Statistical Models for Missing Data in Proteomic Studies of Gastric Cancer

Daniel Dean Kon

June 23, 2019

Thesis submitted for the degree of
Master of Philosophy
in
Statistics
at The University of Adelaide
Faculty of Engineering, Computer and Mathematical Sciences
School of Mathematical Sciences

THE UNIVERSITY
of ADELAIDE

Contents

List of Tables

List of Figures

Abstract

Disease diagnosis is often performed using a blood test for protein biomarkers which exhibit differential expression in diseased subjects as compared to healthy subjects. Discovery of new biomarkers enables cheaper and less invasive diagnosis. A method of biomarker discovery is the statistical analysis of proteomic mass spectrometry data to determine differences in protein concentration between groups of organisms. However, outcome-dependent missingness in proteomic mass spectrometry data hinders the extraction of useful information from the data and results in biased inference about these differences in protein expression. Existing methods of accounting for missing data, used for other, similar datasets such as those from RNA microarray experiments, assume missingness that is less severe and outcome-dependent than that which affects proteomic mass spectrometry data. These methods do not suffice to undo the bias, and new methods of statistical analysis are sought for biomarker discovery.

We develop a joint statistical model for missing and observed data and apply it to a dataset from a gastric cancer experiment that has a large degree of outcome-dependent missingness in order to discover novel candidate biomarkers. A set of candidates is produced using the joint model. This set differs from the set of biomarker candidates produced in earlier work modelling the data without accounting for the outcome-dependent missingness.

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed: . ..... Date: 28/03/2019

Acknowledgements

I thank my supervisors Associate Professor Gary Glonek and Doctor Ty Stanford, as well as my former supervisor Professor Patty Solomon, who have provided me with direction in my research and much guidance in the preparation of this thesis. Thank you for giving your time towards feedback on my writing. My mathematical writing and my communication skills have grown by leaps and bounds while working with all of you.

I also thank my family who have supported me over the course of my Masters program, especially my parents Mirosława and Zbigniew, whose roof I have been living under during my entire University career. I am grateful to them for keeping me on track with working on the manuscript.

I am very grateful to my friends Russell and Peter who have always been there for me when I needed a break from the University. I thank Russell for reading and providing feedback on early drafts of the thesis, and Peter for paying for many rounds at the pub, and both for hours of interesting conversations about math and the world.

Chapter 1

Introduction

1.1 Background to the research project

In human populations, deaths due to gastric cancer are overrepresented relative to other types of cancer (Lin et al., 2012; Penno et al., 2012). The underlying reason for this is that existing cancer detection methods consist mainly of noninvasive imaging methods that do not detect asymptomatic, early-stage tumours (Lin et al., 2012; Terp and Ditzel, 2014). The expense of other detection methods such as endoscopic examinations motivates the search for an easier diagnostic procedure (Humphries et al., 2014). An example of one such procedure is a blood test that detects *biomarkers* for the disease, which are proteins that are differentially expressed in diseased versus healthy subjects.

Novel protein biomarkers are sought for use in routine blood tests for diagnosis of diseases such as breast cancer (Hajduk et al., 2016; Jung, 2016; Ky et al., 2014). Existing blood serum biomarkers for cancer are neither sufficiently sensitive nor specific for practical use (Coghlin and Murray, 2016; Terp and Ditzel, 2014). Proteomic analysis, such as *mass spectrometry* (MS), is a rich source of novel biomarkers (Coghlin and Murray, 2016; Liu et al., 2012). However, proteomic studies face issues such as very wide ranges of protein concentration in biological samples (Callesen et al., 2008; Gianazza et al., 2016), difficulties in reproducibility (Beavis and Chait, 1996; Callesen et al., 2008), and a high degree of missing data, the last problem especially prevalent in proteomic MS studies (Jung et al., 2014; Karpievitch et al., 2009, 2012; Pedreschi et al., 2008).

Missingness is a problem in proteomic MS because the missingness is informative of the outcome that is missing (Davison, 2003; Graham, 2012; Little and Rubin, 2002; Rubin, 1976). As a consequence, statistical inference that does not account for informative missingness is biased, hampering downstream analyses and hindering the discovery of biomarkers (Aittokallio, 2010; Lazar et al., 2016; Li et al., 2011; Webb-Robertson et al., 2015). There is a pressing need to deal with missingness directly and appropriately, and extract as much information out of existing datasets as possible.

The statistical analysis of proteomic datasets bears a superficial similarity to that of RNA microarray data (Jung, 2016; Jung et al., 2006; Li et al., 2011; Webb-Robertson et al., 2010), which is also subject to missingness. Methods of dealing with the missingness exist for such data. The most basic method is to ignore the missingness and perform a complete case analysis. More advanced statistical methods such as the expectation-maximisation algorithm or maximum likelihood estimation can accommodate missing data (Aittokallio, 2010; Kenward and Molenberghs, 1999; Pigott, 2001). Missing values can also be imputed using a variety of methods varying in sophistication. The least sophisticated methods are, for example, assuming that missing values are equal to zero or to the average of the observed data (Pedreschi et al., 2008). Such methods do not account for the biases in statistical inference that arise when missingness is informative of the outcome of interest (Karpievitch et al., 2009, 2012; Webb-Robertson et al., 2010; Wood et al., 2004). More sophisticated methods are to estimate the missing values using a k-Nearest Neighbour or a model-based approach (Aittokallio, 2010; Jung et al., 2005; Webb-Robertson et al., 2015). Finally, multiple imputation is one of the most sophisticated methods (Little and Rubin, 2002; Newman, 2014; Nielsen, 2003).

Proteomic MS data missingness requires *application-specific* methods rather than generic methods because the mechanism that causes values to be missing depends on the outcome values, and this dependence is in different ways for different experimental and technological contexts (Aittokallio, 2010; Lazar et al., 2016). This means that merely replacing missing values with zeroes or with data averages is insufficient to perform unbiased statistical inference. There is much room for improvement in methods of analysis of proteomic MS data that deal with the missingness (Lazar et al., 2016). Explicitly accounting for informative missingness may detect differences in protein expression between diseased versus healthy subjects that are missed in simpler analyses (Jung et al., 2014; Lazar et al., 2016; Li et al., 2011; Webb-Robertson et al., 2010; Wood et al., 2004).

A promising approach is hinted at from the related context of longitudinal and failure time studies, in which informative missingness (commonly due to subject drop-out) is a well-established issue (Wu and Carroll, 1988). The problem of missing data in these studies is attacked by jointly modelling the subjects' responses with linear models and the subjects' drop-out using a logistic model (Diggle and Kenward, 1994; Follmann and Wu, 1995; Gould et al., 2015; Hogan and Laird, 1997; Kenward and Molenberghs, 1999; Little, 1995).

This thesis concerns the investigation of a proteomic MS dataset, called the *gastric cancer* (GC) dataset, which is subject to informative missingness. The goal of the statistical modelling of the GC dataset is to discover peaks pointing to candidate protein biomarkers that are more sensitive and specific in disease diagnosis than presently-known biomarkers. Parametric models are used for this, and parameters that encode differences in peak intensity between disease genotype groups are of greatest interest. We develop and fit joint missing/observed data models for the GC dataset, which explicitly account

for the informative nature of the missingness, ameliorating the biases in statistical inference and extracting information from the data that is overlooked by less sophisticated methods of data analysis.

1.2 Summary of thesis

In Chapter 2, we visualise the numerical intensity responses as well as the pattern of missingness in the GC dataset. Such visualisations help subsequent modelling of the data, and reveal nuances in the observed data that may not be captured in the statistical models but nonetheless point towards protein biomarker candidates. We detail the experimental setup that produced the GC dataset in Section 2.1. In Section 2.2, we investigate in detail the patterns of missingness and the observed data at broad and detailed scales and subsequently the observed intensities at both scales, taking account of how the missingness affects the GC dataset and how the missingness should be modelled. The nature of data collection using the *MALDI-TOF* MS apparatus and processing of the raw data by Stanford (2015) informs the interpretation of the data in terms of how observations come to be missing and the correspondence of measured intensites to the underlying proteins.

In Chapter 3, we investigate parametric statistical models that are fitted to the observed missingness. The model chosen for the missingness yields a set of protein peak m/z values of interest. This set differs from the set obtained by Stanford (2015) in prior work on the GC dataset that focused on the observed data, ignoring the missingness. This difference suggests that taking account of the informative nature of the missingness will extract information from the GC dataset that is not available when analysing the observed data alone. In Section 3.1, we briefly summarise the purpose of modelling the GC dataset, and in Section 3.2, we introduce the relevant mathematical framework and the notation. In Section 3.3, we describe the candidate models for the missingness and explain how the choice of the best model was made. The issue of separation of data, which hindered modelling efforts, was resolved using a Bayesian framework. In Section 3.4 we give details of the chosen model and evaluate the model using a variety of model checking methods, and in Section 3.5 we present results from the fitted models alongside the previous results from Stanford (2015).

In Chapter 4, we investigate parametric statistical models for the joint distribution of the intensity and the missingness that explicitly incorporate the informative nature of the missingness. The joint models may be understood as extensions of the previous, separate models for the missing data and the observed intensities. The results of the joint models provide a refinement of the work done in Chapter 3. In Section 4.1, we give details of the model used by Stanford (2015). In Section 4.2, we introduce the mathematical framework and formulate the joint models, explaining the rationale behind the structure of the joint models. In Section 4.3, we detail the difficulties of estimating parameters in the joint

models as well as the method by which these difficulties were addressed in this thesis. In Section 4.4, we perform model checking, and in Section 4.5 the results from the fitted joint models are presented. The research demonstrates that using the information obtained from the joint model together with the information from the separate observed data and missing data models improves biomarker discovery. In addition, additional investigations of the GC dataset hint at additional biomarker candidates of secondary interest.

In Chapter 5, we summarise the findings of the previous chapters, make concluding comments, and suggest further applications of the joint modelling methodology developed in this thesis.

Chapter 2

The gastric cancer dataset

The experiment from which the GC dataset was derived was a proteomic MS study conducted on genetically modified mice. In this chapter, the genetic and phenotypical traits of the mice are described along with the experimental design used to produce the GC dataset. The proteomic expression observations and the missingness patterns in the GC dataset are visualised in detail. Such visualisations inform the modelling of the data.

2.1 The gastric cancer experiment

The GC dataset analysed in this thesis originated from a proteomic mass spectrometry experiment conducted at the Adelaide Proteomics Centre. The experiment involved transgenic mice belonging to one of five different genotype groups, numbered from 1 to 5, and respectively denoted FF, FFIL6, FFStat3, IL6, and WT. The group WT represents the *wildtype*, the group of genetically healthy mice in a wild population. The remaining groups are mutations of the wildtype. The mutations distinguishing the groups are in the gene for *glycoprotein 130* (gp130), the gene for *signal transducer and activator of transcription 3* (Stat3), and the gene for *interleukin-6* (IL6)[1]. These mutations cause phenotypic responses in the mice.

The mutation in the gp130 gene is a single nucleotide polymorphism which substitutes the amino acid phenylalanine (abbreviated F) for the acid tyrosine (Y) at position 757 of the protein coded by the gene (Penno et al., 2012). The group denoted FF is distinguished by a *homozygous* mutation of this gene, meaning that both of the alleles of the gene are mutant. The mutation in the Stat3 gene causes the inactivation of the Stat3 protein (Jenkins et al., 2005). The group denoted FFStat3 is distinguished by the homozygous mutation in the gp130 gene (as described above) as well as a *heterozygous* mutation of the Stat3 gene, in which one of the alleles of Stat3 is the inactive 'null' form and the other allele is the wildtype form. The mutation in the IL6 gene causes the inactivation of the IL6 cytokine protein. The group denoted IL6 is distinguished by a homozygous mutation of this gene. The group FFIL6 is distinguished by the homozygous mutation in the gp130 gene (as described above) as well as the homozygous mutation of the IL6 gene. Table 2.1 summarises the mutations present in the groups.

Up to two physical characteristics of interest may appear in mice belonging to the five groups. These characteristics are the presence of gastric lesions, and inflammation of the gut. Tebbutt et al. (2002), Jenkins et al. (2005), Judd et al. (2009), and Penno et al. (2012) detail the molecular mechanisms associated with the mutations that lead to the phenotypic changes. There are four phenotypes resulting from these physical characteristics, corresponding to the presence of none, one, or both of the characteristics. Mice belonging to the FF group develop both the gastric cancer and inflamed gut phenotypes. The protein gp130 is a cell receptor for many signal transduction pathways and is activated by the interleukin-6 family of cytokine proteins. This family of proteins is involved in ulcer formation and healing in the gut (Judd et al., 2009). The homozygous mutation in the gp130 gene indirectly upregulates the active form of the Stat3 protein and simultaneously causes the deactivation of the *SHP2-Ras-ERK* pathway, which together lead to the development of gastric cancer and gut inflammation (Tebbutt et al., 2002). Mice belonging to the FFStat3 group develop an inflamed gut phenotype but will not suffer from gastric cancer. The protein Stat3 is a transcription factor, high levels of which

[1]To avoid confusion, the term "IL6" by itself is understood to refer only to the group.

Table 2.1: Summary of mutations.

Gene	Mutation	Present in groups
gp130	Homozygous, gene deactivated by single nucleotide polymorphism (F $\rightarrow$ Y substitution)	FF, FFStat3, FFIL6
Stat3	Heterozygous, gene deactivated	FFStat3
Interleukin-6	Homozygous, gene deactivated	IL6, FFIL6

Table 2.2: Summary of transgenic mouse phenotypes according to groups.

No.	Group	Cancer	Gut inflammation
1	FF	Y	Y
2	FFIL6	Y	
3	FFStat3		Y
4	IL6		
5	WT		

result in gastric cancer in mice carrying the mutant gp130 gene. However, a heterozygous mutation in the Stat3 gene results in reduced levels of active Stat3 protein, arresting the development of cancer. The protein IL6 is a cytokine which plays a role in the inflammatory response of mice to disease or irritation. Mice belonging to the FFIL6 group will develop gastric cancer due to the defective form of the gp130 protein but will not suffer from gut inflammation. Mice belonging to the IL6 and WT groups will develop neither gastric cancer nor gut inflammation. Table 2.2 summarises the phenotypic variations across each group.

2.1.1 Experimental design

The GC experiment was conducted on 40 mice reared to 12 weeks of age. There were eight mice within each of the five groups. The GC experiment may be conceptually split into two stages. The first stage involved the processing of blood serum samples from the mice. The second stage analysed the samples using a MALDI-TOF mass spectrometer to yield protein concentration measurements over a spectrum of *mass-to-charge* ratio (m/z ratio) for each sample. Glish and Vachet (2003); Merchant and Weinberger (2000), and Hajduk et al. (2016) present detailed reviews of the MALDI-TOF MS technology.

In the first stage of the GC experiment, blood serum was extracted from each of the 40 mice prior to euthanisation. Each serum extract was then processed according to the following sequence. First, the extract was split into three aliquots. Each aliquot of blood was then fractionated with three batches of magnetic C8 beads, yielding nine C8 bead batches in total. Each C8 batch was then split into three replicate samples, yielding 27

samples from a single blood serum extract. A total of 1080 samples were taken, with 216 samples from each of the five groups.

The samples were placed on the wells of three MALDI chips which were mounted in the mass spectrometer. Nine samples from each of the 40 mice were placed on each chip, meaning that each chip contained 360 samples. The reason that three chips were used instead of one was a space constraint imposed by the number of wells on a chip; each chip contains 384 wells. The allocation of samples to chips is confounded with the aliquot split, in the sense that the nine samples originating from an aliquot were allocated to the same chip. The 27 samples from each mouse were therefore evenly split across the three chips, and each chip contains 72 samples from each of the five groups.

The sequence of processing stages in the experimental design induced a *hierarchical structure* (Gelman and Hill, 2009; Snijders and Bosker, 2012) in the GC dataset, consisting of sample replicates at the lowest level, within C8 batches, within aliquots, within mice, and within group at the highest level. Accounting for variation at all levels of the hierarchical structure is necessary when modelling the data.

Figure 2.1 depicts the sample preparation stage for a single mouse. Figure 2.2 provides the pattern of allocation of the 360 samples on a single chip according to the group of the originating mouse, and Figure 2.3 details the pattern of allocation of the 72 samples on a single chip and from a single group according to the mouse number within the group.

In the second stage of the experiment, the MALDI chips were loaded into the MALDI-TOF mass spectrometer. Figure 2.4 depicts the operation of the mass spectrometer. The concentrations of proteins in the samples were measured by the machine to produce raw spectra. Figure 2.5 displays an example of raw MS output for a single sample. In the raw spectra, low levels of signal tend to be hidden by noise which arises from matrix ejecta such as fragmented proteins and extraneous pieces of the acid matrix. The raw spectra were processed by Stanford (2015) to remove noise, producing the GC dataset analysed in this research thesis. This processing involved assigning raw peak expressions from each sample to specific m/z locations and removing observations that fell below a threshold value.

2.2 Description and visualisation of the data

The GC dataset consists of observations of peak intensity at 159 locations in the m/z spectrum across the 1080 serum samples, where the intensities are $\log_2$-transformed measurements of the peak concentration of the protein in the blood sample. Metadata information for each of the samples is also included in the GC dataset. There are 171720 observations in total in the GC dataset. The m/z values range from a minimum of 2008 m/z to a maximum of 17976 m/z. Appendix A contains the complete list of peak m/z values in the GC dataset.

The peak intensities range from 6.97 to 15.22 with a median value of 8.19. Many

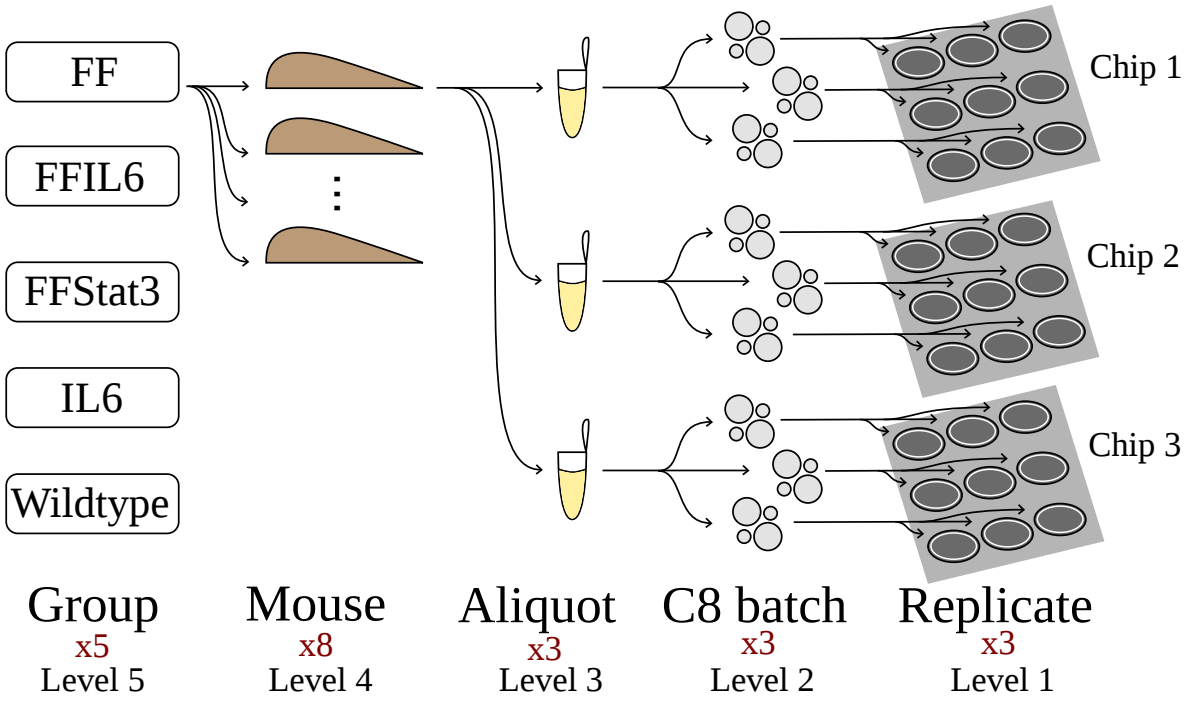

Figure 2.1: Production of 27 replicate samples from a single mouse. Procedure was replicated for all mice in all groups. Note the confounding between aliquot and MALDI chip.

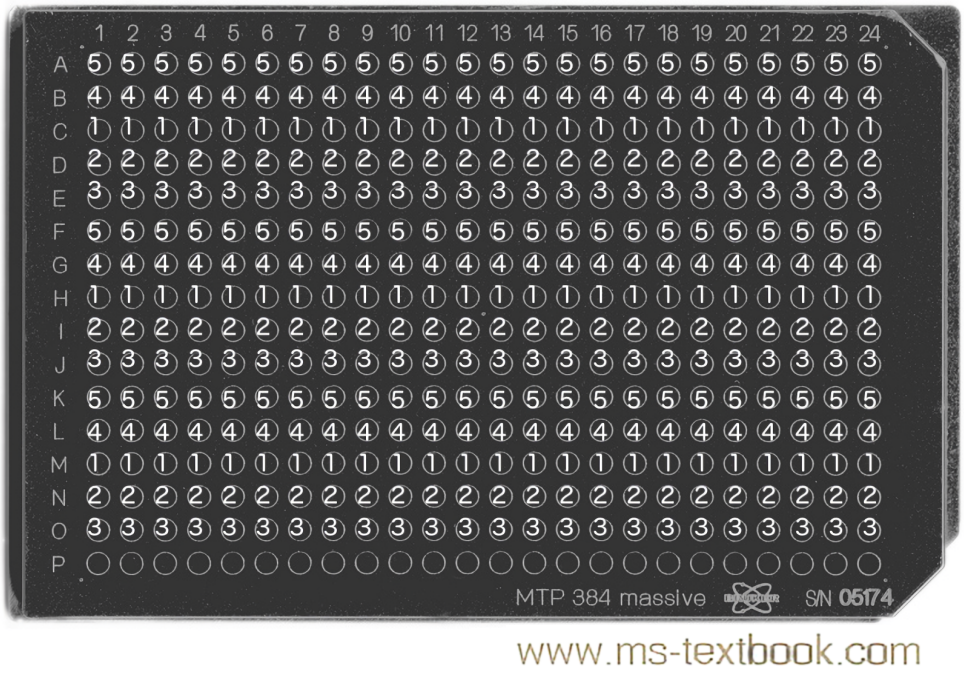

Figure 2.2: Group membership displayed for all samples on a single MALDI chip. This arrangement was replicated for all three chips.

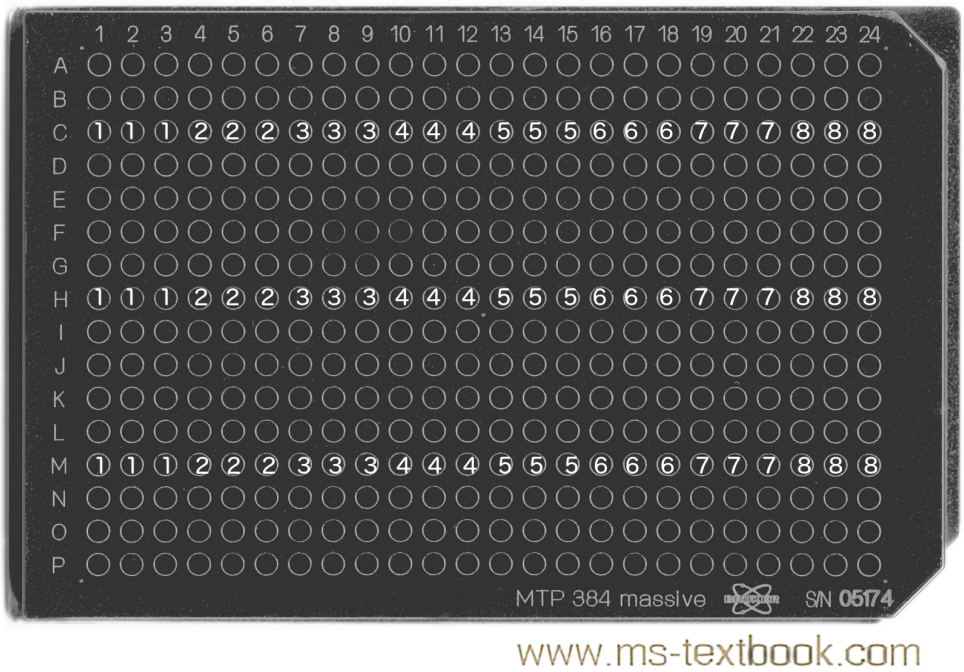

Figure 2.3: Mouse number displayed for all samples from group 1 (FF) on a single MALDI chip. This arrangment was replicated with vertical offset as necessary for groups 2–5 as displayed in Figure 2.2.

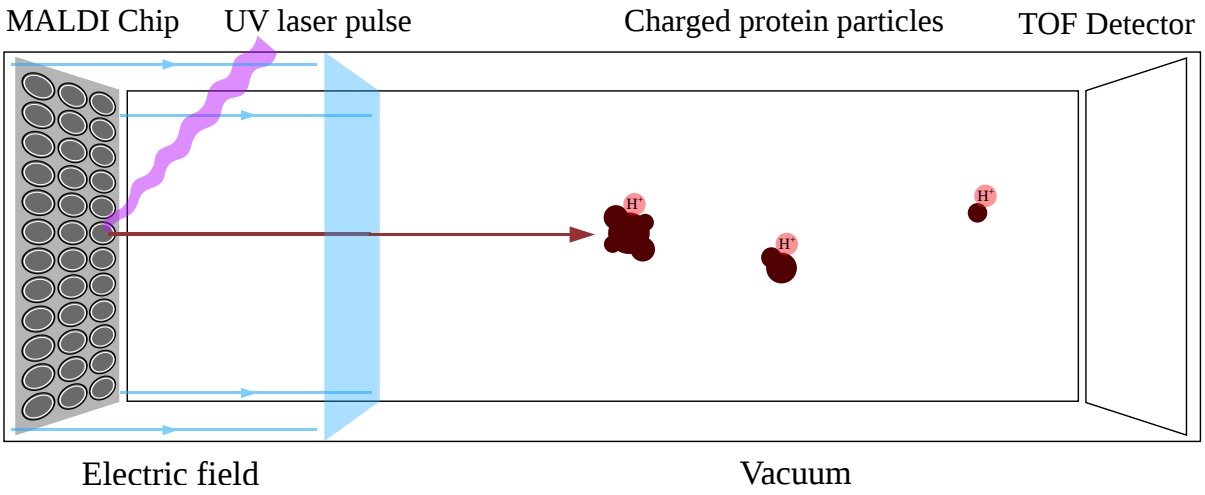

Figure 2.4: Schematic of MALDI-TOF mass spectrometer in operation. A biological extract such as a blood serum sample is mixed with an acidic matrix solution and left to solidify in a spot on a MALDI chip. When a pulse of ultraviolet laser light is shone on the spot, ablation of the solid matrix occurs, ejecting a stream of positively-charged protein particles (Rosa, 2013). The stream of charged particles is accelerated through an electric field in vacuum, with increasing particle m/z ratios corresponding to lesser accelerations and consequently longer particle flight times as measured by the TOF detector, distinguishing between proteins of different masses (Yates III, 2011).

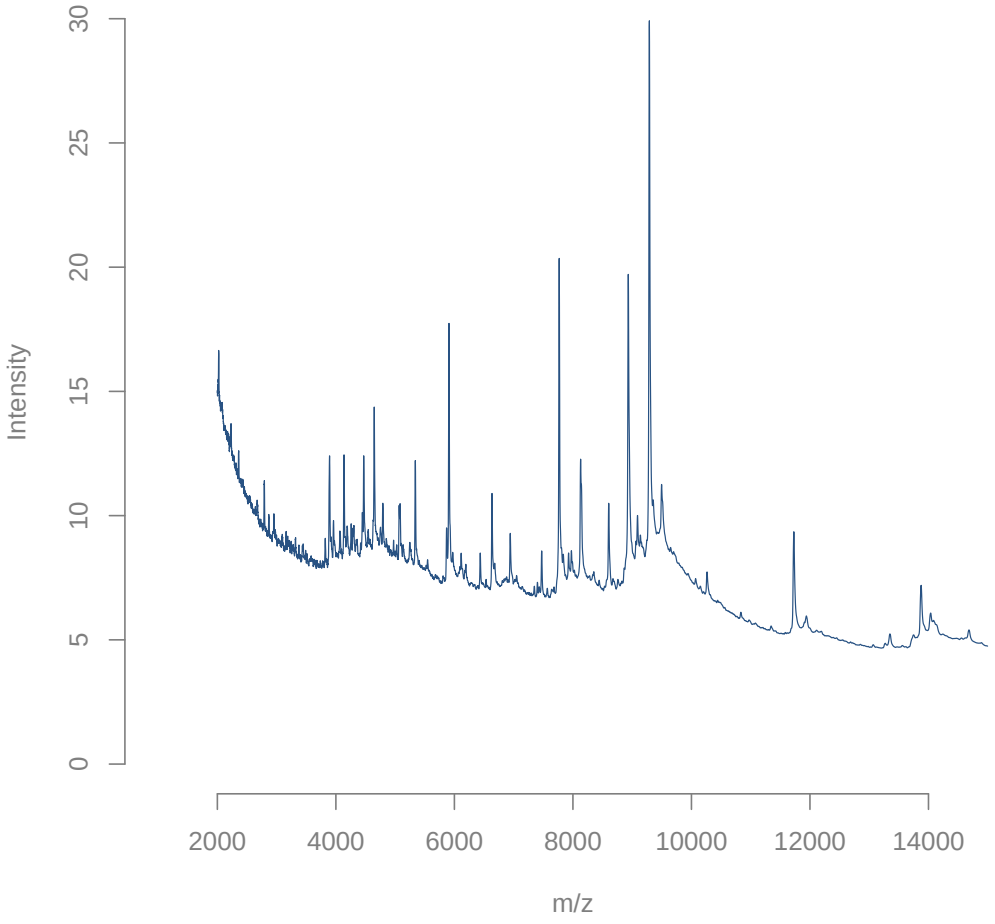

Figure 2.5: A typical example of raw MS data from a single sample. Measurements of intensity over the entire m/z spectrum contain both peak intensities at particular m/z values corresponding to various proteins present in the sample as well as background noise at other m/z values. The GC dataset consists only of intensity measurements from the m/z values that correspond to peak intensities. Source: Adam et al. (2002)

Table 2.3: Subset of `dd` data frame consisting of observations of first ten peaks from first ten samples. Blank entries correspond to missing data.

Sample	Peak m/z									
	2008	2033	2057	2081	2104	2128	2247	2262	2478	2504
1	9.52	9.16	8.80			7.78				7.98
2	9.49	9.21	8.43							
3	8.81	9.03	9.04	7.80	7.65					
4	9.41	9.03	7.90	7.68						
5	9.72	9.23	8.54							8.33
6	9.19	8.73	8.40	8.23						
7	10.23	9.62	8.44	8.17					7.66	8.80
8	10.73	10.46	8.95	8.19						8.75
9	8.72	8.22	8.05	7.86	7.78					
10	9.32	8.62	8.40	8.32			7.96			

observations of peak intensities are missing across the range of the 1080 samples. A missing observation in a sample means that the intensity for a given peak exists but is unknown for that sample. There are 90493 out of 171720 missing observations in the GC dataset, corresponding to a proportion of 0.527. No peak has a proportion of missing values greater than 0.9. Peak 4358 m/z is unique in having zero missing values in its 1080 observations.

2.2.1 Description of GC dataset as `R` objects

The GC dataset exists in the `.RData` format native to the programming language R (R Core Team, 2016). The GC dataset consists of two `data.frame` objects named `dd` and `metaDF`.

The data frame `dd` is a matrix of observations of dimension 1080×159. Each observation is either a numerical value if it is known, or an `NA` value if it is missing. The rows of the matrix correspond to the 1080 samples and the columns of the matrix correspond to the 159 peaks. Table 2.3 displays a subset of `dd` to give an indication of the detailed structure of the observation matrix. The data frame `metaDF` is a 1080×9 matrix of metadata on the 1080 samples from the GC dataset. The first five columns of the matrix contain information about the samples' group numbers, chip numbers, mouse numbers, aliquot numbers, and C8 batch numbers. The next four columns of the matrix contain information about sample placement on the MALDI chip and information from which the sample scanning order of the MALDI-TOF MS laser can be deduced. The information relevant to the statistical modelling of the GC dataset undertaken in this thesis is contained in `dd` and in the first five columns of `metaDF`.

Samples are ordered lexicographically by MALDI chip number, then by genotype group

number, then by mouse number. Within every subset of nine samples that share the same mouse and chip, the ordering of the C8 batch number is 1, 2, 3, 1, 2, 3, 1, 2, 3.

2.2.2 Visualising the missingness pattern

The missingness pattern in the GC dataset is explored and visualised in order to provide an understanding of the missingness, and to inform modelling of the missingness pattern in Chapter 3.

Broad overview

Figure 2.6 depicts the missingness in the GC dataset by displaying a heatmap of the observed data in `dd`, the 1080×159 matrix of observations. Immediately apparent is the vertical striping. This striping arises from the greater variation in missingness proportion across peaks relative to the variation in proportion across samples. Figure 2.7 summarises the distributions of missingness proportions within peaks and samples. Missingness of observations within peaks varies from zero percent up to 90%, whereas missingness of observations within samples is close to 50% in all but a few cases.

Details of peaks and samples

The pattern of missingness in the GC dataset is further explored by considering how missingness varies with the location of samples on the MALDI chips. Figures 2.8a and 2.8b show the missingness counts for each sample on the three chips. Correlations are apparent in missingness counts for samples within the same group and for samples from the same mouse. These correlations show as streaks and spots of similar luminosity in the plots. The patterns of luminosity correspond with the sample layout displayed in Figures 2.2 and 2.3.

The three brightest spots in column 16 of chip 2 from Figure 2.8a are an example of associations in the numbers of missing observations between samples from the same mouse. The bright horizontal bands in rows A, F, and K in chip 3 from Figure 2.8b are an example of correlation in samples from the same group. In addition to the group and mouse variation, chip 1 appears slightly brighter than the other chips, suggesting inter-chip variation in missingness.

Figure 2.9 displays missingness counts within individual peaks for a representative subset of 40 peaks in the GC dataset, spanning the entire m/z spectrum. For each peak, the number of missing observations in each combination of genotype group and MALDI chip number is displayed. There are 15 such combinations with 72 samples per combination. These combinations reveal that variation in missingness between groups and chips are present in the majority of peaks. In most peaks, the primary differences in

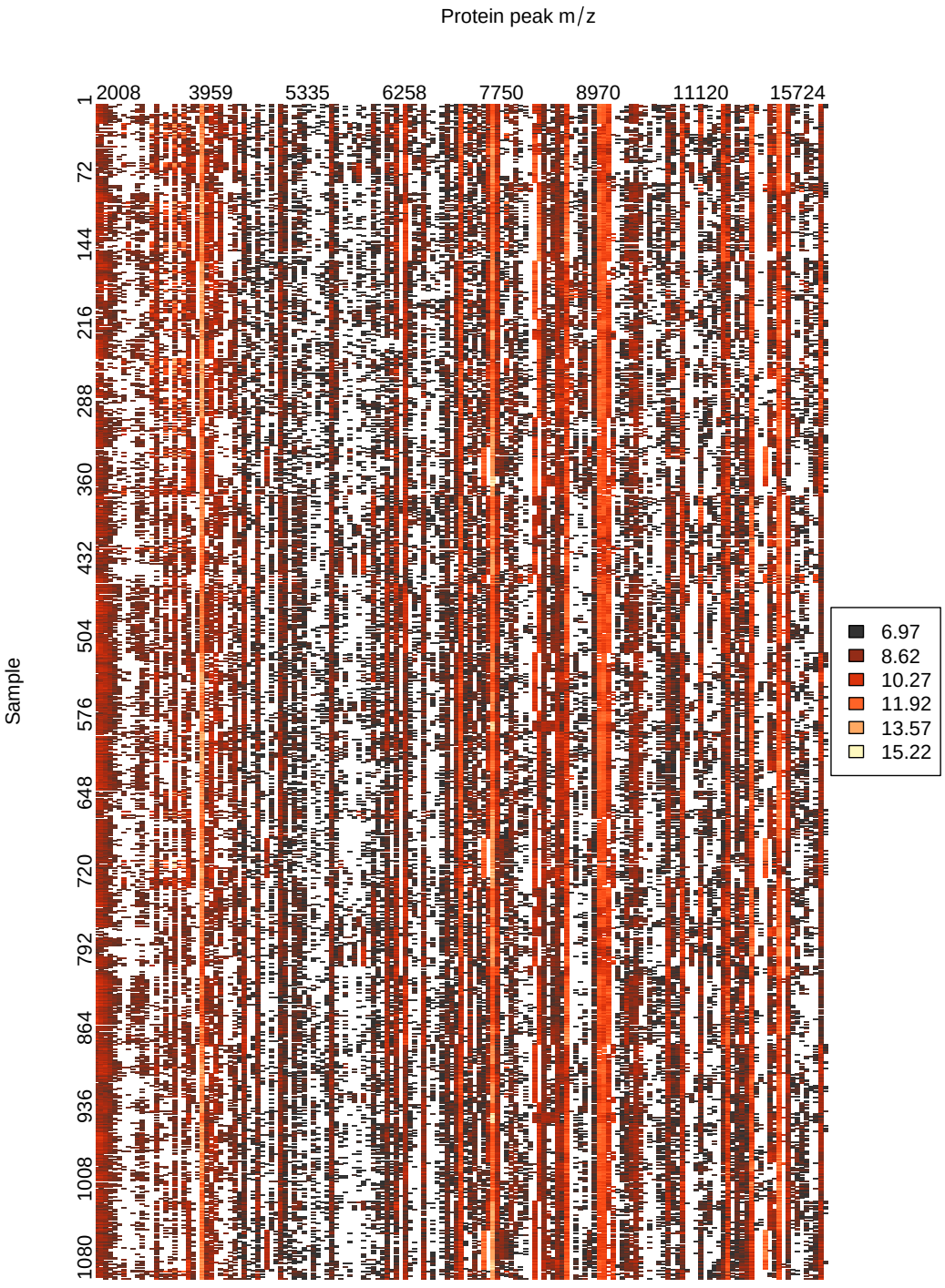

Figure 2.6: Graphical representation of the matrix of observations in the GC dataset. Each entry in the matrix is represented by a cell which is either white or coloured, corresponding to a missing or an observed datum respectively. The brightness of coloured cells indicates the intensity, with dark cells corresponding to low intensities and bright cells corresponding to high intensities. The row position of a cell indicates the observation's sample number, and the column position indicates the observation's peak m/z. The x axis represents peaks in increasing m/z order, and the y axis represents the sample number.

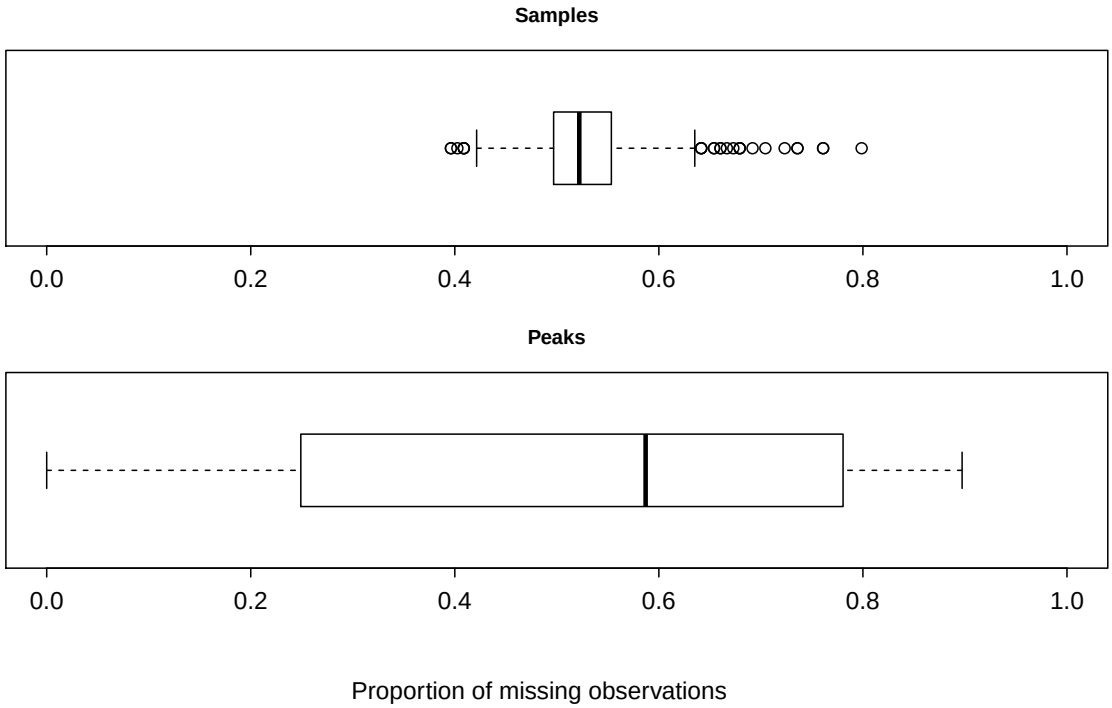

Figure 2.7: Proportion of missing observation values over the set of 1080 samples (top) and the set of 159 peaks (bottom). Variation in missingness proportion is much greater across peaks than across samples.

missingness counts lie between groups. However, some peaks, such as 2793 m/z, exhibit more inter-chip than inter-group variation.

Stark contrasts in the total missingness count between peaks are also apparent in Figure 2.9, even in consecutive peaks, such as the peaks at 5059 m/z and 5189 m/z. Additionally, in a few peaks, group/chip combinations with zero missingness (such as in the peaks at 6602 m/z or 15724 m/z) are found alongside combinations with up to 50% missing observations.

Figure 2.10 displays Cohen's kappa measure of concordance in missingness between peaks (Cohen, 1960). The majority of peak pairs have mild levels of concordance that are expected to arise by chance only (excluding self-pairs, which result in a concordance of one). An exception is the cluster of low m/z peaks having mutual concordance values above the norm, which is visible as a slightly red region in the top left of the plot. The increased level of concordance for these peaks may be an artifact of the increased noise in MALDI-TOF MS systems for low m/z due to matrix ejecta. Another deviation from mild levels of concordance is found in four pairs of peaks each exhibiting relatively extreme negative concordance. These peak pairs are visible as pairs of diagonally adjacent blue cells straddling the diagonal red line. The pairs all consist of adjacent peaks in the m/z spectrum. They are found at 4607 and 4617 m/z, 7738 and 7750 m/z, 8302 and 8337 m/z, and 9305 and 9319 m/z. The missingness patterns for these peak pairs are complementary with respect to group in the sense that if one peak of the pair has a large proportion of missing values for a particular group, then the other peak of the pair tends to have a small proportion. This complementary behaviour may correspond to a single protein being assigned to either one m/z value or the other in the preprocessing stage, an assignment that follows no clear pattern between different samples.

2.2.3 Visualising the intensities

The peak intensity measurements are displayed here in greater detail in order to put the patterns of missingness in context and to provide a clearer picture of the dataset as a whole. A greater degree of missingness in a peak is seen to be associated with low average intensity values in the observed data.

Broad overview

Figure 2.11 provides a summary of all observed intensities in the GC dataset. Note the threshold at intesity values just under 7 below which no observations are found. This threshold is an artefact of preprocessing but may also be apparent in raw MS data (Stanford, 2015). The increased threshold for peaks below 4866 m/z is also an artefact of preprocessing, induced by a greater level of noise in the raw data at low m/z values. For most peaks, a majority of observed values lie within approximately 2 units of the minimum observed value of 6.97. These peaks almost always exhibit positively skewed distributions

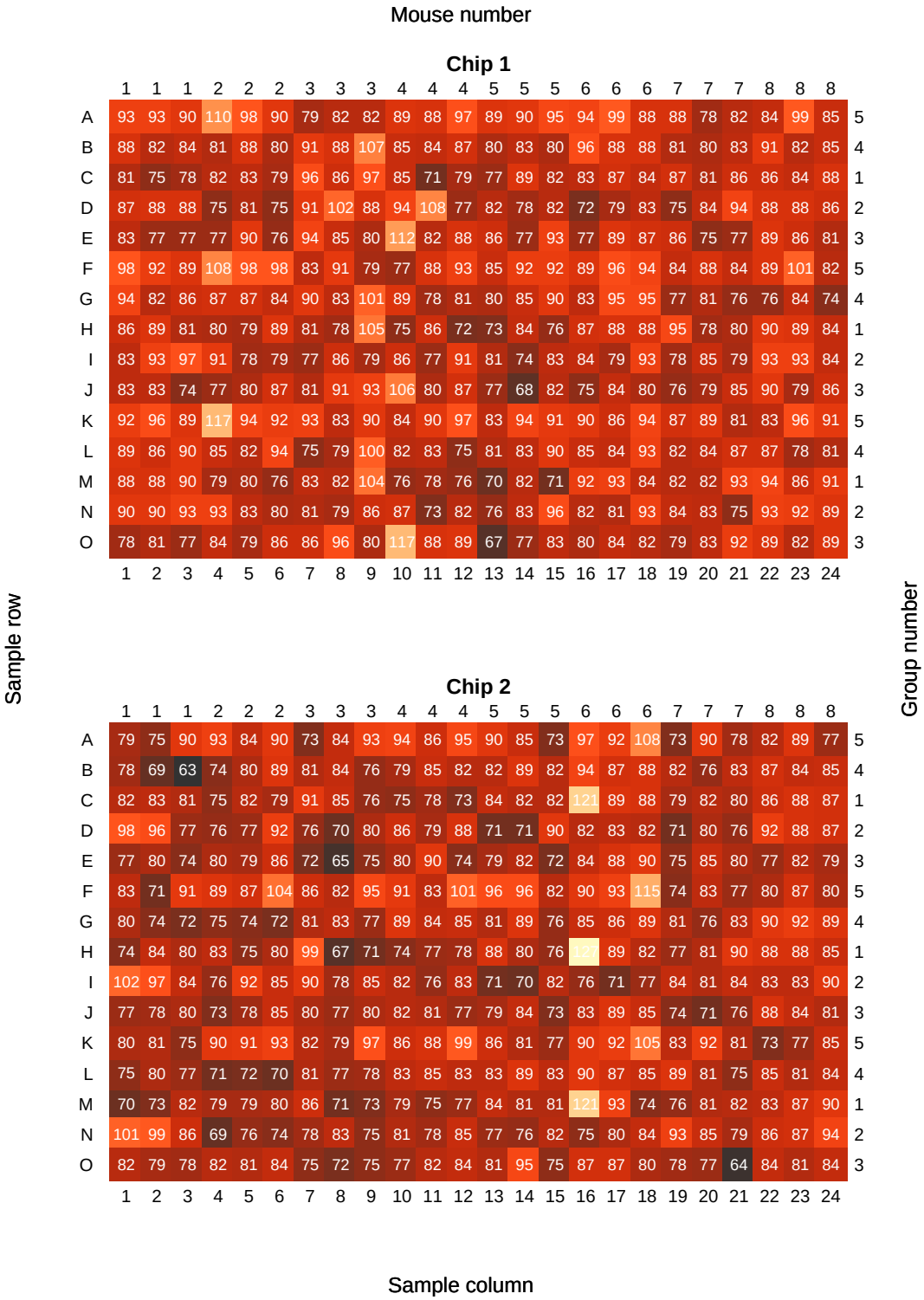

Figure 2.8a: Counts of missing observations (out of a maximum of 159) within each sample on MALDI chips 1 and 2. Each sample is represented by a coloured cell. The missingness counts for each sample are overlaid in white with the visually brighter cells representing larger counts.

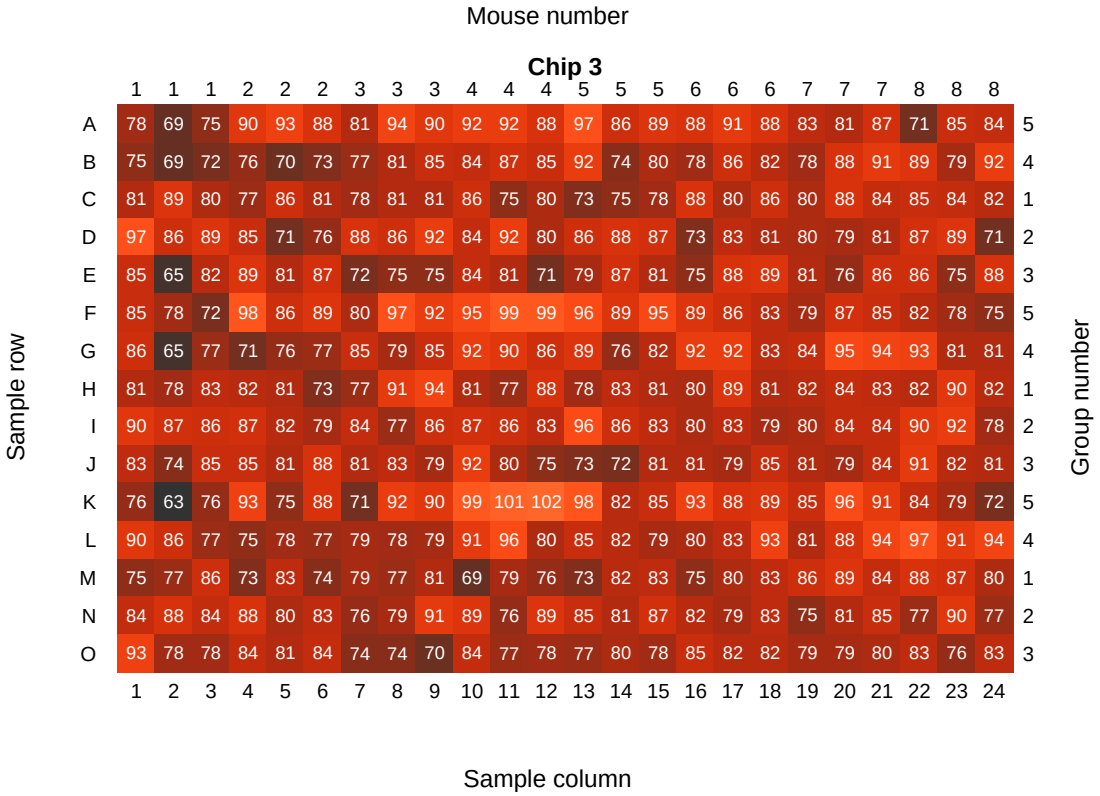

Figure 2.8b: Counts of missing observations (out of a maximum of 159) within each sample on MALDI chip 3. Each sample is represented by a coloured cell. The missingness counts for each sample are overlaid in white with the visually brighter cells representing larger counts.

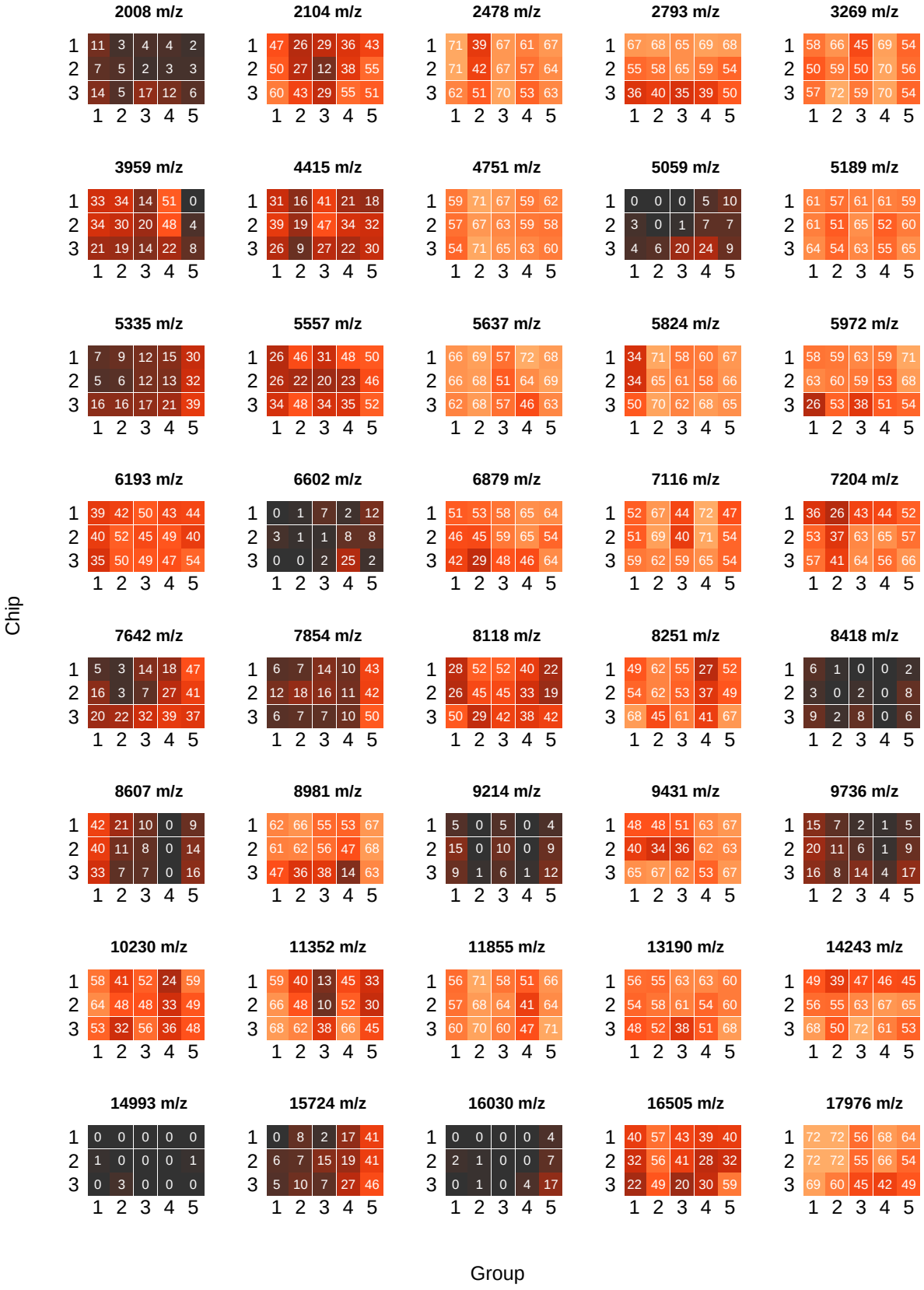

Figure 2.9: Counts of missing observations (out of a maximum of 72) by group number (columns) and chip number (rows) for a selection of peaks. The set of 72 observations for a particular group and chip combination is represented by a coloured cell. The missingness counts for each set are overlaid in white with the visually brighter cells representing larger counts.

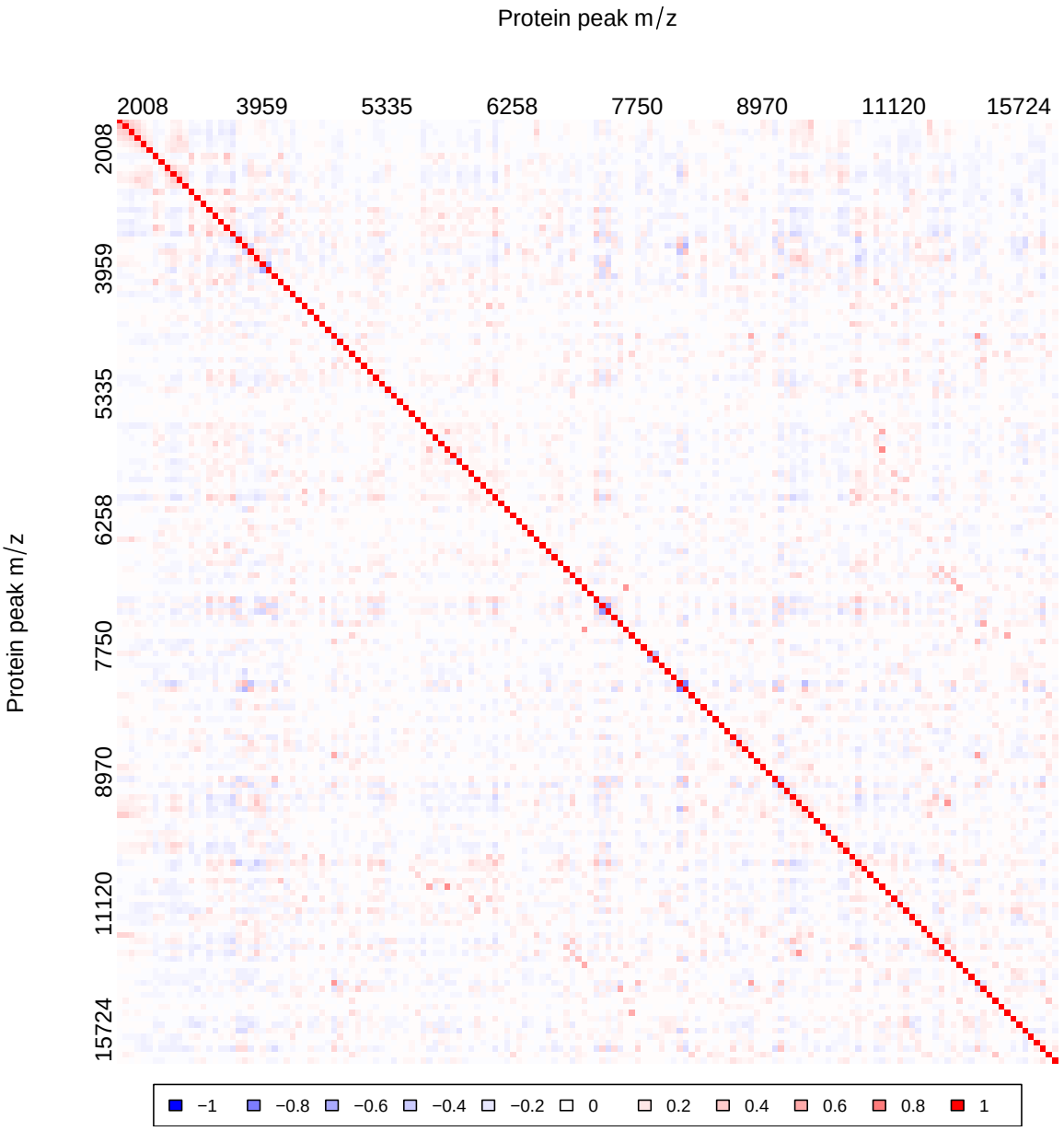

Figure 2.10: Concordance matrix for missingness across peaks in GC dataset. Entries in the matrix, which are sample concordance values between pairs of peaks, are represented as coloured cells. Axes represent the first and second peak of the pair. Peak 4358 m/z is not included as it contains zero missing values.

of intensity values. However, there are some peaks both with median values far above the minimum value and with roughly symmetric distributions of intensity values.

Figure 2.12 plots the medians of observed intensities within peaks versus the missingness proportions of the peaks, revealing an overall negative correlation. There is a complete absence of missingness proportions below 0.5 in peaks with medians below approximately 7.5. Likewise, although it is difficult to make definitive conclusions given that the number of peaks with medians above 9 is low, missingness proportions above 0.5 are rare for such peaks. All of this strongly suggests a dependency of the probability of data missingness on peak intensity in the sense that higher intensities are associated with lower rates of missingness. This dependence cause naive estimates of group differences in intensity based on the observed data to be biased, as the missing low intensities in a group drive estimates of average intensity upwards.

Mouse and genotype group

Subsets of samples from a single mouse and a single group were investigated in order to elucidate the variation in missingness and numerical response at various levels of the hierarchy of the GC dataset.

Figure 2.13 displays intensity values associated with the 27 samples from mouse 1 of group 1 (FF). Most peaks have at least one missing value for these samples, and there is a high degree of variability of observations about their means. A large proportion of variation in both the numerical response and the missingness is present in levels of the hierarchy below the mouse level. This mouse is a typical example of the mice in the GC dataset. A minority of peaks have no observations from the 27 samples. The composition of the set of peaks with zero observations in all 27 samples varies for every mouse. Figure 2.14 displays intensity means for the eight mice from group 1. The variation in intensity *between* mice, while substantial, is less than the variation of intensity *within* mice.

Some peaks contain zero observations for any of the 216 samples in a group, such as peak 4152 m/z which is not represented in any sample from group 2, and peak 11757 m/z which is not represented in any sample from groups 4 or 5.

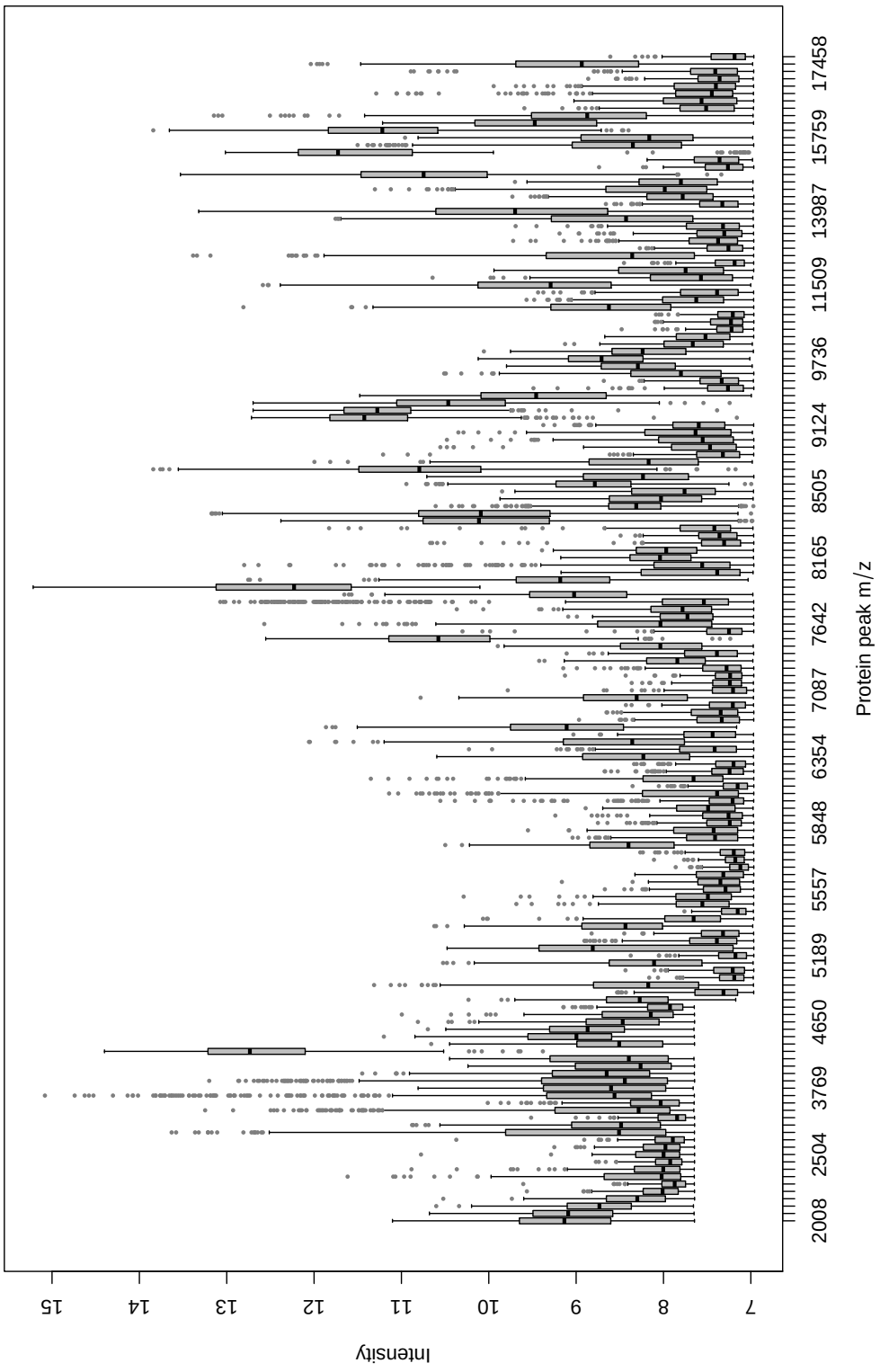

Figure 2.11: Boxplots of all observed values in GC dataset. A box covers the 25–75 percentile range of observations for a peak. The y axis represents intensity. Grey dots represent outlying intensity readings.

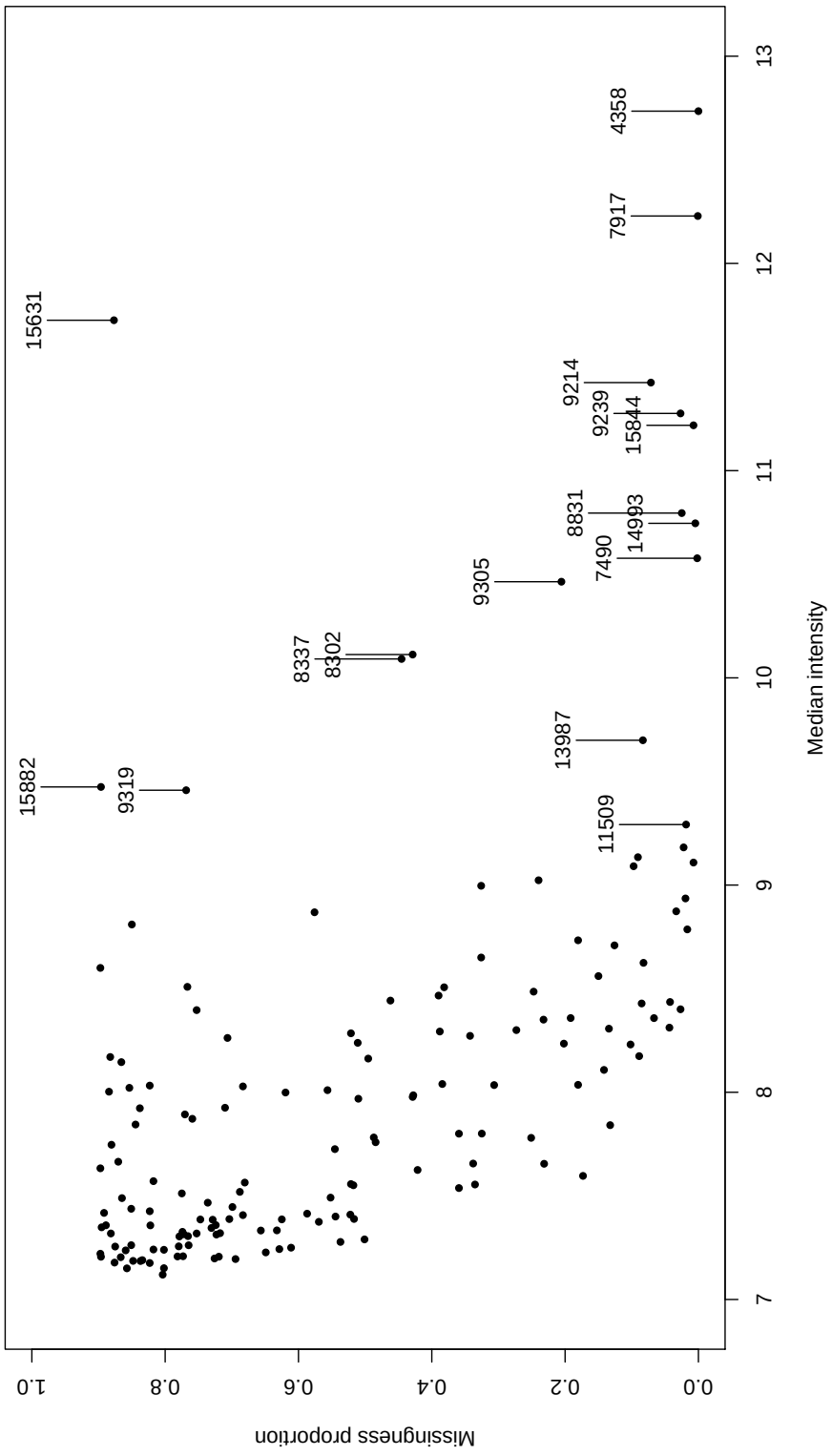

Figure 2.12: Scatterplot of peak m/z intensity median versus proportion of missing values in the peak. Each point represents a peak. The 16 peaks with the largest medians are labelled with their m/z ratio.

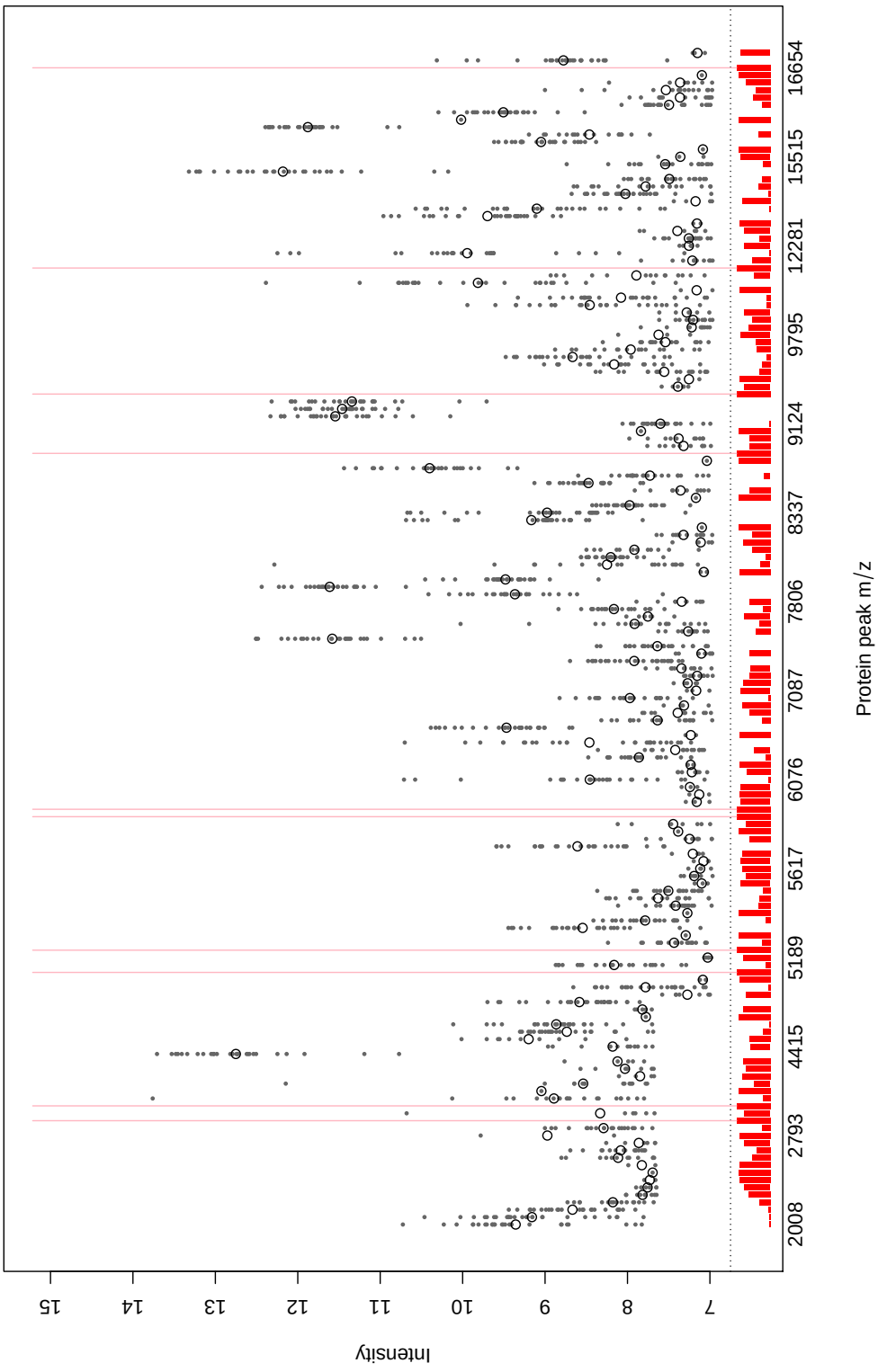

Figure 2.13: Plot of intensity values from a single mouse over the set of peaks. Grey dots represent up to 27 intensity values from one mouse for each peak m/z. Heights of red bars below the dotted line represent counts of missingness up to a maximum of 27. Intensity means are represented by black circles. Complete absences of measurements within a peak are represented by a vertical line.

Figure 2.14: Plot of mean intensity values for mice from a single group over the set of peaks. Black circles represent per-mouse intensity means for group 1 (FF) for each peak m/z. Heights of red bars below the dotted line represent counts of missingness of mouse means up to a maximum of 8.

Individual peaks

Owing to the differing median intensity values across the range of peaks, the peaks are affected by missingness to varying degrees of severity. Figures 2.15 and 2.16 display plots of all observed intensities for a representative subset of eight peaks in the GC dataset. As discussed above, high average intensities are associated with low levels of missingness, and this is readily visible in the plots. Inter-group and inter-chip differences in observed intensity averages are also apparent in the plots.

Figure 2.17 displays the correlation matrix for observations under each peak in the GC dataset. Correlations were calculated using only samples for which *both* observations in the pair of peaks were present. The correlations between most pairs of peaks are weak, with three main exceptions. First, adjacent peaks tend to be positively correlated. These correlations present as red regions adjacent to the diagonal red line that corresponds to self-correlations. Second, in pairs of peaks such that one peak has near double the m/z ratio of the other, correlation appears to be slightly higher than what is typical. These higher correlations appear as faint, irregular red regions either side of the diagonal line. Third, peak pairs with few observations in common sometimes produce extreme correlations due to small sample sizes. These extreme correlations present as isolated, highly saturated red or blue blocks.

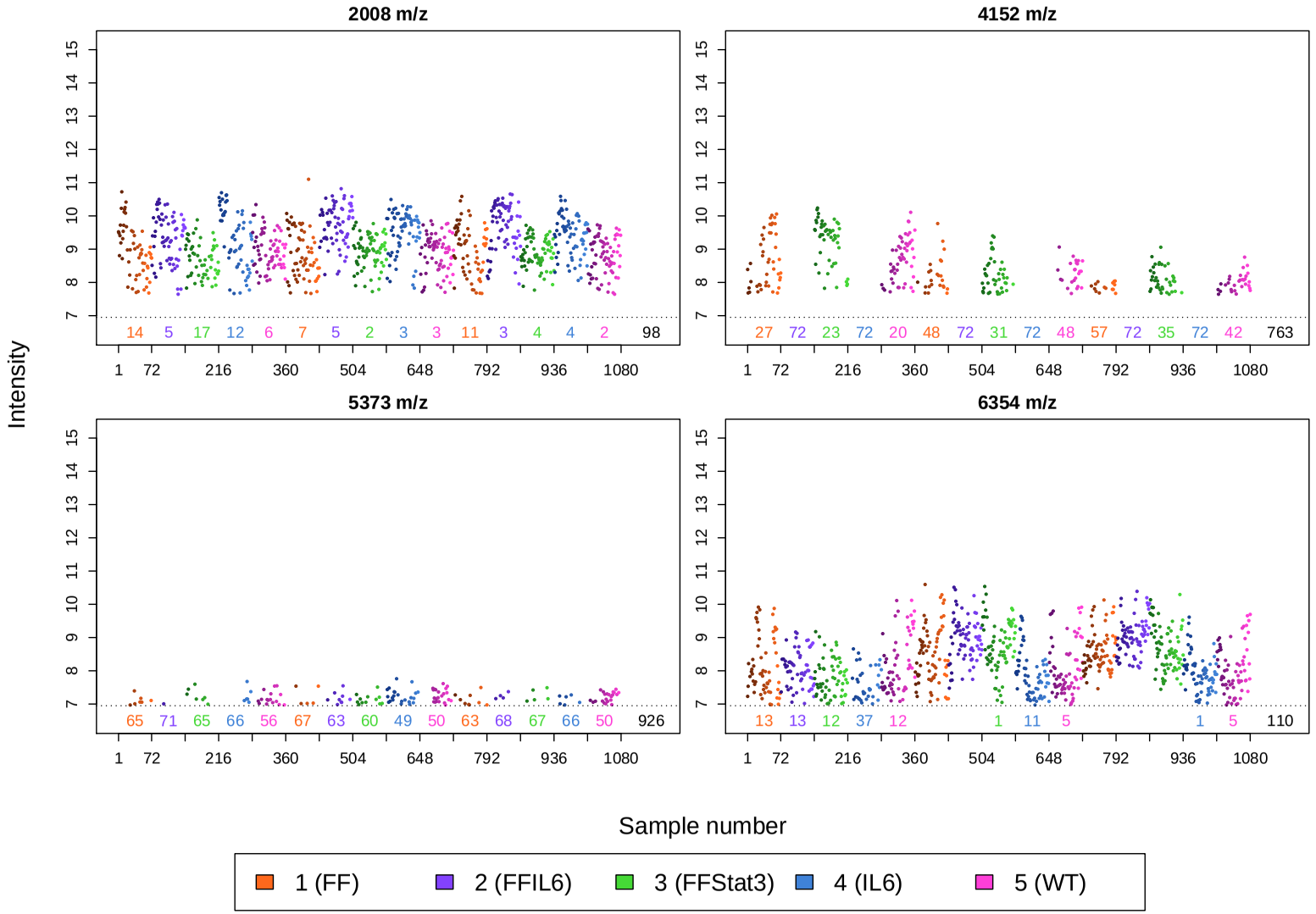

Figure 2.15: Plots of observed data for individual peaks. Coloured points represent intensities plotted against sample index. Colour indicates group membership. Within each group, colour increases in saturation and brightness with increasing mouse number. Coloured numbers below the dotted line indicate the missingness count for observations from a particular group and chip combination (out of 72). The black number indicates the total missingness count (out of 1080).

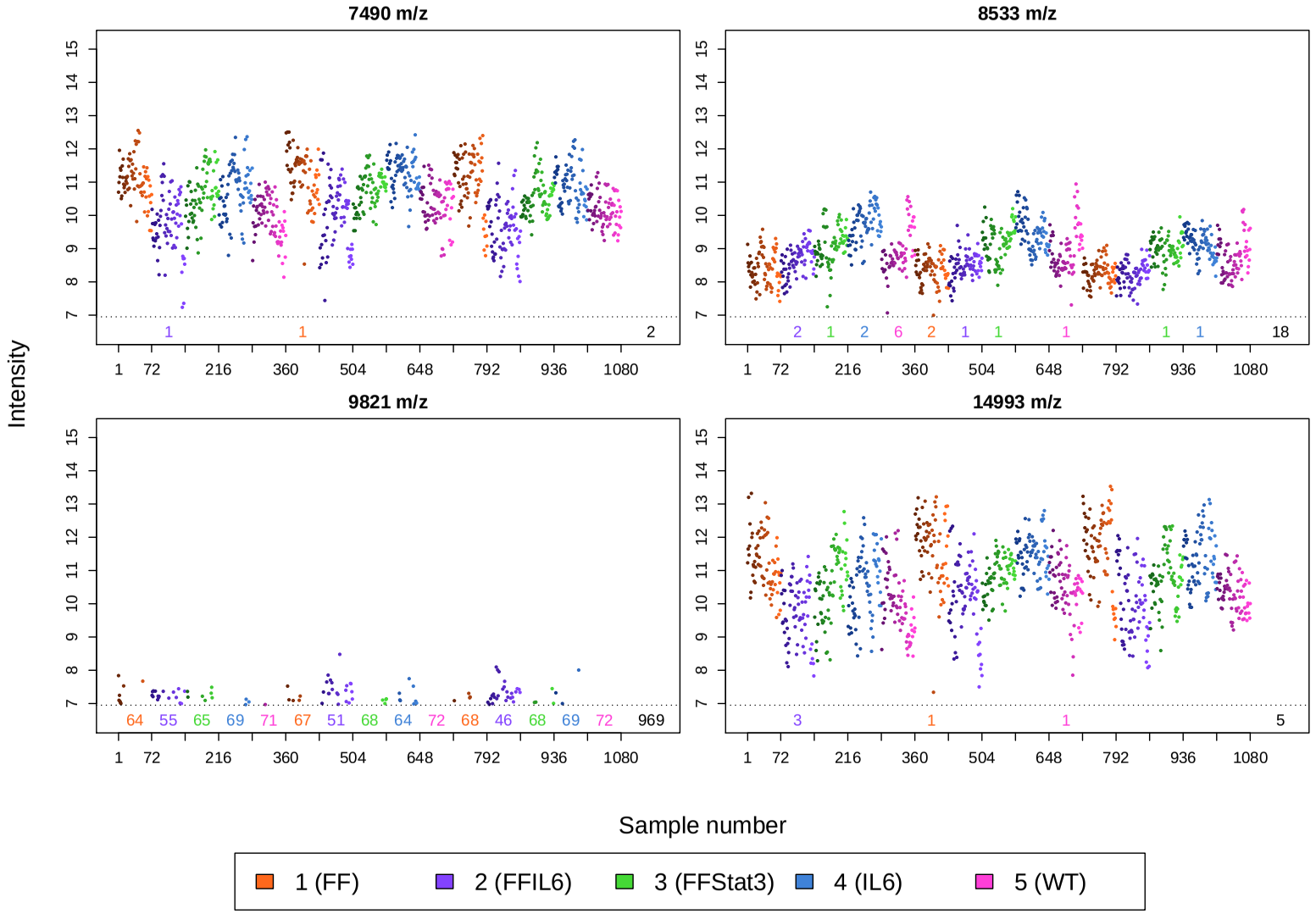

Figure 2.16: Plots of observed data for individual peaks. Coloured points represent intensities plotted against sample index. Colour indicates group membership. Within each group, colour increases in saturation and brightness with increasing mouse number. Coloured numbers below the dotted line indicate the missingness count for observations from a particular group and chip combination (out of 72). The black number indicates the total missingness count (out of 1080).

2.2.4 Correspondences of proteins to peaks

The matching of m/z peaks to protein molecules is beset by ambiguities arising from proteins giving rise to multiple peaks and from distinct proteins with the same atomic mass. In particular, molecules of the same type of protein may pick up single or double charges from the MALDI chip matrix, giving rise to two distinct m/z peaks whose intensities correlate across samples. The GC dataset contains many pairs of proteins which are suspected to be singly and doubly-charged versions of the same molecule on the basis of correlation of intensity values. Table 2.4 provides a listing of pairs gleaned from inspection of the GC dataset. Some additional pairs are suspected to be singly and triply-charged versions of the same molecule for similar reasons. These are the pairs of 5204 and 15631 m/z, of 5275 and 15844 m/z, and of 5335 and 16030 m/z.

Note the subset of four peaks at 4607, 4617, 9214, and 9239 m/z which form three pairs of 4607 and 9214 m/z, 4617 and 9214 m/z, and 4617 and 9239 m/z. It is ambiguous whether all peaks correspond to a single protein performing triple duty, or to two proteins, or possibly to more. However, 4607 and 4617 m/z are likely to correspond to the same protein, based on the complementarity of their missingness patterns as discussed in Section 2.2.2.

2.2.5 Informing future modelling directions

Modelling of the GC dataset is performed on a per-peak basis. Both the missingness and the numerical response in the GC dataset are affected by chip and group effects in the majority of peaks. The hierarchical structure of the dataset must also be accounted for. *Linear mixed models* (LMMs) are a natural choice for modelling the numerical response. This is because the processing steps at the levels of replicates, C8 batches, aliquots, and mice may be understood as samples from hypothetical populations of the same levels, and *random effects* in such models are suitable for modelling the effects of the processing stages. *Generalised linear mixed models* (GLMMs) are suitable for the missingness because missingness is a binary outcome, and are used in Chapter 3 for this purpose. Joint missing/observed data models that account for the response-dependent missingness can extract more information from the GC dataset than separate models for the missingness or the intensity alone. In Chapter 4, joint models are used for the discovery of better biomarker candidates.

An intensity lower threshold cut-off for visibility of observations is one cause of missingness that is consistent with the physical nature of MALDI-TOF MS data acquisition and explains the intensity distributions of the observed data. However, it is likely that the threshold cut-off is not the only reason for missing values, and caution is required in making assumptions about how the data are affected by missingness.

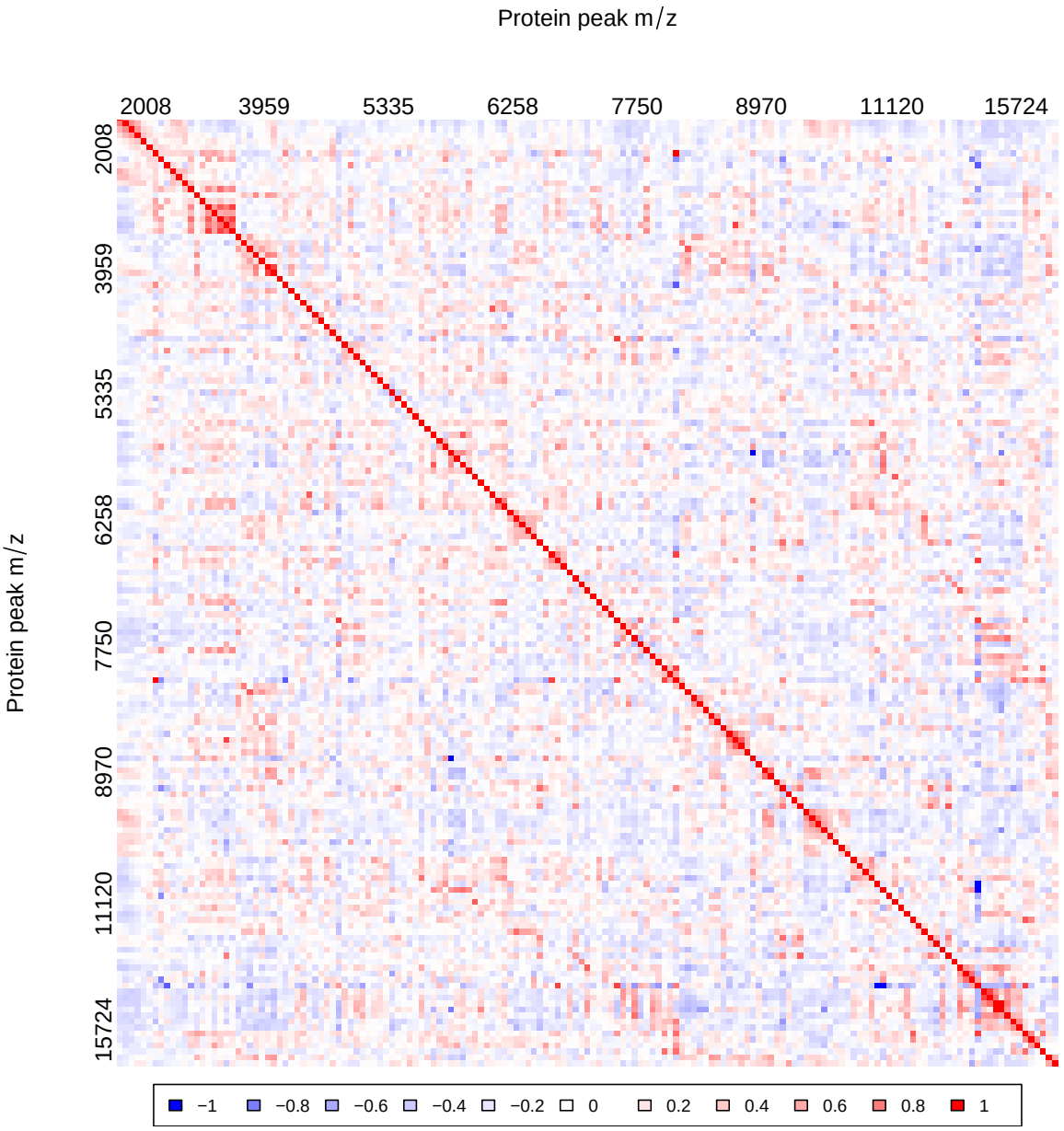

Figure 2.17: Correlation matrix for observations across peaks in the GC dataset. Coloured cells represent the correlation between the observations within samples *common to both peaks* in a pair. Axes represent the m/z ratio of the first and second peak in the pair. Correlation is reported as zero for pairs of peaks with no samples in common.

Table 2.4: List of m/z peak pairs which may correspond to singly and doubly-charged molecules of the same protein molecule. Correlation is measured only from samples with observations in *both* peaks. The number of observations in common is also reported. Pairs are only displayed if the correlation is greater than 0.6 and if the number of observations is at least 100.

Peak m/z	Peak m/z	Correlation	Common observations
3881	7642	0.6	512
3881	8118	0.73	329
3959	7917	0.67	728
4152	8302	0.79	315
4168	8337	0.88	265
4415	8831	0.73	665
4607	9214	0.71	710
4617	9214	0.61	408
4617	9239	0.64	448
4650	9305	0.66	618
5435	10872	0.6	697
5752	11509	0.83	1039
5824	11757	0.72	100
5876	11757	0.68	104
6076	12161	0.87	668
6821	13648	0.83	987
6858	13648	0.61	355
6879	13648	0.62	279
6989	13987	0.75	893
6989	14421	0.66	607
7204	14421	0.69	290
7412	14836	0.72	695
7490	14993	0.89	1075
7806	15631	0.95	132
7854	15724	0.7	740
7854	15759	0.66	426
8007	15759	0.62	537
8007	15844	0.73	1051
8007	15882	0.68	105
8007	16030	0.81	1037
8165	15844	0.64	664
8228	16654	0.82	116
8533	17458	0.69	1046
8607	17458	0.6	857

2.3 Summary

The dataset analysed in this thesis originates from a proteomic mass spectrometry experiment measuring protein concentration in blood samples from transgenic mice. These mice come from five genotype groups defined by mutations in the genes for glycoprotein 130, signal transducer and activator of transcription 3, and interleukin-6. Phenotypic changes from a healthy state associated with these mutations include the formation of gastric tumours and the presence of gut inflammation. The GC dataset has a hierarchical structure induced by the nature of the processing of blood serum samples from the mice. In modelling the data, this hierarchical structure must be accounted for.

The GC dataset consists of a matrix of observations and metadata describing the samples on which the observations were made. Approximately half of the expression values are missing. The pattern of missingness is complex, and highly dependent on the underlying protein concentration values. Visualisation of the GC dataset elucidates the structure of the numerical observations and the missingness pattern and informs statistical modelling of the data.

The hierarchical structure of the GC dataset can be accounted for by using generalised linear mixed models for the missingness pattern in Chapter 3. The association between median intensity value and missingness proportion suggests that the distributions of the intensity and the missingness pattern are not independent, which necessitates a joint modelling approach. Statistical models for the joint distribution of the intensity and missingness pattern are fitted to the GC dataset in Chapter 4.

Chapter 3

Modelling the missingness in the dataset

In this chapter, the goal of the analysis of the missingness in the GC dataset is elaborated upon. The initial statistical framework for the analysis of the missingness in the data is presented. Several candidate models are introduced, and the most appropriate model selected. Model selection was done using cross-validation to estimate the lowest misclassification rates of the models in terms of outcomes of missingness. The problem of separation of data, caused by certain patterns of uniformity in the response variable, affected some models, and a Bayesian solution to the problem was introduced to select a final missingness model. The selected missingness model was checked using simulations and its total misclassification rates were investigated to verify that the model was appropriate for the GC dataset.

3.1 The aim of the missingness modelling

The purpose of the statistical analysis of proteomic MS data in this thesis is the discovery of candidate protein biomarkers that distinguish between diseased and healthy groups of organisms. Biomarker candidates are proteins whose concentrations differ between the groups. The intensity expression measurements in the GC dataset are $\log_2$-transformed measurements of concentration, meaning that two-fold increases or decreases (i.e., doublings or halvings) in concentration are represented by a difference of one unit. Statistical inference for genotype group expression differences is used to discover candidates for such biomarkers. However, without accounting for the informative missingness in the GC dataset, such inference is biased.

Understanding the source of the missingness in the GC dataset, and the extent of its effect on inference on group mean expression, is difficult. In order to elucidate the nature of the informative missingness, the missingness pattern is modelled on its own in this chapter before inference for the protein expression differences between groups is made using joint models in Chapter 4.

3.2 Modelling framework

There are several approaches to modelling that could be followed on the GC dataset. One is to model the observed intensities, disregarding the samples for which the observations in a particular column are missing. This is a complete-case analysis that ignores the missingness. This approach is not performed in this thesis, but it has has been undertaken by Stanford (2015), who obtained a set of candidate biomarkers for gastric cancer, given in Table 3.13.

A second approach is to model the missingness. This approach is investigated in this chapter. The missingness modelling was done using the *indicators* of the missingness (Little and Rubin, 2002), where 1 codes for a missing datum and 0 codes for an observed datum. The expectation of the indicator variable is the probability of missingness. The indicators of missingness are never themselves missing.

Section 2.2 in Chapter 2 provided a detailed description of the GC dataset. Columns of the `dd` matrix, where each column corresponds to the 1080 missing and observed intensities of a single m/z peak, are used to produce the response vector of missingness indicators in the statistical models in this chapter.

The complete case analysis models the intensity values using hierarchical linear models, specifically, LMMs. The missingness modelling is done using hierarchical logistic regression (an example of a GLMM) for the probability of missingness. In both approaches, the hierarchical structure of the GC dataset is handled using *mixed effects*, which are a combination of *fixed effects* and *random effects* in the linear predictor of the model. Snijders and Bosker (2012) and Gelman and Hill (2009) both provide a thorough introduction to

mixed effects models.

Fixed effects were used to model the effects of the genotype groups and the MALDI chips on the responses. Random effects were used to model the populations of mice, aliquots within mice, and C8 batch replicates within aliquots. The reasons that fixed effects were used for the chip and group variables were that the specific group parameters were of interest and that the group and chip variables both contain a small number (5 or fewer) of categories. Random effects were used for modelling the mice, aliquots, and C8 batches as these levels of the hierarchy are best understood as samples from hypothetical populations. Overparametrisation of models was also a concern, as a random effect term for a categorical variable has one parameter, a *variance component*, associated with it, while a fixed effect term for the same variable has as many parameters as there are categories (Searle et al., 2009).

Models were fitted using the `glm` function, the `glmer` function in the R package `lme4` (Bates et al., 2015), and the `bglmer` function in the package `blme` (Chung et al., 2013). Maximum likelihood estimation was used to obtain parameter estimates for GLMs and non-Bayesian GLMMs, and maximum a posteriori estimation was used to obtain parameter estimates for Bayesian GLMMs. Appendix C.1 provides details of the parameter estimation methods for mixed effects models.

3.2.1 Notation for missingness mechanisms

Mechanisms that cause data to be missing may be treated as random processes. Little and Rubin (2002) present a framework for classifying types of missingness with which the GC dataset may be analysed. Suppose that Y is a random variable representing a collection of observations. The variable Y is subject to missingness, and R denotes its missingness indicator. The statistical problem is to make inference for a parameter θ controlling $f(y\,;\,\theta)$, the probability distribution of Y, where it is assumed that the parameter θ does not affect missingness (Davison, 2003).

The *missingness mechanism* is the probability, conditional on the values of Y, that the observation of Y is missing—that is, that $R = 1$. Per Little and Rubin (2002), missingness mechanisms may be classified into one of three types. First, the notion of data being *missing completely at random* (MCAR) corresponds to a missingness mechanism that is fully independent of the response Y. Second, the notion of data being *missing at random* corresponds to a missingness mechanism that depends on the observed components of Y, but not the missing components. Third, the notion of data being *not missing at random* (NMAR) corresponds to missingness mechanisms that are dependent on the missing values of Y.

In the models of the GC dataset, the data consists of a vector of protein expressions $\boldsymbol{Y}$, the vector of missingness indicators $\boldsymbol{R}$, and a matrix of predictor variables X. The missingness is confined to $\boldsymbol{Y}$, meaning that $\boldsymbol{R}$ is the same length as $\boldsymbol{Y}$. A MCAR mechanism for the GC dataset is one where neither X nor $\boldsymbol{Y}$ may affect the probability of

missingness. A MAR mechanism is one where only elements of X or the observed component of $\boldsymbol{Y}$ may affect the probability of missingness. A NMAR mechanism is one such that the missing component of $\boldsymbol{Y}$ affects the probability of missingness. In this thesis, missingness models that assume a MAR mechanism consider the effect only of X, and not of the observed component of $\boldsymbol{Y}$, on $\boldsymbol{R}$. As it is not known in advance of data collection which values of $\boldsymbol{Y}$ are missing in a typical mass spectrometry study, a mechanism that depends on values in $\boldsymbol{Y}$ is likely to be a NMAR mechanism.

NMAR missingness mechanisms are likely to bias inference for the parameter θ (Little and Rubin, 2002). Unbiased statistical inference may require full specification of the missingness mechanism. One way of reducing the bias of parameter estimates is to completely specify the mechanism by modelling the joint distribution of ($\boldsymbol{Y}$, $\boldsymbol{R}$) (Follmann and Wu, 1995; Hogan and Laird, 1997); this is the approach taken in Chapter 4.

The missingness mechanism that affects the GC dataset is considered to be NMAR because of the increased difficulty of detecting low concentrations of protein in a sample as compared to higher concentrations (Hajduk et al., 2016). Furthermore, the data have been preprocessed (Stanford, 2015) to remove noise from the raw data, resulting in additional missingness of low-intensity measurements that fell below a threshold signal/noise ratio.

3.2.2 Notation for the GC dataset and models

The predictor variables in the GC dataset were derived from the metadata for each sample. The predictor variables are categorical variables defining the chip, group, mouse, aliquot, and the C8 batch numbers of each sample.

The response variable, which is the indicator of missingness, is written $R_{ijk\ell mn}$. The indices for the response variable are $i = 1, \ldots, 159$ representing the set of m/z peaks in the dataset, $j = 1, \ldots, 5$ representing group, $k = 1, \ldots, 8$ representing mouse number within a group, $\ell = 1, 2, 3$ representing aliquot number within a mouse, $m = 1, 2, 3$ representing C8 batch number within an aliquot, and $n = 1, 2, 3$ representing sample replicate number within a C8 batch. Because of the confounding between aliquot and chip, the chip number is the same as the aliquot number. For this reason, the chip number does not participate in the indexing.

As modelling was done on a per-peak basis, the term i in the subscript is suppressed in the context of an individual peak and its model, in which case we write the missingness indicator as $r_{jk\ell mn}$.

3.3 Ascertaining the hierarchical structure of the missingness model

Models explored for the missingness in the GC dataset were binary logistic regression models, most of which assumed random effects as well as fixed effects. These models are

specified as

$$p_{jk\ell mn} = \frac{e^{\eta_{jk\ell mn}}}{1 + e^{\eta_{jk\ell mn}}},$$

where $p_{jk\ell mn} = E[R_{jk\ell mn}]$ is the probability of missingness and $\eta_{jk\ell mn}$ is the *linear predictor* containing fixed (and possibly random) effects. The models considered differed only in the random effects structure used to model the hierarchy of the sample processing. Models were built by starting from a logistic regression model with no random effects, and adding random effects term by term.

The peak at 4358 m/z contained zero missing observations and was excluded from the analysis. This means that a 158-peak subset of the original 159 peaks was considered for analysis. The two peaks at 4152 and 11757 m/z have such severe missingness that for at least one group, all 216 observations from that group are missing. This means that parameter estimates could only be obtained from these peaks using Bayesian methods. Comparison of models was performed on a 143-peak subset of the GC dataset due to the issue of *separation of data*, discussed further in Subsection 3.4.1, that affected estimates of the fixed effect parameters. Parameter estimates were reliable only within the subset of peaks for which separation did not occur, meaning that model comparison was performed only within that subset.

3.3.1 List of models under consideration

Four different models were considered for modelling the missingness in the GC dataset. The initial model was the simplest,

$$R_{jk\ell mn} \sim \text{Bern}(p_{jk\ell mn}), \quad p_{jk\ell mn} = \frac{e^{\eta_{jk\ell mn}}}{1 + e^{\eta_{jk\ell mn}}}, \quad \eta_{jk\ell mn} = \mu + \alpha_j + \beta_\ell\,, \tag{3.1}$$

where the indices take the ranges $j = 1, \ldots, 5$, $k = 1, \ldots, 8$, $\ell = 1, 2, 3$, $m = 1, 2, 3$, and $n = 1, 2, 3$. The model was fitted with the `glm` function from the `stats` package in R.

The fixed effect parameter μ represents the log-odds of missingness for samples from group 1 and chip 1. The differences between group 1 and group j are parametrised by the α_j parameters, and the differences between chip 1 and chip ℓ are parametrised by the β_ℓ parameters. Both sets of parameters are fixed effects and are under the reference category constraint.

The 72 samples that share the same group and chip were given the same probability of missingness.

The next model considered was

$$R_{jk\ell mn} \sim \text{Bern}(p_{jk\ell mn}), \quad p_{jk\ell mn} = \frac{e^{\eta_{jk\ell mn}}}{1 + e^{\eta_{jk\ell mn}}}, \quad \eta_{jk\ell mn} = \mu + \alpha_j + \beta_\ell + M_{jk}\,, \tag{3.2}$$

where the fixed effects parameters μ, α_j, and β_ℓ are the same as in model (3.1), and the effect of mouse is represented by the random effect terms M_{jk} which follow the distribu-

tional assumption

$$M_{jk} \sim N(0,\, \sigma_M^2) \quad i.i.d.,$$

where i.i.d. means that the terms M_{jk} are independent and identically distributed. This model and all subsequent models were fitted with the `glmer` function from the `lme4` package in R.

In this model, the nine samples that originate from a single aliquot are given the same probability of missingness.

The third model considered was

$$R_{jk\ell mn} \sim \text{Bern}(p_{jk\ell mn}), \quad p_{jk\ell mn} = \frac{e^{\eta_{jk\ell mn}}}{1+e^{\eta_{jk\ell mn}}}, \quad \eta_{jk\ell mn} = \mu+\alpha_j+\beta_\ell+M_{jk}+C_{jk\ell m}\,. \tag{3.3}$$

The terms in common with model (3.2) have the same interpretation. The additional random effects terms $C_{jk\ell m}$ represent the effect for the C8 batch variation, and the terms satisfy the distributional assumption

$$C_{jk\ell m} \sim N(0,\, \sigma_C^2) \quad i.i.d.,$$

with the $C_{jk\ell m}$ terms also distributed independently of the M_{jk} terms.

The three replicate samples that originate from the same C8 batch are given the same probability of missingness.

The fourth model considered was the most complex, modelling all levels of the sample processing hierarchy. The model was

$$\begin{aligned} R_{jk\ell mn} \sim \text{Bern}(p_{jk\ell mn}), \quad p_{jk\ell mn} &= \frac{e^{\eta_{jk\ell mn}}}{1+e^{\eta_{jk\ell mn}}}, \\ \eta_{jk\ell mn} &= \mu + \alpha_j + \beta_\ell + M_{jk} + A_{jk\ell} + C_{jk\ell m}\,. \end{aligned} \tag{3.4}$$

The additional terms $A_{jk\ell}$ are the random effect terms for the aliquot variation, and the terms satisfy the distributional assumption

$$A_{jk\ell} \sim N(0,\, \sigma_A^2) \quad i.i.d.,$$

with the $A_{jk\ell}$ terms also distributed independently of the other random effects terms. The three replicate samples that originate from the same C8 batch are, again, given the same probability of missingness.

3.3.2 Results of cross-validation

Models were compared by using five-fold cross-validation to estimate the expected out-of-sample *misclassification rate*, where classifications are predicted outputs of R based on the fitted probabilities $\hat{p}$. (If $\hat{p} < 0.5$, then R is predicted to be 0, and if $\hat{p} \geq 0.5$, then R is predicted to be 1.) The out-of-sample misclassification rate is the probability

that a model, fitted to a dataset, produces an incorrect classification on a new data point sampled from the same distribution as that which produced the data to which the model was fitted (Hastie et al., 2009).

In cross-validation, the data are partitioned into a number of folds. The model is fitted to data from all but one fold, and the observations from that fold are predicted using the fitted model. This is repeated, leaving out each fold in turn, providing predictions for all observations. The errors between the predictions and the original data are used to obtain the estimates of the out-of-sample misclassification rate (Hastie et al., 2009). In this thesis, the folds were generated independently for each peak by randomly assigning 216 missingness indicators to each of five folds, ignoring the hierarchical structure of the data. The folds within each peak were the same for all models.

In the context of mixed effects models, cross-validation methods are made more difficult by the additional considerations of out-of-sample predictions being difficult to obtain when random effects are present, and of the selection of the folds being in accordance with the hierarchical structure of the data (Colby and Bair, 2013; Jordan et al., 2005). The method used here does not account for these considerations. In particular, assigning folds completely at random without accounting for the hierarchical structure of the data is liable to underestimate the prediction error (Roberts et al., 2017). While this means the method is not very sophisticated and has room for improvement, the method is merely one source, among multiple sources, of guidance in model selection in this chapter.

The numbers of misclassifications for models (3.1), (3.2), (3.3), and (3.4), averaging over the set of peaks, were 240.2, 210.1, 196.4, and 194.3 respectively. These are the average numbers of missingness indicator observations, out of 1080, that were misclassified. The corresponding rates (proportions) were 0.222, 0.195, 0.182, and 0.180. An estimate of misclassification rate was not able to be obtained for the peaks at 5752 and 8007 m/z for model (3.3), nor for the peak at 8007 m/z for model (3.4). This was due to particular folds in these peaks causing errors in the `glmer` algorithms when they were excluded in the cross-validation procedure. Misclassification rate averages were calculated excluding these peaks.

There is a marked difference in performance between the initial model (3.1), with no random effects and model (3.2), the second model considered, which incorporated the random effect term M_{jk} for the mice. Figure 3.1 displays a comparison of the misclassification rates of these two models over the 143-peak subset. The difference between average misclassification rates is immediately apparent. This was expected due to the clear effects of the sample processing hierarchy on the response of missingness as observed in Chapter 2, and implies that there is significant variation in missingness probabilities between samples from the same mouse but different aliquots and C8 batches.

There is also a marked difference in performance between model (3.2) and models (3.3) and (3.4), the third and fourth models considered. However, the difference between the third and fourth models is small. Figure 3.2 displays differences in the misclassification

Table 3.1: Total misclassifications for the models under consideration, displayed for ten peaks.

Peak m/z	Model (3.1) (None)	Model (3.2) (Mouse)	Model (3.3) (Mouse, C8)	Model (3.4) (Mouse, aliquot, C8)
2008	98	98	94	94
3246	425	355	297	306
5059	96	96	99	99
5590	238	226	235	221
6258	308	281	267	251
7412	199	146	140	140
8441	366	295	298	303
9712	341	201	189	185
12281	230	222	203	202
15882	112	113	113	113

rates of the latter three models. It is clear that the two latter models were superior to the first. Model (3.4), with three random effects, performs almost identically to model (3.3), with two random effects. This may be because allowing for variation at the level of C8 batches without that of aliquots causes the aliquot-level variation to be captured by the C8 batch-level variation.

Table 3.1 displays the numbers of misclassifications from the four models under consideration on a subset of ten peaks. There was much inter-peak variation in misclassification performance for the models. The peaks for which model (3.1) was outperformed by the other models by the greatest margins were those peaks where the missingness indicators *within* mice or C8 batches are closely associated, but *between* C8 batches or mice, are not. In 15 cases, model (3.1) actually performed the best out of all four models. This occurred in peaks such as 2948, 5059, 5275, 11509, and 15882 m/z. Missingness indicators in peaks such as these do not show strong association within mice or C8 batches. However, in these cases, the difference in the number of misclassifications between the best and worst models tended to be low, on the order of 12 or less. For several peaks (such as 2008 m/z), multiple models shared the position of having the best misclassification rate.

On 29 of the peaks in the 143-peak subset, model (3.2) with one random effect performed the best out of the three models that included random effects. The average difference within these 29 peaks in the number of misclassifications between model (3.2) and model (3.3) was around 4.4, and between model (3.2) and model (3.4), around 5. Model (3.3) with two random effects performed best on 62 peaks. The average difference within these 62 peaks in the number of misclassifications between model (3.3) and model (3.2) was around 20.4, and between model (3.3) and model (3.4), around 3. Model (3.4) with three random effects performed best on 71 peaks. The average differ-

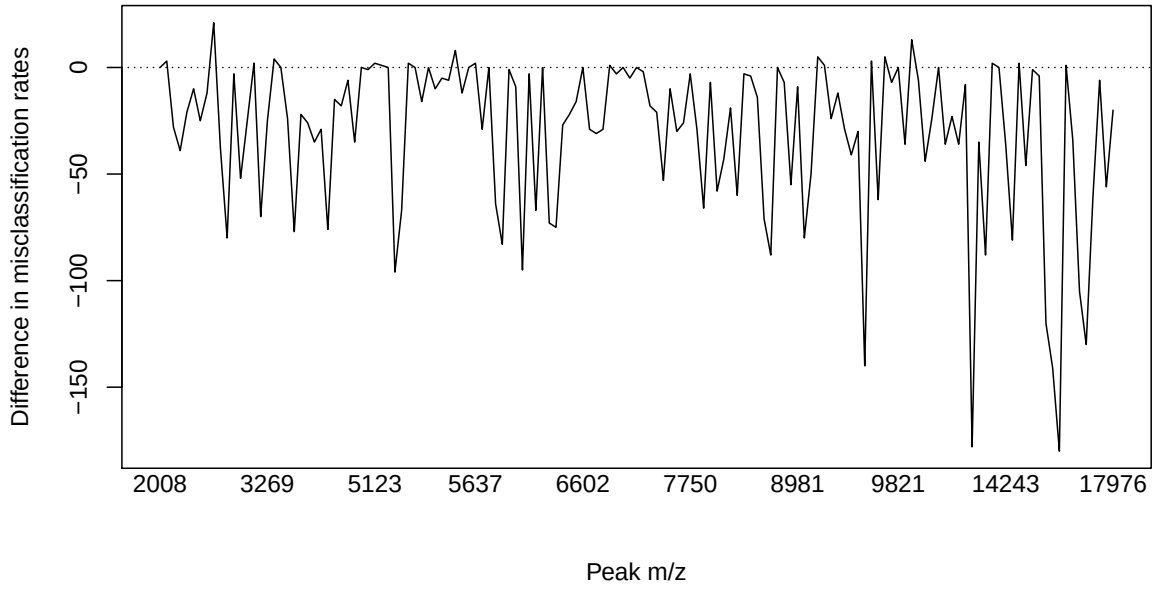

Figure 3.1: Comparison of model (3.1) to model (3.2) in terms of misclassifications made. These are the models with zero and one random effect respectively. The x axis ranges over the 143-peak subset and the y axis represents the number of misclassifications made by model (3.1) subtracted from those made by model (3.2). Negative differences correspond to model (3.2) making fewer misclassifications.

ence within these 71 peaks in the number of misclassifications between model (3.4) and model (3.2) was around 24.1, and between model (3.4) and model (3.3), around 4.6.

The results of the cross-validation suggest that models (3.3) and (3.4) are the best-performing missingness models for the GC dataset. Model (3.3), however, is more attractive on the basis of parsimony.

3.3.3 Variance components

Models (3.3) and (3.4) were very similar in their performance on the GC dataset as estimated by the cross-validation procedure. The additional effect of considering the aliquot variation in the latter model as opposed to excluding it in the former model was investigated through the relative sizes of the variance components σ^2_M , σ^2_A , and σ^2_C for the mouse, aliquot, and C8 batch effects, respectively.

Figure 3.3 displays the relative sizes of the variance components for models (3.3)

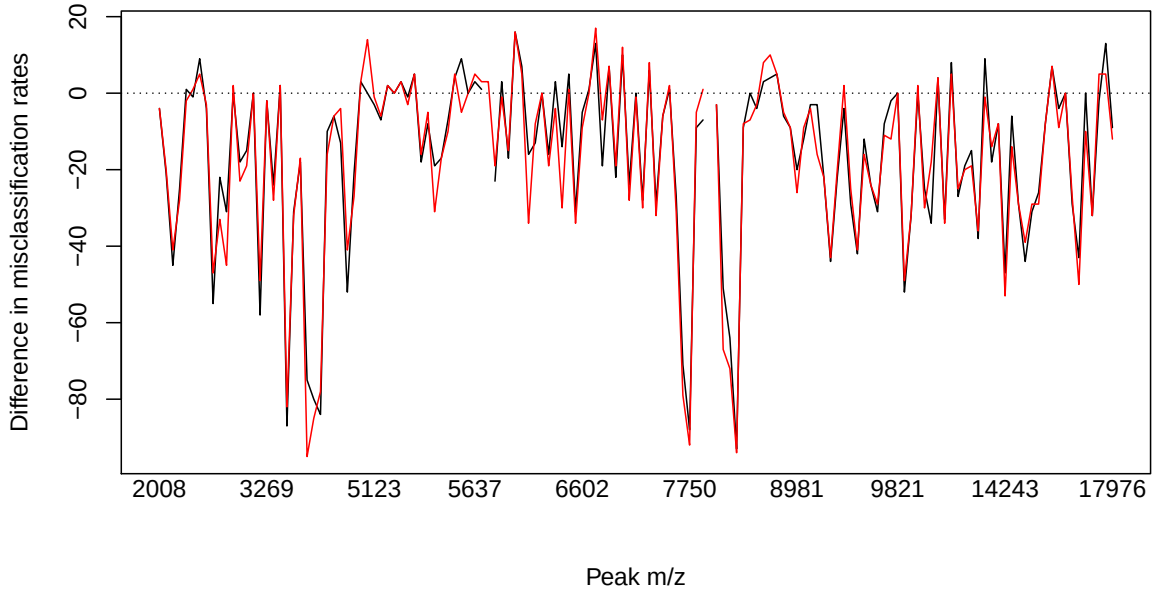

Figure 3.2: Comparison of models (3.2), (3.3), and (3.4) in terms of misclassifications made. These are the equations of the models with one, two, and three random effects respectively. The black line tracks model (3.2)'s numbers of misclassifications subtracted from those of model (3.3), and the red line tracks model (3.2)'s numbers of misclassifications subtracted from those of model (3.4).

Table 3.2: Estimates of variance components under model (3.3) and model (3.4) displayed for ten peaks.

	Model (3.3)		Model (3.4)		
Peak m/z	σ^2_M	σ^2_C	σ^2_M	σ^2_A	σ^2_C
2008	0.328	18.913	0.005	0.454	44.383
3246	1.048	1.675	0.926	0.477	1.321
5059	0.129	13.766	0.016	0.412	9.651
5590	0.34	1.339	0.177	0.632	0.9
6258	0.232	0.371	0.141	0.322	0.144
7412	2.652	2.688	2.653	0.009	2.683
8441	0.386	0.76	0.295	0.392	0.485
9712	3.319	1.854	3.226	0.451	1.575
12281	2.362	1.356	2.156	0.972	0.748
15882	1.102	0	1.102	0	0

and (3.4) fitted to the peaks in the 143-peak subset. Table 3.2 gives the variance component estimates over a subset of ten peaks. The variance component associated with the C8 variation tended to be the greatest in most peaks for both models. The aliquot variation was rarely the largest contribution to the variance components in the latter model.

For model (3.3), the means of the variance components σ^2_M and σ^2_C were 1.970 and 5.515. The standard deviations were 4.450 and 11.25. For model (3.4), the means of the variance components σ^2_M , σ^2_A , and σ^2_C were 1.818, 1.519, and 4.027. The standard deviations were 6.546, 3.635, and 9.645.

The proportional sizes of the aliquot variance component estimates from model (3.4) were small, when taken over the 143-peak subset. In accordance with considerations of parsimony, model (3.3) appears the more suitable of the two models for modelling the GC dataset.

3.3.4 Theoretically minimal versus achieved misclassifications

The lowest misclassification rates theoretically achievable by the missingness models fitted to the entire dataset are greater than zero. This is because not all configurations of missingness indicators are able to be predicted without errors by the models. For example, models (3.3) and (3.4) produce identical fitted probabilities for all 360 triplets of observations originating from the same C8 batch. This means that $p_{jk\ell m1} = p_{jk\ell m2} = p_{jk\ell m3}$. However, if the triplet of observations originating from a single C8 batch are not all missing or not all non-missing, then it is not possible for these models to classify the probabilities perfectly. Every instance of one missingness indicator in a triplet differing from the other two increases the minimum possible misclassification count by one. Similarly, model (3.2) gives identical fitted probabilities for all nine observations sharing the same mouse and

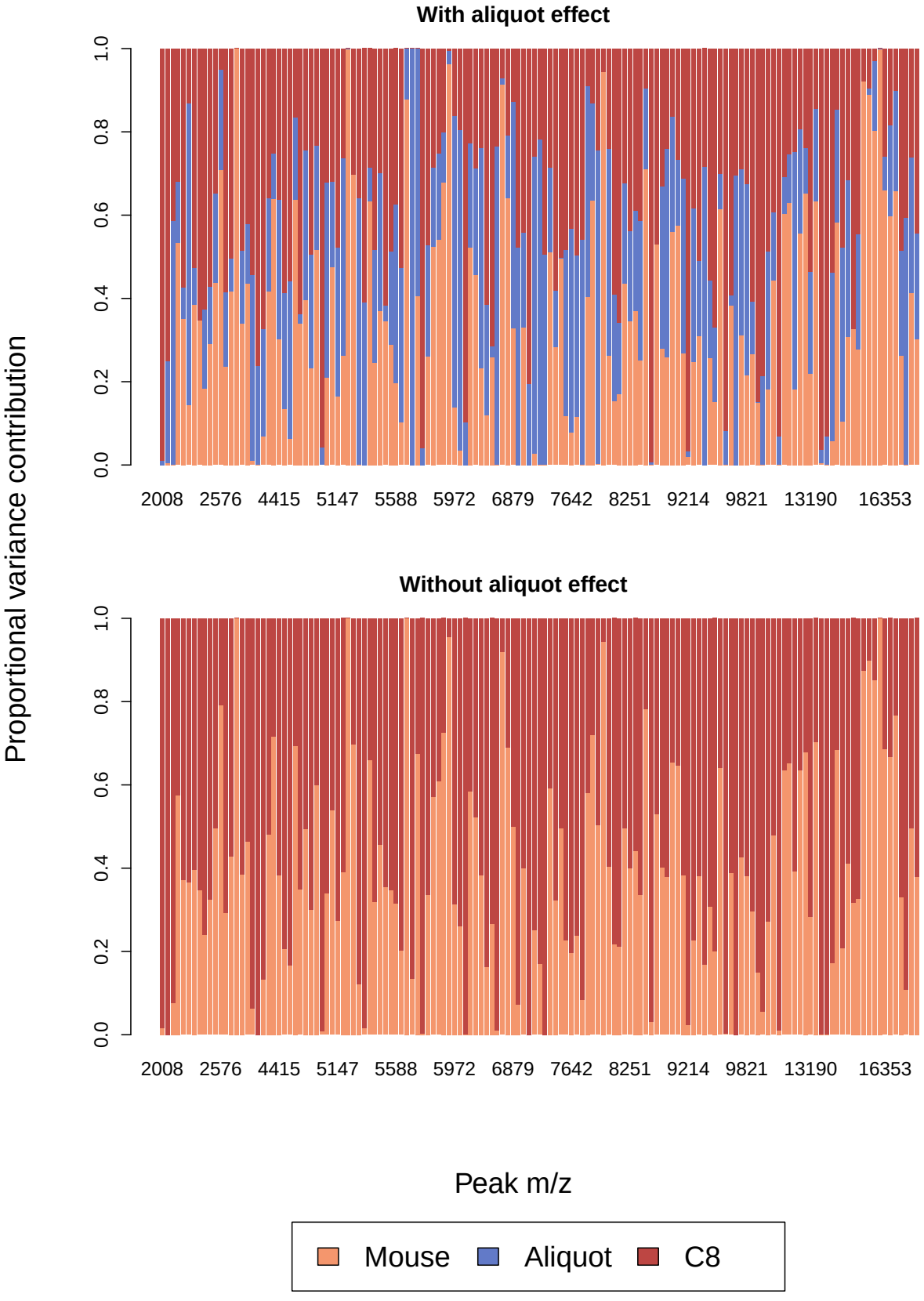

Figure 3.3: Comparison of relative sizes of variance component estimates for model (3.4) (top) and model (3.3) (bottom) fitted to the peaks in the 143-peak subset.

Table 3.3: Average theoretical and achieved numbers of misclassifications by models for the 143-peak GC subset. Numbers in brackets represent average proportions.

	Model (3.1) (None)	Model (3.2) (Mouse)	Model (3.3) (Mouse, C8)	Model (3.4) (Mouse, aliquot, C8)
Theoretical	232.8 (0.216)	168.0 (0.156)	110.4 (0.102)	110.4 (0.102)
Achieved	236.7 (0.219)	194.6 (0.180)	132.3 (0.123)	134.4 (0.124)

chip, and model (3.1) gives identical fitted probabilities for all 72 observations sharing the same group and chip.

As an example, consider peak 3493 m/z. The missingness indicators for the samples from group 1, mouse 1, aliquot 1, and C8 batch 1 are $r_{11111} = 0$, , $r_{11112} = 1$, and $r_{11113} = 1$. The minimum number of misclassifications that model (3.3) could make on this triplet is one, which would occur if $p_{1111n} > 0.5$. (If $p_{1111n} < 0.5$, then there would be two misclassifications, causing the achieved rate to be greater than the theoretically minimal rate.) The missingness indicators for another triplet of samples, such as those from group 1, mouse 1, aliquot 1, and C8 batch 2, are $r_{11121} = 0$, , $r_{11122} = 0$, and $r_{11123} = 0$, and the minimum number of misclassifications for the triplet is zero, occurring if $p_{1112n} < 0.5$. Of the 360 triplets of observations from each C8 batch, 88 triplets will necessarily have at least one observation misclassified, meaning that the theoretically minimum misclassification rate of model (3.3) on peak 3493 m/z is $88/1080 = 0.0815$.

Table 3.3 lists the theoretical and achieved misclassifications by the four models for the 143-peak subset. For the initial model, there is little difference between the theoretical and achieved numbers of misclassifications. For model (3.3), the theoretical minimum misclassification rate is lowest (and equal to that of model (3.4)) and the achieved rate is also the lowest out of the models considered. This supports model (3.3) over model (3.4) as the most suitable model for the GC dataset.

3.4 Final missingness model

Based on the results of the cross-validation and the other considerations discussed above, model (3.3) was used as the model for the vectors of missingness indicators of the GC dataset.

3.4.1 The issue of separation of data and a Bayesian solution

Sensible parameter estimates were not obtained for peaks outside of the 143-peak subset due to the issue of separation of data. For the GLMMs used to model the GC dataset missingness, separation affected the estimates of the fixed effects parameters μ, α_j , and β_ℓ . Separation occurred either when all missingness indicators for the 216 observations

within a genotype group, or for the 360 observations on one MALDI chip, took identical values. When either of these conditions were met, then the maximum likelihood estimate of at least one parameter failed to exist. A deeper explanation of the issue of separation is provided in Appendix B. Sixteen peaks in the GC dataset have missingness indicators that are separated with respect to at least one of the chip or group categories. The remaining 143 peaks are what define the 143-peak subset.

There are several ways of handling the problem of separation, and the most promising solution involves penalising the likelihood function in order to ensure that the likelihood is always maximised at finite values (Heinze and Schemper, 2002; Zorn, 2005). In this thesis, the issue of separation was resolved by casting the chosen missingness model (3.3) in a Bayesian framework in which all parameters were given prior distributions, resulting in a posterior distribution function with finite maximising values even in cases of separated data. The Bayesian missingness model (3.3), with two random effects, was fitted to the 158-peak subset. This was used as the final choice of model. The Bayesian specification improved estimated model performance in terms of misclassification rates. The final model was assessed via its misclassification rate on the dataset and via simulation checks, and the checks showed that the model performed adequately.

The `bglmer` function in the R package `blme` was used for fitting GLMMs in the Bayesian framework. Models were fitted using the default `glmerControl` options for `bglmer`.

3.4.2 Obtaining the prior distributions

Recall that the missingness model (3.3) is specified as

$$R_{jk\ell mn} \sim \text{Bern}(p_{jk\ell mn}), \quad p_{jk\ell mn} = \frac{e^{\eta_{jk\ell mn}}}{1 + e^{\eta_{jk\ell mn}}}, \quad \eta_{jk\ell mn} = \mu + \alpha_j + \beta_\ell + M_{jk} + C_{jk\ell m}\,,$$

with mutually independent random effects

$$M_{jk} \sim N(0,\, \sigma_M^2) \quad i.i.d., \quad C_{jk\ell m} \sim N(0,\, \sigma_C^2) \quad i.i.d.$$

The parameters of the model consist of the vector of fixed effects

$$\boldsymbol{\lambda} = (\mu\,,\, \alpha_2\,,\, \alpha_3\,,\, \alpha_4\,,\, \alpha_5\,,\, \beta_2\,,\, \beta_3)^T$$

and the two variance components σ_M^2 and σ_C^2.

A multivariate normal distribution was assumed for the fixed effects vector and gamma distributions were assumed for the variance components. The hyperparameters of these prior distributions were estimated using the non-Bayesian model (3.3). This model, fitted to the 143-peak subset, produced sets of 143 estimates of each parameter. It is from these sets of estimates that the hyperparameters of the prior distributions were derived.

The set of fixed effects from all of the non-Bayesian models may be written as a 143×7 matrix F, where rows of the matrix are the parameter estimates of each model. In the Bayesian models, the fixed effects vector was given the prior distribution

$$\boldsymbol{\lambda} \sim N_7(\mathbf{0}, \Sigma),$$

where

$$\Sigma = \frac{1}{N-1}(F - \bar{F})^T(F - \bar{F}).$$

$N = 143$ is the number of rows of F and

$$\bar{F} = \begin{bmatrix} \bar{\mu} & \bar{\alpha}_2 & \bar{\alpha}_3 & \bar{\alpha}_4 & \bar{\alpha}_5 & \bar{\beta}_2 & \bar{\beta}_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{\mu} & \bar{\alpha}_2 & \bar{\alpha}_3 & \bar{\alpha}_4 & \bar{\alpha}_5 & \bar{\beta}_2 & \bar{\beta}_3 \end{bmatrix},$$

where $\bar{\mu}$ up to $\bar{\beta}_3$ are the sample mean values of each fixed effect's parameter estimate over the 143 models. The matrix Σ is returned by the `var` function in R applied to the matrix F. Table 3.4 shows the matrix Σ.

The prior mean of the fixed effects was forced to be equal to $\mathbf{0}$ due to a limitation of the `blme` prior objects used by the `bglmer` function.

The two variance components were each given gamma distribution priors on the variance scale. The estimated variances for each of the two random effects in the 143 `glmer` models were recorded, and estimates equal to zero were discarded. For each random effect, a gamma distribution was fitted to the estimates using the `fitdistr` function from the R package `MASS` in order to obtain estimates of the shape and rate parameters. Gamma distributions with the estimated shape and rate parameters were used as the prior distributions. Table 3.5 gives the shape and rate parameters of the gamma priors for the variance components.

3.4.3 Theoretically minimal versus achieved misclassifications

The average minimum possible misclassification proportion across the 158-peak subset with model (3.3) is 0.0945, corresponding to a rate of roughly 102 observations misclassified per peak. This theoretical minimum is lower than that from the 143-peak subset under the same model, because the additional peaks in the 158-peak subset contain peaks with either extremely low or extremely high amounts of missingness, which tend to produce low numbers of misclassifications. The average misclassification proportion of the Bayesian model over the set of peaks is 0.116, or an average of 125.3 observations misclassified per peak. In light of this, the performance of the Bayesian model (3.3) appears good.

Figure 3.4 displays the minimal and actual misclassification rates for the model on each peak. As the proportion of missing observations in a peak approaches 0.5, there is

Table 3.4: Entries of estimated variance-covariance matrix Σ for fixed effects prior distribution used for the Bayesian model (3.3) fitted to the 158-peak subset.

	μ	α_2	α_3	α_4	α_5	β_2	β_3
μ	12.507	−0.627	−1.465	−2.690	−4.337	0.788	1.470
α_2	−0.627	2.948	0.708	1.937	0.809	−0.259	−0.150
α_3	−1.465	0.708	2.399	1.725	2.785	−0.085	−0.289
α_4	−2.690	1.937	1.725	3.971	2.450	−0.555	−0.827
α_5	−4.337	0.809	2.785	2.450	7.336	−0.573	−0.609
β_2	0.788	−0.259	−0.085	−0.555	−0.573	1.589	1.799
β_3	1.470	−0.150	−0.289	−0.827	−0.609	1.799	2.550

Table 3.5: Estimated hyperparameters for random effect variance component distributions used for the Bayesian model (3.3) fitted to the 158-peak subset.

	Shape	Rate
Mouse variance σ^2_M	0.512	0.262
C8 batch variance σ^2_C	0.464	0.082

a trend of increased difficulty in making classifications. This trend is reflected in both the increased minimum rates and the increased divergence between theoretically-minimal and achieved rates. For the vast majority of peaks, the ratio between the minimal and achieved rates is within the range from 1 to 1.6 regardless of the number of missing observations. The exceptions are the peaks at 6821 and 16030 m/z, where the ratios are 2 and 1.6 respectively, and the numbers of missing observations in these peaks are 8 and 24 respectively.

The average difference in the theoretically minimal and the achieved number of misclassifications is 23.2. The maximum difference is 100, and this occurs for the peak at 5675 m/z.

The two outlying peaks with both low misclassification rates and moderate numbers of missing observations are at 8302 m/z and 8337 m/z, respectively with 27 and 11 misclassifications by the model, 26 and 11 misclassifications in the best possible case, and 463 and 481 missing observations. The amounts of missing observations in these peaks differ almost entirely between groups rather than other factors.

3.4.4 Simulation checks of model fit

In a wide variety of situations, two simulation-based methods of assessing statistical models and model fits are available. The first method is to use simulations involving con-

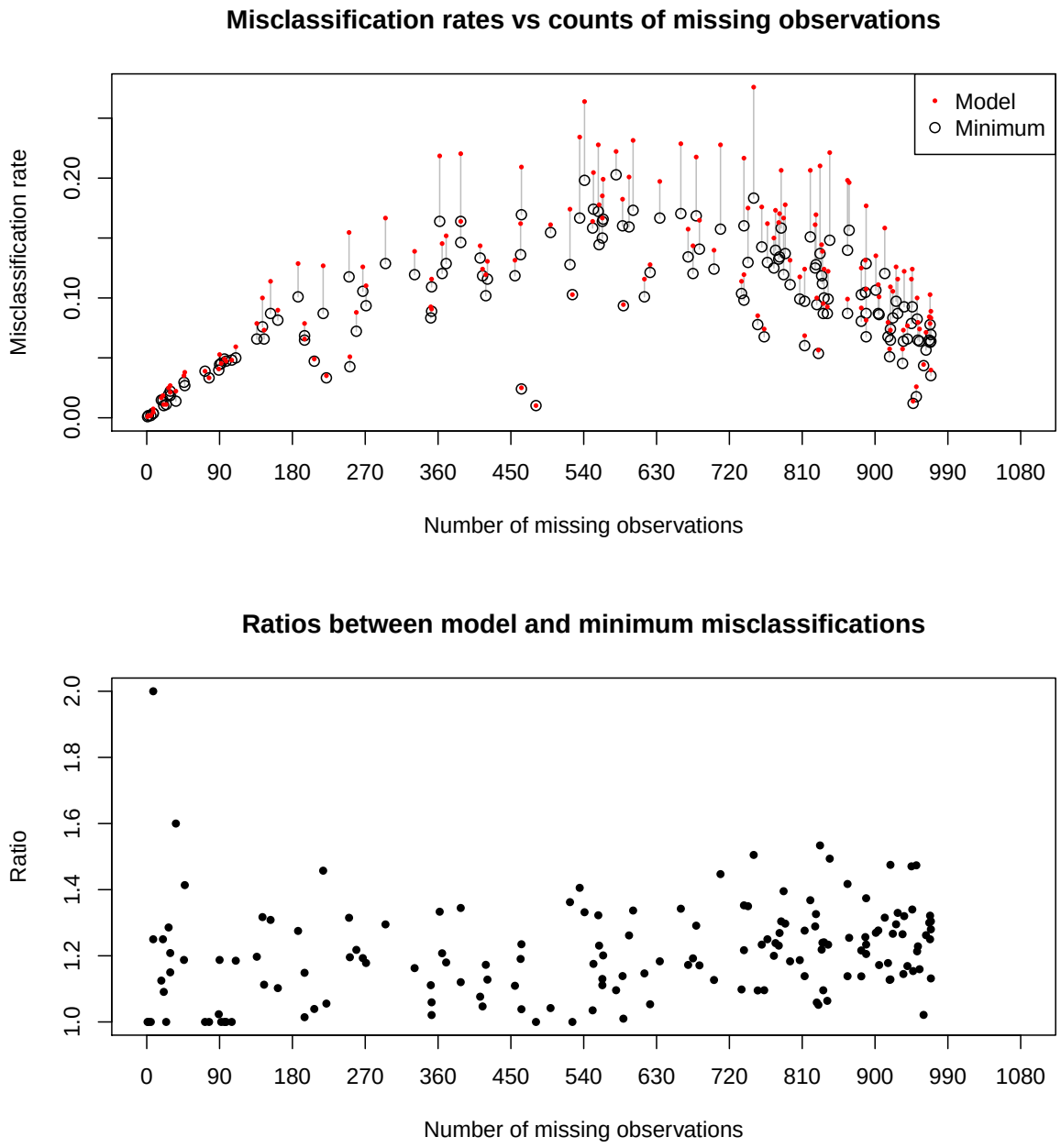

Figure 3.4: Comparison of misclassification rates of model (3.3) to the theoretical minimum, displaying the impact of the proportion of missing observations on both the minimal and actual misclassification rates. Points on the upper plot represent individual peaks, with minimal rates represented by black circles and model rates represented by red points. The x-axis represents the number of missing elements for any particular peak and the y-axis again gives the misclassification rate as a proportion. Grey lines connect the two points that come from a single peak to enable visual comparison of the difference in performance. The lower plot displays the model rates divided by the minimal rates for each peak.

Table 3.6: Means and sample standard deviations of differences between estimated and true parameter values over 1000 replications of constructed data procedure for the Bayesian missingness model (3.3).

Parameter	Mean	S.D.
μ	-0.03	0.36
α_2	-0.18	0.53
α_3	0.03	0.45
α_4	0.19	0.49
α_5	0.18	0.62
β_2	0.05	0.31
β_3	0.05	0.28
σ^2_M	-0.18	0.23
σ^2_C	-0.27	0.34

structed data sets to assess and validate the computational model fitting procedures and the second is to compare model predictions of data to the original dataset (Gelman and Hill, 2009). In the first method, data are generated according to the model using known parameter values, and a comparison of the estimated parameter values with the known values provides an evaluation of the statistical model fitting techniques. In the second method of assessment, the fitted model's parameter estimates are used to generate a new dataset which may be compared directly to the original dataset. Discrepancies between the datasets imply that some quality of the data-generating process is not captured by the model under consideration.

The following is a brief summary of results from the simulation checks, full details of which may be found in Appendix E.1. Table 3.6 summarises the differences between the true and estimated parameter values for 1000 replications of the first simulation method. The results of these simulation checks are that estimates of group effect parameters α_j are slightly shrunk towards zero, that chip effect parameters β_ℓ are estimated fairly accurately, and that estimates of random effect variance components are biased downward.

For the second method, the generated dataset (of missingness indicators) closely resembles the original dataset's missingness pattern.

Overall, model (3.3) fitted using the Bayesian framework performs adequately and appears to be suitable for the GC dataset.

Table 3.7: Summary of means and standard deviations of fixed effect parameter estimates for the 152-peak subset.

Parameter	Mean	S.D.
μ	-0.043	2.964
α_2	0.07	1.662
α_3	0.01	1.295
α_4	0.123	1.832
α_5	0.56	2.081
β_2	-0.268	1.114
β_3	-0.291	1.437

3.5 Results from the Bayesian model

Results for individual parameters are presented first. A contrast between the cancer groups and the non-cancer groups was used to determine a set of peaks of interest. Because of the association of missingness with the peak intensity, these peaks are of interest as candidate biomarkers. However, this interest is in a secondary sense, because the missingness pattern is an imperfect proxy for the actual concentration values that are used in disease diagnosis.

Despite the Bayesian priors that were put on the parameters, some peaks, most of which exhibit separation of data, did not yield sensible parameter estimates. These are the peaks at 7490, 7806, 7917, 9239, and 9319 m/z. The peak at 9712 m/z produced many near-zero entries in the estimated variance-covariance matrix. No model was fitted to the peak at 4358 m/z, as that peak had zero missing elements. These seven peaks are therefore excluded from the analysis, leaving a *152-peak subset* of peaks under consideration.

3.5.1 Individual parameters

Figure 3.5 displays a set of histograms for the parameter estimates over the 152-peak subset. The first seven plots in the figure contain the fixed effect parameter estimates. Table 3.7 summarises the means and standard deviations of the fixed effect parameter estimates. The parameter estimates tend to be centred close to zero. The distributions of estimates appear roughly symmetric, except for the distribution of the estimates of μ, which is negatively skewed. This skewness is due to the fact that peaks with missingness fractions between 0 and 0.1 are present in the GC dataset, while peaks with missingness fractions between 0.9 and 1 are not present, resulting in the existence of more low estimates of μ than high estimates.

These parameters have a linear effect on the logit scale, and therefore nonlinear effects on the probability scale, meaning that care must be taken in the interpretation of the parameters. In general, the effect of a parameter in terms of changing the probability of

Table 3.8: Peaks with the 16 most statistically significant estimates of μ.

Peak m/z	$\hat{\mu}$	S.E.	z statistic
2104	1.659	0.095	17.411
2128	1.902	0.086	22.186
5123	2.22	0.097	22.903
5189	1.993	0.085	23.307
5373	2.959	0.151	19.621
5590	3.141	0.153	20.513
5617	1.533	0.079	19.43
5637	1.765	0.074	23.956
5675	1.463	0.082	17.95
6000	1.506	0.058	25.987
6258	2.593	0.094	27.664
7087	1.786	0.099	18.071
8441	1.628	0.099	16.365
8970	2.654	0.175	15.167
9431	1.94	0.105	18.549
10255	3.005	0.147	20.436

Table 3.9: Peaks with the 16 most statistically significant estimates of α_2.

Peak m/z	$\hat{\alpha}_2$	S.E.	z statistic
2104	-1.344	0.151	-8.929
2128	-1.027	0.119	-8.662
2478	-2.779	0.287	-9.69
2504	-2.349	0.28	-8.405
4617	1.888	0.278	6.782
4993	2.541	0.25	10.161
5453	1.308	0.172	7.587
5557	0.636	0.094	6.745
6000	0.686	0.092	7.492
7566	1.109	0.108	10.246
9795	-2.111	0.24	-8.795
9821	-1.818	0.242	-7.507
10255	-1.51	0.184	-8.186
11120	2.148	0.328	6.547
16505	1.703	0.226	7.528
16525	2.193	0.325	6.742

Table 3.10: Peaks with the 16 most statistically significant estimates of α_3.

Peak m/z	$\hat{\alpha}_3$	S.E.	z statistic
2104	-1.887	0.15	-12.601
2128	-0.769	0.116	-6.619
3881	1.322	0.162	8.182
5453	2.483	0.175	14.161
5637	-1.023	0.094	-10.841
5675	-1.292	0.128	-10.07
6541	1.235	0.165	7.505
7566	-0.641	0.098	-6.543
8441	-1.562	0.153	-10.223
8607	-2.621	0.404	-6.495
8970	-2.272	0.246	-9.23
9608	1.265	0.161	7.875
10255	-1.72	0.179	-9.603
11352	-3.581	0.36	-9.952
14421	1.522	0.186	8.205
14836	-4.324	0.52	-8.315

Table 3.11: Peaks with the 16 most statistically significant estimates of α_4.

Peak m/z	$\hat{\alpha}_4$	S.E.	z statistic
4617	2.102	0.294	7.153
4866	-2.929	0.33	-8.884
5123	-1.161	0.143	-8.109
5453	2.838	0.19	14.959
5617	-0.83	0.121	-6.852
6076	3.534	0.476	7.427
6258	-1.377	0.118	-11.698
7146	0.612	0.069	8.903
7750	3.778	0.407	9.278
8228	-1.567	0.26	-6.033
8251	-1.699	0.262	-6.492
8607	-5.203	0.766	-6.794
9490	-1.708	0.261	-6.53
10230	-2.539	0.392	-6.473
10255	-2.491	0.19	-13.116
14421	1.303	0.194	6.728

Table 3.12: Peaks with the 16 most statistically significant estimates of α_5 .

Peak m/z	$\hat{\alpha}_5$	S.E.	z statistic
3881	1.824	0.178	10.227
4617	2.577	0.312	8.261
5453	2.042	0.184	11.079
5557	1.351	0.101	13.391
6258	-1.232	0.118	-10.427
6541	1.865	0.187	9.993
7146	0.859	0.07	12.225
7566	0.992	0.11	9.035
7642	2.641	0.274	9.63
7750	4.227	0.429	9.856
8441	-1.263	0.161	-7.828
8970	-2.869	0.275	-10.433
9431	1.74	0.184	9.46
10255	-2.032	0.19	-10.671
11120	3.515	0.377	9.315
16654	-4.773	0.581	-8.222

the outcome is largest when the pre-existing probability is near to 0.5 (corresponding to pre-existing log-odds being near zero) and the parameter's effect is to bring the log-odds toward the direction of zero. For example, if the μ parameter is zero, then the probability of missingness for samples from group 1 and chip 1 is 0.5. If $\alpha_2 = 1$ also, this means that the probability for samples from group 2 and chip 1 is 0.731, representing an increase of 0.231. If, additionally, $\beta_2 = 1$, samples from group 2 and chip 2 have probability of missingness 0.881, representing an increase of only 0.15. More generally, for μ parameters far from zero, the effects of the α_j and β_ℓ parameters are dampened on the probability scale.

Tables 3.8 to 3.12 display the sixteen peaks with the most statistically significant estimates of the intercept and group fixed effect parameters. The parameters β_2 and β_3 are the contrasts between chips 2 and 3 relative to that of chip 1. These parameters are not of central importance in the context of biomarker discovery. However, they are noteworthy in that their estimates' standard errors were lower than those of the group parameters α_j , even when the parameter estimates were large in absolute value. This resulted in many z statistics on the order of ± 20. From Table 3.7, the average values of the estimates were below zero, indicating that samples from chip 1 tended to be missing more frequently than on the other two chips. This appears to be in accordance with the between-chip differences visible in Figures 2.8a and 2.8b.

The μ estimates present in Table 3.8 all have positive z statistics. While several peaks

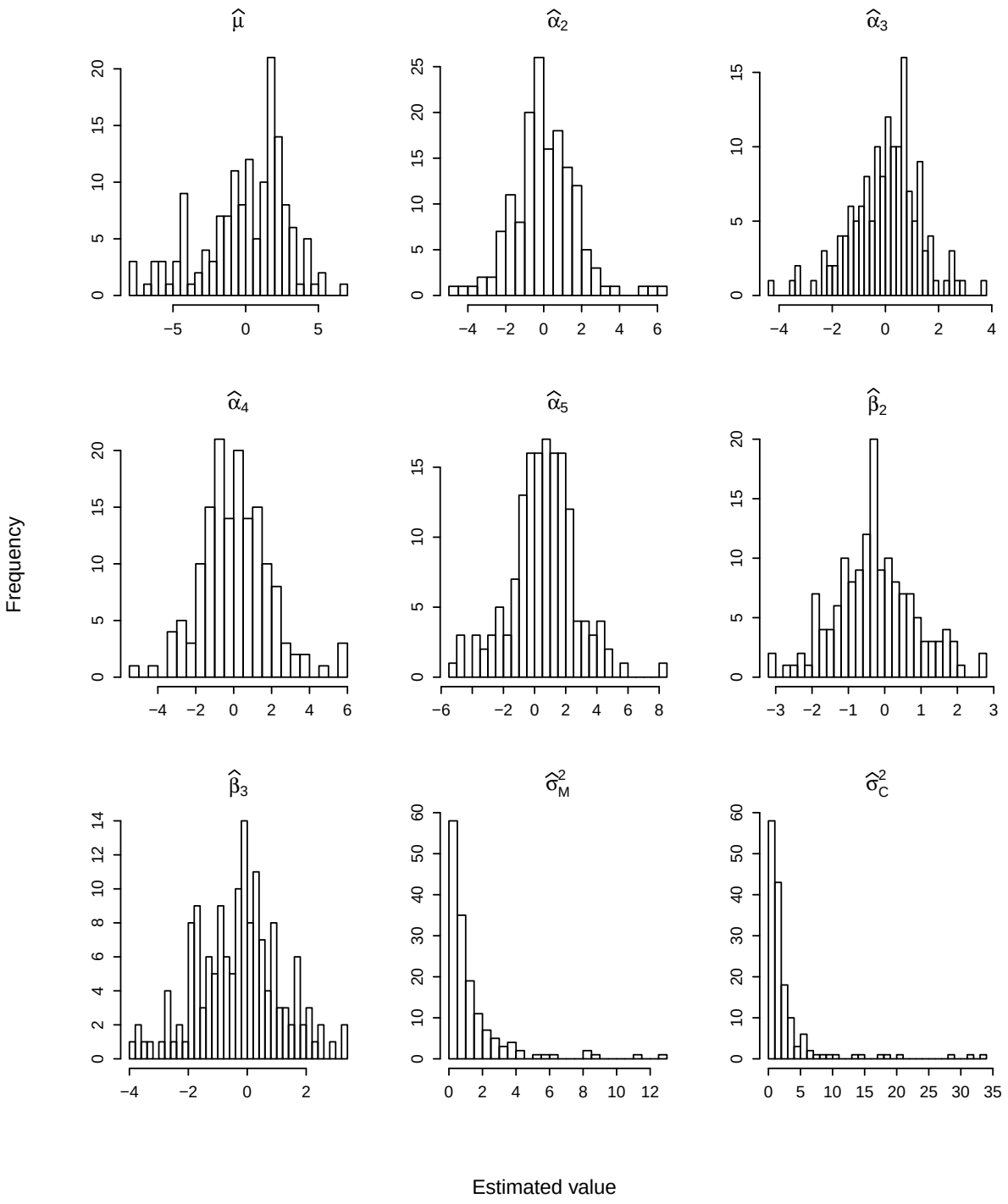

Figure 3.5: Parameter estimates for Model (3.3) fitted with `bglmer` for the 152-peak subset.

exist with z statistics above, for example, 13, there are no peaks with z statistics below -13, despite the fact that the estimates of μ have a negatively skewed distribution. The general pattern that the estimates and standard deviations of the μ parameter follow is that rather than the most extreme estimates of μ, it tends to be estimates with absolute values from 1 to 4 that accompany the largest z statistics. Such z statistics (from Table 3.8, values around 0.1 are typical) are the result of low values of the standard error, which tend to arise when the proportion of missing observations across groups and chips varies little. Peaks with low values of μ tend to exhibit large disparities in missingness proportion between group 1 and the other groups, and such disparities seem to preclude high precision in the estimates of that parameter.

The most biologically relevant parameters are α_4 and α_5 , which pertain to Tables 3.11 and 3.12. These parameters represent the difference in log-odds of missingness between the diseased FF group and, respectively, the healthy IL6 and WT groups. Peaks featuring in these tables tend to be represented in Table 3.14, which contains the set of peaks of secondary interest as biomarker candidates from the contrast analysis of Section 3.5.2. The parameter α_3 is also biologically important, as it represents the difference in the log-odds of missingness for the FFStat3 group compared to the FF group. Both of these groups of mice suffered from an inflamed gut, but the latter group did not suffer from cancer. Table 3.10, which pertains to the parameter α_3 , also has many of its peaks appearing in Table 3.14.

The last two plots in Figure 3.5 contain histograms for the variance components on the set of 152 peaks. The positive skew of these distributions arose firstly due to variances necessarily being non-negative, and secondly due to variance component estimates being low or even zero. Indeed, for 44 out of the 152 peaks, one random effect variance was estimated to be zero, effectively removing the corresponding random effect term from the right hand side of Model 3.3. Twenty-eight peaks had a zero-valued mouse variance estimate and 16 peaks had a zero-valued C8 variance estimate. There were no peaks with both variance estimates equal to zero.

The mean variance component estimates from the 152 models are 1.369 for the mouse variance σ^2_M and 2.894 for the C8 variance σ^2_C . If the means are computed without the zero-valued variances included, then they are equal to 1.678 for the mouse variance and 3.234 for the C8 variance. Likewise, the standard deviations of the sets of estimates are 1.999 and 5.208 for σ^2_M and σ^2_C , or 2.094 and 5.406 when the zero-valued estimates are not included.

3.5.2 Parameter contrast

Identification of peaks of interest using the Bayesian missingness model (3.3) was done by taking a contrast U_m between the cancer groups (groups 1, FF, and 2, FFIL6) and the non-cancer groups (groups 3, FFStat3, 4, IL6, and 5, Wildtype). The chip effect was not considered.

The contrast was formulated with respect to the group means, derived from the model parameters, rather than with respect to the model parameters directly. Suppose the mean of group 1 is denoted $\mu_1^* = \mu$ and the mean of group j is denoted $\mu_j^* = \mu + \alpha_j$ for $j = 2, \ldots, 5$. A contrast between the log-odds of the non-cancer groups and the cancer groups is

$$U_m = \frac{1}{2}(\mu_1^* + \mu_2^*) - \frac{1}{3}(\mu_3^* + \mu_4^* + \mu_5^*). \tag{3.5}$$

Expressed in terms of the original parameters, this is the linear combination

$$U_m = \frac{\alpha_2}{2} - \frac{1}{3}(\alpha_3 + \alpha_4 + \alpha_5). \tag{3.6}$$

Estimates of the contrast for each peak were obtained using the estimated fixed effects parameter values from the fitted model object, extracted using the `fixef` function from the `lme4` package (which may be used on `bglmer` models). The standard error of the contrast for each peak was calculated from the estimated variance-covariance matrix of the fixed effects, which was obtained using the `vcov` function on the fitted model object.

The contrast z statistics were defined as the estimates of the contrast divided by their standard errors. The peaks of interest were selected on the basis of statistically significant p values of the contrast z statistics. The p values were obtained using a Wald Z test for the contrast z statistic (Tuerlinckx et al., 2006) with the null hypothesis

$$H_0 : U_m = 0.$$

Within the Bayesian framework, the prior probability of the null hypothesis being true is zero, as the null hypothesis specifies the single value of 0. Null hypothesis significance testing is performed here with the caveat that the implications of the Bayesian framework are ignored—it is merely a method of obtaining peaks of interest. Appendix D.1 provides a worked example of how the contrast z statistic is obtained for the peak at 7412 m/z.

The type I error rate was set to $\alpha = 0.01$, corresponding to critical thresholds of ± 2.576 for the contrast z statistic. However, the Bayesian missingness model was fitted to 152 peaks, resulting in 152 simultaneous comparisons of the contrast. A correction for multiple comparisons was necessary as the hypotheses were investigated in parallel across peaks. The Benjamini and Hochberg (1995) procedure to control the false discovery rate was used.

The problem of making statistical inference, in terms of hypothesis testing, about parameters in a GLMM is difficult. The sampling distribution of a maximum likelihood estimate is only asymptotically normal, meaning that for insufficiently large n, the sampling distribution is not well-approximated by a normal distribution. For the missingness models fitted to the GC dataset, the value of n is 1080. However, in any given statistical model, it is not immediately clear what value of n is sufficiently large for the sampling distribution to approximate a normal distribution up to some desired degree of error.

Table 3.13: Protein peaks (m/z) of interest as biomarker candidates according to Stanford (2015).

6602	6821	7412	7806	8337	8533	8607	8831
8867	9305	12161	13648	14421	14836	16030	17458

Therefore, the p values of the contrast z statistics obtained in this chapter should not be taken as definitive. Nevertheless, extreme z statistics and contrast estimates are assumed, due to the correction for multiple comparisons, to be the result of true differences in missingness of data from cancer and non-cancer groups, and these differences imply that the groups differ in protein expression levels.

3.5.3 Peaks of interest

The set of peaks from the missingness modelling and the set from the prior modelling of intensities by Stanford (2015) have partial but not complete overlap, implying that information about group differences in protein concentration is present in both the intensity values and the missingness indicators, and that neither approach alone can recover all of the information. Table 3.13 displays the peaks of primary interest as biomarker candidates according to existing work performed by Stanford (2015). Table 3.14 displays the peaks of secondary interest as biomarker candidates on the basis of statistically significant contrast z statistics. There is some degree of overlap between the peaks from Table 3.13, and the peaks from Table 3.14. There were 16 peaks deemed interesting by Stanford (2015), and 45 peaks considered to be of secondary interest from the missingness modelling. However, 7 out of the 16 peaks from the former set are present in the latter set, more than would be expected from random chance given that there are 159 peaks in the dataset, implying that group differences in intensity and missingness probability co-occur. Of the peaks that are present in both tables, a majority exhibited missingness proportions between 18% and 43% with the exception of the peak at 8867 m/z with approximately 88% of values missing. Typically, the peaks found in both tables have observations almost entirely present for one or more groups, with high mean intensity in the observed data, and observations missing up to approximately 50% in other groups alongside a lower mean intensity in the observed data. Although this chapter considered only MAR models for the missingness, the missingness mechanism affecting the GC dataset is thought to be a NMAR mechanism due to the limited capabilities of the measurement apparatus and the data pre-processing (Stanford, 2015). Because of this, the missing observations are likely to correspond to low true intensities, possibly below the minimum intensity observed in the GC dataset.

Of the peaks present in the results of Stanford (2015) (Table 3.13) but not from the missingness modelling (Table 3.14), a majority exhibit missingness proportions below

10% with the exception of the peak at 8337 m/z with approximately 45% of observations missing. Despite the clear between-group differences in missingness present in this peak's data, the peak does not appear in Table 3.14 because there is not a large difference in average missingness *between* the cancer and non-cancer groups.

3.6 Summary

The missingness pattern of the GC dataset was modelled using GLMMs for the probability of missingness. The models were fitted on a per-peak basis. The appropriate fixed effects and random effects structures for the missingness models needed to be determined. From the data visualisations of Chapter 2, it was apparent that the group and chip effects were important and that the hierarchy of sample processing needed to be accounted for. Models with a variety of random effect structures were compared via cross-validation to estimate the out-of-sample misclassification rate. The final model chosen was the GLMM with two random effects. The model development work in this chapter provides a foundation for the specification of the missingness mechanism in the joint model.

Separation of the data with respect to the group and chip variables affected parameter estimates for 15 peaks. Separation manifests as uniformity of the missingness indicator for all observations belonging to a single group or a single chip. A Bayesian approach to model fitting provided a solution to the problem of separation by putting a prior distribution on the parameters in order to ensure that the likelihood function is always maximised at finite values. Prior distributions for the fixed effects and the random effect variance components were obtained using parameter estimates from the non-Bayesian model on a subset of the data. The Bayesian model was assessed by examining the misclassification rate for the whole dataset as well as by use of simulation methods. The model was found to be suitable for the data.

A contrast of the parameters revealed significant differences in missingness probability of observations from cancer versus non-cancer groups in a large number of peaks. A set of peaks of secondary interest as candidate biomarkers was obtained. Such peaks tended to be those with extreme estimates for parameters representing group differences in missingness probability of intensity observations.

In prior work on the GC dataset by Stanford (2015), the peak intensities in the GC dataset were modelled using LMMs and a set of peaks of primary interest as biomarker candidates was obtained. These models did not account for the NMAR nature of the data and it is reasonable to expect additional information about cancer versus non-cancer group intensity differences to lie in the pattern of missingness. Many of the peaks deemed interesting by Stanford (2015) are also present in the set obtained in this chapter. Peaks common to both sets tend to have differences in group means occurring alongside differences in missingness probability. In particular, the lower the average of the observed data within a group, the more likely it is that an observation from that group is missing.

Table 3.14: Display of peaks of secondary interest for cancer/non-cancer contrast based on missingness models[a]. Peaks are arranged in order of the contrast value. The p-values are based on a Wald Z test and are adjusted using a FDR correction for 152 simultaneous comparisons.

Peak m/z	Contrast	S.E.	z statistic	p-value
$9305^{*\dagger 4650}$	-4.013	0.798	-5.03	< 0.001
15724	-2.984	0.937	-3.186	0.001
$4650^{\dagger 9305}$	-2.583	0.514	-5.028	< 0.001
8505	-2.534	0.507	-5.001	< 0.001
12161^{*}	-2.44	0.697	-3.498	< 0.001
7750	-2.289	0.406	-5.631	< 0.001
6899	-2.113	0.39	-5.42	< 0.001
4168	-2.033	0.668	-3.046	0.002
7642	-1.811	0.348	-5.208	< 0.001
5453	-1.8	0.277	-6.507	< 0.001
6076	-1.78	0.435	-4.089	< 0.001
9821	-1.777	0.404	-4.402	< 0.001
9795	-1.775	0.331	-5.361	< 0.001
6989	-1.735	0.51	-3.402	0.001
$14421^{*\dagger 7204}$	-1.717	0.299	-5.749	< 0.001
8265	-1.706	0.538	-3.172	0.002
7087	-1.269	0.27	-4.703	< 0.001
$7204^{\dagger 14421}$	-1.239	0.269	-4.606	< 0.001
9608	-1.214	0.275	-4.417	< 0.001
6879	-1.156	0.335	-3.448	0.001
3246	-1.153	0.364	-3.17	0.002
6541	-0.909	0.271	-3.35	0.001
3881	-0.81	0.268	-3.025	0.002
9431	-0.81	0.253	-3.208	0.001
7146	-0.545	0.17	-3.204	0.001
5675	0.71	0.229	3.096	0.002
5637	0.754	0.244	3.091	0.002
5373	0.978	0.296	3.301	0.001
6258	0.994	0.225	4.426	< 0.001
8441	1.089	0.259	4.2	< 0.001
5248	1.206	0.362	3.332	0.001
6858	1.251	0.336	3.72	< 0.001
10255	1.326	0.271	4.89	< 0.001
8067	1.426	0.472	3.021	0.003
4866	1.458	0.369	3.945	< 0.001
8867^{*}	1.707	0.478	3.569	< 0.001
11352	1.883	0.4	4.71	< 0.001
8970	2.169	0.344	6.305	< 0.001
8607^{*}	2.269	0.47	4.822	< 0.001
16654	2.622	0.486	5.398	< 0.001
17976	2.769	0.552	5.016	< 0.001
7738	2.917	0.531	5.494	< 0.001
$7412^{*\dagger 14836}$	3.306	0.551	6.001	< 0.001
$14836^{*\dagger 7412}$	3.999	0.519	7.699	< 0.001

[a] An asterisk * denotes candidate biomarkers from Stanford (2015). A dagger † with an m/z value denotes peaks belonging to a pair.

The missingness mechanism expected for the GC dataset is one that is thought to depend on the observed and unobserved protein expressions due to the experimental setup and the data pre-processing. Because of this, and the fact that significant group differences in missingness probability occurred in some peaks, we may infer that the missingness indicators carry information about group differences in protein concentration useful for determining which peaks are the best biomarker candidates. The missingness models investigated in this chapter provide an indication of where, precisely, the modelling of the protein expression in the joint models may be improved via accounting for informative missingness.

Chapter 4

Joint missing/observed data models

Statistical models for the joint distribution of the intensities and the missingness indicators in the GC dataset are introduced. An initial MAR joint model that combines models for the missingness indicators and the intensities is formulated. Because the MAR joint model is not sufficient, a NMAR joint model is derived from the MAR joint model according to what is known as a selection model factorisation. Parameters in the NMAR joint model are estimated using the method of Markov chain Monte Carlo *(MCMC). In order to test and validate the MCMC method, the separate models for the missingness indicators and the intensities, as well as the MAR joint model, are fitted using MCMC. The appropriateness of the NMAR joint model to the GC dataset as well as the accuracy and precision of the parameter estimates are checked using simulation studies and standard MCMC diagnostics, and the model is found to be suitable.*
The NMAR joint model, making use of the information in the missingness indicators of the GC dataset, provides a set of m/z peaks corresponding to biomarker candidates that differs from those obtained using LMMs ignoring the missingness.

4.1 Existing models for the intensity response

The joint models for the GC dataset may be understood as extensions of existing models used for the GC dataset. Models for the missingness pattern were introduced in the previous chapter, and models for the protein expression values are provided in the prior work of Stanford (2015), which involved the use of LMMs to model the expression values in the GC dataset for the purpose of biomarker discovery. A key limitation of this work was that the modelling of the intensities assumed MAR data. We have established in Chapter 3 that this assumption is inadequate, because increasing missingness probabilities of observations of protein expression is associated with decreased expression, and missingness probabilities differs between groups. The goal of the joint modelling approach in this chapter is to account for the NMAR nature of the GC dataset in order to better estimate differences in group expression and thereby discover biomarkers. The LMM used by Stanford (2015) is introduced, and the joint model for the GC dataset will be extended from this LMM using the missingness models of Chapter 3.

Suppose that the numerical responses of peak expression are denoted $Y_{ijk\ell mn}$, where the meaning of the subscript is the same as that given in Section 3.2.2. Stanford (2015) proposed a model for the observed intensities of the form

$$y_{jk\ell mn} = \nu + \gamma_j + \delta_\ell + N_{jk} + B_{jk\ell m} + \varepsilon_{jk\ell mn}\,, \tag{4.1}$$

where the indices take the ranges $j = 1, \ldots, 5$, $k = 1, \ldots, 8$, $\ell = 1, 2, 3$, $m = 1, 2, 3$, and $n = 1, 2, 3$.

The fixed effect parameter ν represents the average intensity of samples from group 1 and chip 1. The differences between group 1 and group j are parametrised by the γ_j parameters, and the differences between chip 1 and chip ℓ are parametrised by the δ_ℓ parameters. Both sets of parameters are fixed effects and are under the reference category constraint.

The effects of mouse and C8 batch are represented by the random effects terms N_{jk} and $B_{jk\ell m}$, respectively. The effect corresponding to the aliquot level was not modelled, which meant that the nine C8 batch terms were nested directly within each mouse. The random effects follow the distributional assumptions

$$N_{jk} \sim N(0,\, \sigma_N^2) \quad i.i.d., \quad B_{jk\ell m} \sim N(0,\, \sigma_B^2) \quad i.i.d.$$

The residual variation at the level of the sample replicates is represented by the term $\varepsilon_{jk\ell mn}$, where

$$\varepsilon_{jk\ell mn} \sim N(0,\, \sigma^2) \quad i.i.d.$$

The variance components of the model are σ_N^2, σ_B^2, and σ^2.

This model may be fitted to all peaks in the GC dataset except for the peaks at 4152 and 11757 m/z, which have severe missingness.

Table 3.13 lists the 16 peaks that merited further investigation as candidate biomarkers according to Stanford (2015). These peaks occupy a range of about 68% of the m/z values in the GC dataset. However, peaks with m/z ratios between 2008 and 6602, representing slightly under half of the 159 peaks in the dataset, are not included. Pairs of peaks that are thought to correspond to singly and doubly-charged ions are not always present in the table together. The peaks at 8337, 8831, 9305, and 12161 m/z are included in Table 3.13 but their counterparts at 4168, 4415, 4560 and 6076 m/z are not.

The set of peaks discovered using the joint modelling approach in this chapter, compared to the initial set shown in Table 3.13, will reveal the extra information in the missingness pattern of GC dataset captured when accounting for the NMAR nature of the data.

4.2 Formulating the joint model

The true value of a numerical response in the GC dataset is always assumed to exist regardless of its potential obfuscation by the missingness mechanism. The numerical response vector $\boldsymbol{Y}$ and its respective vector of missingness indicators $\boldsymbol{R}$ can therefore be expressed as an observation of a pair of random variables $(\boldsymbol{Y}, \boldsymbol{R})$ in the space $\mathbb{R}^{1080} \times \{0, 1\}^{1080}$. The *joint distribution* $f(\boldsymbol{y}, \boldsymbol{r}; \theta)$ is a distribution over this space parametrised by $\theta \in \Theta$.

Recall that the index i ranges over the set of 159 peaks and the indices j, k, ℓ, m, and n, collectively, specify the $N = 1080$ samples. It is useful to express the latter subscript indices in a shorter form. Let the index $(jk\ell mn)$ correspond to the lone index s (where s ranges from 1 to 1080) according to

$$s = 360(\ell - 1) + 72(j - 1) + 9(k - 1) + 3(m - 1) + n.$$

Then the vectors of observations for a given peak may be expressed as

$$\boldsymbol{y} = (y_1 \,, \ldots, y_{1080})^T$$

and

$$\boldsymbol{r} = (r_1 \,, \ldots, r_{1080})^T.$$

It is also convenient to sort the vectors of observations such that all observed data appear first and all missing data appear afterwards. We may write

$$\boldsymbol{y} = (\boldsymbol{y}_o \,, \boldsymbol{y}_m)^T,$$

where $\boldsymbol{y}_o$ is the vector of all observed intensities of length N_o and $\boldsymbol{y}_m$ is the vector of all of the missing values of length $N_m = N - N_o$. Likewise,

$$\boldsymbol{r} = (\boldsymbol{r}_o \,, \boldsymbol{r}_m)^T = (\boldsymbol{0}_{N_o} \,, \boldsymbol{1}_{N_m})^T.$$

Models are fitted in this chapter on a per-peak basis.

4.2.1 Selection versus mixture factorisation

Two distinct approaches to modelling NMAR data are available. Both approaches involve factorising the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{R}$ into a product of the marginal distribution of one quantity and the conditional distribution of the other (Little, 1995; Little and Rubin, 2002). One choice of factorisation is the *mixture model factorisation*

$$f(\boldsymbol{y}, \boldsymbol{r}) = f(\boldsymbol{y} \mid \boldsymbol{r}) f(\boldsymbol{r}). \tag{4.2}$$

In this factorisation, the distribution of $\boldsymbol{Y}$ is a mixture of two elementary distributions, and the value of $\boldsymbol{R}$ determines which distribution for $\boldsymbol{Y}$ is chosen. The assumption of NMAR is equivalent to nonequality of the elementary distributions corresponding to missing and observed values. One advantageous feature of the mixture model formulation is that one of the terms in the factorisation is the marginal distribution of $\boldsymbol{R}$, not dependent upon $\boldsymbol{Y}$, for which a model already exists—the GLMM (3.3) developed in Chapter 3, which predicts missingness based only on categorical variables derived from sample metadata. Another positive feature is that even with NMAR data, the missingness mechanism does not necessarily have to be specified precisely for the model to produce a good fit (Little, 2008). The parameters of the elementary distribution corresponding to the observed subset of the data, $\boldsymbol{Y}_o \in \boldsymbol{Y}$, may be estimated from the data. However, by construction, there is no information available in the dataset to obtain the form of the distribution of the missing subset of the data, $\boldsymbol{Y}_m \in \boldsymbol{Y}$, while NMAR data implies that the elementary distributions differ in some way. Extra assumptions relating the elementary distributions to each other must be made to obtain the joint distribution. This is the key difficulty of applying the mixture model factorisation to the GC dataset modelling (Little and Rubin, 2002).

The *selection model factorisation* is

$$f(\boldsymbol{y}, \boldsymbol{r}) = f(\boldsymbol{r} \mid \boldsymbol{y}) f(\boldsymbol{y}). \tag{4.3}$$

This factorisation models the unconditional distribution of $\boldsymbol{Y}$, which is estimated using the subset $\boldsymbol{Y}_o$. The form of the distribution of $\boldsymbol{Y}_m$ is assumed to be the same as the distribution of the observed values $\boldsymbol{Y}_o$. The NMAR nature of the data is accounted for via the term $f(\boldsymbol{r} \mid \boldsymbol{y})$. Ignorable missingness exhibited in MAR or MCAR data is equivalent to the condition that

$$f(\boldsymbol{r} \mid \boldsymbol{y}) = f(\boldsymbol{r}).$$

Selection models are sensitive to both the specification of the marginal distribution of $\boldsymbol{Y}$ and the conditional distribution of $\boldsymbol{R}$ (Hogan and Laird, 1997), whereas mixture models are less stringent about the marginal distribution of $\boldsymbol{R}$.

Despite the positive features of the mixture model, the selection model is the more attractive approach for the joint modelling of the GC dataset. This is because the mixture

model assumes separate distributional forms for a response Y_s depending on the value of R_s. Additionally, a selection model is a natural model to use for numerical measures on natural processes affected by missingness mechanisms after the measurements are performed. Therefore, a selection model is chosen.

4.2.2 Inclusion of random effects

In both the models for the missingness and for the intensities, the hierarchical structure of the GC dataset was modelled using random effects terms. The model for the intensities used by Stanford (2015) was Equation (4.1). The random effects in this model consist of the mouse effect N_{jk} and the C8 batch effect $B_{jk\ell m}$. The model for the missingness was Equation (3.3), reproduced here:

$$R_{jk\ell mn} \sim \text{Bern}(p_{jk\ell mn}), \quad p_{jk\ell mn} = \frac{e^{\eta_{jk\ell mn}}}{1 + e^{\eta_{jk\ell mn}}}, \quad \eta_{jk\ell mn} = \mu + \alpha_j + \beta_\ell + M_{jk} + C_{jk\ell m}\,. \tag{4.4}$$

The random effects consist of the mouse effect M_{jk} and the C8 batch effect $C_{jk\ell m}$. These two models are the initial candidates for the distributions of $\boldsymbol{Y}$ and $\boldsymbol{R}$ in the selection model factorisation (4.3).

This particular choice of distributions for $\boldsymbol{Y}$ and $\boldsymbol{R}$ means that the joint model incorporates multiple random effects terms in both distributions. This is undesirable for two reasons. First, the estimated missingness probabilities $p_{jk\ell mn}$ for samples within the same C8 batch, aliquot, or mouse are correlated, and these correlations are captured by the random effects in model (4.4). However, these correlations are a proxy for underlying correlations in intensity values from those samples, because the outcome of missingness is partially a proxy for low intensity values. A direct dependence on the intensity values of the missingness probability, as would be present in a joint model suitable for NMAR data, causes the random effect terms $(M_{jk}\,,\, C_{jk\ell m})$ for the levels of the sample processing hierarchy in the distribution of $\boldsymbol{R}$ to be redundant when random effects terms $(N_{jk}\,,\, B_{jk\ell m})$ for the same levels are already present in the distribution of $\boldsymbol{Y}$. The second reason is that with a large proportion of missing values, the effective sample size is reduced and a reduction in the number of parameters in the model becomes desirable in order to avoid overfitting the data.

A refinement of the joint model is obtained by removing the random effects from the distribution of $\boldsymbol{R}$. This leads to considering, for the distribution of $\boldsymbol{R}$, Equation (3.1), reproduced here:

$$R_{jk\ell mn} \sim \text{Bern}(p_{jk\ell mn}), \quad p_{jk\ell mn} = \frac{e^{\eta_{jk\ell mn}}}{1 + e^{\eta_{jk\ell mn}}}, \quad \eta_{jk\ell mn} = \mu + \alpha_j + \beta_\ell\,. \tag{4.5}$$

This yields the *MAR joint model*.

4.2.3 The MAR joint model and a NMAR joint model

A joint model that accounts for the informative missingness in the GC dataset is obtained by modifying the MAR joint model to yield the *NMAR joint model* for the GC dataset.

The MAR joint model is specified as a combination of the LMM (4.1) and the GLM (4.5). Let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1\,,\,\boldsymbol{\theta}_2\,,\,\boldsymbol{\theta}_3\}$ where

$$\boldsymbol{\theta}_1 = \boldsymbol{\lambda} = (\mu\,,\,\alpha_2\,,\,\alpha_3\,,\,\alpha_4\,,\,\alpha_5\,,\,\beta_2\,,\,\beta_3),$$

$\boldsymbol{\theta}_2 = \{\boldsymbol{\kappa},\,\sigma^2\}$ with

$$\boldsymbol{\kappa} = (\nu\,,\,\gamma_2\,,\,\gamma_3\,,\,\gamma_4\,,\,\gamma_5\,,\,\delta_2\,,\,\delta_3),$$

and $\boldsymbol{\theta}_3 = \{\sigma_N^2\,,\,\sigma_B^2\}$. Then the likelihood of the data $(\boldsymbol{y},\,\boldsymbol{r})$ under the MAR joint model is

$$L(\boldsymbol{\theta}\,;\boldsymbol{y},\,\boldsymbol{r}) = f(\boldsymbol{r}\mid\boldsymbol{y}\,;\,\boldsymbol{\theta}_1)f(\boldsymbol{y}\mid\boldsymbol{N},\,\boldsymbol{B}\,;\,\boldsymbol{\theta}_2)f(\boldsymbol{N},\,\boldsymbol{B}\,;\boldsymbol{\theta}_3) \tag{4.6}$$

where $\boldsymbol{N}$ and $\boldsymbol{B}$ are independent random effects vectors of length 40 and 360 respectively representing the variation in intensity at the levels of mouse and C8 batch. The vectors are expressed as

$$\boldsymbol{N} = (N_{1,1}\,,\,N_{1,2}\,,\,\ldots\,,\,N_{5,8})^T$$

and

$$\boldsymbol{B} = (B_{1,1,1,1}\,,\,B_{1,1,1,2}\,,\,\ldots\,,\,B_{5,8,3,3})^T$$

as in the LMM (4.1).

The first term in Equation (4.6) is $f(\boldsymbol{r}\mid\boldsymbol{y}\,;\,\boldsymbol{\theta}_1) = f(\boldsymbol{r}\,;\,\boldsymbol{\theta}_1)$ which follows the GLM (4.5). This is equal to

$$f(\boldsymbol{r}\,;\,\boldsymbol{\theta}_1) = \prod_{s=1}^{N}\frac{e^{r_s\eta_s}}{1+e^{\eta_s}}.$$

We may define the set I_o as the set of the s-indices for which $r_s = 0$, and the set I_m as the set of the s-indices for which $r_s = 1$. Then we may write

$$f(\boldsymbol{r}\,;\,\boldsymbol{\theta}_1) = \prod_{s\in I_o}\frac{1}{1+e^{\eta_s}}\cdot\prod_{s\in I_m}\frac{e^{\eta_s}}{1+e^{\eta_s}}$$

where the linear predictor η_s expands to

$$\eta_s = \eta_{jk\ell mn} = \mu + \alpha_j + \beta_\ell$$

in accordance with the formulation of subscript indices described in Section 4.2.

We may additionally write

$$f(\boldsymbol{r}\,;\,\boldsymbol{\theta}_1) = f_{1,o}(\boldsymbol{r}_o\,;\,\boldsymbol{\theta}_1)f_{1,m}(\boldsymbol{r}_m\,;\,\boldsymbol{\theta}_1)$$

by setting

$$f_{1,\,o}(\boldsymbol{r}_o\,;\,\boldsymbol{\theta}_1) = \prod_{s\in I_o}\frac{1}{1+e^{\eta_s}}, \quad f_{1,\,m}(\boldsymbol{r}_m\,;\,\boldsymbol{\theta}_1) = \prod_{s\in I_m}\frac{e^{\eta_s}}{1+e^{\eta_s}}.$$

The second term in Equation (4.6) is $f(\boldsymbol{y}\mid \boldsymbol{N},\,\boldsymbol{B}\,;\,\boldsymbol{\theta}_2)$ which follows the LMM (4.1). This is equal to

$$\begin{aligned} f(\boldsymbol{y}\mid \boldsymbol{N},\,\boldsymbol{B}\,;\boldsymbol{\theta}_2) &= \prod_{s=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp(\text{-}(y_s-\zeta_s)^2/2\sigma^2),\\ &= (2\pi\sigma^2)^{\text{-}N/2}\exp\left(\text{-}\sum_{s=1}^{N}(y_s-\zeta_s)^2/2\sigma^2\right),\\ &= (2\pi\sigma^2)^{\text{-}N_o/2}\exp\left(\text{-}\sum_{s\in I_o}(y_s-\zeta_s)^2/2\sigma^2\right)\\ &\quad\cdot(2\pi\sigma^2)^{\text{-}N_m/2}\exp\left(\text{-}\sum_{s\in I_m}(y_s-\zeta_s)^2/2\sigma^2\right),\\ &= (2\pi\sigma^2)^{\text{-}N_o/2}\exp\left(\text{-}\|\boldsymbol{y}_o-\boldsymbol{\zeta}_o\|^2/2\sigma^2\right)\\ &\quad\cdot(2\pi\sigma^2)^{\text{-}N_m/2}\exp\left(\text{-}\|\boldsymbol{y}_m-\boldsymbol{\zeta}_m\|^2/2\sigma^2\right), \end{aligned} \tag{4.7}$$

where the linear predictor ζ_s expands to

$$\zeta_s = \zeta_{jk\ell mn} = \nu+\gamma_j+\delta_\ell+N_{jk}+B_{jk\ell m}\,.$$

The terms $\boldsymbol{\zeta}_o$ and $\boldsymbol{\zeta}_m$ are vectors of length N_o and N_m whose elements are ζ_s for $s\in I_o$ and $s\in I_m$ respectively.

Equation (4.7) may additionally be written as

$$f(\boldsymbol{y}\mid \boldsymbol{N},\,\boldsymbol{B}\,;\boldsymbol{\theta}_2) = f_{2,\,o}(\boldsymbol{y}_o\mid \boldsymbol{N},\,\boldsymbol{B}\,;\boldsymbol{\theta}_2)f_{2,\,m}(\boldsymbol{y}_m\mid \boldsymbol{N},\,\boldsymbol{B}\,;\boldsymbol{\theta}_2),$$

where

$$\begin{aligned} f_{2,\,o}(\boldsymbol{y}_o\mid \boldsymbol{N},\,\boldsymbol{B}\,;\boldsymbol{\theta}_2) &= (2\pi\sigma^2)^{\text{-}N_o/2}\exp\left(\text{-}\|\boldsymbol{y}_o-\boldsymbol{\zeta}_o\|^2/2\sigma^2\right),\\ f_{2,\,m}(\boldsymbol{y}_m\mid \boldsymbol{N},\,\boldsymbol{B}\,;\boldsymbol{\theta}_2) &= (2\pi\sigma^2)^{\text{-}N_m/2}\exp\left(\text{-}\|\boldsymbol{y}_m-\boldsymbol{\zeta}_m\|^2/2\sigma^2\right). \end{aligned}$$

The third and final term in Equation (4.6) is $f(\boldsymbol{N},\,\boldsymbol{B}\,;\boldsymbol{\theta}_3)$ which, again, follows the LMM (4.1) specifically in the sense of encoding the same distributional assumptions on the random effects vectors as that model. This is equal to

$$f(\boldsymbol{N},\,\boldsymbol{B}\,;\boldsymbol{\theta}_3) = (2\pi\sigma_N^2)^{\text{-}40/2}\exp\left(\text{-}\|\boldsymbol{N}\|^2/2\sigma_N^2\right)(2\pi\sigma_B^2)^{\text{-}360/2}\exp\left(\text{-}\|\boldsymbol{B}\|^2/2\sigma_B^2\right),$$

which may be written as

$$f(\boldsymbol{N},\,\boldsymbol{B}\,;\boldsymbol{\theta}_3) = f_3(\boldsymbol{N}\,;\boldsymbol{\theta}_3)f_4(\boldsymbol{B}\,;\boldsymbol{\theta}_3),$$

where

$$f_3(\boldsymbol{N}\,;\boldsymbol{\theta}_3) = (2\pi\sigma_N^2)^{\text{-}40/2}\exp\left(\text{-}\|\boldsymbol{N}\|^2/2\sigma_N^2\right),$$

$$f_4(\boldsymbol{B}\,;\boldsymbol{\theta}_3) = (2\pi\sigma_B^2)^{\text{-}360/2}\exp\left(\text{-}\|\boldsymbol{B}\|^2/2\sigma_B^2\right).$$

The joint distribution of ($\boldsymbol{Y}$, $\boldsymbol{R}$) may then be expressed as

$$f(\boldsymbol{y},\,\boldsymbol{r}\,;\,\boldsymbol{\theta}) = f_{1,\,o}\cdot f_{1,\,m}\cdot f_{2,\,o}\cdot f_{2,\,m}\cdot f_3\cdot f_4\,, \tag{4.8}$$

with the parameters in each density function suppressed for clarity.

The NMAR joint model is derived by altering the MAR joint model (4.6) such that $\boldsymbol{\theta}_1$ consists of $\{\boldsymbol{\lambda},\,\omega\}$, and that $f(\boldsymbol{r}\mid\boldsymbol{y}\,;\,\boldsymbol{\theta}_1)$, the distribution of $\boldsymbol{R}$, depends explicitly on the observed values of $\boldsymbol{Y}$ through the new parameter ω. The likelihood of the data under the NMAR joint model is

$$L(\boldsymbol{\theta}\,;\boldsymbol{y},\,\boldsymbol{r}) = f(\boldsymbol{r}\mid\boldsymbol{y}\,;\,\boldsymbol{\theta}_1)f(\boldsymbol{y}\mid\boldsymbol{N},\,\boldsymbol{B}\,;\,\boldsymbol{\theta}_2)f(\boldsymbol{N},\,\boldsymbol{B}\,;\boldsymbol{\theta}_3). \tag{4.9}$$

The first term in Equation (4.9) is $f(\boldsymbol{r}\mid\boldsymbol{y}\,;\,\boldsymbol{\theta}_1)$ which is written

$$\begin{aligned} f(\boldsymbol{r}\mid\boldsymbol{y}\,;\,\boldsymbol{\theta}_1) &= \prod_{s=1}^{N}\frac{e^{r_s\eta_s}}{1+e^{\eta_s}}, \\ &= \prod_{s\in I_o}\frac{1}{1+e^{\eta_s}}\cdot\prod_{s\in I_m}\frac{e^{\eta_s}}{1+e^{\eta_s}} \end{aligned}$$

where the linear predictor η_s expands to

$$\eta_s = \eta_{jk\ell mn} = \mu+\alpha_j+\beta_\ell+\omega y_s\,.$$

The dependence of the distribution of $\boldsymbol{R}$ on $\boldsymbol{Y}$ is captured in the additional term ωy_s in the linear predictor η_s. The parameter ω represents the strength of the effect of the expression value y_s on the log-odds of probability of missingness of y_s. If $\omega = 0$, then the NMAR joint model reduces to the MAR joint model.

We may additionally write

$$f(\boldsymbol{r}\mid\boldsymbol{y}\,;\,\boldsymbol{\theta}_1) = f_{1,\,o}(\boldsymbol{r}_o\mid\boldsymbol{y}_o\,;\,\boldsymbol{\theta}_1)f_{1,\,m}(\boldsymbol{r}_m\mid\boldsymbol{y}_m\,;\,\boldsymbol{\theta}_1)$$

by setting

$$f_{1,\,o}(\boldsymbol{r}_o\mid\boldsymbol{y}_o\,;\,\boldsymbol{\theta}_1) = \prod_{s\in I_o}\frac{1}{1+e^{\eta_s}},\quad f_{1,\,m}(\boldsymbol{r}_m\mid\boldsymbol{y}_m\,;\,\boldsymbol{\theta}_1) = \prod_{s\in I_m}\frac{e^{\eta_s}}{1+e^{\eta_s}}.$$

The second and third terms in Equation (4.9) are the same as in Equation (4.6).

4.3 Likelihood inference

Inference with respect to the joint models is concerned with the parameters within $\boldsymbol{\theta}$, which are the fixed effects $\boldsymbol{\lambda}$ relating to the conditional probability of missingness, the parameter ω controlling the strength of the dependence of $\boldsymbol{R}$ on $\boldsymbol{Y}$, the fixed effects $\boldsymbol{\kappa}$ that determine the mean values of the observed *and* unobserved intensities, the random effects variance components σ_N^2 and σ_B^2 relating to the correlations in intensity readings due to the hierarchical processing of samples, and finally the residual error variance σ^2 between replicate sample observations. Of these parameters, the γ_j parameters in $\boldsymbol{\kappa}$, which represent group differences in intensity, are of greatest relevance in discovering peaks of interest as biomarker candidates. Neither the unobservable values of the random effects vector elements in $\boldsymbol{N}$ and $\boldsymbol{B}$ nor the unobserved values $\boldsymbol{y}_m$ are of primary concern.

Estimates of the parameters in the joint models are based on the likelihood function of the data given the model. Likelihood inference is made easier when the likelihood function may be simplified by, for example, marginalising or profiling the likelihood in order to estimate a subset of parameters. The MAR joint model is amenable to such a manipulation. The NMAR joint model is not, necessitating an alternate likelihood inference procedure.

4.3.1 MAR model marginal likelihood

The method of estimating parameters in the MAR joint model is to derive a marginal likelihood function from Equation (4.8) by integrating over the distributions of the unobservable parts of the model, these parts being the terms $\boldsymbol{N}$, $\boldsymbol{B}$, and $\boldsymbol{y}_m$. For the MAR joint model (4.6), the marginal likelihood is written

$$L(\boldsymbol{\theta}\,;\,\boldsymbol{y},\,\boldsymbol{r}) = \int\int\int f_{1,o}\cdot f_{1,m}\cdot f_{2,o}\cdot f_{2,m}\cdot f_3\cdot f_4\,\mathrm{d}\,\boldsymbol{y}_m\,\mathrm{d}\,\boldsymbol{N}\,\mathrm{d}\,\boldsymbol{B}. \tag{4.10}$$

This integral may be simplified by noting that $f_{1,o}$ and $f_{1,m}$ are free of all unobservable terms, and that $f_{2,o}$, f_3, and f_4 are all free of $\boldsymbol{y}_m$. Hence, Equation (4.10) becomes

$$L(\boldsymbol{\theta}\,;\,\boldsymbol{y},\,\boldsymbol{r}) = f_{1,o}\cdot f_{1,m}\int\int\left(\int f_{2,m}\,\mathrm{d}\,\boldsymbol{y}_m\right) f_{2,o}\cdot f_3\cdot f_4\,\mathrm{d}\,\boldsymbol{N}\,\mathrm{d}\,\boldsymbol{B}. \tag{4.11}$$

The marginal likelihood (4.11) is a function of $\boldsymbol{\theta}_1 = \boldsymbol{\lambda}$, $\boldsymbol{\theta}_2 = \{\boldsymbol{\kappa},\,\sigma^2\}$, and $\boldsymbol{\theta}_3 = \{\sigma_N^2\,,\,\sigma_B^2\}$. The term $f_{1,o}\cdot f_{1,m}$ involves $\boldsymbol{\theta}_1$ only, whereas the term

$$\int\int f_{2,o}\cdot f_3\cdot f_4\,\mathrm{d}\,\boldsymbol{N}\,\mathrm{d}\,\boldsymbol{B}$$

involves $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ only. The log-marginal likelihood obtained by taking the logarithm of Equation (4.11) is

$$\ell(\boldsymbol{\theta}\,;\,\boldsymbol{y},\,\boldsymbol{r}) = \ln(f_{1,o}\cdot f_{1,m}) + \ln\left(\int\int\left(\int f_{2,m}\,\mathrm{d}\,\boldsymbol{y}_m\right) f_{2,o}\cdot f_3\cdot f_4\,\mathrm{d}\,\boldsymbol{N}\,\mathrm{d}\,\boldsymbol{B}\right).$$

The problem of maximising the likelihood therefore simplifies to the problem of independently maximising the former term $f_{1,o} \cdot f_{1,m}$ over $\boldsymbol{\theta}_1$ and maximising the latter term

$$\int \int \left(\int f_{2,m} \, \mathrm{d}\, \boldsymbol{y}_m \right) f_{2,o} \cdot f_3 \cdot f_4 \, \mathrm{d}\, \boldsymbol{N} \, \mathrm{d}\, \boldsymbol{B} \tag{4.12}$$

over $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$.

The maximisation of the marginal likelihood over $\boldsymbol{\theta}_1 = \boldsymbol{\lambda}$ concerns only the term $f_{1,o} \cdot f_{1,m}$, and consequently reduces to the problem of estimating parameters in a binary logistic regression model (Firth, 1991).

We turn our attention to the latter term (4.12), which concerns the parameters in $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$. The innermost integral in Equation (4.11) is

$$\int f_{2,m} \, \mathrm{d}\, \boldsymbol{y}_m = \int_{\mathbb{R}^{N_m}} (2\pi\sigma^2)^{-N_m/2} \exp\left(-\|\boldsymbol{y}_m - \boldsymbol{\zeta}_m\|^2 / 2\sigma^2\right) \, \mathrm{d}\, \boldsymbol{y}_m \, .$$

This is none other than the integral of a multivariate normal distribution on $\mathbb{R}^{N_m}$ with mean vector $\boldsymbol{\zeta}_m$ and variance-covariance matrix $\sigma^2 I$. Hence, the integral is equal to 1 and the latter term becomes

$$\begin{aligned}
\int \int f_{2,o} \cdot f_3 \cdot f_4 \, \mathrm{d}\, \boldsymbol{N} \, \mathrm{d}\, \boldsymbol{B} \;\; &= \;\; \int_{\mathbb{R}^{360}} \int_{\mathbb{R}^{40}} (2\pi\sigma^2)^{-N_o/2} \exp\left(-\|\boldsymbol{y}_o - \boldsymbol{\zeta}_o\|^2 / 2\sigma^2\right) (2\pi\sigma_N^2)^{-40/2} \\
&\quad \cdot \exp\left(-\|\boldsymbol{N}\|^2 / 2\sigma_N^2\right) (2\pi\sigma_B^2)^{-360/2} \exp\left(-\|\boldsymbol{B}\|^2 / 2\sigma_B^2\right) \, \mathrm{d}\, \boldsymbol{N} \, \mathrm{d}\, \boldsymbol{B}, \\
&= \;\; (2\pi\sigma^2)^{-N_o/2} (2\pi\sigma_N^2)^{-40/2} (2\pi\sigma_B^2)^{-360/2} \\
&\quad \cdot \int_{\mathbb{R}^{360}} \int_{\mathbb{R}^{40}} \exp\left(-\|\boldsymbol{y}_o - \boldsymbol{\zeta}_o\|^2 / 2\sigma^2\right) \exp\left(-\|\boldsymbol{N}\|^2 / 2\sigma_N^2\right) \\
&\quad \cdot \exp\left(-\|\boldsymbol{B}\|^2 / 2\sigma_B^2\right) \, \mathrm{d}\, \boldsymbol{N} \, \mathrm{d}\, \boldsymbol{B}.
\end{aligned}$$

This integral is the marginal likelihood $L(\boldsymbol{\theta}\,; \boldsymbol{y}_o)$ for a LMM expressible as

$$(\boldsymbol{Y}_o \mid \boldsymbol{b}) \sim N_{N_o}(\boldsymbol{\zeta}_o \,, \, \sigma^2 I)$$

with random effects vector $\boldsymbol{b} = (\boldsymbol{N}, \, \boldsymbol{B})^T$ distributed as

$$\boldsymbol{b} \sim N_{400}(\boldsymbol{0}, \, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma_N^2 I_{40} & 0 \\ 0 & \sigma_B^2 I_{360} \end{bmatrix},$$

where the marginalisation is with respect to the random effects vector elements in $\boldsymbol{N}$ and $\boldsymbol{B}$.

This marginal likelihood can be maximised with the *pseudo-data* approach used by the `lmer` function in R for estimating parameters in LMMs as detailed in Appendix C. This approach yields maximum likelihood estimates of $\boldsymbol{\kappa}$, σ_N^2, σ_B^2, and σ^2 profiled with respect to the random effects variance components σ_N^2 and σ_B^2, .

Through these procedures, maximum likelihood estimates of the parameters $\boldsymbol{\theta}$ in the MAR joint model may be obtained. We now turn our attention to the NMAR joint model.

4.3.2 NMAR model marginal likelihood

Maximum likelihood inference for the NMAR joint model again proceeds with deriving a marginal likelihood function by integrating out the unobservable parts of the model. The marginal likelihood is written

$$L(\boldsymbol{\theta}\,;\,\boldsymbol{y},\,\boldsymbol{r}) = \int\int\int f_{1,o}\cdot f_{1,m}\cdot f_{2,o}\cdot f_{2,m}\cdot f_3\cdot f_4\,\mathrm{d}\,\boldsymbol{y}_m\,\mathrm{d}\,\boldsymbol{N}\,\mathrm{d}\,\boldsymbol{B}, \tag{4.13}$$

this time using the terms $f_{1,o}$ and $f_{1,m}$ from model (4.9).

The difference between the MAR joint model and the NMAR joint model is that $f_{1,m}$ is now a function of the unobservable values in $\boldsymbol{y}_m$. The innermost integral in Equation (4.13) can no longer be simplified by bringing the term $f_{1,m}$ to the front, and we therefore obtain

$$L(\boldsymbol{\theta}\,;\,\boldsymbol{y},\,\boldsymbol{r}) = f_{1,o}\cdot\int\int\left(\int f_{1,m}\cdot f_{2,m}\,\mathrm{d}\,\boldsymbol{y}_m\right) f_{2,o}\cdot f_3\cdot f_4\,\mathrm{d}\,\boldsymbol{N}\,\mathrm{d}\,\boldsymbol{B}. \tag{4.14}$$

Likelihood inference is complicated by the fact that the innermost integral,

$$\int f_{1,m}\cdot f_{2,m}\,\mathrm{d}\,\boldsymbol{y}_m = \int_{\mathbb{R}^{N_m}}\left(\prod_{s\in I_m}\frac{e^{\eta_s}}{1+e^{\eta_s}}\right)(2\pi\sigma^2)^{-N_m/2}\exp\left(-\|\boldsymbol{y}_m-\boldsymbol{\zeta}_m\|^2/2\sigma^2\right)\,\mathrm{d}\,\boldsymbol{y}_m\,,$$

is no longer the integral of a probability density function, and therefore no longer equal to 1. This integral must be either evaluated explicitly or approximated. The outer integrals must then be evaluated or approximated based on the resolution of the innermost integral. The Laplace approximation provides a way forward.

Laplace approximation

Suppose that an integral may be written in the form

$$\int_{\mathbb{R}^q} e^{h(\boldsymbol{b})}\,\mathrm{d}\,\boldsymbol{b}$$

for a scalar function h of some q-vector $\boldsymbol{b}$. Laplace's method may be used to approximate the integral if the function h is smooth and if the value $\tilde{\boldsymbol{b}}$ that maximises h is known (Raudenbush et al., 2000).

The approximation is given by

$$\begin{aligned}\int_{\mathbb{R}^q} e^{h(\boldsymbol{b})}\,\mathrm{d}\,\boldsymbol{b} \;&=\; E\left[\exp\left(\sum_{t=3}^{\infty}\frac{1}{t!}((\overset{t}{\otimes}(\boldsymbol{b}-\tilde{\boldsymbol{b}})^T)h^{(t)}(\tilde{\boldsymbol{b}})(\boldsymbol{b}-\tilde{\boldsymbol{b}}))\right)\right]\\ &\quad\cdot(2\pi)^{q/2}|V|^{1/2}e^{h(\tilde{\boldsymbol{b}})},\end{aligned} \tag{4.15}$$

where

$$h^{(t)}(\tilde{\boldsymbol{b}}) = \left.\frac{\partial \operatorname{vec} h^{(t-1)}(\boldsymbol{b})}{\partial \boldsymbol{b}^T}\right|_{\tilde{\boldsymbol{b}}}$$

and $\overset{t}{\otimes}\boldsymbol{b} = \boldsymbol{b}\otimes\boldsymbol{b}\otimes\cdots\otimes\boldsymbol{b}$, with $\boldsymbol{b}$ repeated t times in the Kronecker product, and $V = (\text{-}h^{(2)}(\tilde{\boldsymbol{b}}))^{\text{-}1}$. The expectation is taken with respect to a multivariate normal distribution of $\boldsymbol{b}$ with mean $\mathbf{0}$ and variance-covariance matrix V. The vec operator acts on a $m \times n$ matrix X to return a vector

$$\operatorname{vec}(X) = (x_{11}\,,\,\ldots\,,\,x_{m1}\,,\,x_{12}\,,\,\ldots\,,\,x_{m2}\,,\,\ldots\,,\,x_{1n}\,,\,\ldots\,,\,x_{mn})^T.$$

Details of how to take the expectation are provided in Raudenbush et al. (2000).

The Laplace approximation method involves expanding a function about its mode. It is suitable for integrating functions that have most of their measure concentrated near their mode. This may be problematic in high dimensional spaces, as the majority of the measure of a probability density function may be concentrated in a so-called *typical set* which is far from the mode (Betancourt, 2017). If a function has relatively large values far from the maximising value $\tilde{\boldsymbol{b}}$, the Laplace approximation becomes less accurate, and this is a second way the application of this method may fail. Another possible point of failure is that $\tilde{\boldsymbol{b}}$ itself may be difficult to obtain.

In theory, then, two repeated applications of the Laplace approximation may be used to resolve the marginal likelihood (4.14) by first computing the approximation for the inner integral

$$\int f_{1,\,m}\cdot f_{2,\,m}\,\mathrm{d}\,\boldsymbol{y}_m$$

to obtain some function $H(\boldsymbol{N},\,\boldsymbol{B})$, and then computing the approximation for the outer integral

$$\int\int H(\boldsymbol{N},\,\boldsymbol{B}) f_{2,\,o}\cdot f_3\cdot f_4\,\mathrm{d}\,\boldsymbol{N}\,\mathrm{d}\,\boldsymbol{B}. \tag{4.16}$$

The inner integral is equal to

$$\int f_{1,\,m}\cdot f_{2,\,m}\,\mathrm{d}\,\boldsymbol{y}_m = \int_{\mathbb{R}^{N_m}}\left(\prod_{s\in I_m}\frac{e^{\eta_s}}{1+e^{\eta_s}}\right)(2\pi\sigma^2)^{\text{-}N_m/2}\exp\left(\text{-}\|\boldsymbol{y}_m-\boldsymbol{\zeta}_m\|^2/2\sigma^2\right)\,\mathrm{d}\boldsymbol{y}_m\,.$$

To write this integral in the form of the approximation given in Equation (4.15), we let $\boldsymbol{b} = \boldsymbol{y}_m$ and then, in integrating $e^{h(\boldsymbol{y}_m)}$,

$$h(\boldsymbol{y}_m) = \sum_{s\in I_m}\eta_s - \sum_{s\in I_m}\ln(1+e^{\eta_s}) - (N_m/2)\ln(2\pi\sigma^2) - \|\boldsymbol{y}_m-\boldsymbol{\zeta}_m\|^2/2\sigma^2.$$

The first derivative of h is a N_m-vector $h^{(1)}(\boldsymbol{y}_m)$ whose entries (corresponding to $s\in I_m$) are

$$\omega - (\omega e^{\eta_s})/(1+e^{\eta_s}) - (y_s-\zeta_s)/\sigma^2.$$

The entry of $h^{(1)}(\boldsymbol{y}_m)$ corresponding to the s^{th} element of $\boldsymbol{y}$ involves only terms relating to that element. This is due to the independence between observations when conditioning on the values of the random effects vector elements $\boldsymbol{N}$ and $\boldsymbol{B}$.

The second derivative of h is a $N_m \times N_m$ diagonal matrix $h^{(2)}(\boldsymbol{y}_m)$ whose entry corresponding to the s^{th} element of $\boldsymbol{y}$ is

$$(\omega^2 e^{\eta_s})/(1+e^{\eta_s})^2 - 1/\sigma^2.$$

To use Laplace's method, it is necessary to find the maximising value $\tilde{\boldsymbol{y}}_m$ of h. This is done by setting the first derivative equal to zero and solving for $\boldsymbol{y}_m$. We obtain the estimating equations

$$\begin{aligned} 0 &= \omega - \omega e^{\eta_s}/(1+e^{\eta_s}) - (y_s - \zeta_s)/\sigma^2 \\ \implies y_s/\sigma^2 + \omega e^{\eta_s}/(1+e^{\eta_s}) &= \omega + \zeta_s/\sigma^2 \\ \implies y_s &= \sigma^2\omega/(1+e^{\eta_s}) + \zeta_s\,. \end{aligned}$$

Each equation is nonlinear in y_s for nonzero ω. The form of the estimating equations means that as ω, the strength of the influence of $\boldsymbol{y}$ on $\boldsymbol{r}$, goes to zero, the function h is maximised at the value $\boldsymbol{\zeta}_m$ which represents the mean value of $\boldsymbol{y}_m$.

If ω is non-zero, then for each $s \in I_m$, the estimating equations may have one, two, or three solutions. Only one of these solutions corresponds to the value of y_s in the vector $\tilde{\boldsymbol{y}}_m$ that maximises h. Further, as the random effects vector elements in $\boldsymbol{N}$ and $\boldsymbol{B}$ vary continuously, the value of y_s that maximises h may jump discontinuously. This means that the value of

$$H(\boldsymbol{N},\, \boldsymbol{B}) \approx \int_{\mathbb{R}^q} e^{h(\boldsymbol{b})}$$

may also jump discontinuously, which violates the condition of smoothness when applying the Laplace approximation to the outer integral of Expression (4.16).

It appears that evaluating the joint distribution in the NMAR case using the Laplace approximation is intractable, and a different method must be used. Fortunately, the method of MCMC may be used to obtain the joint distribution and perform inference for the parameters $\boldsymbol{\theta}$, and this method will be explored in the next section.

4.4 Model fitting with `stan` MCMC procedures

MCMC is a method of obtaining samples from the joint distribution of some set of parameters $\boldsymbol{\theta} \in \Theta$. Samples are obtained by constructing and updating a discrete time Markov chain whose state space is the parameter space Θ and whose stationary distribution is identical to the joint distribution of the parameters. After an initial "burn-in" process, where a subset consisting of initial samples is discarded, the sequence of samples may be used to estimate the mean values of the parameters. MCMC easily accommodates the

Bayesian approach in which case the stationary distribution is the posterior distribution of the parameters given a prior distribution and observed data (Gilks, 2005). Appendix C.2 describes MCMC in detail, including the notion of *overlapping batch means*, a method of estimating the standard error of MCMC parameter estimates that accounts for autocorrelation in the chains.

Some drawbacks of MCMC are that there is never a guarantee that the number of iterations N_M is sufficiently high to explore the stationary distribution of the chain, and that neither the use of multiple chains or a burn-in period guarantees that convergence to the stationary distribution has occurred (Geyer, 2011).

A commonly-used indicator of convergence is the $\hat{R}$ statistic which compares the ratio of variation within chains and variation between chains (Gelman and Rubin, 1992). An $\hat{R}$ value below 1.1 for all parameters is taken as an indicator of convergence and reliability of estimates.

The program `stan`, which may be called from within R, was used to fit models using the Hamiltonian MCMC method described in Appendix C.2.2. The program was called using the `stan` function from the R package `rstan`. The details of setting the mass matrix M and determining the length of time to run the Hamiltonian dynamics, and the intricacies of solving the Hamiltonian partial differential equations, were automatically determined by the `stan` function using default settings.

Eight chains of length 3000 were used in all `stan` model fitting for the GC dataset. Using the default settings of the program, the first 1500 samples from each chain were discarded as part of the burn-in process, resulting in $8N_M = 12000$ total samples from the posterior distribution of the parameters. The options `adapt_delta` and `max_treedepth` were set to 0.85 and 14 respectively.

Parameter estimates were obtained by applying the `summary` function to the fitted `stan` models and extracting the means of the 12000 samples for each parameter. Parameter standard deviations were estimated using the method of overlapping batch means on each of the 8 chains, and averaging the result.

Joint models were fitted to 157 of the 159 peaks in the GC dataset. The excluded peaks are at 4358 and 7917 m/z. The model on the former peak could not be fitted due to the constant response in the missingness indicator, and the model on the latter peak could not be fitted due to a technical limitation of R and `stan` that arises because that peak has exactly one missing observation.

In the modelling performed in this chapter, the intensity values in the GC dataset were recentred at their grand mean of 8.584.

The MAR joint model (4.6) and the NMAR joint model (4.9) were modified to incorporate prior distributions. The posterior distribution for the MAR joint model is

$$f(\boldsymbol{\theta}\,;\boldsymbol{y},\,\boldsymbol{r}) = f(\boldsymbol{r}\,;\,\boldsymbol{\theta}_1)f(\boldsymbol{y}\mid\boldsymbol{N},\,\boldsymbol{B}\,;\,\boldsymbol{\theta}_2)f(\boldsymbol{N},\,\boldsymbol{B}\,;\boldsymbol{\theta}_3)\pi(\boldsymbol{\theta}),$$

where $\pi(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$. As the parameter set $\boldsymbol{\theta}$ consists of disjoint subsets of parameters, the prior distribution on the parameter set was chosen such that

it factorised into a product of prior distributions for the subsets. The prior distribution may be written as

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\lambda})\pi(\boldsymbol{\kappa})\pi(\sigma^2)\pi(\sigma_N^2)\pi(\sigma_B^2),$$

where $\pi(\boldsymbol{\lambda})$ and $\pi(\boldsymbol{\kappa})$ were chosen to be multivariate normal distributions and the prior distributions $\pi(\sigma^2)$, $\pi(\sigma_N^2)$, and $\pi(\sigma_B^2)$ were chosen to be gamma distributions. The hyperparameters for these prior distributions were estimated from the GC dataset using estimates from models (4.1) and (4.5).

The posterior distribution for the NMAR joint model is

$$f(\boldsymbol{\theta}\,; \boldsymbol{y}, \boldsymbol{r}) = f(\boldsymbol{r} \mid \boldsymbol{y}\,; \boldsymbol{\theta}_1) f(\boldsymbol{y} \mid \boldsymbol{N}, \boldsymbol{B}\,; \boldsymbol{\theta}_2) f(\boldsymbol{N}, \boldsymbol{B}\,; \boldsymbol{\theta}_3) \pi(\boldsymbol{\theta}),$$

where $\pi(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$. The prior distribution was

$$\pi(\boldsymbol{\theta}) = \pi(\omega)\pi(\boldsymbol{\lambda})\pi(\boldsymbol{\kappa})\pi(\sigma^2)\pi(\sigma_N^2)\pi(\sigma_B^2).$$

The prior distributions other than $\pi(\omega)$ were the same normal and gamma distributions as used for the MAR joint model. However, the addition of the parameter ω changed the interpretation of the other missingness parameters in $\boldsymbol{\lambda}$ and altered estimates of the parameters in $\boldsymbol{\kappa}$ relevant to the intensity values. The same prior distributions were used, but this was problematic as the prior distributions may have implied a more precise knowledge of the location of parameters than what was justified.

The prior distribution $\pi(\omega)$ was chosen to be a gamma distribution with rate 0.5 and shape 3 on -ω, constraining ω to the non-positive real numbers. The reasoning for this constraint is that negative values of ω correspond to probabilities of missingness rising with decreasing values of y_s. The opposite case, achieved for positive values of ω, does not make sense in light of how missing values in mass spectrometry data occur.

In addition to the two joint models, the intensity model (4.1) and a number of the missingness models from Chapter 3 were fitted using `stan`. This was done because refitting the existing frequentist and Bayesian R models serves as a 'sanity check' for `stan` models in terms of programming error detection and as a diagnostic for Markov chain behaviour. Additional model fitting work is discussed in Appendix F, which concerns the MAR joint model fitted to simulated data.

4.4.1 Preliminary model fitting

A number of preliminary models were fitted before the NMAR joint model in order to gauge the performance of `stan` in fitting the NMAR joint model, and to provide a basis for comparison of parameter estimates.

Existing R models for the GC dataset which were fitted in `stan` begin with Equation (4.1), the LMM for the intensity derived from the work of Stanford (2015). This model, reproduced here, is

$$y_{jk\ell mn} = \nu + \gamma_j + \delta_\ell + N_{jk} + B_{jk\ell m} + \varepsilon_{jk\ell mn}\,.$$

Next was Equation (4.5), the GLM for the missingness, reproduced here:

$$R_{jk\ell mn} \sim \text{Bern}(p_{jk\ell mn}), \quad p_{jk\ell mn} = \frac{e^{\eta_{jk\ell mn}}}{1 + e^{\eta_{jk\ell mn}}}, \quad \eta_{jk\ell mn} = \mu + \alpha_j + \beta_\ell \,.$$

The final R model was Equation (4.4), the Bayesian GLMM for the missingness, reproduced here:

$$R_{jk\ell mn} \sim \text{Bern}(p_{jk\ell mn}), \quad p_{jk\ell mn} = \frac{e^{\eta_{jk\ell mn}}}{1 + e^{\eta_{jk\ell mn}}}, \quad \eta_{jk\ell mn} = \mu + \alpha_j + \beta_\ell + M_{jk} + C_{jk\ell m} \,.$$

For this model, the prior distributions described in Section 3.4.2 were put on the fixed effects and random effects variance components.

The final preliminary model was the MAR joint model (4.6). The MAR joint model differs from the separate models from which it is derived in that prior distributions on the parameters were specified in the joint model, whereas they were not specified in the separate models.

The non-Bayesian GLMMs for the missingness described in Chapter 3 were not considered. All mentions of the R GLMMs are in reference to the Bayesian GLMM.

In the preliminary model fitting, the fixed effects parameter estimates from `stan` models were taken to be the means of the 12000 samples of the parameters. The random effect variance parameter estimates from `stan` models, on the other hand, were taken to be the medians of the 12000 samples. This was because the posterior distributions of the variance parameters tended to be skewed, necessitating the use of an estimate robust to the asymmetries of the distributions. However, it is noteworthy that in the vast majority of cases, the difference between the estimated means and medians of variance parameters across the various models was negligible.

The models with random effects, which are the LMM (4.1) and the GLMM (4.4), did not exhibit an exceedingly close correspondence between R parameter estimates and `stan` parameter estimates of the random effect variances. There are multiple reasons for this. The first is that in these comparisons, `stan` was used to estimate the median values of random effect variance parameters under the models, while R model fitting functions from `lme4`, i.e. `lmer` and `glmer`, obtain estimates of modes of parameter distributions under the same models. The second is that the random effects vector elements are explicitly modelled in `stan` models but not in R models. Another difference between R and `stan` model fits for mixed effects models is that R's estimates of variance component parameters can equal zero, but `stan`'s estimates of those parameters cannot, as `stan` samples avoid boundaries of the parameter space by transforming parameters to $\mathbb{R}$ before sampling (Carpenter et al., 2017; Chung et al., 2013). Because variance components are constrained to be non-negative, Markov chain sampling of the variance components occurs on the log scale. Estimates of variance components equal to zero correspond to negative infinity on the log scale, which does not get explored by the Markov chains.

The random effects, in models that incorporated them, were internally rewritten in the `stan` code to follow a non-centred parametrisation in order to obtain less biased estimates of variance components (Papaspiliopoulos et al., 2007). For the LMM (4.1), this reparametrisation means N_{jk} may be rewritten as $\sigma_N N^*_{jk}$ where

$$N^*_{jk} \sim N(0,\, 1),$$

and likewise $B_{jk\ell m} = \sigma_B B^*_{jk\ell m}$ where

$$B^*_{jk\ell m} \sim N(0,\, 1).$$

The same reparametrisations of these two random effects are made in the MAR joint model (4.6) and the NMAR joint model (4.9). For the GLMM (4.4), the reparametrisations made are $M_{jk} = \sigma_M M^*_{jk}$ where

$$M^*_{jk} \sim N(0,\, 1),$$

and $C_{jk\ell m} = \sigma_C C^*_{jk\ell m}$ where

$$C^*_{jk\ell m} \sim N(0,\, 1).$$

The density functions of the models are invariant under these reparametrisations, and so the models are presented according to the original, centred parametrisations.

The R and `stan` estimates for the GLM (4.5) were very close. This was due to the simplicity of the model and the large sample size for estimating each parameter, which caused the stochastic variation in the MCMC process to be minimal.

The `stan` parameter estimates for the MAR joint model and for the separate models were similar, with differences arising mainly from the use of prior distributions in the joint models but not in the LMM (4.1) or the GLM (4.5).

LMM for the intensity

The parameters in the LMM (4.1) consist of $\boldsymbol{\kappa} = (\nu\,,\, \gamma_2\,,\, \gamma_3\,,\, \gamma_4\,,\, \gamma_5\,,\, \delta_2\,,\, \delta_3)$, the vector of fixed effects, and the random effects variance components $\sigma^2_N\,,\, \sigma^2_B\,,$ and σ^2. Figure 4.1 contrasts the R model parameter estimates with the `stan` model parameter estimates.

There was a large degree of agreement between the R and `stan` model estimates of fixed effects parameters. The differences between the estimates typically were found in the range from 10^{-4} to 10^{-2}. The estimates of the residual variance σ^2 and the C8-level variance component σ^2_B were close, but to a lesser degree than the fixed effects parameters. However, some estimates of σ^2_N differed by more than 0.05 (up to 0.18). This occurred for the peaks at 2793, 7806, 8265, 9319, 15631, 15882, and 16030 m/z, which have more than 75% missing values except for peak 7806 m/z with only 388 missing values and peak 16030 m/z with only 36 missing values.

The vast majority of `stan` estimates of variance components were slightly larger than the corresponding R estimates. All estimates of σ^2_N were larger in the models fitted with `stan`, and consequently, the plot of the estimates of σ^2_N appears to have a slope slightly greater than 1. A reason for this could be that the R estimates of variance components parameters are MLEs, corresponding to the mode of the distribution, while the `stan` MCMC estimates of parameters are the medians of the samples from the posterior distributions of the parameters (Bates et al., 2015; Carpenter et al., 2017). Posterior medians tend to be greater than posterior modes for distributions that are positively skewed. Such skewness of distribution is often the case for variance components as they are constrained to non-negative values.

GLM for the missingness

The parameters in the GLM (4.5) consist of $\boldsymbol{\lambda} = (\mu\,,\, \alpha_2\,,\, \alpha_3\,,\, \alpha_4\,,\, \alpha_5\,,\, \beta_2\,,\, \beta_3)$, the vector of fixed effects. Figure 4.2 contrasts the R and `stan` model estimates of parameters in the model. Apart from a minority of outliers, there was almost perfect correlation ($\rho > 0.998$) between the R model and `stan` model estimates of these parameters. This was due to the simplicity of the model. The outliers correspond to peaks with either high or low proportions of missing values, which were peaks whose parameters tended to take extreme values, and were the most difficult to estimate.

Bayesian GLMM for the missingness

The parameters in the GLMM (4.4) consist of $\boldsymbol{\lambda} = (\mu\,,\, \alpha_2\,,\, \alpha_3\,,\, \alpha_4\,,\, \alpha_5\,,\, \beta_2\,,\, \beta_3)$, the vector of fixed effects, and the random effects variance components σ^2_M and σ^2_C. Figure 4.3 contrasts the R and `stan` model estimates of parameters in the model.

In general, the parameters in a GLMM are estimated with more difficulty than the corresponding parameters in an analogous LMM with the same mixed effects structure. Although all 1080 observations *of the missingness indicator* are always present, each observation contributes a small amount of information, as it is a binary outcome in contrast to the real-valued response in a LMM.

The estimates of fixed effects between the two model fitting approaches displayed large positive correlation ($\rho > 0.96$) for the parameters μ, α_2, β_2, and β_3. Disagreements between R and `stan` parameter estimates tended to occur in cases of complete separation in the missingness indicators, which drove the correlation down. The other fixed effects parameters, which are α_3, α_4, and α_5, had lower correlations between the R and `stan` estimates. These decreases in correlation were largely due to outliers from the peak at 7490 m/z, whose parameters are difficult to estimate due to there being only two missing observations in that peak. There was little to no bias in the `stan` estimates of the fixed effects relative to the R estimates.

The random effect variance component estimates had correlation coefficients of 0.958

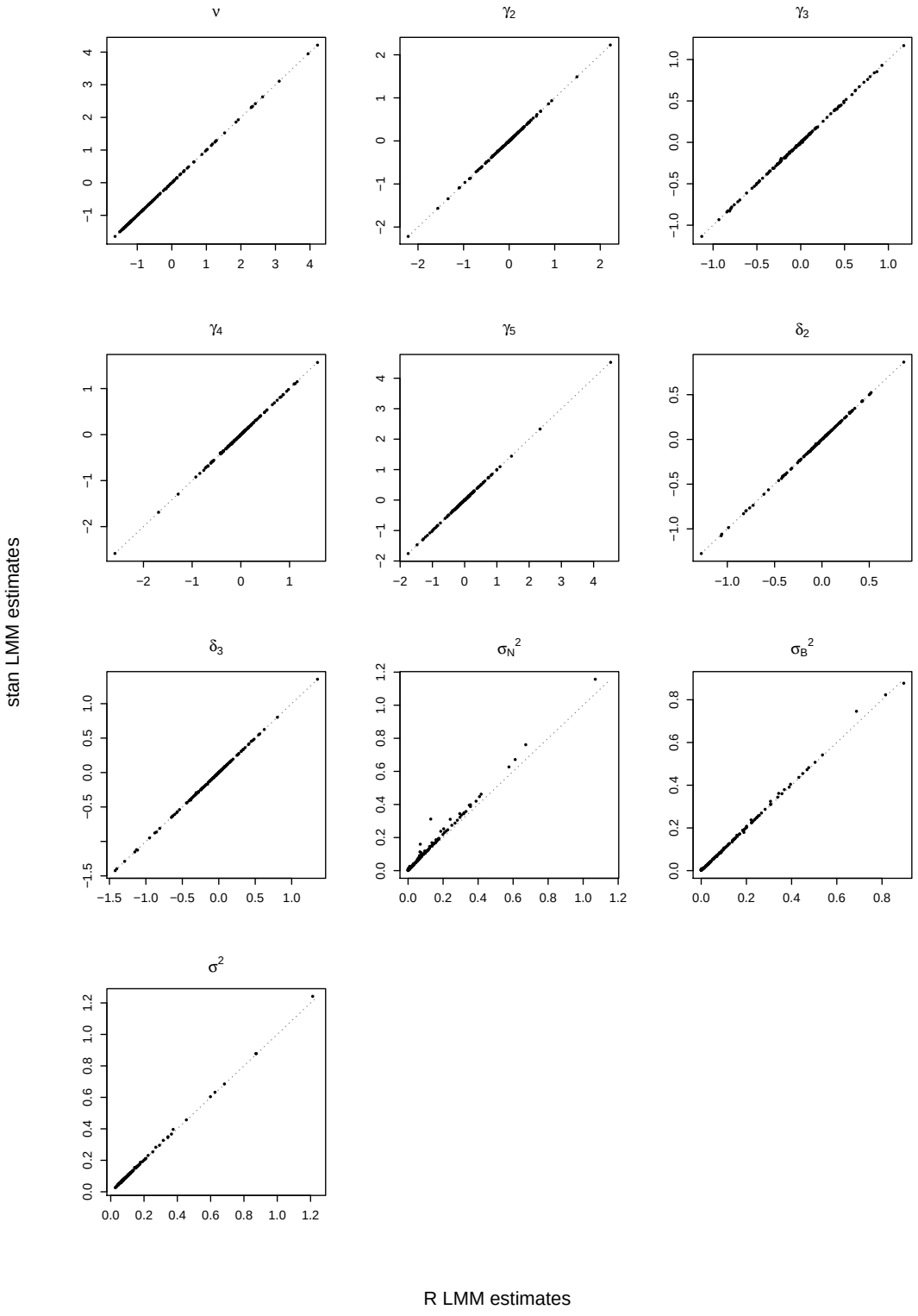

Figure 4.1: Plots of estimates of each parameter from R and `stan` LMMs. Each plot window is devoted to a single parameter and contains the estimates from each peak's models. The x axis of each plot represents the R model estimate, and the y axis represents the `stan` model estimate. Points within plots represent individual peaks. The closer the parameter estimates are for a given peak, the closer the point is to the dotted line $y = x$.

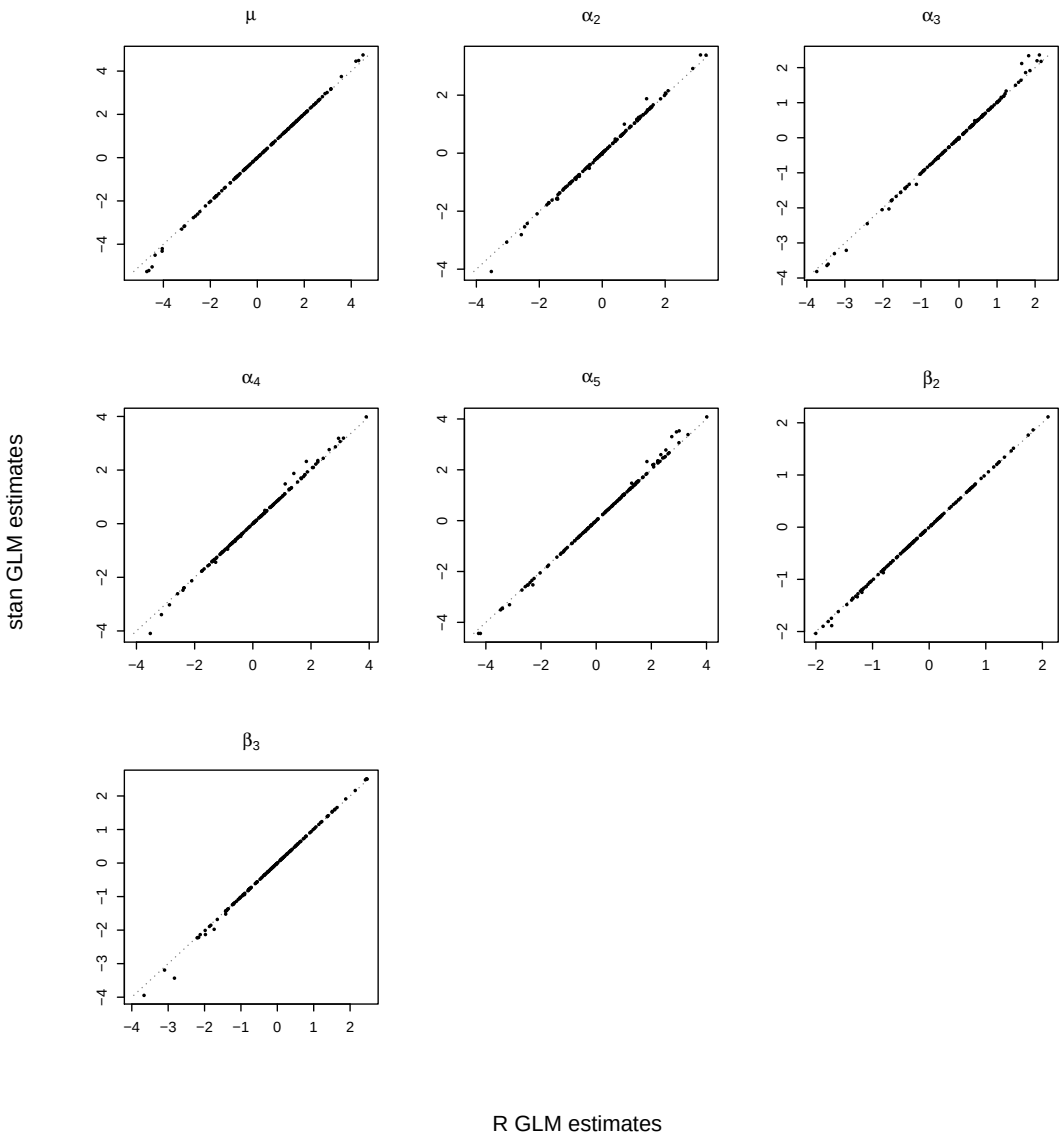

Figure 4.2: Plots of estimates of each parameter from R (x axes) and `stan` (y axes) GLMs. Each plot window is devoted to a single parameter and contains the estimates from each peak's models. Points within plots represent individual peaks. The closer the parameter estimates are for a given peak, the closer the point is to the dotted line $y = x$.

and 0.625 for the estimates of σ_M^2 and σ_C^2, respectively. However, many peaks had R estimates of σ_M^2 equal to zero, and `stan` estimates of that parameter were biased upwards relative to the R estimates. The peaks with the biggest discrepancies in the σ_M^2 parameter followed one of two tendencies. The first was the missingness count being either extremely low or extremely high, and the second was large discrepancies in missingness counts between different mice and their sets of 27 observations. The peaks responsible for the most extreme outliers in the estimates of σ_C^2 were at 2008, 2033, 2057, 5275, 8007, 13648, and 17458 m/z. None of these peaks have more than 20% of their observations missing. In the last two plots in Figure 4.3, which contain the variance component estimates, three features are apparent. The first is the diffuse nature of the set of points, especially for large parameter estimatess, signifying large differences between the R and `stan` estimates. The second is the vertically-aligned cluster of points near the origin of the plot, corresponding to peaks for which a variance component was estimated as zero under the R model but estimated as greater than zero under the `stan` model. The third is the tendency of points to lie above the $y = x$ line. The first feature results from the difficulty of estimating variance components in a GLMM as well as the explicit modelling of random effects vector elements in the `stan` model but not in the R model. The second feature arises because the algorithms for estimating variance component values in the R model will output 0 in some cases, whereas `stan` algorithms avoid boundaries of the parameter space if the parameters are constrained. Finally, the third feature occurs because variance components are nonnegative and consequently have positively skewed probability distributions, inflating the medians more so than the modes.

MAR joint model

Recall that the posterior distribution of the MAR joint model is expressed as

$$f(\boldsymbol{y},\, \boldsymbol{r}\,;\boldsymbol{\theta}) = f(\boldsymbol{r}\,;\,\boldsymbol{\theta}_1) f(\boldsymbol{y} \mid \boldsymbol{N},\, \boldsymbol{B}\,;\,\boldsymbol{\theta}_2) f(\boldsymbol{N},\, \boldsymbol{B}\,;\boldsymbol{\theta}_3)\pi(\boldsymbol{\theta})$$

where $f(\boldsymbol{r}\,;\,\boldsymbol{\theta}_1)$ is the likelihood of $\boldsymbol{r}$ arising from the GLM (4.5), $f(\boldsymbol{y} \mid \boldsymbol{N},\, \boldsymbol{B}\,;\,\boldsymbol{\theta}_2) \cdot f(\boldsymbol{N},\, \boldsymbol{B}\,;\boldsymbol{\theta}_3)$ is the likelihood of $\boldsymbol{y}$ arising from the LMM (4.1), and $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}_1)\pi(\boldsymbol{\theta}_2)\pi(\boldsymbol{\theta}_3)$ is the prior distribution of the parameters.

The prior distribution was obtained via a process that involved fitting the LMMs and GLMs in R. The parameters in $\boldsymbol{\theta}_1$ consist solely of $\boldsymbol{\lambda}$, the vector of fixed effects. Therefore, $\pi(\boldsymbol{\theta}_1) = \pi(\boldsymbol{\lambda})$. To obtain the prior distribution for these parameters, the GLM (4.5) was fitted (using R) to the missingness indicators in the 143-peak subset of the GC dataset. The means of the 143 estimates of each fixed effect, contained in the vector $\hat{\boldsymbol{\lambda}}$, as well as the estimated variance-covariance matrix Σ_m of the fixed effects were obtained using a procedure analogous to that described in Section 3.4.2 for the fixed effects in the missingness model. The prior distribution $\pi(\boldsymbol{\lambda})$ was given by

$$\boldsymbol{\lambda} \sim N_7(\hat{\boldsymbol{\lambda}},\, \Sigma_m).$$

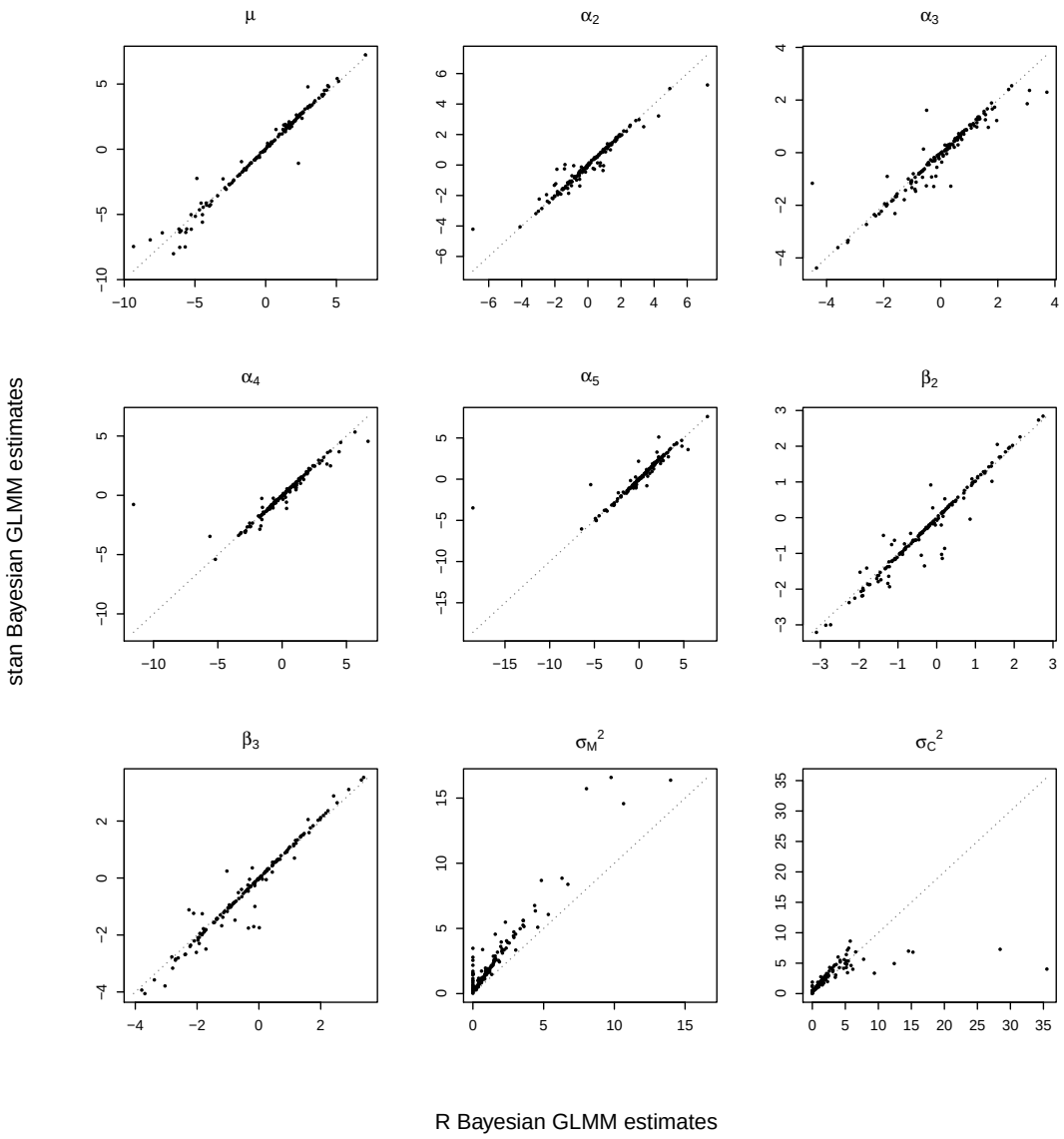

Figure 4.3: Plots of estimates of each parameter from R (x axes) and `stan` (y axes) GLMMs. Each plot window is devoted to a single parameter and contains the estimates from each peak's models. Points within plots represent individual peaks. The closer the parameter estimates are for a given peak, the closer the point is to the dotted line $y = x$.

Table 4.1: Means of estimated values of $\hat{\boldsymbol{\lambda}}$ for the MAR joint models.

Parameter	Mean
μ	0.338
α_2	0.033
α_3	-0.036
α_4	0.047
α_5	0.301
β_2	-0.184
β_3	-0.187

Tables 4.1 and 4.2 list the values of $\hat{\boldsymbol{\lambda}}$ and Σ_m .

The parameters in $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ consist of the vector of fixed effects $\boldsymbol{\kappa}$ and the residual variance σ^2 for $\boldsymbol{\theta}_2$ and the variance components σ_N^2 and σ_B^2 for $\boldsymbol{\theta}_3$. As a consequence, the prior distribution $\pi(\boldsymbol{\theta}_2)$ is equal to $\pi(\boldsymbol{\kappa})\pi(\sigma^2)$, and $\pi(\boldsymbol{\theta}_3) = \pi(\sigma_N^2)\pi(\sigma_B^2)$. To obtain the prior distributions for these parameters, the LMM (4.1) was fitted (using R) to the intensities of 157 peaks in the GC dataset, with the peaks at 4152 and 11757 m/z excluded due to missing data in certain groups. The means of the 157 estimates of each fixed effect, contained in the vector $\hat{\boldsymbol{\kappa}}$, as well as the estimated variance-covariance matrix Σ_o of the fixed effects were obtained using procedures analogous to those described in Section 3.4.2 for the fixed effects in the missingness model. The prior distribution $\pi(\boldsymbol{\kappa})$ is given by

$$\boldsymbol{\kappa} \sim N_7(\hat{\boldsymbol{\kappa}}, \Sigma_o).$$

Tables 4.3 and 4.4 list the values of $\hat{\boldsymbol{\kappa}}$ and Σ_o . The 157 estimates of the variance parameters σ_N^2 , σ_B^2 , and σ^2 were recorded, and a gamma distribution was fitted to each parameter using the `fitdistr` function in order to obtain estimates of the shape and rate parameters. Gamma distributions with the estimated shape and rate parameters were used as the prior distributions $\pi(\sigma_N^2)$, $\pi(\sigma_B^2)$, and $\pi(\sigma^2)$. Table 4.5 lists the estimated shape and rate parameters.

Figures 4.4 and 4.5 compare the MAR joint model estimates with the GLM and LMM `stan` estimates, respectively. The MAR joint model had quite similar estimates of parameters to the separate models fitted in `stan`. The missingness parameters in $\boldsymbol{\lambda}$ appeared to have the most agreement, followed by the intensity parameters in $\boldsymbol{\kappa}$, with the variance components parameter estimates differing the most. The differences largely resulted from the fact that prior distributions were included in the joint model but not in the two separate models.

The effect of the prior distributions on μ (the missingness intercept) and the γ_j parameters (intensity group differences) was to shrink the most extreme estimates of those

Table 4.2: Estimated variance-covariance matrix Σ_m for the MAR joint models.

	μ	α_2	α_3	α_4	α_5	β_2	β_3
μ	3.54	−0.56	−0.87	−1.28	−1.63	0.33	0.63
α_2	−0.56	1.39	0.37	0.84	0.52	−0.14	−0.09
α_3	−0.87	0.37	1.22	0.89	1.23	−0.09	−0.21
α_4	−1.28	0.84	0.89	1.76	1.15	−0.26	−0.37
α_5	−1.63	0.52	1.23	1.15	2.53	−0.29	−0.34
β_2	0.33	−0.14	−0.09	−0.26	−0.29	0.64	0.75
β_3	0.63	−0.09	−0.21	−0.37	−0.34	0.75	1.17

Table 4.3: Means of estimated values of $\hat{\boldsymbol{\kappa}}$ for the MAR joint models.

Parameter	Mean
ν	-0.439
γ_2	-0.020
γ_3	-0.035
γ_4	-0.022
γ_5	-0.029
δ_2	-0.057
δ_3	-0.100

Table 4.4: Estimated variance-covariance matrix Σ_o for the MAR joint models.

	ν	γ_2	γ_3	γ_4	γ_5	δ_2	δ_3
ν	1.21	−0.13	−0.10	−0.09	−0.22	−0.08	−0.09
γ_2	−0.13	0.22	0.05	0.11	0.12	−0.02	−0.00
γ_3	−0.10	0.05	0.14	0.12	0.13	−0.02	−0.03
γ_4	−0.09	0.11	0.12	0.23	0.11	−0.03	−0.04
γ_5	−0.22	0.12	0.13	0.11	0.40	−0.01	−0.00
δ_2	−0.08	−0.02	−0.02	−0.03	−0.01	0.09	0.10
δ_3	−0.09	−0.00	−0.03	−0.04	−0.00	0.10	0.15

Table 4.5: Estimated hyperparameters of random effect variance component distributions for the MAR joint models.

	Shape	Rate
Mouse variance σ_N^2	0.699	6.340
C8 batch variance σ_B^2	0.922	6.801
Residual variance σ^2	1.894	13.823

parameters towards zero. For the random effect variance components, the effect of the prior distributions was less clear. The shrinkage in the fixed effects is apparent in the plots as points being located below the dotted lines for high x values and above the dotted lines for low x values.

The parameter σ_N^2 sometimes took very different estimates from the MAR joint model as compared to the LMM. This occurred, for example, for the peaks at 7806, 8302, 15631, 15724, and 15882 m/z, each of these peaks having a difference of more than 0.15 in the parameter estimates. The discrepancy in the estimates reflected a difficulty of estimating the parameter which is induced by observations from subsets of mice in certain groups having outlying intensities relative to the rest of the mice in the group. Large proportions of missing observations for these peaks also cause parameter estimates to be more informed by the prior distribution than the data.

4.4.2 NMAR joint model

The majority of `stan` estimates and R estimates from the separate intensity and missingness models displayed close associations, and the `stan` estimates from the MAR joint model were likewise close to the `stan` estimates from the separate models. This increases the confidence placed in this method of fitting models. The performance of `stan` in fitting the NMAR joint model, as well as the suitability of the model to the GC dataset, was checked using a number of methods. The first method is particular to the MCMC procedures used to fit the model, and concerns the behaviour of the Markov chain samples. The latter methods are simulation checks. The details of the simulation checks are provided in Appendix E.

The conclusions from the checks were that the models perform well. Two out of 159 peaks in the GC dataset did not admit joint models, and one more peak yielded parameter estimates that could not be trusted due to unmixed Markov chains, but sensible parameter estimates were obtained for the remaining peaks.

In Appendix F, simulation checks are also undertaken for the MAR joint model, revealing that the NMAR joint model is more suitable for the GC dataset than the MAR joint model in terms of precision of parameter estimates and bias reduction due to the inclusion of the ω parameter. The NMAR joint model is also superior to the MAR joint

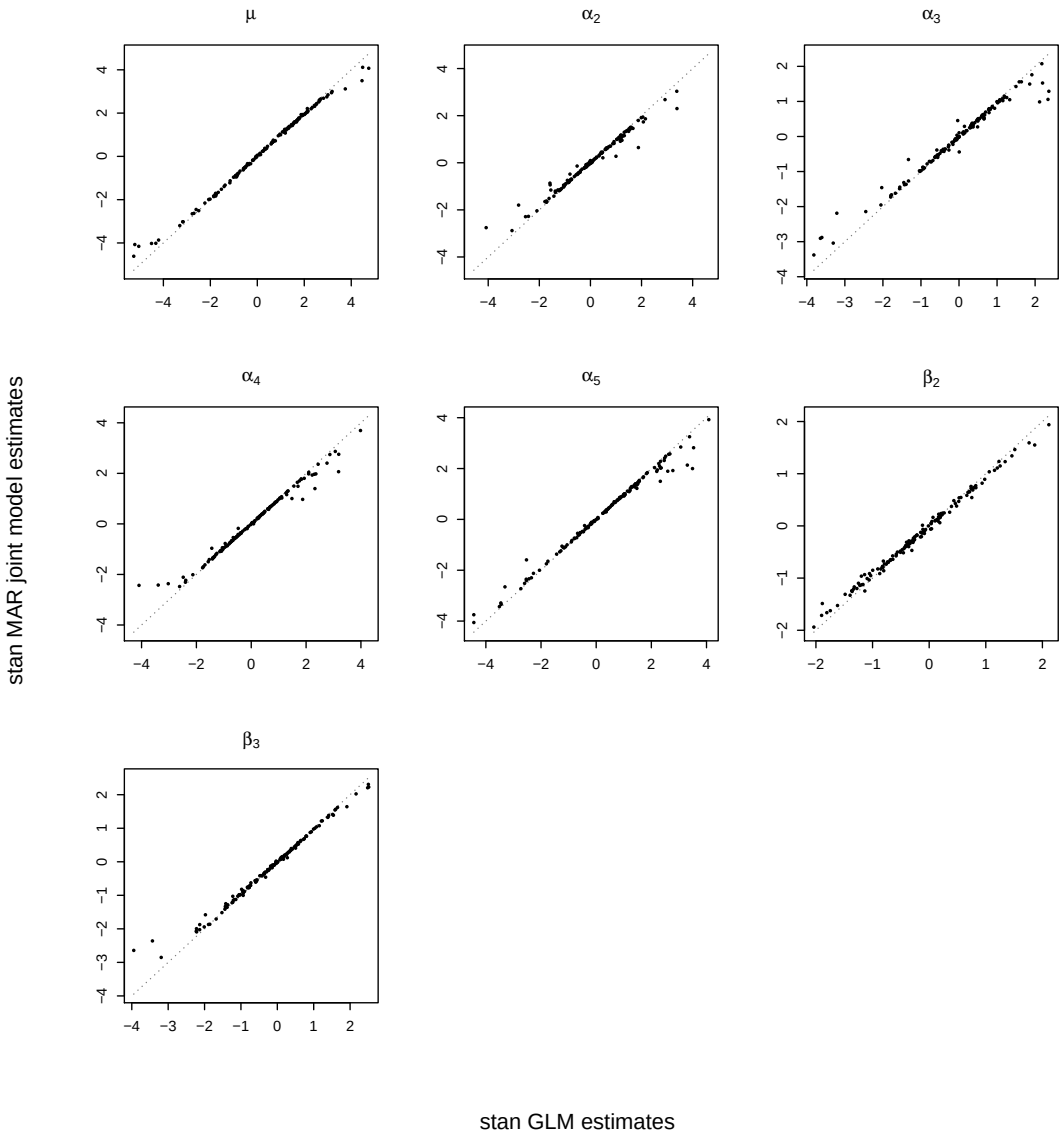

Figure 4.4: Plots of estimates of each parameter from `stan` missingness GLM (x axes) and MAR joint model (y axes). Each plot window is devoted to a single parameter and contains the estimates from each peak's models. Points within plots represent individual peaks. The closer the parameter estimates are for a given peak, the closer the point is to the dotted line $y = x$.

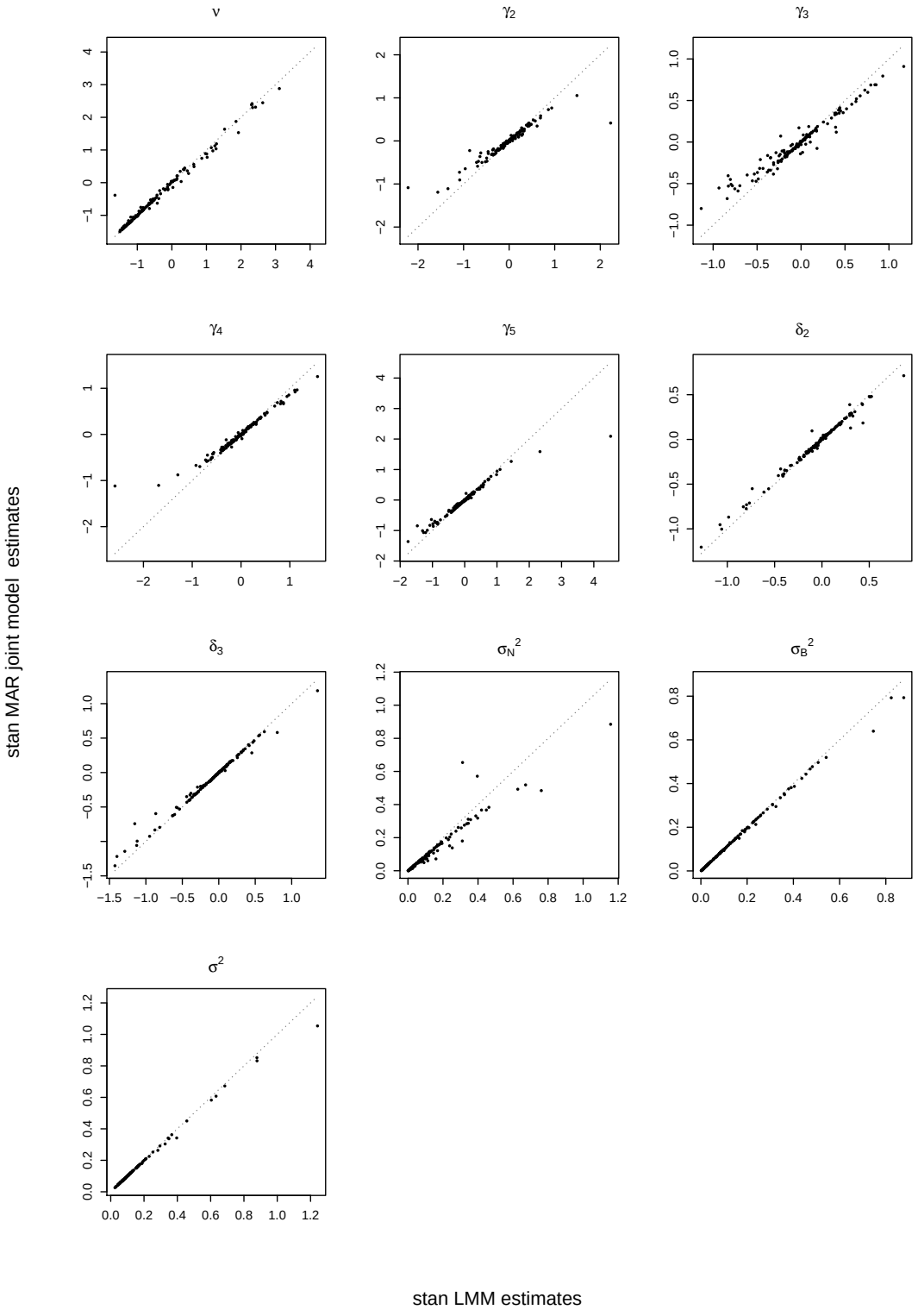

Figure 4.5: Plots of estimates of each parameter from `stan` intensity LMM (x axes) and MAR joint model (y axes). Each plot window is devoted to a single parameter and contains the estimates from each peak's models. Points within plots represent individual peaks. The closer the parameter estimates are for a given peak, the closer the point is to the dotted line $y = x$.

model in predictive simulations.

$\hat{R}$ and n_{eff} statistics

In a MCMC procedure, the distributions of the samples from the Markov chains converges to the stationary distribution. The speed of this convergence is reflected in the mixing of the Markov chains in the sense that slow convergence causes Markov chains with distinct initial values to have trajectories which are not well-mixed, meaning that they are distinct from each other even after many samples are taken, while fast convergence causes chains with distinct initial values to be well-mixed after a short time, meaning that the chains appear similar to each other (Gilks, 2005). The *Gelman-Rubin convergence statistic* $\hat{R}$ of a parameter is a measure of how well-mixed the chains are in terms of the square root of the ratio of estimated inter-chain variance to estimated average intra-chain variance for that parameter (Gelman and Rubin, 1992). The $\hat{R}$ statistic may also be defined for the log posterior density. An $\hat{R}$ statistic below 1.1 indicates sufficient mixing for Markov chain convergence (Gelman and Rubin, 1992; Gelman and Shirley, 2011). However, although $\hat{R} > 1.1$ implies a lack of convergence, convergence does not necessarily follow from $\hat{R} < 1.1$.

The joint models were unable to be fitted to the peaks at 4358 and 7917 m/z as these peaks had zero and one missing observation, respectively. In the NMAR joint model, the peak at 5972 m/z, which has 935 missing observations, had $\hat{R}$ values greater than 1.1 for all parameters except the γ_j parameters. The peak was excluded from the analysis for biomarker candidate discovery as its parameter estimates were not reliable. This resulted in data from 156 peaks out of the original 159 being analysed with the NMAR joint model.

The parameter n_{eff} is a measure of *effective sample size*, which, due to the autocorrelation within the Markov chains, was often less than the total number of iterations $8N_M = 12000$. The typical distribution of n_{eff} values across peaks tended to be roughly bimodal for all parameters, with a large number of peaks achieving low n_{eff} values for any given parameter alongside a number of peaks achieving $n_{\text{eff}} = 12000$. Table 4.6 displays the average values of n_{eff} for all parameters in the NMAR joint models over the set of 156 peaks.

4.4.3 The number of MCMC samples

The `stan` models used 12000 samples to estimate each parameter, but the effective sample sizes for the various parameters were frequently lower than this number. Effective sample sizes that were too low may have adversely affected the accuracy and precision of parameter estimates. While letting the `stan` models run for longer to increase the effective sample sizes is desirable, fitting the NMAR joint model to the 156-peak subset required up to 24 hours of runtime on a typical desktop computer, and increasing the sample size

Table 4.6: Mean effective sample sizes (compared to a maximum of 12000) across 156-peak subset under the NMAR joint models.

Parameter	Mean n_{eff}	Parameter	Mean n_{eff}
μ	4170	γ_3	5259
α_2	7966	γ_4	4723
α_3	7968	γ_5	4312
α_4	7331	δ_2	5118
α_5	7077	δ_3	4730
β_2	9519	σ_N^2	2952
β_3	8522	σ_B^2	2063
ν	2854	σ^2	2657
γ_2	4817	ω	4731

was counterproductive towards the iterative model development that was carried out in the course of this thesis.

The impact of an increased sample size on the accuracy and precision of parameter estimates is investigated here for three peaks. These are the peaks at 5637, 8505, and 16030 m/z, with 946, 523, and 36 missing observations respectively. The NMAR joint model 4.9 was fitted once to these peaks using 12000 samples, and again using 300000 samples. Models were fitted using the same number of burn-in samples (1500 per each chain).

The investigations showed that there cannot be a great benefit to the precision and accuracy from increasing the sample size, because estimated parameter posterior means and posterior standard deviations changed little upon increasing the number of samples. Table 4.7 displays the differences between estimated posterior means and posterior standard deviations of all parameters in model 4.9. Most differences were limited to the third decimal point, if not beyond. Table 4.8 presents the results for the peak at 5637 m/z in more detail. The differences between parameter mean and standard deviation estimates are small relative to the estimated values.

The greater the amount of missing observations, the greater the differences were between the estimates of μ and ω in the two runs of the model on a peak. This suggests that caution is necessary when precise estimates of these parameters are wanted. However, if the parameters relating to intensity fixed effects are of primary interest, as is the case in this thesis, then 12000 samples seems adequate.

Table 4.7: Differences between estimated parameter posterior means and between estimated posterior standard deviations from Model 4.9 using 12000 and 300000 iterations.

	5637 m/z		8505 m/z		16030 m/z	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
μ	0.056	-0.003	0.019	-0.000	-0.002	-0.000
α_2	-0.006	-0.004	-0.001	-0.002	-0.002	0.001
α_3	-0.007	0.001	-0.002	0.003	-0.001	-0.005
α_4	-0.008	-0.001	-0.004	-0.000	0.001	-0.003
α_5	-0.004	-0.006	-0.002	0.002	-0.004	-0.002
β_2	-0.002	0.000	-0.001	-0.002	0.005	0.000
β_3	0.002	-0.001	0.002	-0.003	0.004	-0.001
ν	0.001	-0.000	0.000	-0.001	0.002	0.005
γ_2	0.003	-0.002	-0.001	-0.002	-0.002	-0.004
γ_3	0.001	0.000	0.002	-0.004	0.002	0.002
γ_4	0.002	-0.001	0.004	0.003	0.000	-0.000
γ_5	-0.000	-0.001	0.001	0.001	-0.004	0.002
δ_2	-0.000	-0.001	0.001	-0.000	-0.000	-0.000
δ_3	0.000	-0.000	0.001	0.000	-0.001	0.000
σ_N^2	-0.000	-0.000	-0.001	-0.001	-0.001	0.000
σ_B^2	-0.000	0.000	-0.000	-0.000	0.000	0.000
σ^2	-0.000	-0.000	0.000	-0.000	0.000	-0.000
ω	0.034	-0.002	0.011	-0.000	-0.000	0.003

Table 4.8: Comparison of estimated parameter posterior means, posterior standard deviations, and effective sample sizes from Model 4.9 on peak 5637 m/z using 12000 interations and 300000 iterations.

	12000 iterations			300000 iterations		
Parameter	Mean	S.D.	n_{eff}	Mean	S.D.	n_{eff}
μ	-1.017	1.190	484	-1.073	1.192	12595
α_2	0.786	0.382	4926	0.792	0.386	125798
α_3	-0.883	0.291	4211	-0.876	0.290	106178
α_4	-0.197	0.332	2124	-0.189	0.333	59760
α_5	0.301	0.355	5593	0.305	0.361	135270
β_2	0.445	0.216	12000	0.447	0.216	300000
β_3	0.819	0.260	2400	0.817	0.261	300000
ν	-1.491	0.056	818	-1.492	0.057	20949
γ_2	0.051	0.077	913	0.047	0.078	23768
γ_3	0.077	0.057	1553	0.075	0.056	43934
γ_4	0.165	0.061	1351	0.163	0.062	42819
γ_5	-0.016	0.068	1301	-0.015	0.070	34389
δ_2	-0.013	0.041	1390	-0.012	0.041	40093
δ_3	-0.056	0.049	1051	-0.056	0.049	26079
σ_N^2	0.002	0.002	766	0.002	0.003	17265
σ_B^2	0.004	0.004	282	0.004	0.003	8428
σ^2	0.035	0.005	466	0.035	0.005	12762
ω	-1.919	0.778	487	-1.953	0.779	12573

4.4.4 Comparison of parameter estimates between MAR and NMAR models

The inclusion of the ω parameter in the NMAR joint model affects the interpretations of all other parameters in the model, which causes their estimates to be different compared to the MAR joint model estimates. For example, in the MAR joint model, the fixed effects in $\boldsymbol{\kappa}$ represent the means of the centred observed intensities across the chips and groups, but in the NMAR joint model, the same fixed effects represent the means of the centred observed *and* missing intensities, with missing intensities assumed to be lower, on average, than observed intensities. Figures 4.6a and 4.6b contrast the MAR and NMAR joint model parameter estimates (except for ω, as this parameter is not present in the MAR joint model).

The estimates of μ are completely different between the two joint models. In the MAR joint model, the value $\mu = 0$ corresponds to a 50% chance of missingness for observations in group 1 and chip 1, and an extreme value such as μ = -5 corresponds to approximately 1% of observations being missing. It is therefore very unlikely for a low estimate of μ to be accompanied by high proportions of missing values in group 1. However, in the NMAR joint models, the inclusion of the ω parameter changes the interpretation of μ and allows μ to take more extreme values. The interpretation of μ in the NMAR joint models is the log-odds of missingness probability for samples from group 1 and chip 1 *given that* the expression value of the observation is equal to the grand mean of the observed intensities in the GC dataset. It is plausible for μ to take extreme values such as -10 as long as the ω parameter is low enough to increase the probability of missingness for samples with low intensity. In this way, high levels of missingness are not incompatible with extremely low values of μ. This also accounts for the fact that the bulk of the NMAR joint model estimates of μ were less than the MAR joint model estimates.

The estimates of the α_j parameters had small positive correlations, with correlation coefficients in the range from 0.5 to 0.63. The spreads of estimated values for these parameters was smaller in the NMAR joint models than in the MAR joint models. This reduced spread may be due to the fact that when estimates of ω were low (corresponding to a large dependence of missingness probability on intensity), estimates of the α_j parameters tended be smaller in absolute value. The situation for the estimates of β_2 and β_3 between the two models was analogous.

The estimates of ν from the NMAR joint models are underestimates relative to those from the MAR joint models, with the difference between the two models' estimates tending to increase as the MAR estimates decrease. The estimates of group means from the MAR models are constrained to be near the mean of the observed data, no matter how many observations are missing, but the estimates from the NMAR models are allowed to fall below the observed data means in order to explain the missingness pattern in conjunction with the nonzero value of ω.

The estimates of the γ_j and δ_ℓ parameters displayed fairly high correlations that occu-

pied the range from 0.86 to 0.92. Some peaks have moderate MAR joint model estimates, but relatively extreme NMAR joint model estimates. This occurs for peaks with large proportions of missingness in the relevant groups and chips.

The estimates of σ_N^2, σ_B^2, and σ^2 in the NMAR joint models were almost always greater than or equal to the estimates of these parameters in the MAR joint models. A possible cause of this was the greater range for assumed intensity values for the NMAR joint model as compared to the MAR joint model. In some cases, the NMAR joint model estimates of these parameters were far greater than the MAR joint model estimates. This occurs for peaks with high proportions of missing values.

4.4.5 Checking random effects assumptions for the NMAR joint model

The vectors of random effects in the NMAR joint model, which are $\boldsymbol{N}$ for the mouse-level variation and $\boldsymbol{B}$ for the C8 batch-level variation, are assumed to take multivariate normal distributions with means of $\mathbf{0}$ and variance-covariance matrices respectively equal to $\sigma_N^2 I$ and $\sigma_B^2 I$. These assumptions may be checked using quantile-quantile plots, plotting the vector elements against the quantiles of a normal distribution. As a consequence of how the joint models were written in `stan`, all of the random effect vector elements were internally estimated regardless of whether or not individual elements corresponded to a set of samples whose observations were completely missing within that set. This meant that the assumption of normality could be checked over the entire set of vector elements in all peaks. This is preferable to checking the assumption using only the vector elements with at least one observation per element, because the NMAR missingness mechanism would cause vector elements with negative values to correspond to samples more likely to be all missing, destroying the symmetry around the mean that is assumed by the normal distribution.

Figure 4.7 displays quantile-quantile plots for the mouse effect vector $\boldsymbol{N}$ for a representative subset of 48 peaks. In many peaks where every vector element corresponds to at least one observation (meaning an absence of red disks in the plot), the assumption of normality is justified. There are occasional exceptions, such as the peak at 2906 m/z, in which the most extreme vector element estimates are further from 0 than what the assumption of normality would produce. Peaks where some vector elements correspond to zero observed data points (displayed as red disks) do not seem to violate the assumption of normality moreso than peaks where none do, except in pathological cases such as the peaks at 11757 and 15631 m/z. In such pathological peaks, much of the data is missing and some mice appear to be different to the rest in terms of the average observed expresion within the mice and the proportion of missing observations for the mice.

Figure 4.8 displays quantile-quantile plots for the C8 batch effect vector $\boldsymbol{B}$ for the same representative subset of 48 peaks. Peaks where the number of vector elements that

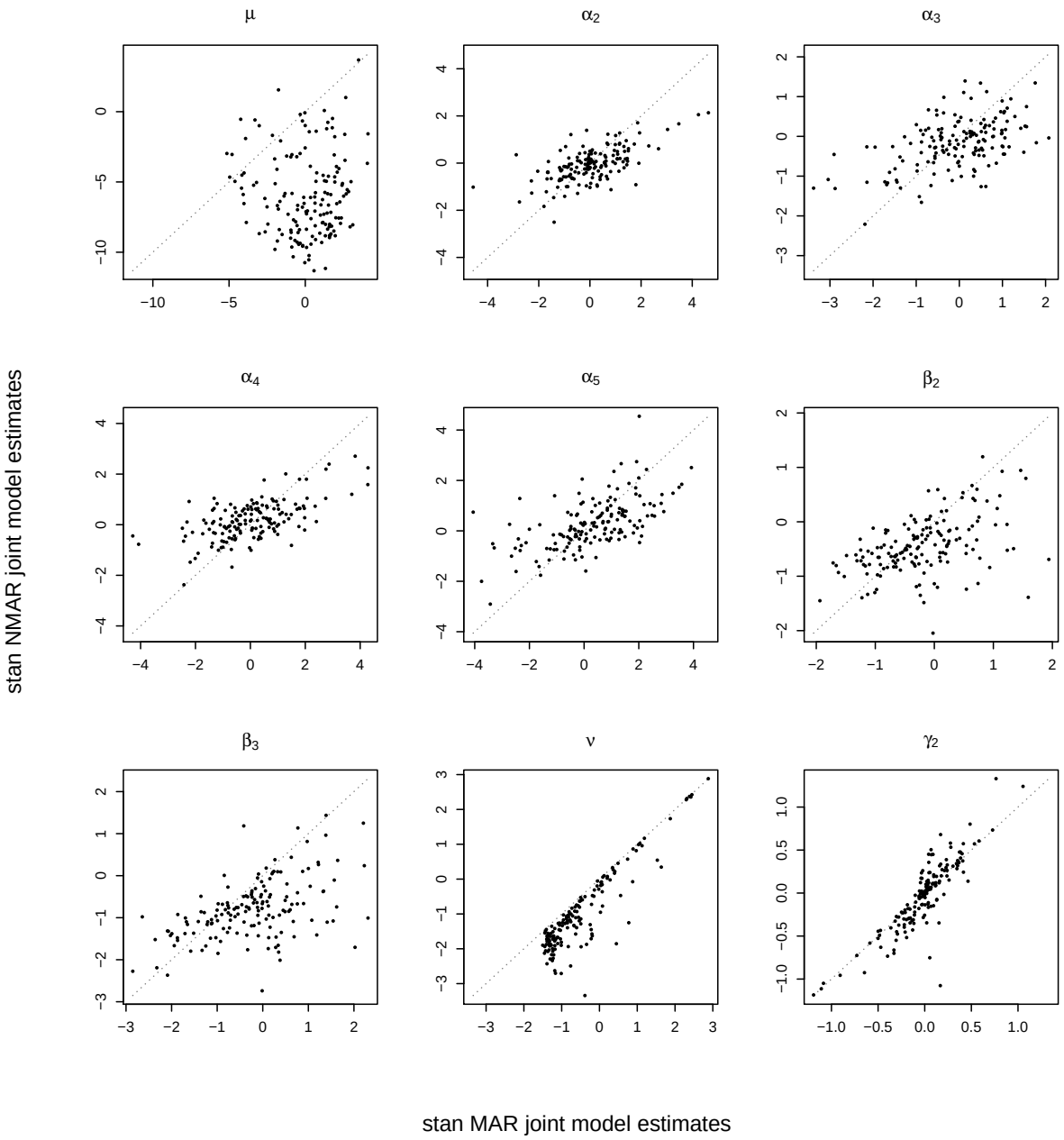

Figure 4.6a: Plots of estimates of each parameter from MAR joint model (x axis) and NMAR joint model (y axis). Each plot window is devoted to a single parameter and contains the estimates from each peak's models. Points within plots represent individual peaks. The closer the parameter estimates are for a given peak, the closer the point is to the dotted line $y = x$.

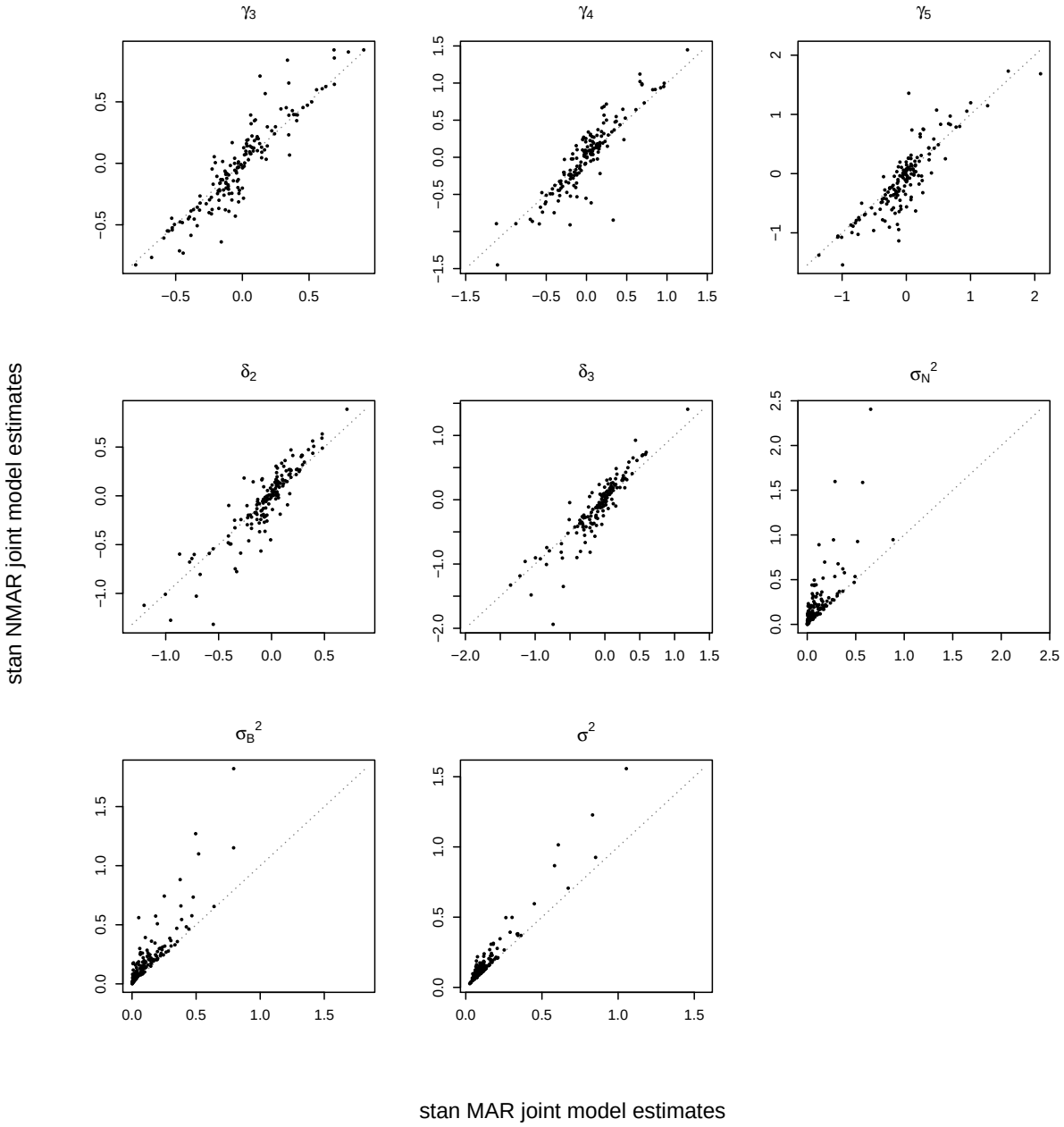

Figure 4.6b: Plots of estimates of each parameter from MAR joint model (x axis) and NMAR joint model (y axis). Each plot window is devoted to a single parameter and contains the estimates from each peak's models. Points within plots represent individual peaks. The closer the parameter estimates are for a given peak, the closer the point is to the dotted line $y = x$.

corresponded to a complete absence of observations was close to zero tended to meet the assumption of normality. The exceptions consisted of occasional outliers caused by the most extreme vector element estimates being further from 0 than what the assumption of normality would produce, and also a tendency towards positive skewness which manifested as points roughly following a concave-up curve and lying above the straight lines at the extremes. The cases where many vector elements corresponded to completely missing sets of observations, such as the peak at 5588 m/z, often exhibited marked departures from normality characterised by an abundance of estimated vector elements below, but very close to, zero. The assumption of normality is questionable in such cases, but the values of the vector elements estimates is likely to be an artefact of the MCMC procedure outputting values concentrated around the mean while no data exists to pull the estimated values away from the mean. This is corroborated by the fact that in the MAR joint model, random effect vector elements that corresponded to completely missing sets of observations were very close to zero in all cases.

Distribution with tails somewhat heavier than those of the normal distribution may be suitable for modelling the random effect vectors, as the increased probability mass in the tails would allow for the appearances of occasional extreme outliers in the element estimates. Additionally, in the cases where the C8 batch vectors contain many elements that are just below zero, the distribution could be scaled to place much of the density in a small interval around zero while allowing estimates far from zero to continue to exist.

4.5 Results for the NMAR model

Results for individual parameters are presented first. A contrast between the cancer and non-cancer groups' mean intensity estimates is later used to determine peaks of primary biological interest as gastric cancer biomarker candidates. The set of peaks determined to be of primary interest is presented together with additional peaks considered to be of secondary interest (which include, among others, the peaks from Stanford (2015)) in order to provide a more complete picture of biomarker candidates in the GC dataset.

4.5.1 Individual parameters

Figures 4.9, 4.10, and 4.11 display sets of histograms for the parameter estimates of the NMAR joint model over the 156-peak subset. The distribution of estimates of the reference category intensity parameter ν is somewhat positively skewed and has a mode near -2, while the distributions of estimates of the group and chip intensity differences γ_j and δ_ℓ are roughly symmetric and centred at or near zero. Given that the numerical data were recentred to the grand mean of the observed values (which was 8.583), this implies that the model tended to estimate peaks' mean intensities as being substantially below the grand mean of the GC dataset, with 61 peaks receiving estimates of ν below -1.618.

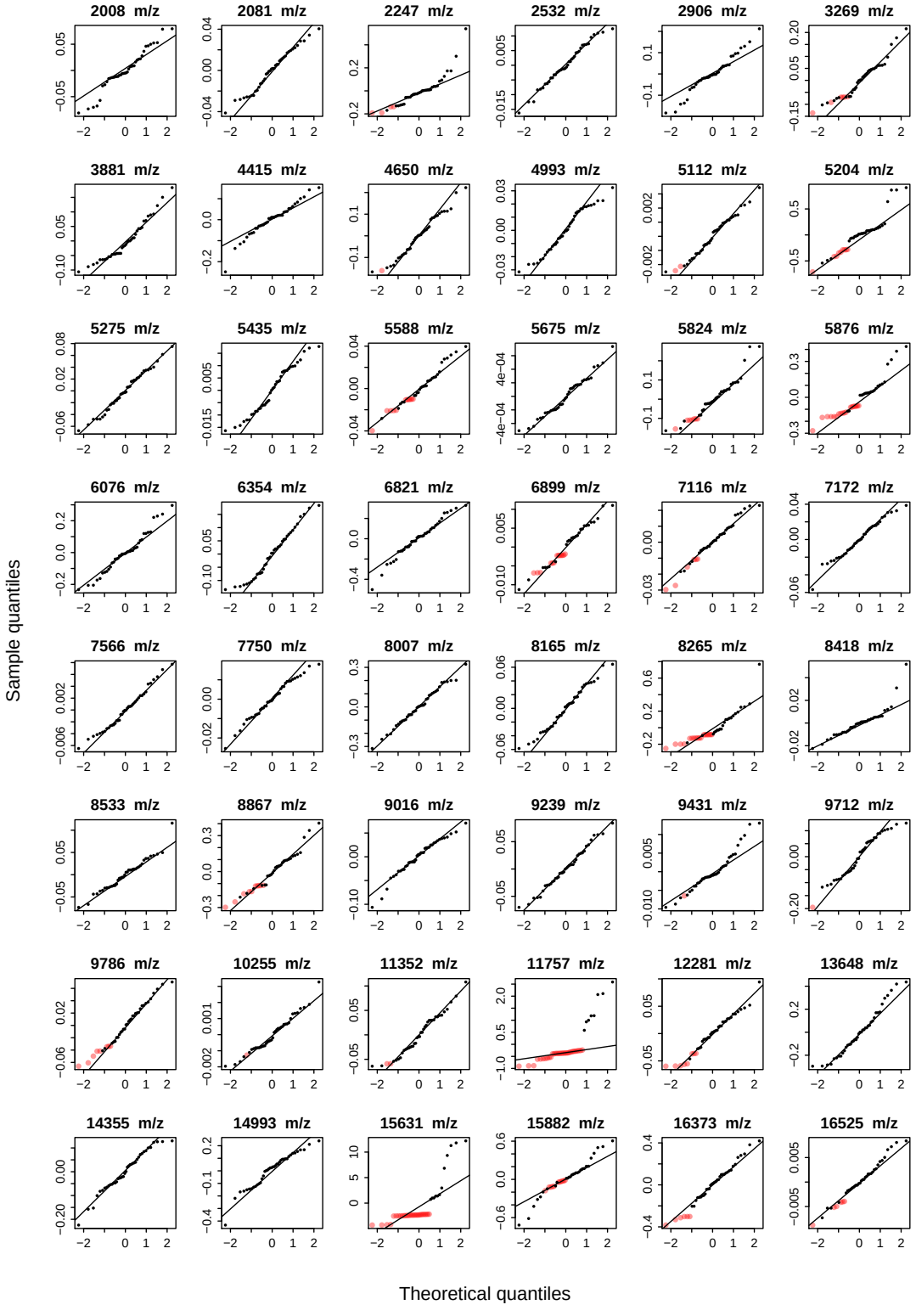

Figure 4.7: Quantile-quantile plots of random effects vector $\boldsymbol{N}$ for the mouse effect for a representative subset of peaks. Points on the plots represent one of the 40 individual vector elements (corresponding to one of the 40 mice) for a particular peak. Points are rendered as translucent red disks if all 27 of the samples from the corresponding mouse are missing.

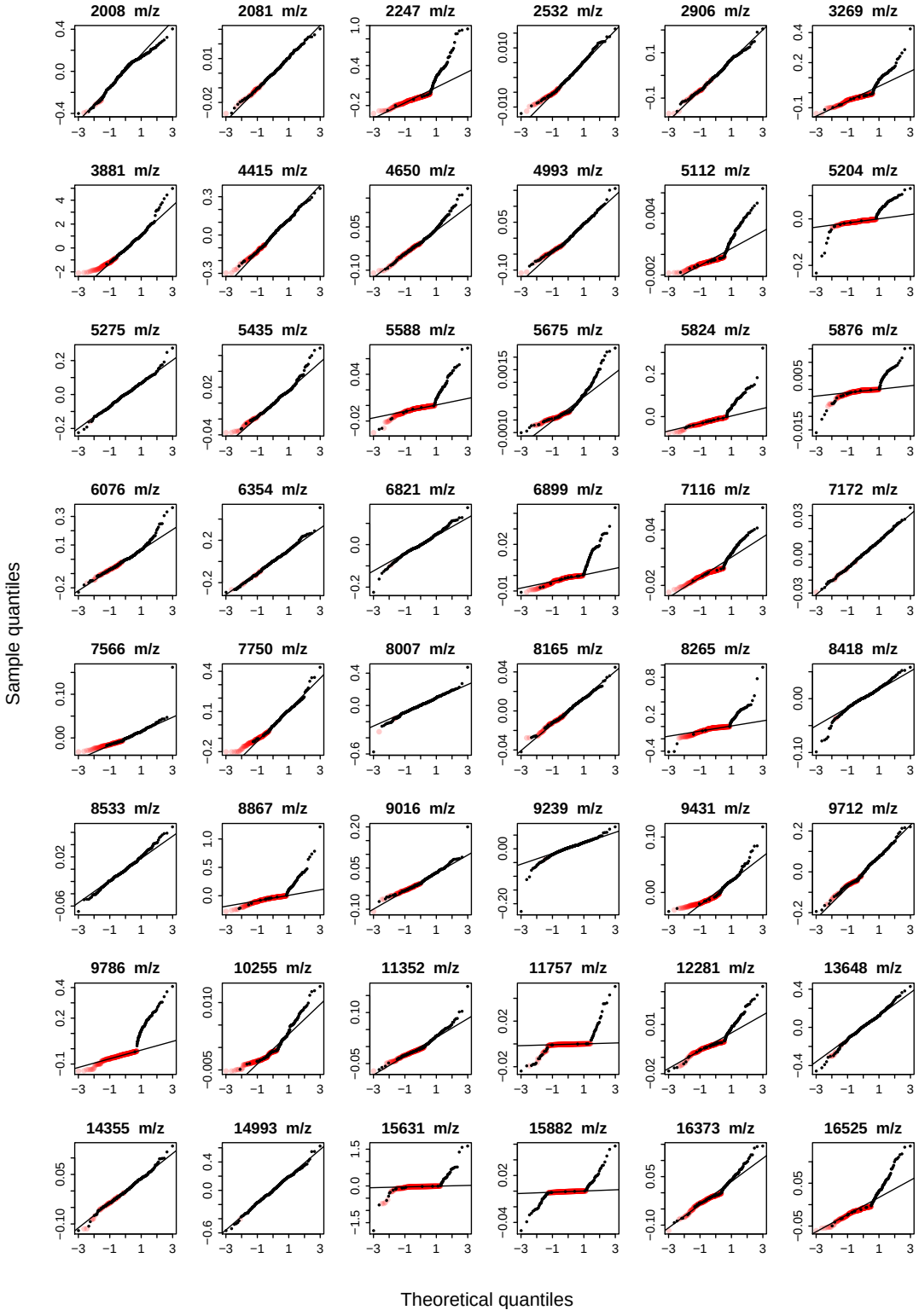

Figure 4.8: Quantile-quantile plots of random effects vector $\boldsymbol{B}$ for the C8 batch effect for a representative subset of peaks. Points on the plots represent one of the 360 individual vector elements (corresponding to one of the 360 C8 batches) for a particular peak. Points are rendered as partially translucent red disks if all three of the samples from the corresponding C8 batch are missing.

This corresponds to a reference category intensity below the minimum observed value in the GC dataset (which was 6.966).

The distributions of the estimates of the group and chip missingness parameters α_j and β_ℓ are roughly symmetric and tend to be centred at values ranging from -1 to 0, with the β_ℓ estimates centred on the lower end of this range. These results are consistent with the low average estimates of β_2 and β_3 from the missingness modelling as discussed in Section 3.5.1 in Chapter 3.

The distributions of the estimates of μ and ω are centred at negative values, positively skewed, and clearly not normal. The shapes and locations of these distributions may be understood in the context of the distribution of the estimates of ν. As discussed in Section 4.4.4, negative values of μ correspond to low missingness specifically for observations from the reference category with an underlying intensity not more than the grand mean of the GC dataset. However, many observations have intensities below that value (reflected by the fact that the median estimate of ν is -1.28). Lower estimated values of ω mean that intensities below the grand mean have greater log-odds of missingness.

No estimate of ω is greater than zero due to the prior distribution assumed for that parameter.

Figure 4.12 displays correlation plots of ω with the other parameters. Most correlations are small, but the correlation between the estimates of μ and ω is around 0.767. As discussed in Section 4.4.4, low values of ω permit low values of μ even in peaks with low missingness proportions. There appear to be two clusters in the plot of μ against ω. These clusters arise from the fact that peaks below 4866 m/z have their intensities bounded below by a greater value than the peaks above 4866 m/z. This is immediately apparent from Figure 2.11. The majority of the peaks below 4866 m/z have low estimates of ω alongside relatively high estimates of μ. This reflects the fact that when intensity observations are greater than the grand mean of the GC dataset, a low value of ω does not necessarily cause the missingness probability of such observations to be high. The same group of peaks also exhibits relatively high estimates of ν, the fixed effect for the intensity of samples from group 1 and chip 1, alongside estimates of ω at or below -8. These relatively high estimates somewhat mask the positive correlation between the estimates of ω and ν, a correlation expected to be present because of the informative missingness affecting the GC dataset.

The estimates of the variance components σ_N^2 , σ_B^2 , and σ^2 are positively skewed because variance estimates are necessarily greater than zero and because the estimates are generally low. As Figure 4.11 demonstrates, the estimates of parameters corresponding to lower levels of the hierarchy tend to be greater than for parameters corresponding to higher levels, meaning that there is the most variation in intensity between samples within C8 batches, then between C8 batches, then between mice. However, the presence of outliers in the parameter estimates causes the means and standard deviations in Table 4.9 to follow the opposite order, with those of σ_N^2 being highest and those of σ^2 being lowest.

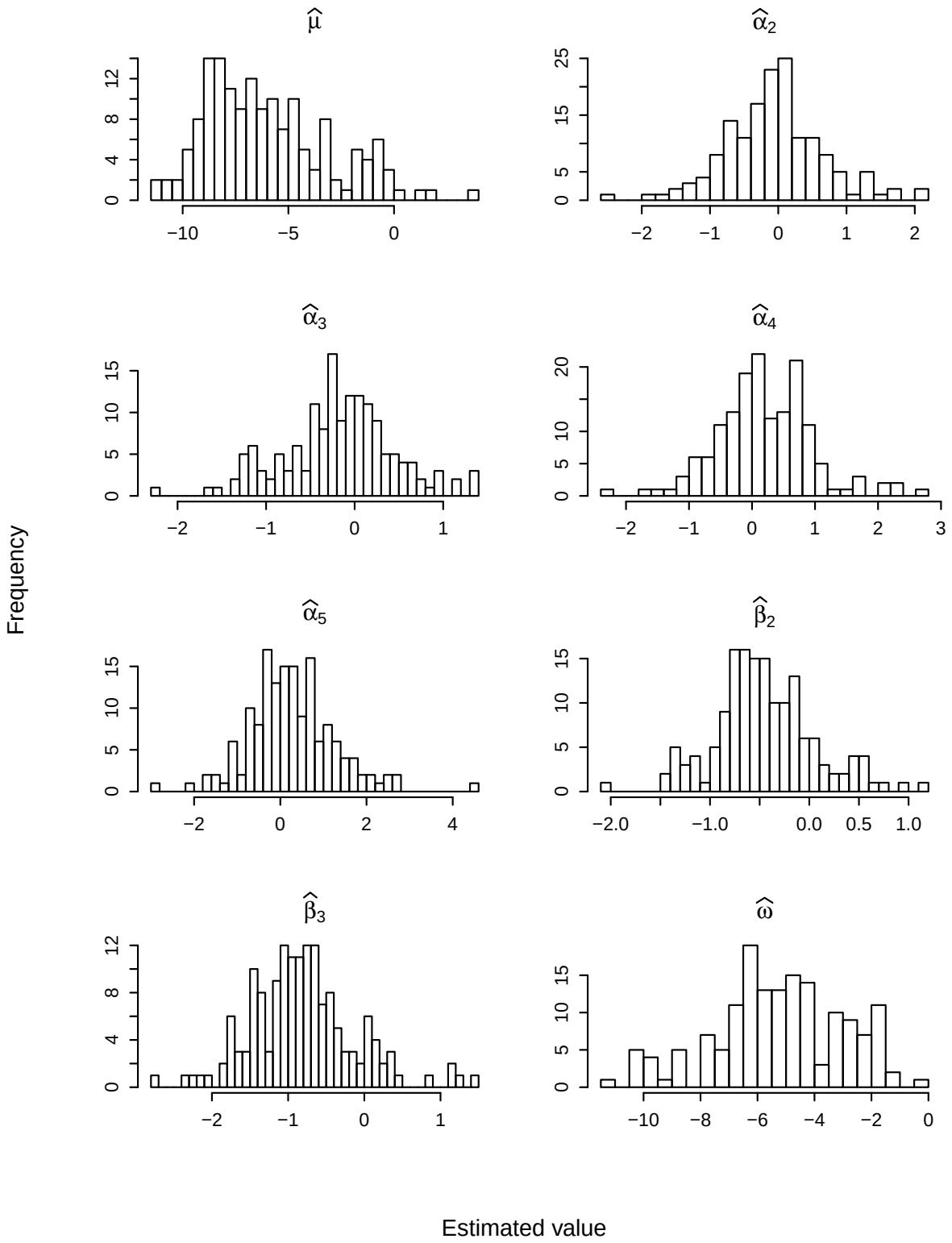

Figure 4.9: Parameter estimates for missingness from the NMAR joint model for the 156-peak subset.

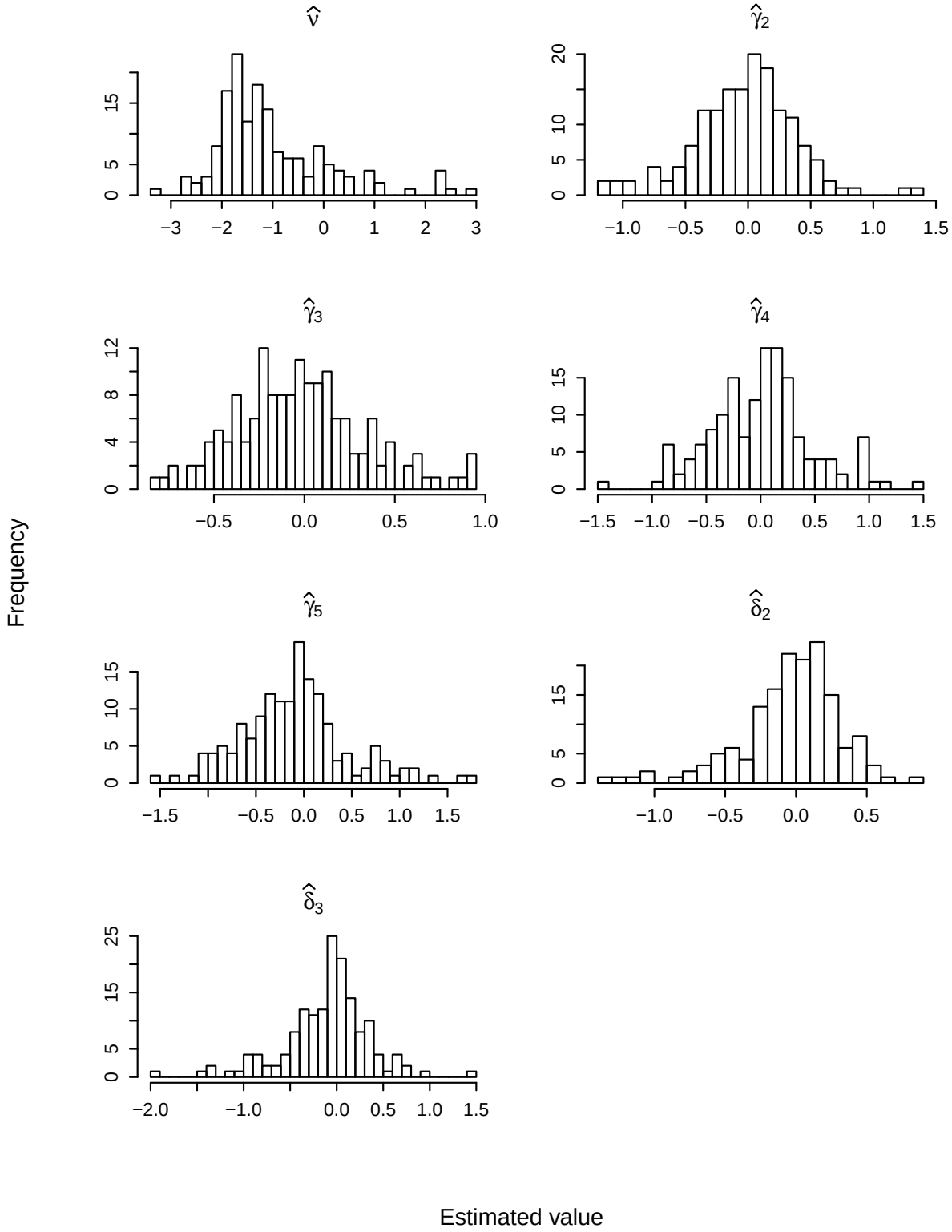

Figure 4.10: Estimates of intensity fixed effects parameter estimates from the NMAR joint model for the 156-peak subset.

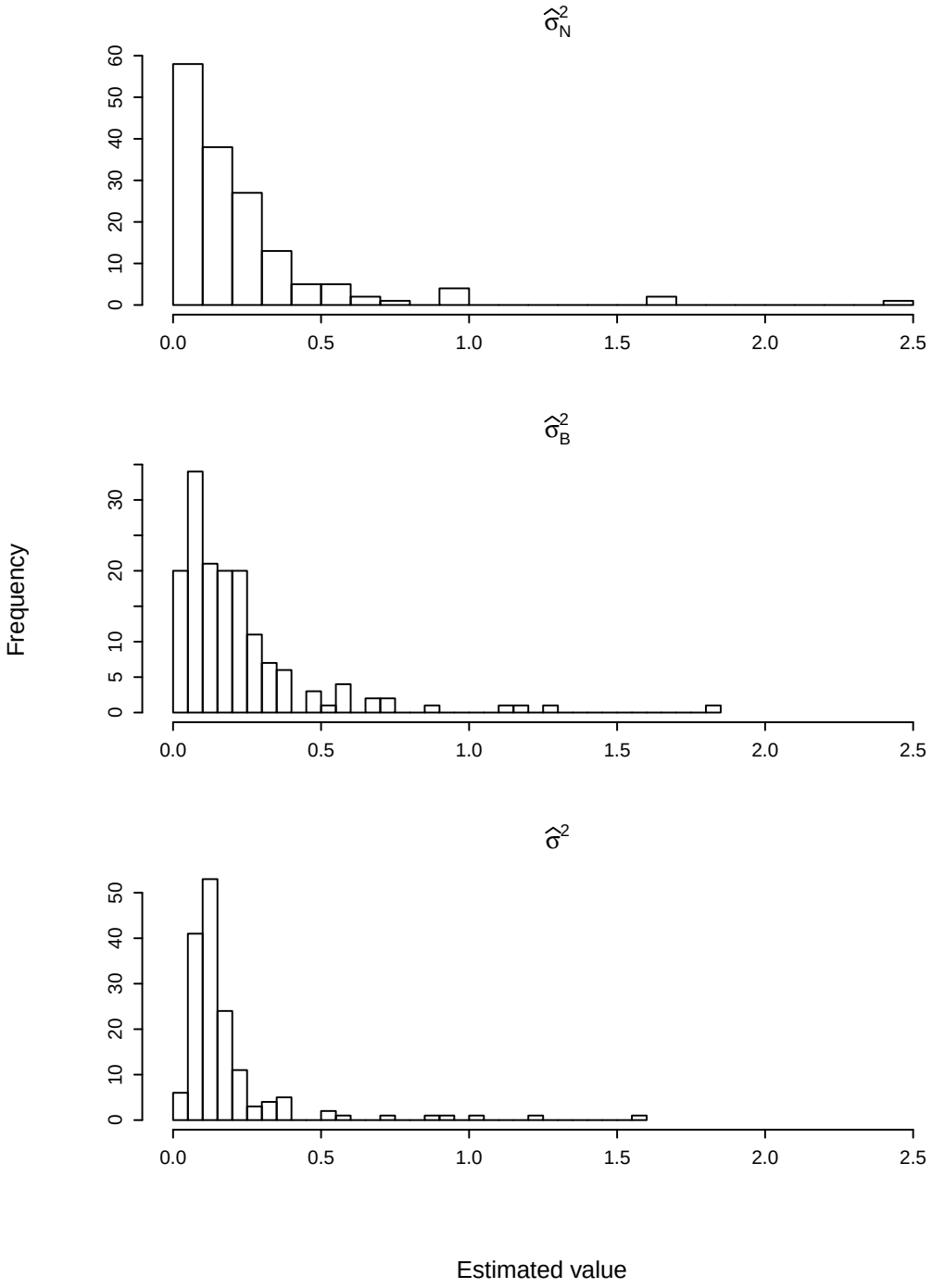

Figure 4.11: Estimates of variance components for intensity from the NMAR joint model for the 156-peak subset.

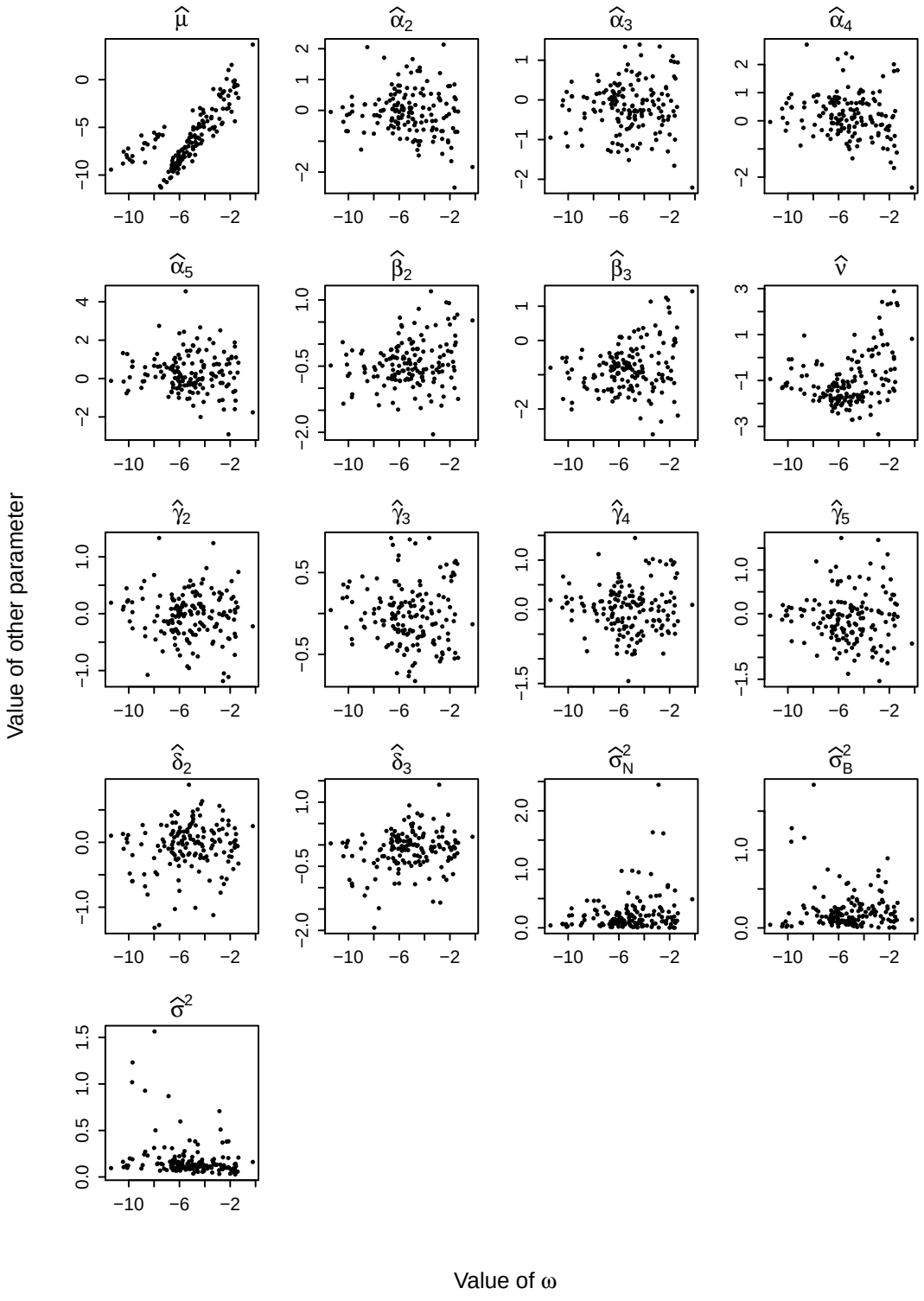

Figure 4.12: Correlations between ω and other parameters. Many pairs show little correlation, except for the pairs involving μ and ν, which show an appreciable positive correlation. The apparent discontinuity in the values of μ near ω = -8, and to a lesser extent in the values of ν, arises from the peaks below 4866 m/z systematically having a greater minimum intensity among their observations.

Table 4.9: Summary of means and standard deviations of all parameter estimates from the NMAR joint model on the 156-peak subset.

Parameter	Mean	S.D.
μ	-5.970	2.949
α_2	-0.060	0.725
α_3	-0.173	0.626
α_4	0.189	0.766
α_5	0.281	1.001
β_2	-0.445	0.509
β_3	-0.807	0.691
ν	-1.013	1.119
γ_2	-0.027	0.406
γ_3	-0.075	0.517
γ_4	0.010	0.467
γ_5	-0.102	0.564
δ_2	-0.046	0.359
δ_3	-0.105	0.459
σ_N^2	0.229	0.309
σ_B^2	0.220	0.248
σ^2	0.183	0.204
ω	-5.228	2.270

Table 4.10: Peaks with the 16 most extreme estimates of ν, the intensity fixed effect reference category parameter. S.E. was estimated via overlapping batch means.

Peak m/z	$\hat{\nu}$	S.E.	z statistic
6788	-2.293	0.480	-4.779
7490	2.314	0.405	5.718
8265	-2.702	0.530	-5.099
8831	2.279	0.352	6.482
9059	-2.629	0.728	-3.610
9214	2.356	0.363	6.488
9239	2.373	0.319	7.438
11352	-2.340	0.408	-5.734
11757	-2.710	0.978	-2.771
14993	2.423	0.401	6.034
15500	-2.175	0.355	-6.118
15515	-2.274	0.513	-4.429
15631	-3.342	1.019	-3.28
15844	2.883	0.383	7.528
16492	-2.493	0.433	-5.758
17976	-2.430	0.525	-4.628

Tables 4.10 to 4.14 display the sixteen peaks with the most extreme estimates of the intercept and group intensity fixed effect parameters. Results for these individual parameters are presented on the basis of observed extreme estimates instead of statistical significance. This is due to the fact that the overlapping batch means estimates of the standard errors of the parameter estimates[1] were often conservative, making it difficult to present results on the basis of statistical significance. For example, the standard Type I error rate of $\alpha = 0.05$ simultaneously caused most estimates of ν to be considered statistically significant while estimates of parameters such as γ_2 or γ_3 were not significant for any of the peaks after adjusting for the 156 simultaneous comparisons using false discovery rate correction (Benjamini and Hochberg, 1995).

Tables 4.12 and 4.14 (corresponding to the parameters γ_3 and γ_5) contain seven and four peaks, respectively, that are deemed to be of primary interest as biomarkers according to the contrast analysis described below, and the other tables (corresponding to ν, γ_2, and γ_4) contain one or zero such peaks.

[1] See Appendix C.2.1 for details.

Table 4.11: Peaks with the 16 most extreme estimates of γ_2, the intensity fixed effect corresponding to the FFIL6 group. S.E. was estimated via overlapping batch means.

Peak m/z	$\hat{\gamma}_2$	S.E.	z statistic
2262	0.680	0.384	1.770
4152	-1.076	0.522	-2.061
4168	1.330	0.367	3.630
4993	-0.663	0.212	-3.120
5824	-0.734	0.306	-2.397
5978	-0.926	0.286	-3.232
7490	-1.113	0.497	-2.240
8302	-1.050	0.448	-2.341
8337	1.240	0.485	2.559
8831	0.733	0.423	1.732
9712	0.802	0.416	1.926
11687	-0.752	0.391	-1.925
11757	-0.701	0.690	-1.015
12161	-0.957	0.420	-2.277
14993	-1.185	0.464	-2.553
15844	-0.725	0.466	-1.555

Table 4.12: Peaks with the 16 most extreme estimates of γ_3, the intensity fixed effect corresponding to the FFStat3 group. S.E. was estimated via overlapping batch means.

Peak m/z	$\hat{\gamma}_3$	S.E.	z statistic
5453	-0.730	0.246	-2.968
6602	-0.607	0.267	-2.279
7412	0.905	0.350	2.585
8533	0.624	0.383	1.630
8607	0.857	0.469	1.826
8970	0.653	0.267	2.447
9239	0.606	0.379	1.598
9760	0.642	0.213	3.009
11352	0.839	0.369	2.274
11757	-0.639	0.502	-1.274
12161	-0.765	0.386	-1.981
13648	-0.826	0.501	-1.648
14421	-0.712	0.284	-2.506
14836	0.922	0.375	2.458
17458	0.923	0.477	1.935
17976	0.709	0.431	1.644

Table 4.13: Peaks with the 16 most extreme estimates of γ_4, the intensity fixed effect corresponding to the IL6 group. S.E. was estimated via overlapping batch means.

Peak m/z	$\hat{\gamma}_4$	S.E.	z statistic
4168	1.121	0.389	2.882
4607	0.911	0.270	3.372
5978	-0.864	0.312	-2.769
6076	-0.898	0.431	-2.081
8302	-0.895	0.458	-1.954
8337	1.021	0.490	2.082
8533	0.953	0.422	2.256
8607	1.447	0.564	2.566
8867	0.977	0.419	2.334
9214	0.998	0.496	2.009
9712	0.989	0.431	2.292
9736	0.913	0.418	2.186
11757	-0.911	0.766	-1.188
12161	-1.450	0.477	-3.037
13648	-0.895	0.585	-1.530
17458	0.937	0.550	1.703

Table 4.14: Peaks with the 16 most extreme estimates of γ_5, the intensity fixed effect corresponding to the wildtype group. S.E. was estimated via overlapping batch means.

Peak m/z	$\hat{\gamma}_5$	S.E.	z statistic
3959	1.196	0.619	1.932
5204	1.147	0.391	2.935
6076	-1.026	0.433	-2.368
7806	1.731	0.870	1.988
7854	-1.136	0.622	-1.827
8337	-0.996	0.635	-1.568
8441	1.055	0.239	4.417
9305	-1.542	0.353	-4.368
9319	1.357	0.529	2.565
11509	-1.072	0.328	-3.269
12161	-1.377	0.511	-2.697
13648	-1.075	0.636	-1.691
14836	0.971	0.403	2.411
14993	-1.054	0.539	-1.955
15631	1.686	0.649	2.598
16654	1.073	0.299	3.591

4.5.2 Parameter contrast

Identification of biomarker candidates using the NMAR joint model (4.9) is done by defining a contrast U_o between the cancer groups 1 (FF) and 2 (FFIL6) and the non-cancer groups 3 (FFStat3), 4 (IL6), and 5 (Wildtype) using the parameters that determine the mean of the intensity vector $\boldsymbol{y}$.

The contrast is formulated with respect to the group means, derived from the model parameters. Suppose the mean of group 1 is denoted $\nu_1^* = \nu$ and the mean of group j is denoted $\nu_j^* = \nu + \gamma_j$ for $j = 2, \ldots, 5$. A contrast between the mean intensity of the non-cancer groups and the cancer groups is

$$U_o = \frac{1}{2}(\nu_1^* + \nu_2^*) - \frac{1}{3}(\nu_3^* + \nu_4^* + \nu_5^*). \tag{4.17}$$

Expressed in terms of the original parameters, this becomes the linear combination

$$U_o = \frac{\gamma_2}{2} - \frac{1}{3}(\gamma_3 + \gamma_4 + \gamma_5) \tag{4.18}$$

The chip effects δ_2 and δ_3 were not considered. Biomarker candidates are defined to be those peaks that have statistically significant ($\alpha = 0.05$) z statistic values for the contrast, where the z statistic is defined as the estimate of the contrast divided by its Monte Carlo standard error as estimated via the method of overlapping batch means. A false discovery rate correction with respect to 156 simultaneous comparisons was made (Benjamini and Hochberg, 1995). To obtain the Monte Carlo estimated standard error, the contrast was formed for each set of parameter draws to yield eight chains of contrast estimates, with each chain of length $N_M = 1500$. The overlapping batch means estimates were computed with $b_n = 38$ for each of the eight chains separately, then averaged, and the square root of the result was taken to be the Markov chain standard error of the contrast estimate. Appendix D.2 gives an example of how the contrast z statistic is obtained for a single peak, 7412 m/z.

4.5.3 Peaks of interest

Peaks with a significant mean cancer/non-cancer group difference according to the NMAR joint model are of primary interest as biomarker candidates. Other peaks are of secondary interest as biomarker candidates. One set of peaks of secondary interest is derived from those detected by the intensity models from Stanford (2015). Another consists of peaks which are related to peaks of primary interest owing to multiply-charged protein ions or to adjacent m/z values being derived from a single species of protein. A third consist of peaks which are of interest based on direct investigation of the GC dataset. Peaks of primary interest that were also deemed to be of secondary interest were simply treated as peaks of primary interest.

Table 4.15: Estimates and z statistics of cancer/non-cancer contrast from the NMAR joint model. The p-values are based on a Wald test and are adjusted using a FDR correction for 156 simultaneous comparisons. Peaks are arranged in order of the contrast value. S.E. was estimated via overlapping batch means.

Peak m/z	Contrast	S.E.	z statistic	p-value
7412	-0.888	0.238	-3.735	< 0.001
14836	-0.875	0.242	-3.614	< 0.001
8970	-0.595	0.195	-3.048	0.002
16654	-0.519	0.178	-2.907	0.004
7738	-0.484	0.149	-3.238	0.001
8441	-0.480	0.144	-3.327	0.001
4607	-0.466	0.151	-3.097	0.002
9760	-0.406	0.141	-2.882	0.004
5453	0.509	0.158	3.230	0.001
6602	0.615	0.202	3.052	0.002
7642	0.628	0.180	3.489	< 0.001
9608	0.671	0.177	3.796	< 0.001
9305	0.929	0.222	4.183	< 0.001
14421	0.948	0.201	4.725	< 0.001

Table 4.15 shows the 14 peaks of primary interest as obtained from the NMAR joint model. Table 4.16 displays the peaks of secondary interest as biomarker candidates alongside the peaks of primary interest. The reasoning behind the inclusion of each peak as a secondary biomarker candidate is discussed below.

LMM and GLMM modelling

Table 4.17 shows the sets of biomarker candidates from the NMAR joint model alongside the sets from the LMM (4.1) of Stanford (2015) for the intensity, and the Bayesian missingness model (4.4) investigated in Chapter 3. The set of peaks of interest discovered using the joint modelling approach in this chapter, compared to the initial set of Stanford (2015) in Table 3.13, reveals the extra information in the GC dataset captured when accounting for the NMAR nature of the data. In particular, the peaks at 5453, 7642, 7738, 8441, 8970, 9608, and 16654 m/z are peaks missed by the LMM, flagged by the GLMM, and deemed significant by the NMAR joint model. What these peaks have in common is that within each peak, the differences between the between-group averages of the observed data are not large, but groups with low values have many more missing values than groups with higher values. This implies that for these groups, the *true* expression value averages are likely to differ by more than what is observed, and this is precisely the situation in which joint modelling is expected to detect group differences in mean intensity.

Table 4.16: Peaks of primary and secondary interest as candidate biomarkers, listed by m/z. An asterisk * denotes candidate biomarkers from Stanford (2015).

Primary	Secondary
4607	3959
5453	4358
6602*	4617
7412*	4650
7642	5204
7738	5876
8441	6076
8970	6821*
9305*	6899
9608	6989
9760	7204
14421*	7750
14836*	7806*
16654	7917
	8302
	8337*
	8533*
	8607*
	8831*
	8867*
	9059
	9214
	9239
	9319
	11352
	11757
	12161*
	13648*
	15631
	16030*
	17458*
	17976

Table 4.17: Display of peaks of interest for separate models and NMAR joint model.

Peak m/z	Missingness GLMM	Stanford LMM	NMAR Joint model
3246	Y		
3881	Y		
4168	Y		
4607			Y
4650	Y		
4866	Y		
5248	Y		
5373	Y		
5453	Y		Y
5637	Y		
5675	Y		
6076	Y		
6258	Y		
6541	Y		
6602		Y	Y
6821		Y	
6858	Y		
6879	Y		
6899	Y		
6989	Y		
7087	Y		
7146	Y		
7204	Y		
7412	Y	Y	Y
7642	Y		Y
7738	Y		Y
7750	Y		
7806		Y	
8067	Y		
8265	Y		
8337		Y	
8441	Y		Y
8505	Y		
8533		Y	
8607	Y	Y	
8831		Y	
8867	Y	Y	
8970	Y		Y
9305	Y	Y	Y
9431	Y		
9608	Y		Y
9760			Y
9795	Y		
9821	Y		
10255	Y		
11352	Y		
12161	Y	Y	
13648		Y	
14421	Y	Y	Y
14836	Y	Y	Y
15724	Y		
16030		Y	
16654	Y		Y
17458		Y	
17976	Y		

There is some overlap between the set of peaks from the LMM of Stanford (2015) and the set of peaks from the joint model, but the two sets are different. This difference may be due to overly-conservative overlapping batch means estimates of the contrast variance that result in a large number of false negatives. Peaks deemed interesting in the prior work of Stanford (2015) but not by the NMAR joint model are of secondary interest as biomarker candidates. These are the peaks which are marked with an asterisk in the second column of Table 4.16.

The set of peaks of interest from the Bayesian missingness model (4.4) is extensive. About half of the peaks deemed interesting using that model are also of interest according either to the LMM (4.1) or the NMAR joint model (4.9). Peaks are not taken to be of secondary interest as biomarker candidates solely due to membership in this set, as the set covers almost a quarter of the peaks in the GC dataset, which is too broad to be useful. Some peaks are of interest under the LMM or the NMAR joint models but not under the GLMM. In all of these cases, group differences in intensity are apparent in the data, but either the missingness proportions in the groups are low (increasing the standard errors of the GLMM parameter estimates) or the missingness differs between groups in a way not aligned with the contrast. For example, in the peak at 9670 m/z, group 5 has high missingness while groups 3 and 4 have low missingness relative to group 1.

The peaks flagged by the missingness GLMM but not by either the NMAR joint model or the work of Stanford (2015) commonly exhibit group differences in missingness probability with little apparent true difference in cancer versus non-cancer group intensities. In contrast to this are the peaks at 4650, 6076, 6899, 6989, 11352, and 17976 m/z, which appear to exhibit true (albeit slight) differences in mean intensity between cancer and non-cancer groups based on direct inspections of their expression values. These peaks are therefore also of secondary interest as biomarkers.

The LMM work did not flag any peaks below 6602 m/z, even though there are peaks in the range from approximately 3200 to 6600 m/z with group differences in missingness according to the Bayesian missingness model. Peaks at low m/z may correspond to double-charged proteins whose singly-charged versions produce high m/z peaks.

The peaks at 4607 and 9760 m/z were deemed interesting by the joint modelling approach but not by either of the separate models. The experimental determination of the corresponding proteins as true candidate biomarkers would validate the joint modelling approach.

Multiply-charged protein ions

Recall from Section 2.2.4 that some pairs of peaks are thought to correspond to the same protein. Identification of one peak in such a pair implicates the other peak of the pair as a candidate biomarker, because it is the underlying proteins that are of interest. One such peak pair listed in Table 2.4, 7412 and 14836 m/z, is such that both peaks in the pair are listed as peaks of interest according to the NMAR joint model. Additional peak pairs

at 4607 and 9214 m/z, at 4607 and 9239 m/z, at 9305 and 4650 m/z, and at 14421 and 7204 m/z have the former, but not the latter peak, listed as a peak of primary interest. This means that the peaks at 4650, 7204, 9214, and 9239 m/z are of secondary interest as biomarker candidates, because these peaks are likely to correspond to doubly-charged versions of the protein corresponding to the other peak in the pair.

Additional investigations of the GC dataset

Recall from Section 2.2.2 that four pairs of adjacent peaks were speculated to be derived from a single species of protein ion being assigned to two adjacent m/z values. These pairs were those at 4607 and 4617 m/z, 7738 and 7750 m/z, 8302 and 8337 m/z, and 9305 and 9319 m/z. Of these pairs, the first, second, and fourth pair are such that one peak in the pair is of primary interest according to the NMAR joint model. Therefore, the other peaks in these pairs are of secondary interest as a biomarker, yielding the peaks at 4617, 7750, and 9319 m/z as peaks of secondary interest. The third pair of peaks contains the peak at 8337 m/z which is of secondary interest as a biomarker according to the LMM, meaning that the corresponding peak at 8302 m/z is also of secondary interest as a biomarker. The peak pair of 4152 and 4168 m/z appears to be an additional pair derived from a single protein ion being assigned to two adjacent m/z values. However, mean intensity differences between groups do not appear to be very large in the combined measurements for the two peaks.

As discussed in Section 4.4.1, the peak at 7806 m/z presented difficulties in obtaining the same parameter estimates when fitting the LMMs in R and `stan`. This peak is anomalous in that most intensity observations are between 7 and 9 except for some observations, belonging to particular mice, which are much higher. A single mouse in group 2, the FFIL6 group, produced samples with intensities up to 11 for that peak. Multiple mice in group 5, the wildtype group, produce samples with intensities up to 13, driving up the average of that group. These anomalies imply that the assumption of normally distributed between-mouse random effects on the intensity does not hold, accounting for the difficulties in estimating the parameter σ_N^2 which is the variance of the mouse effect in the LMM. Due to the intensity difference between the wildtype group and the remaining groups, the peak at 7806 m/z is of secondary interest as a biomarker.

A number of peaks have very high proportions of missingness, which decreases the precision of intensity parameter estimates, and consequently, the constrast z statistic, but the observations that are present imply that there is a difference in intensity between cancer and non-cancer groups due to the fact that groups with many missing observations are likely to have low true means. Such peaks are those at 5204, 5876, 9059, 11757, and 15631 m/z. These peaks are of secondary interest as biomarkers. However, it is possible that individual mice had outlying average expressions of the corresponding proteins in their blood, in which case the pattern of data does not reflect a biomarker for cancer.

There were three peaks for which either the `stan` model was not fitted, or the NMAR

joint model results were discarded due to the diagnostic $\hat{R}$ statistic values indicating a lack of Markov chain convergence. In the latter case, large proportions of missing values hampered the convergence. The data from some of these peaks indicates that differences in protein intensity between cancer and non-cancer groups are present. The peak at 7917 m/z is a clear example of a peak with large group differences, and the peak at 4358 m/z also demonstrates slight group differences in intensity. These two peaks are therefore of secondary interest as biomarkers. The peaks at 3959 and 7917 m/z additionally form a pair corresponding to singly and doubly-charged ions of the same protein, so the peak at 3959 m/z is of secondary interest as a biomarker.

4.6 Summary

Joint observed/missing data models allow for modelling the NMAR nature of the GC dataset by the specification of a joint distribution over the data ($\boldsymbol{Y}$, $\boldsymbol{R}$) from a single peak. A selection model factorisation is used in which the distribution is expressed as the product of the conditional distribution of $\boldsymbol{R}$ given $\boldsymbol{Y}$ and the marginal distribution of $\boldsymbol{Y}$. The explicit dependence of $\boldsymbol{R}$ on $\boldsymbol{Y}$ is where the assumption of NMAR data enters the model. A NMAR joint model is developed by building on simpler models. These simpler models include those developed in Chapter 3 for the missingness, the LMM of Stanford (2015), and a joint model that assumes MAR data.

The NMAR joint model yields an expression for the likelihood that is not amenable to analytic solutions or simpler computational approximations used for the simpler models. Consequently, parameter estimates are obtained using the `stan` MCMC approach. The approach introduces additional considerations about long-term Markov chain behaviour that must be accounted for in order for the parameter estimates to be trustworthy. Using `stan` to fit simpler models informed the use of `stan` to fit the more complex NMAR joint models. Additional model checks such as MCMC statistics and data simulations provide greater confidence in the reliability of `stan` estimates.

The MAR joint model served as a basis for comparison to the NMAR joint model. Parameter estimates differ between the two models, even though the NMAR joint model has only one additional parameter relative to the MAR joint model. This implies that the missingness mechanism is strongly dependent on the peak intensity, which is in line with prior expectations given the limitations of MS technology.

The suitability of the NMAR joint model for the GC dataset was investigated using simulation checks and predicted data. The NMAR joint model was found to be suitable for modelling the data, and yields reasonable estimates of the parameters of interest. However, estimated standard errors of the parameters are more difficult to obtain, which complicates null hypothesis significance testing.

The set of peaks of interest as candidate biomarkers produced by the NMAR joint model differs from the sets obtained from the separate models of the missingness and the

intensity. This has demonstrated that accounting for the missingness mechanism extracts further information from the data than modelling the observed responses alone. If the set of peaks is borne out by future biological research, then the joint modelling approach will be validated. However, research to date has focused on proteins of greater mass than those in the GC dataset.

Chapter 5

Conclusions

5.1 Summary of thesis

Outcome-dependent missingness mechanisms are a key issue in proteomic MS datasets. Informative missingness in data biases statistical inference performed on the data, hampering discovery of biomarkers needed for potentially life-saving disease diagnosis. Existing methods for dealing with missingness are largely imputation methods, which are better suited for data that are missing at random. A joint modelling approach that explicitly accounts for NMAR data from proteomic MS studies is able to debias statistical inference and increase the amount of information that can be extracted from proteomic MS datasets.

Chapter 2 describes in detail the GC dataset with which this thesis is concerned, and provides many visualisations of the data. The GC dataset consists of MALDI-TOF MS-derived measurements of peak intensity (corresponding to concentration) over a range of m/z values associated with the atomic masses of proteins in the blood of mice. Analysis of the GC dataset is complicated by the hierarchical structure of the dataset. The GC dataset is affected by missingness in a way typical to proteomic MS data. That is, a high proportion of values are missing, and while the precise missingness mechanism is not obvious from the data alone, it is clear that missingness is associated with low intensity values.

In Chapter 3, the statistical framework for modelling the missingness in the GC data is introduced. The best model for the missingness accounts for the hierarchical structure of the GC dataset using two random effects terms. A set of peaks of secondary interest as candidate biomarkers is produced using parameter contrasts estimated from the models. The set of peaks deemed interesting by the models for the missingness differs from the set of peaks deemed interesting in prior work on the GC dataset performed by Stanford (2015), in which models were fitted to the intensity values. Due to the informative nature of the missingness, this difference between the lists of peaks indicates that the missingness

indicators carry information about protein biomarkers that is not detected in modelling the intensities alone.

In Chapter 4, the NMAR joint model for the GC dataset is developed and fitted using Markov chain Monte Carlo. The formulation of the joint model is informed by the formulation of the model for the intensities from Stanford (2015). The joint model uses random effects terms only for modelling the intensity, and the random effects that are used are the same as those used in the prior model for the intensities. Random effects for the missingness are not used because missingness acts as a proxy for low intensities, and by explicitly incorporating the intensity as a term affecting the probability of missingness, random effects terms that affect the probability are redundant. The method of parameter contrasts for finding peaks of interest as candidate biomarkers is repeated for the joint model to obtain peaks of primary interest as biomarker candidates. The findings are that while the set of peaks deemed interesting by the NMAR joint model has many peaks in common with the set derived from the previous modelling for the intensities, the additional peaks picked up by the joint model but not the intensity model are of biological interest, validating the joint modelling approach.

5.2 Directions for future work

The research presented in this thesis lends itself to many extensions.

The joint model is a Bayesian model, meaning that prior distributions for parameters in the model must be specified. The prior distributions used were estimated from the data using simpler models. However, many other choices of priors are justifiable. On one hand, weakly informative priors for all parameters are suitable for the purpose of regularisation. On the other hand, outcome-dependent missingness mechanisms call for application-specific models that are informed by experimental and procedural knowledge in order to specify the missingness mechanism as precisely as possible. A sensitivity analysis of prior distributions (possibly based on simulated data) is desirable.

Alternate model specifications for the GC dataset, such as a mixture model, are worth investigating. One possibility for a mixture model is to use truncated normal distributions for the missing and observed data, where the threshold of truncation is the same for both kinds of data and the probability density functions have complementary support. This is a first step that does not account for additional sources of missingness beyond the threshold cut-off. Additional selection models are also worth investigating. In all joint models investigated in this thesis, the probability of missingness was specified as a logit function of the linear predictor. The logit function has horizontal asymptotes at zero and one. A modified logit function with asymptotes at other values within (0, 1) can model a plausible missingness mechanism in which the probability of missingness of an observation does not fall to zero or rise to one as the linear predictor becomes extreme.

Models were fitted to single peaks independently. Not investigated in the modelling

were correspondences between adjacent peaks, or between peaks with m/z values close to a 1:2 or 1:3 ratio. This represents information in the GC dataset that was left unused. The information may be incorporated by consolidating and averaging the values in related peaks and fitting the existing models again, or by specification of a model fitted to multiple peaks' outcome variables simultaneously.

Simulations on constructed data are useful for validating model-fitting procedures. There is room for improvement in the construction of the data to make it even more faithful to the experimental MS setup and data pre-processing pipeline, such as the addition of pairs of peaks corresponding to single proteins. By explicitly declaring a subset of peaks as corresponding to protein biomarkers based on true underlying parameter values before performing simulations, models may be assessed in terms of false positives and false negatives in biomarker detection.

Application of the joint modelling method to other proteomic datasets is an obvious extension. The `stan` platform is easily used for models of the type fitted in this thesis. However, the application-specific nature of missingness mechanisms for proteomic data must be kept in consideration. MALDI-TOF MS data may likely be fitted by adapting the NMAR joint model used in this thesis, but data from other technologies such as liquid chromatography MS, or 2D gel electrophoresis, requires careful thought and investigation to specify the missingness mechanisms and the prior distributions of parameters in the relevant models.

Appendix A

List of peaks

Table A.1: The set of all 159 peaks with missingness count and subset inclusion. The 152-peak subset is the subset of peaks considered in the interpretation of the results of the Bayesian missingness model (3.3). The 156-peak subset is the subset considered in the interpretation of the results of the NMAR joint model (4.9).

Peak m/z	Missing observations	152-peak subset	156-peak subset
2008	98	Y	Y
2033	105	Y	Y
2057	195	Y	Y
2081	295	Y	Y
2104	601	Y	Y
2128	820	Y	Y
2247	922	Y	Y
2262	955	Y	Y
2478	905	Y	Y
2504	669	Y	Y
2532	463	Y	Y
2576	832	Y	Y
2793	828	Y	Y
2906	267	Y	Y
2948	912	Y	Y
3246	563	Y	Y
3269	889	Y	Y
3493	162	Y	Y
3769	969	Y	Y
3881	499	Y	Y
3959	352	Y	Y
4152	763	Y	Y
4168	813	Y	Y
4358	0		
4415	412	Y	Y
4607	352	Y	Y

4617	622	Y	Y
4650	421	Y	Y
4751	935	Y	Y
4866	767	Y	Y
4993	370	Y	Y
5027	781	Y	Y
5059	96	Y	Y
5112	901	Y	Y
5123	844	Y	Y
5147	153	Y	Y
5189	889	Y	Y
5204	918	Y	Y
5236	558	Y	Y
5248	952	Y	Y
5275	46	Y	Y
5335	250	Y	Y
5373	926	Y	Y
5435	362	Y	Y
5453	596	Y	Y
5557	541	Y	Y
5588	967	Y	Y
5590	834	Y	Y
5617	868	Y	Y
5637	946	Y	Y
5675	750	Y	Y
5752	29	Y	Y
5809	564	Y	Y
5824	889	Y	Y
5848	883	Y	Y
5854	945	Y	Y
5876	934	Y	Y
5972	835	Y	
5978	807	Y	Y
6000	866	Y	Y
6076	365	Y	Y
6193	679	Y	Y
6258	784	Y	Y
6354	110	Y	Y
6541	634	Y	Y
6602	72	Y	Y
6788	919	Y	Y
6821	8	Y	Y
6858	709	Y	Y
6879	789	Y	Y
6899	968	Y	Y
6989	145	Y	Y
7087	936	Y	Y
7116	866	Y	Y
7126	928	Y	Y

7146	580	Y	Y
7172	143	Y	Y
7204	760	Y	Y
7412	195	Y	Y
7490	2		Y
7566	660	Y	Y
7642	331	Y	Y
7738	589	Y	Y
7750	526	Y	Y
7806	388		Y
7854	259	Y	Y
7917	1		Y
8007	24	Y	Y
8067	787	Y	Y
8118	563	Y	Y
8165	415	Y	Y
8178	551	Y	Y
8228	827	Y	Y
8251	782	Y	Y
8265	963	Y	Y
8302	463	Y	Y
8337	481	Y	Y
8418	47	Y	Y
8441	738	Y	Y
8505	523	Y	Y
8533	18	Y	Y
8607	218	Y	Y
8831	27	Y	Y
8867	953	Y	Y
8970	775	Y	Y
8981	795	Y	Y
9016	559	Y	Y
9059	969	Y	Y
9124	187	Y	Y
9214	77	Y	Y
9239	29		Y
9305	222	Y	Y
9319	830		Y
9431	826	Y	Y
9490	683	Y	Y
9608	388	Y	Y
9712	419		Y
9736	136	Y	Y
9760	552	Y	Y
9786	940	Y	Y
9795	743	Y	Y
9821	969	Y	Y
10230	701	Y	Y
10255	777	Y	Y

10872	89	Y	Y
11120	455	Y	Y
11352	675	Y	Y
11509	20	Y	Y
11687	883	Y	Y
11757	951	Y	Y
11855	904	Y	Y
12161	207	Y	Y
12281	842	Y	Y
12391	615	Y	Y
13190	841	Y	Y
13231	813	Y	Y
13648	92	Y	Y
13987	90	Y	Y
14243	836	Y	Y
14355	271	Y	Y
14421	462	Y	Y
14836	351	Y	Y
14993	5	Y	Y
15500	919	Y	Y
15515	960	Y	Y
15631	947	Y	Y
15724	251	Y	Y
15759	535	Y	Y
15844	8	Y	Y
15882	968	Y	Y
16030	36	Y	Y
16353	837	Y	Y
16373	735	Y	Y
16492	755	Y	Y
16505	588	Y	Y
16525	888	Y	Y
16654	738	Y	Y
17458	21	Y	Y
17976	916	Y	Y

Appendix B

Separation of data

Separation of data in statistical models refers to situations in which the data are in such a configuration that finite maximum likelihood estimates of model parameters fail to exist (Albert and Anderson, 1984; Heinze and Schemper, 2002). Data that are not in a state of separation are said to be in a state of *overlap*. Logistic regression models fitted to binary outcome data are prone to the issue of separation, although the issue is not limited to such cases.

When data are separated, the likelihood function of the data is not maximised at any finite value of the parameter set. Instead, the likelihood asymptotically approaches its maximum at a point at infinity. Computer estimates of parameters in the case of separated data tend to take extreme values with accompanying extreme standard error estimates due to the flatness of the likelihood function at extreme parameter values (Heinze and Schemper, 2002).

An example clarifies the meaning of separated data. Suppose a simple binary logistic regression model with one continuous predictor variable X is fitted to observations of a binary outcome R. This model is written

$$E[R_i] = p_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}\,, \quad \eta_i = \beta_0 + \beta_1 x_i\,.$$

The parameters β_0 and β_1 are to be estimated. Figure B.1 displays data in states of separation and overlap with this model fitted to the data in both cases. In the case of separated data, all observations for which $x_i > 24$ satisfy $r_i = 1$, and all observations for which $x_i \leq 24$ satisfy $r_i = 0$. This means that the likelihood of the data is maximised by allowing the fitted probabilities to equal 1 if $x_i > 24$, and to equal 0 if $x_i \leq 24$. Achievement of these fitted probabilities in likelihood maximisation routines requires β_0 and β_1 to approach infinity. When predictor variables are categorical instead of continuous, separation occurs when all observations of the response within a single category of a predictor variable have identical values.

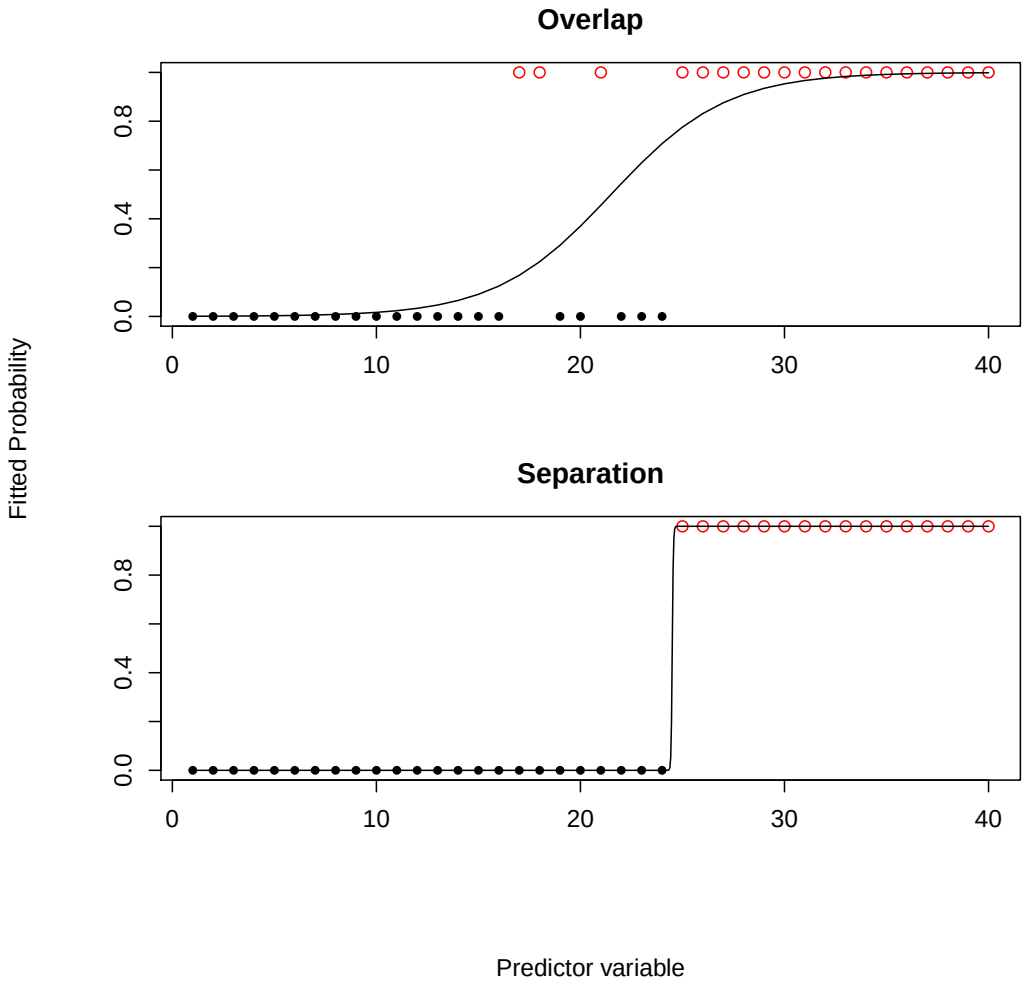

Figure B.1: Data that are in a state of overlap (top) and separation (bottom). The x axis represents values of a lone predictor variable and the y axis represents fitted probability values from a model fitted to binary outcome data. Data are overlaid as black dots for outcomes of 0 and red circles for outcomes of 1.

Appendix C

Estimation of model parameters

C.1 Mixed effects models

In mixed effects models, likelihood maximisation concerns a subset of the parameters in the model, and is not performed over the random effects vector elements (Pinheiro and Bates, 2000; Snijders and Bosker, 2012). The notions of the *profile likelihood* and the *marginal likelihood* are used in maximising the likelihood function over a subset of parameters (Barndorff-Nielsen, 1991). Suppose the set of parameters $\boldsymbol{\theta} \in \Theta = \Omega \times \Lambda$ may be partitioned as $\boldsymbol{\theta} = (\boldsymbol{\omega}, \boldsymbol{\lambda})$ with $\boldsymbol{\omega} \in \Omega$ and $\boldsymbol{\lambda} \in \Lambda$. If $L(\boldsymbol{\theta})$ is the likelihood for $\boldsymbol{\theta}$, then the *profile likelihood* for $\boldsymbol{\omega}$ is

$$L_p(\boldsymbol{\omega}) = \text{argmax}_{\boldsymbol{\omega},\, \boldsymbol{\lambda} \in \Lambda} L(\boldsymbol{\theta}) \tag{C.1}$$

With $\boldsymbol{\theta}$ partitioned as $(\boldsymbol{\omega}, \boldsymbol{\lambda})$, the maximum likelihood estimate of $\boldsymbol{\lambda}$ for a given value of $\boldsymbol{\omega}$ is $\hat{\boldsymbol{\lambda}}_{\boldsymbol{\omega}}$ and the profile likelihood conditional on that value of $\boldsymbol{\omega}$ is expressible as

$$L_p(\boldsymbol{\omega}) = L(\boldsymbol{\omega}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\omega}})$$

Alternatively, suppose that $\boldsymbol{Y}$ depends on both $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$, and $\boldsymbol{\lambda}$ depends in turn on $\boldsymbol{\omega}$. The statistical model for the data may be expressed as

$$f(\boldsymbol{y}\,; \boldsymbol{\omega}, \boldsymbol{\lambda}) = f(\boldsymbol{y} \mid \boldsymbol{\lambda}\,; \boldsymbol{\omega}) f(\boldsymbol{\lambda}\,; \boldsymbol{\omega})$$

The parameters $\boldsymbol{\omega}$ are of interest, whereas the parameters $\boldsymbol{\lambda}$ are *nuisance parameters* which must be accounted for in modelling but whose values are not of interest. If the likelihood for $\boldsymbol{\theta}$ is $L(\boldsymbol{\theta}) = f(\boldsymbol{y}\,; \boldsymbol{\lambda}, \boldsymbol{\omega})$, then the *marginal likelihood* for $\boldsymbol{\omega}$ is

$$L(\boldsymbol{\omega}) = \int_{\Lambda} f(\boldsymbol{y} \mid \boldsymbol{\lambda}\,; \boldsymbol{\omega}) f(\boldsymbol{\lambda}\,; \boldsymbol{\omega}) \, \mathrm{d}\, \boldsymbol{\lambda} \tag{C.2}$$

We consider a mixed effects model expressed in matrix form

$$\boldsymbol{Y} = X\boldsymbol{\beta} + Z\boldsymbol{b} + \boldsymbol{\varepsilon} \tag{C.3}$$

In this model, $\boldsymbol{Y}$ is an n-vector of observations, X is the $n \times p$ fixed effects design matrix, and $\boldsymbol{\beta}$ is the p-vector of fixed effects. The matrix Z is an $n \times q$ random effects design matrix, and $\boldsymbol{b}$ is a q-vector of random effects with a multivariate normal distribution centred at zero. Finally, $\boldsymbol{\varepsilon}$ is the n-vector of residual errors whose elements are independently and identically distributed as $\varepsilon_i \sim N(0,\, \sigma^2)$. The random effects vector and the residual error vector are independent.

The conditional distribution of $\boldsymbol{Y}$ given the random effects is

$$(\boldsymbol{Y} \mid \boldsymbol{b}) \sim N_n(X\boldsymbol{\beta} + Z\boldsymbol{b},\, \sigma^2 I)$$

and the distribution of the random effects vector $\boldsymbol{b}$ is

$$\boldsymbol{b} \sim N_q(\boldsymbol{0},\, \sigma^2 \Sigma)$$

for some symmetric, positive-semidefinite matrix Σ. The unconditional distribution of $\boldsymbol{Y}$ is then

$$\boldsymbol{Y} \sim N_n(X\boldsymbol{\beta},\, H), \quad H = \sigma^2(I_n + Z\Sigma Z^T)$$

The matrix Σ is expressible as $\Delta\Delta^T$ for some upper-triangular matrix Δ. If $\boldsymbol{u} \sim N_q(\boldsymbol{0},\, \sigma^2 I)$, then the vector $\Delta^T\boldsymbol{u}$ has the same distribution as $\boldsymbol{b}$. The mixed-effects model (C.3) is equivalent to

$$\boldsymbol{Y} = X\boldsymbol{\beta} + Z\Delta^T\boldsymbol{u} + \boldsymbol{\varepsilon}$$

The vector $\boldsymbol{u}$ is the *spherical random effects vector*.

The matrix Δ is known as the *relative covariance factor* of the random effects. This matrix may be parametrised by a set of *covariance parameters*, collectively denoted as ϕ. The relative covariance factor is then written Δ_ϕ. The dimension of ϕ is typically much lower than q, the length of the random effects vector (Bates et al., 2015).

The conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{u}$ is

$$(\boldsymbol{Y} \mid \boldsymbol{u}) \sim N_n(X\boldsymbol{\beta} + Z\Delta^T\boldsymbol{u},\, \sigma^2 I)$$

and the corresponding probability density function is

$$f(\boldsymbol{y} \mid \boldsymbol{u}) = (2\pi\sigma^2)^{-n/2}\exp\left(-\|\boldsymbol{y} - X\boldsymbol{\beta} - Z\Delta^T\boldsymbol{u}\|^2/(2\sigma^2)\right) \tag{C.4}$$

The probability density of $\boldsymbol{u}$ is

$$f(\boldsymbol{u}) = (2\pi\sigma^2)^{-q/2}\exp\left(-\|\boldsymbol{u}\|^2/(2\sigma^2)\right) \tag{C.5}$$

The joint probability density function of $\boldsymbol{Y}$ and $\boldsymbol{u}$ may then be expressed as the product of Equations (C.4) and (C.5)

$$f(\boldsymbol{y}) = (2\pi\sigma^2)^{-(n+q)/2}\exp\left(-\|\tilde{\boldsymbol{y}} - \tilde{X}\boldsymbol{\beta} - \tilde{Z}\boldsymbol{u}\|^2/(2\sigma^2)\right), \tag{C.6}$$

where

$$\tilde{\boldsymbol{y}} = (\boldsymbol{y}^T, 0, \ldots, 0)^T$$

is a $(n+q)$-vector,

$$\tilde{X} = \begin{bmatrix} X \\ 0 \end{bmatrix}$$

is a $(n+q) \times p$ matrix, and

$$\tilde{Z} = \begin{bmatrix} Z\Delta^T \\ I_q \end{bmatrix}$$

is a $(n+q) \times q$ matrix.

This is the *pseudo-data* approach used by the `lme4` package in `R` to fit linear mixed-effects models (Bates et al., 2015). The name arises from the fact that the condition on the random effects—that the spherical random effects vector is normally distributed with zero mean and variance $\sigma^2 I_q$—is incorporated by adding q additional 'observations' to $\boldsymbol{y}$, augmenting the design matrices accordingly.

The parameters of interest in a linear mixed-effects model are the fixed effects $\boldsymbol{\beta}$, the residual error variance σ^2, and the covariance parameters ϕ. The random effects vector is considered to be an unobservable set of nuisance parameters. The likelihood of data under the linear mixed-effects model (C.3) is therefore taken to be a marginal likelihood as defined in Equation (C.2) with parameters of interest $\boldsymbol{\omega} = (\boldsymbol{\beta}, \sigma^2, \phi)$ and the spherical random effects vector as the nuisance parameters, i.e. $\boldsymbol{\lambda} = \boldsymbol{u}$ (Pinheiro and Bates, 2000). The likelihood of the data is then

$$L(\boldsymbol{\beta}, \sigma^2, \phi \mid \boldsymbol{y}) = \int_{\mathbb{R}^q} (2\pi\sigma^2)^{-(n+q)/2} \exp\left(-\|\tilde{\boldsymbol{y}} - \tilde{X}\boldsymbol{\beta} - \tilde{Z}\boldsymbol{u}\|^2/(2\sigma^2)\right) \mathrm{d}\,\boldsymbol{u} \tag{C.7}$$

Maximising the likelihood (C.7) as a function of $\boldsymbol{\beta}$, σ^2, and ϕ is prohibitive in terms of computational time, as it requires finding the optimal point in a space whose dimension encompasses all three of $\boldsymbol{\beta}$, σ^2, and ϕ. The likelihood, in practice, is profiled to be a function of ϕ only, and the values of $\boldsymbol{\beta}$ and σ^2 are therefore determined by ϕ.

The exponent in the integrand of the likelihood (C.7) may be expressed as

$$r^2(\phi, \boldsymbol{\beta}, \boldsymbol{u}) = \|\tilde{\boldsymbol{y}} - \tilde{X}\boldsymbol{\beta} - \tilde{Z}\boldsymbol{u}\|^2 = \left\| \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} Z\Delta_\phi & X \\ I & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\beta} \end{bmatrix} \right\|^2$$

which constitutes a least-squares problem for the vector $(\boldsymbol{u}^T, \boldsymbol{\beta}^T)^T$. Maximising the likelihood therefore involves finding the solution $(\hat{\boldsymbol{u}}^T, \hat{\boldsymbol{\beta}}^T)^T$, which depends on ϕ through the $Z\Delta_\phi$ submatrix.

With the spherical random effects vector marginalised out of the likelihood (C.7), the log-likelihood to be maximised may be expressed as

$$\ell(\boldsymbol{\beta}, \sigma^2, \phi \mid \boldsymbol{y}) = -\ln|L_\phi| - \frac{n}{2}\ln(2\pi\sigma^2) - \frac{r^2(\phi)}{2\sigma^2} - \frac{\|R_X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|^2}{2\sigma^2} \tag{C.8}$$

where $r^2(\phi) = r^2(\phi, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}})$, and the matrices L_ϕ and R_X arise from the Cholesky decomposition

$$\begin{bmatrix} \Delta_\phi^T Z^T Z \Delta_\phi + I & \Delta_\phi^T Z^T X \\ X^T Z \Delta_\phi & X^T X \end{bmatrix} = \begin{bmatrix} L_\phi & 0 \\ R_{ZX}^T & R_X^T \end{bmatrix} \begin{bmatrix} L_\phi^T & R_{ZX} \\ 0 & R_X \end{bmatrix}$$

Fixing the value of ϕ, the values of $\boldsymbol{\beta}$ and σ^2 that maximise the log-likelihood (C.8) are $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2 = n^{-1} r^2(\phi)$. Substituting these values into the log-likelihood yields the profiled log-likelihood

$$\ell(\phi \mid \boldsymbol{y}) = -\ln |L_\phi| - \frac{n}{2}\left[1 + \ln(2\pi r^2(\phi)) - \ln(n)\right] \tag{C.9}$$

The maximum likelihood estimates of the parameters ϕ, $\boldsymbol{\beta}$, and σ^2 are found by optimising (C.9) in ϕ to obtain $\hat{\phi}$, then evaluating $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ at $\hat{\phi}$.

The values of the random effects vectors $\boldsymbol{b}$ and $\boldsymbol{u}$ are not estimated through the procedure outlined here due to the marginalisation of the random effects. However, it is possible to predict the values using the *best linear unbiased predictors* (BLUPs) (Searle et al., 2009) which are the conditional modes of the random effects given the observed data. The predictor of the spherical random effects vector is given by

$$\tilde{\boldsymbol{u}} = (L_{\hat{\phi}}^{-1})^T L_{\hat{\phi}}^{-1} \Delta_{\hat{\phi}}^T Z (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})$$

C.1.1 Restricted maximum likelihood

An undesirable property of the maximum likelihood estimates for a linear mixed-effects model is that the estimate $\hat{\sigma}^2$ of the residual error σ^2 is biased, underestimating the residual error by a factor of $n/(n-p)$. In order to correct this bias, parameter estimates may be chosen to maximise an alternative quantity, derived from the likelihood, called the *restricted maximum likelihood* criterion (REML criterion). The REML estimate of σ^2 corrects for the loss in degrees of freedom incurred by estimating the $\boldsymbol{\beta}$ parameters (Corbeil and Searle, 1976; Patterson and Thompson, 1971).

The REML criterion is obtained by integrating the density of $\boldsymbol{y}$ after marginalising out the random effects vector. The integral is

$$\begin{aligned} \int_{\mathbb{R}^p} L(\boldsymbol{\beta}, \sigma^2, \phi \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\beta} &= |L_\phi|^{-1} (2\pi\sigma^2)^{-n/2} \exp\left(-r^2(\phi)/(2\sigma^2)\right) \\ &\quad \cdot \int_{\mathbb{R}^p} \exp\left(-(\|R_X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|^2)/(2\sigma^2)\right) \mathrm{d}\boldsymbol{\beta}. \end{aligned}$$

This integral may be simplified to yield

$$\begin{aligned} L_R(\sigma^2, \phi \mid \boldsymbol{y}) &= \int_{\mathbb{R}^p} L(\boldsymbol{\beta}, \sigma^2, \phi \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\beta}, \\ &= |L_\phi|^{-1} (2\pi\sigma^2)^{-(n-p)/2} \exp\left(-r^2(\phi)/(2\sigma^2)\right) |R_X|^{-1} \end{aligned}$$

The restricted maximum likelihood estimates of σ^2 and ϕ are those that maximise the log-restricted likelihood

$$\ell_R(\sigma^2,\, \phi \mid \boldsymbol{y}) = -\ln|L_\phi| - \frac{n-p}{2}\ln(2\pi\sigma^2) - \frac{r^2(\phi)}{2\sigma^2} - \ln|R_X| \tag{C.10}$$

The log-restricted likelihood is profiled to be a function of ϕ with the substitution $\hat{\sigma}^2 = (n-p)^{-1}r^2(\phi)$ for σ^2. This yields the profiled log-restricted likelihood

$$\ell_R(\phi \mid \boldsymbol{y}) = -\ln|L_\phi| - \ln|R_X| - \frac{n-p}{2}\big(1 + \ln(2\pi r^2(\phi)) - \ln(n-p)\big) \tag{C.11}$$

Maximisation of Equation (C.11) yields estimates of ϕ and σ^2, with the latter estimate corrected for the bias. The estimate of ϕ is, in general, different to the estimate of ϕ obtained via maximising Equation (C.9) as opposed to (C.11). Strictly speaking, the REML method does not provide estimates of the parameter vector of fixed effects $\boldsymbol{\beta}$ because the fixed effects are marginalised out of the likelihood. In practice, the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ at the REML estimate of ϕ may be used (Bates et al., 2015; Pinheiro and Bates, 2000).

C.1.2 Generalised linear mixed models

The parameters which are to be estimated in a generalised linear mixed model consist of the fixed effects $\boldsymbol{\beta}$ and the set of parameters ϕ which determine the distribution of the random effects. Again, obtaining maximum likelihood estimates of $\boldsymbol{\beta}$ and ϕ requires marginalising the random effects terms out of the likelihood. Using the notation for exponential family distributions, the likelihood of observed data $\boldsymbol{y}$ under a generalised linear mixed model with canonical link is expressed as the marginal likelihood

$$\begin{aligned} L(\boldsymbol{\beta},\, \phi\,; \boldsymbol{y}) \;\; = \;\; & \int_{\mathbb{R}^q} \frac{1}{(2\pi)^{q/2}|\Sigma|^{1/2}} \exp\left(\sum_{i=1}^{n} y_i\eta_i - \kappa(\eta_i) + h(y_i)\right) \\ & \cdot\exp\left(-(\boldsymbol{b}^T\Sigma^{-1}\boldsymbol{b})/2\right)\, \mathrm{d}\,\boldsymbol{b}, \end{aligned} \tag{C.12}$$

where

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{i,\,j} + \sum_{\ell=1}^{q} b_\ell z_{i,\,\ell}$$

is the linear predictor for the ith observation of Y and the functions κ and h are particular to the exponential family in use.

Unlike in the theory for linear mixed effects models, the likelihood (C.12) cannot be evaluated exactly in most cases—although exceptions exist if the density of $\boldsymbol{b}$ is conjugate to that of $\boldsymbol{y}$ (Lee and Nelder, 1996)—and it is necessary to approximate the integral in

the likelihood. Multiple methods exist, and a well-performing method is adaptive Gauss-Hermite quadrature, of which the Laplace approximation is a special case (Raudenbush et al., 2000).

After the profiled likelihood is approximated, the likelihood is optimised as a function of both ϕ and $\boldsymbol{\beta}$. The `glmer` function in `lme4` uses the BOBYQA and Nelder-Mead optimisers to achieve this (R Core Team, 2016).

C.2 Markov chain Monte Carlo for more general modelling

Suppose data $\boldsymbol{y}$ are assumed to follow a distribution $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ and the aim is to infer the true values of $\boldsymbol{\theta} = (\theta_1 , \ldots , \theta_q)$. If maximum likelihood approaches do not yield workable estimates of $\boldsymbol{\theta}$, an alternative is to sample a sequence of values

$$(\boldsymbol{\theta}_n)_{n=1}^{N_M} = (\boldsymbol{\theta}_1 , \boldsymbol{\theta}_2 , \ldots , \boldsymbol{\theta}_{N_M})$$

from the distribution of $\boldsymbol{\theta}$ where N_M is the *Monte Carlo sample size* and $\boldsymbol{\theta}_n = (\theta_{n,1} , \ldots , \theta_{n,q})$. With such a sequence, estimates of the parameters $\theta_1 , \ldots , \theta_q$ are then given by

$$\hat{E}[\theta_j] = \sum_{n=1}^{N_M} I_j(\boldsymbol{\theta}_n), \tag{C.13}$$

where j ranges from 1 to q and $I_j(\boldsymbol{\theta}) = \theta_j$ (Flegal and Jones, 2011). Ideally, the samples in the sequence $(\boldsymbol{\theta}_n)_{n=1}^{N_M}$ are independent, but it is still possible to apply Equation (C.13) when the samples are dependent samples from a Markov chain (Nummelin, 2002).

The basic idea of MCMC was first developed by Metropolis et al. (1953) to compute integrals over high-dimensional spaces, and expanded upon by Hastings (1970). Given a set of parameters $\boldsymbol{\theta}_n$, the next set of parameters $\boldsymbol{\theta}_{n+1}$ is chosen by the following method. First, a *candidate value* $\boldsymbol{\theta}^*$ is chosen by sampling from a *proposal distribution* $g(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_n)$, and second, the candidate value is used for the value of $\boldsymbol{\theta}_{n+1}$ with probability

$$\alpha(\boldsymbol{\theta}_n , \boldsymbol{\theta}^*) = \min \left(1, \frac{f(\boldsymbol{\theta}^*) g(\boldsymbol{\theta}_n \mid \boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}_n) g(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_n)} \right),$$

where $f(\boldsymbol{\theta})$ is the distribution function of $\boldsymbol{\theta}$. If the candidate value is not used, then $\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n$. With a suitable proposal distribution, the method ensures that the sequence of $\boldsymbol{\theta}$ values is sampled from a Markov chain whose stationary distribution is equal to f.

The particular choice of acceptance probability α is the reason for the chain having stationary distribution f. The procedure may be seen to work using an ensemble argument (Metropolis et al., 1953). For the sake of clarity, assume that the proposal distribution is symmetric, so

$$g(\boldsymbol{\theta}_n \mid \boldsymbol{\theta}^*) = g(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_n)$$

and consequently

$$g(\boldsymbol{\theta}_n \mid \boldsymbol{\theta}^*)/g(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_n) = 1.$$

Suppose the probability density at $\boldsymbol{\theta}^*$ is greater than the probability density at $\boldsymbol{\theta}'$. Then, starting from $\boldsymbol{\theta}'$, the new value of θ will be equal to $\boldsymbol{\theta}^*$ with probability 1. The reverse motion, that of beginning from $\boldsymbol{\theta}^*$, with the candidate value $\boldsymbol{\theta}'$ having been proposed, and ending by accepting the candidate value, occurs with probability $\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}') < 1$. Let us consider the hypothetical that a vast number of such Markov chains, identical, are constantly evolving and are in equilibrium, meaning that the distribution of present states across all chains reflects the stationary distribution. The quantity of chains moving from $\boldsymbol{\theta}'$ to $\boldsymbol{\theta}^*$ is balanced by the quantity of chains moving from the value $\boldsymbol{\theta}^*$ to $\boldsymbol{\theta}'$ because of the equilibrium of all of the Markov chains. The amount of movement in either direction is the same and the probability of movement from one direction to the other is not, meaning that the number of Markov chains presently making the reverse motion (i.e. starting from $\boldsymbol{\theta}^*$) must be greater than the number of Markov chains presently making the forward motion (from $\boldsymbol{\theta}'$) in order to ensure that the equilibrium is upheld. The ratio of the number of forward-moving chains to the number of reverse-moving chains is in fact precisely $\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}')$. This means that the ratio of time spent by any one instance of such a Markov chain in the state $\boldsymbol{\theta}'$ to the time spent at $\boldsymbol{\theta}^*$ is

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = f(\boldsymbol{\theta}')/f(\boldsymbol{\theta}^*).$$

If $\boldsymbol{\theta}^*$ is the parameter value at which the density function is maximised, then it follows that the stationary distribution of the Markov chain is f.

In a Bayesian setting, the joint distribution $f(\boldsymbol{\theta})$ of the parameters is the posterior distribution given the data, i.e.,

$$f(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\theta}) f(\boldsymbol{y} \mid \boldsymbol{\theta})$$

for some prior distribution $\pi(\boldsymbol{\theta})$. It suffices to set $f(\boldsymbol{\theta})$ equal to the unnormalised function $\pi(\boldsymbol{\theta}) f(\boldsymbol{y} \mid \boldsymbol{\theta})$.

In practice, initial values $\boldsymbol{\theta}_0$ are chosen and the Markov chain is simulated for a number of *burn-in* iterations, after which the Markov chain samples are assumed to be samples from the stationary distribution. Often, multiple chains are run at the same time, each with their own distinct $\boldsymbol{\theta}_0$ values. The burn-in samples are discarded and only the remaining samples are considered for estimation according to Equation (C.13).

There is much freedom in the choice of proposal distribution q. Two pitfalls must be avoided. The first is that if the proposal distribution is too conservative, picking new values $\boldsymbol{\theta}^*$ close to $\boldsymbol{\theta}_n$, the Markov chain explores the state space too slowly, possibly not exploring the entire state space. A rough diagnostic of this pitfall is that candidate values are accepted more than half of the time. On the other hand, the second pitfall is that the proposal distribution is too liberal, picking values $\boldsymbol{\theta}^*$ which have far lower probability

than $\boldsymbol{\theta}_n$. A rough diagnostic of this trap is that candidate values are accepted very rarely, one percent of the time or less. For a large class of problems, candidate values are ideally accepted between 15 and 50 percent of the time (Gilks, 2005).

C.2.1 Precision of parameter point estimates

A point estimate of the mean of a parameter θ using Equation (C.13) varies about the true value with an unknown *Monte Carlo error*. This error is equal to $\hat{E}[\theta] - E[\theta]$, where $E[\theta]$ is the expectation of θ with respect to its stationary distribution. If a Markov chain central limit theorem holds, then

$$\sqrt{N_M}(\hat{E}[\theta] - E[\theta])$$

converges in distribution to $N(0, \sigma^2)$ for some $\sigma^2 \in (0, \infty)$ (Flegal and Jones, 2011). The value σ^2 is generally not equal to the variance of θ with respect to its stationary distribution. With an estimator $\hat{\sigma}^2$ of σ^2 , the *Monte Carlo standard error* $\hat{\sigma}/\sqrt{N_M}$ which approximates the standard deviation of the Monte Carlo error may be reported to indicate the reliability of the parameter estimate and an asymptotically valid confidence interval may also be found for $\hat{E}[\theta]$ as

$$\left(\hat{E}[\theta] - t^* \frac{\hat{\sigma}}{\sqrt{N_M}}, \hat{E}[\theta] + t^* \frac{\hat{\sigma}}{\sqrt{N_M}}\right) \tag{C.14}$$

for some quantile t^*.

The method of *overlapping batch means* (OLBM) is one way of obtaining $\hat{\sigma}^2$ (Flegal and Jones, 2011). In this method, the sequence $(\theta_n)_{n=1}^{N_M}$ is split into batches

$$(\theta_n)_{n=k}^{k+b_n-1} = (\theta_k \, , \, \theta_{k+1} \, , \, \ldots \, , \, \theta_{k+b_n-1} \, , \,)$$

of length b_n indexed by k running from 1 to $N_M - b_n + 1$. Suppose that

$$\bar{\theta}_k(b_n) = \frac{1}{b_n} \sum_{n=0}^{b_n-1} \theta_{k+n}$$

for $k = 0$ ranging up to $k = N_M - b_n + 1$. Then the OLBM estimator of σ^2 is

$$\hat{\sigma}^2_{\mathrm{OLBM}} = \frac{N_M b_n}{(N_M - b_n)(N_M - b_n + 1)} \sum_{n=0}^{N_M - b_n} (\bar{\theta}_k(b_n) - \hat{E}[\theta])^2.$$

If the Markov chain mixes sufficiently quickly and b_n is allowed to grow with N_M , then $\hat{\sigma}^2_{\mathrm{OLBM}}$ is a strongly consistent estimator of σ^2 (Flegal and Jones, 2011). A convenient choice of b_n is $\lfloor \sqrt{N_M} \rfloor$.

When constructing the interval of Equation (C.14), the appropriate quantile t^* is from a t distribution with $N_M - b_n$ degrees of freedom.

When multiple Markov chains were used, as is commonly done in `stan`, σ^2 was estimated by averaging the OLBM estimators of σ^2 obtained from each chain. The standard error σ was estimated by taking the square root of the estimate of σ^2.

C.2.2 Hamiltonian MCMC

A suitable proposal distribution $g(\boldsymbol{\theta})$ is one that permits the Markov chain to explore the parameter space efficiently. Finding such a proposal distribution becomes difficult as the dimension q of the parameter space Θ grows large. One difficulty arises from the fact that in low dimensions, the regions of parameter space near the mode of a distribution provide the main contribution to the location of the mean, but in high dimensions, the region of parameter space influencing the location of the mean is found in the typical set, a set which may be far from the mode (Betancourt, 2017). In high-dimensional spaces, an efficient exploration of the parameter space is almost the same as an efficient exploration of the typical set.

The procedure of *Hamiltonian* MCMC, developed from the method of Hybrid Monte Carlo developed by Duane et al. (1987), is an MCMC method in which new candidate values of the parameters are obtained by computing Hamiltonian dynamics that operate over the parameter space (Neal, 2011).

Hamiltonian dynamics arise in treating the parameter $\boldsymbol{\theta}$ as a q-vector of variables denoting *position* of a particle subject to a *potential energy* field $U(\theta)$. The *momentum* of the particle is then a q-vector of auxiliary variables $\boldsymbol{\phi}$ and the particle carries a *kinetic energy* $K(\boldsymbol{\phi})$. The particle's motion through phase space over time is described by the partial derivatives of the *Hamiltonian* $H(\boldsymbol{\theta}, \boldsymbol{\phi})$ according to

$$\frac{\partial \theta_j}{\partial t} = \frac{\partial H}{\partial \phi_j},$$

$$\frac{\partial \phi_j}{\partial t} = -\frac{\partial H}{\partial \theta_j},$$

for j ranging from 1 to q.

The Hamiltonian function is usually of the form

$$H(\boldsymbol{\theta}, \boldsymbol{\phi}) = U(\boldsymbol{\theta}) + K(\boldsymbol{\phi})$$

and $U(\boldsymbol{\theta})$ is defined as the negative of the log of the probability density of $\boldsymbol{\theta}$. In a Bayesian setting, this probability density is the posterior density of $\boldsymbol{\theta}$. The term $K(\boldsymbol{\phi})$ is usually defined by

$$K(\boldsymbol{\phi}) = \boldsymbol{\phi}^T M^{-1} \boldsymbol{\phi},$$

where M is a $q \times q$ *mass matrix*. This form for the kinetic energy corresponds to a normal distribution with mean zero and variance-covariance matrix M. From this distribution, the value of the momentum vector $\boldsymbol{\phi}$ may be sampled. Thus, at each step of the Markov chain, the current position vector $\boldsymbol{\theta}_n$ and the current, randomly sampled momentum vector $\boldsymbol{\phi}_n$ are input to the Hamiltonian system. Simulating the dynamics of the system for some time yields a new position vector $\boldsymbol{\theta}^*$ which is used as the candidate value of the parameter. The corresponding momentum vector is discarded, as a momentum vector may be sampled anew at every iteration. The Hamiltonian dynamics are commonly simulated via the *leapfrog algorithm* (Betancourt, 2017; Duane et al., 1987).

The form of the mass matrix M, the length of time for which to simulate the Hamiltonian dynamics, and the parameters controlling the leapfrog algorithm need to be determined. These features may be automatically chosen via computer software to ensure maximum efficiency of exploration of the typical set.

Appendix D

Worked examples of contrast

D.1 Missingness model

Consider the peak at 7412 m/z, which was deemed important by the Bayesian missingness model (3.3) on the basis of a statistically significant contrast z statistic. Table D.1 displays the parameter estimates $\hat{\boldsymbol{\lambda}}$ and Table D.2 displays $\widehat{\Sigma}$, the estimated variance-covariance matrix of $\hat{\boldsymbol{\lambda}}$.

The estimated variance of the contrast, $\widehat{\mathrm{Var}}(\widehat{U_m})$, may be found using the result that if a vector $\boldsymbol{X}$ of length p has multivariate normal distribution

$$\boldsymbol{X} \sim N_p(\boldsymbol{\mu},\, \Phi),$$

then $\boldsymbol{b}^T\boldsymbol{X}$, where $\boldsymbol{b}$ is a vector of constants of length p, is normally distributed as

$$\boldsymbol{b}^T\boldsymbol{X} \sim N(\boldsymbol{b}^T\boldsymbol{\mu},\, \boldsymbol{b}^T\Phi\boldsymbol{b}).$$

The vector of parameter estimates $\hat{\boldsymbol{\lambda}}$ is assumed to have the distribution $N_p(\boldsymbol{\lambda},\, \Sigma)$. The contrast (3.5) , which is equivalent to Equation (3.6), may be expressed as $\boldsymbol{b}^T\hat{\boldsymbol{\lambda}}$ where

$$\boldsymbol{b}^T = (0,\, 1/2,\, \text{-}1/3,\, \text{-}1/3,\, \text{-}1/3,\, 0,\, 0).$$

Equivalently, the non-matrix formulation of the variance of the contrast is

$$\begin{aligned}
\widehat{\mathrm{Var}}(\widehat{U_m}) \;=\; & \frac{1}{4}\widehat{\mathrm{Var}}(\widehat{\alpha_2}) \\
& -\frac{1}{3}\left(\widehat{\mathrm{Cov}}(\widehat{\alpha_2}\,,\,\widehat{\alpha_3}) + \widehat{\mathrm{Cov}}(\widehat{\alpha_2}\,,\,\widehat{\alpha_4}) + \widehat{\mathrm{Cov}}(\widehat{\alpha_2}\,,\,\widehat{\alpha_5})\right) \\
& +\frac{2}{9}\left(\widehat{\mathrm{Cov}}(\widehat{\alpha_3}\,,\,\widehat{\alpha_4}) + \widehat{\mathrm{Cov}}(\widehat{\alpha_3}\,,\,\widehat{\alpha_5}) + \widehat{\mathrm{Cov}}(\widehat{\alpha_4}\,,\,\widehat{\alpha_5})\right) \\
& +\frac{1}{9}\left(\widehat{\mathrm{Var}}(\widehat{\alpha_3}) + \widehat{\mathrm{Var}}(\widehat{\alpha_4}) + \widehat{\mathrm{Var}}(\widehat{\alpha_5})\right).
\end{aligned}$$

Table D.1: Estimates of parameters in $\boldsymbol{\lambda}$ from the missingness model on the peak at 7412 m/z.

$\hat{\mu}$	$\widehat{\alpha_2}$	$\widehat{\alpha_3}$	$\widehat{\alpha_4}$	$\widehat{\alpha_5}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$
-1.392	1.180	-3.236	-1.849	-3.065	-0.611	-0.385

Table D.2: Variance-covariance matrix of $\hat{\boldsymbol{\lambda}}$ for the missingness model (4.4) on the peak at 7412 m/z.

	μ	α_2	α_3	α_4	α_5	β_2	β_3
μ	0.351	-0.277	-0.242	-0.286	-0.275	-0.054	-0.055
α_2	-0.277	0.534	0.223	0.293	0.257	-0.007	-0.002
α_3	-0.242	0.223	0.606	0.288	0.316	0.020	0.000
α_4	-0.286	0.293	0.288	0.645	0.312	0.005	-0.003
α_5	-0.275	0.257	0.316	0.312	0.767	0.002	0.011
β_2	-0.054	-0.007	0.020	0.005	0.002	0.116	0.068
β_3	-0.055	-0.002	0.000	-0.003	0.011	0.068	0.123

where $\widehat{\mathrm{Cov}}(\lambda_i\,,\,\lambda_j)$ is the $(i,\,j)$th entry of $\widehat{\Sigma}$. The standard error of the contrast is then $\sqrt{\widehat{\mathrm{Var}}(\widehat{U_m})}$.

For the peak at 7412 m/z, the contrast is

$$\begin{aligned}\widehat{U_m} &= \frac{1}{2}(\widehat{\mu_1}^* + \widehat{\mu_2}^*) - \frac{1}{3}\widehat{\mu_3}^* + \widehat{\mu_4}^* + \widehat{\mu_5}^*)\,,\\ &= \frac{\widehat{\alpha_2}}{2} - \frac{1}{3}(\widehat{\alpha_3} + \widehat{\alpha_4} + \widehat{\alpha_5})\,,\\ &= \frac{1.180}{2} - \frac{1}{3}(\text{-}3.236 - 1.849 - 3.065) = 3.306\end{aligned}$$

The estimated variance of the contrast is then

$$\begin{aligned}\widehat{\mathrm{Var}}(\widehat{U_m}) &= \frac{0.534}{4} - \frac{1}{3}\,(0.223 + 0.293 + 0.257)\\ &\quad + \frac{2}{9}\,(0.288 + 0.316 + 0.312) + \frac{1}{9}\,(0.606 + 0.645 + 0.767)\\ &= 0.303.\end{aligned}$$

Finally, $\sqrt{0.303} = 0.551$, and 3.306 divided by 0.551 yielded a z statistic of approximately 6 which was statistically significant under the FDR-adjusted threshold of $\alpha = 0.01$.

D.2 NMAR joint model

Consider the peak at 7412 m/z. This peak is deemed important by the Bayesian missingness model (4.4) and by the NMAR joint model (4.9). Table D.4 displays the parameter estimates $\hat{\boldsymbol{\kappa}}$. The estimated value of the contrast is equal to

$$\widehat{E}[\widehat{U_o}] = \frac{\widehat{\gamma}_2}{2} - \frac{1}{3}(\widehat{\gamma}_3 + \widehat{\gamma}_4 + \widehat{\gamma}_5) = \frac{1}{2}(\text{-}0.282) - \frac{1}{3}(1.126 + 0.557 + 0.924) = \text{-}1.01.$$

To obtain the MCMC standard error of the contrast, an average was taken of the estimated Monte Carlo error variance for each of the eight Markov chains in the model. The Monte Carlo error variance was estimated using the overlapping batch means standard error estimator $\hat{\sigma}^2_{\text{OLBM}}$ with $N_M = 1500$ and $b_n = 38$. Consider the first chain, whose first 45 draws of the contrast are listed in Table D.3. For this chain, $\widehat{E}[\widehat{U_o}]$ = -1.007. The first batch derived from the chain's draws, $((U_o)_n)_{n=1}^{38}$, is the sequence

(-1.030, -1.022, ..., -0.978).

From this batch, $(\bar{U}_o)_1(38)$ = -1.044. The next batch, $((U_o)_n)_{n=2}^{39}$, is the sequence

(-1.022, -1.075, ..., -1.012).

For this batch, the first draw of -1.030 in the previous batch has been discarded and the new 39th draw of -1.012 has been added. Thus, $(\bar{U}_o)_2(38)$ = -1.043. Likewise, $(\bar{U}_o)_3(38)$ = -1.041. The estimate of the Monte Carlo error, $\hat{\sigma}^2_{\text{OLBM}}$, for this chain is 0.0532. A naive estimate of the Monte Carlo error obtained by taking the sample variance of the 1500 contrast draws is 0.0176. In general, the overlapping batch means estimate of the Monte Carlo error is greater than the naive estimate due to the autocorrelation present in the sequence of draws. However, due to the efficient sampling that may be achieved by Hamiltonian Monte Carlo methods, the ratio between the two estimates may fall as low as 1, and even in some cases slightly below 1. In most cases, the overlapping batch means estimate is substantially higher than the naive estimate, making the former estimate a conservative one.

The overlapping batch means variance estimates from all chains are 0.0532, 0.0797, 0.0992, 0.0787, 0.0657, 0.0774, 0.079, and 0.0966, yielding an average of 0.0787. The square root of this value is 0.281, and this is the estimated Monte Carlo standard error that is used to calculate the z statistic of -1.01/0.281 $\approx$ -3.6 which yielded an adjusted p value below 0.001.

Table D.3: First 45 draws of the contrast U_o from the first Markov chain from the NMAR joint model (4.9) for the peak at 7412 m/z.

1–15	16–30	31–45
-1.030	-1.087	-0.884
-1.022	-1.042	-1.048
-1.075	-1.241	-0.938
-0.978	-1.114	-1.009
-0.956	-1.164	-0.956
-0.942	-1.195	-0.950
-1.017	-1.210	-0.954
-0.978	-1.134	-0.978
-1.048	-1.134	-1.012
-0.987	-1.083	-0.941
-1.135	-1.141	-0.996
-1.219	-1.101	-0.950
-0.905	-1.063	-0.908
-0.957	-1.039	-0.922
-0.969	-0.985	-0.813

Table D.4: Estimates of parameters in $\boldsymbol{\kappa}$ from the NMAR joint model (4.9) on the peak at 7412 m/z.

$\hat{\nu}$	$\widehat{\gamma_2}$	$\widehat{\gamma_3}$	$\widehat{\gamma_4}$	$\widehat{\gamma_5}$	$\widehat{\delta_2}$	$\widehat{\delta_3}$
-1.358	-0.282	1.126	0.557	0.924	0.293	0.094

Appendix E

Model simulation results

E.1 Bayesian GLMM for the missingness

The two simulation checks of Gelman and Hill (2009) are first applied to the Bayesian GLMM (3.4).

E.1.1 Assessing statistical procedures with constructed data

The first of the two simulation checks concerns the ability of model fitting procedures to estimate parameter values. The method is to construct data according to the statistical model using *known* parameters, and then fit the model to the constructed data to obtain parameter estimates in order to compare them to the known parameters. The difference between the true and estimated parameter values in 1000 replications of the procedure was used to assess the appropriateness of the model for the data.

The constructed data generation procedure for the model is described below.

1. Parameters were sampled from distributions
   - $\mu \sim N(0, 1)$,
   - $\alpha_j \sim N(q_j, 1)$, for $j = 2, \ldots, 5$, where
     $$q_2 = 2,\ q_3 = \text{-}0.4,\ q_4 = \text{-}1,\ q_5 = \text{-}2,$$
   - $\beta_\ell \sim N(0, 1)$ for $\ell = 2, 3$,
   - Random effect variance components were each, independently, either set to zero with probability 0.05 or generated according to
     $$\sigma^2_M \sim U(0, 1.2), \quad \sigma^2_A \sim U(0, 0.8), \quad \sigma^2_C \sim U(0, 1.5)$$
     with probability 0.95.

2. Random effect variables M_{jk} , $A_{jk\ell}$, and $C_{jk\ell m}$ were independently sampled as

$$M_{jk} \sim N(0,\, \sigma_M^2), \quad A_{jk\ell} \sim N(0,\, \sigma_A^2), \quad C_{jk\ell m} \sim N(0,\, \sigma_C^2).$$

   If the variance component of a particular random effect was equal to zero, all terms for that random effect were set to zero.

3. Missingness indicators $r_{jk\ell mn}$ were sampled as

$$r_{jk\ell mn} \sim \text{Bern}(p_{jk\ell mn}), \quad p_{jk\ell mn} = \frac{e^{\eta_{jk\ell mn}}}{1 + e^{\eta_{jk\ell mn}}} ,$$

   with

$$\eta_{jk\ell mn} = \mu + \alpha_j + \beta_\ell + M_{jk} + A_{jk\ell} + C_{jk\ell m} .$$

Figure E.1 displays a boxplot summary of the set of 1000 differences between true and estimated values for each parameter. A positive difference implies that the estimated value of a parameter is greater than the true value. Table 3.6 reports the sample means and standard deviations of the sets of differences. The estimates of the group parameters α_2 , α_4 , and α_5 were biased towards zero, which may indicate that the prior distributions assumed for the group parameters were more informative than justified. In rare cases, the discrepancy between a parameter's true value and its estimated value exceeded 1. Such discrepancies arose in cases of separation or near-separation in the generated data, which caused parameter estimates to be more extreme than the actual values despite the regularising effect of the prior distributions. The parameters σ_M^2 and σ_C^2 tend to be underestimated, which may be due to the low rate hyperparameter in the gamma priors for those parameters. Considering the ratios of the means of the differences to the standard deviations of the differences, however, the parameter estimation methods appear to be adequate.

E.1.2 Assessing model fits with predictive simulation

The fitted probabilites from the models on the GC dataset can be used to simulate a *predicted dataset* which may be compared with the missingness pattern of the original dataset. The smaller the differences, the more suitable the model at explaining the data. Comparisons of the original dataset to the predicted dataset are largely ad-hoc judgements that are informed by the particular characteristics of the dataset.

The predicted dataset was generated by sampling $r_{ijk\ell mn}$ as

$$r_{ijk\ell mn} \sim \text{Bern}(\hat{p}_{ijk\ell mn}),$$

where $\hat{p}_{ijk\ell mn}$ is the fitted probability (from the model on the ith peak) for sample replicate n from C8 batch m within aliquot ℓ from mouse k in group j. The vector of fitted probabilities $\hat{\boldsymbol{p}}$ from each peak was obtained using the `fitted` function in R.

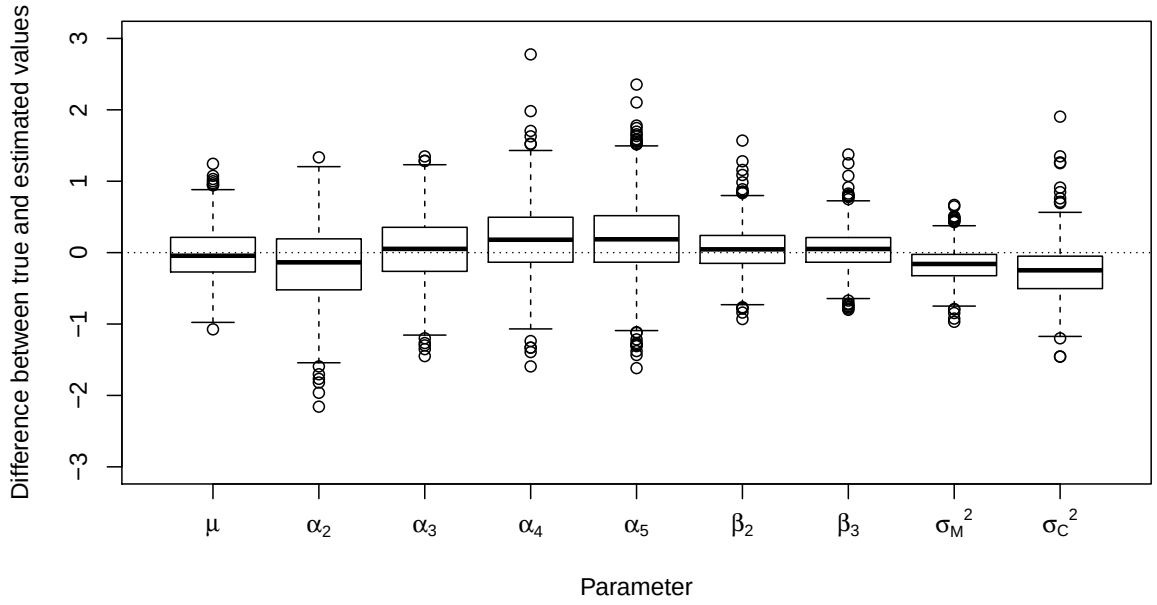

Figure E.1: Summary of differences between estimated and true parameter values over 1000 replications of constructed data procedure.

Table E.1: Mean and sample standard deviation of differences between estimated and true parameter values over 1000 replications of the constructed data procedure.

Parameter	Mean	Standard Deviation
μ	-0.03	0.36
α_2	-0.18	0.53
α_3	0.03	0.45
α_4	0.19	0.49
α_5	0.18	0.62
β_2	0.05	0.31
β_3	0.05	0.28
σ^2_M	-0.18	0.23
σ^2_C	-0.27	0.34

The predicted dataset is compared to the original dataset in several ways. First, Figure E.2 displays the missingness indicators for a subset of 40 peaks in the original dataset and the predicted dataset. The general pattern of missingness appears similar.

Second, Figure E.3 displays the missingness counts for samples on chip 1 in the original and predicted datasets. The predicted dataset diverges from the original dataset in that the former appears to have less variance in missingness, in large part due to an absence of outlying samples with extreme missingness counts.

Third, Figure E.4 displays missingness counts for observations within a subset of peaks for the original and predicted datasets. Patterns of missingness appear to be faithfully reproduced, including cases of isolated group/chip combinations that greatly differ from other combinations that have either chip or the group in common. The combination of chip 3 and group 4 in the peak at 14836 m/z is an example of such a case. The nonlinear effect of chips and groups on the probability of missingness allows this feature of the original dataset to be reproduced.

The Bayesian GLMM of Equation (3.3) is suitable for the data, although some outliers in the original dataset do not appear to be appropriately captured. For example, from Figure E.3, samples on chip 1 with extremely high or low numbers of missing observations are present in the GC dataset but do not appear in the predicted dataset. One possibile cause of the absence of these outliers is that the normal distribution of random effect terms penalises extreme terms too heavily.

E.2 NMAR joint model

The two simulation checks of Gelman and Hill (2009) are applied to the NMAR joint model (4.9).

E.2.1 Assessing statistical procedures with constructed data

The constructed data procedure, which was replicated 60 times, was as follows:

1. Parameters were sampled from distributions

   - $\nu \sim N(0,\, 1)$,
   - $\gamma_j \sim N(q_j\,,\, 1)$ for $j = 2, \ldots, 5$, where

     $$q_2 = 1,\ q_3 = \text{-}0.2,\ q_4 = \text{-}0.5,\ q_5 = \text{-}1,$$

   - $\delta_\ell \sim N(0,\, 1)$ for $\ell = 2,\, 3$,
   - $\sigma_N^2 \sim U(0,\, 2)$ w.p. 0.95, $\sigma_N^2 = 0$ w.p. 0.05,
   - $\sigma_B^2 \sim U(0,\, 2)$ w.p. 0.95, $\sigma_B^2 = 0$ w.p. 0.05,

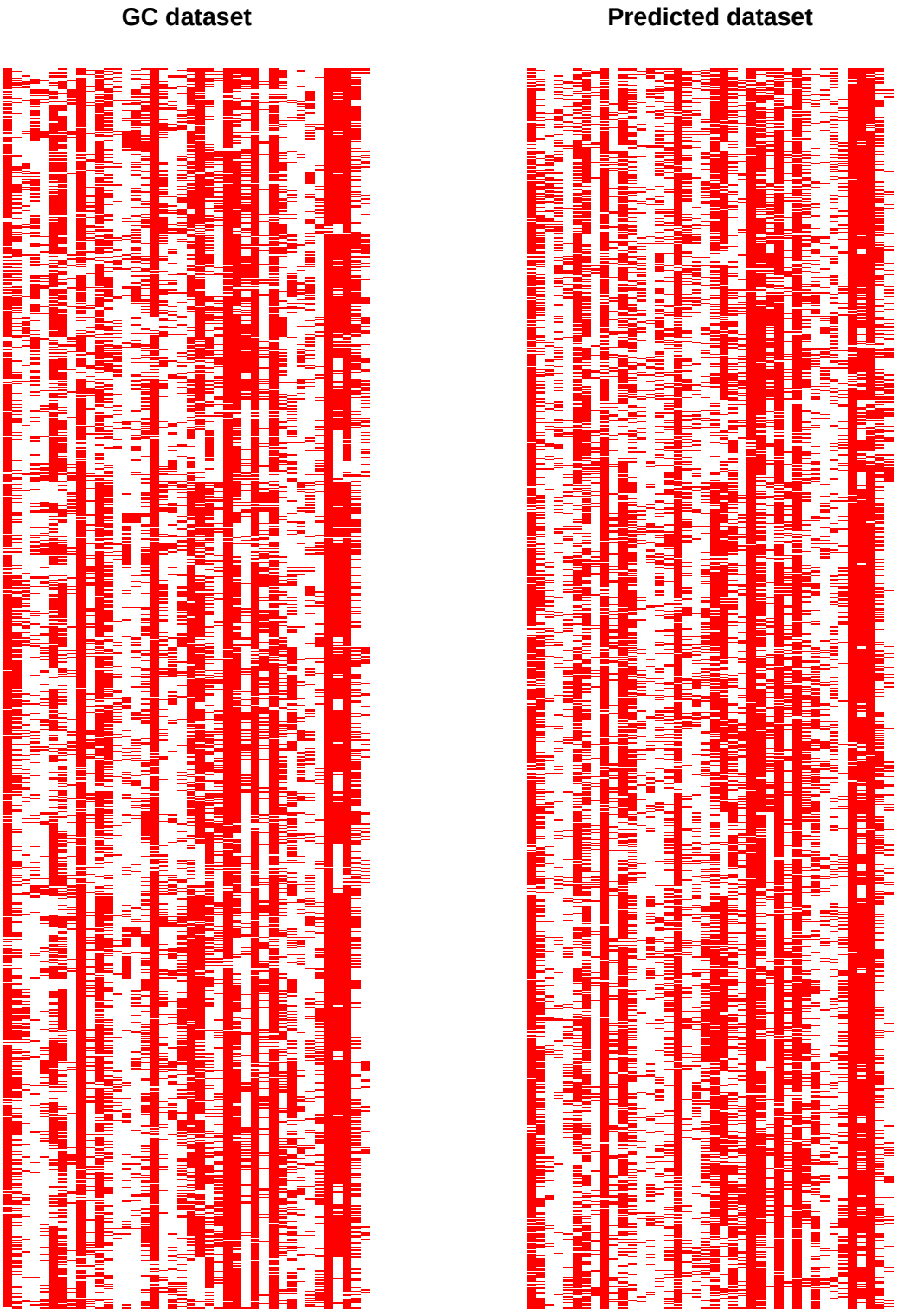

Figure E.2: Array of missingness indicators for all samples within 40 peaks of the original GC dataset (left) and the predicted dataset (right). Red cells correspond to observed data in the matrix and yellow cells to missing data.

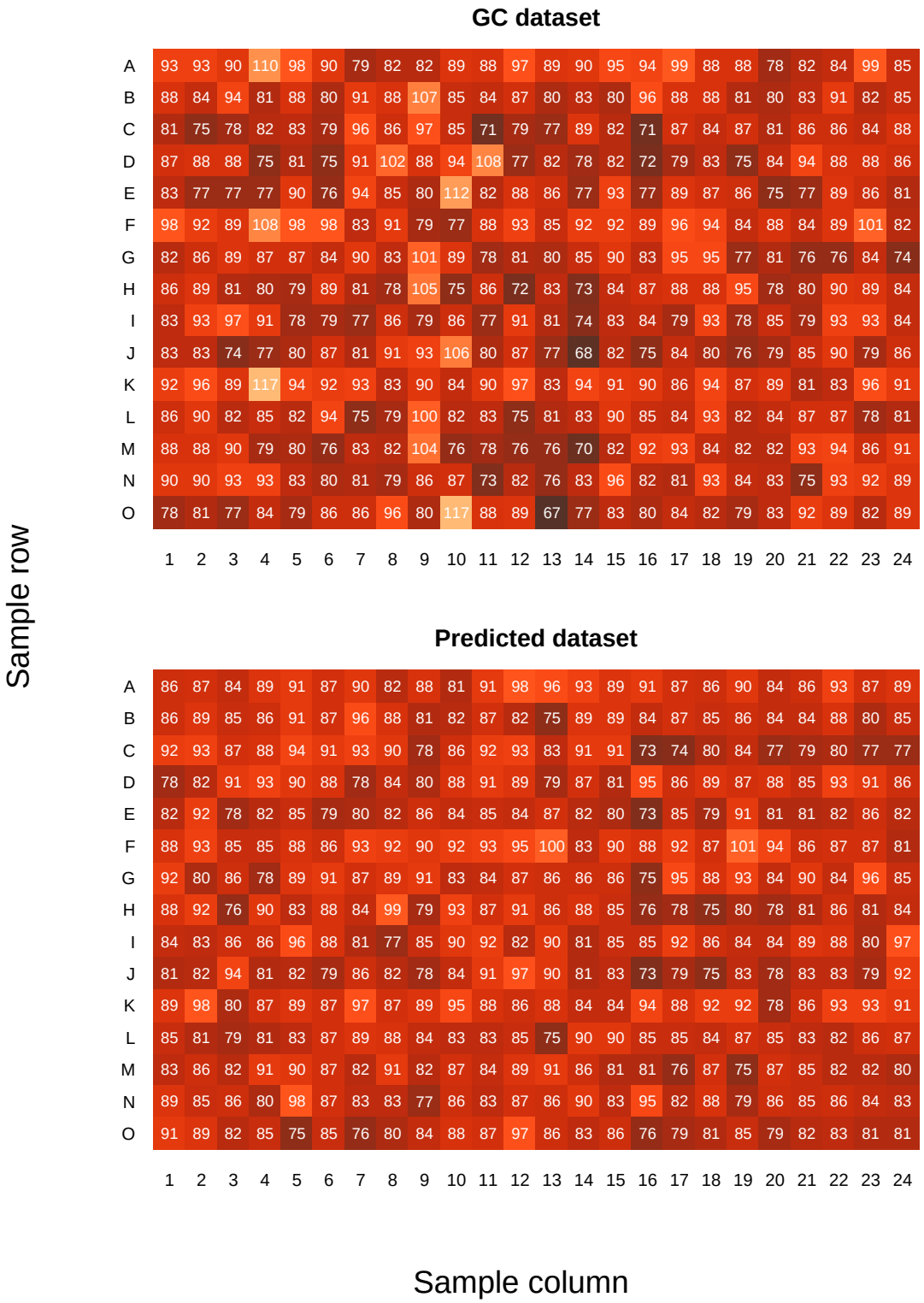

Figure E.3: Counts of missingness for all observations within samples on chip 1 for the original GC dataset (top) and the predicted dataset (bottom). Interpretation follows that of Figure 2.8a.

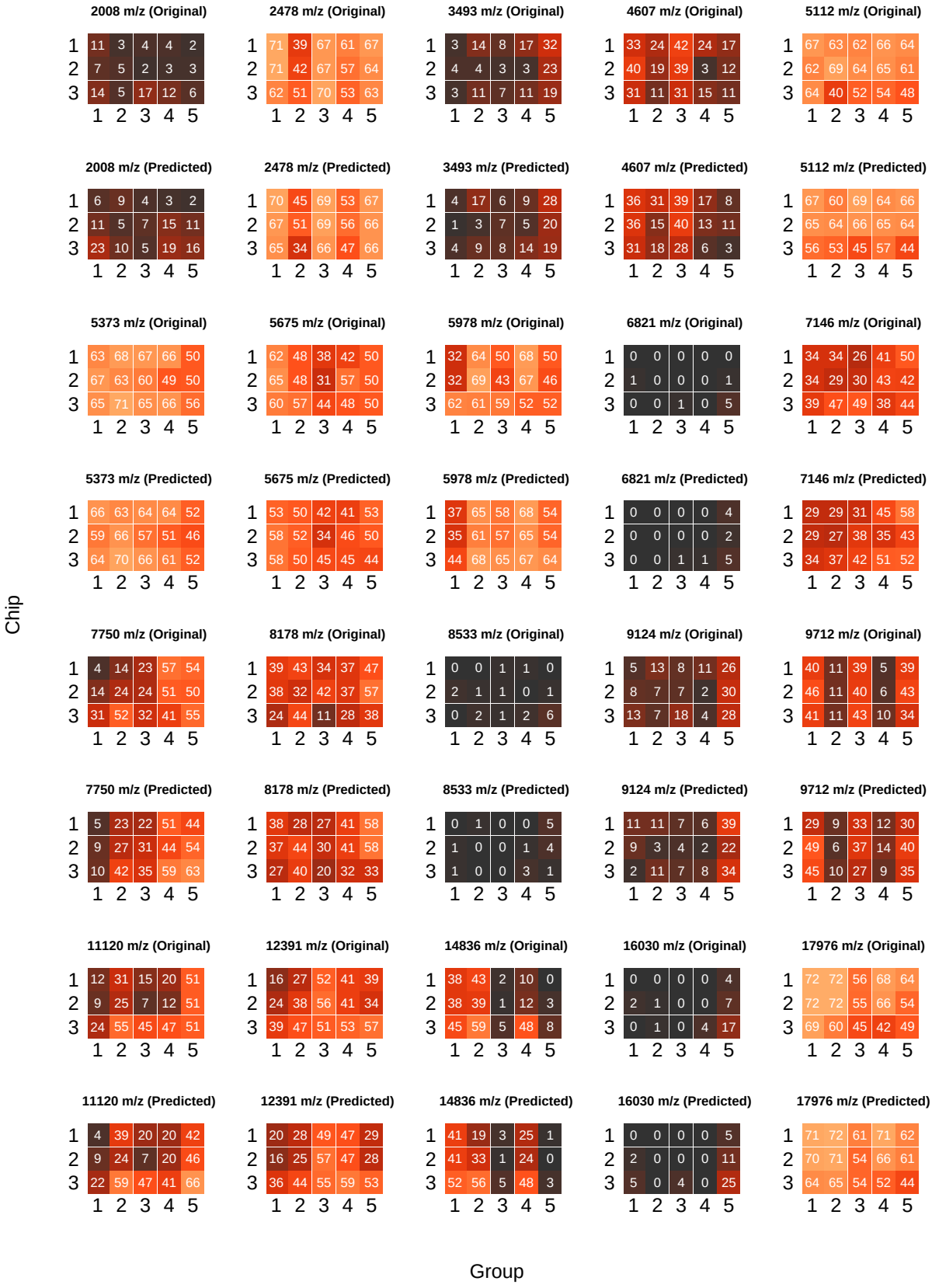

Figure E.4: Counts of missingness for all observations within a subset of peaks. Odd rows correspond to peaks in the original GC dataset, and even rows correspond to peaks in the predicted dataset. Otherwise, interpretation follows that of Figure 2.9.

- $\sigma^2 \sim U(0,\, 2)$,
- $\mu \sim N(\text{-}4,\, 4)$,
- $\alpha_j \sim N(0,\, 4)$ for $i = 2,\, \ldots,\, 5$,
- $\beta_\ell \sim N(0,\, 4)$ for $\ell = 2,\, 3$,
- $\omega \sim U(\text{-}8,\, 0)$.

2. Random effect vector elements of $\boldsymbol{b}_M$ and $\boldsymbol{b}_C$ were generated *i.i.d.* as $N(0,\, \sigma_N^2)$ and $N(0,\, \sigma_B^2)$ respectively, with all elements of either vector equal to zero if the corresponding variance component was equal to zero.

3. Residual vector elements of $\boldsymbol{\varepsilon}$ were generated *i.i.d.* as $N(0,\, \sigma^2)$.

4. Vectors of fixed effects $\boldsymbol{\kappa}$ and $\boldsymbol{\lambda}$ were formed according to

$$\boldsymbol{\kappa} = (\nu\,,\, \gamma_2\,,\, \gamma_3\,,\, \gamma_4\,,\, \gamma_5\,,\, \delta_2\,,\, \delta_3)^T$$

and

$$\boldsymbol{\lambda} = (\mu\,,\, \alpha_2\,,\, \alpha_3\,,\, \alpha_4\,,\, \alpha_5\,,\, \beta_2\,,\, \beta_3)^T.$$

5. Vector of observations $\boldsymbol{y}$ was produced according to

$$\boldsymbol{y} = X\boldsymbol{\kappa} + Z_M\boldsymbol{b}_M + Z_C\boldsymbol{b}_C + \boldsymbol{\varepsilon}$$

and vector of linear predictors $\boldsymbol{\eta}$ for the missingness indicator was produced according to

$$\boldsymbol{\eta} = X\boldsymbol{\lambda} + \omega\boldsymbol{y},$$

where X is the design matrix for the fixed effects and Z_M and Z_C are the design matrices for the random effects of mouse and C8 batch.

6. Vector $\boldsymbol{r}$ was produced by sampling each element r_s as $\text{Bern}(e^{\eta_s}/(1 + e^{\eta_s}))$

7. Elements of $\boldsymbol{y}$ with corresponding element of $\boldsymbol{r}$ equal to 1 were replaced with a placeholder used for missing elements.

This is a procedure in which the process of constructing the data draws parameters from distributions that are not the same as the (relatively informative) prior distributions actually used in fitting the model. This allows for the impact of using possibly-unsuitable priors to be seen when comparing the parameter estimates and the predictive simulations to those of Section E.3.

Figure E.5 displays a boxplot of the distributions of differences between estimated and true values for each parameter. A positive difference implies the estimated value is greater than the true value. Table E.2 summarises the distributions of differences. The

fixed effects parameters for the missingness indicators in $\boldsymbol{\lambda}$ have the greatest spread in their distributions of differences, and the spread for ω is of comparable size. The fixed effects parameters for the intensity in $\boldsymbol{\kappa}$ have a lesser spread, and the parameters σ_N^2 , σ_B^2 , and σ^2 have the lowest spread.

The estimates of the μ and ω parameters are positively biased. A possible cause for the former is that the prior distribution of μ has variance that is too small, biasing the estimates upwards from the true values of that parameter, which were, on average, -4. Estimates of μ that are larger than the true value would have the effect of predicting too low a probability of missingness for certain data points, and an increased estimate of ω (that is closer to 0, as ω must be negative) compensates for these low pedicted probabilities. This may account for the estimates of ω being positively biased.

The parameter ν with the α_j parameters represents estimated group mean intensities. These parameters are estimated with slight biases. The parameter ν was slightly overestimated, which may be due to insufficient compensation for the NMAR missingness mechanism preferentially removing low intensity observations. The parameters γ_4 and γ_5 were generated with mean values of -0.5 and -1 respectively, representing simulated group means below the grand mean of the GC dataset by those amounts (as ν was generated with a mean of 0). These parameters were, on average, overestimated. The parameter γ_2 had an average value of 1, representing a simulated group mean equal to one plus the grand mean of the GC dataset. This parameter tended to be underestimated. These biases may be caused by too high a degree of shrinkage of the estimates towards zero, reflecting a prior distribution with more precision than is justified.

Estimates of variance components tend to have less precision the higher the true parameter values. The σ^2 parameter tended to be underestimated, which may be due to insufficient compensation for the NMAR missingness mechanism.

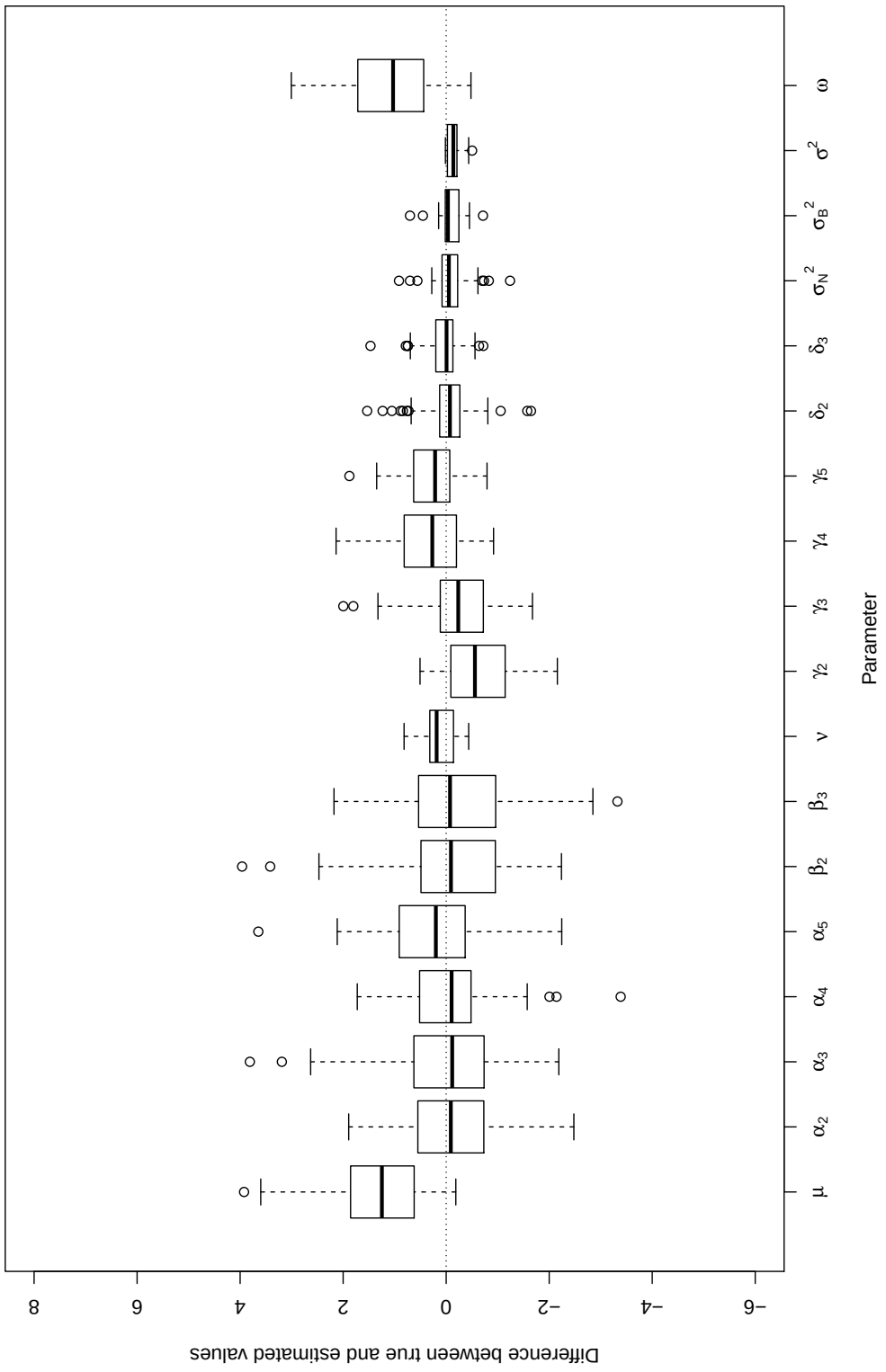

Figure E.5: Boxplots of $N = 60$ differences between estimated and true parameter values from constructed data generation and model-fitting process for the NMAR joint model with priors obtained from the data.

Figure E.6 reveals that the proportion of missing observations has little impact on the ability of the model fitting procedure to recover the parameter values. However, the low sample size of $N = 60$ makes it difficult to be certain that there is no effect.

Figures E.7a and E.7b display the true parameter values plotted against the estimated parameter values. Estimates of the α_j and β_ℓ parameters for the missingness appear to be shrunk towards zero, potentially reflecting prior distributions that are more informative than what is justified. Estimates of the γ_j parameters for the group intensities are similarly shrunk towards zero, making it difficult to trust the estimated values for the purpose of inference of group mean intensity differences. The estimated values of ω differ more from the true values when the true values are low, which implies that the parameter is being shrunk towards zero.

E.2.2 Assessing model fits with predictive simulation

The second method of Gelman (2006) to assess a statistical model is to use the parameter estimates to generate data according to the model. Visual inspection and summary statistics of the data provide insight into the suitability of the model.

The data generation procedure was the same as the constructed data procedure from Section E.2.1 with the exception that each element of the random effect vectors $\boldsymbol{N}$ and $\boldsymbol{B}$ was determined by the estimates from the `stan` model for a particular peak rather than by independently sampling from some pre-specified distribution. Figures E.8 and E.9 display data predicted in this way from four m/z peaks across the range of m/z values in the GC dataset. The simulated peak data appears very similar to the true datasets, which is evidence towards the model's suitability for the data. Figure E.10 displays predicted data from a broader viewpoint finding that the predicted data are very similar to the original data.

One departure from the true data in the predicted data is that in the GC data, there exists a threshold below which intensity values do not occur, but in the predicted data, intensity values may be below this threshold. This is apparent as points below the dotted horizontal lines in Figures E.8 and E.9. This departure suggests that the prior distributions used for ω and the μ_j parameters are insufficiently informative and fail to account for the sharp thresholds induced by the pre-processing of the GC dataset.

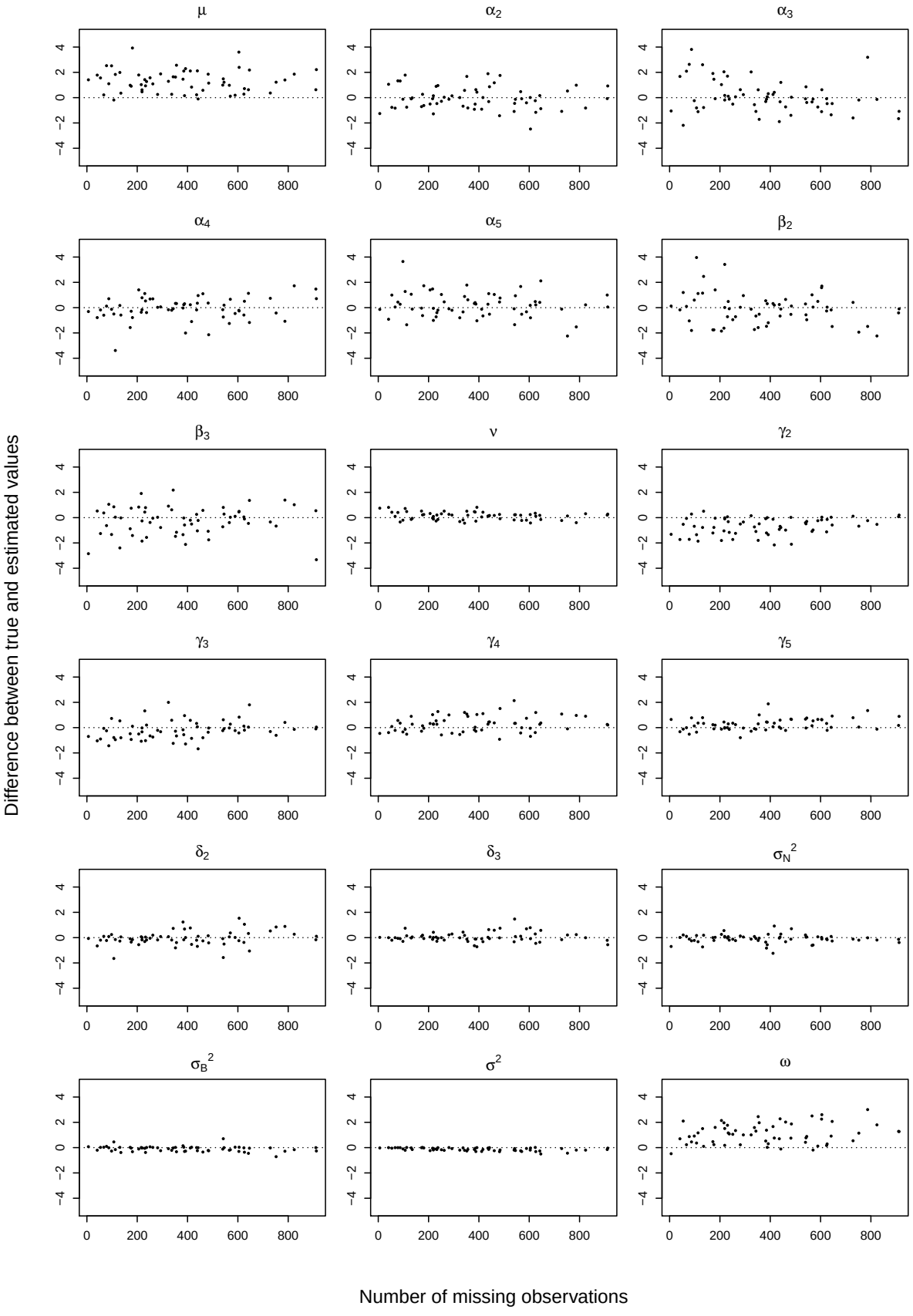

Figure E.6: Plots of $N = 60$ differences between true and estimated parameter values against number of missing observations from constructed data generation and model-fitting process for the NMAR joint model with priors obtained from the data.

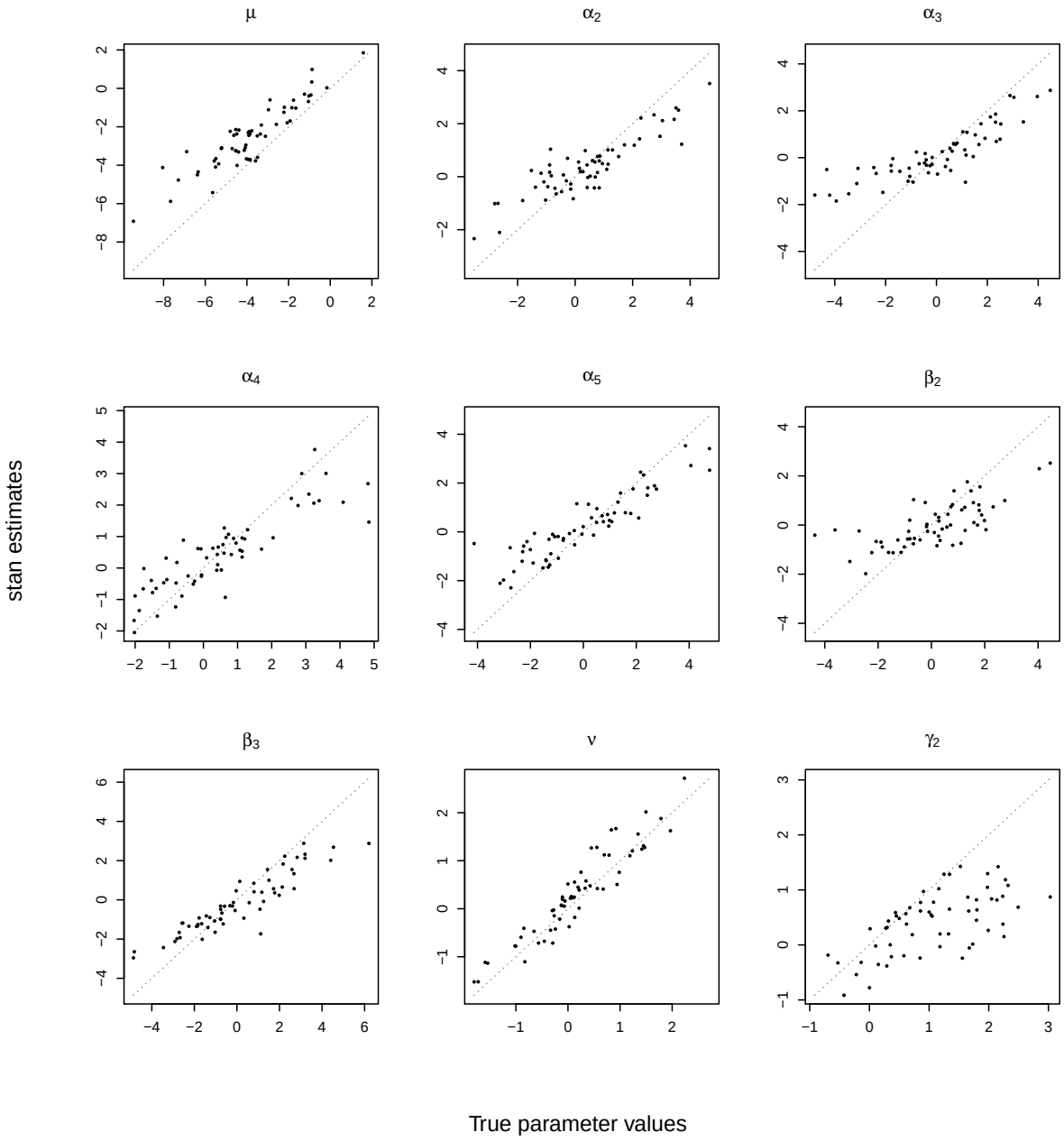

Figure E.7a: Plots of true versus estimated values for parameters across $N = 60$ replications of constructed data generation and model-fitting process for the NMAR joint model with priors obtained from the data. Each plot window is devoted to a single parameter and contains the true and estimated values from each of the simulations. Points within plots represent individual simulations. The closer the values are for a given simulation, the closer the point is to the dotted line $y = x$.

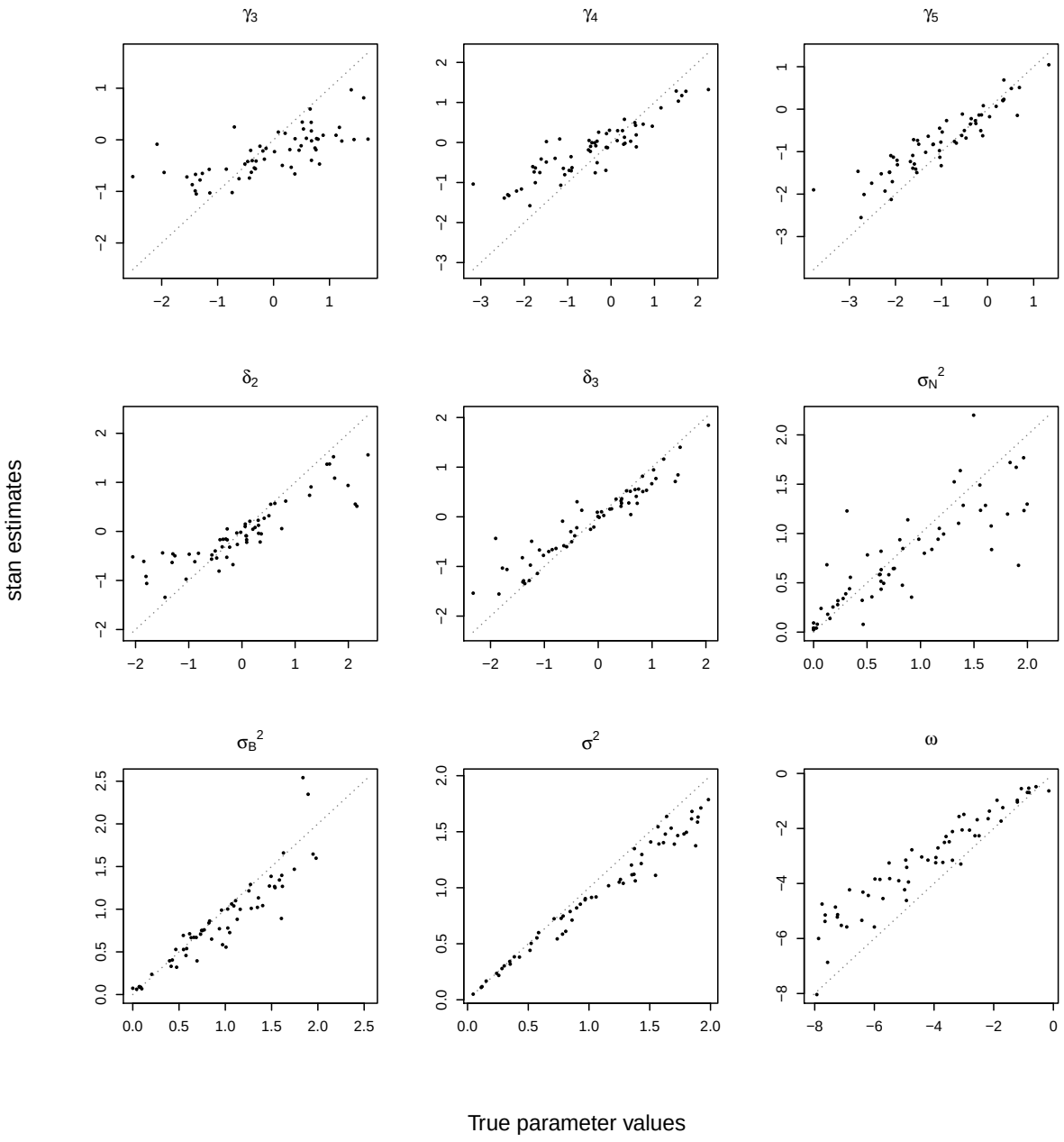

Figure E.7b: Plots of true versus estimated values for parameters across $N = 60$ replications of constructed data generation and model-fitting process for the NMAR joint model with priors obtained from the data. Each plot window is devoted to a single parameter and contains the true and estimated values from each of the simulations. Points within plots represent individual simulations. The closer the values are for a given simulation, the closer the point is to the dotted line $y = x$.

Table E.2: Means and sample standard deviations of differences between estimated and true parameter values over 60 replications of constructed data procedure for the NMAR joint model with priors obtained from the data.

	Mean	SD
μ	1.298	0.869
α_2	-0.040	0.901
α_3	0.112	1.282
α_4	-0.078	0.905
α_5	0.212	0.993
β_2	-0.095	1.250
β_3	-0.242	1.109
ν	0.140	0.311
γ_2	-0.672	0.665
γ_3	-0.199	0.727
γ_4	0.292	0.625
γ_5	0.266	0.466
δ_2	-0.039	0.564
δ_3	0.052	0.377
σ_N^2	-0.088	0.344
σ_B^2	-0.098	0.210
σ^2	-0.134	0.120
ω	1.100	0.803

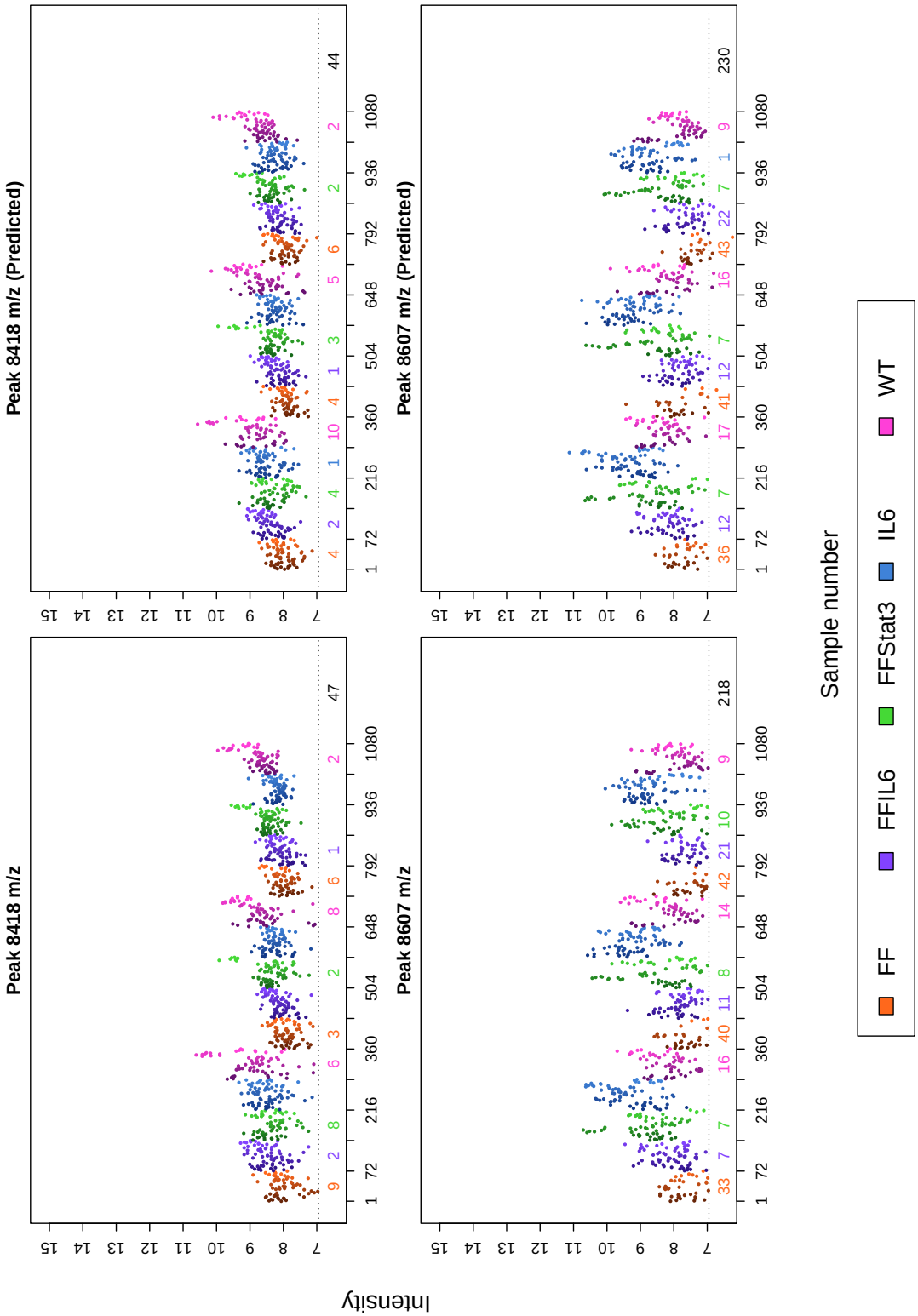

Figure E.8: Comparison of intensities within peaks in the GC dataset (left column) with predicted intensities based on NMAR joint model (4.9) (right).

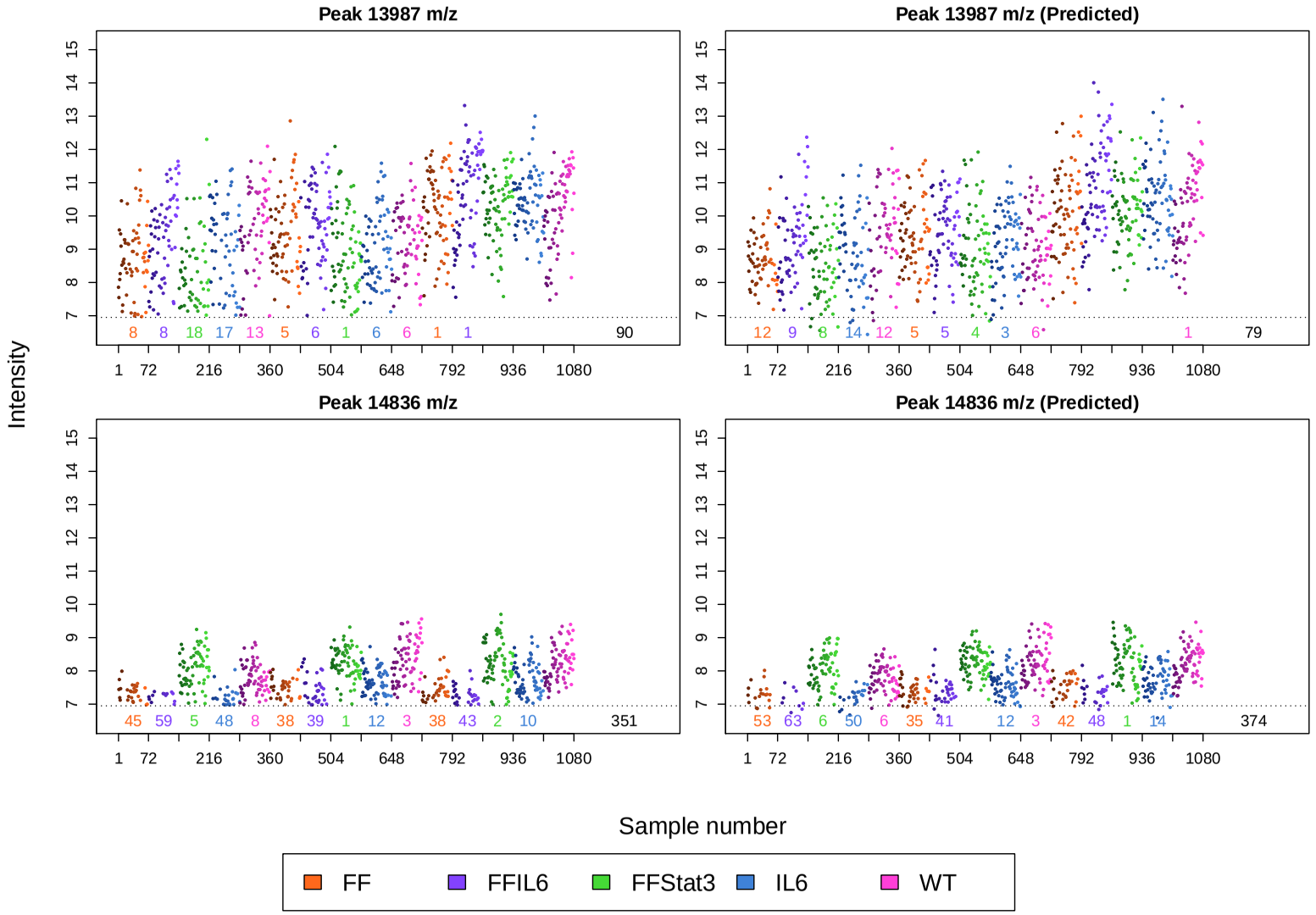

Figure E.9: Comparison of intensities within peaks in the GC dataset (left column) with predicted intensities based on NMAR joint model (4.9) (right).

E.3 NMAR joint model with weak priors

The first of the two simulation checks of Gelman and Hill (2009) for the NMAR joint model is repeated here using weakly-informative prior distributions for the parameters. This was done in order to elucidate the cause of the inaccuracies in recovering the true parameter values that arose in Section E.2.1.

For $\boldsymbol{\kappa}$ and $\boldsymbol{\lambda}$, the prior distributions were normal distributions with mean vector equal to $(0, 0, 0, 0, 0, 0, 0)^T$ and covariance matrix equal to $\text{diag}((5, 5, 5, 5, 5, 5, 5)^T)$. For σ_N^2, σ_B^2, and σ^2, the prior distributions (which were on the variance scale) were gamma distributions with shape parameter 1.2 and rate parameter 0.05. The prior for ω was a gamma distribution with shape equal to 3 and rate equal to 0.5 on -ω.

Figure E.11 displays a boxplot of the distributions of differences between estimated and true values for each parameter. A positive difference implies the estimated value is greater than the true value. Table E.3 summarises the distributions of differences. The variances of estimated values around the true values when using generic priors were less than or approximately equal to the corresponding variances when using priors estimated from the data. The biases in the estimates of μ and ω were greatly improved, and the biases in the estimates of the intensity fixed effects in $\boldsymbol{\kappa}$ were improved too. The biases in the estimates of the α_j and β_ℓ parameters for the group and chip missingness probabilities were not improved on in general.

Figure E.12 reveals that the proportion of missing observations has little impact on the ability of the model fitting procedure to recover the parameter values.

Figures E.13a and E.13b display the true parameter values plotted against the estimated parameter values. When using weak, generic prior distributions for parameters in the NMAR joint model, the issues of shrinkage from the true parameter values towards zero appears to be ameliorated. Variance components (especially σ_N^2 appear to be estimated with less precision as the true parameter values increase.

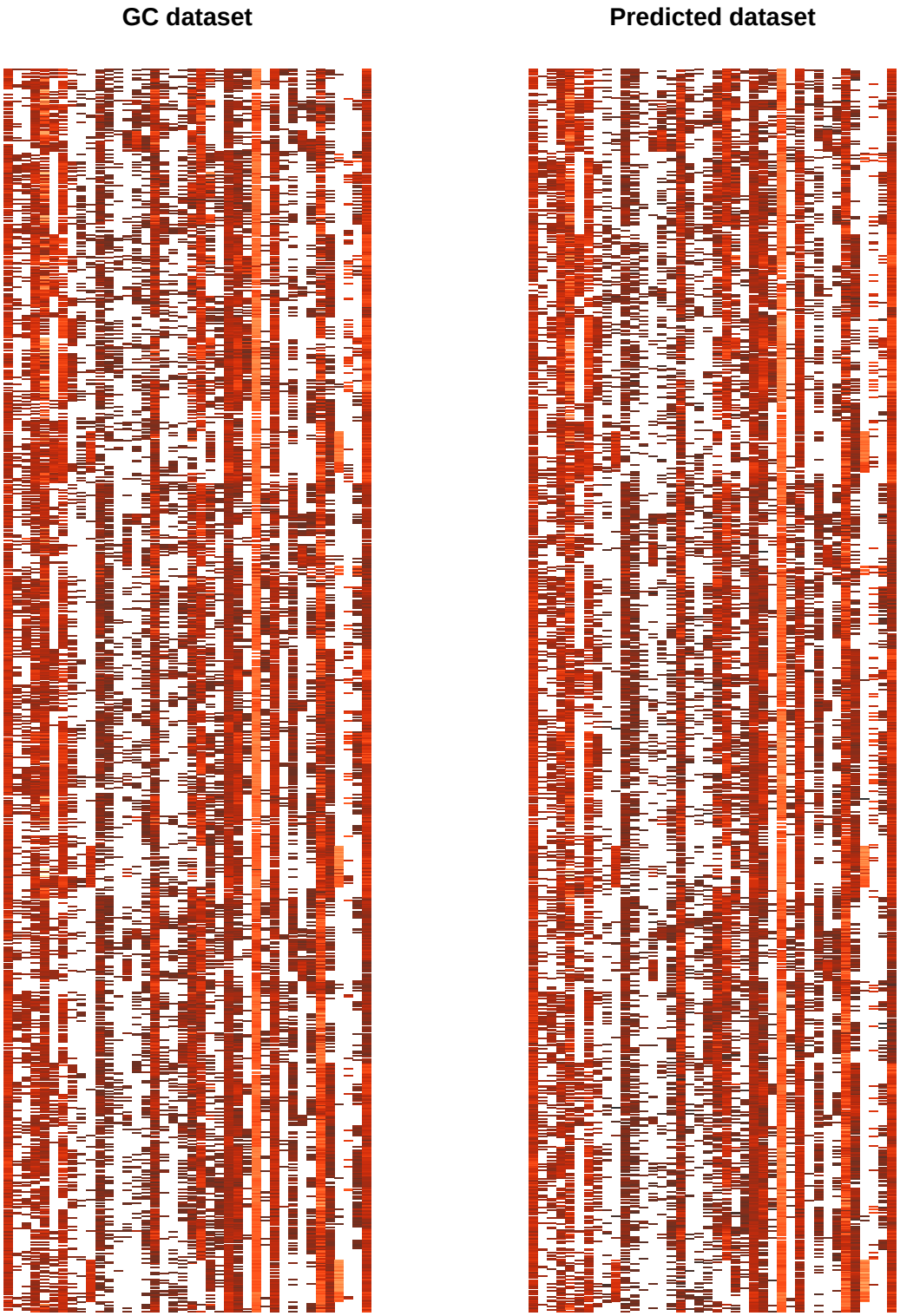

Figure E.10: Comparison of subset of 40 peaks in GC dataset observation matrix (left column) with simulated values based on NMAR joint model (4.9) (right).

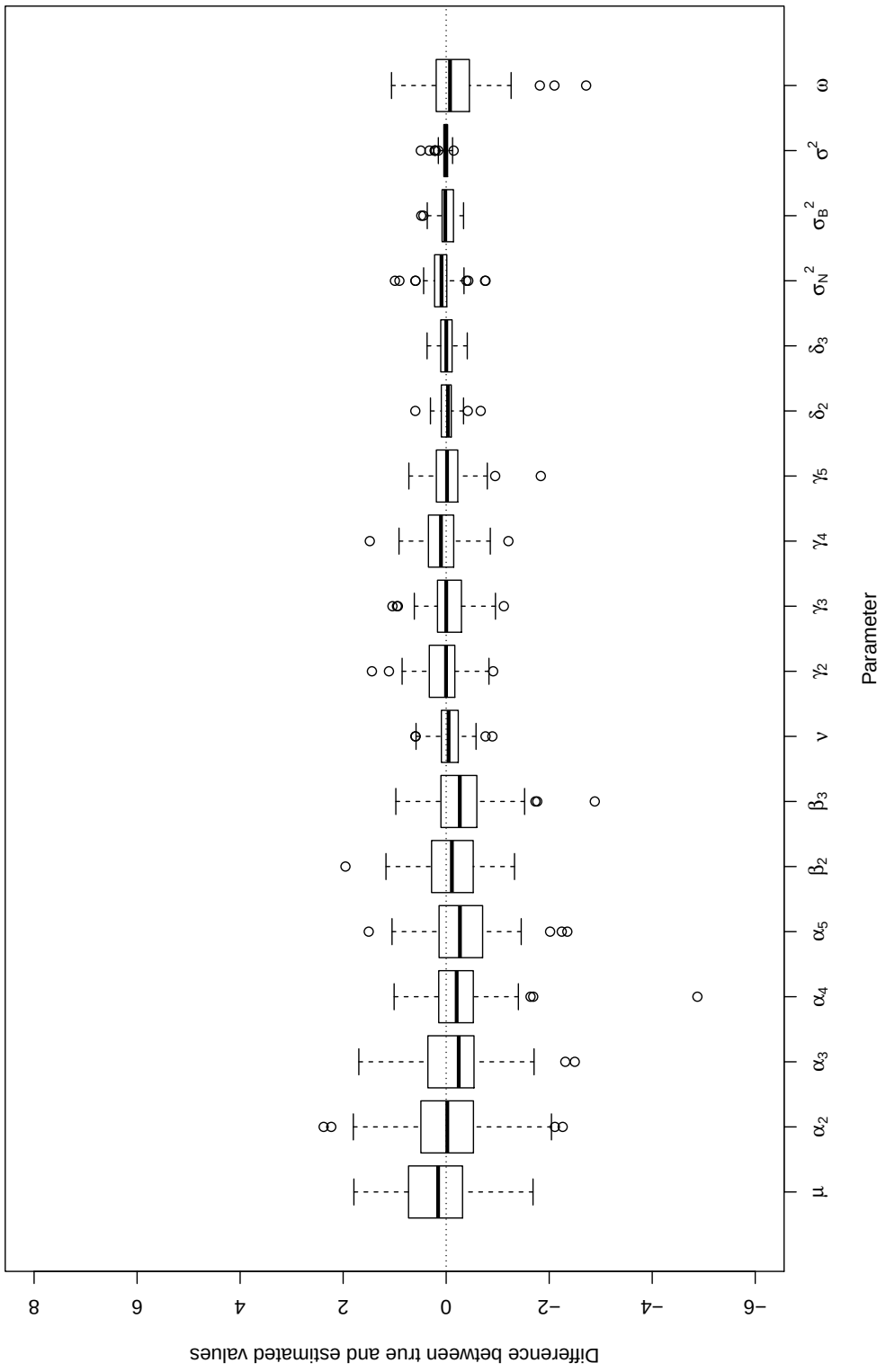

Figure E.11: Boxplots of $N = 60$ differences between estimated and true parameter values from constructed data generation and model-fitting process for the NMAR joint model with generic, weakly informative priors.

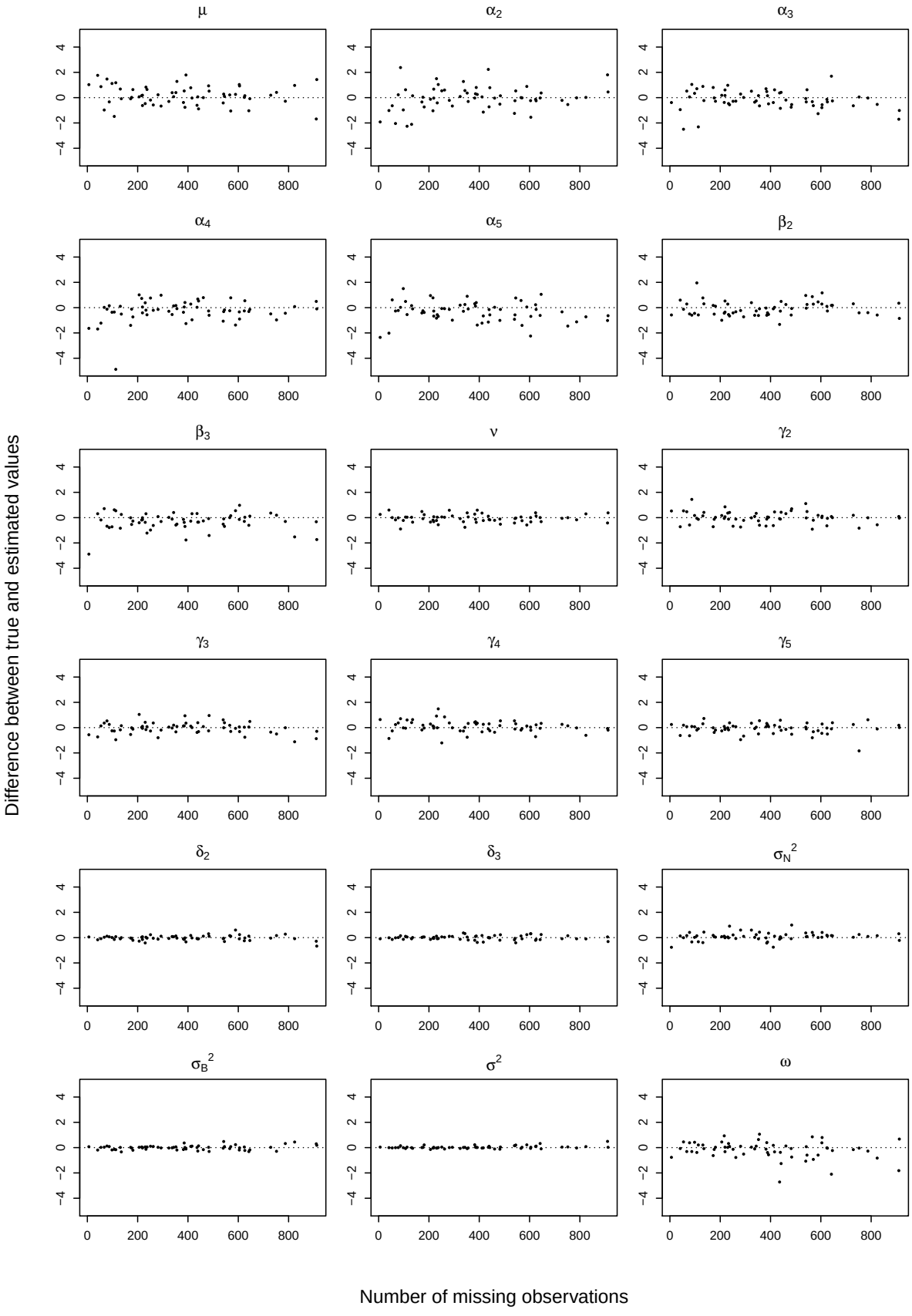

Figure E.12: Plots of $N = 60$ differences between true and estimated parameter values against number of missing observations from constructed data generation and model-fitting process for the NMAR joint model with generic, weakly informative priors.

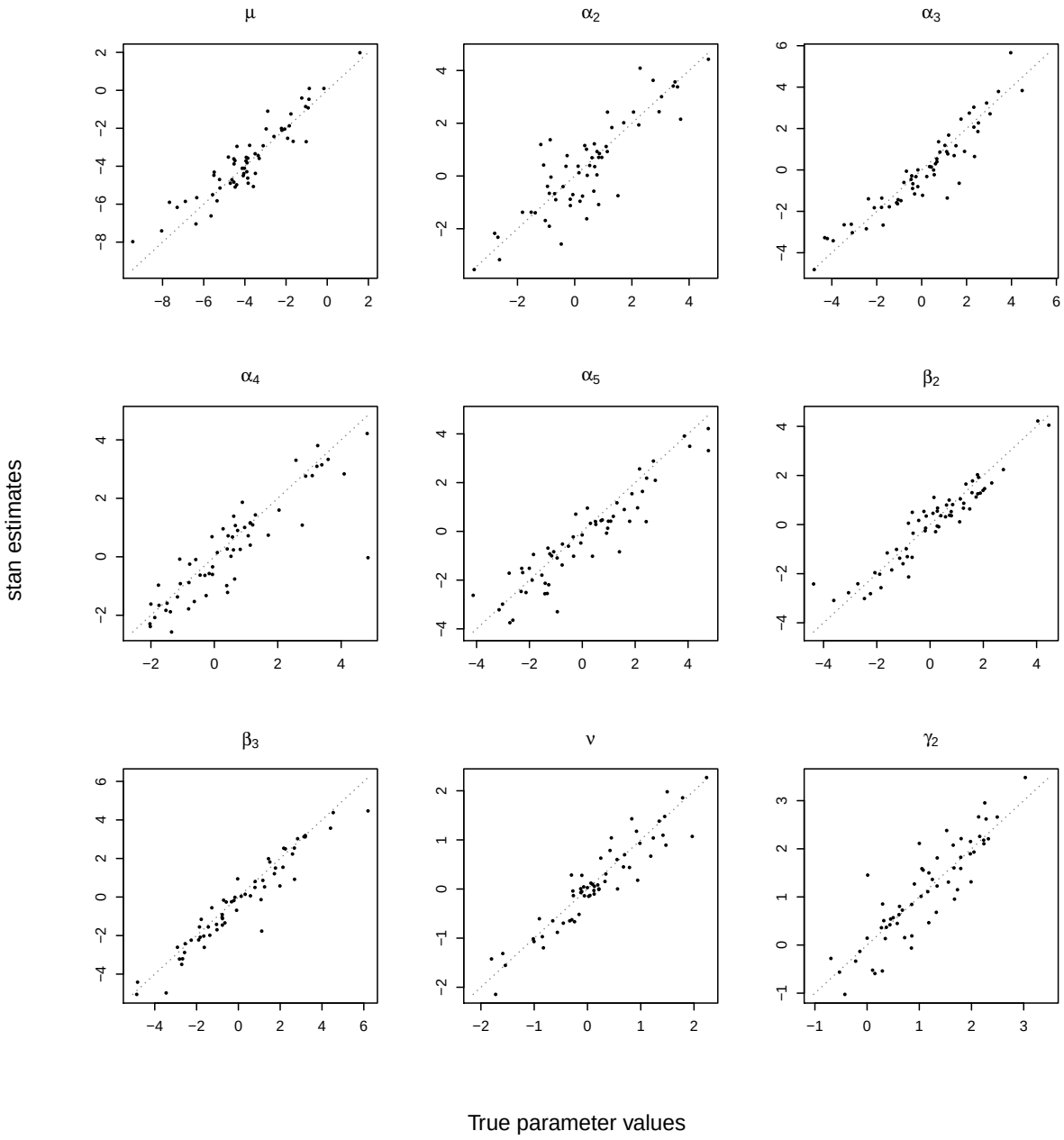

Figure E.13a: Plots of true versus estimated values for parameters across $N = 60$ replications of constructed data generation and model-fitting process for the NMAR joint model with generic, weakly informative priors. Each plot window is devoted to a single parameter and contains the true and estimated values from each of the simulations. Points within plots represent individual simulations. The closer the values are for a given simulation, the closer the point is to the dotted line $y = x$.

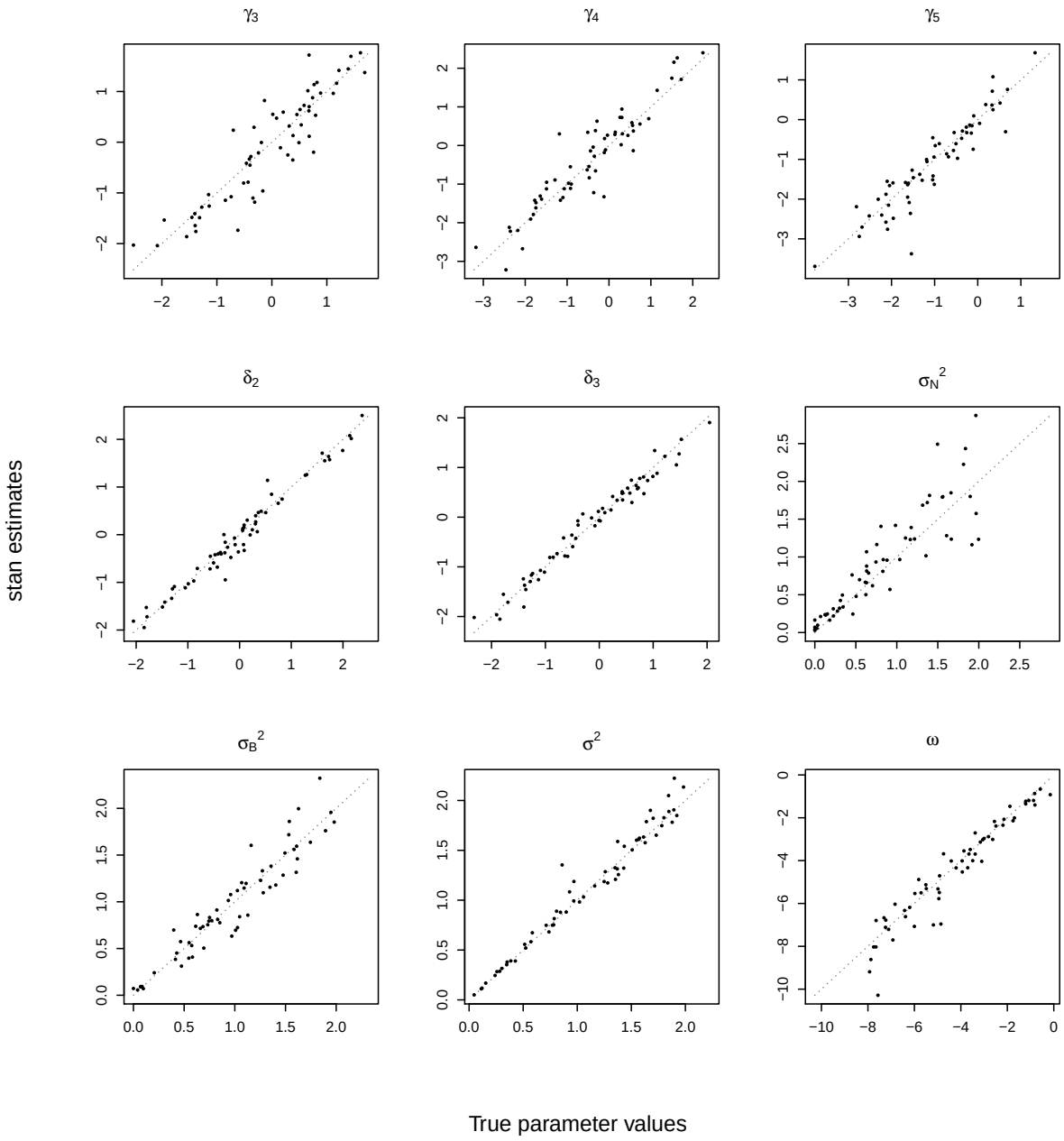

Figure E.13b: Plots of true versus estimated values for parameters across $N = 60$ replications of constructed data generation and model-fitting process for the NMAR joint model with generic, weakly informative priors. Each plot window is devoted to a single parameter and contains the true and estimated values from each of the simulations. Points within plots represent individual simulations. The closer the values are for a given simulation, the closer the point is to the dotted line $y = x$.

Table E.3: Means and sample standard deviations of differences between estimated and true parameter values over 60 replications of constructed data procedure for the NMAR joint model with weakly informative priors.

	Mean	SD
μ	0.177	0.764
α_2	-0.051	0.943
α_3	-0.156	0.744
α_4	-0.281	0.873
α_5	-0.334	0.776
β_2	-0.076	0.570
β_3	-0.301	0.660
ν	-0.058	0.317
γ_2	0.013	0.473
γ_3	-0.037	0.442
γ_4	0.096	0.445
γ_5	-0.064	0.421
δ_2	-0.025	0.191
δ_3	-0.006	0.171
σ_N^2	0.096	0.311
σ_B^2	-0.002	0.176
σ^2	0.031	0.109
ω	-0.180	0.683

Appendix F

Simulation with misspecified MAR joint model

F.1 Assessing statistical procedures with constructed data

The two methods of model assessment recommended by Gelman (2006) are investigated here for the MAR joint model of Equation (4.6). The first method uses constructed data to evaluate the parameter estimation method. The same data construction procedure is used in order to examine the effect of misspecifying the model.

The constructed data procedure was the same as in Appendix E.2.1. The `stan` MCMC method of estimating parameters for the MAR joint model was assessed using 60 replications of the constructed data process. For each replication, the true value subtracted from the estimated value was recorded for each parameter. Figure F.1 displays the distributions of differences between estimated and true values for each parameter as a boxplot. Table F.1 summarises the distributions of differences. The standard deviations of the differences for all parameters are either greater, or very close to equal, in the MAR joint model as compared to the NMAR joint model. For the parameters in $\boldsymbol{\lambda}$, the standard deviations are far greater, but for the parameters in $\boldsymbol{\kappa}$ the standard deviations tend to be only slightly greater.

The parameters μ and ν tend to be overestimated. The estimates of the γ_j parameters are slightly less accurate overall in the MAR joint model as compared to the NMAR joint model. Given that group means are equal to $\nu + \gamma_j$, the true average value of group 2, which was $E[\nu + \gamma_2] = 1$, was estimated with a small overall bias because of the opposite and nearly equal biases of the estimated values of ν and γ_2 . However, for other groups with lower averages, such as groups 4 and 5, the estimates of group means were biased upwards, with more extreme overestimates tending to track with decreasing true group means. This is expected due to the missingness mechanism of the constructed data causing

Table F.1: Means and sample standard deviations of differences between estimated and true parameter values over 60 replications of constructed data procedure using the MAR joint model (4.6).

Parameter	Mean	SD
μ	2.41	1.82
α_2	−1.13	1.54
α_3	0.28	1.92
α_4	0.08	1.83
α_5	1.19	1.74
β_2	−0.08	1.58
β_3	0.05	2.01
ν	0.71	0.50
γ_2	−0.64	0.64
γ_3	−0.05	0.76
γ_4	0.41	0.76
γ_5	0.52	0.72
δ_2	0.02	0.68
δ_3	0.11	0.56
σ_N^2	−0.42	0.44
σ_B^2	−0.44	0.36
σ^2	−0.27	0.25

lower intensities to be missing more frequently, and demonstrates the superiority of the NMAR joint model over the MAR joint model in estimating mean group intensities.

Interestingly, the variance component parameters σ_N^2 , σ_B^2 , and σ^2 were almost always underestimated.

Figure F.2 reveals that varying the proportion of missing observations has little impact on the ability of the model fitting procedure to recover the parameter values, excepting the ν parameter, for which positive bias of parameter estimates correlates with the number of missing observations.

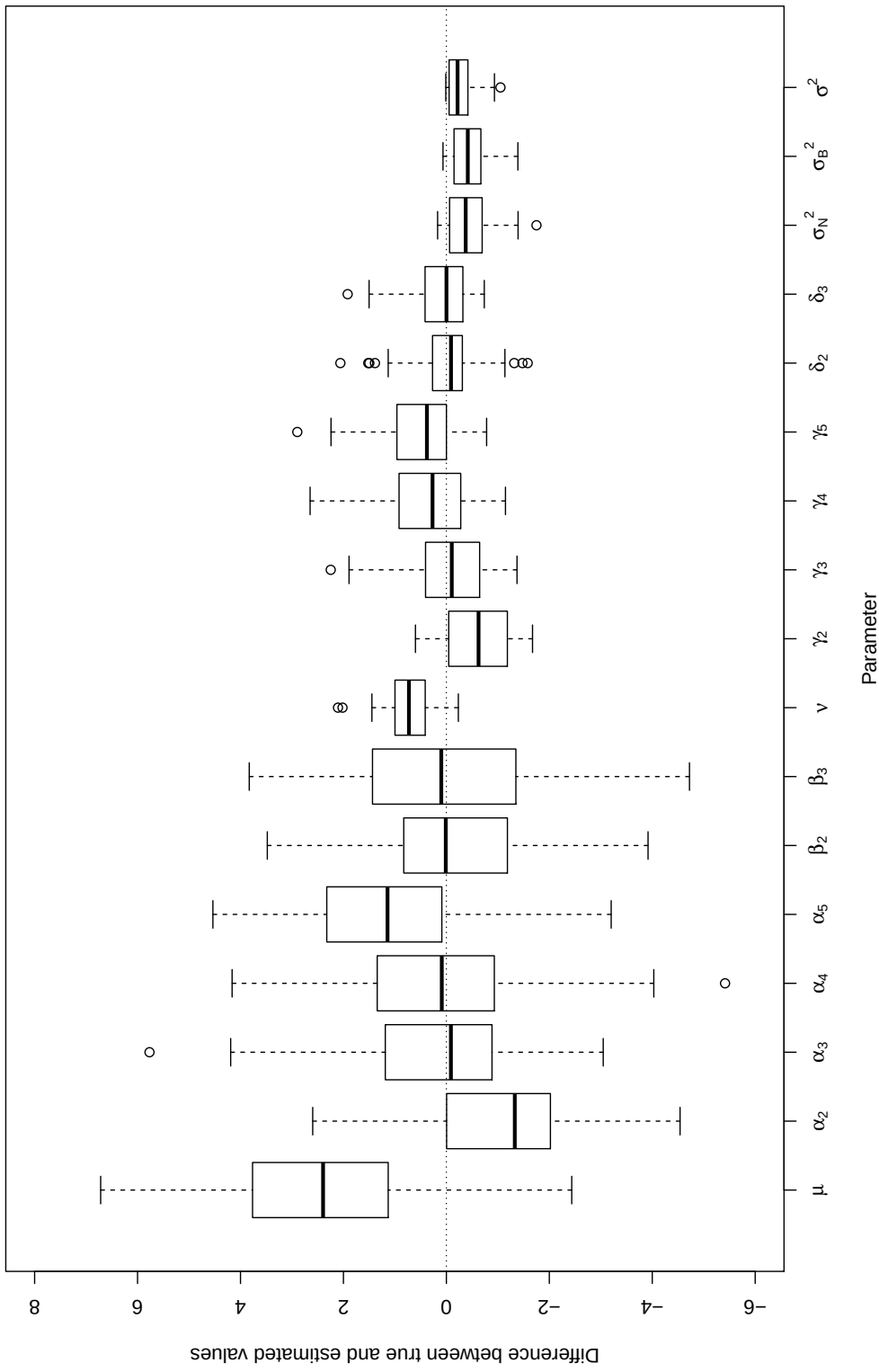

Figure F.1: Boxplots of $N = 60$ differences between estimated and true parameter values from constructed data generation and model-fitting process using the MAR joint model (4.6).

Figures F.3a and F.3b display the correlation of the true parameter values with the estimated parameter values. The estimates of parameters in $\boldsymbol{\lambda}$ show barely any relation with the true values. The estimates of parameters in $\boldsymbol{\kappa}$ tend to be shrunk towards zero. The random effect variance components σ_N^2, σ_B^2, and σ^2 appear to be consistently underestimated, which is consistent with the MAR model not accounting for the fact that the range of data is shrunk by the missingness in the lowest-valued observations.

F.2 Assessing model fits with predictive simulation

The second method of Gelman (2006) to assess a statistical model is to use the parameter estimates to generate data according to the model. Visual inspection and summary statistics of the data provide insight into the suitability of the model.

The data generation procedure was the same as the constructed data procedure from Section F.1 with two exceptions. First, each element of the random effect vectors $\boldsymbol{N}$ and $\boldsymbol{B}$ was determined by the estimates from the `stan` model for a particular peak rather than sampled from a distribution, and second, the vector of missingness indicators $\boldsymbol{r}$ was produced according to

$$\boldsymbol{r} = X\boldsymbol{\lambda}.$$

Figures F.4 and F.5 display data predicted in this way from four m/z peaks across the range of m/z values in the GC dataset. Figure F.6 displays a broader and less detailed comparison of the predicted and true data across a representative subset of 40 peaks. The simulated peak data appear similar to the true datasets. While this similarity is a point in favour of the MAR joint model, the data predicted from the MAR joint model are still not as similar to the true data as the predictions from the NMAR joint model.

One departure from the original GC data that is apparent in the predictions from the MAR joint model is that the missingness mechanism in the predicted datasets does not explicitly act against intensities below the threshold cut-off as it does for the GC data. This is especially visible in the peak at 14836 m/z in Figure F.5. A second departure is that for peaks in Figure F.6 displaying large inter-group and inter-chip differences in missingness, the missingness pattern in the predicted data is slightly more evenly spread across all groups and chips than in the original GC data. This is apparent as a greater amount of contiguous blank and filled spots in the columns of the left array as compared to columns of the right array. This feature of the MAR joint model (4.6) predictions is present to a much smaller degree in the NMAR joint model (4.9) predictions.

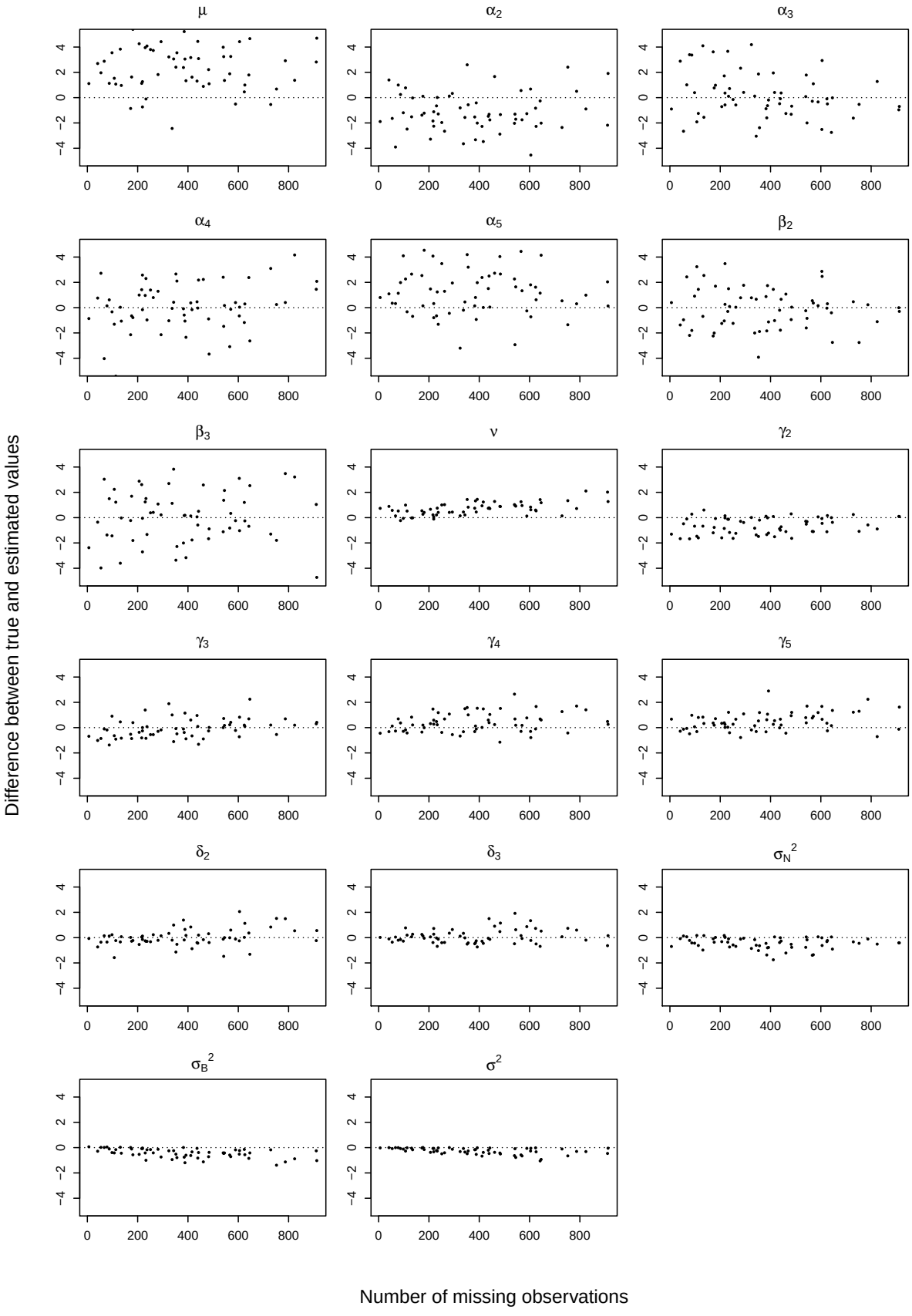

Figure F.2: Plots of $N = 60$ differences between true and estimated parameter values against number of missing observations from constructed data generation and model-fitting process using the MAR joint model (4.6).

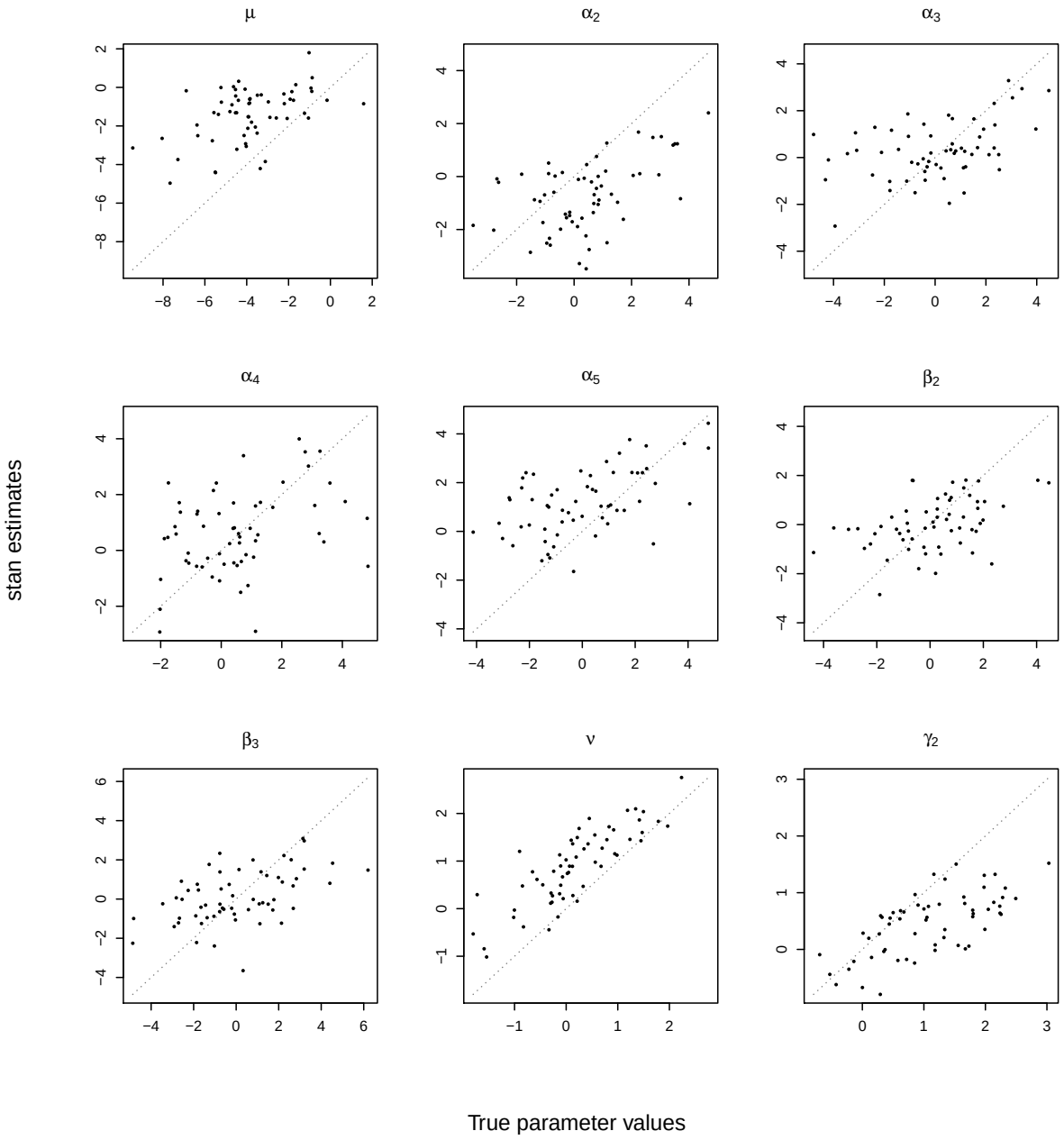

Figure F.3a: Plots of true versus estimated values for parameters across $N = 60$ replications of constructed data generation and model-fitting process for the MAR joint model (4.6) applied to data generated according to the NMAR joint model. Each plot window is devoted to a single parameter and contains the true and estimated values from each of the simulations. Points within plots represent individual simulations. The closer the values are for a given simulation, the closer the point is to the dotted line $y = x$.

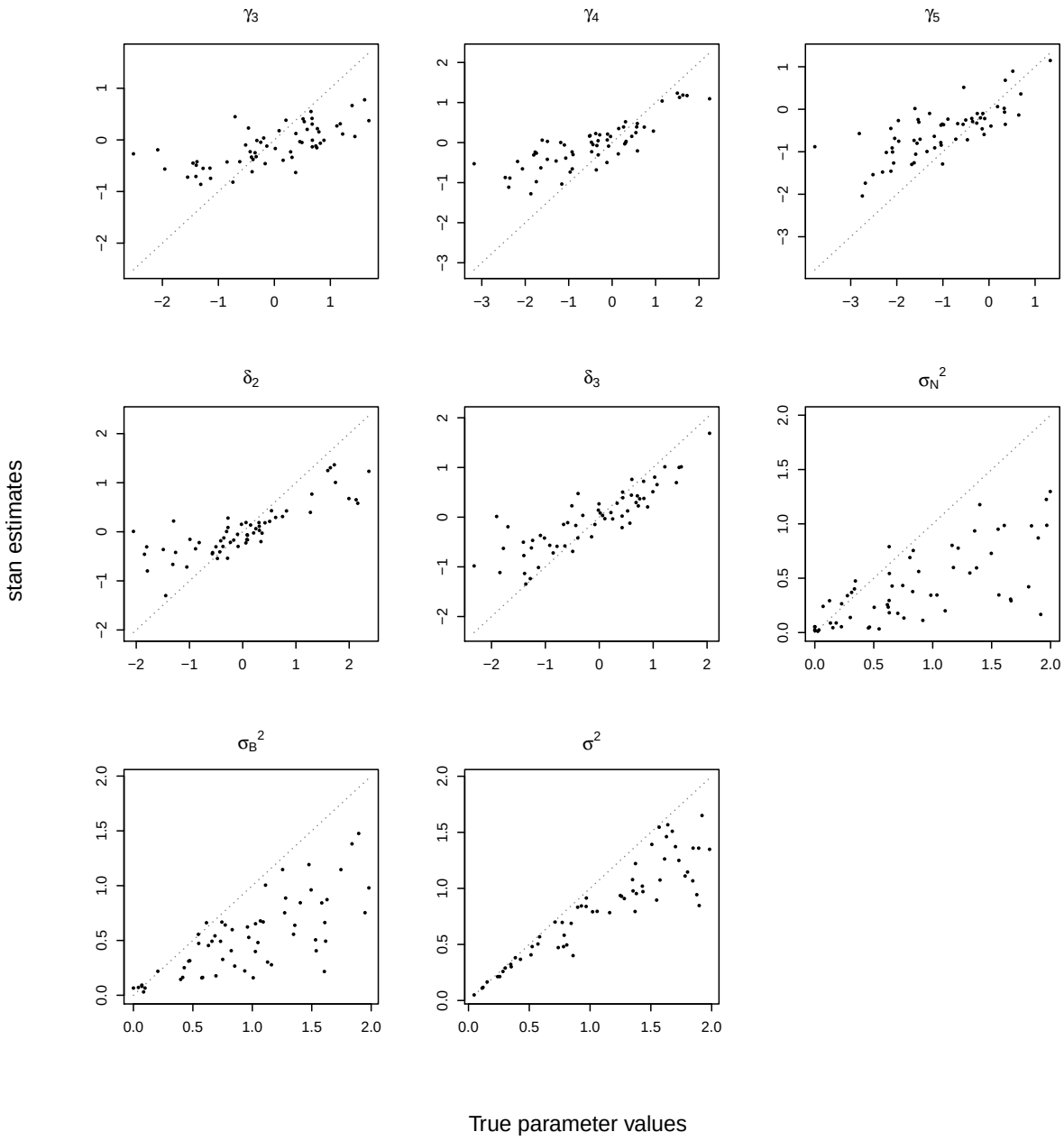

Figure F.3b: Plots of true versus estimated values for parameters across $N = 60$ replications of constructed data generation and model-fitting process for the MAR joint model (4.6) applied to data generated according to the NMAR joint model. Each plot window is devoted to a single parameter and contains the true and estimated values from each of the simulations. Points within plots represent individual simulations. The closer the values are for a given simulation, the closer the point is to the dotted line $y = x$.

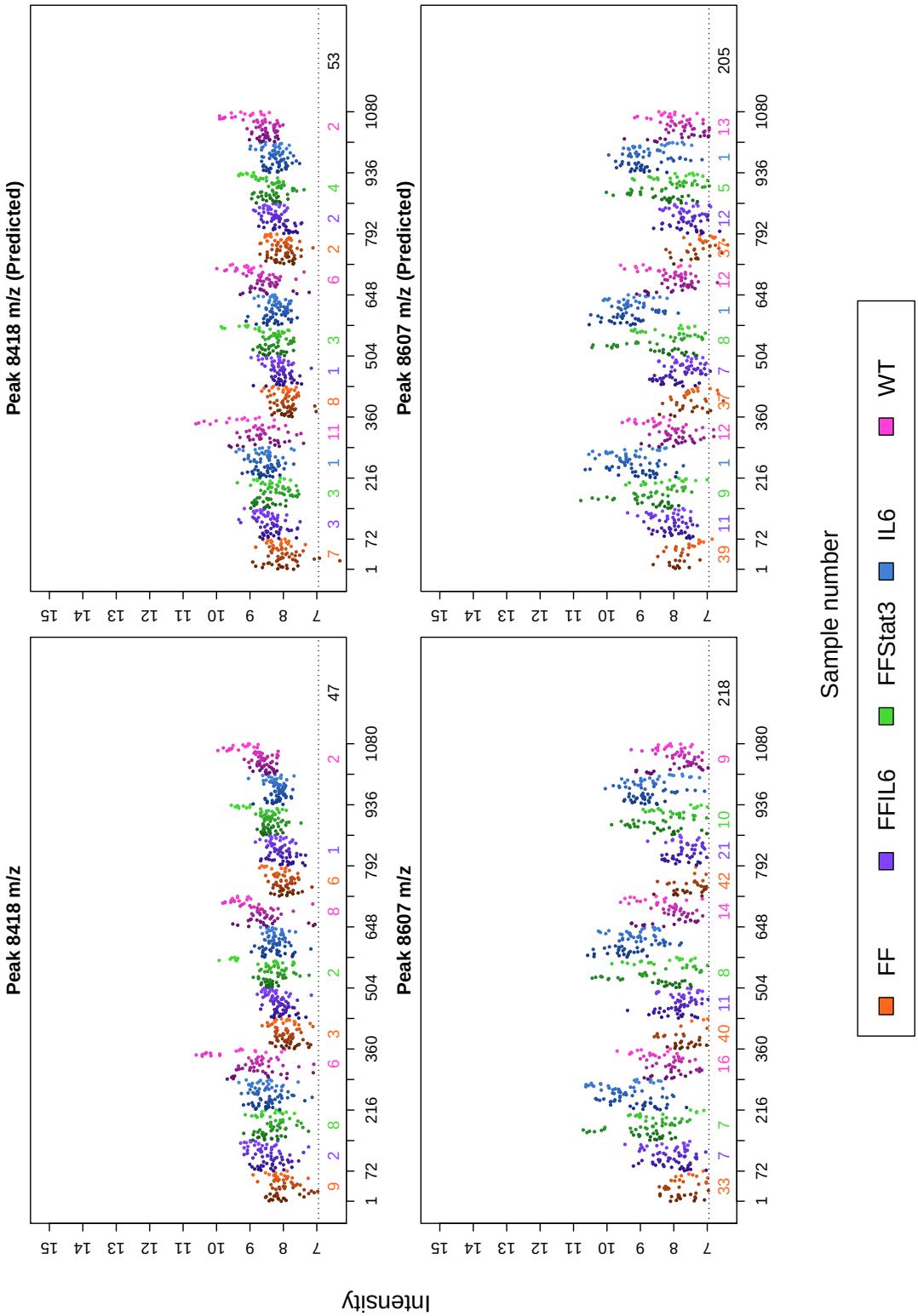

Figure F.4: Comparison of intensities within peaks in the GC dataset (left column) with predicted intensities based on MAR joint model (4.6) (right).

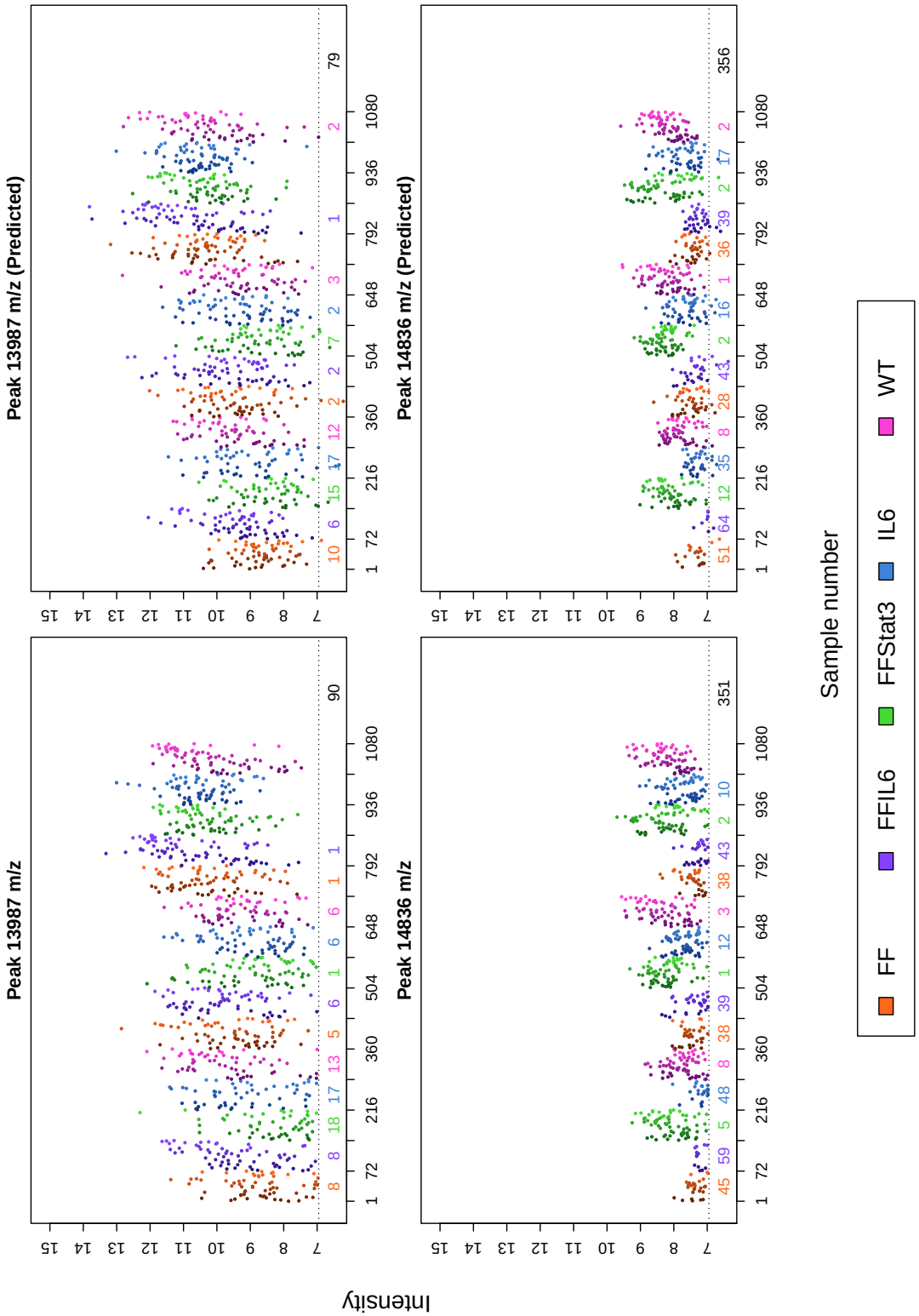

Figure F.5: Comparison of intensities within peaks in the GC dataset (left column) with predicted intensities based on MAR joint model (4.6) (right).

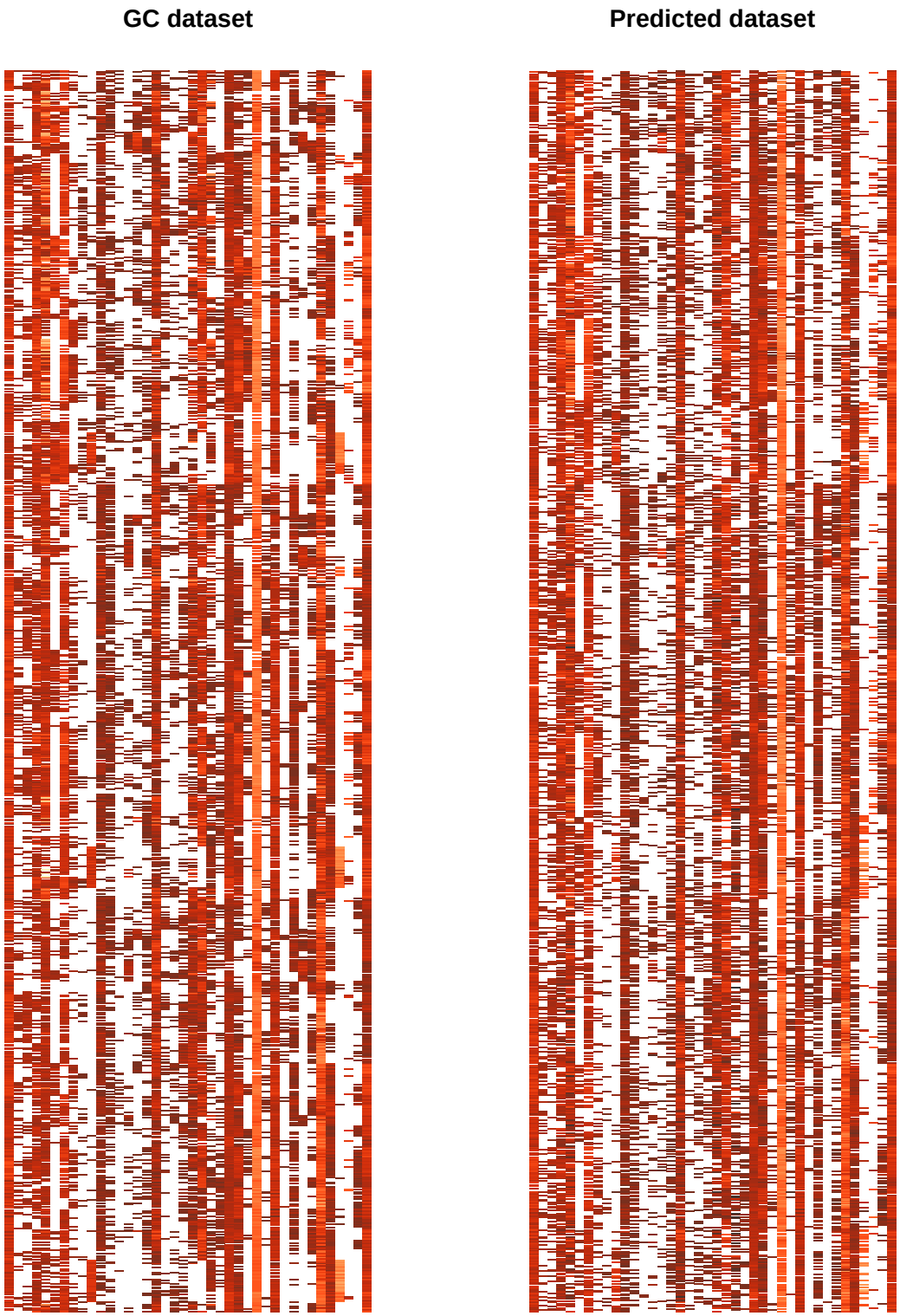

Figure F.6: Comparison of subset of 40 peaks in GC dataset observation matrix (left column) with simulated values based on MAR joint model (4.6) (right).

Bibliography

Adam, B.-L., Qu, Y., Davis, J. W., Ward, M. D., Clements, M. A., Cazares, L. H., Semmes, O. J., Schellhammer, P. F., Yasui, Y., Feng, Z. and others (2002), Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men, *Cancer research* **62**(13), 3609–3614.
`http://cancerres.aacrjournals.org/content/62/13/3609.short` [Accessed 2016-07-14]

Aittokallio, T. (2010), Dealing with missing values in large-scale studies: microarray data imputation and beyond, *Briefings in Bioinformatics* **11**(2), 253–264.
`http://bib.oxfordjournals.org/cgi/doi/10.1093/bib/bbp059` [Accessed 2016-08-09]

Albert, A. and Anderson, J. A. (1984), On the existence of maximum likelihood estimates in logistic regression models, *Biometrika* **71**(1), 1–10.
`http://biomet.oxfordjournals.org/content/71/1/1.short` [Accessed 2016-09-05]

Barndorff-Nielsen, O. E. (1991), Likelihood Theory, *in* D. V. Hinkley, N. Reid and E. J. Snell, eds, 'Statistical Theory and Modelling', 1st edn, Chapman and Hall, pp. 55–82.

Bates, D., Mchler, M., Bolker, B. and Walker, S. (2015), Fitting linear mixed-effects models using lme4, *Journal of Statistical Software* **67**(1).
`http://www.jstatsoft.org/v67/i01/` [Accessed 2017-02-25]

Beavis, R. C. and Chait, B. T. (1996), Matrix-assisted laser desorption ionization mass-spectrometry of proteins, Vol. 270 of *High Resolution Separation and Analysis of Biological Macromolecules Part A: Fundamentals*, Academic Press, pp. 519–551.
`http://www.sciencedirect.com/science/article/pii/S0076687996700241` [Accessed 2016-07-14]

Benjamini, Y. and Hochberg, Y. (1995), Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)* **57**(1), 289–300.
`http://www.jstor.org/stable/2346101` [Accessed 2018-07-09]

Betancourt, M. (2017), A conceptual introduction to Hamiltonian Monte Carlo. http://arxiv.org/abs/1701.02434 [Accessed 2017-05-03]

Callesen, A. K., Christensen, R. d., Madsen, J. S., Vach, W., Zapico, E., Cold, S., Jrgensen, P. E., Mogensen, O., Kruse, T. A. and Jensen, O. N. (2008), Reproducibility of serum protein profiling by systematic assessment using solid-phase extraction and matrix-assisted laser desorption/ionization mass spectrometry, *Rapid Communications in Mass Spectrometry* **22**(3), 291–300. http://doi.wiley.com/10.1002/rcm.3364 [Accessed 2016-07-14]

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P. and Riddell, A. (2017), Stan: A probabilistic programming language, *Journal of Statistical Software* **76**(1). https://www.jstatsoft.org/article/view/v076i01 [Accessed 2018-08-18]

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A. and Liu, J. (2013), A nondegenerate penalized likelihood estimator for variance parameters in multilevel models, *Psychometrika* **78**(4), 685–709. https://link.springer.com/article/10.1007/s11336-013-9328-2 [Accessed 2018-01-24]

Coghlin, C. and Murray, G. I. (2016), Progress in the development of protein biomarkers of oesophageal and gastric cancers, *PROTEOMICS - Clinical Applications* **10**(4), 532–545. http://doi.wiley.com/10.1002/prca.201500079 [Accessed 2017-10-17]

Cohen, J. (1960), A coefficient of agreement for nominal scales, *Educational and psychological measurement* **20**(1). ftp://gis.msl.mt.gov/Maxell/Models/Predictive_Modeling_for_DSS_Lincoln_NE_121510/Modeling_Literature/Cohen_1960.pdf [Accessed 2016-11-24]

Colby, E. and Bair, E. J. (2013), Cross-validation for nonlinear mixed effects models, *Journal of Pharmacokinetics and Pharmacodynamics* **40**(2), 243252. https://link.springer.com/article/10.1007%2Fs10928-013-9313-5 [Accessed 2019-02-24]

Corbeil, R. R. and Searle, S. R. (1976), Restricted maximum likelihood (REML) Eestimation of variance components in the mixed model, *Technometrics* **18**(1), 31. http://www.jstor.org/stable/1267913?origin=crossref [Accessed 2016-11-10]

Davison, A. (2003), *Statistical Models*, Cambridge University Press, Cambridge.

Diggle, P. and Kenward, M. G. (1994), Informative drop-out in longitudinal data analysis, *Applied Statistics* **43**(1), 49.
`http://www.jstor.org/stable/2986113?origin=crossref` [Accessed 2017-04-20]

Duane, S., Kennedy, A. D., Pendleton, B. J. and Roweth, D. (1987), Hybrid monte carlo, *Physics letters B* **195**(2), 216–222.
`http://www.sciencedirect.com/science/article/pii/037026938791197X` [Accessed 2017-05-03]

Firth, D. (1991), Generalised Linear Models, *in* D. V. Hinkley, N. Reid and E. J. Snell, eds, 'Statistical Theory and Modelling', 1st edn, Chapman and Hall, pp. 55–82.

Flegal, J. M. and Jones, G. L. (2011), Implementing MCMC: Estimating with confidence, *in* S. Brooks, A. Gelman and G. Jones, eds, 'Handbook of Markov chain Monte Carlo', Chapman and Hall/CRC handbooks of modern statistical methods, CRC Press.

Follmann, D. and Wu, M. (1995), An approximate generalized linear model with random effects for informative missing data, *Biometrics* **51**(1), 151–168.
`http://www.jstor.org/stable/2533322` [Accessed 2017-03-17]

Gelman, A. (2006), Multilevel (hierarchical) modeling: What it can and cannot do, *Technometrics* **48**(3), 432–435.
`http://www.tandfonline.com/doi/abs/10.1198/004017005000000661` [Accessed 2017-04-04]

Gelman, A. and Hill, J. (2009), *Data analysis using regression and multilevel/hierarchical models*, Cambridge University Press.

Gelman, A. and Rubin, D. B. (1992), Inference from iterative simulation using multiple sequences, *Statistical science* **7**(4), 457–472.
`http://www.jstor.org/stable/2246093` [Accessed 2017-08-14]

Gelman, A. and Shirley, K. (2011), Inference from simulations and monitoring convergence, *in* S. Brooks, A. Gelman, G. Jones and X.-L. Meng, eds, 'Handbook of Markov chain Monte Carlo', Chapman and Hall/CRC handbooks of modern statistical methods, CRC Press.

Geyer, C. J. (2011), Introduction to Markov chain Monte Carlo, *in* S. Brooks, A. Gelman, G. Jones and X.-L. Meng, eds, 'Handbook of Markov chain Monte Carlo', Chapman and Hall/CRC handbooks of modern statistical methods, CRC Press.

Gianazza, E., Miller, I., Palazzolo, L., Parravicini, C. and Eberini, I. (2016), With or without you - Proteomics with or without major plasma/serum proteins, *Journal of Proteomics* **140**, 62–80.

http://linkinghub.elsevier.com/retrieve/pii/S1874391916301166 [Accessed 2017-10-17]

Gilks, W. R. (2005), Markov chain Monte Carlo, *in* P. Armitage and T. Colton, eds, 'Encyclopedia of Biostatistics', John Wiley & Sons, Ltd, Chichester, UK. http://doi.wiley.com/10.1002/0470011815.b2a14021 [Accessed 2017-05-03]

Glish, G. L. and Vachet, R. W. (2003), The basics of mass spectrometry in the twenty-first century, *Nature Reviews Drug Discovery* **2**(2), 140–150. http://www.nature.com/doifinder/10.1038/nrd1011 [Accessed 2016-07-14]

Gould, L., Boye, M. E., Crowther, M. J., Ibrahim, J. G., Quartey, G., Micallef, S. and Bois, F. Y. (2015), Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group, *Statistics in Medicine* **34**(14), 2181–2195. http://onlinelibrary.wiley.com/doi/10.1002/sim.6141/abstract [Accessed 2017-01-24]

Graham, J. W. (2012), *Missing Data*, Springer New York, New York, NY. http://link.springer.com/10.1007/978-1-4614-4018-5 [Accessed 2016-07-14]

Hajduk, J., Matysiak, J. and Kokot, Z. J. (2016), Challenges in biomarker discovery with MALDI-TOF MS, *Clinica Chimica Acta* **458**, 84–98. http://linkinghub.elsevier.com/retrieve/pii/S0009898116301589 [Accessed 2017-01-05]

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning*, 2nd edn, Springer.

Hastings, W. K. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**(1), 97. http://www.jstor.org/stable/2334940?origin=crossref [Accessed 2017-05-04]

Heinze, G. and Schemper, M. (2002), A solution to the problem of separation in logistic regression, *Statistics in medicine* **21**(16), 2409–2419. http://onlinelibrary.wiley.com/doi/10.1002/sim.1047/full [Accessed 2016-09-05]

Hogan, J. W. and Laird, N. M. (1997), Model-based approaches to analysing incomplete longitudinal and failure time data, *Statistics in Medicine* **16**(3), 259–272. http://onlinelibrary.wiley.com.proxy.library.adelaide.edu.au/doi/10.1002/(SICI)1097-0258(19970215)16:3<259::AID-SIM484>3.0.CO;2-S/abstract [Accessed 2017-03-03]

Humphries, J. M., Penno, M. A., Weiland, F., Klingler-Hoffmann, M., Zuber, A., Boussioutas, A., Ernst, M. and Hoffmann, P. (2014), Identification and validation of novel candidate protein biomarkers for the detection of human gastric cancer, *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1844**(5), 1051–1058. `http://linkinghub.elsevier.com/retrieve/pii/S1570963914000211` [Accessed 2016-07-14]

Jenkins, B. J., Grail, D., Nheu, T., Najdovska, M., Wang, B., Waring, P., Inglese, M., McLoughlin, R. M., Jones, S. A., Topley, N., Baumann, H., Judd, L. M., Giraud, A. S., Boussioutas, A., Zhu, H.-J. and Ernst, M. (2005), Hyperactivation of Stat3 in gp130 mutant mice promotes gastric hyperproliferation and desensitizes TGF- signaling, *Nature Medicine* **11**(8), 845–852. `http://www.nature.com/doifinder/10.1038/nm1282` [Accessed 2016-07-14]

Jordan, L., Daniels, R. F., Clark, A. I. and He, R. (2005), Multilevel nonlinear mixed-effects models for the modeling of earlywood and latewood microfibril angle, *Forest Science* **51**(4), 357371. `https://academic.oup.com/forestscience/article/51/4/357/4617605` [Accessed 2019-02-24]

Judd, L. M., Ulaganathan, M., Howlett, M. and Giraud, A. S. (2009), Cytokine signalling by gp130 regulates gastric mucosal healing after ulceration and, indirectly, antral tumour progression, *The Journal of Pathology* **217**(4), 552–562. `http://doi.wiley.com/10.1002/path.2479` [Accessed 2016-07-14]

Jung, K. (2016), Statistical Aspects in Proteomic Biomarker Discovery, *in* K. Jung, ed., 'Statistical Analysis in Proteomics', Vol. 1362 of *Methods in Molecular Biology*, Springer New York, New York, NY. `http://link.springer.com/10.1007/978-1-4939-3106-4` [Accessed 2016-08-09]

Jung, K., Dihazi, H., Bibi, A., Dihazi, G. H. and Beissbarth, T. (2014), Adaption of the global test idea to proteomics data with missing values, *Bioinformatics* **30**(10), 1424–1430. `http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btu062` [Accessed 2016-07-14]

Jung, K., Gannoun, A., Sitek, B., Apostolov, O., Schramm, A., Meyer, H. E., Sthler, K. and Urfer, W. (2006), Statistical evaluation of methods for the analysis of dynamic protein expression data from a tumor study, *RevStat-Statistical Journal* **4**, 67–80. `https://ine.pt/revstat/pdf/rs060104.pdf` [Accessed 2016-08-09]

Jung, K., Gannoun, A., Sitek, B., Meyer, H. E., Sthler, K. and Urfer, W. (2005), Analysis of dynamic protein expression data, *RevStat-Statistical Journal* **3**, 99–111.

Karpievitch, Y., Stanley, J., Taverner, T., Huang, J., Adkins, J. N., Ansong, C., Heffron, F., Metz, T. O., Qian, W.-J., Yoon, H., Smith, R. D. and Dabney, A. R. (2009), A statistical framework for protein quantitation in bottom-up MS-based proteomics, *Bioinformatics* **25**(16), 2028–2034.
`http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp362` [Accessed 2016-08-09]

Karpievitch, Y. V., Dabney, A. R. and Smith, R. D. (2012), Normalization and missing value imputation for label-free LC-MS analysis, *BMC bioinformatics* **13**(16), S5.

Kenward, M. G. and Molenberghs, G. (1999), Parametric models for incomplete continuous and categorical longitudinal data, *Statistical Methods in Medical Research* **8**(1), 51–83.
`http://smm.sagepub.com/content/8/1/51.full.pdf` [Accessed 2017-04-20]

Ky, B., Putt, M., Sawaya, H., French, B., Januzzi, J. L., Sebag, I. A., Plana, J. C., Cohen, V., Banchs, J., Carver, J. R., Wiegers, S. E., Martin, R. P., Picard, M. H., Gerszten, R. E., Halpern, E. F., Passeri, J., Kuter, I. and Scherrer-Crosbie, M. (2014), Early increases in multiple biomarkers predict subsequent cardiotoxicity in patients with breast cancer treated with doxorubicin, taxanes, and trastuzumab, *Journal of the American College of Cardiology* **63**(8), 809–816.
`http://linkinghub.elsevier.com/retrieve/pii/S0735109713061573` [Accessed 2017-01-24]

Lazar, C., Gatto, L., Ferro, M., Bruley, C. and Burger, T. (2016), Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies, *Journal of Proteome Research* **15**(4), 1116–1125.
`http://pubs.acs.org/doi/abs/10.1021/acs.jproteome.5b00981` [Accessed 2016-07-14]

Lee, Y. and Nelder, J. A. (1996), Hierarchical generalized linear models, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(4), 619–678.
`http://www.jstor.org/stable/2346105` [Accessed 2017-06-29]

Li, F., Nie, L., Wu, G., Qiao, J. and Zhang, W. (2011), Prediction and characterization of missing proteomic data in desulfovibrio vulgaris, *International Journal of Genomics* **2011**, 16.
`http://www.hindawi.com/journals/ijg/2011/780973/abs/,http://www.hindawi.com/journals/ijg/2011/780973/abs/` [Accessed 2016-08-09]

Lin, L.-L., Huang, H.-C. and Juan, H.-F. (2012), Discovery of biomarkers for gastric cancer: A proteomics approach, *Journal of Proteomics* **75**(11), 3081–3097.

http://linkinghub.elsevier.com/retrieve/pii/S1874391912001972 [Accessed 2016-07-14]

Little, R. J. A. (1995), Modeling the drop-out mechanism in repeated-measures studies, *Journal of the American Statistical Association* **90**(431), 1112. http://www.jstor.org/stable/2291350?origin=crossref [Accessed 2017-04-20]

Little, R. J. A. (2008), Selection and pattern-mixture models, *in* G. M. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs, eds, 'Longitudinal Data Analysis', Chapman and Hall/CRC handbooks of modern statistical methods, Chapman & Hall/CRC, Boca Raton, FL, pp. 409–431. https://www.hsph.harvard.edu/fitzmaur/lda/C6587_C018.pdf [Accessed 2017-04-21]

Little, R. and Rubin, D. (2002), *Statistical Analysis with Missing Data*, 2nd edn, John Wiley & Sons, Inc.

Liu, W., Liu, B., Cai, Q., Li, J., Chen, X. and Zhu, Z. (2012), Proteomic identification of serum biomarkers for gastric cancer using multi-dimensional liquid chromatography and 2d differential gel electrophoresis, *Clinica Chimica Acta* **413**(13-14), 1098–1106. http://linkinghub.elsevier.com/retrieve/pii/S0009898112001131 [Accessed 2016-07-14]

Merchant, M. and Weinberger, S. R. (2000), Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry, *ELECTROPHORESIS* **21**(6), 1164–1177. http://onlinelibrary.wiley.com.proxy.library.adelaide.edu.au/doi/10.1002/(SICI)1522-2683(20000401)21:6<1164::AID-ELPS1164>3.0.CO;2-0/abstract [Accessed 2016-07-14]

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), Equation of state calculations by fast computing machines, *The Journal of Chemical Physics* **21**(6), 1087–1092. http://aip.scitation.org/doi/10.1063/1.1699114 [Accessed 2017-05-03]

Neal, R. M. (2011), MCMC using Hamiltonian dynamics, *in* 'Handbook of Markov chain Monte Carlo', Chapman and Hall/CRC handbooks of modern statistical methods, CRC Press.

Newman, D. A. (2014), Missing data: Five practical guidelines, *Organizational Research Methods* **17**(4), 372–411. http://journals.sagepub.com/doi/10.1177/1094428114548590 [Accessed 2017-02-14]

Nielsen, S. F. (2003), Proper and improper multiple imputation, *International Statistical Review* **71**(3), 593–607.
http://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.2003.tb00214.x/abstract [Accessed 2016-08-09]

Nummelin, E. (2002), MC's for MCMC'ists, *International Statistical Review / Revue Internationale de Statistique* **70**(2), 215.
http://www.jstor.org/stable/10.2307/1403908?origin=crossref [Accessed 2017-05-02]

Papaspiliopoulos, O., Roberts, G. O. and Skld, M. (2007), A general framework for the parametrization of hierarchical models, *Statistical Science* **22**(1), 59–73.
https://www.jstor.org/stable/27645805 [Accessed 2018-08-21]

Patterson, H. D. and Thompson, R. (1971), Recovery of inter-block information when block sizes are unequal, *Biometrika* **58**(3), 545–554.
https://www.jstor.org/stable/2334389 [Accessed 2018-08-16]

Pedreschi, R., Hertog, M. L. A. T. M., Carpentier, S. C., Lammertyn, J., Robben, J., Noben, J.-P., Panis, B., Swennen, R. and Nicola, B. M. (2008), Treatment of missing values for multivariate statistical analysis of gel-based proteomics data, *PROTEOMICS* **8**(7), 1371–1383.
http://doi.wiley.com/10.1002/pmic.200700975 [Accessed 2016-07-14]

Penno, M. A., Klingler-Hoffmann, M., Brazzatti, J. A., Boussioutas, A., Putoczki, T., Ernst, M. and Hoffmann, P. (2012), 2d-DIGE analysis of sera from transgenic mouse models reveals novel candidate protein biomarkers for human gastric cancer, *Journal of Proteomics* **77**, 40–58.
http://linkinghub.elsevier.com/retrieve/pii/S1874391912004964 [Accessed 2016-07-14]

Pigott, T. D. (2001), A review of methods for missing data, *Educational research and evaluation* **7**(4), 353–383.
http://www.tandfonline.com/doi/abs/10.1076/edre.7.4.353.8937 [Accessed 2016-07-14]

Pinheiro, J. C. and Bates, D. M. (2000), *Mixed-effects models in S and S-PLUS / Jos C. Pinheiro, Douglas M. Bates.*, Statistics and computing, Springer, New York.

R Core Team (2016), 'R: A language and environment for statistical computing'.
https://www.R-project.org [Accessed 2016-07-23]

Raudenbush, S. W., Yang, M.-L. and Yosef, M. (2000), Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace

approximation, *Journal of Computational and Graphical Statistics* **9**(1), 141. http://www.jstor.org/stable/1390617?origin=crossref [Accessed 2016-12-07]

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schrder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F. and Dormann, C. F. (2017), Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, *Ecograph* **40**, 913–929. https://onlinelibrary.wiley.com/doi/epdf/10.1111/ecog.02881 [Accessed 2019-06-15]

Rosa, E.-B. (2013), MALDI: A very useful UV light-induced process that still remains quite obscure, *in* K. Hiraoka, ed., 'Fundamentals of Mass Spectrometry', Springer New York, New York, NY. http://link.springer.com/10.1007/978-1-4614-7233-9 [Accessed 2016-07-14]

Rubin, D. B. (1976), Inference and Missing Data, *Biometrika* **63**(3), 581. http://www.jstor.org/stable/2335739?origin=crossref [Accessed 2017-02-14]

Searle, S. R., Casella, G. and McCulloch, C. E. (2009), *Variance components*, Vol. 391, John Wiley & Sons.

Snijders and Bosker (2012), *Multilevel analysis: an introduction to basic and advanced multilevel modeling*, 2nd edn, SAGE, Los Angeles.

Stanford, T. (2015), Statistical analysis of proteomic mass spectrometry data for the identification of biomarkers and disease diagnosis., PhD thesis, The University of Adelaide. http://hdl.handle.net/2440/97452 [Accessed 2016-07-14]

Tebbutt, N. C., Giraud, A. S., Inglese, M., Jenkins, B., Waring, P., Clay, F. J., Malki, S., Alderman, B. M., Grail, D., Hollande, F., Heath, J. K. and Ernst, M. (2002), Reciprocal regulation of gastrointestinal homeostasis by SHP2 and STAT-mediated trefoil gene activation in gp130 mutant mice, *Nature Medicine* **8**(10), 1089–1097. http://www.nature.com/doifinder/10.1038/nm763 [Accessed 2016-07-14]

Terp, M. G. and Ditzel, H. J. (2014), Application of proteomics in the study of rodent models of cancer, *PROTEOMICS - Clinical Applications* **8**(9-10), 640–652. http://doi.wiley.com/10.1002/prca.201300084 [Accessed 2017-10-17]

Tuerlinckx, F., Rijmen, F., Verbeke, G. and De Boeck, P. (2006), Statistical inference in generalized linear mixed models: A review, *British Journal of Mathematical and Statistical Psychology* **59**(2), 225–255. http://onlinelibrary.wiley.com/doi/10.1348/000711005X79857/abstract [Accessed 2018-02-12]

Webb-Robertson, B.-J. M., McCue, L. A., Waters, K. M., Matzke, M. M., Jacobs, J. M., Metz, T. O., Varnum, S. M. and Pounds, J. G. (2010), Combined statistical analyses of peptide intensities and peptide occurrences improves identification of significant peptides from MS-based proteomics data, *Journal of Proteome Research* **9**(11), 5748–5756. `http://pubs.acs.org/doi/abs/10.1021/pr1005247` [Accessed 2016-07-14]

Webb-Robertson, B.-J. M., Wiberg, H. K., Matzke, M. M., Brown, J. N., Wang, J., McDermott, J. E., Smith, R. D., Rodland, K. D., Metz, T. O., Pounds, J. G. and Waters, K. M. (2015), Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics, *Journal of Proteome Research* **14**(5), 1993–2001.
`http://pubs.acs.org/doi/abs/10.1021/pr501138h` [Accessed 2016-08-09]

Wood, J., White, I. R. and Cutler, P. (2004), A likelihood-based approach to defining statistical significance in proteomic analysis where missing data cannot be disregarded, *Signal Processing* **84**(10), 1777–1788.
`http://linkinghub.elsevier.com/retrieve/pii/S0165168404001410` [Accessed 2016-08-09]

Wu, M. C. and Carroll, R. J. (1988), Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process, *Biometrics* **44**(1), 175.
`http://www.jstor.org/stable/2531905?origin=crossref` [Accessed 2017-10-19]

Yates III, J. R. (2011), A century of mass spectrometry: from atoms to proteomes, *nature methods* **8**(8), 633.
`http://www.imbb.forth.gr/ProFI/pdf/nmeth_1659.pdf` [Accessed 2016-07-14]

Zorn, C. (2005), A solution to separation in binary response models, *Political Analysis* **13**(2), 157–170.
`http://pan.oupjournals.org/cgi/doi/10.1093/pan/mpi009` [Accessed 2016-09-05]