $\bar{\hat{\theta}}$, empirical standard errors $\sigma_{emp}$) and model based standard errors $\sigma_{mb}$.

| Parameter | set5 | | | set6 | | | set7 | | | set8 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ |
| $R_{2_1}$ | 1.2708 | 1.2641 | 0.0048/0.0060 | 1.2708 | 1.1777 | 0.0602/0.0677 | 1.1577 | 1.1589 | 0.0053/0.0045 | 1.1577 | 1.1598 | 0.0098/0.0126 |
| $R_{3_1}$ | 0.8872 | 0.8799 | 0.0042/0.0057 | 0.8872 | 0.7574 | 0.0936/0.0949 | 0.6702 | 0.6724 | 0.0039/0.0054 | 0.6702 | 0.6748 | 0.0082/0.0084 |
| $R_{4_1}$ | 0.4431 | 0.4373 | 0.0033/0.0045 | 0.4431 | 0.3431 | 0.0718/0.0728 | 0.2586 | 0.2615 | 0.0033/0.0038 | 0.2586 | 0.2652 | 0.0066/0.0080 |
| $R_{5_1}$ | 0.1750 | 0.1675 | 0.0029/0.0036 | 0.1750 | 0.1185 | 0.0286/0.0297 | 0.0749 | 0.0798 | 0.0027/0.0027 | 0.0749 | 0.0852 | 0.0044/0.0046 |
| $M_1^*$ | 2000.90 | 2000.901 | 0.0004/0.0006 | 2000.90 | 2000.958 | 0.0578/0.0600 | 2000.90 | 2000.900 | 0.0003/0.0003 | 2000.90 | 2000.900 | 0.0005/0.0006 |
| $R_{2_2}$ | 1.1577 | 1.1677 | 0.0096/0.0117 | 1.1577 | 1.1672 | 0.0627/0.0643 | 0.9763 | 0.9741 | 0.0083/0.0104 | 0.9763 | 0.9746 | 0.0041/0.0049 |
| $R_{3_2}$ | 0.6702 | 0.6822 | 0.0084/0.0107 | 0.6702 | 0.6843 | 0.0104/0.0255 | 0.6385 | 0.6365 | 0.0070/0.0089 | 0.6385 | 0.6372 | 0.0035/0.0049 |
| $R_{4_2}$ | 0.2586 | 0.2682 | 0.0066/0.0079 | 0.2586 | 0.2769 | 0.0098/0.0133 | 0.3125 | 0.3077 | 0.0058/0.0074 | 0.3125 | 0.3103 | 0.0029/0.0035 |
| $R_{5_2}$ | 0.0749 | 0.0877 | 0.0043/0.0045 | 0.0749 | 0.0897 | 0.0057/0.0065 | 0.1277 | 0.1149 | 0.0048/0.0051 | 0.1277 | 0.1231 | 0.0026/0.0030 |
| $M_2^*$ | 2001.06 | 2001.061 | 0.0007/0.0009 | 2001.06 | 2001.366 | 0.3824/0.3957 | 2002.06 | 2002.060 | 0.0005/0.0007 | 2002.06 | 2002.060 | 0.0003/0.0004 |
| $\sigma$ | 10 | 8.0498 | 0.3195/0.3700 | 10 | 16.9136 | 8.9194/9.4351 | 10 | 7.8962 | 0.2247/0.3060 | 10 | 7.9236 | 0.2280/0.2663 |
| $\sigma_s$ | 0.08 | 0.0800 | 0.0002/0.0003 | 0.08 | 0.0959 | 0.0164/0.0173 | 0.08 | 0.0800 | 0.0003/0.0003 | 0.08 | 0.0799 | 0.0003/0.0002 |
| $S$ | 1.0015 | 1.0010 | 0.0004/0.0004 | 1.0015 | 1.0005 | 0.0028/0.0030 | 1.0015 | 1.0018 | 0.0004/0.0003 | 1.0015 | 1.0018 | 0.0002/0.0004 |
| $H_2/H_1$ | 0.5 | 0.4919 | 0.0076/0.0089 | 2 | 1.0602 | 0.8891/0.9006 | 0.5 | 0.4994 | 0.0040/0.0044 | 2 | 2.0028 | 0.0125/0.0139 |

**Table E.7:** Summary statistics of model averaging (mean estimates $\bar{\hat{\theta}}$, empirical standard errors $\sigma_{emp}$) and model based standard errors $\sigma_{mb}$.

| Parameter | set9 | | | set10 | | | set11 | | | set12 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ |
| $R_{2_1}$ | 1.2708 | 1.2664 | 0.0046/0.0070 | 0.9763 | 0.9760 | 0.0039/0.0051 | 1.2708 | 1.2576 | 0.0213/0.0280 | 0.9763 | 0.9822 | 0.0060/0.0068 |
| $R_{3_1}$ | 0.8872 | 0.8816 | 0.0049/0.0085 | 0.6385 | 0.6381 | 0.0033/0.0041 | 0.8872 | 0.8732 | 0.0191/0.0243 | 0.6385 | 0.6452 | 0.0036/0.0039 |
| $R_{4_1}$ | 0.4431 | 0.4384 | 0.0039/0.0068 | 0.3125 | 0.3115 | 0.0029/0.0039 | 0.4431 | 0.4287 | 0.0122/0.0206 | 0.3125 | 0.3180 | 0.0031/0.0036 |
| $R_{5_1}$ | 0.1750 | 0.1693 | 0.0032/0.0051 | 0.1277 | 0.1251 | 0.0027/0.0033 | 0.1750 | 0.1531 | 0.0089/0.0100 | 0.1277 | 0.1298 | 0.0028/0.0029 |
| $M_1^*$ | 2000.90 | 2000.901 | 0.0009/0.0015 | 2000.90 | 2000.900 | 0.0003/0.0003 | 2000.90 | 2000.904 | 0.0225/0.0255 | 2000.90 | 2000.899 | 0.0008/0.0008 |
| $R_{2_2}$ | 0.9763 | 0.9772 | 0.0053/0.0076 | 0.9763 | 0.9750 | 0.0039/0.0051 | 1.1577 | 1.1614 | 0.0110/0.0114 | 1.2708 | 1.1948 | 0.0322/0.0361 |
| $R_{3_2}$ | 0.6385 | 0.6393 | 0.0039/0.0063 | 0.6385 | 0.6382 | 0.0033/0.0042 | 0.6702 | 0.6720 | 0.0044/0.0044 | 0.8872 | 0.8105 | 0.0243/0.0281 |
| $R_{4_2}$ | 0.3125 | 0.3128 | 0.0032/0.0039 | 0.3125 | 0.3118 | 0.0029/0.0036 | 0.2586 | 0.2602 | 0.0033/0.0042 | 0.4431 | 0.3726 | 0.0172/0.0211 |
| $R_{5_2}$ | 0.1277 | 0.1253 | 0.0029/0.0040 | 0.1277 | 0.1264 | 0.0027/0.0033 | 0.0749 | 0.0797 | 0.0026/0.0037 | 0.1750 | 0.1129 | 0.0080/0.0094 |
| $M_2^*$ | 2001.14 | 2001.139 | 0.0007/0.0026 | 2002.14 | 2002.140 | 0.0002/0.0003 | 2001.14 | 2001.139 | 0.0009/0.0017 | 2001.06 | 2001.059 | 0.0025/0.0027 |
| $\sigma$ | 10 | 9.0796 | 0.7211/0.7564 | 10 | 8.1997 | 0.2364/0.3047 | 10 | 9.2103 | 0.8431/0.9449 | 10 | 8.1329 | 0.2391/0.3151 |
| $\sigma_s$ | 0.08 | 0.0809 | 0.0007/0.0009 | 0.08 | 0.0800 | 0.0001/0.0002 | 0.08 | 0.0815 | 0.0105/0.0135 | 0.08 | 0.0801 | 0.0003/0.0004 |
| $S$ | 1.0015 | 1.0013 | 0.0006/0.0011 | 1.0015 | 1.0014 | 0.0002/0.0005 | 1.0015 | 1.0012 | 0.0008/0.0014 | 1.0015 | 1.0025 | 0.0005/0.0006 |
| $H_2/H_1$ | 1 | 1.0565 | 0.0516/0.0591 | 1 | 1.0013 | 0.0039/0.0051 | 2 | 1.9458 | 0.0826/0.1039 | 0.2 | 0.2112 | 0.0080/0.0087 |

**Table E.8:** Summary statistics of model averaging (mean estimates $\bar{\hat{\theta}}$, empirical standard errors $\sigma_{emp}$) and model based standard errors $\sigma_{mb}$.

| Parameter | set13 | | | set14 | | | set15 | | | set16 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ |
| $R_{2_1}$ | 0.9763 | 0.9963 | 0.0213/0.0430 | 1.1577 | 1.1609 | 0.0046/0.0053 | 1.2708 | 1.2198 | 0.0288/0.0348 | 1.1577 | 1.1757 | 0.0196/0.0218 |
| $R_{3_1}$ | 0.6385 | 0.6630 | 0.0174/0.0257 | 0.6702 | 0.6722 | 0.0041/0.0064 | 0.8872 | 0.7701 | 0.0580/0.0587 | 0.6702 | 0.6951 | 0.0133/0.0156 |
| $R_{4_1}$ | 0.3125 | 0.3169 | 0.0129/0.0274 | 0.2586 | 0.2605 | 0.0032/0.0042 | 0.4431 | 0.3283 | 0.0340/0.0457 | 0.2586 | 0.2757 | 0.0094/0.0099 |
| $R_{5_1}$ | 0.1277 | 0.1100 | 0.0073/0.0094 | 0.0749 | 0.0782 | 0.0025/0.0027 | 0.1750 | 0.1092 | 0.0257/0.0194 | 0.0749 | 0.0926 | 0.0030/0.0050 |
| $M_1^*$ | 2000.90 | 2000.913 | 0.0275/0.0449 | 2000.90 | 2000.900 | 0.0004/0.0004 | 2000.90 | 2000.913 | 0.0002/0.0003 | 2000.90 | 2000.900 | 0.0002/0.0003 |
| $R_{2_2}$ | 1.2708 | 1.2652 | 0.0234/0.0261 | 1.1577 | 1.1713 | 0.0219/0.0271 | 1.1577 | 1.2538 | 0.1748/0.1903 | 0.9763 | 0.9626 | 0.0320/0.0362 |
| $R_{3_2}$ | 0.8872 | 0.8616 | 0.0403/0.0665 | 0.6702 | 0.6750 | 0.0163/0.0163 | 0.6702 | 0.6567 | 0.0418/0.0429 | 0.6385 | 0.6297 | 0.0180/0.0213 |
| $R_{4_2}$ | 0.4431 | 0.4250 | 0.0249/0.0288 | 0.2586 | 0.2644 | 0.0113/0.0124 | 0.2586 | 0.2799 | 0.0193/0.0199 | 0.3125 | 0.2900 | 0.0109/0.0112 |
| $R_{5_2}$ | 0.1750 | 0.1641 | 0.0143/0.0232 | 0.0749 | 0.0896 | 0.0060/0.0067 | 0.0749 | 0.0957 | 0.0073/0.0074 | 0.1277 | 0.1066 | 0.0057/0.0056 |
| $M_2^*$ | 2001.06 | 2001.061 | 0.0031/0.0034 | 2002.06 | 2002.061 | 0.0023/0.0024 | 2000.94 | 2003.529 | 1.4493/1.5102 | 2001.94 | 2001.941 | 0.0029/0.0029 |
| $\sigma$ | 10 | 9.2126 | 0.3654/0.5509 | 10 | 7.6929 | 0.2209/0.2912 | 10 | 7.7018 | 0.2180/0.2693 | 10 | 7.6577 | 0.2163/0.3582 |
| $\sigma_s$ | 0.08 | 0.0813 | 0.0051/0.0074 | 0.08 | 0.0801 | 0.0002/0.0004 | 0.08 | 0.0821 | 0.0012/0.0016 | 0.08 | 0.0800 | 0.0002/0.0003 |
| $S$ | 1.0015 | 1.0016 | 0.0007/0.0007 | 1.0015 | 1.0015 | 0.0006/0.0006 | 1.0015 | 1.0003 | 0.0005/0.0007 | 1.0015 | 1.0019 | 0.0004/0.0007 |
| $H_2/H_1$ | 5 | 4.5715 | 0.6324/0.9425 | 0.2 | 0.1967 | 0.0040/0.0051 | 0.5 | 0.0529 | 0.2993/0.3132 | 0.5 | 0.4815 | 0.0170/0.0195 |

**Table E.9:** Summary statistics of model averaging (mean estimates $\bar{\hat{\theta}}$, empirical standard errors $\sigma_{emp}$) and model based standard errors $\sigma_{mb}$.

| Parameter | set17 | | | set18 | | | set19 | | | set20 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ |
| $R_{2_1}$ | 1.1577 | 1.3033 | 0.0612/0.0787 | 1.2708 | 1.1234 | 0.0073/0.0080 | 1.2708 | 1.2546 | 0.0179/0.0197 | 0.9763 | 0.9358 | 0.0217/0.0230 |
| $R_{3_1}$ | 0.6702 | 0.7720 | 0.0427/0.0523 | 0.8872 | 0.6837 | 0.0252/0.0307 | 0.8872 | 0.8695 | 0.0195/0.0244 | 0.6385 | 0.5961 | 0.0175/0.0194 |
| $R_{4_1}$ | 0.2586 | 0.2933 | 0.0150/0.0158 | 0.4431 | 0.2868 | 0.0173/0.0240 | 0.4431 | 0.4297 | 0.0137/0.0154 | 0.3125 | 0.2806 | 0.0111/0.0132 |
| $R_{5_1}$ | 0.0749 | 0.0948 | 0.0067/0.0072 | 0.1750 | 0.0977 | 0.0089/0.0135 | 0.1750 | 0.1652 | 0.0079/0.0124 | 0.1277 | 0.1057 | 0.0065/0.0078 |
| $M_1^*$ | 2000.90 | 2000.900 | 0.0005/0.0007 | 2000.90 | 2000.920 | 0.0002/0.0003 | 2000.90 | 2000.904 | 0.0015/0.0019 | 2000.90 | 2000.900 | 0.0003/0.0003 |
| $R_{2_2}$ | 0.9763 | 0.9989 | 0.0179/0.0228 | 0.9763 | 1.1351 | 0.1034/0.1278 | 0.9763 | 0.9885 | 0.0253/0.0279 | 0.9763 | 0.9790 | 0.0181/0.0209 |
| $R_{3_2}$ | 0.6385 | 0.6694 | 0.0178/0.0231 | 0.6385 | 0.6749 | 0.0382/0.0415 | 0.6385 | 0.6468 | 0.0079/0.0098 | 0.6385 | 0.6450 | 0.0164/0.0199 |
| $R_{4_2}$ | 0.3125 | 0.3255 | 0.0115/0.0144 | 0.3125 | 0.2830 | 0.0177/0.0187 | 0.3125 | 0.3165 | 0.0089/0.0092 | 0.3125 | 0.3213 | 0.0077/0.0087 |
| $R_{5_2}$ | 0.1277 | 0.1289 | 0.0046/0.0044 | 0.1277 | 0.0961 | 0.0070/0.0074 | 0.1277 | 0.1268 | 0.0065/0.0073 | 0.1277 | 0.1324 | 0.0035/0.0042 |
| $M_2^*$ | 2001.94 | 2001.944 | 0.0017/0.0020 | 2000.94 | 2002.815 | 0.4575/0.7387 | 2001.06 | 2001.100 | 0.0774/0.1076 | 2001.94 | 2001.939 | 0.0011/0.0012 |
| $\sigma$ | 10 | 7.8860 | 0.2308/0.3076 | 10 | 7.8285 | 0.3204/0.3687 | 10 | 9.3840 | 0.8085/1.5082 | 10 | 7.8591 | 0.2265/0.2537 |
| $\sigma_s$ | 0.08 | 0.0792 | 0.0004/0.0005 | 0.08 | 0.0824 | 0.0010/0.0012 | 0.08 | 0.0816 | 0.0062/0.0085 | 0.08 | 0.0801 | 0.0002/0.0002 |
| $S$ | 1.0015 | 1.0004 | 0.0007/0.0009 | 1.0015 | 0.9989 | 0.0034/0.0038 | 1.0015 | 1.0004 | 0.0012/0.0020 | 1.0015 | 1.0013 | 0.0003/0.0005 |
| $H_2/H_1$ | 2 | 1.8655 | 0.0586/0.0737 | 1 | 0.3820 | 0.3782/0.4836 | 1 | 0.9445 | 0.0958/0.1373 | 1 | 1.0403 | 0.0209/0.0216 |

**Table E.10:** Summary statistics of model averaging (mean estimates $\bar{\hat{\theta}}$, empirical standard errors $\sigma_{emp}$) and model based standard errors $\sigma_{mb}$.

| Parameter | set21 | | | set22 | | | set23 | | | set24 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ |
| $R_{2_1}$ | 1.1577 | 1.1574 | 0.0051/0.0060 | 1.1577 | 1.1638 | 0.0200/0.0248 | 0.9763 | 0.9766 | 0.0041/0.0053 | 0.9763 | 0.9829 | 0.0260/0.0308 |
| $R_{3_1}$ | 0.6702 | 0.6699 | 0.0034/0.0045 | 0.6702 | 0.6742 | 0.0150/0.0183 | 0.6385 | 0.6392 | 0.0032/0.0035 | 0.6385 | 0.6402 | 0.0182/0.0218 |
| $R_{4_1}$ | 0.2586 | 0.2593 | 0.0034/0.0041 | 0.2586 | 0.2668 | 0.0107/0.0109 | 0.3125 | 0.3131 | 0.0027/0.0034 | 0.3125 | 0.3227 | 0.0110/0.0110 |
| $R_{5_1}$ | 0.0749 | 0.0906 | 0.0041/0.0046 | 0.0749 | 0.0964 | 0.0011/0.0074 | 0.1277 | 0.1308 | 0.0024/0.0027 | 0.1277 | 0.1448 | 0.0035/0.0062 |
| $M_1^*$ | 2000.90 | 2000.900 | 0.0004/0.0004 | 2000.90 | 2000.900 | 0.0009/0.0012 | 2000.90 | 2000.900 | 0.0003/0.0003 | 2000.90 | 2000.900 | 0.0012/0.0014 |
| $R_{2_2}$ | 1.1577 | 1.1993 | 0.0324/0.0344 | 1.1577 | 1.1612 | 0.0044/0.0055 | 1.2708 | 1.2943 | 0.0207/0.0245 | 1.2708 | 1.2726 | 0.0045/0.0057 |
| $R_{3_2}$ | 0.6702 | 0.7041 | 0.0181/0.0210 | 0.6702 | 0.6724 | 0.0033/0.0043 | 0.8872 | 0.8988 | 0.0154/0.0196 | 0.8872 | 0.8883 | 0.0040/0.0048 |
| $R_{4_2}$ | 0.2586 | 0.2782 | 0.0104/0.0122 | 0.2586 | 0.2602 | 0.0027/0.0032 | 0.4431 | 0.4544 | 0.0110/0.0112 | 0.4431 | 0.4438 | 0.0032/0.0040 |
| $R_{5_2}$ | 0.0749 | 0.0939 | 0.0037/0.0064 | 0.0749 | 0.0785 | 0.0024/0.0028 | 0.1750 | 0.1918 | 0.0032/0.0062 | 0.1750 | 0.1783 | 0.0024/0.0024 |
| $M_2^*$ | 2004.94 | 2004.942 | 0.0013/0.0016 | 2004.94 | 2004.940 | 0.0003/0.0003 | 2006.94 | 2006.940 | 0.0009/0.0013 | 2006.94 | 2006.940 | 0.0003/0.0004 |
| $\sigma$ | 10 | 7.8233 | 0.2760/0.3332 | 10 | 7.8989 | 0.2235/0.2968 | 10 | 7.9357 | 0.2251/0.2872 | 10 | 7.9129 | 0.3160/0.3615 |
| $\sigma_s$ | 0.08 | 0.0800 | 0.0002/0.0002 | 0.08 | 0.0800 | 0.0003/0.0003 | 0.08 | 0.0800 | 0.0002/0.0002 | 0.08 | 0.0801 | 0.0002/0.0002 |
| $S$ | 1.0015 | 1.0015 | 0.0005/0.0006 | 1.0015 | 1.0015 | 0.0005/0.0005 | 1.0015 | 1.0016 | 0.0005/0.0005 | 1.0015 | 1.0015 | 0.0005/0.0006 |
| $H_2/H_1$ | 0.2 | 0.1851 | 0.0041/0.0047 | 5 | 4.9978 | 0.0674/0.0801 | 0.2 | 0.1963 | 0.0027/0.0033 | 5 | 5.0197 | 0.0849/0.1005 |

**Table E.11:** Summary statistics of model averaging (mean estimates $\bar{\hat{\theta}}$, empirical standard errors $\sigma_{emp}$) and model based standard errors $\sigma_{mb}$.

| Parameter | set25 | | | set26 | | | set27 | | | set28 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\bar{\hat{\theta}}$ | $\sigma_{emp}/\sigma_{mb}$ |
| $R_{2_1}$ | 1.1577 | 1.1588 | 0.0041/0.0055 | 1.1577 | 1.1596 | 0.0083/0.0103 | 1.2708 | 1.2671 | 0.0044/0.0058 | 1.2708 | 1.2591 | 0.0088/0.0099 |
| $R_{3_1}$ | 0.6702 | 0.6706 | 0.0033/0.0043 | 0.6702 | 0.6709 | 0.0065/0.0083 | 0.8872 | 0.8840 | 0.0037/0.0045 | 0.8872 | 0.8758 | 0.0072/0.0090 |
| $R_{4_1}$ | 0.2586 | 0.2593 | 0.0028/0.0036 | 0.2586 | 0.2608 | 0.0054/0.0061 | 0.4431 | 0.4403 | 0.0030/0.0032 | 0.4431 | 0.4327 | 0.0059/0.0069 |
| $R_{5_1}$ | 0.0749 | 0.0801 | 0.0028/0.0032 | 0.0749 | 0.0870 | 0.0040/0.0047 | 0.1750 | 0.1704 | 0.0028/0.0030 | 0.1750 | 0.1570 | 0.0054/0.0060 |
| $M_1^*$ | 2000.90 | 2000.900 | 0.0004/0.0004 | 2000.90 | 2000.900 | 0.0004/0.0005 | 2000.90 | 2000.900 | 0.0002/0.0003 | 2000.90 | 2000.900 | 0.0004/0.0005 |
| $R_{2_2}$ | 0.9763 | 0.9738 | 0.0079/0.0095 | 0.9763 | 0.9769 | 0.0039/0.0049 | 1.1577 | 1.1534 | 0.0084/0.0090 | 1.1577 | 1.1566 | 0.0042/0.0049 |
| $R_{3_2}$ | 0.6385 | 0.6383 | 0.0065/0.0081 | 0.6385 | 0.6395 | 0.0033/0.0042 | 0.6702 | 0.6657 | 0.0065/0.0068 | 0.6702 | 0.6704 | 0.0033/0.0036 |
| $R_{4_2}$ | 0.3125 | 0.3109 | 0.0056/0.0063 | 0.3125 | 0.3121 | 0.0029/0.0031 | 0.2586 | 0.2581 | 0.0054/0.0061 | 0.2586 | 0.2587 | 0.0028/0.0037 |
| $R_{5_2}$ | 0.1277 | 0.1179 | 0.0048/0.0055 | 0.1277 | 0.1253 | 0.0027/0.0032 | 0.0749 | 0.0836 | 0.0036/0.0041 | 0.0749 | 0.0780 | 0.0025/0.0027 |
| $M_2^*$ | 2005.06 | 2005.061 | 0.0006/0.0007 | 2005.06 | 2005.060 | 0.0004/0.0004 | 2007.06 | 2007.059 | 0.0004/0.0005 | 2007.06 | 2007.060 | 0.0002/0.0003 |
| $\sigma$ | 10 | 8.0878 | 0.2298/0.3078 | 10 | 8.0974 | 0.2320/0.2952 | 10 | 8.2322 | 0.2416/0.3065 | 10 | 8,2474 | 0.2408/0.2866 |
| $\sigma_s$ | 0.08 | 0.0799 | 0.0002/0.0002 | 0.08 | 0.0798 | 0.0002/0.0002 | 0.08 | 0.0801 | 0.0002/0.0002 | 0.08 | 0.0801 | 0.0002/0.0002 |
| $S$ | 1.0015 | 1.0015 | 0.0005/0.0005 | 1.0015 | 1.0015 | 0.0002/0.0005 | 1.0015 | 1.0016 | 0.0002/0.0005 | 1.0015 | 1.0017 | 0.0005/0.0005 |
| $H_2/H_1$ | 0.5 | 0.4991 | 0.0032/0.0040 | 2 | 1.9986 | 0.0124/0.0156 | 0.5 | 0.5016 | 0.0031/0.0031 | 2 | 1.9848 | 0.0122/0.0139 |

**Table E.12:** Summary statistics of model averaging (mean estimates $\bar{\hat{\theta}}$, empirical standard errors $\sigma_{emp}$) and model based standard errors $\sigma_{mb}$.

| Parameter | set29 | | | set30 | | |
|---|---|---|---|---|---|---|
| | TRUE | $\hat{\theta}$ | $\sigma_{emp}/\sigma_{mb}$ | TRUE | $\hat{\theta}$ | $\sigma_{emp}/\sigma_{mb}$ |
| $R_{2_1}$ | 1.2708 | 1.2669 | 0.0047/0.0056 | 1.2708 | 1.2674 | 0.0046/0.0054 |
| $R_{3_1}$ | 0.8872 | 0.8838 | 0.0038/0.0051 | 0.8872 | 0.8840 | 0.0038/0.0044 |
| $R_{4_1}$ | 0.4431 | 0.4400 | 0.0032/0.0033 | 0.4431 | 0.4402 | 0.0031/0.0038 |
| $R_{5_1}$ | 0.1750 | 0.1680 | 0.0030/0.0037 | 0.1750 | 0.1696 | 0.0029/0.0034 |
| $M_1^*$ | 2000.90 | 2000.900 | 0.0002/0.0002 | 2000.90 | 2000.900 | 0.0002/0.0003 |
| $R_{2_2}$ | 1.2708 | 1.2672 | 0.0047/0.0059 | 0.9763 | 0.9753 | 0.0040/0.0043 |
| $R_{3_2}$ | 0.8872 | 0.8839 | 0.0039/0.0043 | 0.6385 | 0.6372 | 0.0034/0.0045 |
| $R_{4_2}$ | 0.4431 | 0.4397 | 0.0032/0.0034 | 0.3125 | 0.3116 | 0.0030/0.0034 |
| $R_{5_2}$ | 0.1750 | 0.1691 | 0.0029/0.0031 | 0.1277 | 0.1244 | 0.0028/0.0037 |
| $M_2^*$ | 2005.14 | 2005.140 | 0.0002/0.0002 | 2007.14 | 2007.140 | 0.0002/0.0003 |
| $\sigma$ | 10 | 8.6752 | 0.2633/0.3499 | 10 | 8.5103 | 0.2521/0.3422 |
| $\sigma_s$ | 0.08 | 0.0802 | 0.0001/0.0002 | 0.08 | 0.0801 | 0.0001/0.0002 |
| $S$ | 1.0015 | 1.0015 | 0.0005/0.0005 | 1.0015 | 1.0016 | 0.0005/0.0005 |
| $H_2/H_1$ | 1 | 1.0013 | 0.0041/0.0054 | 1 | 0.9988 | 0.0040/0.0045 |