Interuniversity Institute for Biostatistics
and statistical Bioinformatics

Dissertation submitted in fulfilment of the requirements for the degree of
Doctor of Science: Statistics

Trishanta Padayachee

DOCTORAL DISSERTATION

Statistical Methods for Transcriptomic and Metabolomic Data Analysis

Supervisor: Prof. Dr. Tomasz Burzykowski | UHasselt
Co-supervisor: Prof. Dr. Ziv Shkedy | UHasselt

September 2018

Acknowledgements

During this experience, many people gave colour and direction to my path. Although I cannot thank them all, I would like to seize this opportunity to thank a few.

I would like to express my sincere gratitude to my promoter Prof. dr. Tomasz Burzykowski. Tomasz, I appreciate the time you dedicated towards this research. Our meetings always provided clarity on the path ahead. I am grateful for your academic guidance and enthusiasm. Your continued support and the flexibility you permitted was invaluable for the completion of this dissertation. I am also grateful to my co-promoter Prof. dr. Ziv Shkedy. Ziv, thank you for sharing your academic insights on the various statistical problems I encountered along the way.

I very much appreciate the interactions I had with Prof. dr. Marc Aerts and Prof. dr. Geert Molenberghs. Despite the brevity of our meetings, they were extremely insightful. I'd also like to extend my gratitude to Dr. Tatsiana Khamiakova. Tatsiana, thank you for your guidance and support at the beginning of this journey. I appreciate the knowledge you shared on identifying metabolites in ^{1}H-NMR spectra and our further discussions on the ^{1}H-NMR analysis. Thanks also to Prof. dr. Peter Adriaenssens for his explanations on the technical aspects of ^{1}H-NMR spectroscopy. Many thanks to Dr. Jurgen Claesen for his support and for helping with the Dutch translation of the summary.

I would like to extend my gratitude to every member of my jury: Prof. dr. Jeanine Houwing-Duistermaat, Prof. dr. Przemyslaw Biecek, Prof. dr. Peter Adriaensens, and Prof. dr. Marc Aerts, for taking the time to read my dissertation and for their constructive comments and suggestions.

I gratefully acknowledge the funding I received from the MIMOmics grant of the European Union's Seventh Framework Programme (FP7-Health-F5-2012) and the support of the IAP Research Network of the Belgian state (Belgian Science Policy).

I appreciate the help of the CenStat secretariat and members of the HR department for their administrative support. I am also thankful to my colleagues in CenStat (Fatemeh, Leacky, Theophile, Belay, Marijke, Ewoud, Nolen, Kate, Rudradev, Thao, Thang, Alvaro, Joris, Oana, Annelies, Daniel, Evangelina, ...) and JOSS for the fun and friendly interactions.

I have the utmost gratitude for my dear friends: Renata, Martin, Yimer, Eva, Jimmy, Mohammed, and Tarylee. Your friendship through these years has made this experience exceedingly more valuable. Tarylee, embarking on this journey with you was a wonderful coincidence. It was always incredibly exciting when you visited Belgium and I thoroughly enjoyed our adventures together. Thank you for your constant support and encouragement.

I dedicate this dissertation to my parents, Umsha Padayachee and Dr. Krish Padayachee. I am eternally grateful to you. The knowledge and wisdom you have imparted, and the efforts you have made, set the foundation for achieving this goal. I am also particularly thankful for your assistance and encouragement during the final stretch. I am grateful to my siblings - Avashni, Dasevan, and Arushin - for our hilarious conversations which offered welcome distractions.

I wish to thank my extended family in South Africa and Belgium for their support.

Finally, and most importantly, I am extremely appreciative of my partner. Koen, thank you for your unwavering support and assistance through the tough times. The help you provided on a daily basis has been invaluable for the completion of this dissertation.

Trishanta Padayachee
Hasselt, September 2018

List of Publications

The contents of this dissertation are based on the following publications:

Padayachee, T., Khamiakova, T., Shkedy, Z., Perola, M., Salo, P., Burzykowski, T. (2016) The Detection of Metabolite-Mediated Gene Module Co-Expression Using Multivariate Linear Models. *PLoS ONE* 11(2). doi:10.1371/journal.pone.0150257

Padayachee, T., Khamiakova, T., Shkedy, Z., Salo, P., Perola, M., Burzykowski, T. A multivariate linear model for investigating the association between gene-module co-expression and a continuous covariate. [Under revision for resubmission to *Statistical Applications in Genetics and Molecular Biology*]

Padayachee, T., Khamiakova, T., Louis, E., Adriaensens, P., Burzykowski, T. The impact of the method of extracting metabolic signal from ^{1}H-NMR data on the classification of samples: a case study of binning and BATMAN in lung cancer. [Under revision for resubmission to *PLoS ONE*]

Padayachee, T., Shkedy, Z., Salo, P., Perola, M., Burzykowski, T. A copula-based pseudo-likelihood approach for investigating the association between gene-module co-expression and a continuous covariate. [In preparation]

Contents

List of Figures

List of Tables

1 Overview of the dissertation

1.1 Outline of the dissertation

A way to enhance our understanding of the development and progression of complex diseases is to investigate the influence of cellular environments on gene co-expression (i.e., gene-pair correlations). Investigating whether metabolites regulate the co-expression of a predefined gene module (a set of co-expressed (correlated) genes belonging to the same biological pathway) is one of the relevant questions posed in the integrative analysis of metabolomic and transcriptomic data (Inouye et al., 2010a). In Part I of this dissertation, three statistical models are described for investigating the association between gene-module co-expression and metabolite concentrations. The suitability and versatility of the proposed models are investigated through simulation studies and an application to real-life data. Specifically, using a subset of the DILGOM (**DI**etary, **L**ifestyle, and **G**enetic determinants of **O**besity and **M**etabolic syndrome) study data (Inouye et al., 2010a), the proposed models are used to study the association between the co-expression of the core Lipid-Leukocyte (LL) gene module (Inouye et al., 2010b), which is of relevance to coronary artery disease, and serum-metabolite concentrations.

An introduction to conditional co-expression analysis and the motivation behind this research is provided in Chapter 2. The DILGOM data are described in Chapter 3. In Chapter 4, we propose a multivariate linear model for studying the dependence between categorised metabolite concentrations and gene-module co-expression. Performance of statistical tests for the inference of conditional co-expression are evaluated through a simulation study. The proposed methodology is applied to the gene-expression data of the core lipid-leukocyte gene module.

Often, changes in gene co-expression are investigated across two or more biological conditions defined by categorizing a continuous covariate. However, the selection of arbitrary cut-off points may have an influence on the results of an analysis. To address this issue, in Chapter 5, a multivariate linear model for investigating the association between gene-module co-expression and a continuous covariate is proposed. The versatility of the model is illustrated by using a real-life example and a simulation

study.

Background information on pseudo-likelihood estimation and copulas is provided in Chapter 6. In Chapter 7, a pseudo-likelihood approach incorporating conditional copulas is considered for investigating the association between gene-module co-expression and a continuous covariate. As for the previous two models, the copula-based pseudo-likelihood approach is applied to the DILGOM data to investigate the conditional co-expression of the core lipid-leukocyte gene module.

In Part II of this dissertation, the impact of the method for extracting metabolic signal from proton nuclear magnetic resonance (^{1}H-NMR) data on the classification of lung cancer samples is studied. Extracting metabolic information from NMR spectra is complex due to the fact that an immense amount of detail on the chemical composition of a biological sample is expressed through a single spectrum. The simplest approach to quantify the signal is through spectral binning which involves subdividing the spectra into regions along the chemical shift axis and integrating the peaks within each region (Louis et al., 2015). However, due to overlapping resonance signals, the integration values do not always correspond to the concentrations of specific metabolites. An alternate, more advanced statistical approach is spectral deconvolution. BATMAN (**B**ayesian **A**u**T**omated **M**etabolite **A**nalyser for **N**MR data) (Astle et al., 2012; Hao et al., 2014) performs spectral deconvolution using prior information on the spectral signatures of metabolites. In this way, BATMAN estimates relative metabolic concentrations. In this study, both spectral binning and spectral deconvolution using BATMAN were applied to 400 MHz and 900 MHz NMR spectra of blood plasma samples from lung cancer patients and control subjects. The relative concentrations estimated by BATMAN were compared with the binning integration values in terms of their ability to discriminate between lung cancer patients and controls.

An introduction to the ^{1}H-NMR study is provided in Chapter 8. Background information on ^{1}H-NMR spectroscopy is provided in Chapter 9. The data and preprocessing steps are described in Chapter 10. A description of spectral binning and spectral deconvolution using BATMAN is described in Chapter 11. Finally, details and results of the classification analysis appear in Chapter 12.

1.2 Aims of the research

Technological advances have brought about a rapid increase in the high-throughput analysis of biological molecules (genes, mRNA, proteins, metabolites etcetera). The widespread availability of omics (genomics, proteomics, metabolomics, transcriptomics, glycomics, and lipidomics) data has revolutionised medical research (Hasin et al., 2017). The analysis of omics data can lead to the identification of molecular profiles that are associated with disease status, susceptibility, or progression, or it

may provide insight into biological pathways or processes that differ in diseased and control patients. Biological processes are, however, extremely intricate and obtaining biologically meaningful information from this mass of data is a non-trivial task. To capture the complexity of biological processes, research is now centering on the integrative analysis of omics data. In this context, methodological development is lacking, leading to complex data being analysed in rather simple ways that do not capture the complexity of the biological problem. Methodological frameworks for the integrative analysis of multilevel omics datasets are required. The aim of Part I of this dissertation is to improve on the methods currently available for the analysis of omics datasets. The focus is on statistical methods for the integrative analysis of transcriptomic and metabolomic data for investigating the association between gene-module co-expression and metabolite concentrations.

High-throughput techniques enable the measurement of the chemical composition of cells, tissues, or, biofluids. The reproducibility, precision, and inherent noise of the measurements vary between techniques. In some instances, the biological signal may constitute only a small portion of the collected measurements. Efficient extraction of the biological signal is required before the data can be analysed with the aim of gaining insights into complex pathological processes. Various approaches exist to extract biological signal. The approach adopted for extracting the biological signal can have an impact on downstream analyses. The aim of Part II of this dissertation is to investigate the impact of the method of extracting metabolic signal from ^{1}H-NMR spectra on the classification of lung cancer samples.

Part I

Statistical models to investigate the conditional co-expression of a gene module

2 Introduction

Omics technologies have rapidly advanced giving rise to an extensive amount of omics data with widespread availability. *Genomics* is a study of the *genome*, i.e., the collection of the genetic material (including all the genes) of an organism. In contrast to genetics, which focuses on the role of individual genes in, e.g., inheritance of diseases, genomics aims at the characterization and quantification of all the genes involved in a genome. *Transcriptomics*, which is also referred to as expression profiling, is the study of the mRNA molecules arising from the expression of genes in a particular biological sample at a given moment. Thus, the *transcriptome*, i.e., the total mRNA of an organism, reflects the genes that are actively expressed at a given moment. The transcriptome acts as a template for protein synthesis. *Proteomics* studies the *proteome*, i.e., the collection of proteins that are produced by an organism, while *metabolomics* focuses on the *metabolome*, i.e., the collection of metabolites (small molecules), which are the products of cellular processes. As can be seen from these examples, the *omics* suffix indicates a collective analysis of molecules included in the particular study subject, which is described by the *omes* suffix. Figure 2.1 from Euceda et al. (2015) depicts the omics cascade. Metabolites are the final products in the omics cascade. Thus, changes in the metabolome reflect changes in the transcriptome and the proteome. The metabolome is closest molecular measure of the phenotype of the biological system (Horgan and Kenny, 2011). Given the complexity of biological processes, integrative analyses of multiple omics datasets can lead to a better understanding of the molecular basis of complex diseases.

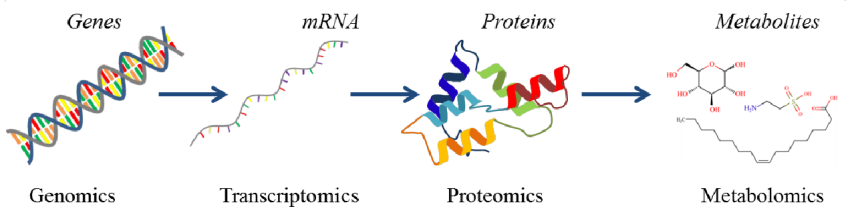

Figure 2.1: Illustration of the omics cascade from genes to metabolites. The source of this Figure is Euceda et al. (2015).

In this part of the dissertation, we focus on the integrative analysis of metabolomic and transcriptomic data specifically for investigating the association between the co-expression of a gene module (a set of co-expressed (correlated) genes belonging to the same biological pathway) and metabolic concentrations.

Genome-wide gene expression data, often obtained through microarray experiments, were initially analysed (and continue to be analysed) for changes in the expression level of individual genes across biological conditions. However, genes do not function in singularity. It is well known that genes are not only intricately related to one-another, but are also largely influenced by biological products (e.g., proteins, metabolites, and glycans) in their cellular environments. Investigating the influence of cellular environments on gene co-expression is an important step in the search for gene regulatory mechanisms and the pathways which contribute to the development and progression of complex diseases.

The dependence of the correlation(s) (or other measures of association) of gene expression levels on the values of a covariate is termed *conditional co-expression*. In this dissertation, we use the term conditional co-expression, though the term differential co-expression is also often used to describe the phenomenon of regulated co-expression (Kayano et al., 2014). The covariate, which is investigated as a potential mediator of co-expression, can be categorical (e.g., SNPs) or continuous (e.g., metabolite concentrations). Changes in the co-expression of gene pairs or gene modules are often investigated across discrete biological conditions such as diseased and healthy, young and old, male and female, or between two species such as humans and chimpanzees (Tesson et al., 2010). A gene pair with gene expression values that are, for example, strongly correlated in healthy samples and weakly correlated in diseased samples (or vice-versa) exhibits a pattern of conditional co-expression. Similarly, in the case of gene modules, if significant differences are observed in a gene-module's co-expression (i.e., gene-pair correlations) at high concentrations of a particular metabolite, then the gene module exhibits conditional co-expression.

Conditional co-expression studies can be described as either *targeted* or *untargeted* (Tesson et al., 2010). An untargeted study considers all genes and attempts to identify conditionally co-expressed gene pairs or gene modules. A targeted study investigates whether a predefined gene pair or gene module is conditionally co-expressed. A wide range of methods have been proposed for the detection of conditionally co-expressed gene pairs and gene sets (i.e., untargeted studies), particularly across two biological conditions. Kayano et al. (2014) review the methods for the detection of conditionally co-expressed gene pairs characterized by cross, i.e., a biological phenomenon in which two genes are positively correlated under one condition and negatively correlated under the other condition. Methods to detect gene sets with positive correlations under one condition and random gene-pair correlations under the other condition are

also reviewed. In the review, the need for more efficient techniques is highlighted. Differential co-expression network analysis is one of the more commonly implemented techniques for the detection of conditional co-expression (de la Fuente, 2010; Southworth et al., 2009). Fewer methodologies have been proposed for the investigation of co-expression across multiple groups. Gillis and Pavlidis (2009) analyzed co-expression across multiple-ordered groups (defined by age categories). Chen et al. (2011) proposed a penalized-likelihood approach for bivariate conditional normal models to identify variables that mediate the co-expression of a gene pair (i.e., a targeted study).

We focus on a targeted conditional co-expression analysis, Specifically, we investigate an a priori defined gene module with the aim of identifying variables that mediate its co-expression. Our research is motivated by the conditional co-expression analysis presented in Inouye et al. (2010a). Inouye et al. (2010a) provide a proof-of-concept paper for the integrative analysis of metabolomic, transcriptomic, and genomic data. In particular, they explore the serum-metabolite mediation of the recently characterized core Lipid-Leukocyte (LL) gene-module's (Inouye et al., 2010b) co-expression. Toward this aim, they fit a simple linear regression model to Spearman's correlation coefficients for all pairs of genes of the core LL module for five subsets of samples formed by using quintiles of the metabolite concentrations. In this way, the dependence of the correlation (co-expression) on metabolic concentrations can be detected and quantified.

The method applied by Inouye et al. (2010a), although innovative, is limited in several aspects:

1. It does not allow for the adjustment of the gene-expression values for potential confounding factors. As a consequence, relevant correlations can be missed or spurious correlations can be detected.

2. The simple linear model framework incorrectly treats the correlation coefficients as independent. In addition, the estimation error in the coefficients is ignored.

3. The approach focuses only on linear trends in co-expression by metabolic concentrations.

4. The results may depend on the definition of the metabolic subsets. Categorisation assumes a flat relationship between the predictor (metabolite concentrations) and the response (correlations) within categories. For a linear association, this may not be too problematic. However, consider a non-linear parabola association. A binary split of the trend at the axis of symmetry would result in two categories with almost identical correlations, i.e., it would be difficult to detect

the association. As another example, consider a non-linear 'hockey-stick' association, i.e., the slope is zero except at the largest values of the predictor, where it increases. By using a quintile split, the fifth subset may have a correlation that is rather different to that of the first four subsets but having four subsets positioned in the flat region of the trend may diminish the power to detect the association.

In this part of the dissertation, we consider various modeling approaches to address points 1–4 from the aforementioned list. In particular, in the first approach discussed in Chapter 4, we use a general linear model (GLM) for correlated data (Verbeke and Molenberghs, 2011; Galecki and Burzykowski, 2013) to analyze the dependence structure of gene expression measurements for different metabolic subsets. Statistical tests for the inference of conditional co-expression are proposed. A simulation study is conducted to evaluate the Type I error probability and the sensitivity of the test statistics for different co-expression dynamics. We apply the model to a subset of the DILGOM (**DI**etary, **L**ifestyle, and **G**enetic determinants of **O**besity and **M**etabolic syndrome) study data collected in Helsinki, Finland to study the serum metabolite-induced conditional co-expression of the core Lipid-Leukocyte (LL) module. This dataset is described in Chapter 3.

The methodology proposed in Chapter 4 is directed towards investigating the association between the co-expression of a gene module and a categorical covariate. Some covariates are often difficult to categorize and the selection of arbitrary cut-off points may have an influence on the results of the analysis. To address this, in Chapter 5 we describe a second approach for targeted conditional co-expression investigations involving continuous mediators. We consider modeling the gene-pair correlations (co-expression) of a gene module as a function of a continuous covariate. In particular, transcriptomic and metabolomic data is used to investigate the metabolite-co-expression association of a gene module by specifying a multivariate model which assumes the correlation coefficients as a function of the metabolite concentrations. The model can be seen as a more general version of the bivariate model described in Wilding et al. (2011). The versatility of the model is illustrated using a subset of the DILGOM study data and a simulation study.

In Chapter 7, pseudo-likelihood estimation and conditional copulas are employed for investigating the association between gene-module co-expression and a continuous covariate. In particular, the multivariate density described in Chapter 5 is replaced by the product of all pairwise densities over the set of all possible gene pairs within the gene module. Additionally, the bivariate densities are modelled by using conditional copulas that specify the gene-pair correlations as functions of the continuous covariate. In this way, the computational burden is reduced and the use of conditional copulas facilitates the estimation of non-parametric measures of the association. A simulation

study is conducted to investigate the Type I error probability and power of the pseudo-likelihood ratio test statistic before using the model to investigate the conditional co-expression of the core LL module.

3

Data

In this part of the dissertation, various models are described for investigating the metabolite-mediated co-expression of a gene module. The data described in this chapter are analyzed in Chapters 4, 5, and 7 to illustrate the versatility of the models proposed.

A note on the composition of blood. A sample of blood can be separated into its main constituents: plasma, white blood cells (WBCs), and red blood cells (RBCs). Plasma is the medium of blood in which the WBCs, RBCs, and other blood constituents are suspended. It is mostly comprised of water, and contains proteins, hormones, metabolites, antibodies etc. Blood serum has a similar composition to blood plasma but excludes the clotting factors of blood. WBCs, also called leukocytes, play an important role in the bodies immune system. There are numerous types of WBCs including lymphocytes, neutrophils, monocytes, eosinophils, and basophils. RBCs, also called erythrocytes, are involved in the transportation of oxygen throughout the body.

3.1 Data

We consider a subset of the transcriptomic and metabolomic data from the Finnish population-based cohort, DILGOM (**DI**etary, **L**ifestyle, and **G**enetic determinants of **O**besity and **M**etabolic syndrome). In particular, we investigate the co-expression dynamics of the core Lipid-Leukocyte (LL) gene module (Inouye et al., 2010b) conditional on serum-metabolite concentrations for 466 subjects. Of the 466 individuals analysed, 215 correspond to males and 251 to females, with ages ranging from 25 to 74 years.

3.1.1 Metabolomic data

Proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy was used to determine the serum-metabolite concentrations. Metabolomic data were available on 137 serum metabolites inclusive of amino acids, lipids, and sugars. The full metabolomic dataset

is analysed in Chapter 4. However, for illustration, we primarily focus on six metabolites: 3-hydroxybutyrate, linoleic acid, large HDL particles, small HDL particles, small LDL particles, and total cholesterol in large HDL. Histograms of the observed values of these metabolites are shown in Figure 3.1, with summary statistics listed in Table 3.1.

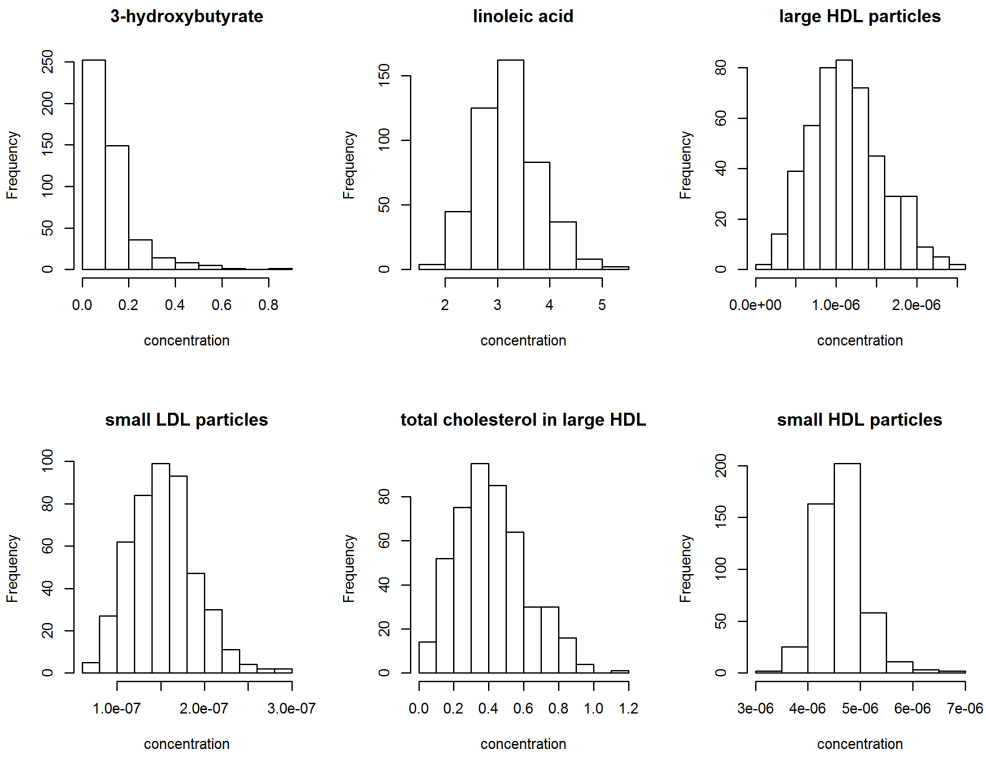

Figure 3.1: Histograms of the observed values for 3-hydroxybutyrate, linoleic acid, large HDL particles, small HDL particles, small LDL particles, and total cholesterol in large HDL.

3.1.2 Transcriptomic data

Gene expression data were obtained from blood lymphocytes using the Illumina HT-12 expression array (Illumina Inc., San Diego, CA, USA). The LL gene module is comprised of 11 highly correlated genes. The module harbors key immune response mediators and has been shown to be strongly associated with serum-lipid concentrations (Inouye et al., 2010a) linking it to the two main contributors of coronary artery disease (CAD), namely, lipid concentrations (such as, high density lipoprotein

Table 3.1: Summary statistics of the observed concentrations for the six metabolites selected for illustration (N = 466).

Metabolite	Mean	Standard deviation	1st Quartile	Median	3rd Quartile
3-hydroxybutyrate	0.1290	0.0970	0.0768	0.0955	0.1363
linoleic acid	3.2141	0.5879	2.8233	3.1735	3.5635
large HDL particles ($\times 10^{-6}$)	1.1334	0.4531	0.8076	1.1080	1.4133
small LDL particles ($\times 10^{-6}$)	0.1524	0.0373	0.1249	0.1504	0.1745
total cholesterol in large HDL	0.4157	0.2008	0.2696	0.3961	0.5418
small HDL particles ($\times 10^{-6}$)	4.6213	0.4505	4.3520	4.6075	4.8668

(HDL) and low density lipoprotein (LDL)) and inflammation (Libby et al., 2002). The seven genes – HDC, FCER1A, GATA2, CPA3, MS4A2, SPRYD5, and SLC45A3 – which form the core LL gene module (Inouye et al., 2010b) are of interest. Summary statistics of the gene-expression values for the seven core LL-module genes appear in Table 3.2. Genes forming the core LL module have heterogeneous variances (see Figure 3.2) and are highly correlated (see Figure 3.3), with Pearson's correlation coefficients larger than 0.6.

Table 3.2: Summary statistics of the gene-expression values for the seven core LL-module genes ($N = 466$).

Gene	Mean	Standard deviation	1st Quantile	Median	3rd Quantile
CPA3	8.2216	0.4215	7.9318	8.2311	8.5038
FCER1A	10.6465	0.6318	10.1987	10.6922	11.0609
GATA2	7.8863	0.3693	7.6198	7.8702	8.1374
HDC	8.9903	0.6608	8.5553	9.0726	9.4352
MS4A2	7.7648	0.2771	7.5838	7.7527	7.8988
SLC45A3	8.1344	0.3556	7.8670	8.1296	8.3708
SPRYD5	8.0550	0.3171	7.8217	8.0433	8.2638

Using this data, we illustrate three methodologies for investigating the co-expression dynamics of the core LL gene module conditional on serum-metabolite concentrations. The association of the co-expression of the core LL gene module with serum-metabolite concentrations was initially investigated by Inouye et al. (2010a), also using a subset of the DILGOM study data.

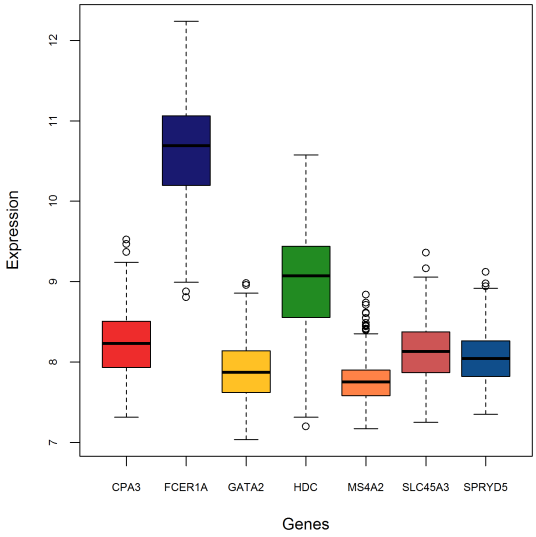

Figure 3.2: Box-plots of the core LL module expression. Heterogeneous mean expression values and variances are observed.

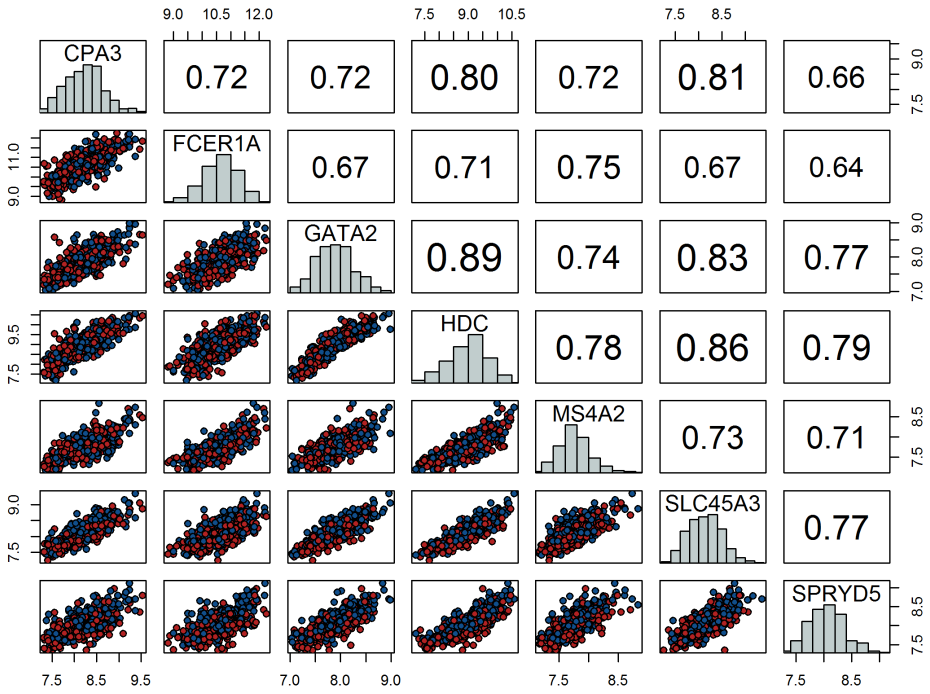

Figure 3.3: Scatter-plot matrix of the core LL module gene-expression values. Scatter-plots of the expression values for each gene pair appear in the lower triangular matrix. Points are colour coded by gender: red represents males and blue represents females. Pairwise correlation coefficients are indicated in the upper triangular matrix. The distribution of gene-expression values for each gene is illustrated on the main diagonal.

4

A multivariate linear model for investigating the association between gene-module co-expression and a categorical covariate

4.1 Introduction

In Chapter 2, the limitations of the simple linear regression approach of Inouye et al. (2010a) for investigating the conditional co-expression of the core LL module was discussed. To recapitulate, the approach of Inouye et al. (2010a) involves fitting a simple linear regression model to the Spearman's correlation coefficients of all pairs of genes forming the gene module for five subsets of samples defined by quintiles of the metabolite concentrations. A more detailed description of this methodology is provided in Section 4.2.2. Limitations of the simple linear regression approach include:

1. It does not allow for the adjustment of gene-expression values for potential confounding factors.
2. The framework incorrectly treats the correlation coefficients as independent.
3. The focus is only on linear trends in co-expression by metabolic concentrations.
4. The results may depend on the definition of the metabolic subsets.

In this chapter, we consider a modelling approach that addresses points [1]-[3] of the aforementioned list. In particular, we propose a multivariate linear model that models the dependence between adjusted gene-expression values through a block-diagonal variance-covariance structure formed by metabolic-subset specific general variance-covariance blocks. Statistical tests for the inference of conditional co-expression are described. The Type I error probabilities and power of the tests are investigated through a simulation study. The models (simple linear regression and multivariate

normal) are then applied to the DILGOM data (described in Chapter 3) to investigate the metabolite-mediated co-expression of the core LL module. The chapter is organized as follows. Section 4.2 includes a description of the statistical methodology and the workflow of the analysis. Results of the simulation study and the application of the models to the DILGOM data appear in Section 4.3. The chapter concludes with a discussion of the results in Section 4.4.

4.2 Statistical methodology

4.2.1 Exploratory analysis

To get a general idea of the co-expression dynamics as a function of metabolic concentrations, we estimate sliding-window correlations. In preparation, for a specific metabolite, the data are sorted in ascending order of the observed metabolic concentrations and a window size (expressed as a proportion, represented by w, of the total sample size) is selected. The procedure begins by computing the correlation coefficient between pairs of genes for the first $w \times N$ individuals, together with the corresponding mean metabolite value. Then, the window is shifted so that it starts from the second ordered metabolite measurement, and the window-specific correlation coefficients and mean metabolite value are estimated. The procedure continues until the window includes the last (ordered) metabolite measurement. The obtained correlation coefficients are plotted against the mean metabolite values. The smoothness of the plot depends on the window size: selecting a large window results in a smoother estimate of the correlation trajectory.

4.2.2 Simple linear regression of Spearman's correlation coefficients

The conditional co-expression analysis by Inouye et al. (2010a) is performed per metabolite. For a given metabolite, the data are split into five subsets based on quintiles of the metabolite's concentration. For each subset, Spearman's rank correlation coefficients are computed for all pairs of genes in the core LL module. A linear regression model is used to relate the estimated correlation coefficients to the quintiles upon which the metabolic subsets are defined.

Using a formal notation, the following model is fitted:

$$Y_{sp} = \alpha + \beta x_s + \varepsilon_{sp}, \tag{4.1}$$

where s ($s = 1, \ldots, S$) indexes the metabolic subsets ($S = 5$ for our case study), p ($p = 1, \ldots, G(G-1)/2$) indexes the gene pairs with G denoting the number of genes

in the gene module ($G = 7$ for the core LL gene module), Y_{sp} is the Spearman's correlation coefficient for the p-th gene pair in the s-th metabolic subset, and x_s is the value of the s-th quintile of the metabolic concentration. As in classical linear regression, ε_{sp} are residual errors that are assumed to be independent and normally distributed with mean zero and variance σ_e^2.

To determine whether there is a relationship between the module co-expression and the metabolite concentrations, the null hypothesis of a zero slope, $H_0 : \beta = 0$, is tested against the alternative hypothesis, $H_A : \beta \neq 0$.

4.2.3 Multivariate linear model for gene-expression measurements

In accordance with the simple linear-regression approach, this analysis is performed per metabolite. For a given metabolite, the data are split into five metabolic-subsets based on quintiles of the metabolite's concentration. Gene-expression values are modeled using a general linear model (GLM) allowing for a correlation between an individual's gene-expression values. A general variance-covariance structure of within-individual gene-expression measurements is assumed for each metabolic subset.

In a formal notation, the following model is considered:

$$\boldsymbol{y}_{si} = \boldsymbol{X}_{si}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{si}, \tag{4.2}$$

where $\boldsymbol{y}_{si} = (y_{si1}, \ldots, y_{siG})^T$ is the vector of gene-expression measurements for the i-th individual ($i = 1, \ldots, n_s$) in the s-th subset, $\boldsymbol{X}_{si}$ is a $G \times R$-dimensional matrix of R covariates (an example of the design matrix $\boldsymbol{X}_{si}$ is included in Appendix A.2), $\boldsymbol{\beta}$ is an R-dimensional vector of coefficients corresponding to the R covariates, and $\boldsymbol{\varepsilon}_{si}$ is a G-dimensional vector of residual errors which are normally distributed with zero mean and variance-covariance matrix $\boldsymbol{\Sigma}_s$. In particular,

$$\boldsymbol{\Sigma}_s = \begin{pmatrix} \sigma_{s,1}^2 & \rho_{s,12}\sigma_{s,1}\sigma_{s,2} & \cdots & \rho_{s,1G}\sigma_{s,1}\sigma_{s,G} \\ \rho_{s,12}\sigma_{s,1}\sigma_{s,2} & \sigma_{s,2}^2 & \cdots & \rho_{s,2G}\sigma_{s,2}\sigma_{s,G} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{s,1G}\sigma_{s,1}\sigma_{s,G} & \rho_{s,2G}\sigma_{s,2}\sigma_{s,G} & \cdots & \sigma_{s,G}^2 \end{pmatrix}, \tag{4.3}$$

where $\sigma_{s,g}^2$ is the variance of the g-th gene for the s-th subset and ρ_{s,g_1g_2} is the correlation between genes g_1 and g_2 for the s-th subset.

The null hypothesis of no metabolite-dependent co-expression can be seen as

corresponding to the following variance-covariance structure:

$$\mathbf{\Sigma}_s^{(0)} = \begin{pmatrix} \sigma_{s,1}^2 & \rho_{12}\sigma_{s,1}\sigma_{s,2} & \cdots & \rho_{1G}\sigma_{s,1}\sigma_{s,G} \\ \rho_{12}\sigma_{s,1}\sigma_{s,2} & \sigma_{s,2}^2 & \cdots & \rho_{2G}\sigma_{s,2}\sigma_{s,G} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1G}\sigma_{s,1}\sigma_{s,G} & \rho_{2G}\sigma_{s,2}\sigma_{s,G} & \cdots & \sigma_{s,G}^2 \end{pmatrix}, \tag{4.4}$$

in which the correlation coefficients $\rho_{g_1 g_2}$ do not depend on the metabolic-subset. In correspondence with $\mathbf{\Sigma}_s$, the gene variances $\sigma_{s,g}^2$ are metabolic-subset specific.

4.2.4 Inference

Likelihood ratio test The null hypothesis of no metabolite-dependent co-expression can be tested by using the likelihood-ratio (LR) test comparing the null model specified by (4.2) and (4.4) with the alternative model defined by (4.2) and (4.3). Wilks (1938) showed that the asymptotic distribution of the LR test is a $\chi_{(k)}^2$ distribution where k is the difference in the number of parameters estimated between the alternative model and the null model. However, there is evidence suggesting that the approximation to a chi-squared distribution may be rather poor for small sample sizes (Pooi, 2003; Gill, 2004).

Larntz and Perlman The statistical test proposed by Larntz and Perlman (1985) is a possible alternative to the LR test for testing the equality of correlation matrices. In the Larntz & Perlman approach, each of the $G(G-1)/2$ hypotheses of equal correlations (i.e., $H_{g_1 g_2} : \rho_{1,g_1 g_2} = \rho_{2,g_1 g_2} = \ldots = \rho_{S,g_1 g_2}$ for all $g_1 \neq g_2$ $(g_1, g_2 = 1, \ldots, G)$) is tested by using the statistic

$$S_{g_1 g_2} = \sum_{i=1}^{S} (n_s - 3) z_{s,g_1 g_2}^2 - \frac{\left[\sum_{i=1}^{S} (n_s - 3) z_{s,g_1 g_2}\right]^2}{\sum_{i=1}^{S} (n_s - 3)}, \tag{4.5}$$

where $z_{s,g_1 g_2}$ is the Fisher's z-transformed correlation between genes g_1 and g_2 for the s-th subset. To test the equality of the correlation matrices, the composite test statistic T, defined as the maximum of the $G(G-1)/2$ test statistics, is computed:

$$T = \max_{1 \leq g_1 < g_2 \leq G} S_{g_1 g_2}. \tag{4.6}$$

Under the null hypothesis, T has an asymptotic χ^2 distribution with $S-1$ degrees of freedom. The Sidák inequality is used to control the probability of committing a Type I error. As such, the null hypothesis of no metabolite-dependent co-expression

is rejected if

$$T > \chi^2_{S-1,\alpha'}, \tag{4.7}$$

where $\alpha' = 1 - (1-\alpha)^{2/G(G-1)}$ is the Sidák-adjusted significance level. The Larntz & Perlman approach has been reported to have good small-sample properties as it relies on the univariate normality of the Fisher's z-transformed correlations (Larntz and Perlman, 1985).

Cole and Jennrich Other possible statistical approaches for testing the equality of correlation matrices include the statistical tests proposed by Cole (1968) and Jennrich (1970), which are based on a quadratic form of deviations from the mean that has an asymptotic χ^2 distribution with $(S-1)G(G-1)/2$ degrees of freedom (Modarres and Jernigan, 1992). We consider the formulation of the Cole and Jennrich test statistics reported in Modarres and Jernigan (1992).

Let $\boldsymbol{\Sigma}_s$ denote the s-th sample correlation matrix, and $\boldsymbol{P}_s$ denote the s-th population correlation matrix. Let $\text{vec}(\boldsymbol{\Sigma}_s)$ denote the vector of correlation coefficients constructed by placing the sub-diagonal elements of $\boldsymbol{\Sigma}_s$ underneath each other in a column-wise order. The asymptotic distribution of $\sqrt{n_s}\text{vec}(\boldsymbol{\Sigma}_s - \boldsymbol{P}_s)$ is multivariate normal with variance-covariance matrix $\boldsymbol{\Gamma}$. The elements of $\boldsymbol{\Gamma}$ are defined as

$$\begin{aligned}\text{Cov}(\rho_{jk}, \rho_{hl}) = \gamma_{jk,hl} = \rho_{jkhl} + \frac{1}{4}\rho_{jk}\rho_{hl}(\rho_{jjhh} + \rho_{kkhh} + \rho_{jjll} + \rho_{kkll}) \\ - \frac{1}{2}\rho_{jk}(\rho_{jjhl} + \rho_{kkhl}) - \frac{1}{2}\rho_{hl}(\rho_{jkhh} + \rho_{jkll}),\end{aligned} \tag{4.8}$$

where j, k, h, and l index genes, i.e., $j, k, h, l = 1, \ldots, G$, and

$$\mu_j = E(\boldsymbol{y}_j), \tag{4.9}$$

$$\sigma_{jk} = E\{(\boldsymbol{y}_j - \mu_j)(\boldsymbol{y}_k - \mu_k)\}, \tag{4.10}$$

$$\sigma_{jkhl} = E\{(\boldsymbol{y}_j - \mu_j)(\boldsymbol{y}_k - \mu_k)(\boldsymbol{y}_h - \mu_h)(\boldsymbol{y}_l - \mu_l)\}, \tag{4.11}$$

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}}, \tag{4.12}$$

$$\rho_{jkhl} = \frac{\sigma_{jkhl}}{\sqrt{\sigma_{jj}\sigma_{kk}\sigma_{hh}\sigma_{ll}}}. \tag{4.13}$$

This is the Steiger-Hakstian (Steiger and Hakstian, 1982) expression for $\boldsymbol{\Gamma}$. In the case of multivariate normal populations, a simplified expression for the elements of $\boldsymbol{\Gamma}$ is given by

$$\begin{aligned}\gamma_{jk,hl} = \frac{1}{2}\rho_{jk}\rho_{hl}(\rho^2_{jh} + \rho^2_{jl} + \rho^2_{kh} + \rho^2_{kl}) + \rho_{jh}\rho_{kl} + \rho_{jl}\rho_{kh} \\ - \rho_{jk}(\rho_{kh}\rho_{kl} + \rho_{jh}\rho_{jl}) - \rho_{hl}(\rho_{kh}\rho_{jh} + \rho_{kl}\rho_{jl}),\end{aligned} \tag{4.14}$$

and is obtained by substituting $\rho_{jkhl} = \rho_{jk}\rho_{hl} + \rho_{jh}\rho_{kl} + \rho_{jl}\rho_{kh}$ in (4.8). This result is referred to as the Pearson-Filon (Pearson and Filon, 1898) expression for $\mathbf{\Gamma}$.

In the two-sample case, to test the equality of the correlation matrices $\mathbf{\Sigma}_{s_1}$ and $\mathbf{\Sigma}_{s_2}$ for the samples s_1 and s_2, respectively, the test statistic of Cole (1968) is given by

$$Q_c = \frac{n_{s_1} n_{s_2}}{n_{s_1} + n_{n_2}} \text{vec}(\mathbf{\Sigma}_{s_1} - \mathbf{\Sigma}_{s_2}) \hat{\mathbf{\Gamma}}_c^{-1} \text{vec}(\mathbf{\Sigma}_{s_1} - \mathbf{\Sigma}_{s_2}), \tag{4.15}$$

and the test statistic of Jennrich (1970) is defined as

$$Q_j = \frac{n_{s_1} n_{s_2}}{n_{s_1} + n_{s_2}} \text{vec}(\mathbf{\Sigma}_{s_1} - \mathbf{\Sigma}_{s_2}) \hat{\mathbf{\Gamma}}_j^{-1} \text{vec}(\mathbf{\Sigma}_{s_1} - \mathbf{\Sigma}_{s_2}). \tag{4.16}$$

The difference between the methods of Cole (1968) and Jennrich (1970) is the way in which $\mathbf{\Gamma}$ is estimated. Cole (1968) proposed to estimate $\mathbf{\Gamma}$ by pooling the estimated covariance matrices. In particular, $\hat{\mathbf{\Gamma}}_c$ in equation (4.15) is defined as

$$\hat{\mathbf{\Gamma}}_c = \frac{(n_{s_1} - 1)\hat{\mathbf{\Gamma}}_{s_1} + (n_{s_2} - 1)\hat{\mathbf{\Gamma}}_{s_2}}{n_{s_1} + n_{s_2} - 2}, \tag{4.17}$$

where $\hat{\mathbf{\Gamma}}_{s_1}$ and $\hat{\mathbf{\Gamma}}_{s_2}$ are obtained by evaluating the Steiger-Hakstian expression (4.8) at $\mathbf{\Sigma}_{s_1}$ and $\mathbf{\Sigma}_{s_2}$, respectively. The approach of Jennrich (1970) involves pooling the estimated correlation matrices. In particular, $\hat{\mathbf{\Gamma}}_j$ is obtained by evaluating the Pearson-Filon expression (4.14) at $\hat{\mathbf{\Sigma}}$ where,

$$\hat{\mathbf{\Sigma}} = \frac{(n_{s_1} - 1)\hat{\mathbf{\Sigma}}_{s_1} + (n_{s_2} - 1)\hat{\mathbf{\Sigma}}_{s_2}}{n_{s_1} + n_{s_2} - 2}. \tag{4.18}$$

In the case of S samples, to test the null hypothesis of no metabolite-dependent co-expression (i.e., $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_S$), the test statistic Q_a, which is the sum of $S(S-1)/2$ pairwise comparisons of the S correlation matrices, can be used. Specifically,

$$Q_a = \sum_{s_1 < s_2} \frac{n_{s_1} n_{s_2}}{n_{s_1} + n_{s_2}} \text{vec}(\mathbf{\Sigma}_{s_1} - \mathbf{\Sigma}_{s_2})' \hat{\mathbf{\Gamma}}^{-1} \text{vec}(\mathbf{\Sigma}_{s_1} - \mathbf{\Sigma}_{s_2}), \tag{4.19}$$

where $s_1, s_2 = 1, \ldots, S$. Modarres and Jernigan (1992) show that, under the null hypothesis, Q_a is distributed as a weighted sum of independent χ^2 random variables each with $r = G(G-1)/2$ degrees of freedom. For $u, v, w, z = 1, \ldots, S$, the weights

of the null distribution are the eigenvalues of the matrix $\boldsymbol{M}$ defined by

$$\boldsymbol{M}_{uv,wz} = \begin{cases} 1 & \text{for } u = w, v = z \\ 0 & \text{for } u \neq v \neq w \neq z \\ \lambda_{vz} & \text{for } u = w, v \neq z \\ \lambda_{uw} & \text{for } u \neq w, v = z \\ -\lambda_{uz} & \text{for } v = w, u \neq z \\ -\lambda_{vw} & \text{for } v \neq w, u = z \end{cases} \tag{4.20}$$

where

$$\lambda_{s_1 s_2} = \sqrt{\frac{n_{s_1} n_{s_2}}{(n_u + n_v)(n_w + n_z)}}. \tag{4.21}$$

As in the two-sample case, $\hat{\boldsymbol{\Gamma}}$ can be estimated as either a pooled estimate of the covariance of the correlation matrices which is the method of Cole (1968), or one can use a pooled estimate of the correlation matrices, i.e., the method proposed by Jennrich (1970).

4.2.5 Multiple comparisons p-value adjustment

The simple linear regression approach (Section 4.2.2) and the GLM approach (Section 4.2.3) both entail fitting a separate model per metabolite. Hence, a multiple testing adjustment should be considered to control either the family-wise error rate (FWER) or the false discovery rate (FDR). FWER-controlling procedures restrict the probability of committing a Type I error (i.e., falsely rejecting the null hypothesis for any of the tests conducted). Controlling the FDR is a less stringent, and hence more powerful approach that instead controls the proportion of discoveries that are allowed to be false. Given the correlated nature of our hypothesis tests (i.e., due to the correlation within the metabolomics data), we use the Benjamini and Yekutieli FDR-controlling procedure (Benjamini and Yekutieli, 2001). It is an extension of Benjamini and Hochberg's correction for cases where the independence of hypothesis tests cannot be assumed (Benjamini and Yekutieli, 2001). Lin et al. (2012) discuss an assortment of FDR-controlling procedures and their implementation using the R statistical programming language.

In some sense, the investigation of conditional co-expression may resemble a dose-response study with, for instance, a placebo and five doses. However, there is no connection from an inferential point of view. In dose-response studies one is often interested in identifying which doses are effective, or the minimal effective dose. The analysis may involve multiple pairwise comparisons and an adaptive strategy to control the FWER. In the conditional co-expression setting, one is interested in the existence of an overall association between gene-module co-expression and metabolite

concentrations. It is not the goal to identify an optimal concentration.

4.2.6 Simulation study

To assess the Type I error probability and the power of the proposed GLM methodology for different co-expression dynamics, we simulate data reflecting six variations in metabolite-co-expression dependence (Figure 4.1). Specifically, we simulate:

- data characterised by no metabolite-co-expression dependence,
- data based on an approximately linear positive association between co-expression and metabolic concentrations and another dataset based on an approximately linear negative metabolite-co-expression association,
- data based on two variations of non-linear dependencies, and
- data exhibiting a weak positive metabolite-co-expression association.

For each of the six co-expression dynamics, we create 1000 datasets of 125, 450, and 800 observations each. Metabolic concentrations are sampled from a normal distribution with mean 3.2141 and variance 0.3456 (i.e., the distribution of linoleic acid in the DILGOM subset). Gene-expression values are sampled from a multivariate normal distribution with means and variances corresponding to that of the CPA3, FCER1A, GATA2, HDC, MS4A2, SLC45A3, and SPRYD5 expression values in the DILGOM data. Gene-pair correlations vary with the metabolite concentration in a manner defined by one of the six metabolite-co-expression associations listed above. These co-expression dynamics are illustrated in Figure 4.1. To investigate the Type I error probability, null-hypothesis data (i.e., characterised by no metabolite-co-expression dependence) are simulated for a four-, five-, and seven-gene module. Data for the power investigation are simulated for a module of four genes.

To define the explicit functional forms of each of the association patterns shown in Figure 4.1, let $\hat{\rho}_{\text{DILGOM},g_1g_2}$ denote the gene-pair correlation coefficient between genes g_1 and g_2 for the DILGOM subset. Additionally, let ρ_{SIM,g_1g_2} denote the simulated correlation coefficient between genes g_1 and g_2, and let $\rho_{\text{SIM},s,g_1g_2}$ represent the simulated subset-specific gene-pair correlation coefficient between genes g_1 and g_2.

The data characterised by no metabolite-co-expression association was simulated assuming that

$$\rho_{\text{SIM},g_1g_2} = \hat{\rho}_{\text{DILGOM},g_1g_2}. \tag{4.22}$$

The simulated data for a four-gene module utilizes the gene-pair correlation coefficients between CPA3, FCER1A, GATA2, and HDC. The gene-pair correlations between these four genes plus MS4A2 are utilized to simulate the data for the five-gene

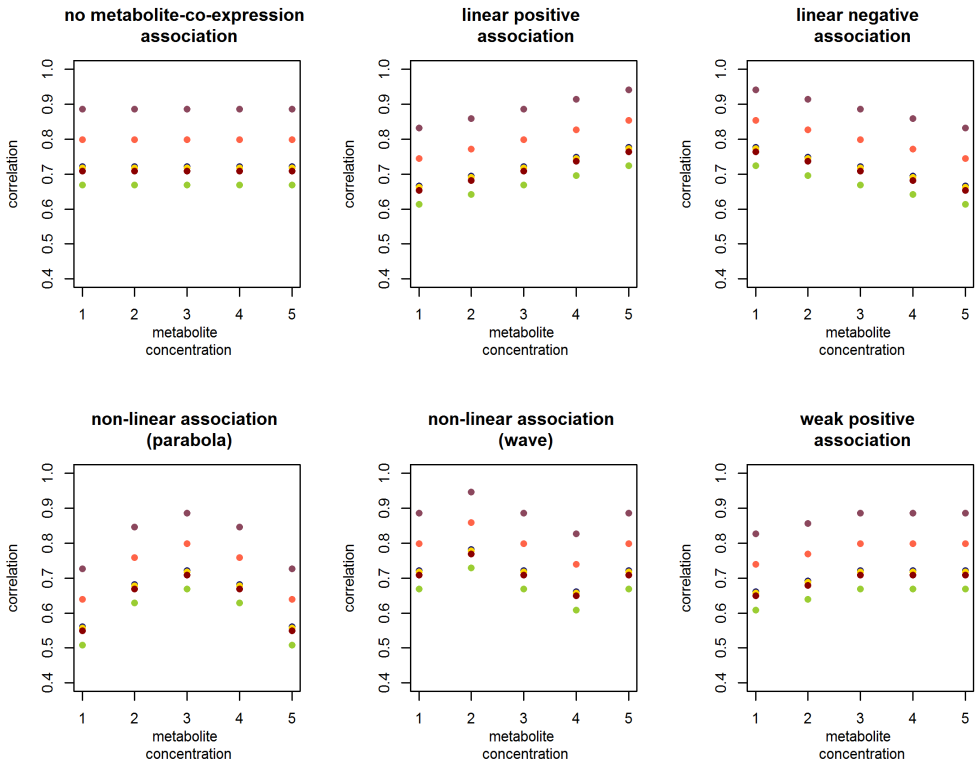

Figure 4.1: Simulated co-expression dynamics for a gene module of four genes. The four genes of the simulated module generate six gene-pair correlations. Each trajectory of dots captures the metabolite-co-expression association for one of the module gene pairs.

module. The simulated data for the seven-gene module is based on the gene-pair correlations between all seven core LL-module genes.

The functional form of the approximate linear associations is

$$\rho_{\text{SIM},s,g_1g_2} = \hat{\rho}_{\text{DILGOM},g_1g_2} + a \times (s-3), \tag{4.23}$$

where $a = -0.0275$ for the approximate linear negative association and $a = 0.0275$ for the approximate linear positive association.

The parabola metabolite-co-expression dynamic was defined as follows:

$$\rho_{\text{SIM},s,g_1g_2} = \hat{\rho}_{\text{DILGOM},g_1g_2} + a \times (s-3)^2, \tag{4.24}$$

where $a = -0.04$.

For the simulated wave co-expression dynamic,

$$\rho_{\text{SIM},s,g_1g_2} = \begin{cases} \hat{\rho}_{\text{DILGOM},g_1g_2} & \text{for } s \in \{1,3,5\}, \\ \hat{\rho}_{\text{DILGOM},g_1g_2} + a & \text{for } s = 2, \\ \hat{\rho}_{\text{DILGOM},g_1g_2} - a & \text{for } s = 4, \end{cases} \tag{4.25}$$

where $a = 0.06$.

The simulated weak positive association had the following form:

$$\rho_{\text{SIM},s,g_1g_2} = \begin{cases} \hat{\rho}_{\text{DILGOM},g_1g_2} & \text{for } s \in \{1,2,3\}, \\ \hat{\rho}_{\text{DILGOM},g_1g_2} + a \times (s-3) & \text{for } s \in \{4,5\}, \end{cases} \tag{4.26}$$

where $a = 0.03$.

The linear regression model and the GLM together with the LR, Larntz & Perlman, Jennrich, and Cole tests were applied to the simulated data (see Section 4.2.3 and Section 4.2.4).

4.2.7 DILGOM analysis

Using the DILGOM data, described in Chapter 3, we study the metabolite-co-expression association by means of the GLM for gene-expression values (Section 4.2.3) and the linear-regression approach of Inouye et al. (2010a) (Section 4.2.2). Due to the non-normality of the metabolite distributions (see Figure 3.1), metabolic concentrations were transformed using the two-parameter Box-Cox transformation (Box and Cox, 1964). The normalized metabolite distributions were then corrected for age, gender, and their two-way interaction using metabolite-specific ANOVA models. The mean structure of the GLM, defined in (4.2), included the four-way interaction between gene, the adjusted metabolite concentration, age, and gender. Figure A.1.2 includes plots of the gene-expression values by age and the gene-expression values by the adjusted concentration of linoleic acid, for a subset of genes. The p-values of the metabolite-specific tests were adjusted by using the Benjamini and Yekutieli (Benjamini and Yekutieli, 2001) FDR-controlling procedure to control the FDR at 0.05.

4.2.8 Implementation

The GLMs were fitted using PROC GLIMMIX of SAS 9.4. The COVTEST statement of PROC GLIMMIX enables the statistical inference on covariance parameters. The LR test is implemented by specifying constraints in the COVTEST statement that, when applied to the variance-covariance structure of the alternate model (4.3), defines the null model's variance-covariance structure (4.4). The generic SAS code is provided

in Appendix A.3. For ease of illustration, the included code is for a module of three genes. Functions to implement the Larntz & Perlman (1985), Jennrich (1970), and Cole (1968) tests were coded in the R programming language. The Benjamini and Yekutieli adjustment was performed using R 3.1.1 and the R-package `multtest`.

4.3 Results

4.3.1 Simulation study

Table 4.1 presents the estimated Type I error probabilities for the GLM-based LR, Larntz & Perlman, Jennrich, and Cole test statistics by module size and sample size. We have found that the Larntz & Perlman test statistic outperforms the Jennrich and Cole statistics with regard to the proper control of the Type I error probability. For the simulated data, Jennrich's approach has better control over the Type I error probability than Cole's approach. However, both Jennrich's and Cole's test statistics are more liberal than the Larntz & Perlman test statistic and struggle to control the Type I error probability, at the nominal level of 0.05, when the sample size is small relative to the number of parameters estimated. Thus, in what follows, we will focus on the linear-regression approach, the GLM-based LR test, and the GLM-based Larntz & Perlman test. The power results of the GLM-based Jennrich and Cole statistics are shown in Table A.4.1 of Appendix A.4.

Table 4.2 integrates the simulation results for the investigation of the Type I error probability. The linear-regression approach fails to control the Type I error probability. When the sample size is small ($n = 125$), the Type I error probability becomes unacceptably high. On the other hand, for large sample sizes (relative to the number of estimated correlation coefficients), the linear regression becomes too conservative. Due to these extreme fluctuations in the Type I error probability, the linear regression approach cannot be deemed a reliable analysis method, as it is difficult to know in a practical setting whether the regression-based test will be liberal or conservative. The GLM-based LR test provides better control of the Type I error probability than the linear-regression approach, particularly for large sample sizes (i.e., when the asymptotic properties of the LR test come into effect). However, the probability is inflated for small sample sizes. The Larntz & Perlman approach properly controls the Type I error probability, with a slight tendency to become conservative for large sample sizes. Hence, combining the Larntz & Perlman test with a suitable multiple-testing procedure should result in a testing framework that properly controls the FWER or the FDR.

Table 4.1: Type I error probabilities for the GLM-based test statistics by module size and sample size.

Module size	Sample size (n)	Type I error probability			
		LR test*	Larntz & Perlman*	Jennrich*	Cole*
4	125	0.109 [0.089, 0.129]	0.045 [0.032, 0.058]	0.091 [0.073, 0.109]	0.198 [0.173, 0.223]
4	450	0.062 [0.047, 0.077]	0.043 [0.030, 0.056]	0.065 [0.049, 0.081]	0.082 [0.064, 0.100]
4	800	0.053 [0.039, 0.067]	0.035 [0.023, 0.047]	0.050 [0.036, 0.064]	0.063 [0.047, 0.079]
5	125	0.141 [0.119, 0.163]	0.048 [0.034, 0.062]	0.091 [0.073, 0.109]	0.288 [0.259, 0.317]
5	450	0.067 [0.051, 0.083]	0.037 [0.025, 0.049]	0.055 [0.040, 0.070]	0.088 [0.070, 0.106]
5	800	0.066 [0.050, 0.082]	0.048 [0.034, 0.062]	0.054 [0.039, 0.069]	0.073 [0.056, 0.090]
7	125	0.314 [0.285, 0.343]	0.035 [0.023, 0.047]	0.104 [0.085, 0.123]	0.572 [0.541, 0.603]
7	450*	0.083 [0.065, 0.100]	0.036 [0.024, 0.048]	0.068 [0.051, 0.084]	0.089 [0.071, 0.107]
7	800**	0.070 [0.054, 0.087]	0.029 [0.018, 0.040]	0.058 [0.043, 0.073]	0.080 [0.062, 0.097]

$\star$ estimate [95% confidence interval]

$*$ convergence rate of GLM: 0.991

$**$ convergence rate of GLM: 0.993

Table 4.2: Type I error probabilities for the linear regression and the GLM-based test statistics by module size and sample size.

Module size	Sample size (n)	Linear regression*	GLM-based LR test*	GLM-based Larntz & Perlman*
4	125	0.205 [0.179, 0.231]	0.109 [0.089, 0.129]	0.045 [0.032, 0.058]
4	450	0.056 [0.041, 0.071]	0.062 [0.047, 0.077]	0.043 [0.030, 0.056]
4	800	0.020 [0.011, 0.029]	0.053 [0.039, 0.067]	0.035 [0.023, 0.047]
5	125	0.197 [0.172, 0.222]	0.141 [0.119, 0.163]	0.048 [0.034, 0.062]
5	450	0.064 [0.048, 0.080]	0.067 [0.051, 0.083]	0.037 [0.025, 0.049]
5	800	0.022 [0.012, 0.032]	0.066 [0.050, 0.082]	0.048 [0.034, 0.062]
7	125	0.461 [0.430, 0.492]	0.314 [0.285, 0.343]	0.035 [0.023, 0.047]
7	450	0.314 [0.285, 0.343]	0.083* [0.065, 0.100]	0.036* [0.024, 0.048]
7	800	0.212 [0.186, 0.238]	0.070** [0.054, 0.087]	0.029** [0.018, 0.040]

$\star$ estimate [95% confidence interval]

$*$ convergence rate of GLM: 0.991

$**$ convergence rate of GLM: 0.993

Table 4.3 shows the results of the power investigation. In view of the problems with the control of the Type I error probability for the linear-regression test and the GLM-based LR test, we focus on the sensitivity of the test statistics to detect the co-expression dynamics in the case of a four-gene module and a sample size of $n = 450$ observations. This is because for this case the Type I error probability, shown in Table 4.2, did not differ significantly from 0.05 for the three approaches. Table 4.3 indicates that the power of the GLM-based LR test and the Larntz & Perlman test is comparable. The GLM-based tests are clearly more powerful than the linear-regression-based test in detecting linear trends and are substantially more powerful in the case of non-linear trends. The only case when the linear-regression-based approach shows some advantage is a weak positive association.

In view of these results, we choose to use the GLM-based Larntz & Perlman test in the DILGOM analysis.

Table 4.3: Power of the linear regression and GLM-based test statistics for different co-expression dynamics and sample sizes.

Co-expression dynamics	Sample size (n)	Linear regression*	GLM-based LR test*	GLM-based Larntz & Perlman*
linear positive association	125	0.408 [0.377, 0.439]	0.314 [0.285, 0.343]	0.188 [0.163, 0.213]
linear positive association	450	**0.635** [0.605, 0.665]	**0.826** [0.802, 0.850]	**0.797** [0.772, 0.822]
linear positive association	800	0.712 [0.683, 0.741]	0.990 [0.983, 0.997]	0.989 [0.982, 0.996]
linear negative association	125	0.451 [0.420, 0.482]	0.300 [0.271, 0.329]	0.184 [0.159, 0.209]
linear negative association	450	**0.621** [0.590, 0.652]	**0.838** [0.815, 0.861]	**0.819** [0.795, 0.843]
linear negative association	800	0.723 [0.695, 0.751]	0.988 [0.981, 0.995]	0.987 [0.979, 0.995]
non-linear association (parabola)	125	0.219 [0.193, 0.245]	0.293 [0.264, 0.322]	0.243 [0.216, 0.270]
non-linear association (parabola)	450	**0.051** [0.037, 0.065]	**0.759** [0.732, 0.786]	**0.856** [0.834, 0.878]
non-linear association (parabola)	800	0.010 [0.003, 0.017]	0.969 [0.958, 0.980]	0.993 [0.987, 0.999]
non-linear association (wave)	125	0.253 [0.226, 0.280]	0.348 [0.318, 0.378]	0.193 [0.168, 0.218]
non-linear association (wave)	450	**0.152** [0.129, 0.175]	**0.863** [0.841, 0.885]	**0.841** [0.818, 0.864]
non-linear association (wave)	800	0.108 [0.088, 0.128]	0.992 [0.986, 0.998]	0.990 [0.983, 0.997]
weak positive association	125	0.278 [0.250, 0.306]	0.143 [0.121, 0.165]	0.072 [0.055, 0.089]
weak positive association	450	**0.257** [0.229, 0.285]	**0.182** [0.158, 0.206]	**0.183** [0.159, 0.207]
weak positive association	800	0.235 [0.208, 0.262]	0.316 [0.287, 0.345]	0.351 [0.321, 0.381]

Data simulated for a four-gene module.
* estimate [95% confidence interval]

4.3.2 DILGOM analysis

Figure 4.2 illustrates the changes in co-expression as a continuous function of the metabolic concentrations for the six metabolites: 3-hydroxybutyrate, linoleic acid, large HDL particles, small HDL particles, small LDL particles, and total cholesterol in large HDL; these are the results of the sliding-window procedure (Section 4.2.1). Evidently, the metabolite-co-expression relationship is not always monotonic as seen, for instance, in the plots for 3-hydroxybutyrate, linoleic acid, or large HDL particles.

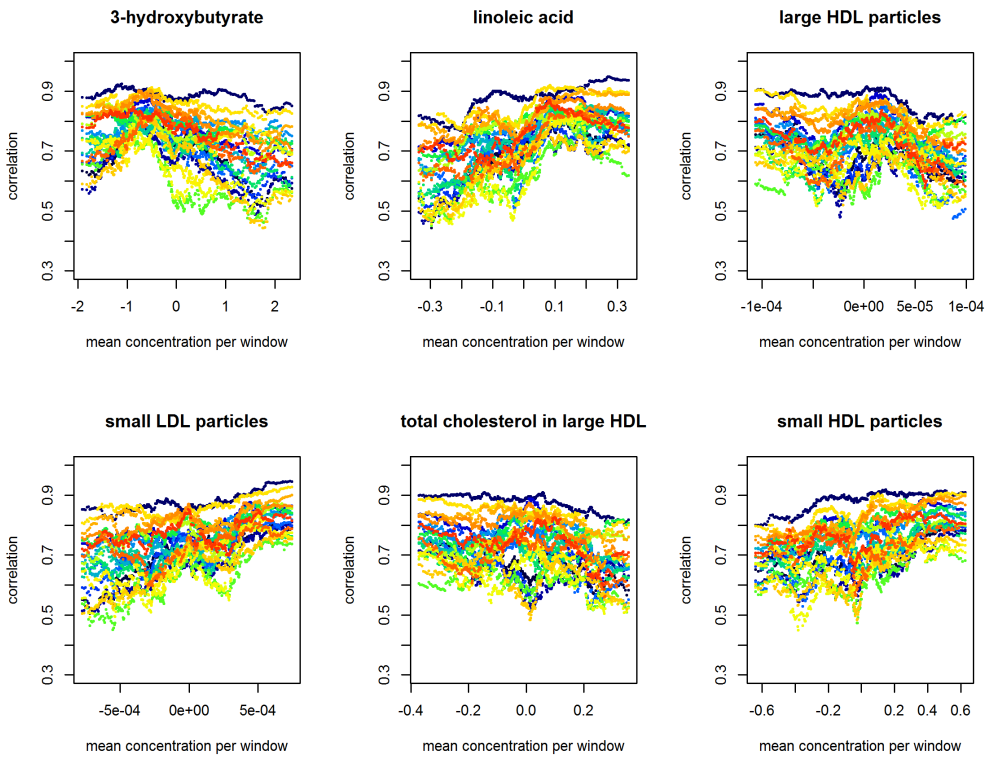

Figure 4.2: Co-expression dynamics by mean metabolic concentration based on sliding-window correlation estimates ($w = 0.2$). The $G = 7$ genes of the core LL module result in 21 gene-pair correlations. Each trajectory roughly captures the co-expression dynamics of one of the module's gene pairs.

Figure 4.3 presents the results obtained by using the simple linear regression model for the six metabolites chosen for illustration. The adjusted p-values for all six metabolites suggest a statistically significant relationship between the correlation coefficients and the metabolite levels. Assuming a FDR of 5%, there are 80 metabolites (including the six presented in Figure 4.3) for which a metabolite-dependent

co-expression could be concluded. However, given the results shown in Table 4.2, it is plausible that the linear-regression-based test is liberal in this case. Thus, in turn, we cannot be sure that the FDR is indeed controlled at the 5% level.

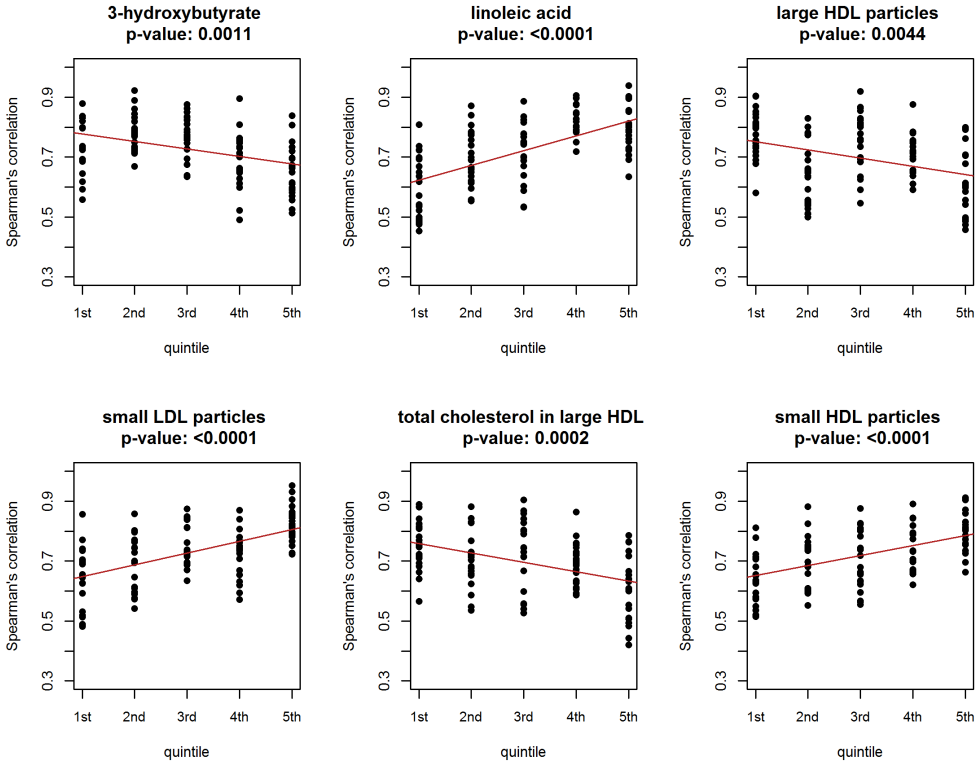

Figure 4.3: Results of the linear-regression-based investigation of conditional co-expression. Dots represent the estimated Spearman's correlation coefficients for the five metabolic subsets (defined by quintiles of the metabolite); the fitted regression line is drawn in red. Benjamini and Yekutieli adjusted p-values are reported.

Figure 4.4 shows the metabolic-subset specific correlation between gene-pairs estimated using the GLM defined by (4.2) and (4.3). Based on the multiplicity-adjusted p-values of the Larntz & Perlman test, a statistically significant relationship between the co-expression and metabolite levels cannot be concluded for any of the metabolites. Given that the Larntz & Perlman test provides a proper control of the Type I error probability, we can expect that, in the analysis, the FDR is controlled at the 5% level.

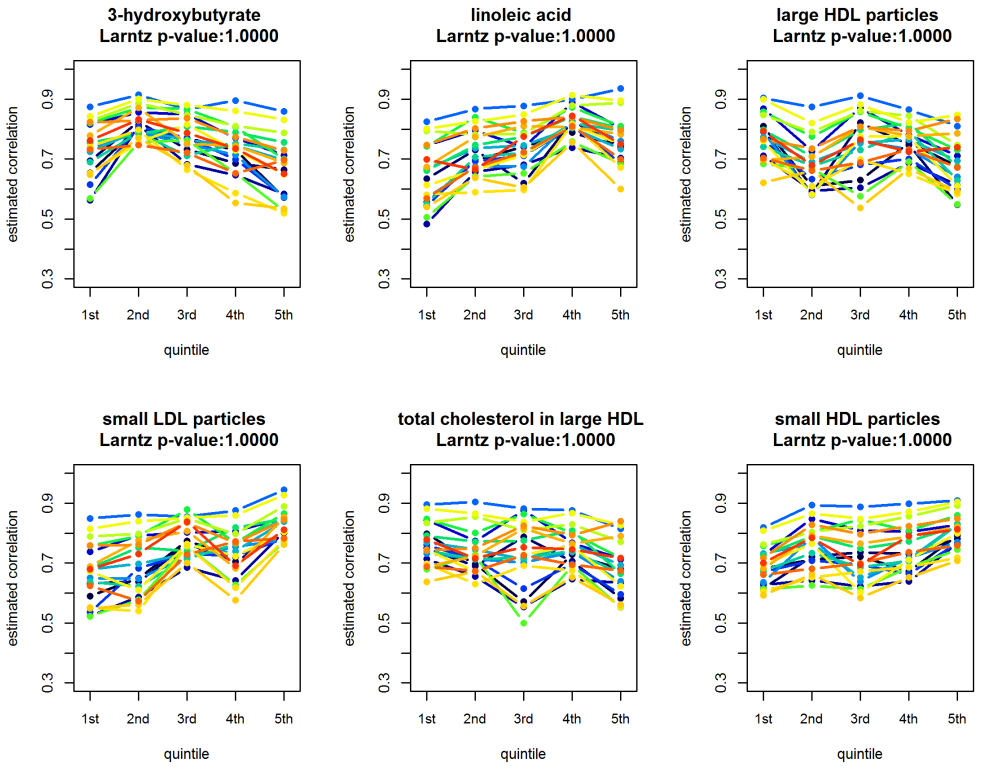

Figure 4.4: GLM based gene-pair correlation estimates for the five metabolic subsets. The estimates corresponding to a particular pair of genes are connected by a line. Benjamini and Yekutieli adjusted Larntz & Perlman test p-values are reported. The unadjusted p-values are $p = 0.0722$ for 3-hydroxybutyrate, $p = 0.0710$ for linoleic acid, $p = 0.0048$ for large HDL particles, $p = 0.0370$ for small LDL particles, $p = 0.2494$ for total cholesterol in large HDL, and $p = 0.6069$ for small HDL particles.

The GLM-framework is flexible in that it allows, for instance, the testing of a variety of hypotheses regarding the variance-covariance structure. To illustrate this aspect of the model, we use the concentration of apolipoprotein B as a potential mediator of the core LL module co-expression. The left-hand-side plot of Figure 4.5 presents the estimated correlation coefficients obtained using the GLM with the variance-covariance structure defined in (4.3) with $S = 5$. We can see that the coefficients seem to only slightly deviate from a common value across the first three subsets (quintiles of the metabolite), while they seem to increase for the last two subsets. Using the Larntz and Perlman statistic, we can formally test whether a common correlation-coefficient could be assumed for the first three subsets. To this aim, we test each hypothesis of $H_{g_1g_2} : \rho_{1,g_1g_2} = \rho_{2,g_1g_2} = \rho_{3,g_1g_2}$, for all $g_1 \neq g_2$ $(g_1, g_2 = 1, \ldots, G)$. The result of

the Larntz & Perlman test is not statistically significant ($p = 0.9950$), suggesting that the simpler variance-covariance structure might be adopted. The plot in the middle column of Figure 4.5 presents the estimated correlation coefficients based on the simplified model. In turn, one could compare the correlation matrices of the simpler model to test for a difference between metabolic subsets, i.e., to determine whether the GLM with the variance-covariance structure defined in (4.4) can be adopted. The right-hand-side plot of Figure 4.5 presents the estimates of the correlation coefficients obtained for the GLM defined by (4.2) and (4.4). The result of the corresponding Larntz & Perlman test is statistically significant ($p = 0.0079$), suggesting that the observed increase of the correlation coefficients across the last two subsets cannot be attributed to a chance variation. The aforementioned results are data-driven and do not take into account the multiple-testing adjustment, but they do illustrate the potential of the GLM in testing various hypotheses that might be of interest.

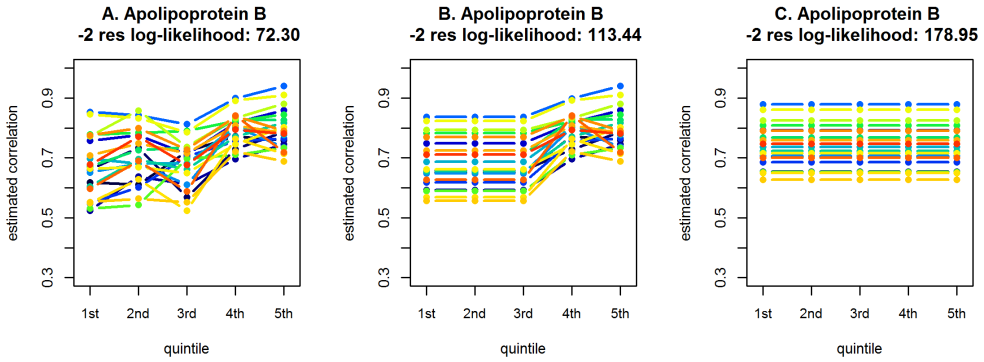

Figure 4.5: Estimated correlation coefficients obtained using the general linear model with different variance-covariance structures, for the five metabolic subsets defined for apolipoprotein B. A. GLM with metabolic-subset specific correlation coefficients defined by (4.2) and (4.3) ; B. GLM with common correlation coefficients across the first three metabolic-subsets; C. GLM with no metabolic-subset dependent correlation coefficients, i.e., the null model defined by (4.2) and (4.4).

4.4 Discussion & Conclusions

The use of the GLM offers a formal, flexible framework to investigate the co-expression-mediation of a gene module. The model facilitates the adjustment of gene-expression values for any potential confounding factors. Questions regarding the conditional co-expression can be formulated as hypotheses about the variance-covariance structure of gene expression measurements and formally tested by using the Larntz & Perlman

test or the LR test (provided that, for the latter, an adequate sample size is available). The model can be fitted using existing software like SAS (PROC MIXED or PROC GLIMMIX) (Verbeke and Molenberghs, 2011; Galecki and Burzykowski, 2013).

As compared to the approach proposed by Inouye et al. (2010a), the GLM-based analysis requires the assumption of normality of the gene-expression measurements. One can see it as a drawback. However, models based on such an assumption (often, on the logarithmic scale) have already been considered in the literature (Wolfinger et al., 2001; Haldermans et al., 2007; Furlotte et al., 2011). Assessing all aspects of multivariate normality is difficult. However, investigating univariate normality, though it will not guarantee multivariate normality, can detect cases of multivariate non-normality. Quantile-quantile plots of the GLM residuals were used to assess the univariate normality (see Appendix A.5). In this way, the plausibility of the assumption can be checked. In return, the GLM-based approach removes the limitations (1–3) of the linear-regression-based analysis mentioned in Section 4.1.

The advantages of using a formal modeling framework were illustrated in the simulation study and in the analysis of the metabolite-mediated conditional co-expression of the core LL gene module. Worth noting is the fact that we did not identify any statistically significant metabolite-co-expression associations. The linear-regression approach results in 80 such associations. This large discrepancy is not surprising in light of the simulation study. For a seven-gene module and a sample size of $n = 450$ observations, the simulation study indicated that the linear-regression approach fails to control the Type I error probability (SLR: 0.314 [0.285, 0.343] vs. GLM-based Larntz & Perlman test: 0.036 [0.024, 0.048]). In a linear-regression model, inconsistent standard error estimates may arise as a consequence of ignoring any estimation error inherent in the dependent variable (Lewis, 2000). The regression approach ignores the estimation error in the observed correlation coefficients. In addition, the coefficients estimated for the same metabolic subset are treated as independent, though they are not. Consequently, the precision of the estimation of the linear regression coefficients may be overestimated, resulting in too small raw p-values and an excess of "false positive" findings even after a multiple-testing correction.

A potential issue in the use of the GLM approach is the number of parameters. Besides the coefficients used in the mean-structure (4.2), the most general variance-covariance structure (4.3) involves SG variances and $SG(G-1)/2$ correlation coefficients, i.e., $SG(G+1)/2$ parameters. Depending on the size of the gene module and the number of metabolic subsets, the number can be very large. For instance, for the core LL gene module with $G = 7$ genes and $S = 5$ subsets, the number of variance-covariance parameters is equal to 140. Thus, estimation of the model requires a considerable sample size. Note, however, that the same remark applies to the linear-regression approach, as it also requires estimation of the $SG(G-1)/2$

correlation coefficients (105 in the case of the core LL gene module).

Another drawback shared by the linear-regression and GLM approaches is that they require the splitting of the metabolite measurements into subsets. Naturally, this implies that the results may depend on the definition of the subsets. A possible solution to this problem would be to model the correlation coefficients as a function of metabolite values. One could imagine using a suitable class of functions, capturing the trends seen in Figure 4.2, to model the correlation coefficients in the variance-covariance matrix (4.3). Such a solution would obviate the need for defining metabolic subsets. This is the topic of Chapter 5.

5

A multivariate linear model for investigating the association between gene-module co-expression and a continuous covariate

5.1 Introduction

In Chapter 4, we described a multivariate linear model for gene-expression values to identify changes in a gene-module's co-expression associated with categorized metabolite levels. The correlations between adjusted gene-expression values are captured through a block-diagonal variance-covariance structure with blocks specific to the categorized levels of the metabolite's concentration. In practice, some covariates may be difficult to categorize in a meaningful way. Moreover, the selection of arbitrary cut-off points may have an influence on the results of the analysis. To address this issue, we consider modeling the gene-pair correlations (co-expression) of a gene module as a function of a continuous covariate. The transcriptomic and metabolomic data is used to investigate the dependence of a gene-module's co-expression on metabolite concentrations by specifying a multivariate model which assumes the gene-pair correlation coefficients as a function of the metabolite concentrations. The model can be seen as a more general version of the bivariate model described in Wilding et al. (2011). A simulation study is conducted to investigate the Type I error probability and power of the likelihood ratio (LR) test for inferring conditional co-expression. The proposed model is applied to the DILGOM data, described in Chapter 3, to investigate the metabolite-co-expression association of the core LL module.

The chapter is organised as follows. Section 5.2 introduces the statistical methodology. Results of the simulation study and analysis of the DILGOM data are presented in Section 5.3. The chapter concludes with a discussion of the results in Section 5.4.

5.2 Statistical methodology

5.2.1 Multivariate linear model with metabolite dependent correlation function

The general linear model (GLM) for correlated data is employed to model the association between gene-pair correlation coefficients and the metabolite concentrations. In mathematical notation, the following model is considered:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \tag{5.1}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iG})^T$ is the vector of gene-expression values for G genes of the i-th individual ($i = 1, \ldots, N$), $\boldsymbol{X}_i$ is the $G \times R$-dimensional matrix of covariates, $\boldsymbol{\beta}$ is an R-dimensional vector of coefficients corresponding to the R covariates, and $\boldsymbol{\varepsilon}_i$ is a G-dimensional vector of residual errors which are normally distributed with zero mean and variance-covariance matrix $\boldsymbol{\Sigma}_i$. In particular,

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma_1^2 & \rho_{i,12}\sigma_1\sigma_2 & \cdots & \rho_{i,1G}\sigma_1\sigma_G \\ \rho_{i,12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{i,2G}\sigma_2\sigma_G \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{i,1G}\sigma_1\sigma_G & \rho_{i,2G}\sigma_2\sigma_G & \cdots & \sigma_G^2 \end{pmatrix}, \tag{5.2}$$

where σ_g^2 is the variance of the g-th gene and ρ_{i,g_1g_2} is the correlation between genes g_1 and g_2 for the i-th individual.

The correlation coefficients ρ_{i,g_1g_2} are linked to the continuous metabolite concentrations via the Fisher-z transformation:

$$\ln\left(\frac{1+\rho_{i,g_1g_2}}{1-\rho_{i,g_1g_2}}\right) = \gamma_{g_1g_2} + \delta_{g_1g_2} f(m_i), \tag{5.3}$$

where $f(\cdot)$ is a known function of the metabolite value m_i, while $\gamma_{g_1g_2}$ and $\delta_{g_1g_2}$ are unknown coefficients.

Note that, instead of the Fisher-z transformation, other transformations based on established cumulative distribution functions (CDF), such as the probit transformation, could be used in (5.3) as well.

From (5.3) it follows that

$$\rho_{i,g_1g_2} = \frac{\exp\{\gamma_{g_1g_2} + \delta_{g_1g_2} f(m_i)\} - 1}{\exp\{\gamma_{g_1g_2} + \delta_{g_1g_2} f(m_i)\} + 1}. \tag{5.4}$$

Many choices for function $f(\cdot)$ are possible. A natural one is to use $f(m_i) = m_i$, which leads to a linear dependence of the Fisher-z transformation of the correlation

coefficient on m_i. Another possible choice is $f(m_i) = \ln(m_i)$. One could use the exploratory plots, as described in Section 4.2.1, as a guide for the selection of the form of $f(\cdot)$.

Figure 5.1 illustrates the relationship (5.4) for various choices of $f(\cdot)$. Note that the graph in the left-hand-side panel was obtained by using $\gamma_{g_1g_2} = 0.5$ and $\delta_{g_1g_2} = 1$, while the graph in the right-hand-side panel was obtained by using $\gamma_{g_1g_2} = -0.5$ and $\delta_{g_1g_2} = 1$. Interestingly, for functions other than $f(m) = \ln(m)$, the lowest possible value of the correlation coefficient is larger than -1. For instance, for $\gamma_{g_1g_2} = 0.5$, $\delta_{g_1g_2} = 1$, and $f(m) = m^2$, the value of the correlation coefficient at $m = 0$ is equal to 0.2449. On the other hand, for $f(m) = m^{-2}$, the value of the correlation coefficient tends to 0.2449 when $m \to \infty$. For $\gamma_{g_1g_2} = -0.5$, the minimum value of the correlation coefficient for $f(m) = m^2$ and $f(m) = m^{-2}$ is approximately -0.2449. For $f(m) = \ln(m)$, the minimum value of the correlation coefficient implied by (5.4) is -1 and the coefficient tends to 1 when $m \to \infty$; the convergence rate is larger for $\gamma_{g_1g_2} = 0.5$ than for $\gamma_{g_1g_2} = -0.5$. Figure 5.1 suggests that, regarding the form of $f(\cdot)$, the most important decision is whether to use the logarithmic function or not.

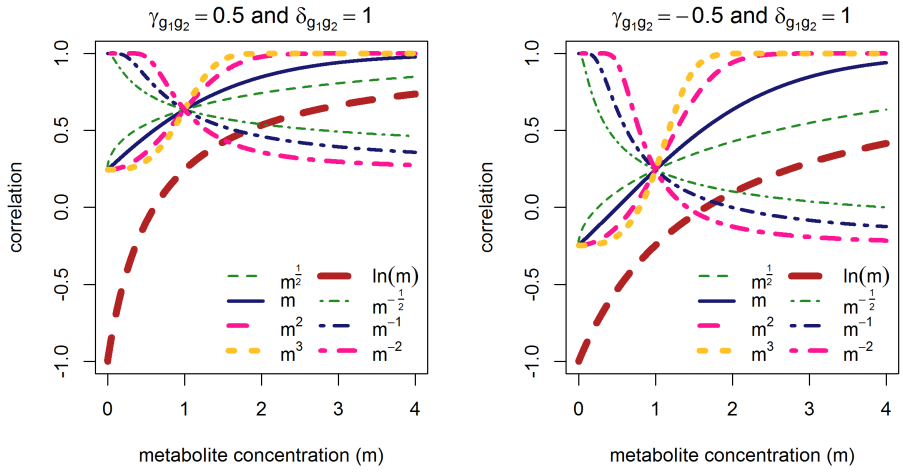

Figure 5.1: Correlation dynamics by metabolite concentration for various transformations $f(\cdot)$ while keeping the intercept $\gamma_{g_1g_2}$ and slope $\delta_{g_1g_2}$ coefficients constant. Left: $\gamma_{g_1g_2} = 0.5$ and $\delta_{g_1g_2} = 1$. Right: $\gamma_{g_1g_2} = -0.5$ and $\delta_{g_1g_2} = 1$.

The parameters of the GLM are estimated by maximizing the restricted likelihood function (Verbeke and Molenberghs, 2011) given by

$$L_{REML} = C \left| \sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \right|^{-\frac{1}{2}} L_{ML}, \tag{5.5}$$

where C is a constant, $|\boldsymbol{A}|$ denotes the determinant of matrix $\boldsymbol{A}$, and L_{ML} is the likelihood function of the model defined as

$$L_{ML} = \prod_{i=1}^{N} \left\{ (2\pi)^{-\frac{G}{2}} |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \right) \right\}. \tag{5.6}$$

The null hypothesis of no metabolite-co-expression dependence implies that $\rho_{1,g_1g_2} = \rho_{2,g_1g_2} = \ldots = \rho_{N,g_1g_2}$ for all $g_1 \neq g_2$. In terms of parameterization (5.3), it corresponds to

$$H_0 : \quad \delta_{12} = \delta_{13} = \ldots = \delta_{G-1,G} = 0. \tag{5.7}$$

Thus, the null model of no metabolite-dependent co-expression corresponds to the following variance-covariance structure:

$$\boldsymbol{\Sigma}_i^{(0)} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1G}\sigma_1\sigma_G \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2G}\sigma_2\sigma_G \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1G}\sigma_1\sigma_G & \rho_{2G}\sigma_2\sigma_G & \cdots & \sigma_G^2 \end{pmatrix}. \tag{5.8}$$

The null hypothesis (5.7) can be tested by using the likelihood-ratio (LR) test comparing the null model, defined by (5.1) and (5.8), with the alternative model, defined by (5.1)–(5.3).

Other hypotheses related to the nature of the metabolite-co-expression dependence can be tested by using the LR test as well. For instance, one can test the hypothesis about uniformity of a metabolite effect across gene pairs,

$$H_0^u : \quad \delta_{12} = \delta_{13} = \ldots = \delta_{G-1,G} = \delta, \tag{5.9}$$

by comparing model (5.1)–(5.3) with the model obtained by using the constraint specified in (5.9).

Wilks (1938) showed that under some regularity conditions, the LR test statistic has an asymptotic chi-squared distribution $\chi^2_{(k)}$, where k is the difference in the number of degrees of freedom between the alternate model and the null model. However, in some situations, the true distribution of the LR statistic can substantially differ from the $\chi^2_{(k)}$ distribution. For instance, if a sample size is not large enough, assuming the asymptotic $\chi^2_{(k)}$ distribution can result in an inflated Type I error probability (as observed in the simulation study described in Chapter 4). As such, while the LR test can be used to perform inference on the correlation parameters of nested models with the same mean structure and different covariance structures, the suitability of the asymptotic chi-square distribution should be investigated.

5.2.2 Simulation study

In order to investigate the Type I error probability (with the nominal level set at 0.05) and the power of the LR test, 1000 datasets of 450 observations (size of the sample is similar to the one available in the case study) were simulated. For each dataset, metabolite concentrations were sampled from a standard normal distribution. Gene-expression values for a seven-gene module were sampled from a multivariate normal distribution with means and variances corresponding to those observed for the core LL-module genes in the DILGOM subset (see Table 3.2). Gene-pair correlations vary with metabolite concentrations according to one of the six metabolite-co-expression association patterns illustrated in Figure 5.2.

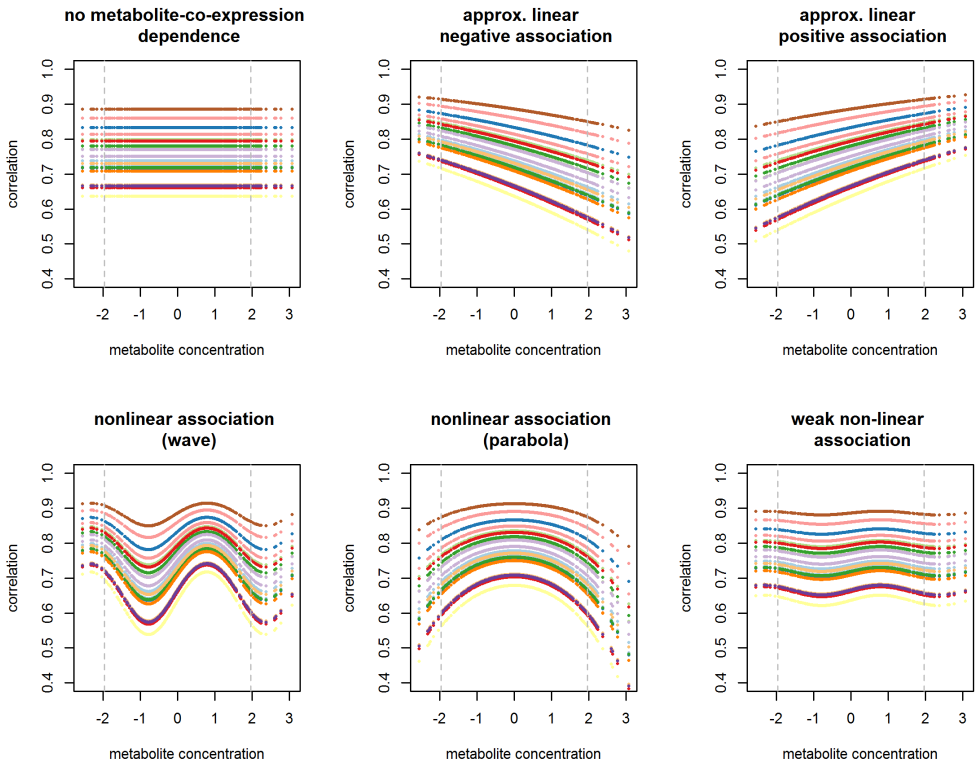

Figure 5.2: Simulated co-expression dynamics for a gene module of seven genes.

The explicit functional form of each of the association patterns shown in Figure 5.2 is given below using the following notation. Let $\hat{\rho}_{\mathrm{DILGOM},g_1g_2}$ denote the gene-pair correlation coefficient between genes g_1 and g_2 for the DILGOM subset (as shown in Figure 3.3), and let $\hat{z}_{\mathrm{DILGOM},g_1g_2}$ denote the corresponding Fisher-z-transformed correlation. Additionally, let $\rho_{\mathrm{SIM},g_1g_2}$ denote the simulated correlation coefficient be-

tween genes g_1 and g_2, and let $\rho_{\text{SIM},i,g_1g_2}$ represent the simulated individual-specific gene-pair correlation coefficient between genes g_1 and g_2. The simulated metabolite concentration for the i-th individual is denoted by $m_{\text{SIM},i}$.

No metabolite-co-expression association was simulated assuming that

$$\rho_{\text{SIM},g_1g_2} = \hat{\rho}_{\text{DILGOM},g_1g_2}. \tag{5.10}$$

For the approximate linear associations,

$$\rho_{\text{SIM},i,g_1g_2} = \frac{\exp\left(\hat{z}_{\text{DILGOM},g_1g_2} + a \times m_{\text{SIM},i}\right) - 1}{\exp\left(\hat{z}_{\text{DILGOM},g_1g_2} + a \times m_{\text{SIM},i}\right) + 1}, \tag{5.11}$$

where $a = -0.15$ for the approximate linear negative association and $a = 0.15$ for the approximate linear positive association.

The functional form for the simulated wave co-expression dynamic and the weak non-linear association was defined as follows:

$$\rho_{\text{SIM},i,g_1g_2} = \frac{\exp\left(\hat{z}_{\text{DILGOM},g_1g_2} + b \times \sin(a \times m_{\text{SIM},i})\right) - 1}{\exp\left(\hat{z}_{\text{DILGOM},g_1g_2} + b \times \sin(a \times m_{\text{SIM},i})\right) + 1} \tag{5.12}$$

where $a = 2$ and $b = 0.3$ for the wave dynamic, and $a = 2$ and $b = 0.05$ for the weak non-linear association.

The parabola metabolite-co-expression dynamic was defined as follows:

$$\rho_{\text{SIM},i,g_1g_2} = \frac{\exp\left(b \times \hat{z}_{\text{DILGOM},g_1g_2} + a \times m^2_{\text{SIM},i}\right) - 1}{\exp\left(b \times \hat{z}_{\text{DILGOM},g_1g_2} + a \times m^2_{\text{SIM},i}\right) + 1}, \tag{5.13}$$

where $a = -0.1$ and $b = 1.1$.

5.2.3 DILGOM analysis

The versatility of the GLM defined by (5.1)–(5.3) is illustrated by studying the co-expression dynamics of the core LL gene module conditional on serum-metabolite concentrations. The following forms of the model were fitted for each of the six considered metabolites (described in Section 3.1.1):

model A with unrestricted intercepts $\gamma_{g_1g_2}$ and slopes $\delta_{g_1g_2}$ (as in (5.3));

model B with $\delta_{12} = \delta_{13} = \ldots = \delta_{G-1,G}$ (as in null hypothesis (5.9));

model C with $\gamma_{12} = \gamma_{13} = \ldots = \gamma_{G-1,G}$;

model D with $\delta_{12} = \delta_{13} = \ldots = \delta_{G-1,G} = 0$ (as in null hypothesis (5.7)).

The mean structure (5.1) included the two-way interaction between gene and each of the following covariates: gender, age, and metabolite concentration. To avoid instability in the estimation procedure due to the exceptionally small observed metabolite concentrations (see Table 3.1), the models were fitted by using standardized metabolite concentrations. In particular,

$$f(m_i) = \frac{m_i - \mu_m}{\sigma_m} \tag{5.14}$$

was used in (5.3), where μ_m and σ_m are the mean and standard deviation of the metabolite concentrations, respectively. Subtracting by the mean in (5.14) ensures that the intercept parameters ($\gamma_{g_1 g_2}$) are estimated at the mean of the metabolite concentrations, μ_m. Inference was based on the LR tests comparing models A and B, A and C, A and D, and B and D.

5.2.4 Implementation

The models were implemented by using the R v.3.2.3 statistical programming language. The restricted log-likelihood defined by (5.5) and (5.6) was optimized by using the Newton-Raphson algorithm through the R package `maxLik` (Henningsen and Toomet, 2011). An analytical gradient was supplied to accelerate convergence (Lindstrom and Bates, 1988; Lin et al., 2013). The starting values of the models were based on the parameter estimates of the null model, defined by (5.1) and (5.8), as obtained by using the `gls` function of the R package `nlme`. For instance, the starting values of the general model, defined by (5.1)–(5.3), were the estimates of the null model rounded off to either one or six decimal places (the starting values of the slopes $\delta_{g_1 g_2}$ were zero).

5.3 Results

5.3.1 Simulation study

The results of the simulation study are displayed in Table 5.1. A convergence rate higher than 0.9 was achieved for each of the co-expression dynamics. Non-convergent samples were excluded from Table 5.1. For a seven-gene module and a sample size of 450 observations, the LR test controls the Type I error probability (0.061, with the 95% confidence interval, CI, $[0.046, 0.076]$) and has a reasonably high power (i.e., greater than 0.850) to detect the approximate linear co-expression dynamics. The power to detect the non-linear associations is lower: it is equal to 0.365, 0.141, and 0.075 for the wave, parabola, and weak association, respectively.

Table 5.1: Simulation study results: Estimated Type I error probability and power of the LR test for a seven-gene module and a sample size of 450 observations.

Co-expression dynamics	Convergence rate	Type I error / Power *
no metabolite-co-expression association	0.949	0.061 [0.046, 0.076]
approx. linear negative association	0.937	0.859 [0.837, 0.881]
approx. linear positive association	0.952	0.863 [0.842, 0.885]
non-linear association (wave)	0.951	0.365 [0.334, 0.395]
non-linear association (parabola)	0.979	0.141 [0.119, 0.163]
weak non-linear association	0.950	0.075 [0.058, 0.091]

∗ point estimate [95% confidence interval]

5.3.2 DILGOM analysis

Figures 5.3 to 5.8 illustrate the estimated metabolite-co-expression dynamics for the models based on total cholesterol in large HDL, linoleic acid, large HDL particles, small LDL particles, 3-hydroxybutyrate, and small HDL particles, in that order. Each plot is comprised of a series of curves representing the estimated co-expression (correlation) dynamic in function of the metabolite concentration for each of the core LL module gene-pairs. The estimated metabolite-co-expression association based on the sliding-window estimation (see Section 4.2.1) is presented in the top-left plot of each figure. The other three graphs in each figure presents results obtained for models A, B, and either C or D (see Section 5.2.3). It is worth noting that model C did not converge successfully for 3-hydroxybutyrate, linoleic acid, total cholesterol in large HDL, and small HDL particles; therefore, these results are not shown in the figures and, instead, results for model D are presented. However, based on the sliding-window correlation estimates, we expect that model C is unlikely to be suitable for the data. Table 5.2 summarizes the LR test results of the analysis.

Table 5.2: DILGOM analysis results: Likelihood-ratio test results.

Metabolite	LR test p-values[a]			
	A vs. B	A vs. C	A vs. D	B vs. D
3-hydroxybutyrate	0.3982	-	0.3188	0.1138
linoleic acid	0.3107	-	0.1765	0.0388
large HDL particles	0.1302	< 0.0001	0.1090	0.1543
small LDL particles	0.3800	< 0.0001	0.2724	0.0763
total cholesterol in large HDL	0.0334	-	0.0244	0.1114
small HDL particles	0.7771	-	0.3860	0.0071

[a] Unadjusted p-values are reported.

Standardised metabolite concentrations, as defined by (5.14), were used to fit the

models for linoleic acid, large HDL particles, small LDL particles, and total cholesterol in large HDL. For total cholesterol in large HDL (Figure 5.3), the LR test comparing model A with model B yields the unadjusted $p = 0.0334$. Thus, conditional on gene-pair-specific intercepts, one cannot assume the same metabolite effect for each gene-pair correlation coefficient. The LR test comparing models A and D results in $p = 0.0244$. This is a test of conditional co-expression and indicates that total cholesterol in large HDL is associated with the core LL module co-expression.

By comparing models A and B for linoleic acid (Figure 5.4), large HDL particles (Figure 5.5), and small LDL particles (Figure 5.6), we find that there is insufficient evidence of a difference in the metabolite effect for each gene-pair correlation. The LR tests comparing model A with model C yield $p < 0.0001$ for large HDL particles, and $p < 0.0001$ for small LDL particles implying that the same overall level of correlation cannot be assumed for each gene-pair. This is consistent with our expectations based on the exploratory analysis. Based on these results, LR tests comparing models B and D are used to test for conditional co-expression. It is concluded that linoleic acid is associated with core LL module co-expression ($p = 0.0388$). There is insufficient evidence of a metabolite-co-expression association for large HDL particles and small LDL particles.

For 3-hydroxybutyrate and small HDL particles, the models using standardized metabolite concentrations as the covariate did not converge. To overcome this issue, the natural logarithmic transformation was applied to the highly skewed distributions of these metabolites (see Figure 3.1). In particular,

$$f(m_i) = \frac{\ln(m_i) - \mu_{\ln(m)}}{\sigma_{\ln(m)}}, \tag{5.15}$$

where $\mu_{\ln(m)}$ and $\sigma_{\ln(m)}$ are the mean and standard deviation of the logarithmically transformed concentrations, was used in (5.3).

For 3-hydroxybutyrate (Figure 5.7), comparison of models B and D indicate no significant metabolite-co-expression association.

For small HDL particles (Figure 5.8), there is a lack of evidence of gene-pair differences in metabolite effect (conditional on the inclusion of gene-pair-specific intercepts). The LR test comparing models B and D yields $p = 0.0071$. We therefore conclude a metabolite-mediated co-expression association for small HDL particles and the core LL module.

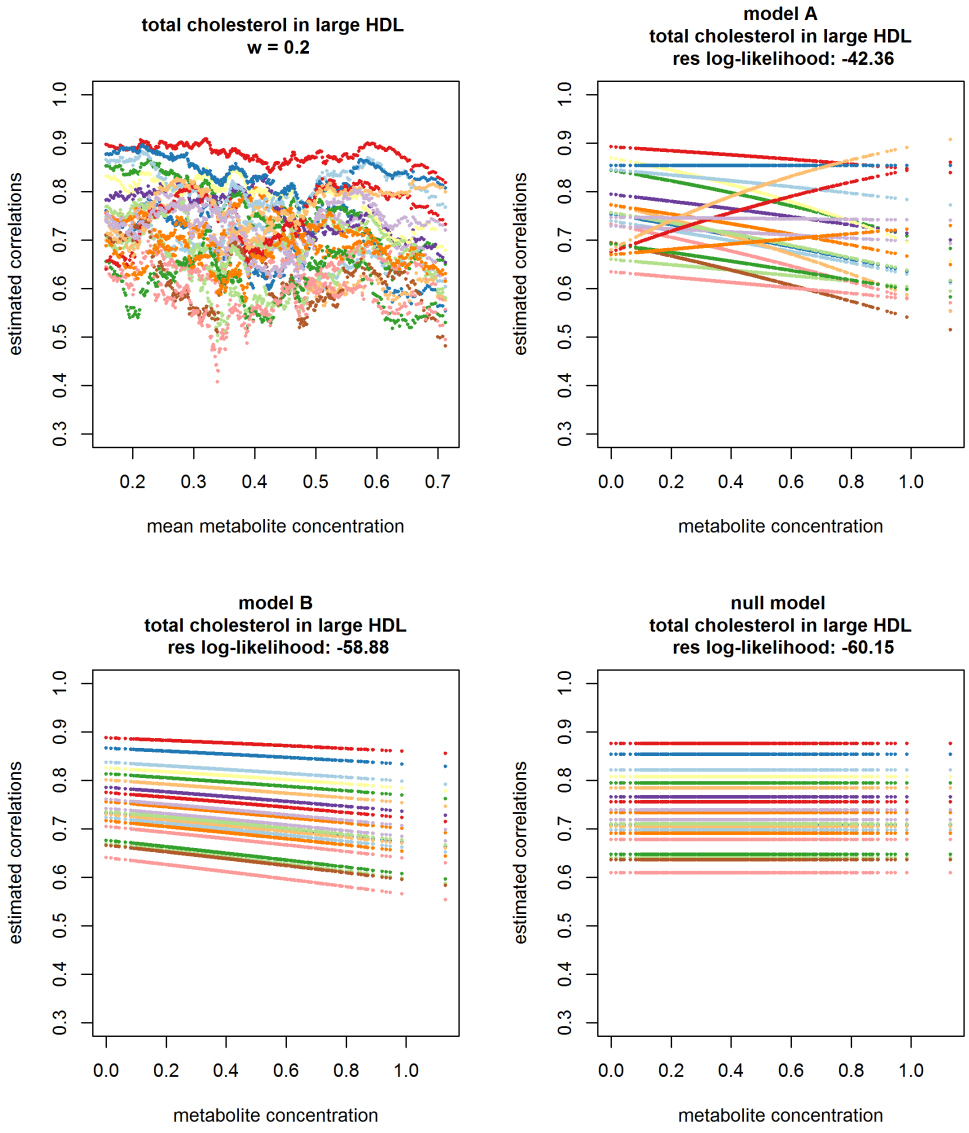

Figure 5.3: Estimated correlations by the concentration of total cholesterol in large HDL. Top-left: sliding-window correlation estimates. Top-right: unrestricted model i.e., the model with gene-pair-specific intercepts and slopes. Bottom-left: model with equal metabolite effect across gene pairs. Bottom-right: the null model i.e., the model with no metabolite effect in the variance-covariance structure.

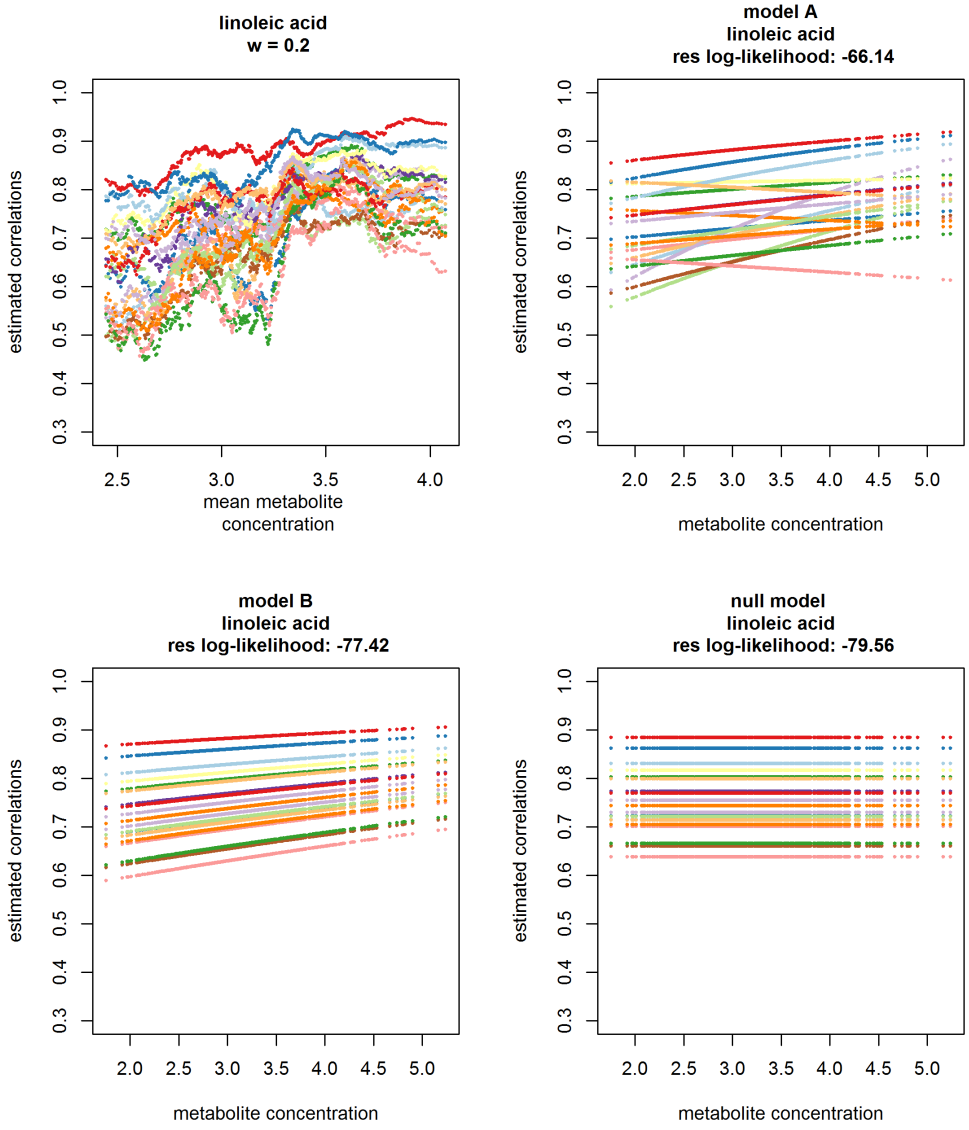

Figure 5.4: Estimated correlations by the concentration of linoleic acid. Top-left: sliding-window correlation estimates. Top-right: unrestricted model i.e., the model with gene-pair-specific intercepts and slopes. Bottom-left: model with equal metabolite effect across gene pairs. Bottom-right: the null model i.e., the model with no metabolite effect in the variance-covariance structure.

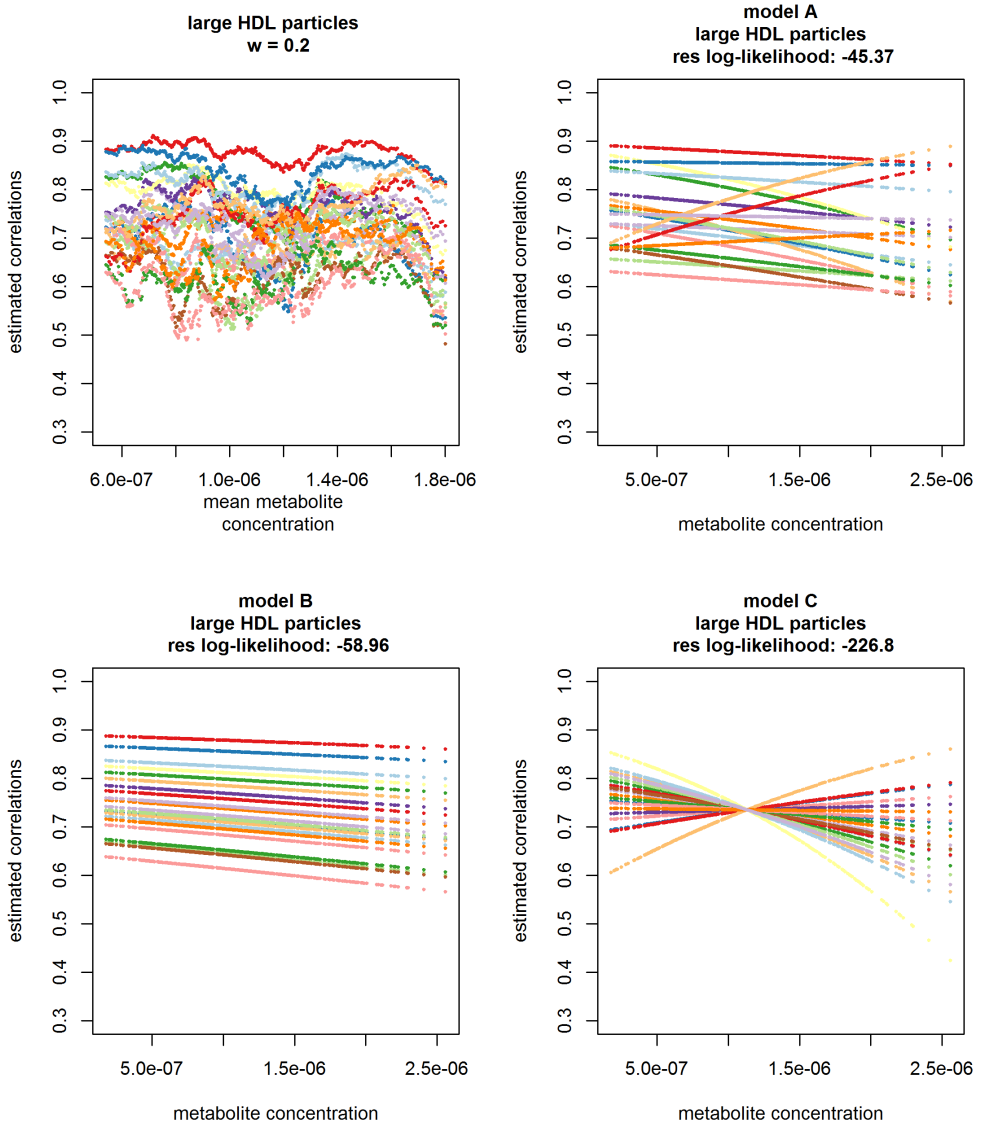

Figure 5.5: Estimated correlations by the concentration of large HDL particles. Top-left: sliding-window correlation estimates. Top-right: unrestricted model i.e., the model with gene-pair-specific intercepts and slopes. Bottom-left: model with equal metabolite effect across gene pairs. Bottom-right: model with the same overall level of correlation for all gene pairs.

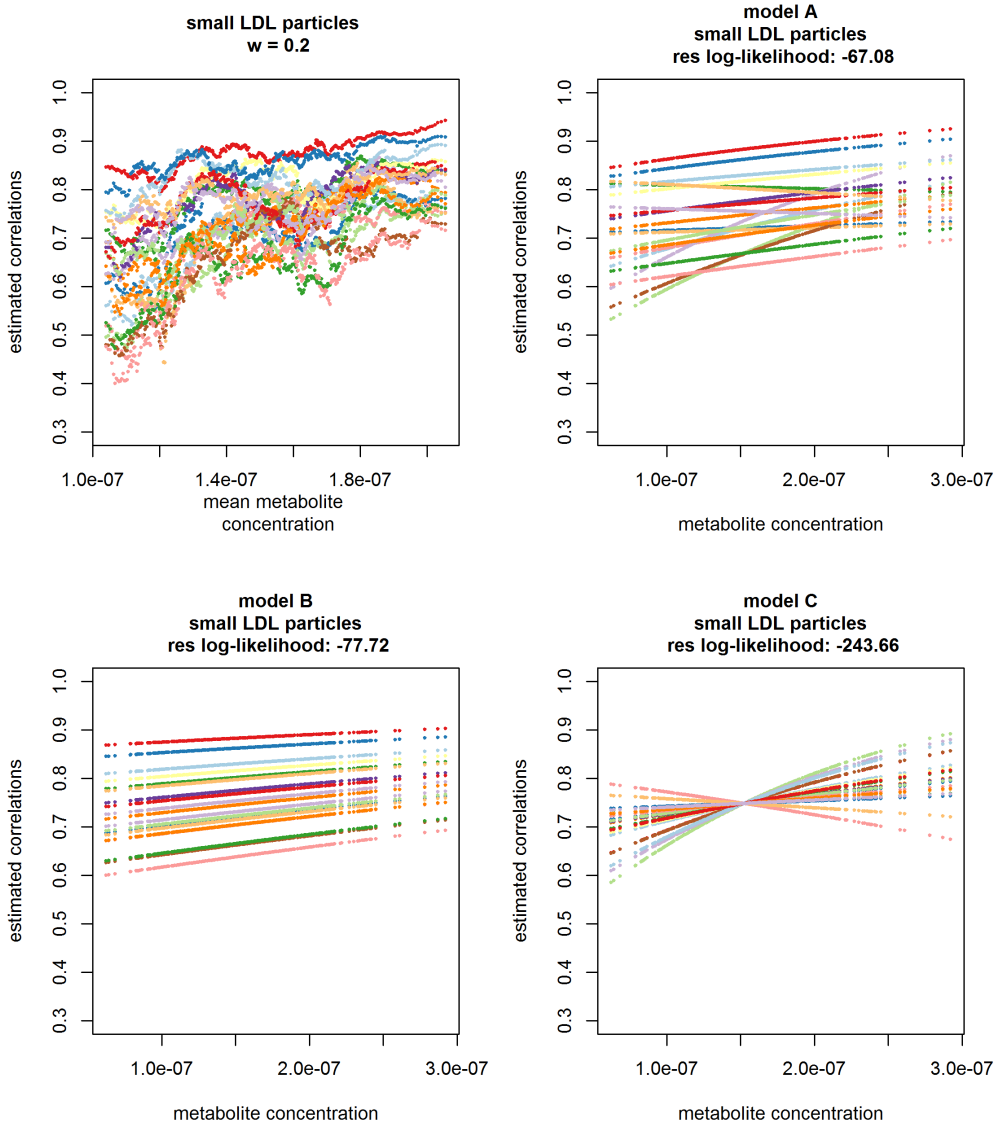

Figure 5.6: Estimated correlations by the concentration of small LDL particles. Top-left: sliding-window correlation estimates. Top-right: unrestricted model i.e., the model with gene-pair-specific intercepts and slopes. Bottom-left: model with equal metabolite effect across gene pairs. Bottom-right: model with the same overall level of correlation for all gene pairs.

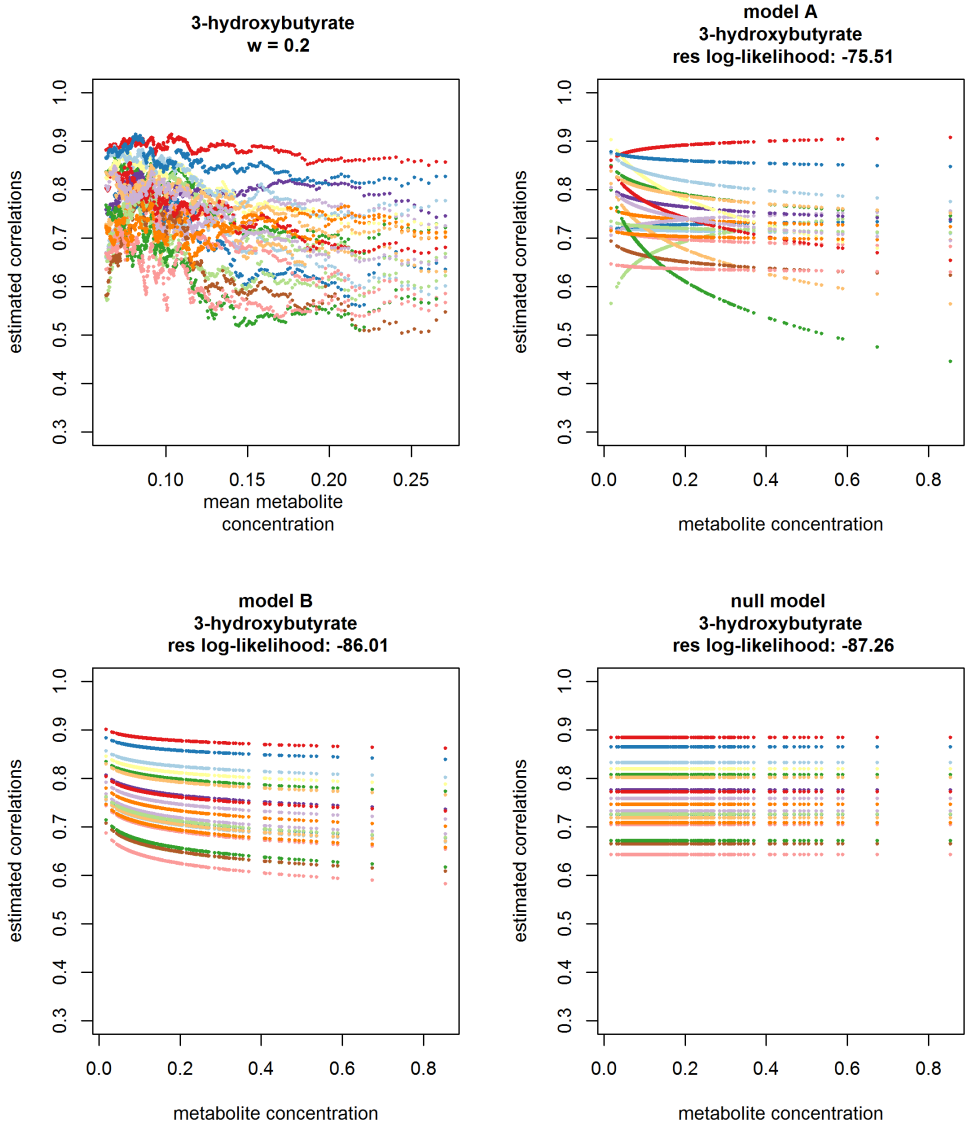

Figure 5.7: Estimated correlations by the concentration of 3-hydroxybutyrate. Top-left: sliding-window correlation estimates. Top-right: unrestricted model i.e., the model with gene-pair-specific intercepts and slopes. Bottom-left: model with equal metabolite effect across gene pairs. Bottom-right: the null model i.e., the model with no metabolite effect in the variance-covariance structure.

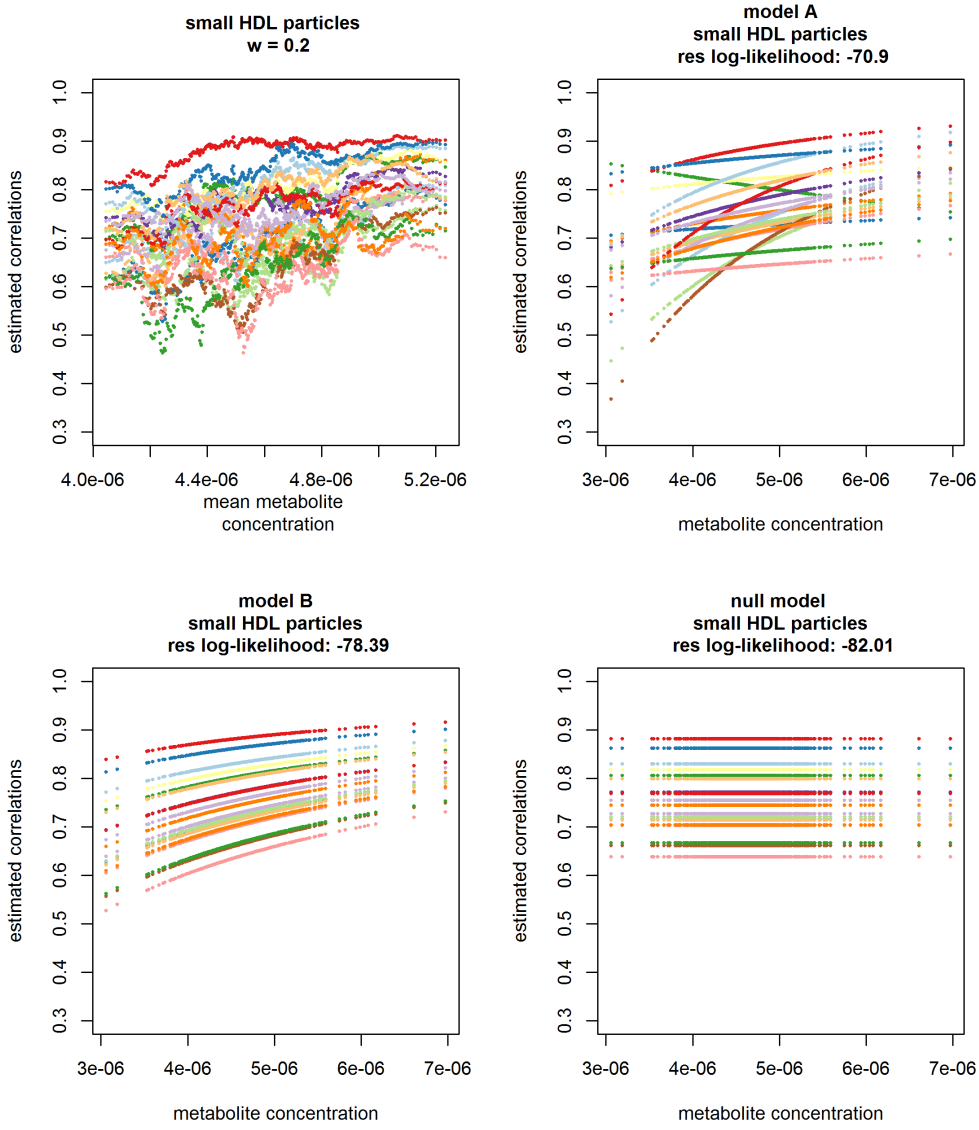

Figure 5.8: Estimated correlations by the concentration of small HDL particles. Top-left: sliding-window correlation estimates. Top-right: unrestricted model i.e., the model with gene-pair-specific intercepts and slopes. Bottom-left: model with equal metabolite effect across gene pairs. Bottom-right: the null model i.e., the model with no metabolite effect in the variance-covariance structure.

A visual comparison of the model-based estimates with the sliding-window estimates shows that for the most part the model is capable of capturing the general trend in the co-expression change in function of metabolite concentrations (see Figure 5.3 to Figure 5.8). However, for some gene pairs it seems that the assumption of a linear relation between the Fisher z-transformed correlation coefficient and the metabolite concentration may be too restrictive. In those cases, a non-linear function of the metabolite concentration may provide a better fit. Figure 5.9 shows the model-A estimated correlation coefficients with point-wise confidence intervals estimated by the delta method for three gene pairs (FCER1A and CPA3, MS4A2 and GATA2, and SLC45A3 and FCER1A) in function of the linoleic acid concentration. The sliding-window correlation estimates are shown in red and the model-estimated correlation coefficients are shown in black. The model-based results in the top row were obtained by using standardized metabolite concentrations. The plots suggest that the model overestimates the co-expression at lower concentrations of linoleic acid. The plots in the bottom row of Figure 5.9 show the estimates obtained for the model with the standardized reciprocal of the squared metabolite concentration:

$$f(m_i) = \frac{m_i^{-2} - \mu_{m^{-2}}}{\sigma_{m^{-2}}}. \tag{5.16}$$

In that case, the estimated trajectories are more comparable with the sliding-window results in that they exhibit a more curvilinear shape than the trajectories shown in the top row of Figure 5.9.

The residuals of each of the models fitted were examined. They did not show any substantial heteroscedasticity. However, we observed a few potential outlying observations in the residuals of the 3-hydroxybutyrate models. The univariate Q-Q plots indicated that there were some, though not substantial, deviations from normality in the tails of the residual distributions for gene MS4A2 (see, for instance, Figure 5.10).

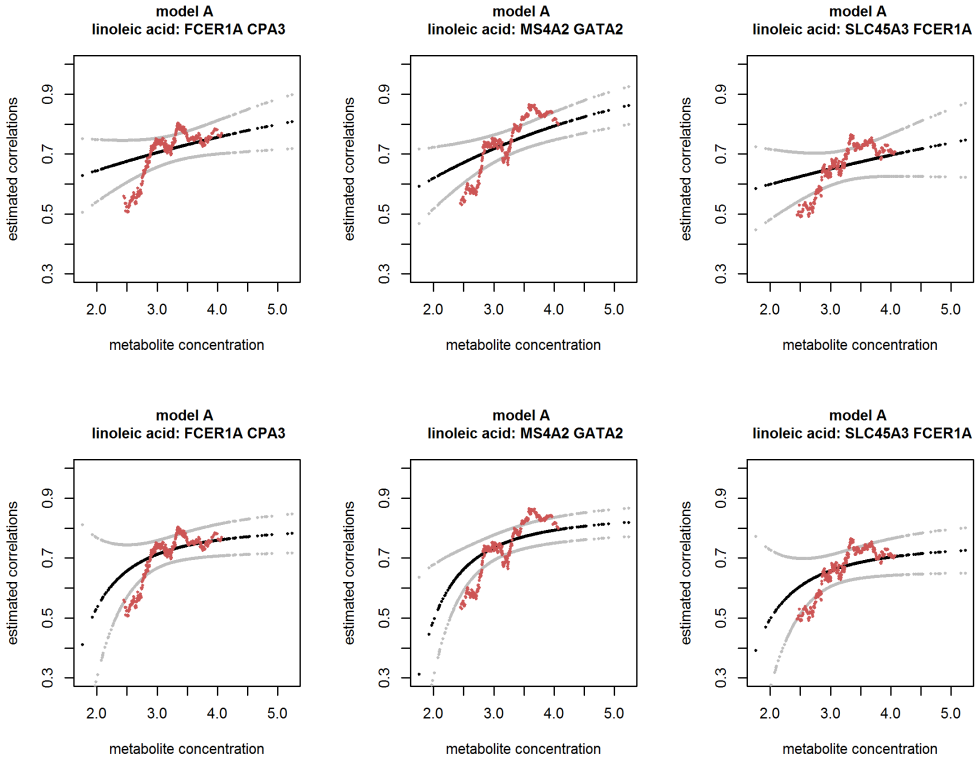

Figure 5.9: Sliding-window correlation estimates together with the model-estimated co-expression dynamics for three gene pairs of the core LL module by linoleic acid concentration. Top row: unrestricted model with standardized linoleic acid concentrations. Bottom row: unrestricted model with standardized reciprocal of squared linoleic acid concentrations. Sliding-window estimated dynamics (window incorporates 20% of the data, i.e., $w = 0.2$) are shown in red and the model-estimated co-expression dynamics are shown in black. The point-wise confidence intervals are shown in grey.

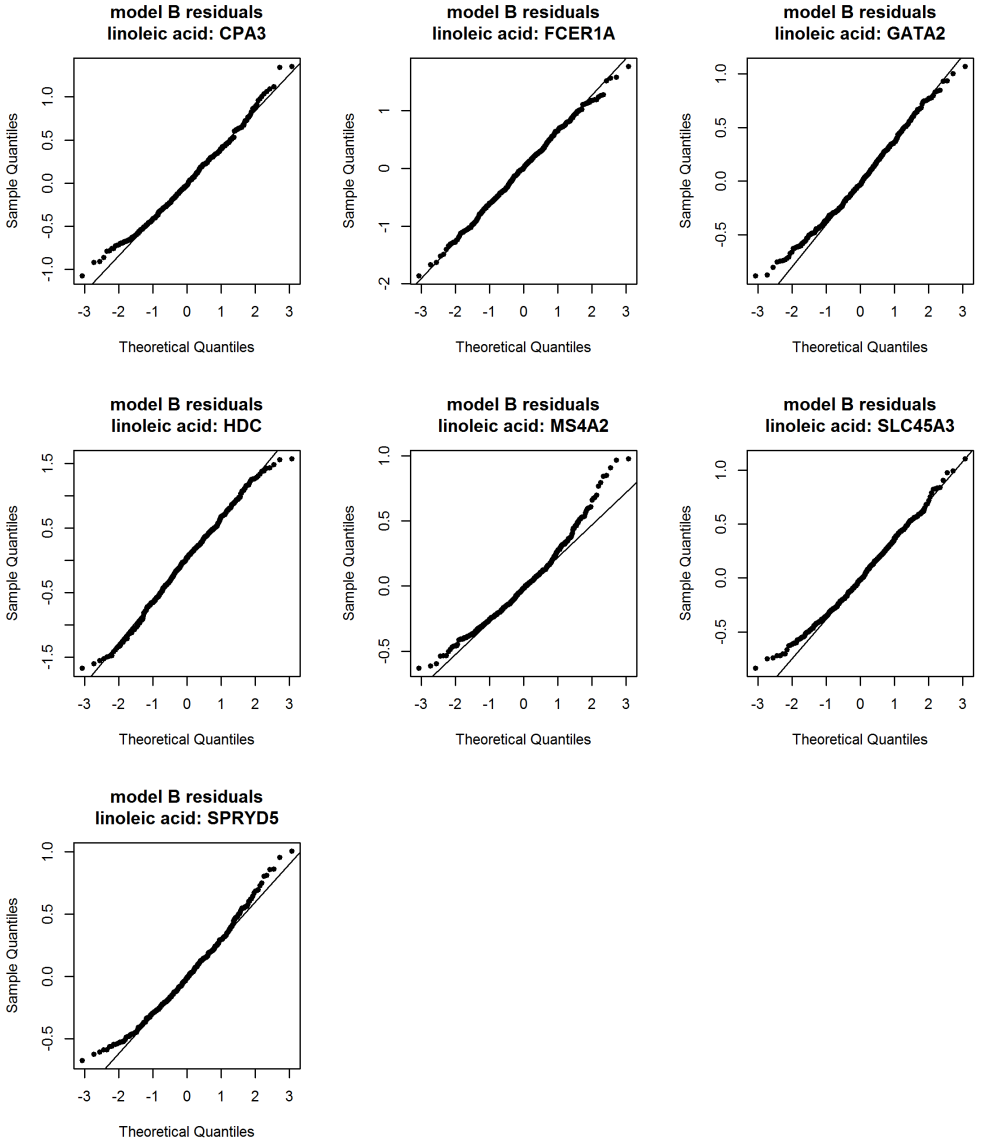

Figure 5.10: Univariate quantile-quantile plots of GLM residuals by gene for the linoleic acid model with the restriction of an equal metabolite effect for all gene pairs.

5.4 Discussion & Conclusions

As illustrated by the simulation study and the analysis of the DILGOM data, the proposed model offers a formal, flexible framework for investigating different (linear and non-linear) patterns of gene-module co-expression. The model avoids the need to categorize the values of a continuous covariate. A variety of hypotheses pertaining to the form of the gene-module's co-expression dependence on the covariate can be tested in a straightforward manner by using the LR test comparing models with appropriately constrained intercept and/or slope parameters.

Despite the similarity in the simulated metabolite-co-expression dynamics of Chapter 4 and Chapter 5, the explicit functional forms of the association patterns, given in Section 4.2.6 and Section 5.2.2, are different. To compare the power of the two approaches for the various simulated metabolite-co-expression dynamics, it is necessary to apply the categorical approach, defined in Section 4.2.3, to the simulated data described in Section 5.2.2. The results of this application are presented in Table B.1.1. The categorical approach of Chapter 4 achieves a slightly higher convergence rate than the continuous approach proposed in this chapter. A power comparison indicates that, for the simulated data, the LR test of the continuous-co-expression approach (power: 0.863 and 0.859 for the positive and negative association, respectively), described in Section 5.2.1, is substantially higher than the power of the categorical-co-expression approach (power: 0.215 and 0.218 for the positive and negative association, respectively), described in Section 4.2.3. For the simulated non-linear associations, the continuous approach (power: 0.365, 0.141, and 0.075 for the wave, parabola, and weak associations, respectively) has a slight power advantage over the categorical approach (power: 0.335, 0.124, and 0.031 for the wave, parabola, and weak associations, respectively). Although, the increase isn't significant for the non-linear wave and parabola associations. The same observation is made when performing the comparison using a merged convergence rate (i.e., only considering the model runs which converged in both approaches) (see Table B.1.2).

The transformations and adjustments applied to the metabolite concentrations and gene-expression values of the DILGOM data differed between Chapter 4 and Chapter 5. Consequently, the DILGOM results reported in Section 4.3.2 cannot be directly compared with the results reported in Section 5.3.2. A comparable analysis of the DILGOM data using the categorical approach results in the unadjusted Larntz & Perlman $p = 0.0111$ for 3-hydroxybutyrate, $p = 0.0147$ for linoleic acid, $p = 0.2496$ for the concentration of large HDL particles, $p = 0.0744$ for the concentration of small LDL particles, $p = 0.5182$ for total cholesterol in large HDL, and $p = 0.1950$ for small HDL particles. Based on these unadjusted p-values, metabolite-co-expression associations are identified for 3-hydroxybutyrate and linoleic acid. In comparison, the LR tests comparing model A with the null-model D identified a metabolite-co-

expression association for total cholesterol in large HDL ($p = 0.0244$, see Table 5.2). The LR test comparison of model B with model D identified co-expression associations with linoleic acid ($p = 0.0388$) and the concentration of small HDL particles ($p = 0.0071$).

Computational difficulties are common in fitting multivariate models (Fieuws and Verbeke, 2006). Convergence issues are amplified as the number of parameters and the dimension of the variance-covariance matrix increases (see Table 4.2). Excluding the parameters included in the mean structure, the general model, defined by (5.1)–(5.3), involves the estimation of G variances and $G(G-1)$ correlation coefficients, i.e., G^2 parameters in total. Thus, for the seven-gene core LL module ($G = 7$), the number of variance-covariance parameters to be estimated is 49. This can be seen as an advantage over the model defined by (4.2) and (4.3), in Chapter 4, which required the estimation of $SG(G+1)/2 = 140$ variance-covariance parameters for the seven-gene core LL module ($G = 7$) and $S = 5$ subsets of metabolite concentrations. It is difficult to visualize the likelihood surface and zoom in on the exact cause of non-convergence or convergence to a solution with a non-zero gradient with the correlation function defined by (5.3). However, some of the strategies we adopted to aid convergence included standardizing and/or transforming (e.g., logarithmically) the continuous covariate and selecting suitable starting values.

An important numerical challenge is ensuring that the estimated variance-covariance matrix is positive definite. This is often achieved by performing unconstrained optimization with a parameterization that enforces the positive-definiteness of the resulting matrices (Pinheiro and Bates, 1996). However, the parameterization does not offer a simple way of expressing the correlation coefficients as known functions of a covariate. Thus, in our implementation of the model, we simply used a general form of an unstructured variance-covariance matrix, as in (5.2), and discarded non-positive-definite solutions.

While the model allows for different functions (linear or non-linear) of the continuous covariate to be used in (5.3), it does not automatically select the most suitable transformation based on the observed co-expression dynamic. In practice, one could consider fitting a variety of models with different covariate functions and select the best fitting model based on, for instance, the lowest value of Akaike's Information Criterion (AIC). In practice, as seen with the DILGOM data, difficulties may arise in achieving model convergence for some of the functions.

The multivariate normality assumption of the GLM may be considered as a potential drawback. Assessing all aspects of multivariate normality is not a straightforward process. In our real-life example, quantile-quantile plots of the GLM residuals were used to assess univariate normality. While univariate normality does not guarantee multivariate normality, it can detect cases of multivariate non-normality.

To get around some of the difficulties associated with fitting the multivariate models, one could also consider implementing the model, defined by (5.1)–(5.3), by using a pairwise modeling approach similar to the one described by Fieuws and Verbeke (2006) for linear mixed-effects models. This is the topic of Chapter 7.

6

Statistical background on pseudo-likelihood and copulas

This chapter provides a brief introduction to pseudo-likelihood estimation and copulas. These concepts are utilized in Chapter 7.

6.1 Pseudo-likelihood

Optimizing a full multivariate likelihood can become increasingly challenging as the number of parameters and the size of the variance-covariance matrix increases. Pseudo-likelihood estimation is a way of approximating a numerically complex, full multivariate likelihood by a simpler, less computationally burdensome function. Typically, a pseudo-likelihood function is formed by taking the product of a series of lower-dimensional marginal and/or conditional densities. Through the use of pseudo-likelihood functions one can obtain valid point and precision estimates without the computational difficulties associated with optimizing a full multivariate likelihood.

The definition by Arnold and Strauss (1991) is used to formally define a pseudo-likelihood function. See also Geys et al. (1999) and Aerts et al. (2002). Let $\boldsymbol{Y}_i$ denote the random variable corresponding to the responses for subject i with $i = 1, \ldots, N$. We assume that $\boldsymbol{Y}_i$ is of constant length n for all $i = 1, \ldots, N$. However, the theory can be extended to include $\boldsymbol{Y}_i$ of variable lengths.

Assume $\boldsymbol{y}_i$ is the response vector for subject i. Consider the set S consisting of $2^n - 1$ vectors of length n. Each vector $\boldsymbol{s} \in S$ is comprised of zeros and ones, and has at least one non-zero entry. Let $\boldsymbol{y}_i^{(\boldsymbol{s})}$ denote the subvector of $\boldsymbol{y}_i$ corresponding to the components of $\boldsymbol{s}$ that are non-zero. The associated joint density is written as $f_{\boldsymbol{s}}(\boldsymbol{y}_i^{(\boldsymbol{s})}|\boldsymbol{\Theta}_i)$ where $\boldsymbol{\Theta}_i$ represents the vector of parameters which characterise the density. To define a pseudo-likelihood function, a set $\delta = \{\delta_{\boldsymbol{s}}|\boldsymbol{s} \in S\}$ of real numbers with at least one non-zero component is chosen.

Definition: The log of the pseudo-likelihood is defined as

$$\ln pl = \sum_{i=1}^{N} \sum_{\boldsymbol{s} \in S} \delta_{\boldsymbol{s}} \ln f_{\boldsymbol{s}} \left(\boldsymbol{y}_i^{(\boldsymbol{s})} | \boldsymbol{\Theta}_i \right) \tag{6.1}$$

The classical full multivariate log-likelihood can be attained by setting $\delta_{\boldsymbol{s}} = 1$ for the vector $\boldsymbol{s} \in S$ that consists solely of ones, and $\delta_{\boldsymbol{s}} = 0$ otherwise.

Regularity conditions The regularity conditions on the density functions $f_{\boldsymbol{s}}(\boldsymbol{y}^{(\boldsymbol{s})}|\boldsymbol{\Theta})$ listed below are necessary to ensure that the log of the pseudo-likelihood function, defined by (6.1), can be maximised by solving the system of equations obtained by setting the derivative of (6.1) equal to zero.

A0 The densities $f_{\boldsymbol{s}}(\boldsymbol{y}^{(\boldsymbol{s})}|\boldsymbol{\Theta})$ are distinct for different values of the parameter vector $\boldsymbol{\Theta}$.

A1 The densities $f_{\boldsymbol{s}}(\boldsymbol{y}^{(\boldsymbol{s})}|\boldsymbol{\Theta})$ have a common support that does not depend on $\boldsymbol{\Theta}$.

A2 The parameter space Ω contains an open region ω of which the true parameter value $\boldsymbol{\Theta}_0$ is an interior point.

A3 The region ω is such that for all $\boldsymbol{s} \in S$, and for almost all $\boldsymbol{y}^{(\boldsymbol{s})}$ in the support of $\boldsymbol{Y}^{\boldsymbol{s}}$, the densities $f_{\boldsymbol{s}}(\boldsymbol{y}^{(\boldsymbol{s})}|\boldsymbol{\Theta})$ admit third derivatives

$$\frac{\partial^3 f_{\boldsymbol{s}}(\boldsymbol{y}^{(\boldsymbol{s})}|\boldsymbol{\Theta})}{\partial \Theta_{p_1} \partial \Theta_{p_2} \partial \Theta_{p_3}}, \tag{6.2}$$

for all $\boldsymbol{\Theta} \in \omega$.

A4 The first and second logarithmic derivatives of $f_{\boldsymbol{s}}$ satisfy the equations

$$E_{\boldsymbol{\Theta}} \left(\frac{\partial \ln f_{\boldsymbol{s}}(\boldsymbol{y}^{(\boldsymbol{s})}|\boldsymbol{\Theta})}{\partial \Theta_p} \right) = 0, \qquad p = 1, \ldots, P, \tag{6.3}$$

and

$$0 < E_{\boldsymbol{\Theta}} \left(\frac{-\partial^2 \ln f_{\boldsymbol{s}}(\boldsymbol{y}^{(\boldsymbol{s})}|\boldsymbol{\Theta})}{\partial \Theta_{p_1} \partial \Theta_{p_2}} \right) < \infty, \qquad p_1, p_2 = 1, \ldots, P. \tag{6.4}$$

A5 The matrix $\boldsymbol{J}(\boldsymbol{\Theta})$ defined by

$$J_{p_1 p_2}(\boldsymbol{\Theta}) = -\sum_{\boldsymbol{s} \in S} \partial_{\boldsymbol{s}} E_{\boldsymbol{\Theta}} \left(\frac{\delta^2 \ln f_{\boldsymbol{s}}(\boldsymbol{y}^{(\boldsymbol{s})}|\boldsymbol{\Theta})}{\partial \Theta_{p_1} \partial \Theta_{p_2}} \right), \tag{6.5}$$

is positive definite.

A6 There exist functions $M_{p_1p_2p_3}$ such that

$$\sum_{\boldsymbol{s}\in S}\delta_{\boldsymbol{s}}E_{\boldsymbol{\Theta}}\left|\frac{\partial^3 f_{\boldsymbol{s}}(\boldsymbol{y}^{(\boldsymbol{s})}|\boldsymbol{\Theta})}{\partial\Theta_{p_1}\partial\Theta_{p_2}\partial\Theta_{p_3}}\right| < M_{p_1p_2p_3}(y) \tag{6.6}$$

for all $\boldsymbol{y}$ in the support of f, for all $\Theta\in\omega$, and $m_{p_1p_2p_3} = E_{\boldsymbol{\Theta}_0}[M_{p_1p_2p_3}(Y)] < \infty$.

The following theorem by Arnold and Strauss (1991) ensures the existence of at least one solution to the pseudo-likelihood (score) equations that is consistent and asymptotically normal.

Theorem 3: Consistency and asymptotic normality.
Let $(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_N)$ be independent and identically distributed with a common density that depends on $\boldsymbol{\Theta}_0$. Then under regularity conditions (A1)-(A6), the maximiser of the logarithm of the pseudo-likelihood denoted by $\hat{\boldsymbol{\Theta}}_N$, has the following properties:

1. The pseudo-likelihood estimator $\hat{\boldsymbol{\Theta}}_N$ converges in probability to the true parameter value, $\boldsymbol{\Theta}_0$.
2. $\sqrt{N}(\hat{\boldsymbol{\Theta}}_N - \boldsymbol{\Theta}_0)$ converges in distribution to

$$N_q\left(\boldsymbol{0}, \boldsymbol{J}(\boldsymbol{\Theta}_0)^{-1}\boldsymbol{K}(\boldsymbol{\Theta}_0)\boldsymbol{J}(\boldsymbol{\Theta}_0)^{-1}\right), \tag{6.7}$$

where $\boldsymbol{J}(\boldsymbol{\Theta})$ is defined by

$$J_{p_1p_2}(\boldsymbol{\Theta}) = -\sum_{\boldsymbol{s}\in S}\delta_{\boldsymbol{s}}E_{\boldsymbol{\Theta}}\left(\frac{\partial^2\ln f_{\boldsymbol{s}}(\boldsymbol{y}^{(\boldsymbol{s})}|\boldsymbol{\Theta})}{\partial\Theta_{p_1}\partial\Theta_{p_2}}\right), \tag{6.8}$$

and $\boldsymbol{K}(\boldsymbol{\Theta})$ is defined by

$$K_{p_1p_2}(\boldsymbol{\Theta}) = \sum_{\boldsymbol{s},\boldsymbol{t}\in S}\delta_{\boldsymbol{s}}\delta_{\boldsymbol{t}}E_{\boldsymbol{\Theta}}\left(\frac{\partial\ln f_{\boldsymbol{s}}(\boldsymbol{y}^{(\boldsymbol{s})}|\boldsymbol{\Theta})}{\partial\Theta_{p_1}}\frac{\partial\ln f_{\boldsymbol{t}}(\boldsymbol{y}^{(\boldsymbol{t})}|\boldsymbol{\Theta})}{\partial\Theta_{p_2}}\right). \tag{6.9}$$

The asymptotic normality result provides a way to consistently estimate the asymptotic covariance matrix. The matrix $\boldsymbol{J}$ is found by evaluating the second derivative of the log of the pseudo-likelihood function at the pseudo-likelihood estimate $\hat{\boldsymbol{\Theta}}_N$ and the expectation in $\boldsymbol{K}$ can be replaced by the cross-products of the observed scores. $\boldsymbol{J}^{-1}$ is referred to as the model-based variance estimator, which should not be used as it overestimates the precision. $\boldsymbol{K}$ is referred to as the empirical correction, and $\boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}$ as the empirically-corrected variance estimator.

As discussed by Arnold and Strauss (1991), the Cramér-Rao inequality implies that $\boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}$ is greater than the inverse of $\boldsymbol{I}$ (i.e., the Fisher information matrix

for the maximum likelihood case), in the sense that $\boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1} - \boldsymbol{I}^{-1}$ is positive semi-definite. Asymptotically, a pseudo-likelihood estimator is always less efficient than the corresponding maximum-likelihood estimator. By using pseudo-likelihood estimation you gain computational simplicity at the expense of some efficiency loss. Aerts et al. (2002) show that in many settings asymptotic efficiency losses are minor. Additionally, Aerts et al. (2002) investigate the small-sample relative efficiency of pseudo-likelihood estimators versus maximum likelihood estimators, through a simulation study, and observe relative efficiencies that fluctuate about one, showing instances in which the the pseudo-likelihood is more favourable than the maximum likelihood estimator.

6.1.1 Pseudo-likelihood inference

Given the close connection between a pseudo-likelihood and a likelihood, Geys et al. (1999) were able to extend the likelihood-ratio test statistic to the pseudo-likelihood framework. Our main interest lies in the pseudo-likelihood-ratio (PLR) test statistic. However, other statistics such as the Wald and score statistics have also been extended to the pseudo-likelihood framework (Geys et al., 1999).

Pseudo-likelihood-ratio test statistic Suppose we are interested in testing the null hypothesis $H_0 : \boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ where $\boldsymbol{\gamma}$ is a sub-vector of the vector of regression parameters denoted by $\boldsymbol{\beta}$. Let $\boldsymbol{\beta} = (\boldsymbol{\gamma}^T, \boldsymbol{\delta}^T)^T$. Furthermore, let $\boldsymbol{J}^{\gamma\gamma}$ denote the $r \times r$ sub-matrix of the inverse of $\boldsymbol{J}$ corresponding to the parameters $\boldsymbol{\gamma}$, and let $\boldsymbol{\Sigma}_{\gamma\gamma}$ denote the $r \times r$ sub-matrix of the empirically-corrected variance-covariance matrix $\boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}$. The PLR test statistic can be computed, in a similar way to the LR test statistic, as twice the difference between the log of the pseudo-likelihood of the full and reduced models:

$$G^{*2} = 2\left[\ln pl(\hat{\boldsymbol{\beta}}_N) - \ln pl(\boldsymbol{\gamma}_0, \hat{\boldsymbol{\delta}}(\boldsymbol{\gamma}_0))\right]. \tag{6.10}$$

Geys et al. (1999) show that the asymptotic distribution of G^{*2} is the weighted sum $\sum_{j=1}^{r} \lambda_j \chi^2_{1(j)}$ where $\chi^2_{1(j)}$ are independently distributed as χ^2_1 variables and the weights $\lambda_1 \geq \ldots \geq \lambda_r$ are the eigenvalues of $(\boldsymbol{J}^{\gamma\gamma})^{-1}\boldsymbol{\Sigma}_{\gamma\gamma}$. Alternatively, the adjusted PLR test statistic, defined by

$$G^{*2}_a = G^{*2}/\bar{\lambda}, \tag{6.11}$$

where $\bar{\lambda}$ is the mean of the eigenvalues λ_j, is approximately χ^2_r distributed. As Geys et al. (1999) point out, it can be argued that G^{*2}_a can be evaluated under both the null and the alternative hypotheses. The adjusted statistics are therefore denoted by $G^{*2}_a(H_0)$ and $G^{*2}_a(H_1)$ when evaluated under the null and the alternative hypotheses, respectively.

6.2 Two-dimensional (bivariate) copulas

Studying the dependence between variables is often of major interest. One way to model the dependence between two random variables is through the use of a two-dimensional (bivariate) copula function.

The definitions and theorems described in this chapter are given as provided by (Nelsen, 2006).

Simply, a two-dimensional (bivariate) copula is a bivariate cumulative distribution function (CDF) with univariate margins that are uniformly distributed on the interval $[0, 1]$. More formally,

Definition 1: A two-dimensional copula is a function $C: [0,1]^2 \mapsto [0,1]$ with the following properties:

1. For every $u, v \in [0,1]$,
$$C(u, 0) = 0 = C(0, v); \tag{6.12}$$

2. For every $u, v \in [0,1]$,
$$C(u, 1) = u \text{ and } C(1, v) = v; \tag{6.13}$$

3. For every u_1, u_2, v_1, v_2 in $[0,1]$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$,
$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0. \tag{6.14}$$

The first property implies that a copula C is grounded. If the observed value of one variable has a marginal probability of zero, then the joint probability of the two responses is zero. Based on the second property, should the observed value of one variable have a marginal probability of one, then the joint probability is equal to the marginal probability of the less certain variable. The third property implies that the copula C is 2-increasing. This property ensures that the joint probability is non-negative since the volume of the rectangle $[u_1, u_2] \times [v_1, v_2]$ is non-negative.

6.2.1 Sklar's theorem

Copulas are of interest as a way of studying scale-free measures of dependence. The theorem of Sklar (1959) forms the theoretical foundation necessary for copula based dependence modelling. Consider the two-dimensional random vector (Y_1, Y_2). Suppose interest lies in the association between the continuous random variables Y_1 and Y_2. Let $F(y_1, y_2)$ denote the joint CDF of Y_1 and Y_2 with marginal CDFs $F_1(y_1)$ and

$F_2(y_2)$. Mathematically,

$$F(y_1, y_2) = Pr(Y_1 \leq y_1, Y_2 \leq y_2), \tag{6.15}$$

$$F_1(y_1) = Pr(Y_1 \leq y_1), \quad \text{and} \quad F_2(y_2) = Pr(Y_2 \leq y_2). \tag{6.16}$$

Theorem 1: Sklar's Theorem. Let F be a joint cumulative distribution function with margins F_1 and F_2. Then there exists a copula C such that for all y_1, y_2 in $[-\infty, \infty]$,

$$F(y_1, y_2) = C(F_1(y_1), F_2(y_2)). \tag{6.17}$$

If F_1 and F_2 are continuous, then C is unique; otherwise, C is uniquely determined on $\text{Ran}F_1 \times \text{Ran}F_2$, i.e., the ranges of F_1 and F_2, respectively. Conversely, if C is a copula and F_1 and F_2 are distribution functions, then the function F defined by (6.17) is a joint distribution function with margins F_1 and F_2.

According to Sklar's theorem, the joint distribution of two continuous random variables can be fully characterised by its univariate marginal distributions and a copula that captures the dependence between the two variables. Several different bivariate distributions can be formed by coupling any two univariate distributions, not necessarily of the same type, with any copula.

6.2.2 Measures of association

The strength of the association modelled by the bivariate copula is commonly expressed in terms of scale-free measures of association such as Kendall's tau and Spearman's rho. The definitions and formulas presented in this section were obtained from Nelsen (2006).

Kendall's tau Kendall's tau, denoted by τ, measures the difference in the probability of concordance and the probability of discordance between two independent realisations, say (Y_{1_1}, Y_{2_1}) and (Y_{1_2}, Y_{2_2}), of the random variables Y_1 and Y_2. In mathematical notation,

$$\tau = Pr[(Y_{1_1} - Y_{1_2})(Y_{2_1} - Y_{2_2}) > 0] - Pr[(Y_{1_1} - Y_{1_2})(Y_{2_1} - Y_{2_2}) < 0]. \tag{6.18}$$

Kendall's tau takes on values in the interval $[-1, 1]$. A value of zero implies independence between Y_1 and Y_2.

Definition 2: Kendall's tau in terms of a copula function is given by

$$\tau = 4 \int_0^1 \int_0^1 C(u,v) dC(u,v) - 1. \tag{6.19}$$

Kendall's tau depends only on the copula function and is independent of the marginal distributions of Y_1 and Y_2.

Spearman's rho Spearman's rho, denoted by ρ_s, is also a measure of concordance. Let (Y_{1_1}, Y_{2_1}), (Y_{1_2}, Y_{2_2}), and (Y_{1_3}, Y_{2_3}) be three realisations of the random variables Y_1 and Y_2. Spearman's rho is proportional to the difference in the probability of concordance and the probability of discordance between (Y_{1_1}, Y_{2_1}) and (Y_{1_2}, Y_{2_3}). Mathematically,

$$\rho_s = 3Pr[(Y_{1_1} - Y_{1_2})(Y_{2_1} - Y_{2_3}) > 0] - Pr[(Y_{1_1} - Y_{1_2})(Y_{2_1} - Y_{2_3}) < 0]. \tag{6.20}$$

Similar to Kendall's tau, Spearman's rho takes on values in the interval $[-1, 1]$. A value of zero implies independence between the two random variables Y_1 and Y_2.

Definition 3: Spearman's rho in terms of a copula function is given by

$$\rho_s = 12 \int_0^1 \int_0^1 C(u,v) du dv - 3. \tag{6.21}$$

Spearman's rho is independent of the marginal distributions of Y_1 and Y_2.

Tail dependence Measures of association, such as Kendall's tau and Spearman's rho, are global measures of the association between two variables. Another concept is tail dependence which measures the dependence between the variables in the tails of the data, i.e., at extremely large values (in the joint upper tail of the bivariate distribution) or at extremely small values (in the joint lower tail of the bivariate distribution). The choice of copula has an influence on which parts of the distributions of the variables are more strongly associated. Some examples are given in Section 6.2.4.

6.2.3 Estimation

Our interest lies in fully parametric copula models i.e., models in which both the copula and its marginal distributions have a parametric form. In this situation, estimation can be performed by maximising the log-likelihood function. Let θ denote the copula parameter, and ϕ_1 and ϕ_2 denote the parameters that characterise the marginal distributions of Y_1 and Y_2. The bivariate density of $f(y_1, y_2|\theta, \phi_1, \phi_2)$ is obtained

as follows:

$$\begin{aligned} f(y_1, y_2|\theta, \phi_1, \phi_2) &= \frac{\partial F(y_1, y_2|\theta, \phi_1, \phi_2)}{\partial y_1 \partial y_2} \\ &= \frac{\partial C\left(F_1(y_1|\phi_1), F_2(y_2|\phi_2)\right)}{\partial F_1(y_1|\phi_1)\partial F_2(y_2|\phi_2)} \times \frac{\partial F_1(y_1|\phi_1)}{\partial y_1} \times \frac{\partial F_2(y_2|\phi_2)}{\partial y_2} \\ &= c\left(F_1(y_1|\phi_1), F_2(y_2|\phi_2)|\theta\right) \times f_1(y_1|\phi_1) \times f_2(y_2|\phi_2). \end{aligned} \tag{6.22}$$

The log-likelihood function is

$$L(\theta, \phi_1, \phi_2) = \sum_{i=1}^{N} \ln c\left(F_1(y_{1i}|\phi_1), F_2(y_{2i}|\phi_2)|\theta\right) + \sum_{i=1}^{N} \ln f_1(y_{1i}|\phi_1) + \sum_{i=1}^{N} \ln f_2(y_{2i}|\phi_2). \tag{6.23}$$

These results can be found in Joe (2014).

6.2.4 Examples of copulas

Example 1: Minimum and maximum copulas The minimum and maximum copulas denoted by W and M, respectively, are defined as

$$W(u, v) = max(u + v - 1, 0) \quad \text{and} \quad M(u, v) = min(u, v). \tag{6.24}$$

For every copula $C(u, v)$, with $(u, v) \in [0, 1] \times [0, 1]$, the following inequality applies:

$$W(u, v) \leq C(u, v) \leq M(u, v). \tag{6.25}$$

W and M are also referred to as the Fréchet-Hoeffding lower and upper bounds, respectively.

Example 2: Product copula The product copula is denoted by Π. It has independent margins and has the form

$$\Pi(u, v) = uv. \tag{6.26}$$

Example 3: Gaussian copula The Gaussian copula is part of the Elliptical class of copulas. Here, we denote the marginal distributions by u_1 and u_2, instead of u and v, to be consistent with the notation used in Chapter 7. The CDF of the Gaussian copula is given by

$$C(u_1, u_2|\rho) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2)|\rho), \tag{6.27}$$

where Φ^{-1} is the quantile function, i.e., the inverse of the CDF of a $N(0, 1)$ (standard normal) random variable, and $\Phi_2(\cdot, \cdot|\rho)$ is the CDF of two standard-normally distributed random variables with correlation given by the copula parameter ρ.

The density of the Gaussian copula is defined as

$$c(u_1, u_2|\rho) = \frac{1}{\sqrt{1-\rho^2}} \exp\left\{-\frac{\rho^2(x_1^2+x_2^2) - 2\rho x_1 x_2}{2(1-p^2)}\right\}, \tag{6.28}$$

where $x_1 = \Phi^{-1}(u_1)$ and $x_2 = \Phi^{-1}(u_2)$.

The copula parameter ρ, which corresponds to the Pearson correlation coefficient between u_1 and u_2, is restricted to the interval $[-1, 1]$. Closed-form expressions exist for the relationship between Kendall's tau and ρ, as well as for the relationship between Spearman's rho and ρ. In particular,

$$\tau = \frac{2}{\pi}\arcsin(\rho), \tag{6.29}$$

and

$$\rho_s = \frac{6}{\pi}\arcsin\left(\frac{\rho}{2}\right). \tag{6.30}$$

The Gaussian copula is a symmetric copula. It has no tail dependence, i.e., the same correlation is modeled for entire span of the distributions of u_1 and u_2.

The t-copula is an example of another copula that belongs to the Elliptical class of copulas.

Example 4: Gumbel-Hougaard copula As with the Gaussian copula, we use u_1 and u_2 to denote the marginal distributions. The Gumbel-Hougaard copula is from the Archimedian class of parametric copulas and is defined as

$$C(u_1, u_2|\theta) = \exp\left[-\left\{(-\ln u_1)^\theta + (-\ln u_2)^\theta\right\}^{\frac{1}{\theta}}\right], \tag{6.31}$$

where θ is the copula parameter that is restricted to the interval $[1, \infty)$.

The density of the Gumbel-Hougaard copula is defined as

$$\begin{aligned} c(u_1, u_2|\theta) = &\exp\left[-\left\{(-\ln u_1)^\theta + (-\ln u_2)^\theta\right\}^{\frac{1}{\theta}}\right] \frac{1}{u_1 u_2} \\ &\times \left\{(-\ln u_1)^\theta + (-\ln u_2)^\theta\right\}^{-2+\frac{2}{\theta}} (\ln u_1 \ln u_2)^{\theta-1} \\ &\times \left[1 + (\theta-1)\left\{(-\ln u_1)^\theta + (-\ln u_2)^\theta\right\}^{-\frac{1}{\theta}}\right]. \end{aligned} \tag{6.32}$$

The Gumbel-Hougaard copula represents the case of independence and positive dependence. When θ approaches 1, the marginals become independent, and when θ tends to infinity the Gumbel-Hougaard copula approaches the Fréchet-Hoeffding upper bound (i.e., the marginals become positively dependent). The relationship

between Kendall's tau and θ is given by

$$\tau = \frac{\theta - 1}{\theta}. \tag{6.33}$$

There is no closed-form expression for the relation between Spearman's rho and the copula parameter θ for the Gumbel-Hougaard copula. The Gumbel-Hougaard copula has upper-tail dependence for $\theta \neq 1$ and zero lower-tail dependence.

Example 5: Clayton copula The Clayton copula is from the Archimedian class of parametric copulas and is defined as

$$C(u_1, u_2|\theta) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}}, \tag{6.34}$$

where θ is the copula parameter that is restricted to the interval $(0, \infty)$.

The density of the Clayton copula is defined as

$$c(u_1, u_2|\theta) = \frac{(1+\theta)(u_1 u_2)^{-1-\theta}}{(u_1^{-\theta} + u_2^{-\theta} - 1)^{\frac{1}{\theta}+2}}. \tag{6.35}$$

For the Clayton copula, when θ approaches 0, the marginals become independent, and when θ tends to infinity the marginals become positively dependent. The relationship between Kendall's tau and θ is given by

$$\tau = \frac{\theta}{\theta + 2}. \tag{6.36}$$

The Clayton copula has lower-tail dependence and zero upper-tail dependence.

Other examples of copulas belonging to the Archimedian class include the Frank and Joe copulas.

6.2.5 Conditional copulas

The dependence between two random variables is often characterised by a single number. However, sometimes it is of interest to investigate whether the dependence between two random variables, Y_1 and Y_2, changes depending on a covariate, M. A way to investigate this is by using a conditional copula.

Definition 4: The conditional copula of $(Y_1, Y_2)|M = m$, where $Y_1|M = m \sim F_{1|M}(\cdot|m)$ and $Y_2|M = m \sim F_{2|M}(\cdot|m)$, is the conditional joint distribution function of $U = F_{1|M}(Y_1|m)$ and $V = F_{2|M}(Y_2|m)$ given $M = m$.

Patton (2006) extended Sklar's Theorem to conditional distributions.

Theorem 2: Extension of Sklar's Theorem to conditional distributions. Let $F_{1|M}(\cdot|m)$ be the conditional distribution of $Y_1|M = m$, $F_{2|M}(\cdot|m)$ be the conditional distribution of $Y_2|M = m$, $F_M(\cdot|m)$ be the joint distribution of $(Y_1, Y_2)|M = m$, and $\mathcal{M}$ be the support of M. If $F_{1|M}(\cdot|m)$ and $F_{2|M}(\cdot|m)$ are continuous in y_1 and y_2, for all $m \in \mathcal{M}$, then there exists a unique conditional copula $C(\cdot|m)$, such that

$$F_M(y_1, y_2|m) = C(F_{1|M}(y_1|m), F_{2|M}(y_2|m|m)), \tag{6.37}$$

$\forall\, (y_1, y_2) \in (-\infty, \infty) \times (-\infty, \infty)$ and each $m \in \mathcal{M}$. Conversely, if $F_{1|M}(y_1|m)$ is the conditional distribution of $Y_1|M = m$, $F_{2|M}(y_1|m)$ is the conditional distribution of $Y_2|M = m$, and $C(\cdot|m)$ is a family of conditional copulas measurable in m, then the $F_M(\cdot|m)$ defined in (6.37) is a conditional bivariate distribution function with conditional marginal distributions $F_{1|M}(\cdot|m)$ and $F_{2|M}(\cdot|m)$.

7

A copula-based pseudo-likelihood approach for investigating the association between gene-module co-expression and a continuous covariate

7.1 Introduction

Fitting a multivariate model that completely captures the dependence structure of several dependent variables is often a complex task in multivariate statistical modelling. In Chapter 5, a multivariate normal linear model was specified for investigating the metabolite-co-expression association of a gene module. The model utilizes a complex correlation structure to model the correlation between adjusted gene-expression values in function of the continuous covariate (i.e., metabolite concentrations). Maximising the full multivariate likelihood may not always be feasible due to computational challenges that arise as the number of parameters and the number of genes that constitute the gene module increases. Adopting a pseudo-likelihood approach can significantly reduce the challenge of constructing the multivariate distribution.

In this chapter, we propose to estimate the parameters of the multivariate normal linear model described in Section 5.2.1 with a pseudo-likelihood function. In particular, the multivariate density is replaced by the product of all pairwise densities over the set of all possible gene pairs within the gene module. Additionally, the bivariate densities are modelled by using Gaussian conditional copulas that specify the gene-pair correlations as functions of the metabolite concentration. In this way, the computational burden is reduced. Moreover, the Gaussian conditional copula facilitates the estimation of other non-parametric measures of the association (for instance, Kendall's tau and Spearman's rho). Pseudo-likelihood ratio (PLR) tests can be employed to infer conditional co-expression. The proposed model is applied to the

DILGOM data, described in Chapter 3, to investigate the metabolite-co-expression association of the core LL module. The Gaussian copula-based pseudo-likelihood approach opens the possibility to consider a wide range of copulas to study the conditional co-expression of a gene module. As an alternative to assuming a Gaussian distribution, we consider modelling the bivariate densities using Gumbel-Hougaard and Clayton conditional copulas. A simulation study is conducted to investigate the Type I error probability and power of the PLR test.

The chapter is organised as follows. Section 7.2 describes the copula-based pseudo-likelihood approach, simulation study, and DILGOM study. Results of the simulation study and DILGOM study are reported on in Section 7.3. A discussion of the results and conclusions are provided in Section 7.4.

7.2 Statistical methodology

7.2.1 Copula-based pseudo-likelihood approach

In this section, we first describe a pseudo-likelihood function with Gaussian copulas for approximating the multivariate normal model described in Section 5.2.1. The pseudo-likelihood is then adapted, through the use of non-Gaussian copulas, for a multivariate model with an 'unspecified' distribution. The estimation of the parameters of the pseudo-likelihood and an approach for inferring conditional co-expression is described towards the end of the section.

Pseudo-likelihood with Gaussian copulas for a multivariate normal model

Pseudo-likelihood estimation using a pairwise-likelihood function and conditional copulas is used to approximate the multivariate normal likelihood defined by (5.6) in Section 5.2.1.

Consider the set S of indices corresponding to all possible pairs of genes (g_1, g_2) in a gene module, where $1 \leq g_1 < g_2 \leq G$ (G = 7 for the core LL module). Let Y_{g_1} and Y_{g_2} denote the gene-expression values for genes g_1 and g_2, respectively. Let $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,G})^T$ denote the response vector for the i-th individual, where $i = 1, \ldots, N$.

The pseudo-likelihood function, denoted by pl, is obtained by evaluating the product of the bivariate densities of all possible pairs of genes in the gene module. In mathematical notation,

$$pl = \prod_{i=1}^{N} \prod_{(g_1,g_2)\in S} f_{g_1 g_2|X}(y_{i,g_1}, y_{i,g_2}|\boldsymbol{X}_i, \boldsymbol{\Theta}^*_{g_1 g_2}), \tag{7.1}$$

where $\boldsymbol{X}_i$ is a $2 \times R$-dimensional matrix of covariates for the i-th individual, $f_{g_1g_2|X}(y_{i,g_1}, y_{i,g_2}|\boldsymbol{X}_i, \boldsymbol{\Theta}^*_{g_1g_2})$ is the value of the bivariate density of genes g_1 and g_2 conditional on $\boldsymbol{X}$ evaluated for the respective responses of individual i, and $\boldsymbol{\Theta}^*_{g_1g_2}$ is a vector comprised of the parameters that characterise the bivariate density of genes g_1 and g_2. It is a sub-vector of the full vector of parameters $\boldsymbol{\Theta}^*$. The model parameters $\boldsymbol{\Theta}^*$ can be chosen in the same way as the parameters of the full-multivariate likelihood, defined by (5.6), in order to retain their meaning.

The bivariate densities of (7.1) are modeled using conditional copulas. Let the full vector of parameters

$$\boldsymbol{\Theta}^* = (\boldsymbol{\beta}^*_1, \ldots, \boldsymbol{\beta}^*_G, \sigma^{2*}_1, \ldots, \sigma^{2*}_G, \gamma^*_1, \ldots, \gamma^*_{G(G-1)/2}, \delta^*_1, \ldots, \delta^*_{G(G-1)/2})^T,$$

and let the gene-pair specific vector of parameters

$$\boldsymbol{\Theta}^*_{g_1g_2} = (\boldsymbol{\beta}^*_{g_1}, \boldsymbol{\beta}^*_{g_2}, \sigma^{2*}_{g_1}, \sigma^{2*}_{g_2}, \gamma^*_{g_1g_2}, \delta^*_{g_1g_2})^T.$$

By utilizing result (6.22), we obtain

$$\begin{aligned} pl &= \prod_{i=1}^{N} \prod_{(g_1,g_2)\in S} f_{g_1g_2|X}(y_{ig_1}, y_{ig_2}|\boldsymbol{X}_i, \boldsymbol{\beta}^*_{g_1}, \boldsymbol{\beta}^*_{g_2}, \sigma^{2*}_{g_1}, \sigma^{2*}_{g_2}, \gamma^*_{g_1g_2}, \delta^*_{g_1g_2}), \\ &= \prod_{i=1}^{N} \prod_{(g_1,g_2)\in S} c(u_{i,g_1}, u_{i,g_2}|\boldsymbol{x}_i, \boldsymbol{\beta}^*_{g_1}, \boldsymbol{\beta}^*_{g_2}, \sigma^{2*}_{g_1}, \sigma^{2*}_{g_2}, \gamma^*_{g_1g_2}, \delta^*_{g_1g_2}) \\ &\qquad \times f_{g_1|X}(y_{i,g_1}|\boldsymbol{x}_i, \boldsymbol{\beta}^*_{g_1}, \sigma^{2*}_{g_1}) \\ &\qquad \times f_{g_2|X}(y_{i,g_2}|\boldsymbol{x}_i, \boldsymbol{\beta}^*_{g_2}, \sigma^{2*}_{g_2}), \end{aligned} \tag{7.2}$$

where $c\left(\cdot|\boldsymbol{\Theta}^*_{g_1g_2}\right)$ denotes the density of the bivariate conditional copula for genes g_1 and g_2, and u_{i,g_1} and u_{i,g_2} are the values of the conditional marginal distributions of Y_{g_1} and Y_{g_2}, respectively, evaluated for the i-th individual. In particular,

$$u_{i,g_1} = F_{g_1|X}(y_{i,g_1}|\boldsymbol{x}_i, \boldsymbol{\beta}^*_{g_1}, \sigma^{2*}_{g_1}), \tag{7.3}$$

and

$$u_{i,g_2} = F_{g_2|X}(y_{i,g_2}|\boldsymbol{x}_i, \boldsymbol{\beta}^*_{g_2}, \sigma^{2*}_{g_2}). \tag{7.4}$$

Furthermore, in equation (7.2), $f_{g_1|X}(y_{i,g_1}|\boldsymbol{x}_i, \boldsymbol{\beta}^*_{g_1}, \sigma^{2*}_{g_1})$ and $f_{g_2|X}(y_{i,g_2}|\boldsymbol{x}_i, \boldsymbol{\beta}^*_{g_2}, \sigma^{2*}_{g_2})$ correspond to the values of the conditional marginal densities of Y_{g_1} and Y_{g_2}, respectively, for the i-th individual, and $\boldsymbol{x}_i$ is a vector of covariates of length R for individual i. The parameters $\boldsymbol{\beta}^*_{g_1}$, $\boldsymbol{\beta}^*_{g_2}$, $\sigma^{2*}_{g_1}$, and $\sigma^{2*}_{g_2}$ characterise the conditional marginal distributions (densities), and $\gamma^*_{g_1g_2}$ and $\delta^*_{g_1g_2}$ characterise the copula parameter. Specifically, $\boldsymbol{\beta}^*_{g_1}$ and $\boldsymbol{\beta}^*_{g_2}$ are vectors of regression parameters corresponding to the

conditional marginal distributions (densities) of Y_{g_1} and Y_{g_2}, respectively, while $\sigma^{2*}_{g_1}$ and $\sigma^{2*}_{g_2}$ denote the variance of gene g_1 and g_2, respectively. Analogous to $\gamma_{g_1g_2}$ and $\delta_{g_1g_2}$ in (5.3), $\gamma^*_{g_1g_2}$ and $\delta^*_{g_1g_2}$ are intercept and slope parameters which define the correlation between genes g_1 and g_2 as a function of the metabolite concentrations, denoted by m_i.

To approximate the multivariate normal likelihood, defined by (5.6), the Gaussian copula density function, defined by (6.28), with normal margins is used to model the bivariate densities.

When using the Gaussian conditional copula, the Fisher-z-transformed correlations are modelled as a linear function of the metabolite concentrations m_i:

$$\ln\left(\frac{1+\rho_{i,g_1g_2}}{1-\rho_{i,g_1g_2}}\right) = \gamma^*_{g_1g_2} + \delta^*_{g_1g_2} g(m_i), \tag{7.5}$$

where ρ_{i,g_1g_2} is the Gaussian copula parameter (corresponding to the Pearson correlation coefficient) between genes g_1 and g_2 for the i-th individual. This is the same relationship that was used in the multivariate normal model. From (7.5) it can be deduced that

$$\rho_{i,g_1g_2} = \frac{\exp\left\{\gamma^*_{g_1g_2} + \delta^*_{g_1g_2} g(m_i)\right\} - 1}{\exp\left\{\gamma^*_{g_1g_2} + \delta^*_{g_1g_2} g(m_i)\right\} + 1}. \tag{7.6}$$

As described in Section 5.2.1, there are many choices for the function $g(\cdot)$. For instance, one can assume $g(m_i) = m_i$ or $g(m_i) = \ln(m_i)$.

The pseudo-likelihood function, defined by (7.2)–(7.5), which incorporates the Gaussian copula density, defined by (6.28), approximates the multivariate normal linear likelihood defined by (5.2), (5.3), and (5.6).

A benefit of using copulas is that scale-free measures of association such as Kendall's tau, denoted by τ, or Spearman's rho, denoted by ρ_s, can be used to assess the strength of the dependence between the variables.

An alternative to (7.5) is to model the logit-transformed Kendall's tau correlations as a linear function of the metabolite concentrations. In particular,

$$\text{logit}(\tau_{i,g_1g_2}) = \ln\left(\frac{\tau_{i,g_1g_2}}{1-\tau_{i,g_1g_2}}\right) = \gamma^*_{g_1g_2} + \delta^*_{g_1g_2} g(m_i), \tag{7.7}$$

where τ_{i,g_1g_2} is Kendall's tau between genes g_1 and g_2 for the i-th individual. From (7.7), it follows that

$$\tau_{i,g_1g_2} = \frac{\exp\left\{\gamma^*_{g_1g_2} + \delta^*_{g_1g_2} g(m_i)\right\}}{\exp\left\{\gamma^*_{g_1g_2} + \delta^*_{g_1g_2} g(m_i)\right\} + 1}. \tag{7.8}$$

As stated in Section 6.2.4, the relationship between Kendall's tau and the Gaussian

copula parameter is given by

$$\tau_{i,g_1g_2} = \frac{2}{\pi}\arcsin\left(\rho_{i,g_1g_2}\right). \tag{7.9}$$

Thus, when using (7.7), the assumed relationship between the copula parameter (i.e., Pearson's correlation coefficient) and the metabolite concentration is

$$\rho_{i,g_1g_2} = \sin\left[\frac{\pi}{2}\frac{\exp\left\{\gamma^*_{g_1g_2} + \delta^*_{g_1g_2}g(m_i)\right\}}{\exp\left\{\gamma^*_{g_1g_2} + \delta^*_{g_1g_2}g(m_i)\right\}+1}\right]. \tag{7.10}$$

Pseudo-likelihood with non-Gaussian copulas

The pseudo-likelihood function described up until this point approximates a multivariate normal model by a product of bivariate normals through the use of Gaussian copulas. Apart from the Gaussian copula, there are many other copulas that can be used to model the bivariate densities in (7.2). The pseudo-likelihood function that arises from using an alternative copula, ceases to be an approximation of a multivariate normal likelihood. Rather, it corresponds to an 'unspecified' multivariate model. As alternatives to the Gaussian copula, we consider modelling the bivariate densities using the Gumbel-Hougaard copula, defined by (6.31), with corresponding density defined by (6.32), and the Clayton copula, defined by (6.34), with corresponding density defined by (6.35) .

For both the Gumbel-Hougaard and the Clayton copula-based pseudo-likelihoods, we assume that Kendall's tau is linked to the metabolite concentrations via the logit transformation, as defined in (7.7). The relationship between Kendall's tau and the Gumbel-Hougaard copula parameter is given by

$$\tau_{i,g_1g_2} = \frac{\theta_{i,g_1g_2}-1}{\theta_{i,g_1g_2}}, \tag{7.11}$$

where θ_{i,g_1g_2} is the Gumbel-Hougaard copula parameter between genes g_1 and g_2 for the i-th individual. The relationship between Kendall's tau and the Clayton copula parameter is given by

$$\tau_{i,g_1g_2} = \frac{\theta_{i,g_1g_2}}{\theta_{i,g_1g_2}+2}, \tag{7.12}$$

where θ_{i,g_1g_2} is the Clayton copula parameter between genes g_1 and g_2 for the i-th individual.

Estimation

The parameters of the pseudo-likelihood can be estimated by maximising the logarithm of the pseudo-likelihood function given by

$$\begin{aligned}\ln pl = \sum_{i=1}^{N} \sum_{(g_1,g_2)\in S} & \ln c(u_{i,g_1}, u_{i,g_2}|\boldsymbol{X}_i, \boldsymbol{\beta}^*_{g_1}, \boldsymbol{\beta}^*_{g_2}, \sigma^{2*}_{g_1}, \sigma^{2*}_{g_2}, \gamma^*_{g_1g_2}, \delta^*_{g_1g_2}) \\ & + \ln f_{g_1|X}(y_{i,g_1}|\boldsymbol{X}_i, \boldsymbol{\beta}^*_{g_1}, \sigma^{2*}_{g_1}) \\ & + \ln f_{g_2|X}(y_{i,g_2}|\boldsymbol{X}_i, \boldsymbol{\beta}^*_{g_2}, \sigma^{2*}_{g_2}). \end{aligned} \tag{7.13}$$

The pseudo-likelihood estimator $\hat{\boldsymbol{\Theta}}^*$ is defined as the maximiser of (7.13). The model-based variances of the parameters must be adjusted to account for the assumed independence amongst the terms forming the pseudo-likelihood. The empirically-corrected variance-covariance matrix is obtained from the asymptotic normality result discussed in Section 6.1 (i.e., Theorem 3), and is given by

$$\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Theta}}^*\right) = \boldsymbol{J}\left(\hat{\boldsymbol{\Theta}}^*\right)^{-1} \boldsymbol{K}\left(\hat{\boldsymbol{\Theta}}^*\right) \boldsymbol{J}\left(\hat{\boldsymbol{\Theta}}^*\right)^{-1} \tag{7.14}$$

where $\boldsymbol{J}\left(\hat{\boldsymbol{\Theta}}^*\right)$ can be estimated as minus the second derivative of the pseudo-likelihood function evaluated at $\hat{\boldsymbol{\Theta}}^*$, and $\boldsymbol{K}\left(\hat{\boldsymbol{\Theta}}^*\right)$ is the cross-product of the observed scores.

Inference

The null hypothesis model which corresponds to no metabolite-co-expression association can be obtained by setting the slope parameters $\delta^*_{g_1g_2}$, in (7.5) or (7.7), to zero. The null hypothesis can be tested using the PLR test statistic described in Section 6.1.1.

7.2.2 Simulation study

A simulation study is conducted to investigate the performance of the pseudo-likelihood approach in the multivariate normal case. The data simulated to investigate the Type I error probability and power of the LR test in Chapter 5 are used in this chapter to investigate the Type I error probability and the power of the PLR test and adjusted PLR test statistics described in Section 6.1.1. In particular, 1000 datasets of 450 observations each were simulated for each of the metabolite-co-expression association dynamics illustrated in Figure 5.2. For each dataset, metabolite concentrations were sampled from a $N(0,1)$ distribution. Gene-expression values for a seven-gene module were sampled from a multivariate normal distribution with means and variances corresponding to those observed for the core LL-module genes in the DILGOM subset.

The gene-pair correlations were simulated according to one of the six metabolite-co-expression association patterns illustrated in Figure 5.2. The explicit functional forms of each of the association patterns shown in Figure 5.2 were specified in Section 5.2.2.

The multivariate normal model defined by (5.1)–(5.3), and the null-hypothesis model defined by (5.1) and (5.8), were applied to the simulated data. Parameters were estimated by maximising the likelihood, defined by (5.6), as well as by maximising the pseudo-likelihood, defined by (7.2)–(7.5), with the Gaussian copula density, defined by (6.28), and normal margins.

This simulation study is based on 1000 replicates. This may seem insufficient. However, the decision was made taking into account the numerical complexity of the task, in particular, for fitting the multivariate model.

7.2.3 DILGOM analysis

The DILGOM data, described in Chapter 3, is analysed to illustrate the use of the pseudo-likelihood function, defined by (7.2)–(7.4), with bivariate densities modeled using Gaussian copulas. The data is also used to explore the use of alternative copulas such as the Gumbel-Hougaard and Clayton copulas. The pseudo-likelihood function is applied to the DILGOM subset to study the co-expression dynamics of the core LL-module conditional on serum-metabolite concentrations. The following forms of the pseudo-likelihood were maximized for each of the six considered metabolites (described in Section 3.1.1):

model A(H_1) with the Gaussian copula density (as in 6.28) and the Fisher-z relationship (as in 7.5);

model A(H_0) with the Gaussian copula density and $\delta^*_{12} = \delta^*_{13} = \ldots = \delta^*_{G-1,G} = 0$ in the Fisher-z relationship (i.e., the null-hypothesis model);

model B(H_1) with the Gaussian copula density and the logit relationship (as in 7.7);

model B(H_0) with the Gaussian copula density and $\delta^*_{12} = \delta^*_{13} = \ldots = \delta^*_{G-1,G} = 0$ in the logit relationship;

model C(H_1) with the Gumbel-Hougaard density (as in 6.32) and the logit relationship;

model C(H_0) with the Gumbel-Hougaard density and $\delta^*_{12} = \delta^*_{13} = \ldots = \delta^*_{G-1,G} = 0$ in the logit relationship.

model D(H_1) with the Clayton density (as in 6.35) and the logit relationship;

model D(H_0) with the Clayton density and $\delta^*_{12} = \delta^*_{13} = \ldots = \delta^*_{G-1,G} = 0$ in the logit relationship.

Using multiple linear regression, the gene-expression values of each gene were adjusted for gender, age, and metabolite concentration. The metabolite concentrations were standardized to avoid instability in the estimation procedure caused by the exceptionally small observed metabolite concentrations (see Table 3.1). In particular,

$$g(m_i) = \frac{m_i - \mu_m}{\sigma_m}, \tag{7.15}$$

was used in (7.5) and (7.7), where μ_m and σ_m are the mean and standard deviation of the metabolite concentrations, respectively. Inference of conditional co-expression was based on the PLR test statistics comparing the alternative and null-hypothesis models, i.e., $A(H_1)$ vs. $A(H_0)$, $B(H_1)$ vs. $B(H_0)$, $C(H_1)$ vs. $C(H_0)$, and $D(H_1)$ vs. $D(H_0)$.

A data-driven approach based on the Akaike Information Criterion (AIC) of each gene-pair's contribution to the pseudo-likelihood was used to select the best copula (from the Gaussian, Gumbel-Hougaard, and Clayton copulas) for each gene pair. The AIC is defined as

$$AIC = -2\,(\text{log-likelihood}) + 2k, \tag{7.16}$$

where k is the number of model parameters. Since the parameters of model $B(H_1)$, model $C(H_1)$, and model $D(H_1)$ are selected in the same way and have the same interpretation, the number of parameters corresponding to each gene-pair's contribution will be the same for the two models. As such, we can select the best copula for each gene pair by simply comparing their corresponding likelihood contributions to the pseudo-likelihood function. The most suitable copula for a particular gene pair, is the one with the largest likelihood contribution to the pseudo-likelihood function.

7.2.4 Implementation

The maximisation of the pseudo-likelihood functions was accomplished by using the R v.3.4.0 statistical programming language. The logarithm of the pseudo-likelihood defined by (7.13) was optimized by using the Newton-Raphson algorithm through the R package `maxLik` (Henningsen and Toomet, 2011). An analytical gradient was supplied to accelerate convergence. Derivatives of the Gaussian copula, Gumbel-Hougaard copula, and Clayton copula can be found in Appendix C.1. The starting values for the optimization routine were obtained by fitting linear models of the gene-expression values adjusted for age, gender, and metabolite concentrations. The regression parameters of the linear models were rounded off to one or six decimal places and used as starting values for the regression coefficients of the pseudo-likelihood function. Initial estimates of the variances and intercept parameters $\gamma_{g_1 g_2}$ were based on the variances and correlations of the linear-model residuals (also rounded off to

one or six decimal places), respectively. Initial estimates of zero were used for the slope parameters $\delta_{g_1 g_2}$.

7.3 Results

7.3.1 Simulation study: Gaussian copula

The Type I error probability of the PLR test statistic G^{*2}, defined by (6.10), and the adjusted PLR test statistics $G_a^{*2}(H_0)$ and $G_a^{*2}(H_1)$, defined by (6.11), was investigated. The Type I error probabilities of the statistics are reported in Table 7.1 for a seven-gene module and a sample size of 450 observations. Of the three test statistics, only G^{*2} (i.e., the unadjusted PLR test statistic) controls the Type I error probability at the nominal level of 0.05.

Table 7.1: Simulation study results: Estimated Type I error probability of the PLR test and adjusted PLR test statistics for a seven-gene module and a sample size of 450 observations.

Test statistic	Type I error probability*
$G^{*2}(H_0)$	0.048 [0.035, 0.061]
$G_a^{*2}(H_1)$	0.118 [0.098, 0.138]
$G_a^{*2}(H_0)$	0.113 [0.093, 0.133]

* point estimate [95% confidence interval]

Given the results of the Type I error investigation, the remaining simulation study results pertain to the power of the PLR test statistic G^{*2}. However, the estimated power of the adjusted PLR statistics, $G_a^{*2}(H_1)$ and $G_a^{*2}(H_0)$, based on the simulated distribution of the test statistics under the null hypothesis are given in Table C.2.1 and Table C.2.2. The simulation study results for G^{*2} are shown in Table 7.2. No convergence difficulties were encountered when optimising the pseudo-likelihood function for the six metabolite-co-expression dynamics. For the simulated seven-gene module and a sample size of 450 observations, G^{*2} has a Type I error probability of 0.048 (95% CI: $[0.035, 0.061]$). The power of the PLR test to detect the approximately linear negative and positive associations is large (0.984 and 0.988, respectively). Smaller estimates of the power of G^{*2} were obtained for the non-linear associations. In particular, for the wave association, a power of 0.580 is obtained; for the weak non-linear association, a power of 0.063 is obtained; and the lowest power of 0.044 was estimated for the parabola association. This is to be expected given that, in those cases, the simulated patterns were non-linear while the assumed relationship

between the correlations and the metabolite concentrations were monotonic functions of the concentrations.

Table 7.2: Simulation study results: Estimated Type I error probability and power of the PLR test for a seven-gene module and a sample size of 450 observations.

Co-expression dynamics	Type I error or power*
no metabolite-co-expression association	0.048 [0.035, 0.061]
approx. linear negative association	0.984 [0.976, 0.992]
approx. linear positive association	0.988 [0.981, 0.995]
non-linear association (wave)	0.580 [0.549, 0.611]
non-linear association (parabola)	0.044 [0.031, 0.057]
weak non-linear association	0.063 [0.048, 0.078]

∗ point estimate [95% confidence interval]

A comparison of the PLR and the LR test results is shown in Table 7.3. The maximisation of the full-likelihood did not achieve a hundred percent convergence, in contrast to the pseudo-likelihood. Table 7.3 only reflects the cases in which both the optimisation of the pseudo-likelihood and the optimisation of the full-likelihood converged. The results show that, in the considered setting, the pseudo-likelihood approach does not lead to any considerable power loss as compared to the full-likelihood maximization. In fact, for the approximate linear negative and positive association, and the wave association, we observe an increase in the power of the PLR test statistic compared to the LR test statistic. It appears that the closer the estimated co-expression dynamic is to the null hypothesis, the greater the loss in power of the PLR test when compared to the LR test.

Density plots of the PLR and LR test statistics are shown in Figure 7.1 and Figure 7.2, respectively, for the six co-expression dynamics. Higher powers are obtained for the co-expression dynamics with observed (PLR or LR test statistic) distributions that have the least overlap with the asymptotic distributions of the test statistics (shown in green). The empirical CDF plots of the PLR and LR test statistics are shown in Figure C.3.1 and Figure C.3.2, respectively, for the six co-expression dynamics.

Table 7.3: Simulation study results: Comparison of the PLR test and the LR test in terms of estimated Type I error probability and power for a seven-gene module and a sample size of 450 observations.

Co-expression dynamics	MVN convergence	Type I error or power* PLR test	LR test
no metabolite-co-expression association	918	0.041 [0.029, 0.054]	0.061 [0.046, 0.076]
approx. linear negative association	912	0.985 [0.977, 0.993]	0.863 [0.841, 0.885]
approx. linear positive association	923	0.988 [0.981, 0.995]	0.872 [0.851, 0.894]
non-linear association (wave)	922	0.573 [0.541, 0.605]	0.374 [0.343, 0.405]
non-linear association (parabola)	966	0.043 [0.031, 0.056]	0.142 [0.120, 0.164]
weak non-linear association	924	0.056 [0.041, 0.071]	0.070 [0.054, 0.087]

$*$ point estimate [95% confidence interval]

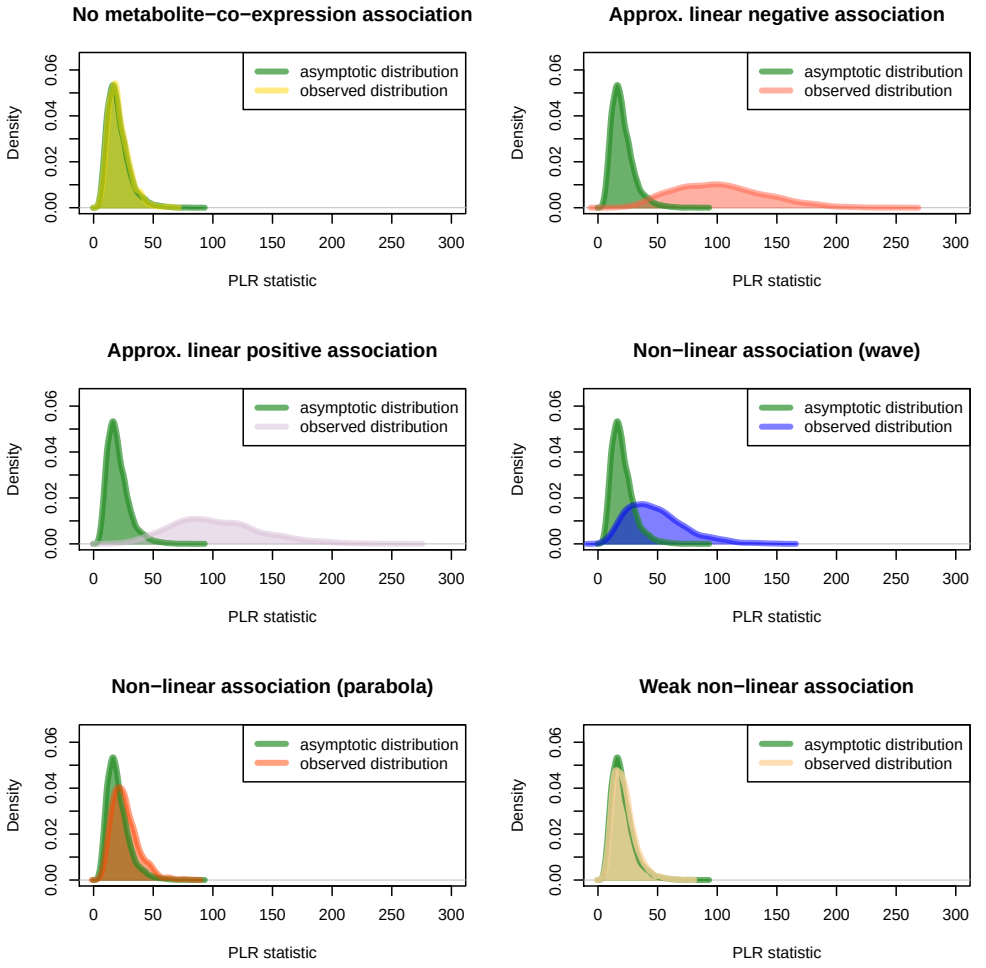

Figure 7.1: Density plots of the observed PLR test statistics for each of the six co-expression dynamics together with the asymptotic distribution of the PLR test statistic. In each plot, the asymptotic distribution of the PLR test statistic is shown in green.

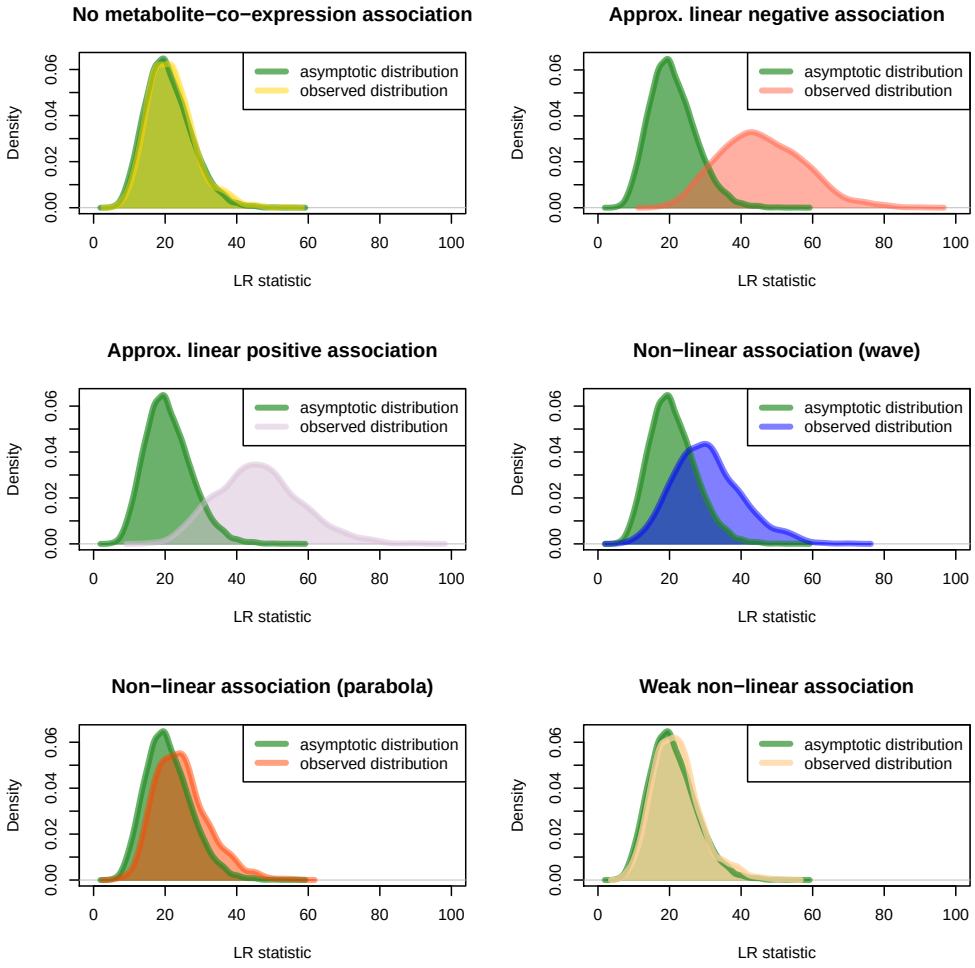

Figure 7.2: Density plots of the observed LR test statistics for each of the six co-expression dynamics together with the asymptotic distribution of the LR test statistic. In each plot, the asymptotic distribution of the LR test statistic is shown in green.

7.3.2 DILGOM analysis: Gaussian copula

Table 7.4 presents the results of the DILGOM analysis, conducted using the pseudo-likelihood function with Gaussian copulas, for the six considered metabolites. In particular, the table shows the unadjusted p-values of the PLR tests for the comparison of the alternative models with the null-hypothesis model of no metabolite-co-expression association.

Table 7.4: DILGOM analysis: Gaussian copula pseudo-likelihood-ratio test results.

Metabolite	Pseudo-likelihood-ratio test p-values $A(H_1)$ vs. $A(H_0)$[a]	$B(H_1)$ vs. $B(H_0)$[b]
3-hydroxybutyrate	0.5549	0.5573
linoleic acid	0.0571	0.0595
large HDL particles	0.0357	0.0334
small LDL particles	0.0789	0.0799
total cholesterol in large HDL	0.0223	0.0212
small HDL particles	0.0154	0.0156

[a] model $A(H_1)$: with Gaussian copula and Fisher-z relationship defined by (7.5)
model $A(H_0)$: with Gaussian copula and $\delta^*_{12} = \delta^*_{13} = \ldots = \delta^*_{G-1,G} = 0$ in (7.5)
[b] model $B(H_1)$: with Gaussian copula and logit relationship defined by (7.7)
model $B(H_0)$: with Gaussian copula and $\delta^*_{12} = \delta^*_{13} = \ldots = \delta^*_{G-1,G} = 0$ in (7.7)

Based on the PLR tests of models $A(H_1)$ vs. $A(H_0)$ for the six metabolites, we find that the core LL-module exhibits a significant metabolite-co-expression association with the concentration of large HDL particles ($p = 0.0357$), total cholesterol in large HDL ($p = 0.0223$), and small HDL particles ($p = 0.0154$). Figure 7.3 illustrates the estimated metabolite-co-expression dynamics based on model $A(H_1)$ for the six considered metabolites. Each curve corresponds to the estimated metabolite-co-expression dynamic of a particular gene pair of the core LL-module.

Figure 7.4 illustrates the estimated co-expression dynamic for 3-hydroxybutyrate based on model $A(H_1)$ and $A(H_0)$ in terms of Pearson's correlation, Spearman's rho, and Kendall's tau. The top row and the bottom row of each figure present the results of the alternate model $A(H_1)$ and null-hypothesis model $A(H_0)$, respectively. The closed form expressions for the relationship between Spearman's rho and the Gaussian copula parameter ρ (i.e., the Pearson correlation coefficient), defined by (6.30), as well as between Kendall's tau and ρ, defined by (6.29), are used to translate the estimated correlation trajectories from the Pearson correlation scale to the Spearman's rho and Kendall's tau scale.

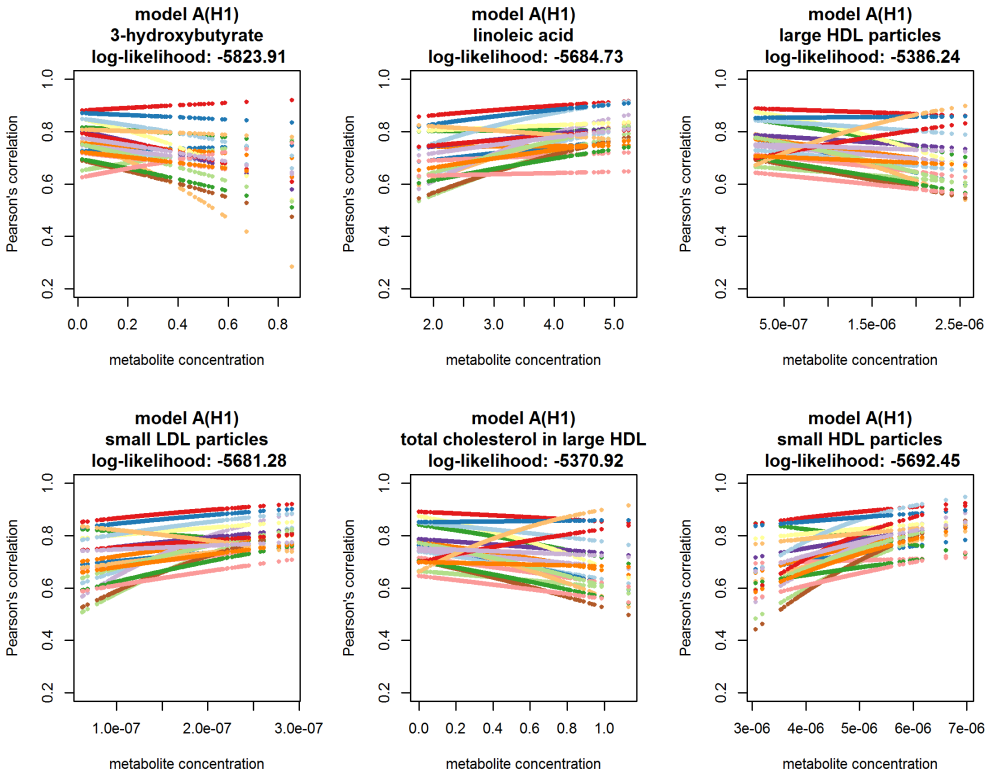

Figure 7.3: Estimated correlation dynamics for the considered metabolites based on model A(H_1) (i.e., the pseudo-likelihood function with a Gaussian copula and Fisher-z correlation-metabolite relationship defined by (7.5)).

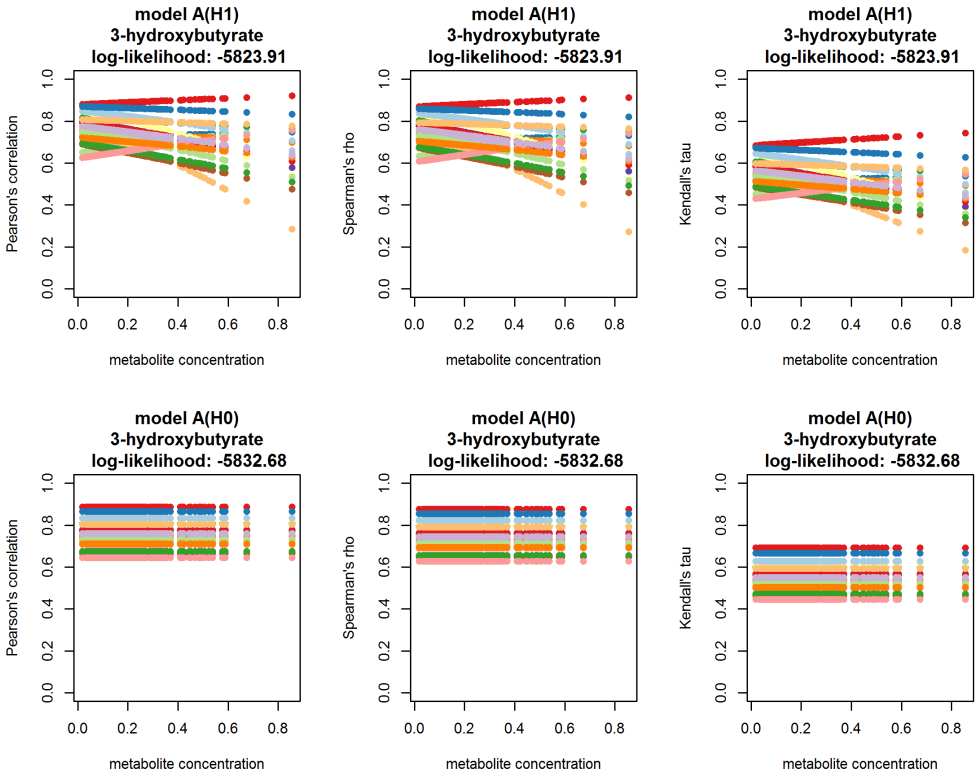

Figure 7.4: Estimated trajectories for the 3-hydroxybutyrate-co-expression association based on model A(H_1) and model A(H_0) in terms of Pearson's correlations, Spearman's rho, and Kendall's tau. The top row presents the results of the alternate model A(H_1) and the bottom row displays the results of the null-hypothesis model A(H_0).

Based on the PLR tests comparing models $B(H_1)$ and $B(H_0)$, for the six metabolites, a significant co-expression-metabolite effect is identified for the concentration of large HDL particles ($p = 0.0334$), total cholesterol in large HDL ($p = 0.0212$), and small HDL particles ($p = 0.0156$).

7.3.3 DILGOM analysis: Non-Gaussian copulas

Table 7.5 presents the results of the DILGOM analysis, using the pseudo-likelihood function with non-Gaussian copulas, for the six considered metabolites. Based on the PLR tests comparing model $C(H_1)$ vs. $C(H_0)$, there is insufficient evidence to conclude a metabolite-co-expression association for any of the six metabolites. The PLR test p-values for $D(H_1)$ vs. $D(H_0)$ indicate that the core LL-module exhibits a significant metabolite-co-expression association with the concentration of large HDL particles ($p = 0.0089$) and total cholesterol in large HDL ($p = 0.0097$).

Table 7.5: DILGOM analysis: Non-Gaussian copula pseudo-likelihood-ratio test results.

Metabolite	Pseudo-likelihood-ratio test p-values	
	$C(H_1)$ vs. $C(H_0)$[c]	$D(H_1)$ vs. $D(H_0)$[d]
3-hydroxybutyrate	0.7062	0.5832
linoleic acid	0.2958	0.2421
large HDL particles	0.3011	0.0089
small LDL particles	0.3750	0.0867
total cholesterol in large HDL	0.2312	0.0097
small HDL particles	0.1215	0.2240

[c] model $C(H_1)$: with Gumbel-Hougaard copula and logit relationship defined by (7.7)
model $C(H_0)$: with Gumbel-Hougaard copula and $\delta^*_{12} = \delta^*_{13} = \ldots = \delta^*_{G-1,G} = 0$ in (7.7)
[d] model $D(H_1)$: with Clayton copula and logit relationship defined by (7.7)
model $D(H_0)$: with Clayton copula and $\delta^*_{12} = \delta^*_{13} = \ldots = \delta^*_{G-1,G} = 0$ in (7.7)

Figure 7.5 illustrates the co-expression dynamics in terms of Kendall's tau, as estimated by model $B(H_1)$, model $C(H_1)$, and model $D(H_1)$ for 3-hydroxybutyrate. Slight differences can be seen in the estimated co-expression dynamics of the three models. Comparisons of the log-likelihood contributions of each gene pair to the pseudo-likelihood of model $B(H_1)$ (i.e., with the Gaussian copula), the pseudo-likelihood of model $C(H_1)$ (i.e., with the Gumbel-Hougaard copula), and the pseudo-likelihood of model $D(H_1)$ (i.e., with the Clayton copula) indicate that the Gaussian copula consistently provides a better fit for the bivariate densities.

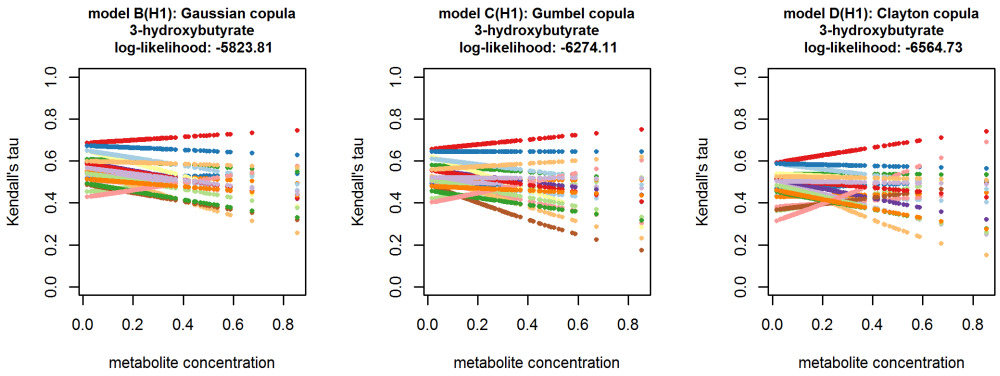

Figure 7.5: Estimated trajectories for the 3-hydroxybutyrate-co-expression association based on model $B(H_1)$ (i.e., with the Gaussian copula), model $C(H_1)$ (i.e., with the Gumbel-Hougaard copula), and model $D(H_1)$ (i.e., with the Clayton copula).

7.4 Discussion & Conclusions

In this chapter, we proposed using pseudo-likelihood estimation, in particular, a pairwise-likelihood function with conditional copulas, to approximate the full-multivariate likelihood (5.6) of model (5.1)–(5.3) described in Chapter 5. By using the pseudo-likelihood function, the computational burden is reduced. The use of conditional copulas provides increased flexibility. By using conditional copulas, co-expression can be modelled in terms of scale-free measures of association, such as Kendall's tau and Spearman's rho, in function of the metabolite concentration. Furthermore, the pseudo-likelihood can be implemented using a non-Gaussian copula; thereby, avoiding the assumption of bivariate normally-distributed random variables. Inference of conditional co-expression can be based on PLR test statistics comparing the the null model of no metabolite-dependent co-expression with the alternate model with unrestricted intercept $\gamma_{g_1 g_2}$ and slope $\delta_{g_1 g_2}$ parameters.

Often, the price-to-pay for the computational simplicity gained through the use of a pseudo-likelihood approach is a loss in efficiency. In the simulation study, when the estimated co-expression dynamic is relatively close to the null hypothesis (i.e., for the non-linear parabola association and weak non-linear association), the PLR test experiences a drop in power compared to the LR test. This result is not surprising in the context of Molenberghs et al. (2011) who investigate the asymptotic efficiency of a pseudo-likelihood approach for partitioned samples (PPL) relative to maximum likelihood (ML). The asymptotic relative efficiency is defined as the variance of the ML to the PPL. In the case of dependent sub-samples of a compound-symmetry multivariate normal sample, they find that full efficiency is sometimes, but not always, reached. In fact, they illustrate that the efficiency loss depends on the parameter under consideration (i.e., the mean, variance, or correlation), and, for variance-components, possibly on the value of the parameter. Situations in which the pseudo-likelihood outperforms a maximum-likelihood approach, as observed in our simulation study for the approximate linear negative (positive) association and the non-linear wave association, have been reported in the literature. For instance, Andor and Parmeter (2017) examine the mean square errors of parameters estimated using PL and ML estimation, and find that in their scenario, PL outperforms ML when the sample size is small.

In the DILGOM analysis, the PLR test comparing the alternate model $\mathrm{A}(\mathrm{H}_1)$ and null-hypothesis model $\mathrm{A}(\mathrm{H}_0)$ for the six considered metabolites indicated that three metabolites, i.e., large HDL particles, total cholesterol in large HDL, and small HDL particles, are associated with core LL-module co-expression. In Section 5.3.2, the LR test statistics comparing the alternate model A and the null-hypothesis model D only indicated an association between total cholesterol in large HDL and core LL-module co-expression (see Table 5.2). Thus, in comparison to the full-likelihood maximisation, a greater number of metabolite-co-expression associations were identified through

maximising the pseudo-likelihood.

In the DILGOM analysis, the bivariate densities of the pseudo-likelihood are modeled using Gaussian copulas in model $B(H_1)$, Gumbel-Hougaard copulas in model $C(H_1)$, and Clayton copulas in model $D(H_1)$. A multitude of other copulas can be selected to model the bivariate densities. The Gumbel-Hougaard and Clayton copulas were selected for ease of illustration as an alternative to the Gaussian copula. The likelihood contribution of each gene pair towards the pseudo-likelihood of model $B(H_1)$, model $C(H_1)$, and model $D(H_1)$ was compared in order to determine which copula is better suited for modelling each pair of genes of the core LL-module. A formal approach for pseudo-likelihood model selection was not addressed in this chapter and is a topic for further research.

Part II

The impact of the method of extracting metabolic signal from ^{1}H-NMR data on the classification of samples: a case study for lung cancer

8

Introduction to metabolic data analysis

Metabolomics is a comprehensive analysis in which the small molecule or metabolite composition of cells, tissues, or biofluids (e.g., urine, cerebrospinal fluid, or blood plasma) is identified and quantified. The metabolome (i.e., the complete set of metabolites present in a biological sample, in a particular physiological state) is comprised of a diverse group of small molecules with molecular masses less than 1500 Da. It includes compounds belonging to various chemical classes including amino acids, sugars, organic acids, and lipids amongst others.

Metabolites are the intermediates or the end products of virtually all biological processes. The metabolome is located further down the cascade from gene to function than the transcriptome and proteome. Thus, changes that occur in the genome, transcriptome, or proteome are reflected in the metabolome (Stringer et al., 2016). The metabolome is the closest measurable representation of the phenotype currently available (Beisken et al., 2015). Analyzing the metabolic composition of biological samples has considerable potential for disease diagnosis (Gowda et al., 2008; Louis et al., 2016a,b). Metabolic profiling also provides information about patient heterogeneity that could play a pivotal role in personalized medicine (Stringer et al., 2016).

Proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy and mass spectrometry (MS) are two analytical techniques that are routinely used for profiling metabolites. MS is more sensitive than ^{1}H-NMR spectroscopy, but requires an extraction step to separate the hydrophilic from the hydrophobic metabolites. ^{1}H-NMR spectroscopy is a popular choice as it requires minimal sample preparation and is non-destructive (i.e., the biological sample remains intact).

In this study, the spectra obtained by applying ^{1}H-NMR spectroscopy to blood plasma are analysed. ^{1}H-NMR spectroscopy exploits the magnetic properties of hydrogen nuclei; that is, in a strong external magnetic field, a short radiofrequency (RF) pulse causes hydrogen nuclei to absorb and subsequently emit electromagnetic (EM) radiation. The frequency of RF radiation that is required to bring hydrogen nuclei into resonance (i.e., the frequency of absorbed and re-emitted radiation), is called the resonance frequency (MHz) and it is influenced by the strength of the magnetic field

and the chemical environment of the hydrogen nuclei. Resonating hydrogen nuclei produce an NMR response which is called the free induction decay (FID). The FID (time domain signal) is Fourier transformed to obtain a ^{1}H-NMR spectrum (frequency domain signal) that is visualized as a series of peaks along a chemical shift axis (see Figure 8.1). The peaks correspond to the resonating hydrogen nuclei. The unit of the chemical shift axis is parts per million (ppm), i.e., the difference between the resonance frequency of the hydrogen nucleus of the metabolite and the hydrogen nucleus of a reference compound, divided by the resonance frequency of the reference compound. Each metabolite in the biological sample produces a characteristic spectral signature that is formed by a combination of peaks not necessarily adjacent to each other along the chemical shift axis. The resonances of each metabolite present in the biological sample appears in the spectrum with an intensity proportional to the concentration of the corresponding metabolite in the sample.

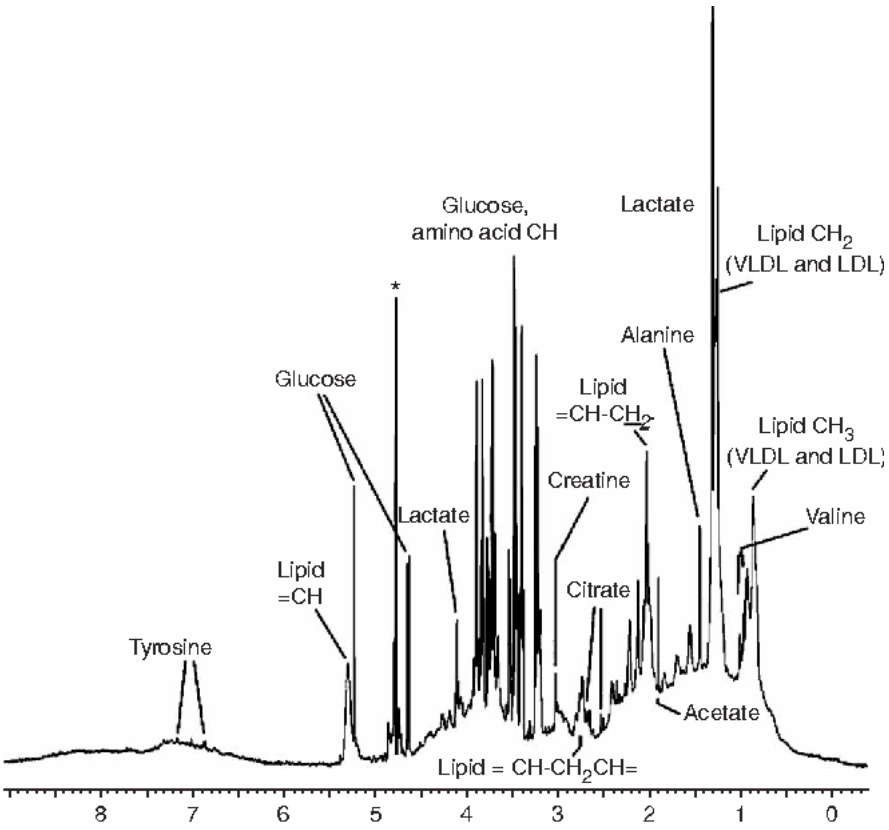

Figure 8.1: An example of a ^{1}H-NMR spectrum of a blood-serum sample. The spectrum shows resonances corresponding to metabolites as well as large molecules such as lipoproteins which have broader resonances. The source of this Figure is Beckonert et al. (2007).

The identification and quantification of blood plasma metabolites based on ^{1}H-NMR spectra is a challenge for the following reasons:

1. ^{1}H-NMR spectrometers have detection limits that allow for only a very small subset (approximately 40-50) of existing plasma metabolites to be reliably detected and quantified.

2. Baseline (i.e., the intensity in regions not containing peaks) distortions and broad resonances from lipoproteins and glycoproteins might mask the less prominent signals arising from low concentration metabolites (de Graaf and Behar, 2003).

3. Experimental conditions (e.g., pH and temperature) might influence the chemical shift position of metabolic peaks. In addition, more than one metabolite can contribute to a signal at a specific location which further complicates peak identification and metabolite quantification.

Typically, ^{1}H-NMR metabolomics of blood plasma is conducted using spectrometers with magnetic field strengths ranging from 9.4 Tesla to 14.1 Tesla, i.e., with proton resonance frequencies ranging from 400 MHz to 600 MHz. Higher-field spectrometers (e.g., 900 MHz spectrometers) produce spectra with improved resolution. The ability to resolve peaks with different chemical shifts increases with field strength. However, higher-field spectrometers are also far more costly (Louis et al., 2017).

Spectral binning (Louis et al., 2015) is a simple and commonly used technique for extracting metabolic signal from NMR spectra. It involves subdividing the spectra into regions along the chemical shift axis and computing the area under the curve within each integration region. However, peak overlap and variation in the chemical shift positions of the peaks across spectra often prevents a one-to-one mapping between integration regions and metabolites. This may be especially problematic in the context of sample classification. In particular, an integration region may fail to show potential for classification if it includes metabolites which show opposite behavior (under- and over-expression) in patients versus controls. For instance, assume an integration region encompasses signal coming from two discriminative metabolites. On average, one metabolite has a higher concentration in patients than in controls while the second metabolite has a lower concentration in patients than in controls. Despite the fact that the integration region contains signal from two discriminative metabolites, the opposite behavior of the two metabolites diminishes the classification potential of the integration region, potentially resulting in a non-differential integrated spectral region (ISR). On the other hand, for an ISR that shows classification potential it may be difficult to uniquely assign its effect to a single metabolite.

Since overlapping molecular resonances complicate the extraction of metabolic information from ^{1}H-NMR data, spectral deconvolution techniques are currently the

state of the art. BATMAN (Bayesian AuTomated Metabolite Analyser for NMR data) (Astle et al., 2012; Hao et al., 2014) is a Bayesian model for ^{1}H-NMR spectral deconvolution which resolves resonance peaks to obtain relative concentration estimates for a set of metabolites in an automated manner. It exploits extensive prior information on the characteristic resonance signatures of each metabolite and combines this information with the intensities observed in the actual spectrum to model the metabolic signal. Other deconvolution models include Bayesil and the commercially available software package Chenomx amongst others (Alonso et al., 2015; Misra and der Hooft, 2016). The advantage of BATMAN is its flexibility and adaptability to the problem at hand. The prior information on peak shape and relative intensity plays an important role in any spectral deconvolution and signal extraction model. Flexibility in setting up the prior information is desirable especially when ^{1}H-NMR spectroscopy is performed on a spectrometer different than the one used to create the spectral deconvolution software.

In this part of the dissertation, the application of the widely-used spectral binning approach is compared with the automated spectral deconvolution technique, BATMAN, for extracting metabolic signal from the ^{1}H-NMR spectra of blood plasma samples for the purposes of sample classification. The two approaches were applied to 400 MHz (medium-field) and 900 MHz (high-field) ^{1}H-NMR spectra of blood plasma samples from lung cancer patients and control subjects. The extracted features, i.e., the ISRs and the BATMAN estimated relative concentrations of the metabolites, were compared in terms of their ability to correctly classify lung cancer and control samples. This was performed separately for the 400 MHz and 900 MHz spectra.

A series of pre-processing steps were required to reduce the noise, external sources of variation, and artifacts which result during the process of NMR data acquisition before the metabolic signal could be extracted. Different pre-processing protocols were applied to the 400 MHz and 900 MHz ^{1}H-NMR spectra. In particular, the use of a more automated approach for pre-processing the 900 MHz ^{1}H-NMR spectra was investigated.

The content of the next couple of chapters is as follows: background information on ^{1}H-NMR spectroscopy is provided in Chapter 9. The data and pre-processing steps are described in Chapter 10. Chapter 11 provides a description of spectral binning and BATMAN, and details on the classification analysis can be found in Chapter 12.

9

Proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy

In this chapter, various aspects of ^{1}H-NMR spectroscopy are discussed. The technical details of ^{1}H-NMR spectroscopy, as described in this chapter, are based on Macomber (1998), Rousseau (2011), and Louis (2015). Section 9.1 describes the magnetic properties of hydrogen nuclei which form the basis of ^{1}H-NMR spectroscopy. In Section 9.2, an ^{1}H-NMR experiment is described, from the sample preparation to the detection of the time domain signal, and its transformation to the frequency domain. Further details on the functional form of the time domain signal (i.e., the free induction decay) is provided in Section 9.3, and details on the parameters which characterise an ^{1}H-NMR experiment are included in Section 9.4. Section 9.5 discusses the steps involved in pre-processing the time domain signal. Lastly, Section 9.6 introduces the parameters used to characterise the peaks of an ^{1}H-NMR spectrum (i.e., the frequency domain signal).

9.1 Basic principles of ^{1}H-NMR spectroscopy

^{1}H, ^{13}C, ^{15}N, and ^{31}P are amongst the most important nuclei for biomolecular NMR studies (Markley et al., 2017). However, the proton (^{1}H) is the most sensitive, and has near 100% natural abundance (Markley et al., 2017). An NMR experiment is sensitive to only one particular isotope of one particular element. One-dimensional (1D) ^{1}H-NMR spectroscopy is the most widely used NMR approach in metabolomics.

^{1}H-NMR spectroscopy exploits the magnetic properties of hydrogen nuclei. Nuclei with an odd atomic number, such as the nuclei of ^{1}H particles, rotate around an axis in a movement called nuclear spin. The spin of a nucleus (i.e., a charged particle) generates a magnetic field along the spin axis called the magnetic moment (see Figure 9.1).

In the absence of an external magnetic field B_0, the spin of the nuclei are randomly oriented (see left-side of Figure 9.2). Once a magnetic field is applied, the spin axis aligns with the external magnetic field (see right-side of Figure 9.2). The mag-

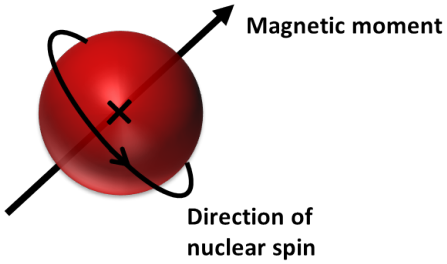

Figure 9.1: Illustration of the spin and magnetic moment of a positively charged hydrogen nucleus.

netic moments of the nuclei can either align with the field B_0 (i.e., become parallel or spin aligned) or align against it (i.e., become anti-parallel or spin opposed). Protons in parallel are in the more populated, lower-energy state, and are more stable than anti-parallel protons.

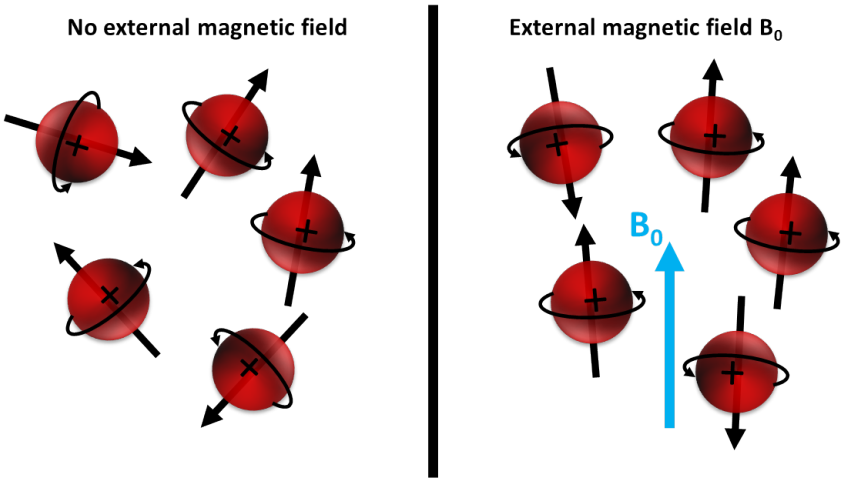

Figure 9.2: Orientation of the magnetic moments of positively charged hydrogen nuclei in the absence of an external magnetic field (left) and in the presence of an external magnetic field B_0 (right).

In fact, due to the interaction between the magnetic moment and the external magnetic field, the spin axis does not align perfectly with B_0, but rather precesses around it at an angle θ with a precession rate called the Larmor frequency denoted by ν (see Figure 9.3).

Protons can transition from the low-energy state to the high-energy state through the absorption of energy. The difference in the energy of the spin states depends on the strength of the external magnetic field (measured in tesla, T). The energy

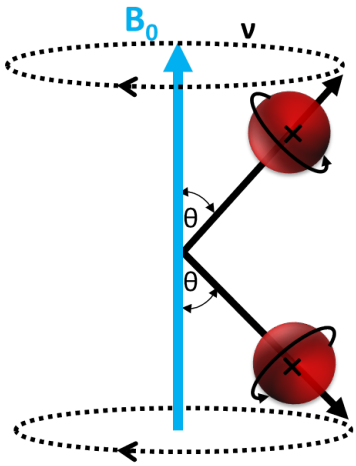

Figure 9.3: Precession of the magnetic moments of protons around the external magnetic field B_0. ν is the precession rate and θ is the angle of the proton's magnetic moment to B_0.

difference increases with an increase in magnetic field strength. However, this energy difference remains quite small, even for very strong magnetic fields (for example, of 21.1 tesla), and corresponds to the radio frequency (RF) range of the electromagnetic (EM) spectrum. In ^{1}H-NMR spectroscopy, energy is supplied to the protons via an RF pulse. If the frequency of the applied RF and the Larmor frequency coincide, resonance occurs. The spin absorbs the energy and shifts the proton to the higher-energy state. The higher-energy state is less stable and upon removal of the RF pulse, the nucleus returns to the initial low-energy state, emitting the previously absorbed energy at a frequency equal to that of the Larmor frequency. This produces a current in the detection coil of the NMR spectrometer. Based on the frequency of emission, one can determine which chemical group the proton belongs to.

9.2 An ^{1}H-NMR experiment

Before a sample is analysed by using ^{1}H-NMR spectroscopy, several products are added to the sample including a magnetic field lock signal and a reference compound. For the blood plasma samples analysed in this dissertation, deuterium oxide was added as a magnetic field lock signal, and trimethylsilyl-2,2,3,3-tetradeuteropropionic acid (TSP) was employed to act as a reference compound.

The solution of the sample is placed between the poles of a strong magnet. Both the energy difference and the population difference between the spin states increase with an increase in magnetic field strength. At equilibrium, the number of spins in

the low-energy state is slightly higher than the number of spins in the high-energy state (see Figure 9.4). Thus, the sum of the individual magnetic moments of the protons belonging to the same chemical group (i.e., having the same neighbourhood), results in a net magnetic moment M_0 that aligns with the direction of the external magnetic field B_0. Using a three-dimensional co-ordinate system, at equilibrium, the z-component of magnetization, called the longitudinal magnetization, equals M_0, while the x and y components of magnetization, called the transversal magnetization, equals 0.

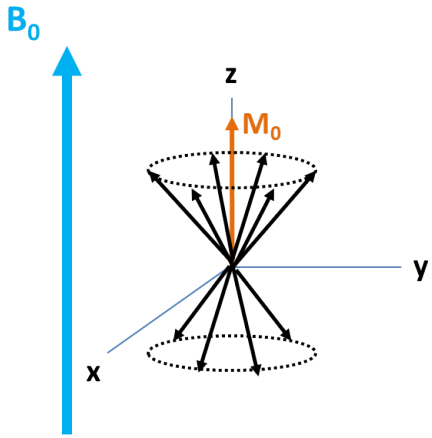

Figure 9.4: Distribution of the magnetic moments at equilibrium resulting in a longitudinal magnetization vector M_0 in the direction of the external magnetic field B_0. The transverse magnetization (i.e., in the xy-plane) is zero.

The NMR signal is obtained by applying an RF pulse perpendicular to the external magnetic field at the Larmor frequency ν of the observed protons. The time during which the RF pulse is applied is called the 90° pulse width (measured in μs). During the 90° pulse, the magnetic moments of the protons enter a state of phase coherence characterised by the clustering of the magnetic moments into a precessing bundle (see Figure 9.5). This produces a shift of the net magnetization moment M_0 from the longitudinal z-axis towards the y-axis of the transverse xy-plane. The energy transmitted by the RF pulse is absorbed by the protons causing some to flip from the low-energy state to the high-energy state. At the end of the 90° pulse, the longitudinal magnetization is zero, and the transversal magnetization equals M_0. The transversal magnetization vector rotates about the the z-axis which induces a current in the receiver coil of the NMR spectrometer.

The NMR signal is detected once the RF pulse is switched off. At this point, the nuclei relax and return to their equilibrium positions. The relaxation process has two components: a longitudinal component that corresponds to the recovery

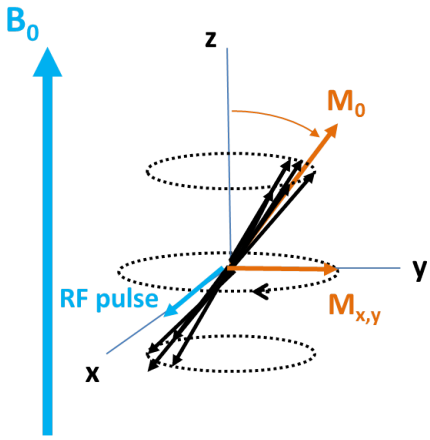

Figure 9.5: Illustration of the shift of the magnetic moments upon applying an RF-pulse perpendicular to the external magnetic field (i.e., along the x-axis). The net magnetic moment M_0 shifts away from the z-axis and now has a component in the transverse xy-plane.

of the longitudinal magnetization and a transverse component which corresponds to the transverse magnetization decay. Each relaxation process is described by an exponential function characterised by the time constants T_1 and T_2, respectively. During the relaxation period, the transversal magnetization induces a fluctuating electric current in the receiver coil of the NMR spectrometer which forms a time-dependent signal (see Figure 9.6). The signal decays with time as the magnetic moments move out of phase. Each signal is characterized by it's amplitude, frequency ν, and decay time T_2.

During an ^{1}H-NMR experiment several RF pulses of different frequencies are applied and several time-dependent signals are produced from protons of different Larmor frequencies. The recorded current, called the free induction decay (FID), is the sum of the time domain signals emitted by protons from different chemical groups that entered into resonance (see Figure 9.7). The FID is Fourier-transformed to produce an ^{1}H-NMR spectrum (i.e., a signal in the frequency domain).

The Fourier transformation of the FID (i.e., time domain signal) results in an ^{1}H-NMR spectrum that is visualized as a series of peaks in the frequency domain (see Figure 9.8). The peaks correspond to the resonating hydrogen nuclei. The position of the peaks in the frequency-domain spectrum corresponds to the frequency of each component forming the FID. During the pre-processing of the FID, the frequency scale is translated into a chemical shift scale expressed in parts per million (ppm), i.e., the difference between the resonance frequency of the hydrogen nucleus of a compound in the the sample and the hydrogen nucleus of a reference compound, divided by the

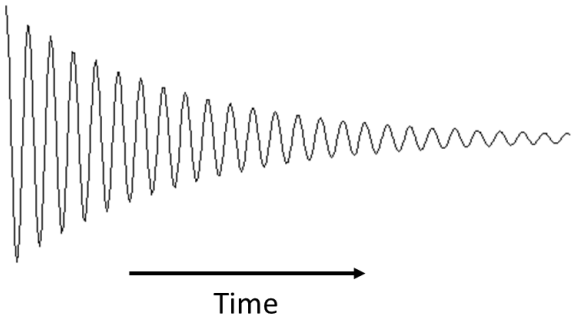

Figure 9.6: A single time domain signal.

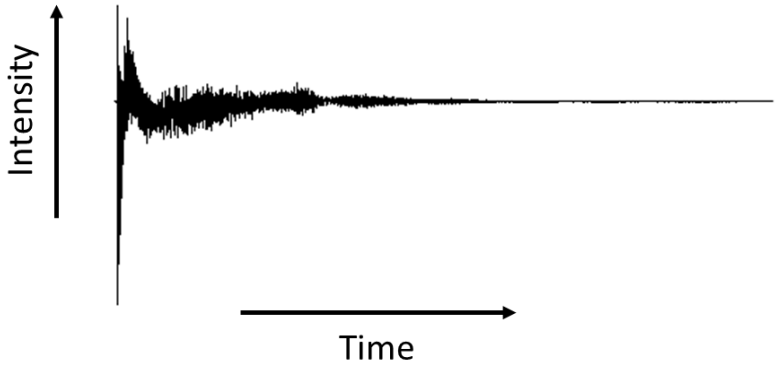

Figure 9.7: Illustration of a free induction decay (FID).

resonance frequency of the reference compound. Note that the chemical shift axis typically extends from 0 ppm on the right to larger values (for instance, 10 ppm) on the left, as illustrated in Figure 9.8. The area under the peak corresponds to the amplitude of the FID component. The width of a peak at half-height is inversely proportional to the time constant of the transverse magnetization decay, i.e., T_2. The faster the decay (i.e., smaller T_2), the broader the peak. Note that the area under the

peak remains constant, so a slower decay producing a sharper peak will also result in an increase in the peak height.

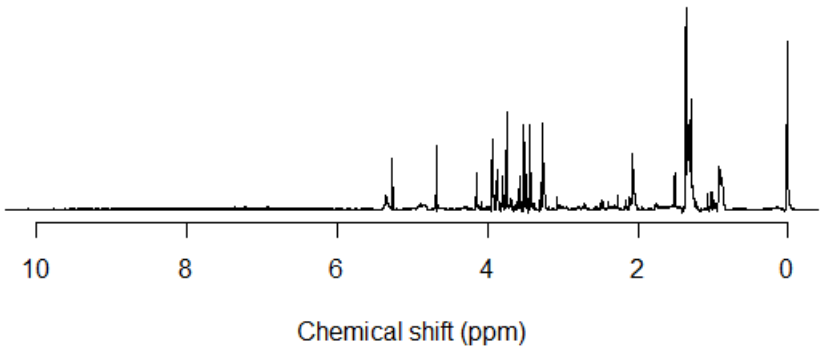

Figure 9.8: A Fourier transformed FID.

9.2.1 CPMG pulse sequence

The ^{1}H-NMR spectra of blood plasma samples display broad signals from macro molecules, such as proteins and polysaccharides, with sharp peaks from metabolites superimposed on them. Additionally, a large signal arising from water protons can be seen in ^{1}H-NMR spectra of biofluids. These signals obscure a large part of the ^{1}H-NMR spectrum if left unsuppressed. Generally, in order to improve the visibility of the metabolite signal, a method is applied to suppress the high-intensity signal caused by water molecules and to attenuate the broad signals from the macromolecules (i.e., proteins and polysaccharides). In this dissertation, slightly T_2-weighted spectra that were acquired using the Carr-Purcell-Meiboom-Gill (CPMG)-pre-saturation pulse sequence are analysed. The CPMG pulse sequence uses the faster transversal relaxation times (T_2) of protons in macromolecules to suppress the macromolecular signals.

9.3 Functional form of the free induction decay (FID)

As mentioned previously, an FID arises from the decay of the transverse magnetization which induces a fluctuating current in the receiver coil once the RF pulse is removed. The FID is composed of the signals of multiple protons from different chemical groups.

Consider a single FID component arising from the decay of the transverse magnetization of protons belonging to a specific chemical group (e.g. CH_3, CH_2, or OH) with a specific Larmor frequency. The net magnetization vector precesses in the xy-plane in either a clockwise (positive frequency: ν) or anti-clockwise (negative frequency: -ν). Assume that once the RF pulse is removed, the net magnetization vector is situated along the x-axis and is precessing in the clockwise direction as illustrated in Figure 9.9.

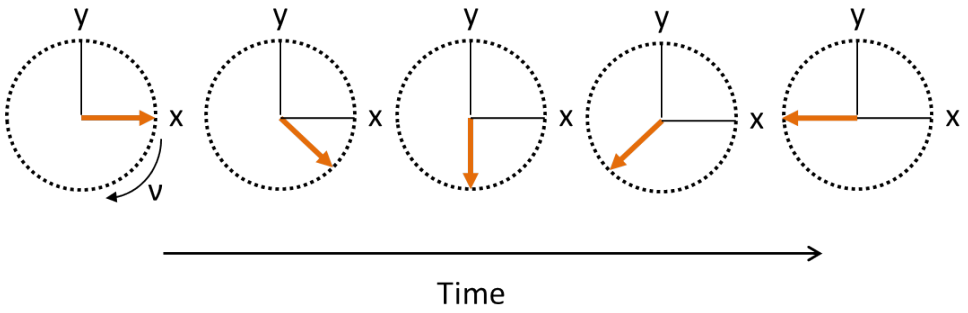

Figure 9.9: Precession of the transverse magnetization vector in the xy-plane.

In this situation, the x-component of magnetization is proportional to a $\cos(2\pi\nu t)$ wave and the y-component of magnetization is proportional to $\sin(2\pi\nu t)$. It is convenient to think of the detected signal as having two components, an x-component and a y-component that are defined as

$$s_x(t) = s_0 \cos(2\pi\nu t) \exp\left(-\frac{t}{T_2}\right), \tag{9.1}$$

$$s_y(t) = s_0 \sin(2\pi\nu t) \exp\left(-\frac{t}{T_2}\right), \tag{9.2}$$

where s_0 is the overall amplitude of the signal, ν is the Larmor frequency, t is the time, and T_2 is a time constant corresponding to the decay of the transverse magnetization.

As a consequence of the Fourier transformation, it is also convenient to consider $s_x(t)$ and $s_y(t)$ as the Real and Imaginary parts of the detected signal $s(t)$:

$$\begin{aligned} s(t) &= s_x(t) + i s_y(t) \\ &= s_0 \cos(2\pi\nu t) \exp\left(-\frac{t}{T_2}\right) + i s_0 \sin(2\pi\nu t) \exp\left(-\frac{t}{T_2}\right) \\ &= s_0 \exp(i2\pi\nu t) \exp\left(-\frac{t}{T_2}\right) \end{aligned} \tag{9.3}$$

The last equality results from Euler's formula: $\exp(ix) = \cos(x) + i\sin(x)$.

The FID is the sum of the signals arising from protons belonging to different chemical groups. Similar to the individual components $s(t)$, the FID is a function of time with Real and Imaginary components.

9.4 Understanding the parameters of an ^{1}H-NMR experiment

A typical ^{1}H-NMR experiment involves repeating a pulse sequence, consisting of three stages, several times in order to improve the signal-to-noise ratio of the acquired data. The pulse sequence begins with a preparation (or recycle) delay, which is followed by a period called the pulse width, and then by the acquisition period. The pulse sequence can be characterised by three parameters: (a) preparation delay, (b) pulse width, and (c) acquisition time. A brief description of each of these parameters and other parameters that describe an ^{1}H-NMR experiment is provided below.

Pulse width During the pulse width (duration measured in microseconds, μs), the RF pulse is applied causing the net magnetization vector to shift from the z-axis into the xy-plane.

Acquisition time The RF pulse is then switched off and the acquisition period (duration measured in seconds, s) begins. During the acquisition period, the transverse magnetization decays induces a current in the receiver coils which forms the NMR signal.

Preparation delay Although the preparation delay (duration measured in seconds, s) appears at the beginning of a pulse sequence, it can be thought of as occurring after the acquisition period. The preparation delay ensures that protons have sufficient time to return to their equilibrium state after the removal of the RF pulse.

Spectrometer frequency ^{1}H-NMR spectrometers are named after the frequency at which protons resonate in the magnetic field of the spectrometer. For instance, protons in a spectrometer with a magnetic field strength of 9.4 T will resonate at a frequency of 400 MHz. The spectrometer frequency corresponds to the center of the acquired ^{1}H-NMR spectrum.

Spectral width The spectral width corresponds to the range of the frequency domain spectrum. The spectral width measured in ppm is independent of the spectrometer frequency. However, the spectral width measured in Hz depends on the spectrometer frequency. A spectral width of 10 ppm is usually necessary for ^{1}H-NMR spectroscopy.

Thus, for a spectrometer of 400 MHz, 1 ppm corresponds to 400 Hz. Thus, a spectral width of at least 10 ppm $\times$ 400 Hz/ppm $=$ 4000 Hz would be required to measure a 10 ppm frequency range. Similarly, a spectral width of 9000 Hz would be required to measure a 10 ppm frequency range when using a 900 MHz spectrometer.

Number of data points recorded This is the number of data points used to record an FID during the acquisition time.

9.5 Pre-processing of ^{1}H-NMR data

Before the ^{1}H-NMR signal is quantified, the FID is pre-processed. Pre-processing the FID transforms the signal from the time domain to the frequency domain, and attempts to remove noise and artefacts that may otherwise interfere with subsequent statistical analyses. Several steps are involved in pre-processing an FID. A brief outline of the steps involved in pre-processing the ^{1}H-NMR data analysed in this dissertation, is provided in this section. The content of this section is predominantly based on Rousseau (2011), and the reader is referred to this text for further details on the pre-processing of ^{1}H-NMR spectra. Two pre-processing protocols were employed on the data analysed in this dissertation, a manual pre-processing protocol and an automated pre-processing protocol based on the R statistical software package `PepsNMR`.

The FID The FID is the sum of the ^{1}H-NMR signals that arise from resonating protons of different chemical groups. These individual FID components have the form defined by (9.3), and have real and imaginary terms defined by $s_x(t)$ (9.1) and $is_y(t)$ (9.2), respectively. The FID, denoted by $S(t)$, is a complex decaying function of time, also comprised of a real and an imaginary term:

$$S(t) = S_x(t) + iS_y(t), \tag{9.4}$$

where, $S_x(t)$ is the real term and $S_y(t)$ is the imaginary term.

Phase correction The phase of a spectrum depends on the position (or intensity) of the corresponding FID at time zero (i.e., $S(0)$). A perfectly phased FID has a real component, $S_x(t)$, that achieves its maximum value at time zero, and an imaginary component, $S_y(t)$, that is equal to zero at time zero. The Fourier transform of a phased FID corresponds to a real spectrum consisting of only positive intensities and an imaginary spectrum that has both positive and negative intensities.

In practice, FID's are not perfectly phased and instead exhibit a phase shift. Phase shifts can be split into two categories: those that are the same for all frequencies (i.e., frequency independent) called zero-order phase shifts, and those that are

frequency dependent called first-order phase shifts. Zero-order phase shifts may arise from the inability of a spectrometer to detect the correct x and y components of the transverse magnetization. First-order phase errors could arise due to a delay between the cessation of the RF pulse and the start of the acquisition period. Correcting the phase involves identifying the magnitude of the phase shift and correcting for it.

In manual phase correction, a spectroscopist visually identifies the magnitude of the phase shift. The FID is first Fourier-transformed to obtain a spectrum in the frequency domain. Then, a large peak is selected at one end of the spectrum and the phase of the selected peak is optimised. This corresponds to a zero-order (frequency independent) phase correction. A peak is then selected on the opposite end of the spectrum and its phase is optimised corresponding to a first-order phase correction. After phase correction the spectrum is back-transformed using an inverse-Fourier transformation.

In `PepsNMR`, the phase correction is conducted in two non-consecutive steps beginning with the first-order (frequency-dependent) phase correction. If τ is the delay between the end of the RF pulse and the start of the acquisition period, correcting the first-order phase correction corresponds to multiplying the spectrum by $\exp(-i2\pi\nu\tau)$. This is a frequency-dependent linear phase shift. The FID is Fourier-transformed, then multiplied by the linear phase shift before being back-transformed by use of the inverse-Fourier transform.

The `PepsNMR` zero-order (frequency-independent) phase correction is conducted in an automated way. A range of phase corrections between -180° and 180° are tested and the angle which achieves the largest positive real part of the spectrum is selected. The positiveness can be measured as the ratio of the sum of squares of the positive intensities and the total sum of squares of all intensities forming the spectrum.

Solvent suppression A solvent is a non-informative but major component of the solution analysed. It appears as a high intensity peak in the spectrum and has the potential to mask informative peaks of compounds that are of interest. Water is the solvent in metabolomics studies. The water molecules produce an intense peak in the spectrum that, if not suppressed, masks the signals from metabolites that are of interest.

During the pre-saturation phase of an ^{1}H-NMR experiment, the water signal is reduced. However, a residual water signal remains and the suppression of this residual signal is one of the steps involved in the pre-processing of an FID. Essentially, the residual water signal is suppressed by modeling the component of the FID that corresponds to the water molecules and subtracting it from the original FID. Since water is the main component of the original FID, the FID component of water is modelled as a smoothed function of the FID.

Apodization During the acquisition period of an ^{1}H-NMR experiment, noise arising from, for instance, the electronics in the spectrometer or from the receiver coil, is recorded together with the ^{1}H-NMR signal. Although the ^{1}H-NMR signal decays over time, the noise recorded remains constant. Thus, the signal-to-noise ratio decreases over time and is lowest in the tail of the FID. The goal of apodization is to increase the signal-to-noise ratio by multiplying the FID by a function that emphasizes the initial portion of the FID and down weighs the latter noisy portion. This is achieved by multiplying the FID by a decaying (negative) exponential function of the form

$$h(t) = \exp\left(-\frac{t}{d}\right), \tag{9.5}$$

where d is the decay parameter. The decay parameter should be carefully selected. If d is too small, too much of the initial portion of the signal can be lost. After apodization, the FID will have a shorter T_2 time-constant, implying a faster decay which results in shorter, broader peaks.

Zero-filling The digital resolution of a spectrum is related to the distance (in Hz) between two data points in a spectrum. The smaller the distance, the better the resolution. In order to increase the digital resolution of a spectrum, one might consider increasing the number of data points by increasing the acquisition time. However, too large an increase of the acquisition time would result in a decrease in the signal-to-noise ratio.

Zero-filling is a way to increase the digital resolution (i.e., reduce the number of Hz per data point) in a spectrum without adding additional noise. It involves simply adding data points with zero intensities to the end of the FID. Zero-filling in the time domain corresponds to interpolation in the frequency domain and can therefore elucidate fine coupling that may not be visible at a low digital resolution.

Fourier transform The Fourier transform is used to convert the ^{1}H-NMR signal from the time domain (FID) to a spectrum in the frequency domain. The Fourier transform extracts the components of the FID and utilizes the information on the amplitude of the signal, Larmor frequency of the proton, and the transverse relaxation time (T_2) to convert the signal to peaks with a specific height, position, and width at half-height in the frequency domain.

For $S(t)$ of n_t complex data points, the discrete Fourier transform is used:

$$F(\nu_j) = \sum_{t=0}^{n_t-1} S(t) \exp\left(\frac{-i2\pi jt}{n_t}\right), \tag{9.6}$$

where $j = 0, \ldots, n_f - 1$, and $n_f = n_t$ is the number of data points forming the

spectrum $F(\nu)$ in the frequency domain.

Baseline correction Technically, in an ^{1}H-NMR spectrum, frequencies at which no signal is emitted by the solution should correspond to a zero intensity. However, even after phase correction, due to multiple sources of noise, baseline artefacts arise (i.e., the recorded intensity is not zero). A convenient way to correct the baseline is to estimate the baseline distortions and subtract them from the observed spectrum.

Manual baseline correction traditionally involves selecting several points along the baseline and interpolating a polynomial fit between the points.

`PepsNMR` utlizes asymmetric least squares with a roughness penalty to estimate the baseline distortion. The baseline estimator Z is found by minimizing Q defined as follows:

$$Q = \sum_{j=0}^{n_f-1} \omega_j \left(F_j - Z_j\right)^2 + \lambda \sum_{j=2}^{n_f-1} \left[(Z_j - Z_{j-1}) - (Z_{j-1} - Z_{j-2})\right], \tag{9.7}$$

where F_j is the intensity at frequency ν_j in the spectrum, and Z_j is the intensity at the frequency ν_j in the estimated baseline. For $p \in [0, 1]$ and typically close to zero, the weight $\omega_j = p$ when $F_j > Z_j$ and $\omega_j = 1 - p$ when $F_j \leq Z_j$. This asymmetric weighting puts less weight on positive deviations of the spectrum from the estimated baseline, and therefore favours a positive corrected spectrum. Finally, λ is the roughness penalty with larger values of λ corresponding to a smoother estimated baseline.

Chemical shift conversion The frequency at which a proton resonates depends on its chemical environment, as well as on the external magnetic field strength. When a hydrogen atom is placed in an external magnetic field, neighbouring electrons shield the nucleus and prevent it from experiencing the full extent of the magnetic field. As a result, protons within different chemical environments experience different field strengths and, therefore, have different resonance frequencies. Furthermore, two protons belonging to the same chemical group (i.e., having the same chemical environment) can have different resonance frequencies depending on the magnetic field strength of the spectrometer used.

The aim of this pre-processing step is to translate the frequency scale of the ^{1}H-NMR spectra from Hertz to a scale that is independent of the magnetic field strength. Such a scale is the chemical shift scale and is measured in parts per million (ppm). The conversion uses a reference compound that does not change substantially with changes in the external magnetic field strength. In this dissertation, the compound trimethylsilyl-2,2,3,3-tetradeuteropropionic acid (TSP) is used as a chemical shift reference compound and is conventionally situated at 0 ppm. The chemical shift of a

data point j is calculated as the difference between the resonance frequency ν_j (in Hz) of the proton considered and resonance frequency ν_{TSP} (in Hz) of the reference compound (TSP), divided by the resonance frequency of the reference compound:

$$\text{ppm}_j = \frac{\nu_j - \nu_{\text{TSP}}}{\nu_{\text{TSP}}} \times 10^6. \tag{9.8}$$

Spectral alignment Due to experimental conditions such as differences in the pH, temperature, and concentrations of biological samples, peaks corresponding to the same compound may appear at slightly different chemical shift locations across different spectra. `PepsNMR` corrects this misalignment through the use of a warping algorithm that aligns the spectra to a reference spectrum. The choice of reference spectrum can be based on the square distances of a spectrum to all other spectra either before or after warping.

Window selection The entire spectral width does not necessarily contain peaks. This step trims the spectral window to the region of interest, leaving the spectrum with fewer data points. The domain is also reversed so that the spectrum is read from 0 ppm on the right to, for instance, 10 ppm on the left.

Normalization Normalization is necessary to remove the variation in the concentration of the samples. Integral or constant sum normalization is most commonly used and involves dividing each spectral intensity by the total area, or mean intensity, of the spectrum. Median normalization amongst others is also an option.

9.6 Peaks of an ^{1}H-NMR spectrum

The Fourier transformation of an FID (time domain signal) results in an ^{1}H-NMR spectrum (frequency domain signal) that is visualized as a series of peaks along a chemical shift axis. The peaks have the form of a Lorentzian curve. In mathematical notation, the form of the zero-centered Lorentzian function is

$$l_\gamma(\delta) = \frac{2}{\pi} \frac{\gamma}{4\delta^2 + \gamma^2}, \tag{9.9}$$

where γ is the peak width at half height. Each metabolite in the biological sample produces a characteristic signature in the ^{1}H-NMR spectrum that is formed by a combination of peaks not necessarily adjacent to each other along the chemical shift axis. Each signature appears with an intensity proportional to the concentration of the corresponding metabolite in the sample. The peaks of an ^{1}H-NMR spectrum are characterised by their chemical shift, signal intensity, and J-coupling patterns.

Chemical shift Each peak in an ^{1}H-NMR spectrum corresponds to the resonance of a proton belonging to a specific chemical group. As described in Section 9.5, the chemical shift (measured in ppm) expresses a proton's resonance relative to a reference compound in a magnetic field. Most proton resonance signals are situated between 0 ppm and 12 ppm.

Signal intensity The area under the peak at a specific chemical shift is proportional to the concentration of the proton in the sample that produced the peak.

J-coupling Nuclei with different chemical environments, within the same molecule may interact with each other producing a phenomenon known as J-coupling. The interaction of a specific proton with other protons of a different chemical group, arising from the same molecule, may lead to the proton's signal appearing as a multiplicity of peaks in the ^{1}H-NMR spectrum. The pattern with which the signal appears in the ^{1}H-NMR spectrum is called a J-coupling pattern and can correspond to a singlet, doublet, triplet etc., as illustrated in Figure 9.10. The J-coupling constant is the distance between two adjacent peaks of a split signal.

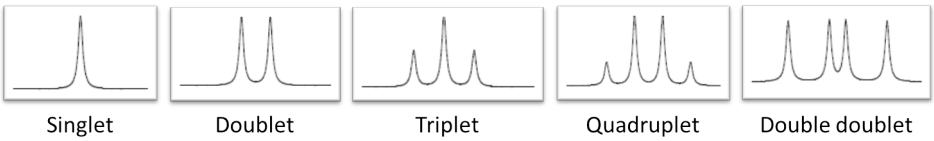

Figure 9.10: Examples of J-coupling patterns.

10

Data and pre-processing

10.1 Data

To investigate the impact of the method of extracting metabolic signal from ^{1}H-NMR data on the classification of samples, the previously analyzed (Louis et al., 2017), ^{1}H-NMR spectra of blood plasma samples obtained from lung cancer patients (n_{cases}= 69), included in the Limburg Positron Emission Tomography center (Hasselt, Belgium) from March 2011 to January 2012, and control subjects (n_{controls}= 74), attending Ziekenhuis Oost-Limburg (Genk, Belgium) between December 2011 and April 2012, were used. The ^{1}H-NMR data were acquired by analyzing the blood plasma samples at 21.2°C on a 400 MHz spectrometer (9.4 Tesla; 54 mm bore (i.e., the hollow center of the magnet)-size; Varian Inova; Agilent Technologies Inc.; VnmrJ 3.2 RevisionA) and on a 900 MHz spectrometer (21.1 Tesla; 54 mm bore-size; Bruker Avance; Bruker Biospin). The 400 MHz spectrometer was equipped with an Agilent OneNMR 5mm probe, whereas the 900 MHz spectrometer had a triple resonance cryoprobe. Slightly T_2-weighted spectra (i.e., spectra that have been weighted based on the transversal relaxation times of molecules, see Section 9.2.1) were acquired using the Carr-Purcell-Meiboom-Gill pulse sequence (total spin-echo time of 32 ms; interpulse delay of 0.1 ms), preceded by an initial preparation delay of 0.5 s, and 3 s for water suppression presaturation. Other parameters for acquiring the 400 MHz/900 MHz data, respectively, were: a spectral width of 6000 Hz/14423 Hz, a 90° pulse length of 6.35/9.15 μs, an acquisition time of 1.2 s, a preparation delay of 3.5 s, and 96/64 scans (7min 44sec/5min 9sec on 400 MHz/900 MHz).

10.1.1 Spiking experiments

The chemical shift of metabolite peaks depend on the cell, tissue or biofluid under study, as well as the experimental conditions (e.g., the temperature, pH, and sample concentration). Spiking experiments are routinely conducted to accurately determine the chemical shift values of the metabolites of interest for a specific experiment. In a metabolite spiking experiment, a relevant concentration of a known metabolite is added to reference plasma and the chemical composition of the sample is measured

through ^{1}H-NMR spectroscopy. The characteristic spectral signature (i.e., the peak locations, peak shape, and multiplicity) of the spiked metabolite is clearly visible in the resulting spectrum due to its increased relative concentration in the sample (Louis et al., 2015).

Spiking experiments (Louis et al., 2015) were conducted to determine the chemical shift positions of the blood plasma metabolites. Spiked spectra were acquired on the 400 MHz and 900 MHz spectrometers for 37 metabolites: alanine, arginine, asparagine, aspartate, cysteine, glutamine, glutamate, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, valine, glucose, myo-inositol, acetate, acetoacetate, α-ketoglutarate, β-hydroxybutyrate, citrate, lactate, pyruvate, succinate, creatine, creatinine, acetone, betaine, choline, glycerol, and methanol.

10.2 Spectral pre-processing

A manual pre-processing protocol was applied to the 400 MHz and the 900 MHz spectra. The 900 MHz spectra were also pre-processed using a more automated protocol.

10.2.1 Manual pre-processing

The 400 MHz spectra were pre-processed using the Varian/Agilent software. The pre-processing steps included zero-filling and multiplication by an exponential apodization function of 0.7 Hz prior to the Fourier transformation. The spectra were manually phased, automatically baseline corrected using polynomials (or splines), and referenced to trimethylsilyl-2,2,3,3-tetradeuteropropionic acid (TSP) at 0.015 ppm (Louis et al., 2016b). The final step of the spectral pre-processing was normalization by the total area under the curve, without accounting for the water and TSP signal.

The 900 MHz Bruker files were first transformed to the Varian format for compatibility with the Varian pre-processing software before being manually pre-processed in the same way as the 400 MHz data.

10.2.2 Automated pre-processing

The 900 MHz spectra were also automatically pre-processed using the R statistical software package `PepsNMR` (Rousseau, 2011). `PepsNMR` was applied to the raw Bruker FIDs. Pre-processing included a first-order and zero-order phase correction, solvent (i.e., water) suppression, apodization, zero-filling, Fourier transformation, baseline correction, spectral alignment, and median normalization (see Figure 10.1).

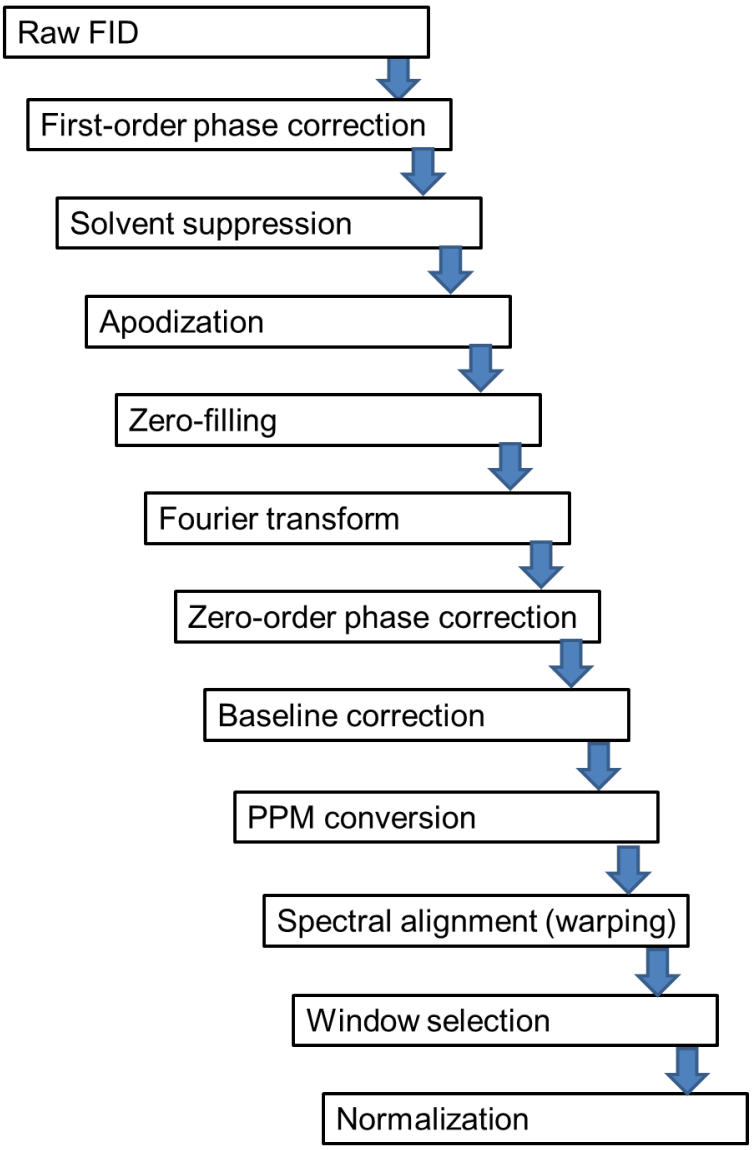

Figure 10.1: The sequence of PepsNMR pre-processing steps.

The default PepsNMR settings were utilized for all steps prior to the baseline correction. A more stringent penalty (on negative intensities) was selected for the baseline correction in order to keep the number of spectral points with negative intensities to a minimum. That is, the PepsNMR baseline correction asymmetry parameter was set to 0.01 (see Figure 10.2).

The reference spectrum chosen for spectral alignment was the spectrum that

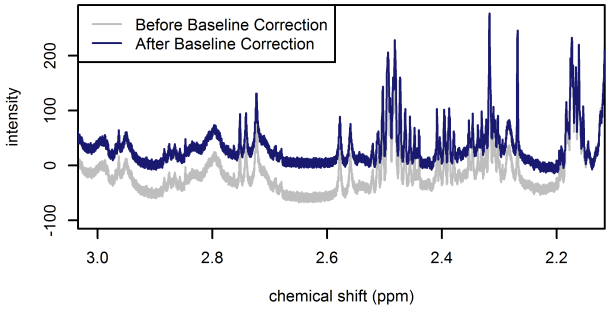

Figure 10.2: Illustration of a portion of a 900 MHz spectrum before (grey spectrum) and after (blue spectrum) baseline correction.

achieved the smallest sum of squared differences between itself and all other spectra after warping. This corresponds to setting the reference choosing parameter of the `PepsNMR` warping function to *after* (see Figure 10.3). Since we expect differences in the metabolic profile between lung cancer patients and control subjects, warping was performed separately for the two groups.

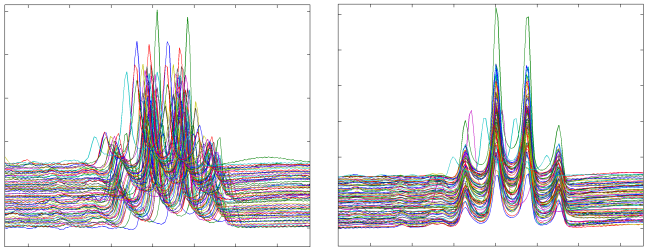

Figure 10.3: Illustration of warping in the region of the lactate signal. Left: a portion of a 900 MHz spectrum before warping. Right: a portion of a 900 MHz spectrum after warping.

The main reason for applying two different pre-processing protocols to the 900 MHz spectra was that the manual pre-processing of the 900 MHz spectra did not provide data of sufficient quality to perform the BATMAN analysis (see Figure 10.4). It was necessary to use the raw FID data to improve the manual baseline correction and to avoid numerous manual steps in phasing. With `PepsNMR`, the pre-processing steps and parameter settings can be clearly defined which improves the reproducibility of the analysis.

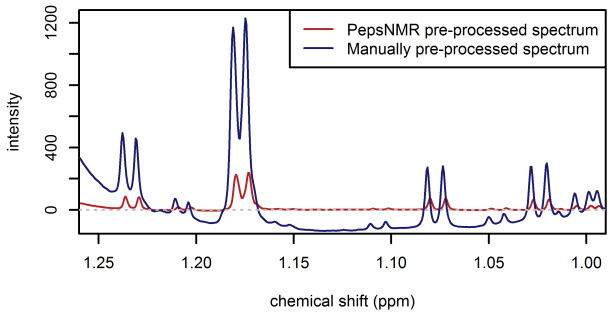

Figure 10.4: Illustration of a portion of a 900 MHz `PepsNMR` automatically pre-processed spectrum (red) and a 900 MHz manually pre-processed spectrum (blue).

11

Spectral binning and BATMAN for extracting metabolic signal from ^{1}H-NMR spectra

Spectral binning and spectral deconvolution using BATMAN were employed to quantify the signal in the 400 MHz and 900 MHz ^{1}H-NMR spectra of lung cancer patients and controls. See Chapter 10 for a description of the data. In this chapter, a description of these two techniques and details of their implementation is provided. In particular, Section 11.1 introduces spectral binning, and Section 11.2 focuses on BATMAN.

11.1 Spectral binning

Spectral binning (Louis et al., 2015) is a simple and commonly used technique for extracting metabolic signal from NMR spectra. Spectral binning involves partitioning the ^{1}H-NMR spectra into regions along the chemical shift axis. The regions can either be fixed- or variable-sized. The disadvantage of fixed binning is that peaks are often split across adjacent bins. In the lung-cancer study, variable-sized regions were selected. In that case, the limits of the integration regions are defined by spectroscopists in a way that best accommodates the metabolite peaks of interest. The resonance peaks encompassed by each region are integrated. The resulting integrated spectral regions (ISRs) constitute a set of features that represent the NMR signal. Reliable information on the chemical shift of metabolite peaks is essential for the identification of biologically meaningful spectral regions.

Using the chemical shift information acquired through spiking experiments, the 400 MHz spectra were subdivided into 110 ISRs (Louis et al., 2015) of varying widths excluding the water region and the TSP region (see Table 11.1). Similarly, the manually pre-processed 900 MHz spectra were partitioned into 105 ISRs (Louis et al., 2017) and the `PepsNMR` automatically pre-processed 900 MHz spectra were partitioned into 103 ISRs (see Table 11.1).

Table 11.1: Spectral binning regions for the 400 MHz and 900 MHz spectra.
ABBREVIATIONS: NI: not identified, FAC: fatty acid chain, NAG: N-acetylated glycoproteins, PC: phosphatidylcholine, PL: phospholipids, SM: sphingomyelins, TG: triglycerides, Ala: Alanine, Arg: Arginine, Asn: Asparagine, Asp: Aspartate, Cys: Cysteine, Gln: Glutamine, Glu: Glutamate, Gly: Glycine, His: Histidine, Ile: Isoleucine, Leu: Leucine, Lys: Lysine, Met: Methionine, Phe: Phenylalanine, Pro: Proline, Ser: Serine, Thr: Threonine, Trp: Tryptophan, Tyr: Tyrosine, Val: Valine.

400 MHz (manually pre-processed spectra)				900 MHz (manually pre-processed spectra)				900 MHz (PepsNMR pre-processed spectra)			
Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End
				1	NI	8.4914	8.4796	1	NI	8.4980	8.4750
				2	Formate	8.3702	8.3602	2	Formate	8.3810	8.3550
				3	NI	8.2601	8.2500	3	NI	8.2670	8.2480
				4	NI	8.2300	8.2050	4	NI	8.2340	8.2000
1	NI	7.9500	7.8200	5	NI	7.8561	7.8104	5	NI	7.8600	7.8200
2	His	7.8200	7.7890								
3	NI	7.7890	7.7780	6	His	7.7812	7.7544	6	His	7.8000	7.7644
4	His	7.7780	7.7480								
5	NI	7.7480	7.7200								
6	NI	7.6800	7.5920								
7	NI	7.5920	7.5480								
8	Phe	7.4840	7.3620	7	Phe	7.4677	7.4380	7	Phe	7.4750	7.4380
				8	Phe, NI	7.4162	7.3755	8	Phe, NI	7.4210	7.3755
9	Phe, NI	7.3620	7.3300	9	Phe	7.3675	7.3484	9	Phe	7.3750	7.3510
10	NI	7.3300	7.2820	10	NI	7.3484	7.3227	10	NI	7.3510	7.3227
11	NI	7.2820	7.2550								
12	Tyr, NI	7.2550	7.2390								
13	Tyr, NI	7.2390	7.2000	11	Tyr	7.2327	7.2046	11	Tyr	7.2400	7.2046
				12	NI	7.1894	7.1591	12	NI	7.1894	7.1591
14	His	7.1070	7.0656	13	His	7.0792	7.0597	13	His	7.0880	7.0600
				14	NI	7.0201	6.9652	14	NI	7.0201	6.9600
15	Tyr	6.9430	6.9050	15	Tyr	6.9355	6.9056	15	Tyr	6.9440	6.9056

400 MHz (manually pre-processed spectra)				900 MHz (manually pre-processed spectra)				900 MHz (PepsNMR pre-processed spectra)			
Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End
16	NI	6.9050	6.8810								
17	NI	6.7445	6.7020	16	NI	6.7460	6.7004	16	NI	6.7600	6.7004
18	Lipids:–C**H**=C**H**–in FAC	5.4300	5.2752	17	Lipids:–C**H**=C**H**–in FAC	5.4422	5.2833	17	Lipids:–C**H**=C**H**–in FAC	5.4422	5.2900
19	Glucose	5.2752	5.2526	18	Glucose	5.2751	5.2542	18	Glucose	5.2770	5.2560
20	C_2H in glycerol backbone of PL and TG	5.2526	5.2030	19	C_2H in glycerol backbone	5.2542	5.2301	19	C_2H glycerol backbone	5.2560	5.2301
				20	of PL and TG	5.2186	5.2038	20	of PL and TG	5.2260	5.2038
				21	NI	5.1525	5.1187	21	NI	5.1550	5.1187
21	Glucose	4.6940	4.6620	22	Glucose	4.7088	4.6421	22	Glucose	4.7088	4.6421
22	NI	4.5560	4.5380								
23	NI	4.5380	4.4100								
24	C_1H and C_3H in glycerol backbone of TG	4.4100	4.3159	23	C_1H and C_3H in glycerol backbone of TG	4.3579	4.2902	23	C_1H and C_3H in glycerol backbone of TG	4.3579	4.2940
25	O–C**H**$_2$–CH$_2$–N$^+$(CH$_3$)$_3$ of PC and SM, Thr	4.3159	4.2332	24	O–C**H**$_2$–CH$_2$–N$^+$(CH$_3$)$_3$ of PC and SM, Thr	4.2852	4.2536	24	O–C**H**$_2$–CH$_2$–N$^+$(CH$_3$)$_3$ of PC and SM, Thr	4.2940	4.2500
26	β-hydroxybutyrate, Pro	4.2000	4.1885								
27	β-hydroxybutyrate, Pro, Lactate	4.1885	4.1750	25	β-hydroxybutyrate	4.2000	4.1607	25	β-hydroxybutyrate	4.2000	4.1607
28	C_1H and C_3H glycerol backbone of PL and TG, lactate	4.1750	4.1260	26	C_1H and C_3H glycerol backbone of PL and TG, lactate	4.1570	4.1276	26	C_1H and C_3H glycerol backbone of PL and TG, lactate	4.1600	4.1276
29	NI	4.1260	4.1110	27	NI	4.1276	4.0942	27	NI	4.1276	4.0950
30	NI	4.1110	4.1032								
31	Creatinine	4.1032	4.0700	28	Creatinine	4.0904	4.0780	28	Creatinine	4.0950	4.0770
32	NI	4.0700	4.0570								

400 MHz (manually pre-processed spectra)				900 MHz (manually pre-processed spectra)				900 MHz (PepsNMR pre-processed spectra)			
Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End
33	His, Ser	4.0570	4.0310								
34	Asn, His, Phe, Ser	4.0310	4.0136	29	C_3H_2 in glycerol backbone of PL, Asn, His, Phe, Ser	4.0400	3.9913	29	C_3H_2 in glycerol backbone of PL, Asn, His, Phe, Ser	4.0420	3.9920
35	C_3H_2 in glycerol backbone of PL, Asn, His, Phe, Ser	4.0136	4.0010								
36	C_3H_2 in glycerol backbone of PL, Asn, His, Phe, Ser	4.0010	3.9810								
37	Creatine, Asn, His, Tyr, Ser	3.9810	3.9590	30	Asn, His, Ser, Tyr	3.9903	3.9644	30	Asn, His, Ser, Tyr	3.9920	3.9680
				31	Creatine	3.9644	3.9586	31	Creatine	3.9680	3.9600
				32	Tyr	3.9586	3.9527	32	Tyr	3.9600	3.9527
38	Glucose, Asp, Met, Ser, Tyr	3.9590	3.8330	33	Glucose	3.9527	3.9120	33	Glucose	3.9527	3.9150
				34	Glucose	3.9120	3.8957	34	Glucose	3.9150	3.8920
				35	Glucose	3.8881	3.8306	35	Glucose	3.8920	3.8410
39	Glucose, Ala, Ser	3.8330	3.8100	36	Glucose, Ala, Gln, Glu, Ser	3.8286	3.8097	36	Glucose, Ala, Gln, Glu, Ser	3.8410	3.8140
40	Glucose, Ala, Gln, Glu	3.8100	3.7956								
41	Glucose, Ala, Gln, Glu, Leu, Lys	3.7956	3.7820	37	Glucose, Ala, Gln	3.8097	3.7794	37	Glucose, Ala, Gln	3.8140	3.7794
42	Glucose, Ala, Gln, Glu, Leu, Lys	3.7820	3.7550	38	Glucose	3.7776	3.7275	38	Glucose	3.7776	3.7275
43	Glucose, Ala, Leu	3.7550	3.7390								
44	Glucose	3.7390	3.7141								
45	O–C**H**$_2$–CH$_2$–N$^+$(CH$_3$)$_3$ of PC and SM, glycerol, Ile	3.7141	3.6680	39	Glycerol	3.7204	3.6453	39	Glycerol	3.7240	3.6500

400 MHz (manually pre-processed spectra)				900 MHz (manually pre-processed spectra)				900 MHz (PepsNMR pre-processed spectra)			
Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End
46	Glycerol	3.6680	3.6500								
47	Glycerol, Val	3.6500	3.6376								
48	Val	3.6376	3.6240	40	Val	3.6453	3.6212	40	Val	3.6500	3.6250
49	Thr	3.6240	3.6097	41	Thr	3.6163	3.5861	41	Thr	3.6163	3.5930
50	Thr	3.6097	3.5914								
51	Glucose, glycerol, Gly, Thr	3.5914	3.5649	42	Glycerol	3.5861	3.5771	42	Glycerol	3.5930	3.5810
52	Glucose	3.5649	3.5510	43	Glucose	3.5771	3.5481	43	Glucose	3.5810	3.5481
53	Glucose, acetoacetate, Pro	3.5510	3.5360								
53, 54	Glucose, acetoacetate, Pro	3.5360	3.3980	44	Glucose	3.5355	3.4798	44	Glucose	3.5481	3.4798
				45	Pro	3.4772	3.4576	45	Pro	3.4798	3.4600
				46	Glucose	3.4576	3.4093	46	Glucose	3.4600	3.4093
55	Methanol, NI	3.3980	3.3765	47	Methanol	3.3964	3.3924	47	Methanol	3.4004	3.3924
				48	NI	3.3924	3.3746	48	NI	3.3924	3.3770
56	Pro	3.3765	3.3430	49	Pro	3.3746	3.3465	49	Pro	3.3770	3.3465
57	Phe, Pro	3.3430	3.3230								
58	O–C**H**$_2$–CH$_2$–N$^+$(CH$_3$)$_3$ of PC and SM, glucose, His, Phe, Tyr	3.3230	3.2186	50	Phe	3.3256	3.3132	50	Phe	3.3300	3.3160
				51	Phe, NI	3.3132	3.3030	51	Phe, NI	3.3160	3.3050
				52	NI	3.3030	3.2956	52	NI	3.3050	3.2950
				53	NI	3.2956	3.2909	53	NI	3.2950	3.2920
				54	Glucose	3.2909	3.2616	54	Glucose	3.2920	3.2646
				55	O–C**H**$_2$–CH$_2$–N$^+$(CH$_3$)$_3$ of PC and SM	3.2616	3.2085	55	O–C**H**$_2$–CH$_2$–N$^+$(CH$_3$)$_3$ of PC and SM	3.2640	3.2085
59	Tyr, NI	3.2186	3.1930	56	Tyr, NI	3.1972	3.1895	56	Tyr, NI	3.2000	3.1900
60	NI	3.1930	3.1760	57	NI	3.1881	3.1821	57	NI	3.1900	3.1800

400 MHz (manually pre-processed spectra)				900 MHz (manually pre-processed spectra)				900 MHz (PepsNMR pre-processed spectra)			
Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End
				58	NI	3.1821	3.1724	58	NI	3.1800	3.1724
61	NI	3.1760	3.1462	59	NI	3.1707	3.1571	59	NI	3.1724	3.1600
62	His, Phe	3.1462	3.1090	60	His, Phe	3.1541	3.1378	60	His, Phe	3.1600	3.1300
63	Lys, Tyr	3.1090	3.0860	61	Tyr	3.0921	3.0769	61	Tyr	3.0940	3.0785
64	Creatinine, Lys, Tyr	3.0860	3.0716	62	Creatinine	3.0769	3.0699	62	Creatinine	3.0785	3.0720
65	Creatinine, creatine, Lys	3.0716	3.0640	63	Creatine	3.0699	3.0635	63	Creatine	3.0720	3.0655
66	α-ketoglutarate, Lys	3.0640	2.9950	64	α-ketoglutrate, Lys	3.0635	3.0047	64	α-ketoglutrate, Lys	3.0655	3.0047
67	Lipids: =CH–C**H**$_2$–CH= in FAC, Asn	2.9950	2.8860	65	Lipids: =CH–C**H**$_2$–CH= in FAC	3.0047	2.9655	65	Lipids: =CH–C**H**$_2$–CH= in FAC	3.0047	2.9655
				66	Asn	2.9597	2.9201	66	Asn	2.9655	2.9201
68	Lipids: =CH–C**H**$_2$–CH= in FAC, Asn, Asp	2.8860	2.8550	67	Lipids: =CH–C**H**$_2$–CH= in FAC, Asn, Asp	2.8874	2.8465	67	Lipids: =CH–C**H**$_2$–CH= in FAC, Asn, Asp	2.8874	2.8450
69	Lipids: =CH–C**H**$_2$–CH= in FAC, Asn, Asp	2.8550	2.7500	68	Lipids: =CH–C**H**$_2$–CH= in FAC	2.8465	2.7623	68	Lipids: =CH–C**H**$_2$–CH= in FAC	2.8465	2.7623
70	Citrate, Asp	2.7500	2.7360	69	Citrate	2.7571	2.7493	69	Citrate	2.7530	2.7250
				70	NI	2.7472	2.7390				
71	Citrate, Asp, Met	2.7360	2.6600	71	Citrate	2.7368	2.7251				
				72	Asp	2.7237	2.6768	70	Asp	2.7250	2.6900
72	Met	2.6600	2.6300	73	Met	2.6768	2.6597	71	Met	2.6900	2.6597
73	Citrate	2.5960	2.5340	74	Citrate	2.5865	2.5426	72	Citrate	2.5865	2.5426
74	NI	2.5340	2.5150								
75	Gln	2.5150	2.4920	75	Gln	2.5183	2.4428	73	Gln	2.5250	2.4440
76	β-hydroxybutyrate, α-ketoglutarate, Gln	2.4920	2.4500								

400 MHz (manually pre-processed spectra)				900 MHz (manually pre-processed spectra)				900 MHz (`PepsNMR` pre-processed spectra)			
Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End
77	β-hydroxybutyrate, α-ketoglutarate, succinate	2.4500	2.4324	76	β-hydroxybutyrate	2.4428	2.4280	74	β-hydroxybutyrate	2.4440	2.4280
78	β-hydroxybutyrate, Pro	2.4324	2.4148								
79	β-hydroxybutyrate, Glu, Pro	2.4148	2.4050								
80	Pyruvate, Pro, Glu	2.4050	2.3990	77	Pyruvate	2.4060	2.3978	75	Pyruvate	2.4110	2.3980
81	β-hydroxybutyrate, Pro, Glu	2.3990	2.3640	78	Glu	2.3978	2.3648	76	Glu	2.4000	2.3600
82	β-hydroxybutyrate, Pro, Glu	2.3640	2.3500								
83	β-hydroxybutyrate, Pro, Val	2.3500	2.3380	79	β-hydroxybutyrate	2.3540	2.3194	77	β-hydroxybutyrate	2.3540	2.3194
84	β-hydroxybutyrate , acetoacetate, Pro, Val	2.3380	2.3170								
85	β-hydroxybutyrate, acetoacetate, Val	2.3170	2.3040	80	Acetoacetate	2.3134	2.3067	78	Acetoacetate	2.3194	2.3055
86	Lipids: $-C\mathbf{H}_2-C=O$ or $-C\mathbf{H}_2-CH=CH-$in FAC, Val, β-hydroxybutyrate	2.3040	2.2915	81	Lipids: $-C\mathbf{H}_2-C=O$ or $-C\mathbf{H}_2-CH=CH-$in FAC	2.3067	2.2630	79	Lipids: $-C\mathbf{H}_2-C=O$ or $-C\mathbf{H}_2-CH=CH-$in FAC	2.2990	2.2680
87	Lipids: $-C\mathbf{H}_2-C=O$ or $-C\mathbf{H}_2-CH=CH-$in FAC, Met, Val	2.2915	2.2690								
88	Lipids: $-C\mathbf{H}_2-C=O$ or $-C\mathbf{H}_2-CH=CH-$, acetone, Met, Val	2.2690	2.2300	82	Acetone	2.2630	2.2563	80	Acetone	2.2680	2.2563
89	Glu, Met	2.2180	2.1970								
90	Gln, Glu, Pro, Met	2.1970	2.1230	83	NI	2.1975	2.1814	81	NI	2.1975	2.1930

400 MHz (manually pre-processed spectra)				900 MHz (manually pre-processed spectra)				900 MHz (PepsNMR pre-processed spectra)			
Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End
				84	Gln	2.1777	2.1670	82	Gln	2.1930	2.1700
				85	Met	2.1670	2.1919	83	Met	2.1700	2.1650
				86	Gln	2.1619	2.1311	84	Gln	2.1650	2.1311
91	Lipids: –C**H**$_2$–CH=CH–in FAC, CH_3 of NAG, Glu, Ile, Met, Pro	2.1230	1.9720	87	Lipids: –C**H**$_2$–CH=CH–in FAC	2.1289	2.0993	85	Lipids: –C**H**$_2$–CH=CH–in FAC	2.1300	2.0975
				88	Lipids: –C**H**$_2$–CH=CH–in FAC, CH_3 of NAG	2.0993	1.9889	86	Lipids: –C**H**$_2$–CH=CH–in FAC, CH_3 of NAG	2.0985	1.9889
92	Acetate, Ile, Lys	1.9720	1.9240								
				89	Acetate	1.9547	1.9421	87	Acetate	1.9547	1.9450
92				90	Lys	1.9421	1.9028	88	Lys	1.9450	1.9100
93	Ile, Lys	1.9240	1.8800								
94	Leu, Lys	1.8060	1.6860	91	Leu	1.8006	1.6758	89	Leu	1.8006	1.6758
95	Lipids: –C**H**$_2$–CH_2–C=O or –C**H**$_2$–CH_2–CH=CH–in FAC, Lys	1.6860	1.5600	92	Lipids: –C**H**$_2$–CH_2–C=O or –C**H**$_2$–CH_2–CH=CH–in FAC	1.6530	1.5770	90	Lipids: –C**H**$_2$–CH_2–C=O or –C**H**$_2$–CH_2–CH=CH–in FAC	1.6530	1.5770
96	Ala, Ile, Lys	1.5400	1.4900	93	Ala	1.5226	1.4919	91	Ala	1.5226	1.4850
97	Ile, Leu, Lys	1.4900	1.4200	94	Lys	1.4587	1.4201	92	Lys	1.4587	1.4201
98	Lactate	1.4200	1.3740	95	Lactate	1.4169	1.3675	93	Lactate	1.4169	1.3730
99	Lactate, Thr	1.3740	1.3450	96	Lactate	1.3675	1.3516	94	Lactate	1.3730	1.3516
100	Lipids:–CH_3–(C**H**$_2$)$_n$–in FAC, Ile, Thr	1.3450	1.2458	97	Lipids: –CH_3–(C**H**$_2$)$_n$–in FAC	1.3516	1.2366	95	Lipids: –CH_3–(C**H**$_2$)$_n$–in FAC	1.3500	1.2500
101	β-hydroxybutyrate, Ile	1.2458	1.2180	98	β-hydroxybutyrate	1.2366	1.2240	96	β-hydroxybutyrate	1.2500	1.2250
102	NI	1.2180	1.1300	99	NI	1.2240	1.1766	97	NI	1.2200	1.1700
103	Val	1.0930	1.0610	100	Val	1.0860	1.0592	98	Val	1.0950	1.0620
104	Ile	1.0610	1.0400	101	Ile	1.0513	1.0340	99	Ile	1.0620	1.0370

400 MHz (manually pre-processed spectra)				900 MHz (manually pre-processed spectra)				900 MHz (`PepsNMR` pre-processed spectra)			
Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End	Region	Contributing metabolites	Start	End
105	Ile, Val	1.0400	1.0220	102	Val	1.0396	1.0106	100	Val	1.0370	1.0150
106	Ile, Leu, Val	1.0220	1.0020								
107	Ile, Leu	1.0020	0.9860	103	Leu	1.0083	0.9766	101	Leu	1.0150	0.9800
108	Ile, Leu	0.9860	0.9760								
109	Ile	0.9760	0.9660	104	Ile	0.9766	0.9663	102	Ile	0.9800	0.9550
110	Lipids: C$\mathbf{H}_3$–$(CH_2)_n$–in FAC	0.9660	0.8000	105	Lipids: C$\mathbf{H}_3$–$(CH_2)_n$–in FAC	0.9663	0.7961	103	Lipids: C$\mathbf{H}_3$–$(CH_2)_n$–in FAC	0.9663	0.7961

11.2 BATMAN

The Bayesian state-of-the-art spectral deconvolution technique, BATMAN, was developed by Astle et al. (2012). BATMAN resolves the resonance peaks of NMR spectra in order to estimate the relative concentrations of a pre-specified set of metabolites. BATMAN is a two-component model. The first component models the metabolic signal (i.e., the signal assigned to specific metabolites) while the second component models the residual signal. BATMAN exploits extensive prior information on the characteristic spectral signatures of each metabolite and combines this information with the observed intensities to model the metabolic signal. The second component uses wavelets to capture the residual signal. The residual signal includes the signal that arises from other uncatalogued chemical constituents such as lipids. When the metabolic signal has been properly extracted, the wavelet signal can be divided into carefully selected broad ISRs to approximate lipid concentrations. In this way, a set of relative metabolite concentrations and lipid features can be obtained. In addition to providing point estimates of the metabolic concentrations per spectrum, BATMAN also provides 95% credible intervals for each estimate which can be used to assess the degree of uncertainty in the estimated concentrations.

Astle et al. (2012) and Hao et al. (2014) describe BATMAN and its implementation in great detail. The next two sections provide an overview of the methodology as described in Astle et al. (2012) and Hao et al. (2014) with additional details pertaining to our implementation of BATMAN.

11.2.1 Specification of BATMAN

In this section, the model specification of BATMAN is provided using the notation of Astle et al. (2012) and Hao et al. (2014). BATMAN analyses the pre-processed NMR spectra (i.e., the frequency domain signals). Let y denote the $n_f \times 1$ dimensional vector of intensities. The NMR spectral intensities y are modeled as the sum of the catalogued metabolites y^c, the uncatalogued metabolites y^u, and the noise ϵ. In mathematical notation,

$$y = y^c + y^u + \epsilon, \quad \epsilon \sim N\left(0, \frac{I}{\lambda}\right), \tag{11.1}$$

where I is the $n_f \times n_f$ identity matrix, and λ is a scalar precision parameter.

Catalogued metabolites The catalogued metabolite component y^c is formed by taking a weighted sum of the metabolite signatures of each catalogued metabolite $m = 1, \ldots, M$. The signatures of each catalogued metabolite are specified by a template function denoted by t_m. In mathematical notation, the catalogued metabolite

component is modeled as

$$y^c = \sum_{m=1}^{M} \beta_m t_m(\delta), \tag{11.2}$$

where δ is the chemical shift parameter. The coefficient β_m is of main interest and takes on a value that is proportional to the concentration of metabolite m in the sample.

The NMR signature for a particular metabolite m, which is modeled by the template function t_m, is expressed as a linear combination of U_m multiplet functions. The mathematical form of the template function is given by

$$t_m(\delta) = \sum_{u=1}^{U_m} z_{mu} g_{mu}(\delta - \delta^\star_{mu}). \tag{11.3}$$

In equation (11.3), z_{mu} corresponds to the number of protons in the molecule of m that contribute resonance signal to the multiplet u. The multiplet curve is modeled by g_{mu} where $\delta^\star_{mu}$ is the chemical shift at the center of mass of the u^{th} multiplet of the m^{th} catalogued metabolite. It is assumed that $\int_0^\infty g_{mu}(\delta)d\delta = \int_{-\infty}^0 g_{mu}(\delta)d\delta$ for all m and u. Furthermore, $\int_{-\infty}^\infty g_{mu}(\delta)d\delta$ is constant over m and u. Thus, the area under t_m is proportional to the number of protons resonating in the molecule. The multiplet curve g_{mu} is the sum of horizontally shifted and vertically scaled Lorentzian functions $l(\delta)$, as defined in (9.9), with peak width at half height denoted by γ_m. Formally,

$$g_{mu}(\delta) = \sum_{v}^{V_{mu}} \omega_{muv} l_{\gamma_m}(\delta - c_{muv}), \tag{11.4}$$

where ω_{muv} represents the relative intensities (heights) of the peaks of the multiplet, and c_{muv} determines the horizontal offset of the peaks from the center of mass of the multiplet $\delta^\star_{mu}$. The peak widths γ_m are allowed to vary between metabolites according to

$$\ln(\gamma_m) = \mu + \nu_m, \tag{11.5}$$

where μ is the average peak width across the spectrum and ν_m is the deviation from the average for metabolite m which is assumed to be normally distributed.

Uncatalogued signal The uncatalogued component y^u is modeled as a linear combination of wavelet basis functions with θ denoting the vector of wavelet coefficients. In particular, Daubechie's least asymmetric wavelets with six vanishing moments are used as a wavelet basis. These wavelets have a similar shape to Lorentzian peaks.

Prior specification The prior specification of the parameters is provided below. The numerical values of the constants, used to define the priors, are discussed in Section 11.2.2.

μ **and** ν_m Gaussian priors are assumed for the average peak width (μ) across the spectrum, as well as for the deviation from the average peak width (ν_m) for a metabolite m.

c_{muv} **and** ω_{muv} The parameters c_{muv} (i.e., the horizontal offsets) and ω_{muv} (i.e., the relative intensities which sum to one) characterize the multiplet shapes and vary very slightly between spectra. As such, they are assumed to be constant. The values of c_{muv} and ω_{muv} are computed using estimates of the J-coupling constants.

$\delta^{\star}_{mu}$ Due to differences in experimental conditions, the center of mass of a multiplet, denoted by $\delta^{\star}_{mu}$, fluctuates slightly between spectra. Smaller fluctuations are more probable than larger ones. To account for this, a truncated Gaussian prior distribution is assigned to each $\delta^{\star}_{mu}$.

β_m A truncated Gaussian distribution confined to positive values with low prior information is assumed for the coefficients β_m.

θ **and** λ In order to distinguish between the parametric and wavelet components of the model, the wavelet component is penalized. A truncated Gaussian prior distribution with probability mass concentrated at zero is imposed on the wavelet coefficients θ. To ensure a mostly positive signal, a strong penalty is imposed on the parts of the wavelet component that lie below a small negative threshold h. The prior distribution on the wavelet coefficients are characterised by λ, and the hyperparameters ψ and τ. The hyperparameter vector $\psi \sim \text{Gamma}(c, d/2)$ allows the prior precision associated with each wavelet to deviate from the global precision $\lambda \sim \text{Gamma}(a, b/2)$ and induces shrinkage of the wavelet coefficients. Smaller values of a and b correspond to increased uncertainty in the value of λ. The values of c and d control the degree of penalization imposed on the wavelet coefficients. τ is a truncation limit vector associated with the spectral data points $i = 1, \ldots, n$. Each τ_i has a Gaussian distribution that is right-truncated at a small negative intensity h.

11.2.2 Implementation of BATMAN

BATMAN was implemented by using the R statistical software package `batman`, as detailed in the protocol by Hao et al. (2014). In this section, the implementation of the model is described and some of the steps that are crucial for improving the extraction of the metabolic signal are summarized.

The standard BATMAN inputs are the NMR spectroscopy data (NMRdata.txt), the parameter options file (batmanOptions.txt), the library of characteristic metabolic signatures (multi_data.csv or multi_data_user.csv), and a list of the metabolites of interest (metabolitesList.csv).

BATMAN options file The parameter settings (batmanOptions.txt) used for the 400 MHz and the 900 MHz analysis are shown in Table 11.2. The parameters used to fit the BATMAN model were selected based on the properties of the spectra and the recommendations of Hao et al. (2014). The truncation threshold for negative intensities was lower for the 400 MHz analysis compared to the 900 MHz analysis in order to accommodate the negative intensities (due to minor phasing issues) in some of the 400 MHz spectra. As per the recommendations of Hao et al. (2014), the parameters controlling the precision parameter λ (i.e., shape (a) and scale (b) in Table 11.2) were left at their default values. Peaks were allowed to shift more in the 400 MHz analysis as greater variation was observed in the location of the multiplets across the 400 MHz spectra.

Template file Prior information about the spectral signatures of each metabolite is specified in the default BATMAN template file multi_data.csv. The default template file can be modified by constructing the template file multi_data_user.csv. The fit of the BATMAN model can be improved considerably by providing prior information that more accurately describes the observed peaks. Each resonance is described in the BATMAN template file in terms of its chemical shift position (in ppm), multiplicity (i.e., the J-coupling pattern), J-coupling constants, and the relative intensities of the peaks. Multiplets with well-defined coupling patterns and known coupling constants, which exhibit second order effects (i.e., leaning effects), can be modeled empirically by specifying the observed intensity ratios of the peaks. Further details on this aspect are provided in the note on empirical multiplets below. Complex multiplets (e.g., multiplets that are not well-defined or those that exhibit higher-order coupling patterns), for which elucidation would require a substantial amount of input from a spectroscopist, may be modeled as raster multiplets by providing a corresponding section of a pure compound spectrum (see the note on raster multiplets below).

Given the complexity of NMR resonances, ill-defined chemical shift positions is the recipe for a poor fit. Prior information about the peak locations was determined

Table 11.2: Parameters used to run BATMAN.

BATMAN options file parameters	400 MHz spectra	900 MHz spectra
General parameters		
Truncation threshold for negative intensities (h)	-0.5	-0.05
Intensity scale factor	100	100
Down sampling factor	3	3
Number of burn-in iterations	3500	3500
Number of post-burn-in iterations	1500	1500
Spectrometer frequency (MHz)	399.793	900.229863
Uncatalogued (wavelet) component		
Shape (a)	0.00001	0.00001
Scale (b)	0.000000001	0.000000001
Catalogued metabolite component		
Mean of prior on global peak width (μ) in ln(Hz)	0	0
Variance of prior on global peak width (μ) in ln(Hz)	0.01	0.01
Variance of prior on peak width offset (ν_m) in ln(Hz)	0.0025	0.0025
Wavelet truncation		
Mean of the prior on τ	-0.05	-0.05
Inverse of variance of prior on τ	2	1
Peak shift		
Truncation of prior on peak shift (ppm)	0.01	0.005

by using the *splineFit* routine (Hao et al., 2014) implemented in Matlab, and the details on the ^{1}H-NMR chemical shift locations of plasma metabolites reported by Louis et al. (2015).

A note on empirical multiplets For the user-defined empirical multiplets, the accurate specification of relative intensities is subject to the availability of pure compound spectra (i.e., NMR spectra obtained by analyzing a sample containing only the target metabolite). For the multiplets in regions of no overlap, baseline-corrected spiked spectra were used as a substitute for the pure compound spectra. To compute the relative intensities, the number of resonating protons should be taken into account. For a particular multiplet, a simple numeric solution is to take the intensities of the peaks observed in the pure compound spectra and to normalize them to sum to the actual number of protons. That is, the relative intensity of a multiplet's i^{th} peak is computed using the following formula:

$$h_i = p \times \frac{y_i}{\sum_i y_i} \tag{11.6}$$

where y_i is the observed intensity based on pure compound spectra and p is the number of protons associated with the multiplet.

Empirical templates can be defined to model multiplets that exhibit leaning effects. If, for instance, two protons (with different chemical shifts) are coupled together, the signal of the one proton splits the signal of the other proton and vice-versa, resulting in a doublet for each proton. Sometimes, depending on the distance between the signals in the spectrum and the strength of the coupling, the patterns lean towards each other resulting in the outer peaks having a lower intensity than the inner peaks. Figure 11.1 shows the leaning effect of two doublets of citrate. Each doublet was produced by two resonating protons. Using equation (11.6), the relative intensity of the peak at 2.586 ppm is $h_1 = 1.2$ and the relative intensity of the peak at 2.547 ppm is $h_2 = 0.8$.

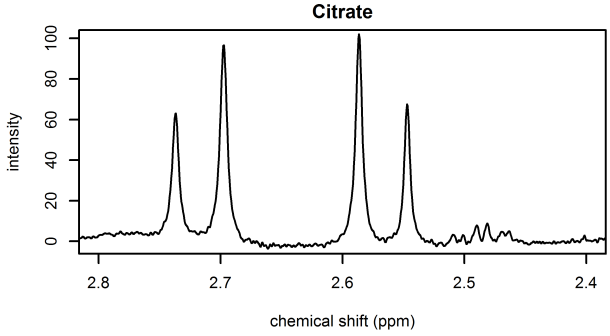

Figure 11.1: Illustration of a 400 MHz spectrum with two doublets of citrate at 2.717 and 2.566 ppm. Each doublet arises from a CH_2-group and thus from two protons.

In addition to the relative intensities, the offset of the peaks should be specified (in Hz). Offsets are specified from a point of origin. For convenience, the center of the multiplet can be taken as the origin. The offsets can be determined from pure compound spectra. Alternatively, the J-coupling information of ^{1}H-NMR plasma metabolites reported by Louis et al. (2015) and public databases like the Human Metabolome Database (HMDB) can be used. This is depicted in Figure 11.2, where the offset of the leftmost peak from the center of the double doublet of aspartate is half of the sum of the two J-coupling constants.

A note on raster multiplets Raster multiplets can be modeled from pure compound spectra. Due to the lack of pure compound spectra, spiked spectra were used for the

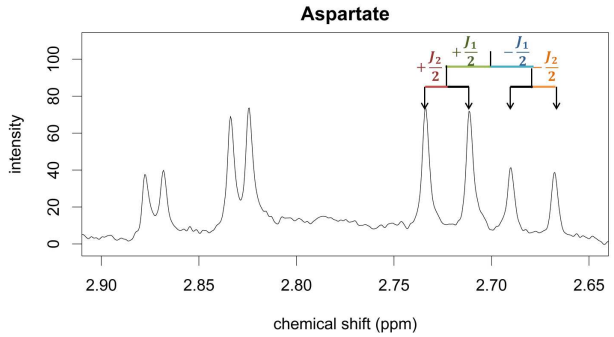

Figure 11.2: A portion of a 400 MHz spectrum illustrating the identification of peak offsets for the double doublet of aspartate at 2.702 ppm. Coupling constants J_1 and J_2 can be used to obtain the location of the four peaks from the center of the multiplet

multiplets that are found in regions where there is no significant overlap with other metabolites. Examples of raster multiplets for the 400 MHz and 900 MHz spectra are shown in Figure 11.3.

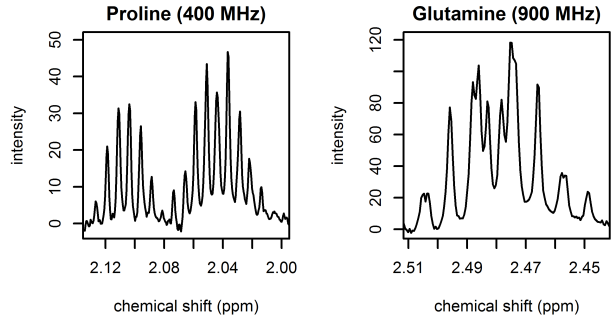

Figure 11.3: Illustration of raster multiplets. Left: 400 MHz raster multiplet for proline. Right: 900 MHz raster multiplet for glutamine.

Target metabolites The BATMAN model was applied to estimate the relative concentrations of the following metabolites (together with their Chemical Entities of Biological Interest (ChEBI) codes) in the 400 MHz spectra: alanine (CHEBI: 57972), arginine (CHEBI: 32682), asparagine (CHEBI: 58048), aspartate (CHEBI: 29991), cysteine (CHEBI: 35235), glutamine (CHEBI: 58359), glutamate (CHEBI: 29985), glycine (CHEBI: 57305), histidine (CHEBI: 57595), isoleucine (CHEBI: 58045), leucine (CHEBI: 57427), lysine (CHEBI: 32551), methionine (CHEBI: 57844), phenylalanine (CHEBI: 58095), proline (CHEBI: 60039), serine (CHEBI: 33384), threonine (CHEBI: 57926), tryptophan (CHEBI: 57912), tyrosine (CHEBI: 58315), valine

(CHEBI: 57762), α-D-glucopyranose (CHEBI: 17925), β-D-glucopyranose (CHEBI: 15903), myo-inositol (CHEBI: 17268), acetate (CHEBI: 30089), acetoacetate (CHEBI: 13705), α-ketoglutarate (CHEBI: 16810), β-hydroxybutyrate (CHEBI: 10983), citrate (CHEBI: 16947), lactate (CHEBI: 16651), pyruvate (CHEBI: 15361), succinate (CHEBI: 30031), creatine (CHEBI: 57947), and creatinine (CHEBI: 16737). In the 900 MHz spectra, betaine (CHEBI: 17750) and choline (CHEBI: 133341) were added to the above list of metabolites.

Verifying the goodness of the BATMAN fit The goodness of fit of the modeled metabolic signal can be checked by using the built-in tools of the R `batman` package. In particular, for multiple spectra, the fit can be assessed by examining a plot comparing the integrated bin intensities with the BATMAN metabolite fit and wavelet fit for each multiplet (see Figure 11.4). Note that the bin is placed over the modeled position of the multiplet. A lack of correlation between the integrated bin intensity and the BATMAN-estimated intensity may reveal a poor fit. Conversely, a correlation between the integrated bin intensity and the BATMAN wavelet fit may be an indication that the metabolite fit is failing to capture signal that should be modeled by a template.

It is worthwhile to note that comparing the integrated bin intensities with the BATMAN metabolite fit for multiplets in crowded-peak regions is less informative (i.e., it is not a solution for evaluating the BATMAN fit or for identifying problem spectra). To illustrate this point, consider Figure 11.4 showing the diagnostic scatterplot for alanine. While the integration values are aligned for the doublet at 1.509 ppm, they are somewhat scattered for the quadruplet at 3.810 ppm. This is primarily due to the glucose resonances that lie in the vicinity of the quadruplet and which contribute to the integrated bin intensity.

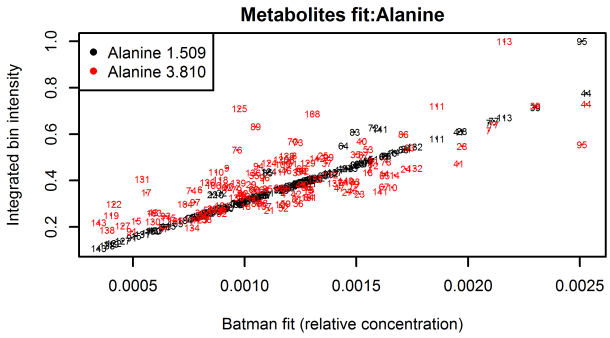

Figure 11.4: BATMAN diagnostic plot for alanine. Each number corresponds to a specific spectrum

Post-processing of spectral fits Once the metabolic signal has been correctly assigned, the residual signal captured by the wavelet component of the BATMAN model can be used to estimate the lipid concentrations. Towards this aim, integration regions that encompass lipid resonances were specified. Lipid resonances typically appear as broad peaks in NMR spectra. In general, the lipid integration regions selected for the BATMAN analysis are broader than those used for spectral binning (see Figure 11.5 and Table 11.3). The defined integration regions aim to capture the following broad lipid resonances: C**H**$_3$–(CH$_2$)$_n$– in the fatty acid chain (FAC), –CH$_3$–(C**H**$_2$)$_n$– in the FAC (captured using two integration regions in the 900 MHz analysis), –C**H**$_2$–CH$_2$–C=O or –C**H**$_2$–CH$_2$–CH=CH– in the FAC, –C**H**$_2$–CH=CH– in the FAC and C**H**$_3$ in N-acetylated glycoproteins (NAG), –C**H**$_2$–C=O or –C**H**$_2$–CH=CH– in FAC, =CH–C**H**$_2$–CH= in FAC, lysyl, and –C**H**=C**H**– in FAC. In this way, a set of lipid-specific features were obtained in addition to the relative metabolic concentrations. This approach only works when the metabolic signal has been sufficiently extracted. Should this not be the case, the residual signal will be contaminated by other metabolites resonating in the area.

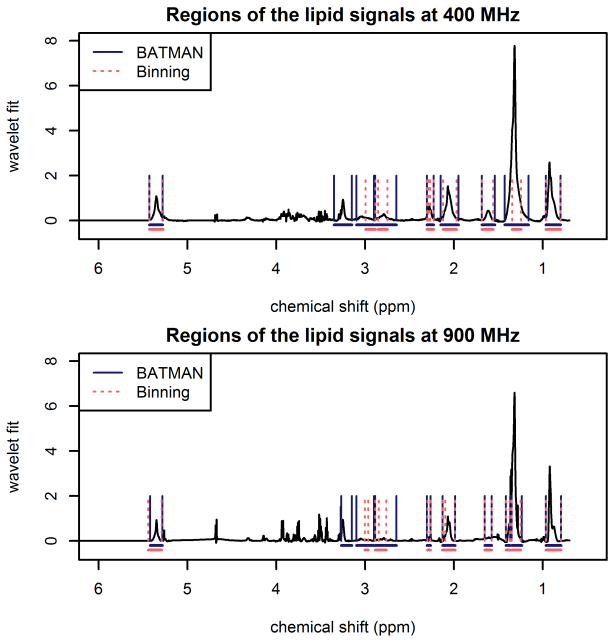

Figure 11.5: Illustration of the BATMAN wavelet-fit showing lipid integration regions for a 400 MHz (top) and 900 MHz (bottom) spectrum. The BATMAN integration regions that capture the broad lipid resonances are delimited by blue solid lines. The narrower spectral binning integration regions (delimited by red dashed lines) capture lipid signals, but not necessarily exclusively.

Table 11.3: Comparison of the lipid integration regions for the BATMAN and spectral binning analyses.

Lipid features	Manually pre-processed 400 MHz spectra Spectral binning*		Manually pre-processed 400 MHz spectra BATMAN		PepsNMR pre-processed 900 MHz spectra Spectral binning		PepsNMR pre-processed 900 MHz spectra BATMAN	
	Start	End	Start	End	Start	End	Start	End
–C**H**=C**H**– in FAC	5.4300	5.2752	5.4300	5.2800	5.4422	5.2900	5.4200	5.2833
albumin (lysyl)			3.3500	3.1500			3.2700	3.1500
=CH–C$\mathbf{H}_2$–CH= in FAC (b)	2.9950 2.8860	2.8860 2.8550	3.1000	2.9000	3.0047 2.8874	2.9655 2.8465	3.1000	2.9000
=CH–C$\mathbf{H}_2$–CH= in FAC (a)	2.8550	2.7500	2.8880	2.6500	2.8465	2.7623	2.8880	2.6500
–C$\mathbf{H}_2$–C=O or –C$\mathbf{H}_2$–CH=CH– in FAC	2.3040 2.2915 2.2690	2.2915 2.2690 2.2300	2.3060	2.2300	2.2990	2.2680	2.3060	2.2630
–C$\mathbf{H}_2$–CH=CH– in the FAC and C$\mathbf{H}_3$ in NAG	2.1230	1.9720	2.1500	1.9500	2.1300 2.0985	2.0975 1.9889	2.1289	1.9889
–C$\mathbf{H}_2$–CH_2–C=O or –C$\mathbf{H}_2$–CH_2–CH=CH– in the FAC	1.6860	1.5600	1.6860	1.5400	1.6530	1.5770	1.6530	1.5770
–CH_3–(C$\mathbf{H}_2$)$_n$– in FAC	1.3450	1.2458	1.4300	1.1600	1.4169 1.3516	1.3730 1.2500	1.4169 1.3516	1.3675 1.2366
C$\mathbf{H}_3$–(CH_2)$_n$– in FAC	0.9660	0.8000	0.9660	0.8000	0.9663	0.7961	0.9660	0.7961

Abbreviations: FAC: fatty acid chain, NAG: N-acetylated glycoproteins

* As reported by Louis et al. (2015)

12

Classification analysis

In this chapter, the classification methodology used to investigate the impact of spectral binning and BATMAN on the classification of lung cancer samples and controls is described (see Section 12.1). Section 12.2 presents the results of the classification analysis. The chapter concludes with a discussion of the results in Section 12.3.

12.1 Methodology

Spectral binning and spectral deconvolution by BATMAN were applied to the manually pre-processed (mp) 400 MHz and the `PepsNMR` automatically pre-processed (ap) 900 MHz ^{1}H-NMR spectra of lung cancer patients and control subjects (see Chapter 10 and Chapter 11). Spectral binning was also applied to the manually pre-processed 900 MHz ^{1}H-NMR spectra. As a result, the following five sets of predictors were obtained:

1. 110 integrated spectral regions (ISRs) based on the mp 400 MHz spectra.
2. Relative concentrations obtained using BATMAN for 33 metabolites and 9 lipid features based on the mp 400 MHz spectra.
3. 103 ISRs based on the ap 900 MHz spectra.
4. Relative concentrations obtained using BATMAN for 33 metabolites and 10 lipid features based on the ap 900 MHz spectra.
5. 105 ISRs based on the mp 900 MHz spectra.

Classifiers were built by using each set of predictors. The predictive performance of the classifiers was assessed by using a three-fold cross-validation (CV) scheme (see Figure 12.1). CV works by dividing the dataset in two parts, a training set and a test set. In our implementation, we ensure that the training and test set have the same proportion of cases and control subjects as the full dataset. In K-fold CV, the data are split into K roughly equal parts. In the k^{th} iteration, where $k = 1, \ldots, K$, the k^{th} part of the data forms the test set and the remaining $K - 1$ parts form the training

set. Thus, in three-fold CV, one-third of the data forms the test set and the remaining two-thirds of the data (i.e., the training set) are used to build the classifier. At each iteration, the performance of the classifier is evaluated in terms of the proportion of misclassifications and the sensitivity and specificity of the classifier when applied to the test set. Since the splitting is not uniquely determined (Slawski et al., 2008), the cross validation procedure was repeated 333 times. The overall performance is based on the mean classification error rate, the mean sensitivity, and the mean specificity of the 999 classifiers.

For the classification analysis involving the binning features, variable selection was based on the discriminative power of the individual bins between the two conditions. This was assessed by using *limma* (Smyth et al., 2003), as indicated by the asterisk in Figure 12.1. In a two-class setting, *limma* is a moderated t-statistic based on a hybrid frequentist empirical Bayes linear model. Let $\sigma_1^2, \ldots, \sigma_m^2$ denote the ISR-specific variances. *Limma* assumes a scaled inverse chi-square prior density for $\sigma_1^2, \ldots, \sigma_m^2$ with hyperparameters s_0^2 for the prior variance and ν_0 for the prior degrees of freedom. The hyperparameters are estimated by applying an empirical Bayes function to the sample variances $s_1^2, \ldots, s_m^2$. Moderated t-statistics are produced by dividing the standard frequentist numerator of the t-test statistic by a denominator in which the sample variance s_i^2 with ν degrees of freedom is replaced with $\tilde{s}_i^2 = (\nu_0 s_0^2 + \nu s_i^2)/(\nu_0 + \nu)$, i.e., the posterior mean of $\sigma_i^2 | s_i^2$. The moderated t-statistic follows a t-distribution with $\nu_0 + \nu$ degrees of freedom under the null hypothesis.

The classification analysis involving the BATMAN estimated features proceeded using the following three subsets of the features: (1) all the BATMAN-estimated relative metabolite concentrations, (2) all the relative lipid concentrations, and (3) all the BATMAN-estimated relative metabolic concentrations together with all the relative lipid concentrations.

Five classification methods that are appropriate for the analysis of large, complex datasets were used to build the classifiers, namely, elastic net, lasso, orthogonal partial least squares-discriminant analysis (OPLS-DA), support vector machines (SVMs), and random forests (RF). The selected range of classifiers by no means encompasses all the possible classifiers that could be considered. It is not our goal to investigate all the classifiers or to identify the most optimal classification approach. A brief description of each of the selected classification methods used is provided below. The reader is referred to Hastie et al. (2009) and Bylesjö et al. (2006) for further details.

Lasso and elastic net are both regularized regression procedures (i.e., penalty terms are added to the regression framework, which is logistic regression in this case). Lasso utilizes the L1 penalty which constrains the sum of absolute values of the regression coefficients. Lasso enables variable selection as L1 regularization allows for some regression coefficients to be shrunk to zero. Ridge regression, also a regularized

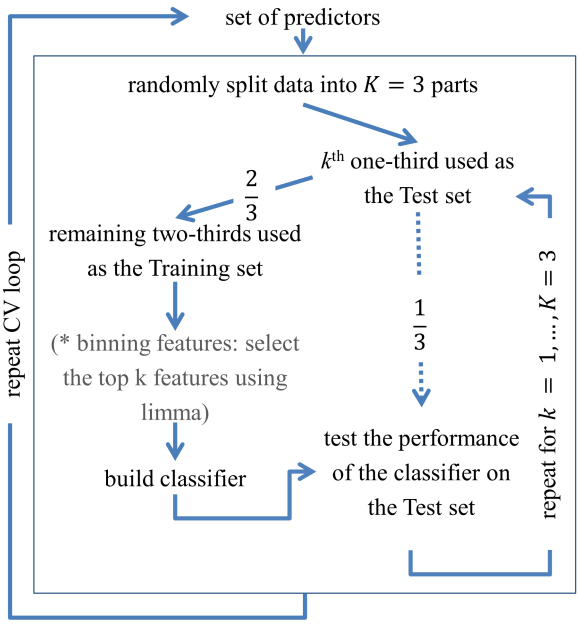

Figure 12.1: The three-fold cross-validation procedure.

regression procedure, utilizes the L2 penalty which constrains the sum of squared regression coefficients. Regularization with the L2 penalty is good for handling correlated predictors. Elastic net combines the L1 penalty of lasso with the L2 penalty of ridge regression. Thus, elastic net enables variable selection and allows for strongly correlated predictors to either enter or be left out of the model together.

Partial least squares-discriminant analysis (PLS-DA) and orthogonal partial least squares-discriminant analysis (OPLS-DA) are well-known multivariate-regression approaches, used for performing classification, in metabolomics. Given two matrices, a response matrix Y containing class information and a matrix of predictors X, PLS-DA models X and Y simultaneously with the aim of maximizing the covariance between X and Y. The procedure identifies latent variables (a.k.a. PLS components) in X that are predictive of Y. In OPLS-DA, the information contained in Y is used to split the X matrix into blocks that are correlated with Y, capturing the predictive variation, and orthogonal to Y, capturing the non-predictive variation. Thus, the variation in X that is not correlated with Y is removed. This reduces the complexity of the final model making the results more interpretable. SVMs search for a hyperplane that maximizes the margin of separation between the two classes. RF classifiers combine a number of decision trees each based on bootstrapped samples of the training dataset.

Although *limma* was not used to build the BATMAN feature-based classifiers, it

was applied (in 333 iterations of three-fold cross-validation) as a univariate approach to identify the top 15 variables of each of the five sets of predictors. For each training dataset in the three-fold cross-validation procedure, the predictors are ranked according to their associated *limma* t-test statistics. The top ranking k = 15 features of each dataset are selected. The frequency with which each feature appears in the collection of top 15 feature lists is computed and those features appearing most frequently in the top 15 lists across the iterations are selected. These variables were identified to check whether there were any similarities in the most discriminative variables selected from each set of predictors.

The classification analysis was conducted by using the R statistical software (version 3.2.3, R Development Core Team, 2015). Classification methods were implemented by using the default options of the R Bioconductor package `CMA` (Slawski et al., 2008).

12.2 Results

A single list of top ranking features was obtained at every step of the repeated CV based on the ranking of *limma* t-test statistics (see Section 12.1). This resultant in a total of 999 top 15 feature lists for each of the five sets of predictors. Tables 12.1 to 12.5 list the features appearing most frequently across the 999 top 15 feature lists together with their frequency of appearance. There are similarities in the top ISRs of the 400 MHz and 900 MHz (ap and mp) ISRs. In particular, the following ISRs appear in the top 15 lists for all three sets of spectral binning features (see Table 12.1, Table 12.3, and Table 12.5): isoleucine (around 0.97 ppm), threonine (around 3.61 ppm), glycerol (around 3.66 ppm), glucose (around 3.90 ppm), and asparagine, histidine, serine, and tyrosine (around 3.98 ppm). Unidentified ISRs also appear in the top 15 lists. However, the chemical shift of these unidentified ISRs differ across the three tables. Of the common ISRs listed above, only the metabolites glucose and asparagine appear in the top 15 list of the 400 MHz BATMAN analysis (see Table 12.1). Glucose, serine, histidine, threonine, tyrosine, and asparagine appear in the top 15 list of the 900 MHz BATMAN univariate analysis (see Table 12.4).

Table 12.1: Top integration regions for the 400 MHz spectral binning analysis (based on repeated three-fold cross-validation using the *limma* t-test).

Top 15 integration regions			
Region	Chemical shift	Signal found in region	Top 15 frequency
36	$[4.0010, 3.9810]$	C_3H_2 in glycerol backbone of PL, Asparagine, Histidine, Phenylalanine, Serine	999
37	$[3.9810, 3.9590]$	Creatine, Asparagine, Histidine, Tyrosine, Serine	999
48	$[3.6376, 3.6240]$	Valine	998
29	$[4.1260, 4.1110]$	Not identified	997
49	$[3.6240, 3.6097]$	Threonine	969
91	$[2.1230, 1.9720]$	Lipids: –C$\mathbf{H}_2$–CH=CH– in FAC, C$\mathbf{H}_3$ of NAG, Glutamate, Isoleucine, Methionine, Proline	902
45	$[3.7141, 3.6680]$	O–C$\mathbf{H}_2$–CH_2–$N^+(CH_3)_3$ of PC and SM, Glycerol, Isoleucine	865
28	$[4.1750, 4.1260]$	C_1H and C_3H in glycerol backbone of PL and TG, Lactate	856
46	$[3.6680, 3.6500]$	Glycerol	803
23	$[4.5380, 4.4100]$	Not identified	766
50	$[3.6097, 3.5914]$	Threonine	729
109	$[0.9760, 0.9660]$	Isoleucine	690
2	$[7.8200, 7.7890]$	Histidine	653
73	$[2.5960, 2.5340]$	Citrate	534
38	$[3.9590, 3.8330]$	Glucose, Aspartate, Methionine, Serine, Tyrosine	521

Abbreviations: FAC: fatty acid chain, NAG: N-acetylated glycoproteins, PC: phosphatidylcholine, PL: phospholipids, SM: sphingomyelins, TG: triglycerides

Table 12.2: Top metabolite/lipid features for the 400 MHz BATMAN analysis (based on repeated three-fold cross-validation using the *limma* t-test).

Top 15 metabolite/lipid features	Top 15 frequency
Lipids: –C**H**$_2$–CH=CH– in FAC and C**H**$_3$ in NAG	999
Lactate	996
Lipids: =CH–C**H**$_2$–CH= in FAC (a)	993
α-Ketoglutarate	987
Cysteine	934
Lipids:=CH–C**H**$_2$–CH= in FAC (b)	923
Acetoacetate	887
α-D-glucopyranose	867
Lysyl	815
Lipids: –C**H**=C**H**– in FAC	778
Proline	730
Alanine	625
Lipids: C**H**$_3$–(CH$_2$)$_n$– in FAC	597
Methionine	584
Asparagine	342

Abbreviations: FAC: fatty acid chain, NAG: N-acetylated glycoproteins

Table 12.3: Top integration regions for the `PepsNMR` pre-processed 900 MHZ spectral binning analysis (based on repeated three-fold cross-validation using the *limma* t-test).

Top 15 integration regions			
Region	Chemical shift	Signal found in region	Top 15 frequency
31	$[3.9680, 3.9600]$	Creatine	999
80	$[2.2680, 2.2563]$	Acetone	999
82	$[2.1930, 2.1700]$	Glutamine	999
36	$[3.8410, 3.8140]$	Glucose, Alanine, Glutamine, Glutamate, Serine	991
71	$[2.6900, 2.6597]$	Methionine	991
34	$[3.9150, 3.8920]$	Glucose	975
30	$[3.9920, 3.9680]$	Asparagine, Histidine, Serine, Tyrosine	974
61	$[3.0940, 3.0785]$	Tyrosine	944
16	$[6.7600, 6.7004]$	Not identified	893
39	$[3.7240, 3.6500]$	Glycerol	853
102	$[0.9800, 0.9550]$	Isoleucine	767
94	$[1.3730, 1.3516]$	Lactate	727
41	$[3.6163, 3.5930]$	Threonine	713
10	$[7.3510, 7.3227]$	Not identified	462
83	$[2.1700, 2.1650]$	Methionine	433

Table 12.4: Top metabolite/lipid features for the `PepsNMR` pre-processed 900 MHz BATMAN analysis (based on repeated three-fold cross validation using the *limma* t-test).

Top 15 metabolite/lipid features	Top 15 frequency
α-D-glucopyranose	999
α-Ketoglutarate	999
Serine	999
Histidine	992
Citrate	986
Glycine	977
Glutamate	977
Threonine	926
Tryptophan	884
Lipids: C**H**$_3$–(CH$_2$)$_n$–in FAC	708
Tyrosine	662
Myo-inositol	651
Lipids: –C**H**$_2$–C=O or –C**H**$_2$–CH=CH– in FAC	628
Lipids: =CH–C**H**$_2$–CH= in FAC (b)	598
Asparagine	554

Abbreviations: FAC: fatty acid chain

Table 12.5: Top integration regions for the manually pre-processed 900 MHz spectral binning analysis (based on repeated three-fold cross-validation using the *limma* t-test).

Top 15 integration regions			
Region	Chemical shift	Signal found in region	Top 15 frequency
39	$[3.7204, 3.6453]$	Glycerol	999
87	$[2.1289, 2.0993]$	Lipids: –C**H**$_2$–CH=CH– in FAC, C**H**$_3$ of NAG	999
34	$[3.9120, 3.8957]$	Glucose	991
88	$[2.0993, 1.9889]$	Lipids: –C**H**$_2$–CH=CH– in FAC, C**H**$_3$ of NAG	988
20	$[5.2186, 5.2038]$	C_2H in glycerol backbone of PL and TG	982
73	$[2.6768, 2.6597]$	Methionine	973
94	$[1.4587, 1.4201]$	Lysine	931
104	$[0.9766, 0.9663]$	Isoleucine	917
41	$[3.6163, 3.5861]$	Threonine	846
99	$[1.2240, 1.1766]$	Not identified	815
30	$[3.9903, 3.9644]$	Asparagine, Histidine, Serine, Tyrosine	749
13	$[7.0792, 7.0597]$	Histidine	597
16	$[6.7460, 6.7004]$	Not identified	595
101	$[1.0513, 1.0340]$	Isoleucine	510
40	$[3.6453, 3.6212]$	Valine	484

Abbreviations: FAC: fatty acid chain, NAG: N-acetylated glycoproteins, PL: phospholipids, TG: triglycerides

Figure 12.2 illustrates the fit of the BATMAN model in the region extending from 2.99 to 3.11 ppm for both the 400 MHz and 900 MHz spectrum of a particular plasma sample. This region contains resonances from creatine (singlet), creatinine (singlet), lysine (triplet), and tyrosine (double doublet), as well as a part of the lipid =CH–C**H**$_2$–CH= resonance. The resonances are more distinguishable in the 900 MHz spectrum compared to the 400 MHz spectrum. For the 400 MHz spectrum, the four integration regions from left to right aim to capture the signal corresponding to (1) cysteine, lysine, and tyrosine; (2) cysteine, lysine, tyrosine, and creatinine; (3) cysteine, lysine, tyrosine, creatinine, and creatine; and (4) cysteine, lysine, tyrosine, and α-ketoglutarate. For the 900 MHz spectrum, the four integration regions from left to right, beginning at 3.0921 ppm, correspond to (1) tyrosine, (2) creatinine, (3) creatine, and (4) lysine and α-ketoglutarate.

Box plots of the misclassification errors, sensitivities, and specificities of the various classifiers (elastic net, lasso, OPLS-DA, SVMs, and RF) for each of the considered sets of predictors are shown in Figure 12.3, Figure 12.4, and Figure 12.5, respectively. An overview of these results are provided in Table D.1.1 and Figure 12.6. Note that the spectral binning classifiers were built using the top k ISRs where k is a series of numbers extending from three to the total number of ISRs forming the set of predictors. The misclassification error, sensitivity and specificity of each of these spectral binning classifiers developed using elastic net are shown in Figure D.1.1 (for the 400 MHz ISRs), Figure D.1.2 (for the ap 900 MHz ISRs), and Figure D.1.3 (for the mp 900 MHz ISRs). Of the classifiers built using the top k ISRs, only the best performing classifiers are reported in Table D.1.1. For most of the sets of features, the classification approaches (elastic net, lasso, OPLS-DA, RF, and SVMs) performed more or less similarly and none stand out as being consistently better than the others. However, based on the classification performance, i.e., the misclassification errors, sensitivity, and specificity, of the classifiers, the decision was made to proceed with elastic net which performed reasonably well for all the sets of features that will be focused on.

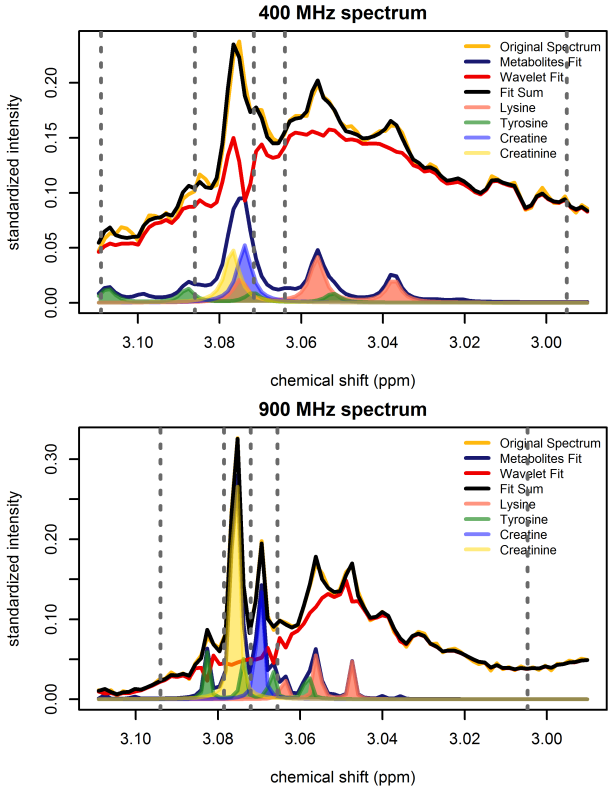

Figure 12.2: BATMAN fit in the region extending from 2.99 to 3.11 ppm for the 400 MHz spectrum (top) and the 900 MHz spectrum (bottom) of a plasma sample. The original spectrum is shown in yellow. The two components of the BATMAN model fit, that is, the component modeling the metabolic signal (metabolites fit) and the component capturing the residual signal (wavelet fit) are indicated by blue and red curves, respectively. The fit sum which is the sum of the metabolite fit and the wavelet fit is shown in black. The shaded regions show the resonances from creatine (blue), creatinine (yellow), lysine (pink), and tyrosine (green) that are captured by the metabolite fit. The broad lipid $=$CH–C$\mathbf{H}_2$–CH$=$ resonance in the region is captured by the wavelet fit. Binning integration region limits for the region are delimited by grey dotted lines.

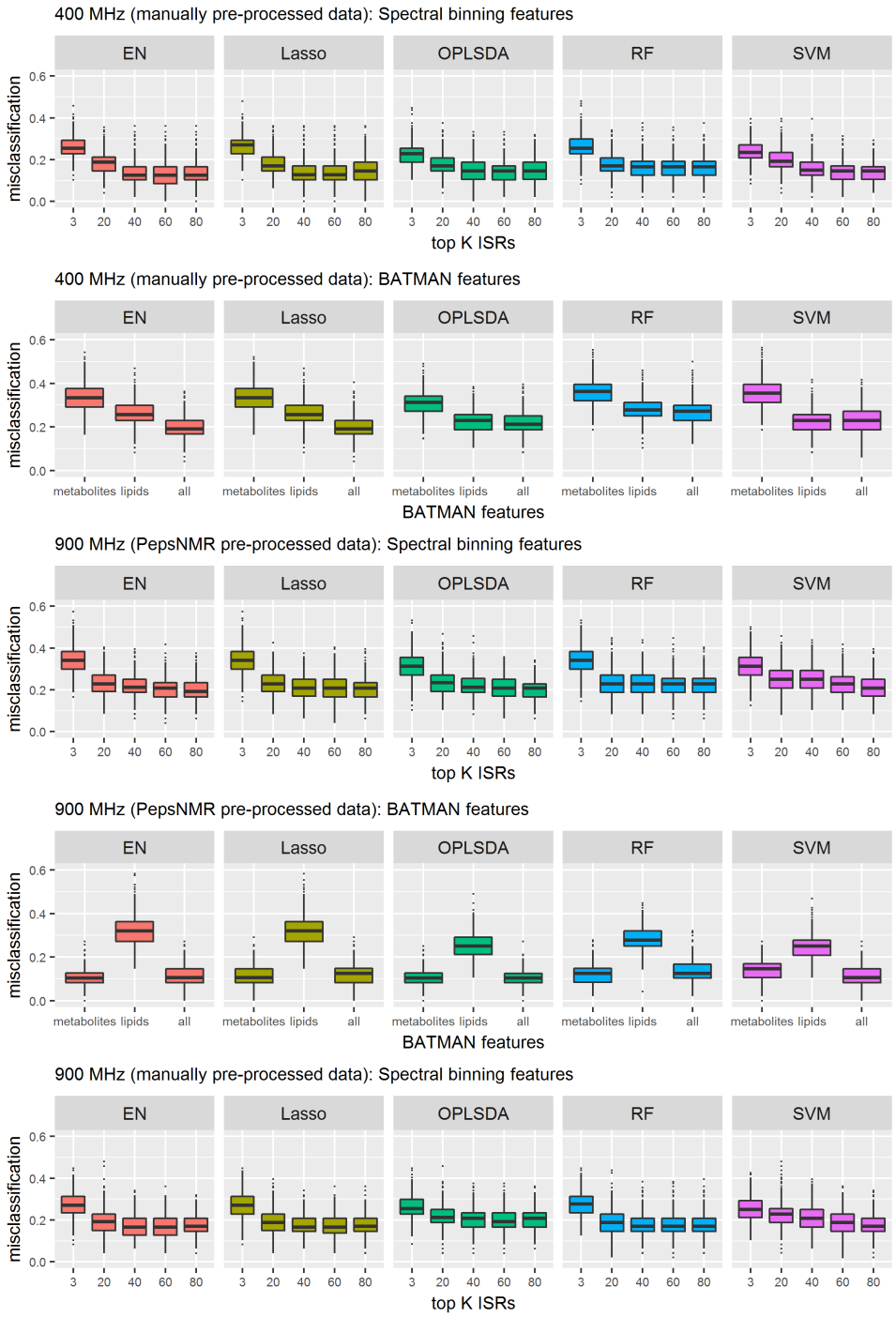

Figure 12.3: Box plots of the misclassification errors of the elastic net, lasso, orthogonal partial least squares-discriminant analysis (OPLS-DA), random forest (RF), and support vector machine (SVM) classifiers.

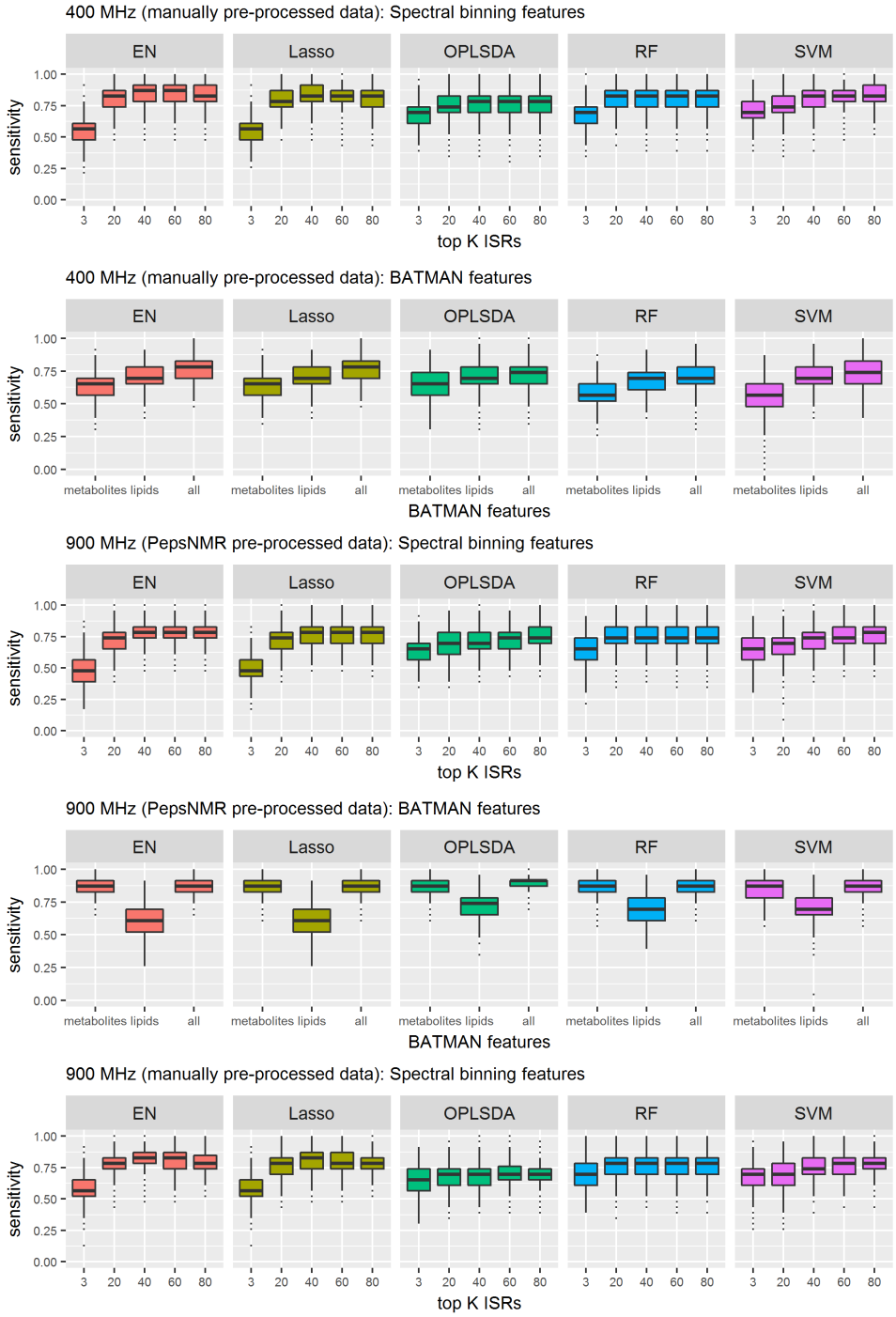

Figure 12.4: Box plots of the sensitivity of the elastic net, lasso, orthogonal partial least squares-discriminant analysis (OPLS-DA), random forest (RF), and support vector machine (SVM) classifiers.

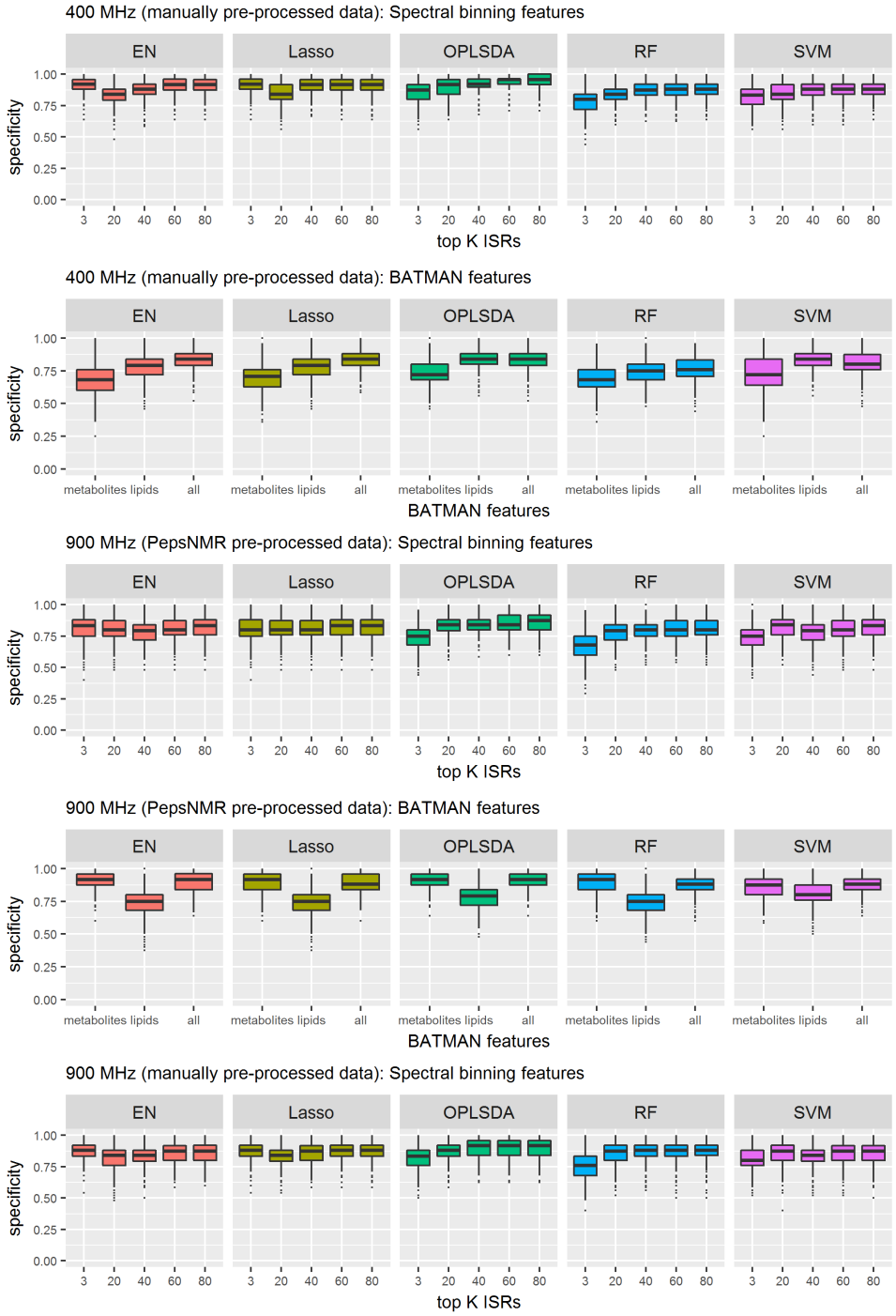

Figure 12.5: Box plots of the specificity of the elastic net, lasso, orthogonal partial least squares-discriminant analysis (OPLS-DA), random forest (RF), and support vector machine (SVM) classifiers.

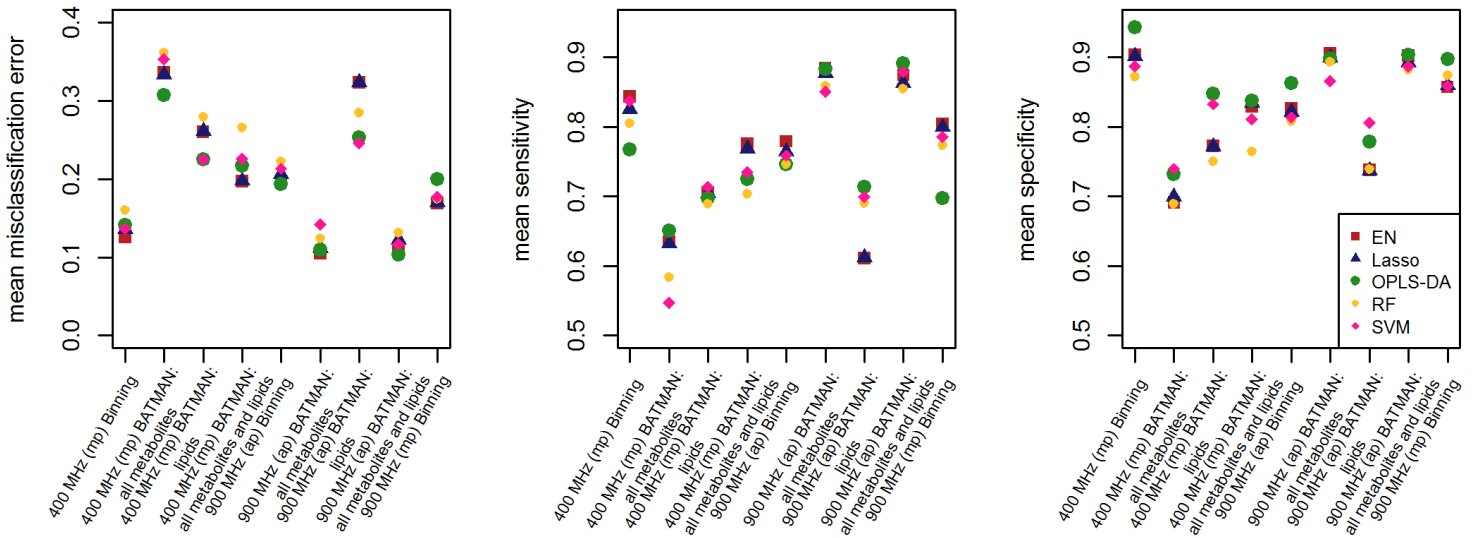

Figure 12.6: Classification performance in terms of mean misclassification error, mean sensitivity and mean specificity of the elastic net, lasso, orthogonal partial least squares-discriminant analysis (OPLS-DA), random forest (RF), and support vector machine (SVM) classifiers.

Table 12.6 presents the mean cross-validated classification error, sensitivity, and specificity of the 400 MHz and 900 MHz elastic net classifiers. The 400 MHz classification results indicate that the ISRs (misclassification rate: 0.125, sensitivity: 0.844, specificity: 0.904) had greater predictive power than the relative metabolic and lipid concentrations obtained by using BATMAN (classification error: 0.197, sensitivity: 0.775, specificity: 0.829). For the 900 MHz classification analysis, the relative metabolic concentrations estimated by BATMAN (misclassification rate: 0.105, sensitivity: 0.884, specificity: 0.906) had greater predictive power than the ISRs of the 900 MHz spectral bins (classification error rate: 0.170 (mp), 0.197 (ap); sensitivity: 0.813 (mp), 0.779 (ap); specificity: 0.846 (mp), 0.826 (ap)). Note that for the `PepsNMR` automatically pre-processed 900 MHz spectra, an additional spectral alignment step was carried out to improve the homogeneity of the bins (in terms of the signal captured) across spectra.

Table 12.6: Elastic net classification results (standard errors in parentheses).

Features	Misclassification	Sensitivity	Specificity
400 MHz (manually pre-processed data)			
Binning: top integrated spectral regions[a]	0.125 (0.002)	0.844 (0.003)	0.904 (0.002)
BATMAN: all metabolites and lipids[a]	0.197 (0.002)	0.775 (0.003)	0.829 (0.002)
900 MHz (`PepsNMR` automatically pre-processed data)			
Binning: top integrated spectral regions[a]	0.197 (0.002)	0.779 (0.003)	0.826 (0.003)
BATMAN: all metabolites[a]	0.105 (0.001)	0.884 (0.002)	0.906 (0.002)
900 MHz (manually pre-processed data)[b]			
Binning: top integrated spectral regions	0.170 (0.002)	0.813 (0.003)	0.846 (0.002)

[a] Features utilized in Figure 12.7.

[b] The 900 MHz manually pre-processed spectra were not of sufficient quality to fit the BATMAN model.

Histograms of the probability of lung cancer for the different sets of features are presented in Figure 12.7. Each histogram is based on the classifiers developed using the subset of features indicated by the letter 'a' in Table 12.6. Assuming that a probability greater than 0.5 implies the presence of lung cancer, the ISRs of the 400 MHz spectral bins and the 900 MHz relative metabolic concentrations estimated by BATMAN produced the best classifiers in terms of lowest misclassification error and highest sensitivity and specificity.

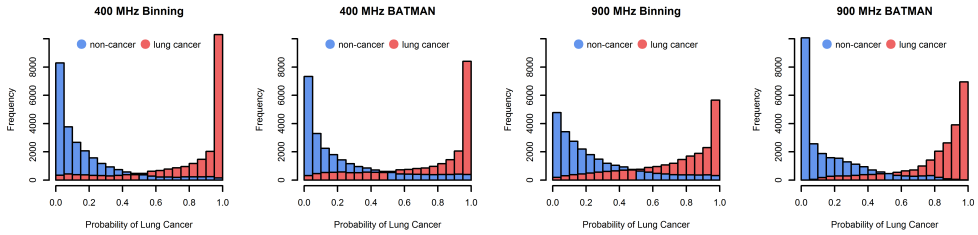

Figure 12.7: Histograms of the probability of lung cancer based on 333 iterations of three-fold cross-validation. Blue corresponds to the control samples and red represents the lung cancer samples.

Figure 12.8 illustrates the receiver operating characteristic (ROC) curves for the elastic net classifiers (grey curves) based on the subset of features indicated by the letter 'a' in Table 12.6. The threshold averaged ROC curve is indicated in red. The mean area under the curve for the 400 MHz Binning, 400 MHz BATMAN, 900 MHz Binning, and 900 MHz BATMAN features is 0.932 (0.001), 0.885 (0.001), 0.880 (0.002), and 0.963 (0.001), respectively. Standard errors are reported in parentheses. Based on the ROC curves, it is once again evident that the classifiers based on the ISRs of the 400 MHz spectral bins and the 900 MHz relative metabolic concentrations estimated by BATMAN are more capable of distinguishing between the lung cancer and control samples compared to the 400 MHz BATMAN features and the 900 MHz (`PepsNMR` pre-processed) ISRs, respectively.

12.3 Discussion & Conclusions

In this study, spectral binning and spectral deconvolution using BATMAN were applied in order to extract metabolic signal from ^{1}H-NMR spectra of different spectral resolutions (400 MHz versus 900 MHz spectra).

Implementation Both spectral binning and spectral deconvolution using BATMAN require expert knowledge of the characteristic spectral signatures (i.e., the peak locations and coupling patterns) of different metabolites. For spectral binning, this

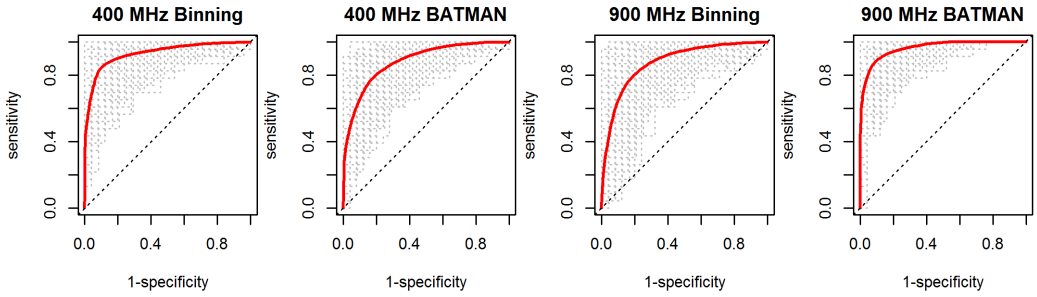

Figure 12.8: Receiver operating characteristic (ROC) curves of the elastic net classifiers at each iteration of the repeated three-fold cross-validation for the feature sets labeled with the letter 'a' in Table 12.6 (grey curves). The threshold averaged ROC curve is indicated in red.

insight is necessary to select meaningful integration regions. For spectral deconvolution using BATMAN, this information is required to accurately specify and refine the prior information on each multiplet of interest.

Despite BATMAN's description as an automated metabolite analyzer, an extensive amount of time was spent on developing and fine-tuning the template file in order to improve signal extraction. Although metabolites have characteristic resonances, experimental parameters and pre-processing steps influence the resultant chemical shift positions, identifiable coupling patterns, and relative peak intensities. Note that a single template file is specified for a large number of spectra which exhibit between-spectrum variation in peak shift and peak definition. Thus, template adjustments made to improve the fit of some spectra or peaks may have an opposite effect on others. Updating the template file is a repetitious task which is extremely time-consuming, especially for crowded spectral regions, but it is essential. Once the template database is developed, the process is automated.

Though selecting the integration regions for spectral binning is a manual task, spectral binning is a relatively fast and straightforward method for ^{1}H-NMR signal extraction.

The magnetic field strength of the NMR spectrometer influences the resolution of the metabolic peaks. In higher resolution spectra, peaks appear with greater definition, exhibit fewer higher-order effects, and show less overlap. This is beneficial for both spectral binning and spectral deconvolution using the BATMAN model. Fewer overlapping regions imply a greater one-to-one mapping between spectral bins and metabolites (Louis et al., 2017) and the increased signal-to-noise ratio in the higher resolution spectra is advantageous for metabolic signal extraction using BATMAN (see Figure 12.2).

Classification and Clinical Relevance An abundance of detail pertaining to biological functions is contained within the metabolome. There is a strong desire to eventually utilize these data to make informed clinical decisions about disease status, susceptibility, and progression. It is expected that metabolomics will be of vital importance in reaching the goal of providing healthcare that is customized for individual patients. Therefore, obtaining interpretable, reliable, and reproducible results is essential.

The variation in chemical shift locations across spectra is a challenge for spectral binning. Therefore, the inclusion of a spectral alignment step in the pre-processing of NMR data is important in order to obtain reliable and interpretable features. However, even with good spectral alignment, overlapping peaks often prevent a one-to-one mapping between integration regions and metabolites. Integration regions, especially those of lower resolution spectra, may contain signals from two or more metabolites in conjunction with an unidentified signal (for illustration, see Tables 12.1 to 12.5). Thus, a drawback of the simplicity surrounding spectral binning is the lack of biological interpretability of the resultant features. Nonetheless, for the 400 MHz analysis, the classifier based on the binning features performed better than the one using BATMAN-estimated features.

Spectral deconvolution, particularly the BATMAN model, provides the means to obtain a single concentration estimate for each metabolite of interest. The residual signal captured by wavelets can be divided into integration regions in order to capture for instance, broad lipid resonances. In the end, clinically relevant features are extracted from the ^{1}H-NMR spectra. The benefit obtained from the effort put into running BATMAN is biological interpretability. In addition, although not the focus of this manuscript, the reliability of BATMAN estimated relative concentrations can also be assessed by using the 95% credible intervals. For the 900 MHz spectra, the relative metabolic concentrations estimated by BATMAN excelled, producing the best performing classifier in terms of mean misclassification error.

13

Concluding remarks and further research

In this dissertation, a variety of statistical methods were explored for the analysis of transcriptomic and metabolomic data. In Part I, three statistical approaches were proposed for investigating the metabolite-co-expression association of a gene module (Chapter 4, Chapter 5, and Chapter 7). The motivation behind this research was to improve on a previously implemented approach for investigating the conditional co-expression of a gene module. Part II focused on metabolic data analysis. The steps involved in pre-processing ^{1}H-NMR data for metabolic signal extraction were described (Chapter 10). Metabolic features were extracted from ^{1}H-NMR spectra using two signal extraction procedures and the resulting features were compared in terms of their ability to classify lung cancer patients and control subjects (Chapter 11 and Chapter 12).

This chapter of the dissertation provides an overall discussion, concluding remarks, and suggestions for further research.

13.1 Conditional co-expression analysis of a gene module

Part I begins by introducing the simple linear regression approach implemented by Inouye et al. (2010a) for investigating the metabolite-co-expression association of a gene module. The approach involves the simple linear regression of Spearman's correlation coefficients for all pairs of genes of a gene module for five subsets of samples formed by using quintiles of the metabolite concentrations. Attention was drawn to several limitations of the approach.

To improve on the linear-regression-based approach, three comprehensive statistical models that facilitate the inference of conditional co-expression for a gene module were presented (see Chapter 4, Chapter 5, and Chapter 7). Each of the approaches were investigated through simulation studies and illustrated using a subset of the

DILGOM study data described in Chapter 3.

In Chapter 4, a multivariate linear model for investigating the association between gene-module co-expression and a categorical covariate was described. The model uses a block-diagonal variance-covariance structure consisting of metabolic-subset specific general variance-covariance blocks to capture the dependence between adjusted gene-expression values. Inference is based on the Larntz & Perlman test statistic. The model addresses the limitations of the simple linear regression approach and can be easily implemented using existing statistical software like SAS (PROC GLIMMIX).

A model for investigating the association between gene-module co-expression and a continuous covariate was described in Chapter 5. The model avoids the arbitrary categorisation of metabolite concentrations by modelling the gene-pair correlations as a function of the metabolite concentrations. The likelihood ratio test statistic is used to infer conditional co-expression. For a specific gene-module size, the number of variance-covariance parameters that must be estimated with the continuous approach (i.e., the Chapter 5 model) is substantially smaller than for the categorical approach (i.e., the Chapter 4 model). Moreover, the power to detect the simulated approximately linear associations is significantly greater for the model using continuous concentrations (power: 0.863 and 0.859 for the positive and negative association, respectively) than for the model using metabolic subsets (power: 0.215 and 0.218 for the positive and negative association, respectively).

Fitting a multivariate model that fully captures the dependence structure of several variables can become increasingly challenging as the number of parameters and the size of the variance-covariance matrix increases. Chapter 7 detailed a more computationally feasible solution, than the multivariate model of Chapter 5, in the form of a copula-based pseudo-likelihood approach for investigating the conditional co-expression of a gene module. In addition to reducing the computational burden, the approach facilitated the estimation of non-parametric measures of association such as Kendall's tau and Spearman's rho. Furthermore, the copula-based pseudo-likelihood approach using the pseudo-likelihood ratio test had greater power to detect metabolite-co-expression associations that lie further away from the null hypothesis compared to the multivariate approach of Chapter 5. A formal investigation into this phenomenon is of interest for future work.

The versatility of the three statistical methodologies, proposed in Chapter 4, Chapter 5, and Chapter 7, was illustrated. Additional steps could be taken to increase the flexibility of the models. Instead of using a fixed transformation of the metabolite concentrations (see equation 5.4), one could consider using a Box-Cox transformation. A Box-Cox transformation could also be applied to the gene-expression values. In order to apply this in an automated way, the selection of the optimal Box-Cox transformation parameter should be embedded into the modelling process. This would

increase the flexibility of the model, although also the complexity of it.

The DILGOM data was mainly used for illustration and we didn't strive to identify the optimal model for each metabolite. However, majority of the residual plots are satisfactory. The application of a Box-Cox transformation to the expression values of MS4A2 could be further explored.

In Chapter 7, Gaussian, Gumbel-Hougaard, and Clayton copulas were used to model the DILGOM data. Based on a comparison of the AICs of the bivariate likelihoods forming each pseudo-likelihood, it was concluded that the Gaussian copula was most suitable. The identification of a formal approach for selecting the overall best fitting copula-based pseudo-likelihood model is a topic for further research.

13.2 Metabolic data analysis

Metabolic data analysis was the focus of Part II of this dissertation. The impact of spectral binning and spectral deconvolution using BATMAN (Bayesian AuTomated Metabolite Analyser for NMR data) for extracting metabolic signal from proton nuclear magnetic resonance (^{1}H-NMR) data on the classification of lung cancer samples was studied. The ^{1}H-NMR data of blood plasma samples, extracted from lung cancer patients and control subjects, attained using a 400 MHz and a 900 MHz ^{1}H-NMR spectrometer, were analysed using the two metabolic signal extraction approaches. A comparison of the classification performance of the extracted features, i.e., the integrated spectral regions (ISRs) arising from spectral binning and the BATMAN estimated relative concentrations, was performed separately for the 400 MHz and 900 MHz spectra. For the 400 MHz data, the spectral binning approach provided greater discriminatory power. However, for the 900 MHz data, the relative metabolic concentrations obtained by using BATMAN provided greater predictive power. Spectral binning is computationally advantageous and less laborious. However, BATMAN estimated features correspond directly with specific metabolites and therefore have a simpler interpretation.

Although the aim of this study was to compare the overall performance of the features in terms of classification accuracy, it would be of value to further investigate the features selected by the classifiers in an attempt to understand the reason for the differences in classification performance of the ISRs and BATMAN features.

In this study, two resolutions of ^{1}H-NMR spectra, i.e., 400 MHz and 900 MHz spectra, were analysed. Further research could focus on conducting the same investigation using samples analysed with, for instance, a 600 MHz spectrometer. This could contribute to the establishment of more concrete conclusions regarding the influence of the technological platform on the classification performance of the ISRs and BATMAN features.

BATMAN was used to estimate the relative concentrations of metabolites. Estimating absolute metabolite concentrations and exploring the performance of these features on the ability to distinguish between samples could be another topic for further research.

Bibliography

Aerts, M., Molenberghs, G., Ryan, L. M., and Geys, H. (2002). *Topics in modelling of clustered data*. CRC Press.

Alonso, A., Marsal, S., and Julià, A. (2015). Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in Bioengineering and Biotechnology*, 3.

Andor, M. and Parmeter, C. (2017). Pseudolikelihood estimation of the stochastic frontier model. *Applied Economics*, 49(55):5651–5661.

Arnold, B. C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 233–243.

Astle, W., De Iorio, M., Richardson, S., Stephens, D., and Ebbels, T. (2012). A bayesian model of nmr spectra for the deconvolution and quantification of metabolites in complex biological mixtures. *Journal of the American Statistical Association*, 107(500):1259–1271.

Beckonert, O., Keun, H. C., Ebbels, T. M., Bundy, J., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for nmr spectroscopy of urine, plasma, serum and tissue extracts. *Nature protocols*, 2(11):2692.

Beisken, S., Eiden, M., and Salek, R. M. (2015). Getting the right answers: understanding metabolomics challenges. *Expert Review of Molecular Diagnostics*, 15(1):97–109.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252.

Bylesjö, M., Rantalainen, M., Cloarec, O., Nicholson, J. K., Holmes, E., and Trygg, J. (2006). Opls discriminant analysis: combining the strengths of pls-da and simca classification. *Journal of Chemometrics*, 20(8-10):341–351.

Chen, J., Xie, J., and Li, H. (2011). A penalized likelihood approach for bivariate conditional normal models for dynamic co-expression analysis. *Biometrics*, 67(March):299–308.

Cole, N. (1968). On testing equality of correlation matrices. Technical Report 66, University of North Carolina.

de Graaf, R. A. and Behar, K. L. (2003). Quantitative 1h nmr spectroscopy of blood plasma metabolites. *Analytical Chemistry*, 75(9):2100–2104.

de la Fuente, A. (2010). From 'differential expression' to 'differential networking'-identification of dysfunctional regulatory networks in diseases. *Cell - Trends in Genetics*, 26(7):326–333.

Euceda, L. R., Giskeødegård, G. F., and Bathen, T. F. (2015). Preprocessing of nmr metabolomics data. *Scandinavian Journal of Clinical and Laboratory Investigation*, 75(3):193–203.

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62(2):424–431.

Furlotte, N. a., Kang, H. M., Ye, C., and Eskin, E. (2011). Mixed-model coexpression: Calculating gene coexpression while accounting for expression heterogeneity. *Bioinformatics*, 27:288–294.

Galecki, A. and Burzykowski, T. (2013). *Linear mixed-effects models using R, A step-by-step approach*. Springer.

Geys, H., Molenberghs, G., and Ryan, L. M. (1999). Pseudolikelihood modeling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, 94(447):734–745.

Gill, P. S. (2004). Small-sample inference for the comparison of means of log-normal distributions. *Biometrics*, 60(2):525–527.

Gillis, J. and Pavlidis, P. (2009). A methodology for the analysis of differential coexpression across the human lifespan. *BMC Bioinformatics*, 10:306.

Gowda, G. N., Zhang, S., Gu, H., Asiago, V., Shanaiah, N., and Raftery, D. (2008). Metabolomics-based methods for early disease diagnostics. *Expert Review of Molecular Diagnostics*, 8(5):617–633.

Haldermans, P., Shkedy, Z., Van Sanden, S., Burzykowski, T., and Aerts, M. (2007). Using linear mixed models for normalization of cdna microarrays. *Statistical Applications in Genetics and Molecular Biology*, 6:19.

Hao, J., Liebeke, M., Astle, W., De Iorio, M., Bundy, J. G., and Ebbels, T. M. (2014). Bayesian deconvolution and quantification of metabolites in complex 1d nmr spectra using batman. *Nature protocols*, 9(6):1416.

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome biology*, 18(1):83.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer.

Henningsen, A. and Toomet, O. (2011). MaxLik: A package for maximum likelihood estimation in R. *Computational Statistics*, 26:443–458.

Horgan, R. P. and Kenny, L. C. (2011). 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist*, 13(3):189–195.

Inouye, M., Kettunen, J., Soininen, P., Silander, K., and Ripatti, S. (2010a). Metabonomic, transcriptomic, and genomic variation of a population cohort. *Molecular Systems Biology*, 6(441).

Inouye, M., Silander, K., Hamalainen, E., Salomaa, V., and Harald, K. (2010b). An immune response network associated with blood lipid levels. *PloS Genetics*, 6(9).

Jennrich, R. I. (1970). An asymptotic $\chi 2$ test for the equality of two correlation matrices. *Journal of the American Statistical Association*, 65(330):904–912.

Joe, H. (2014). *Dependence modeling with copulas*. Chapman and Hall/CRC.

Kayano, M., Shiga, M., and Mamitsuka, H. (2014). Detecting differentially coexpressed genes from labeled expression data: A brief review. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11:154–167.

Larntz, K. and Perlman, M. D. (1985). A simple test for the equality of correlation matrices. *Rapport technique, Department of Statistics, University of Washington*, page 141.

Lewis, J. (2000). Estimating regression models in which the dependent variable is based on estimates with application to testing key's racial threat hypothesis. *Mimeograph, Princeton University*.

Lin, B., Pang, Z., and Jiang, J. (2013). Fixed and random effects selection by reml and pathwise coordinate optimization. *Journal of Computational and Graphical Statistics*, 22(2):341–355. Available from: `http://dx.doi.org/10.1080/10618600.2012.681219`.

Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D., and Bijnens, L. (2012). *Modeling dose-response microarray data in early drug development experiments using R: order-restricted analysis of microarray data*. Springer Science & Business Media.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022.

Louis, E. (2015). *Study of the plasma metabolic phenotype by NMR spectroscopy and evaluation of its ability to detect lung cancer*. PhD dissertation, Univeristeit Hasselt.

Louis, E., Adriaensens, P., Guedens, W., Bigirumurame, T., Baeten, K., Vanhove, K., Vandeurzen, K., Darquennes, K., Vansteenkiste, J., Dooms, C., et al. (2016a). Detection of lung cancer through metabolic changes measured in blood plasma. *Journal of Thoracic Oncology*, 11(4):516–523.

Louis, E., Adriaensens, P., Guedens, W., Vanhove, K., Vandeurzen, K., Darquennes, K., Vansteenkiste, J., Dooms, C., De Jonge, E., Thomeer, M., et al. (2016b). Metabolic phenotyping of human blood plasma: a powerful tool to discriminate between cancer types? *Annals of Oncology*, 27(1):178–184.

Louis, E., Bervoets, L., Reekmans, G., De Jonge, E., Mesotten, L., Thomeer, M., and Adriaensens, P. (2015). Phenotyping human blood plasma by 1h-nmr: a robust protocol based on metabolite spiking and its evaluation in breast cancer. *Metabolomics*, 11(1):225–236.

Louis, E., Cantrelle, F.-X., Mesotten, L., Reekmans, G., Bervoets, L., Vanhove, K., Thomeer, M., Lippens, G., and Adriaensens, P. (2017). Metabolic phenotyping of human plasma by 1h-nmr at high and medium magnetic field strengths: a case study for lung cancer. *Magnetic Resonance in Chemistry*, 55(8):706–713.

Macomber, R. S. (1998). *A complete introduction to modern NMR spectroscopy*. Wiley New York.

Markley, J. L., Brüschweiler, R., Edison, A. S., Eghbalnia, H. R., Powers, R., Raftery, D., and Wishart, D. S. (2017). The future of nmr-based metabolomics. *Current Opinion in Biotechnology*, 43:34–40.

Misra, B. B. and der Hooft, J. J. (2016). Updates in metabolomics tools and resources: 2014–2015. *Electrophoresis*, 37(1):86–110.

Modarres, R. and Jernigan, R. W. (1992). Testing the equality of correlation matrices. *Communications in Statistics-Theory and Methods*, 21(8):2107–2125.

Molenberghs, G., Verbeke, G., and Iddi, S. (2011). Pseudo-likelihood methodology for partitioned large and complex samples. *Statistics & Probability Letters*, 81(7):892–901.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer Science & Business Media.

Pearson, K. and Filon, L. N. G. (1898). Mathematical contributions to the theory of evolution. iv. on the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 191:229–311.

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6(3):289–296.

Pooi, A. (2003). Performance of the likelihood ratio test when fitting logistic regression models with small samples. *Communications in Statistics - Simulation and Computation*, 32(2):411–418.

Rousseau, R. (2011). *Statistical contribution to the analysis of metabonomics data in 1H NMR spectroscopy*. PhD dissertation, Université Catholique de Liège.

Schepsmeier, U. and Stöber, J. (2014). Derivatives and fisher information of bivariate copulas. *Statistical Papers*, 55(2):525–542.

Slawski, M., Daumer, M., and Boulesteix, A.-L. (2008). Cma–a comprehensive bioconductor package for supervised classification with high dimensional data. *BMC Bioinformatics*, 9(1):439.

Smyth, G. K., Yang, Y. H., and Speed, T. (2003). Statistical issues in cdna microarray data analysis. In *Functional Genomics*, pages 111–136. Springer.

Southworth, L. K., Owen, A. B., and Kim, S. K. (2009). Aging mice show a decreasing correlation of gene expression within genetic modules. *PLoS Genetics*, 5(12).

Steiger, J. H. and Hakstian, A. R. (1982). The asymptotic distribution of elements of a correlation matrix: Theory and application. *British Journal of Mathematical and Statistical Psychology*, 35(2):208–215.

Stringer, K. A., McKay, R. T., Karnovsky, A., Quémerais, B., and Lacy, P. (2016). Metabolomics and its application to acute lung diseases. *Frontiers in Immunology*, 7.

Tesson, B. M., Breitling, R., and Jansen, R. C. (2010). Diffcoex: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics*, 11(497).

Verbeke, G. and Molenberghs, G. (2011). *Linear mixed models for longitudinal data*. Springer.

Wilding, G. E., Cai, X., Hutson, A., and Yu, Z. (2011). A linear model-based test for heterogeneity of conditional correlations. *Journal of Applied Statistics*, 38(10):2355–2366.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001). Assessing gene significance from cdna microarray expression data via mixed models. *Journal of Computational Biology*, 8:625–637.

A

Appendix for Chapter 4

A.1 Plots of the gene-expression values

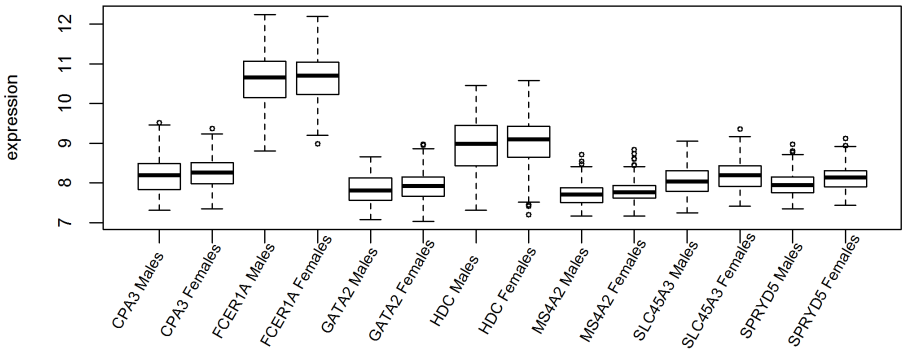

Figure A.1.1: Box plots of gene-expression values by gender.

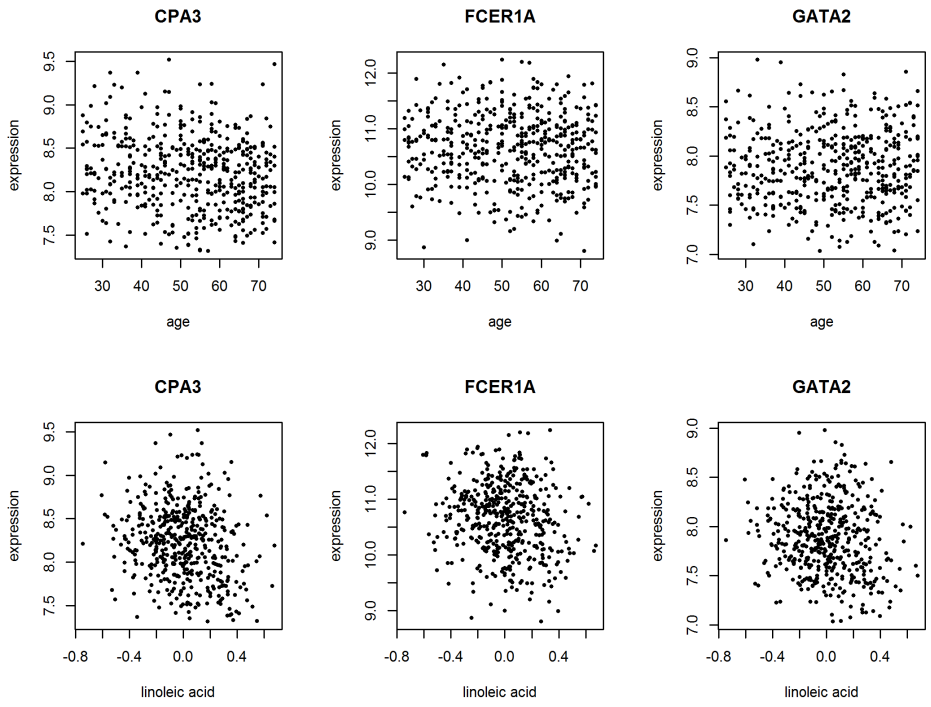

Figure A.1.2: Scatter plot of gene-expression values against age (top row) and gene-expression values against the concentration of linoleic acid (bottom row).

A.2 Design matrix $\boldsymbol{X}_{si}$

An example of the design matrix, X_{si}, for the i-th individual in the s-th subset (see Equation 4.2 of Section 4.2.3) is provided on the next page. This representation is based on a model including the covariates: gene, metabolic concentration, age, gender and the two-way interaction between gene and metabolic concentrations.

$$
X_{si} = \begin{array}{c}
\begin{array}{ccccccccccc}
\mu & g_1 & g_2 & \cdots & g_{G-1} & conc. & age & gender & g_1 * conc. & g_2 * conc. & \cdots \quad g_{G-1} * conc.
\end{array} \\
\begin{bmatrix}
1 & 1 & 0 & \cdots & 0 & 0.3 & 30 & 1 & 0.3 & 0 & \cdots \quad 0 \\
1 & 0 & 1 & \cdots & 0 & 0.3 & 30 & 1 & 0 & 0.3 & \cdots \quad 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \quad \vdots \\
1 & 0 & 0 & \cdots & 1 & 0.3 & 30 & 1 & 0 & 0 & \cdots \quad 0.3 \\
1 & 0 & 0 & \cdots & 0 & 0.3 & 30 & 1 & 0 & 0 & \cdots \quad 0
\end{bmatrix}
\end{array}
$$

A.3 SAS code

SAS procedure GLIMMIX GLM code for a gene module comprised of three genes.

```
proc GLIMMIX data = DATA;
* DATA is in long format, sorted by m_INDEX, SUBJECT and GENE;
* GE is the vector of gene expression values;
* adjMETABOLITE is the adjusted metabolic concentrations;

  ods select none;
  by m_INDEX;
  nloptions TECHNIQUE = NRRIDG;
  class SUBJECT GENE GENDER m_QUINTILE;
  model GE = AGE|GENE|adjMETABOLITE|GENDER;
  random _residual_ / subject = SUBJECT type = unr group = m_QUINTILE;

 * OPTIONAL: Input initial values using the PARMS statement to solve
    convergence problems;

  covtest "test for homogeneous correlations" general
  0 0 0 1 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ,
  0 0 0 0 1 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ,
  0 0 0 0 0 1 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ,
  0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 ,
  0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 ,
  0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 ,
  0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 ,
  0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 ,
  0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 ,
  0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 ,
  0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 ,
  0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 ,
  0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 ,
  0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 ,
  0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 ;

  ods output FitStatistics = Fit
             CovParms = Estimates
             CovTests = TestResults;
run;
```

A.4 Power of the GLM-based test statistics

Table A.4.1: Power of the GLM-based test statistics for different co-expression dynamics and sample sizes.

Co-expression	Sample size	LR test⋆	Larntz & Perlman⋆	Jennrich⋆	Cole⋆
approx. linear positive association	125	0.314 [0.285, 0.343]	0.188 [0.163, 0.213]	0.239 [0.212, 0.266]	0.374 [0.344, 0.404]
approx. linear positive association	450	0.826 [0.802, 0.850]	0.797 [0.772, 0.822]	0.711 [0.682, 0.740]	0.690 [0.661, 0.719]
approx. linear positive association	800	0.990 [0.983, 0.997]	0.989 [0.982, 0.996]	0.962 [0.950, 0.974]	0.953 [0.939, 0.967]
approx. linear negative association	125	0.300 [0.271, 0.329]	0.184 [0.159, 0.209]	0.228 [0.201, 0.255]	0.358 [0.328, 0.388]
approx. linear negative association	450	0.838 [0.815, 0.861]	0.819 [0.795, 0.843]	0.716 [0.688, 0.744]	0.693 [0.664, 0.722]
approx. linear negative association	800	0.988 [0.981, 0.995]	0.987 [0.979, 0.995]	0.960 [0.947, 0.973]	0.949 [0.935, 0.963]
non-linear association	125	0.293 [0.264, 0.322]	0.243 [0.216, 0.270]	0.240 [0.213, 0.267]	0.457 [0.426, 0.488]
non-linear association	450	0.759 [0.732, 0.786]	0.856 [0.834, 0.878]	0.744 [0.716, 0.772]	0.748 [0.721, 0.775]
non-linear association	800	0.969 [0.958, 0.980]	0.993 [0.987, 0.999]	0.963 [0.951, 0.975]	0.960 [0.947, 0.973]
non-linear association	125	0.348 [0.318, 0.378]	0.193 [0.168, 0.218]	0.228 [0.201, 0.255]	0.371 [0.341, 0.401]
non-linear association	450	0.863 [0.841, 0.885]	0.841 [0.818, 0.864]	0.693 [0.664, 0.722]	0.648 [0.618, 0.678]
non-linear association	800	0.992 [0.986, 0.998]	0.990 [0.983, 0.997]	0.956 [0.943, 0.969]	0.943 [0.928, 0.958]
weak positive association	125	0.143 [0.121, 0.165]	0.072 [0.055, 0.089]	0.131 [0.110, 0.152]	0.267 [0.239, 0.295]
weak positive association	450	0.182 [0.158, 0.206]	0.183 [0.159, 0.207]	0.192 [0.167, 0.217]	0.205 [0.179, 0.231]
weak positive association	800	0.316 [0.287, 0.345]	0.351 [0.321, 0.381]	0.342 [0.312, 0.372]	0.340 [0.310, 0.370]

Data simulated for a four-gene module.
⋆ estimate [95% confidence interval]

A.5 Plots of the GLM residuals

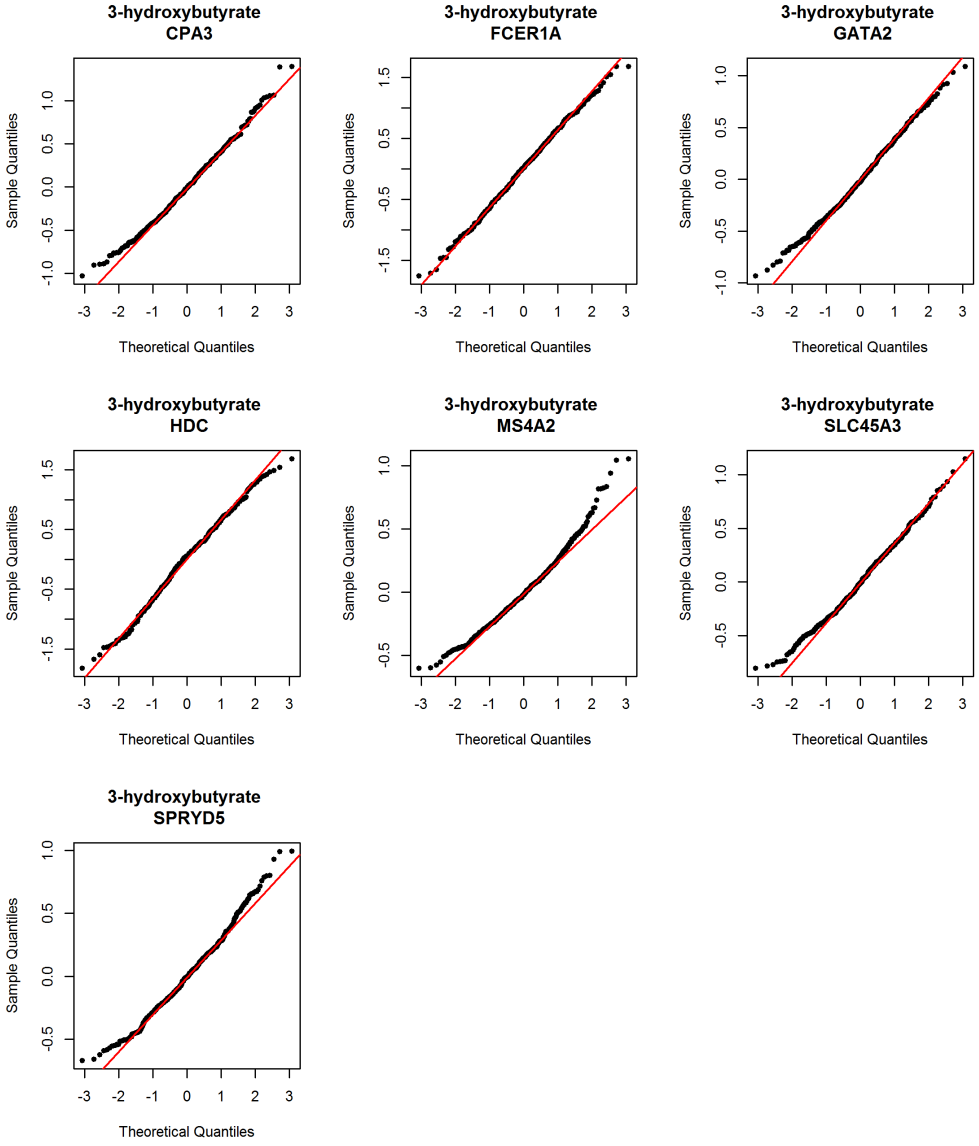

Figure A.5.3: Univariate quantile-quantile plots of the GLM residuals for 3-hydroxybutyrate.

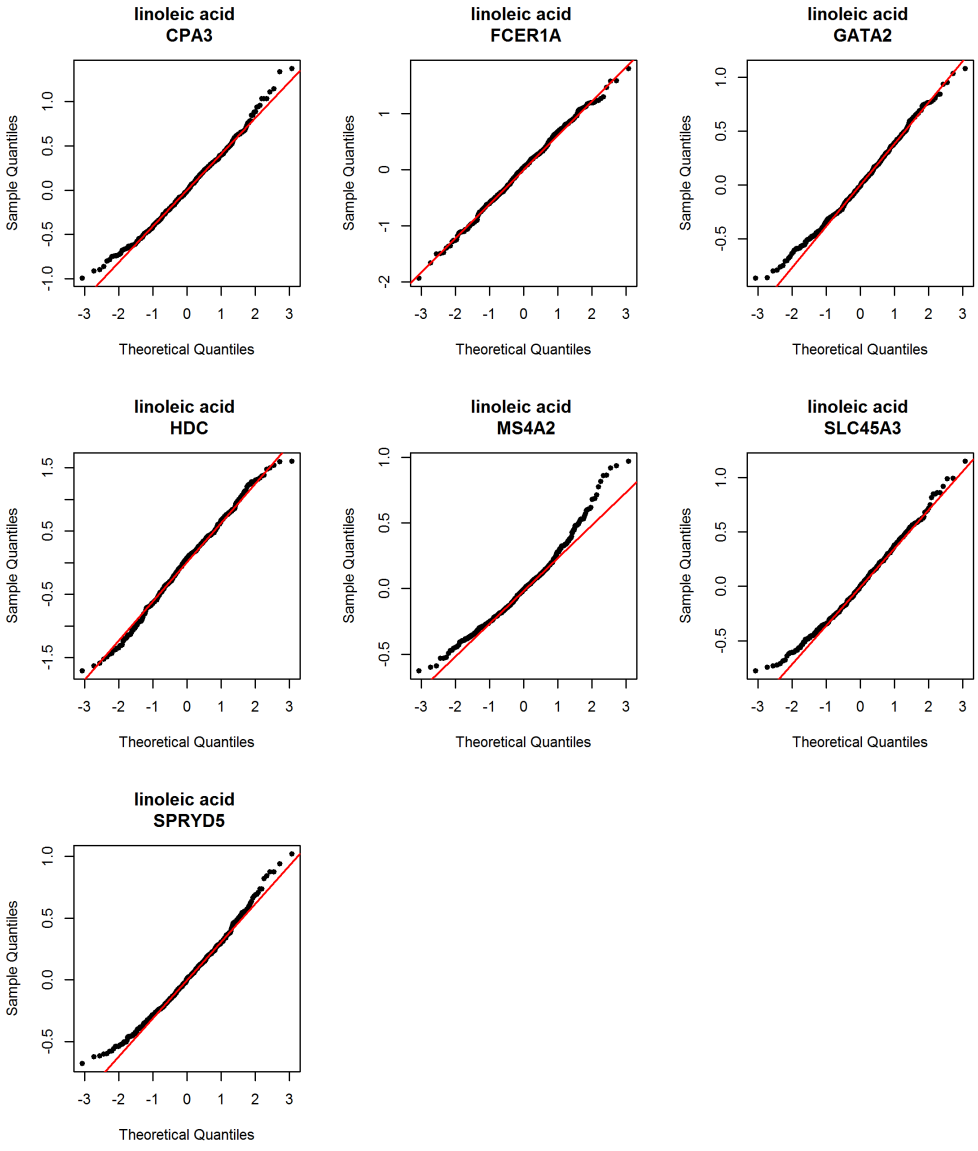

Figure A.5.4: Univariate quantile-quantile plots of the GLM residuals for linoleic acid.

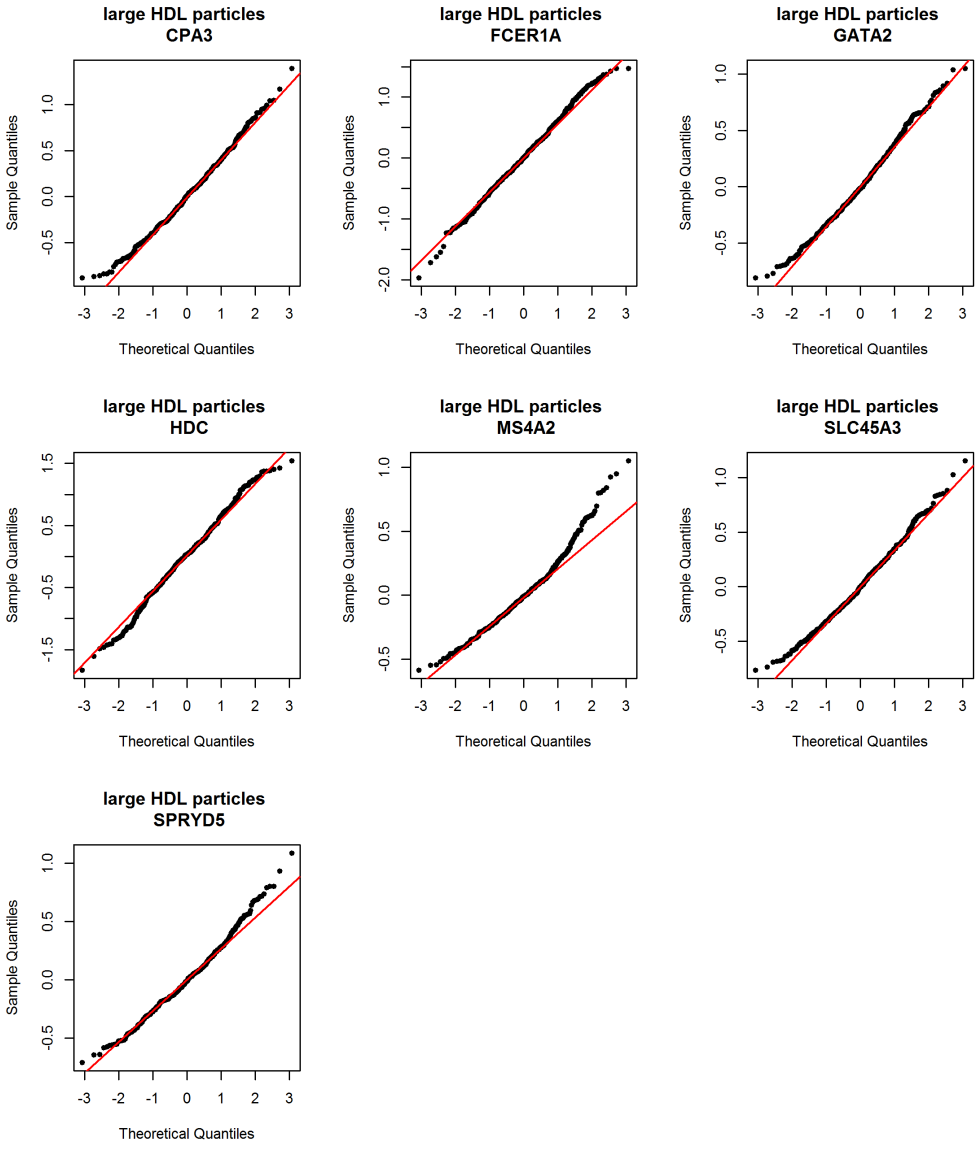

Figure A.5.5: Univariate quantile-quantile plots of the GLM residuals for large HDL particles.

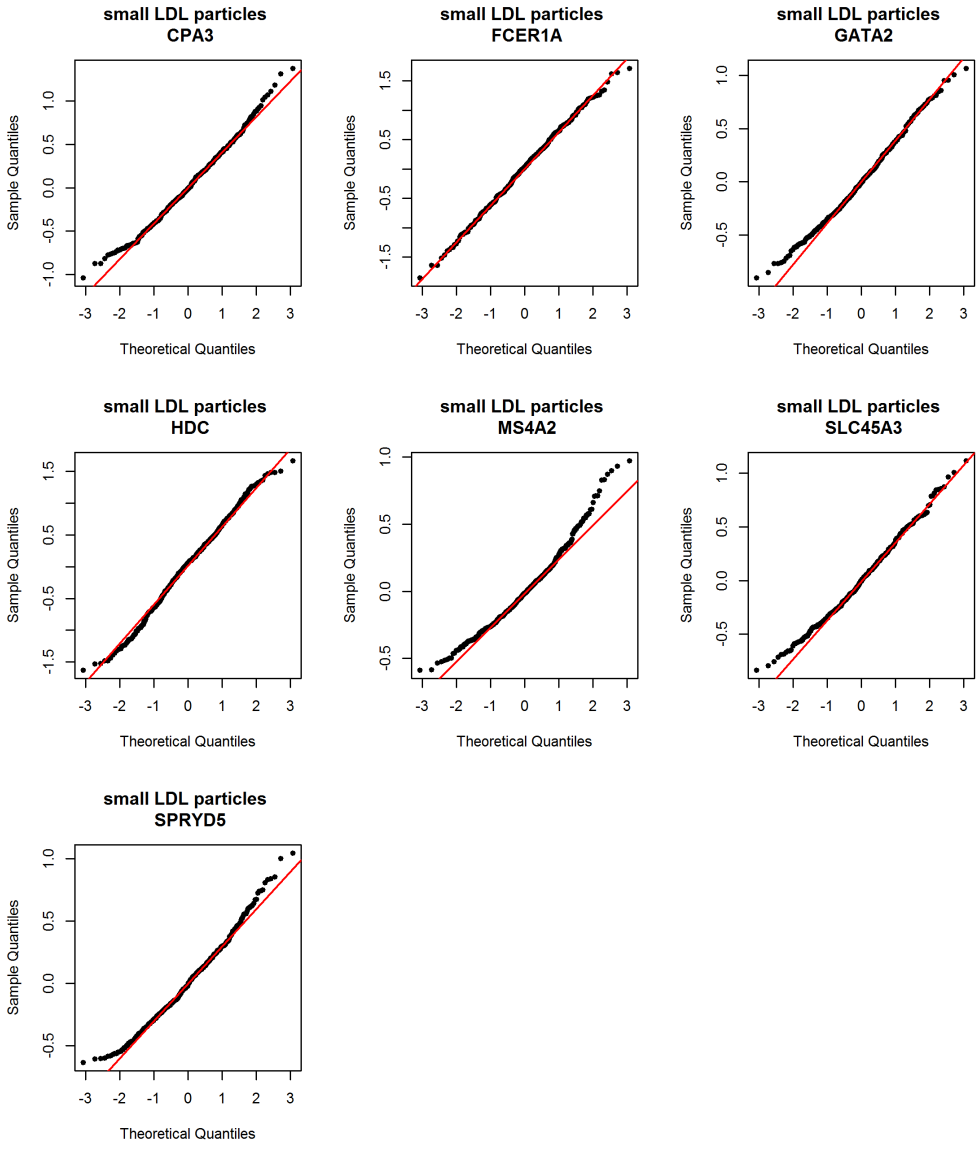

Figure A.5.6: Univariate quantile-quantile plots of the GLM residuals for small LDL particles.

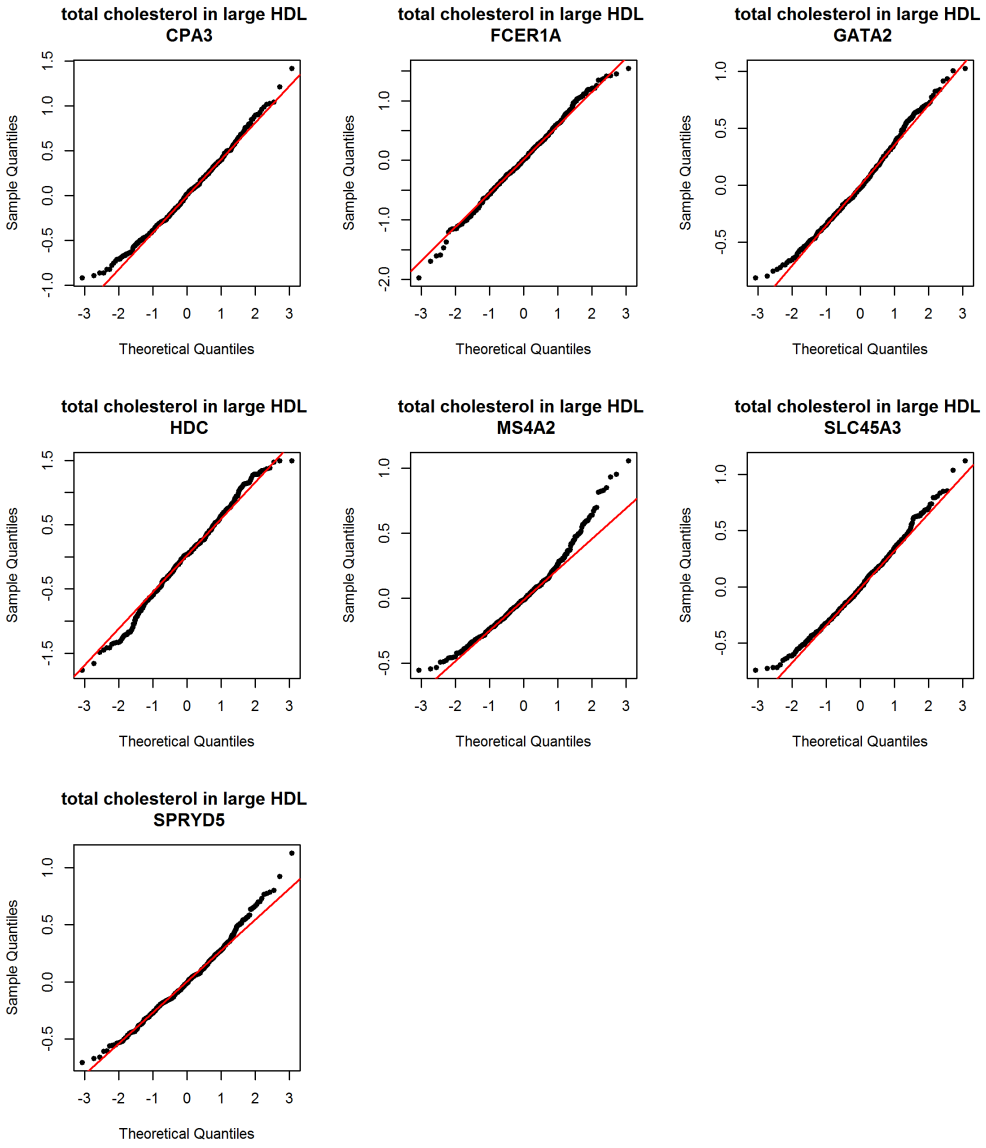

Figure A.5.7: Univariate quantile-quantile plots of the GLM residuals for total cholesterol in large HDL.

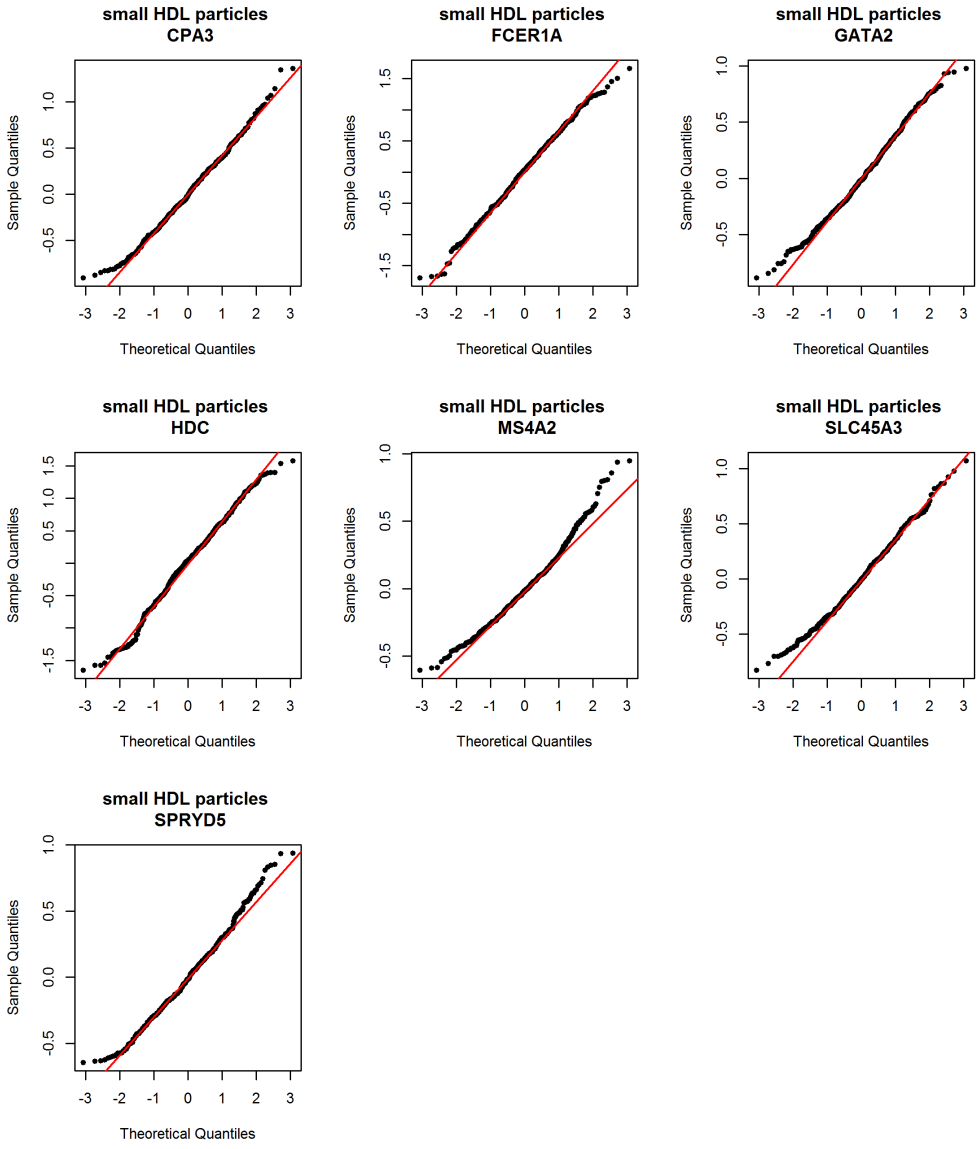

Figure A.5.8: Univariate quantile-quantile plots of the GLM residuals for small HDL particles.

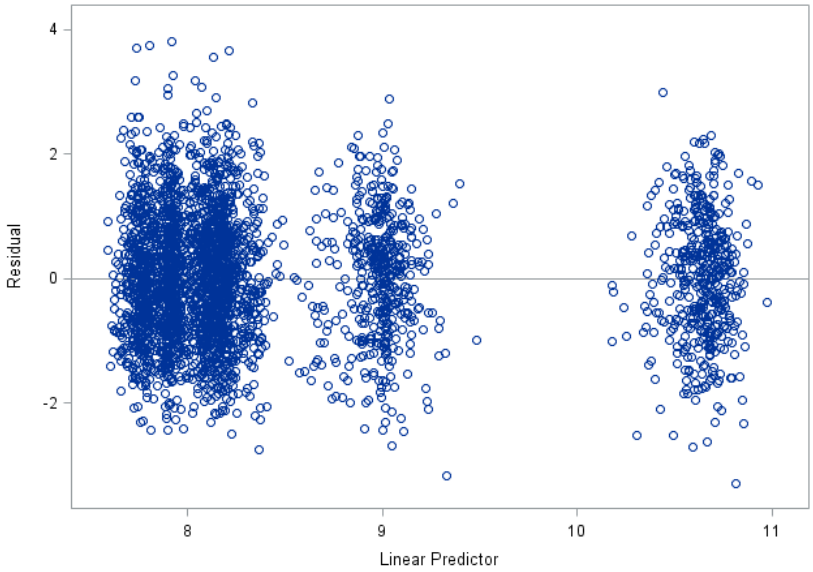

Figure A.5.9: Studentised residuals for the linoleic acid GLM.

B

Appendix for Chapter 5

B.1 Simulation study results

Table B.1.1: Simulation study results: Estimated Type I error probability and power of the Larntz & Perlman test and LR test for a seven-gene module and a sample size of 450 observations.

Co-expression dynamics	Convergence rate	Type I error / Power * Larntz & Perlman (Chapter 4)	Convergence rate	Type I error / Power * LR test (Chapter 5)
no metabolite-co-expression association	0.980	0.036 [0.024, 0.047]	0.949	0.061 [0.046, 0.076]
approx. linear negative association	0.958	0.218 [0.192, 0.244]	0.937	0.859 [0.837, 0.881]
approx. linear positive association	0.955	0.215 [0.189, 0.241]	0.952	0.863 [0.842, 0.885]
non-linear association (wave)	0.958	0.335 [0.305, 0.365]	0.951	0.365 [0.334, 0.395]
non-linear association (parabola)	0.994	0.124 [0.103, 0.144]	0.979	0.141 [0.119, 0.163]
weak non-linear association	0.968	0.031 [0.020, 0.042]	0.950	0.075 [0.058, 0.091]

$*$ point estimate [95% confidence interval]

Table B.1.2: Simulation study results: Estimated Type I error probability and power of the Larntz & Perlman test and LR test for a seven-gene module and a sample size of 450 observations.

Co-expression dynamics	Merged convergence rate	Type I error / Power * Larntz & Perlman (Chapter 4)	Type I error / Power * LR test (Chapter 5)
no metabolite-co-expression association	0.929	0.037 [0.025, 0.049]	0.060 [0.045, 0.076]
approx. linear negative association	0.897	0.219 [0.191, 0.246]	0.857 [0.834, 0.880]
approx. linear positive association	0.911	0.211 [0.184, 0.237]	0.863 [0.840, 0.885]
non-linear association (wave)	0.910	0.333 [0.302, 0.364]	0.360 [0.329, 0.392]
non-linear association (parabola)	0.973	0.125 [0.105, 0.146]	0.142 [0.120, 0.164]
weak non-linear association	0.920	0.028 [0.018, 0.039]	0.076 [0.059, 0.093]

$*$ point estimate [95% confidence interval]

C

Appendix for Chapter 7

C.1 Derivatives of the Gaussian, Gumbel-Hougaard, and Clayton copulas

Gaussian copula CDF, density, and derivatives.

The CDF of the Gaussian copula is given by

$$C(u_1, u_2|\rho) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2)|\rho).$$

The density of the Gaussian copula is given by

$$c(u_1, u_2|\rho) = \frac{1}{\sqrt{1-\rho^2}} \exp\left\{-\frac{\rho^2(x_1^2 + x_2^2) - 2\rho x_1 x_2}{2(1-p^2)}\right\},$$

where $x_1 = \Phi^{-1}(u_1)$, $x_2 = \Phi^{-1}(u_2)$, and Φ^{-1} is the inverse of CDF of a $N(0,1)$ (standard normal) variable. The first derivatives of the Gaussian copula density are provided below based on Schepsmeier and Stöber (2014). By taking the first derivative of the density with respect to ρ, we obtain

$$\frac{\partial c}{\partial \rho} = -\frac{(\rho^3 - x_1 x_2 \rho^2 + x_1^2\rho + x_2^2\rho - \rho - x_1 x_2) \exp\left\{\frac{\rho(x_1^2\rho + x_2^2\rho - 2x_1x_2)}{2(\rho-1)(\rho+1)}\right\}}{(1-\rho^2)^{\frac{5}{2}}}. \tag{C.1}$$

The first derivative with respect to u_1 is

$$\frac{\partial c}{\partial u_1} = c(u_1, u_2|\rho)\left\{-\frac{\left(2\rho^2 x_1 \frac{\partial x_1}{\partial u_1} - 2\rho x_2 \frac{\partial x_1}{\partial u_1}\right)}{2(1-\rho^2)}\right\}, \tag{C.2}$$

and,

$$\frac{\partial c}{\partial u_2} = c(u_1, u_2|\rho)\left\{-\frac{\left(2\rho^2 x_2 \frac{\partial x_2}{\partial u_2} - 2\rho x_1 \frac{\partial x_2}{\partial u_2}\right)}{2(1-\rho^2)}\right\}, \tag{C.3}$$

where,

$$\frac{\partial x_i}{\partial u_i} = \frac{\sqrt{(2\pi)}}{\exp\left\{-\Phi^{-1}(u_i)^2/2\right\}}, \qquad \text{for } i = 1, 2. \tag{C.4}$$

Gumbel-Hougaard copula CDF, density, and derivatives.

The CDF of the Gumbel-Hougaard copula is defined as

$$C(u_1, u_2|\theta) = \exp\left[-\left\{(-\ln u_1)^\theta + (-\ln u_2)^\theta\right\}^{\frac{1}{\theta}}\right].$$

Let $t_1 = (-\ln u_1)^\theta$ and $t_2 = (-\ln u_2)^\theta$, then

$$C(u_1, u_2|\theta) = \exp\left\{-(t_1 + t_2)^{\frac{1}{\theta}}\right\}.$$

The density of the Gumbel-Hougaard copula is given by

$$\begin{aligned} c(u_1, u_2|\theta) = & C(u_1, u_2|\theta)\frac{1}{u_1 u_2}\left(t_1 + t_2\right)^{-2+\frac{2}{\theta}}\left(\ln u_1 \ln u_2\right)^{\theta-1} \\ & \times\left\{1 + (\theta - 1)\left(t_1 + t_2\right)^{-\frac{1}{\theta}}\right\}. \end{aligned} \tag{C.5}$$

As report by Schepsmeier and Stöber (2014), the first derivative of the Gumbel-Hougaard copula density with respect to θ is given by

$$\begin{aligned} \frac{\partial c}{\partial \theta} = & c(u_1, u_2)\left[-(t_1 + t_2)^{\frac{1}{\theta}}\left\{-\frac{\ln(t_1 + t_2)}{\theta^2} + \frac{t_1 \ln(-\ln u_1) + t_2 \ln(-\ln u_2)}{\theta(t_1 + t_2)}\right\}\right. \\ & + \left\{-2\frac{\ln(t_1 + t_2)}{\theta^2} + \left(-2 + \frac{2}{\theta}\right)\frac{t_1 \ln(-\ln u_1) + t_2 \ln(-\ln u_2)}{t_1 + t_2}\right\} \\ & \left. + \ln(\ln u_1 \ln u_2)\right] \\ & + C(u_1, u_2)(t_1 + t_2)^{-2+\frac{2}{\theta}}\frac{(\ln u_1 \ln u_2)^{\theta-1}}{u_1 u_2}\left[(t_1 + t_2)^{-\frac{1}{\theta}}(\theta - 1)(t_1 + t_2)^{-\frac{1}{\theta}}\right. \\ & \left.\times\left\{\frac{\ln(t_1 + t_2)}{\theta^2} - \frac{t_1 \ln(-\ln u_1) + t_2 \ln(-\ln u_2)}{\theta(t_1 + t_2)}\right\}\right]. \end{aligned} \tag{C.6}$$

The derivative with respect to u_1 is

$$\frac{\partial c}{\partial u_1} = c(u_1,u_2)\left\{-(t_1+t_2)^{\frac{1}{\theta}-1}\frac{t_1}{u_1\ln u_1} - \frac{1}{u_1} + (t_1+t_2)^{-1}\frac{(-2+\frac{2}{\theta})t_1\theta}{u_1\ln u_1} + \frac{(\theta-1)}{u_1\ln u_1}\right\}$$
$$- C(u_1,u_2)(t_1+t_2)^{-2+\frac{2}{\theta}}\frac{(\ln u_1\ln u_2)^{\theta-1}}{u_1u_2}(\theta-1)(t_1+t_2)^{-\frac{1}{\theta}-1}\frac{t_1}{u_1\ln u_1}. \tag{C.7}$$

The derivative with respect to u_2 is

$$\frac{\partial c}{\partial u_2} = c(u_1,u_2)\left\{-(t_1+t_2)^{\frac{1}{\theta}-1}\frac{t_2}{u_2\ln u_2} - \frac{1}{u_2} + (t_1+t_2)^{-1}\frac{(-2+\frac{2}{\theta})t_2\theta}{u_2\ln u_2} + \frac{(\theta-1)}{u_2\ln u_2}\right\}$$
$$- C(u_1,u_2)(t_1+t_2)^{-2+\frac{2}{\theta}}\frac{(\ln u_1\ln u_2)^{\theta-1}}{u_1u_2}(\theta-1)(t_1+t_2)^{-\frac{1}{\theta}-1}\frac{t_2}{u_2\ln u_2}. \tag{C.8}$$

Note that the latter two derivatives of the Gumbel-Hougaard copula density differ to those reported by Schepsmeier and Stöber (2014).

Clayton copula CDF, density, and derivatives.

The CDF of the Clayton copula is defined as

$$C(u_1,u_2|\theta) = (u_1^{-\theta}+u_2^{-\theta}-1)^{-\frac{1}{\theta}}.$$

The density of the Clayton is given by

$$c(u_1,u_2|\theta) = \frac{(1+\theta)(u_1u_2)^{-1-\theta}}{(u_1^{-\theta}+u_2^{-\theta}-1)^{\frac{1}{\theta}+2}}. \tag{C.9}$$

The first derivative of the Clayton copula density with respect to θ is given by

$$\frac{\partial c}{\partial\theta} = \frac{(u_1u_2)^{-1-\theta}}{(u_1^{-\theta}+u_2^{-\theta}-1)^{\frac{1}{\theta}+2}} - c(u_1,u_2)\left[\ln(u_1u_2) - \left\{\frac{\ln(u_1^{-\theta}+u_2^{-\theta}-1)}{\theta^2}\right.\right.$$
$$\left.\left. + \frac{(-2-\frac{1}{\theta})(-u_1^{-\theta}\ln(u_1)-u_2^{-\theta}\ln(u_2))}{(u_1^{-\theta}+u_2^{-\theta}-1)}\right\}\right]. \tag{C.10}$$

As report by Schepsmeier and Stöber (2014), the derivative with respect to u_1 is

$$\frac{\partial c}{\partial u_1} = -\frac{c(u_1,u_2)(\theta+1)}{u_1} + \frac{c(u_1,u_2)(2+\frac{1}{\theta})\theta}{u_1^{\theta+1}(u_1^{-\theta}+u_2^{-\theta}-1)}. \tag{C.11}$$

C

The derivative with respect to u_2 is

$$\frac{\partial c}{\partial u_2} = -\frac{c(u_1,u_2)(\theta+1)}{u_2} + \frac{c(u_1,u_2)(2+\frac{1}{\theta})\theta}{u_2^{\theta+1}(u_1^{-\theta}+u_2^{-\theta}-1)}. \tag{C.12}$$

C.2 Simulation study results for the adjusted PLR test statistics

Table C.2.1: Simulation study results: Estimated power of the PLR and adjusted PLR tests for a seven-gene module and a sample size of 450 observations.

Co-expression dynamics	Power*		
	PLR test: $G^{*2}(H_0)$	adjusted PLR test: $G_a^{*2}(H_1)^a$	adjusted PLR test: $G_a^{*2}(H_0)^a$
approx. linear negative association	0.984 [0.976, 0.992]	0.987 [0.980, 0.994]	0.987 [0.980, 0.994]
approx. linear positive association	0.988 [0.981, 0.995]	0.990 [0.984, 0.996]	0.989 [0.983, 0.995]
non-linear association (wave)	0.580 [0.549, 0.611]	0.558 [0.527, 0.589]	0.600 [0.570, 0.630]
non-linear association (parabola)	0.044 [0.031, 0.057]	0.064 [0.049, 0.079]	0.055 [0.041, 0.069]
weak non-linear association	0.063 [0.048, 0.078]	0.065 [0.050, 0.080]	0.066 [0.051, 0.081]

* point estimate [95% confidence interval]

[a] using the empirical distribution of the test statistic.

Table C.2.2: Simulation study results: Estimated power of the PLR test, the adjusted PLR tests, and the LR test for a seven-gene module and a sample size of 450 observations.

Co-expression dynamics	MVN convergence	Type I error or power*			
		PLR test: $G^{*2}(H_0)$	adjusted PLR test: $G_a^{*2}(H_1)^a$	adjusted PLR test: $G_a^{*2}(H_0)^a$	LR test
approx. linear negative association	912	0.985 [0.977, 0.993]	0.988 [0.980, 0.994]	0.988 [0.980, 0.994]	0.863 [0.841, 0.885]
approx. linear positive association	923	0.988 [0.981, 0.995]	0.990 [0.984, 0.996]	0.989 [0.983, 0.995]	0.872 [0.851, 0.894]
non-linear association (wave)	922	0.573 [0.541, 0.605]	0.561 [0.527, 0.589]	0.604 [0.570, 0.630]	0.374 [0.343, 0.405]
non-linear association (parabola)	966	0.043 [0.031, 0.056]	0.064 [0.049, 0.079]	0.055 [0.041, 0.069]	0.142 [0.120, 0.164]
weak non-linear association	924	0.056 [0.041, 0.071]	0.067 [0.050, 0.080]	0.068 [0.051, 0.081]	0.070 [0.054, 0.087]

* point estimate [95% confidence interval]

[a] using the empirical distribution of the test statistic.

C.3 Empirical CDF plots

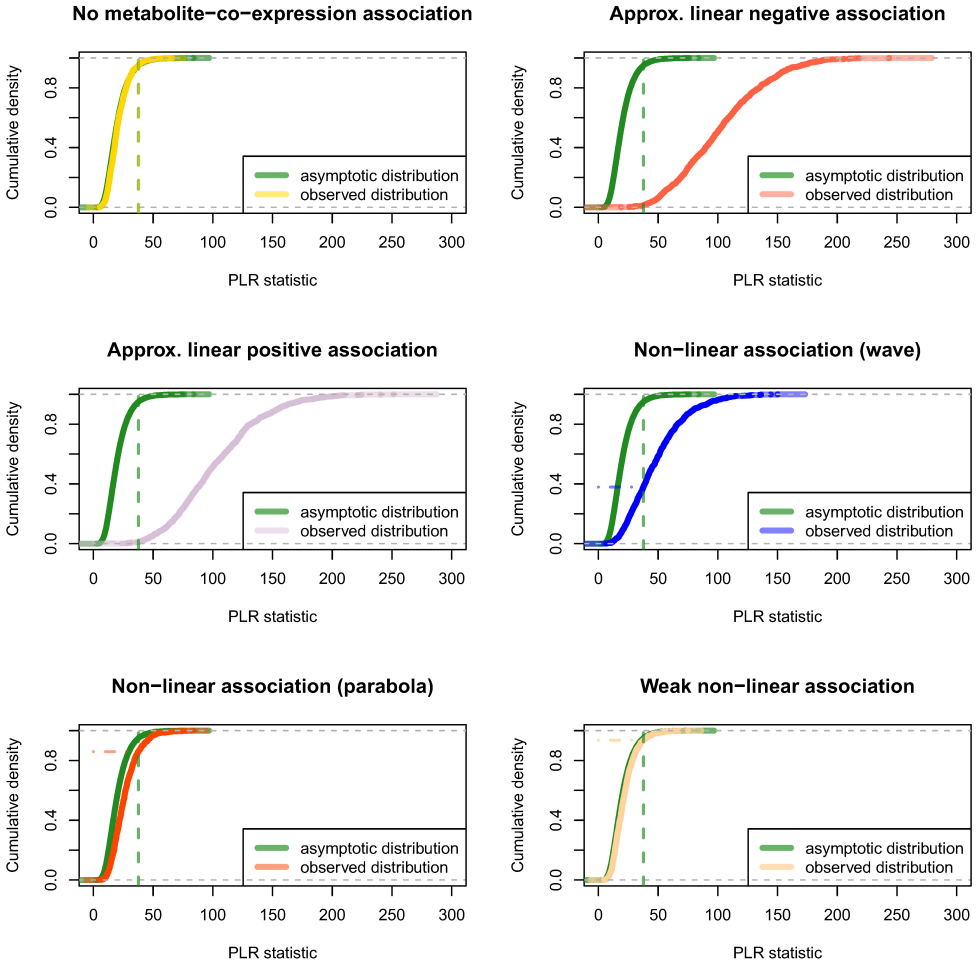

Figure C.3.1: Empirical CDF of the observed PLR test statistics for each of the six co-expression dynamics together with the asymptotic CDF of the PLR test statistic. In each plot, the asymptotic CDF of the PLR test statistic is shown in green. Dashed lines indicate the critical value of the asymptotic distribution (green) and the observed distribution (yellow) on the x-axis. Dot-dashed lines indicate the cumulative density of the empirical CDFs at the asymptotic critical value.

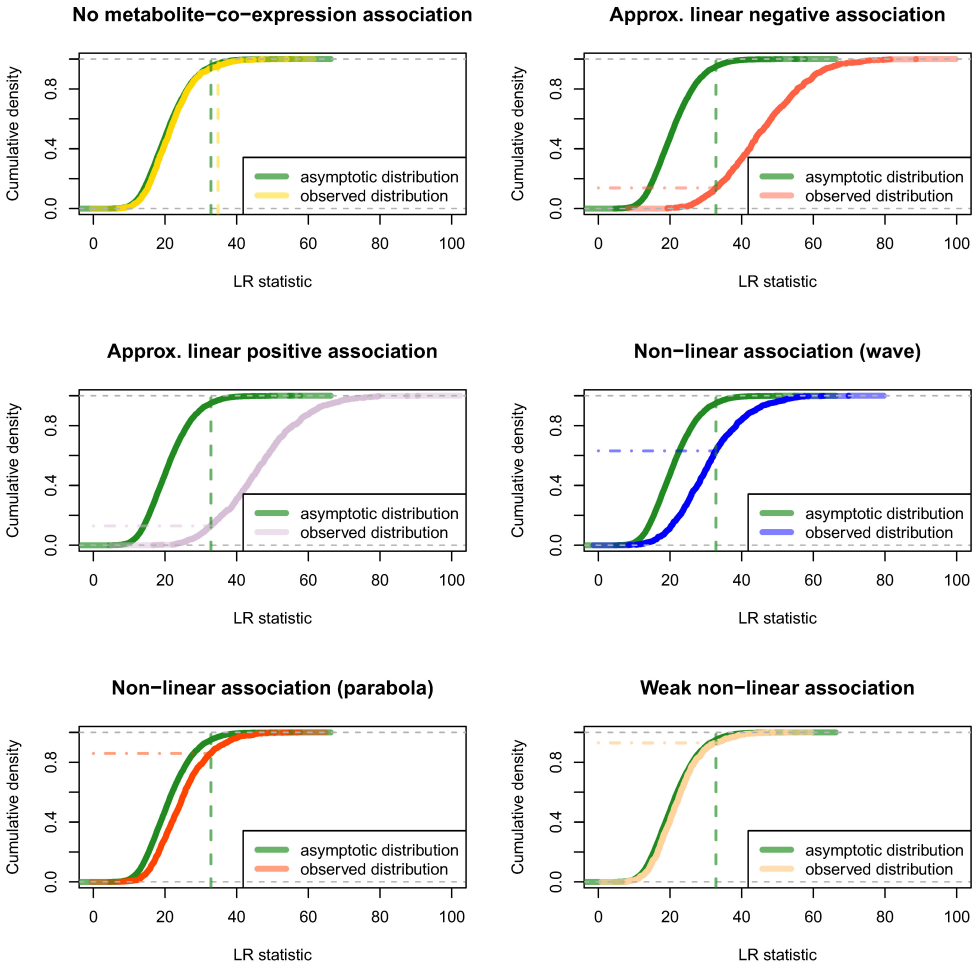

Figure C.3.2: Empirical CDF of the observed LR test statistics for each of the six co-expression dynamics together with the asymptotic CDF of the LR test statistic. In each plot, the asymptotic CDF of the PLR test statistic is shown in green. Dashed lines indicate the critical value of the asymptotic distribution (green) and the observed distribution (yellow) on the x-axis. Dot-dashed lines indicate the cumulative density of the empirical CDFs at the asymptotic critical value.

D

Appendix for Chapter 12

D.1 Classification results

Table D.1.1: Classification results.

Features	Misclassification	Sensitivity	Specificity
Elastic net			
400 MHz (manually pre-processed data)			
Binning: top 50 integration regions	0.125 (0.002)	0.844 (0.003)	0.904 (0.002)
BATMAN: all metabolites	0.336 (0.002)	0.634 (0.003)	0.691 (0.003)
BATMAN: lipids	0.260 (0.002)	0.705 (0.003)	0.772 (0.003)
BATMAN: all metabolites and lipids	0.197 (0.002)	0.775 (0.003)	0.829 (0.002)
900 MHz (`PepsNMR` automatically pre-processed data)			
Binning: top 90 integration regions	0.197 (0.002)	0.779 (0.003)	0.826 (0.003)
BATMAN: all metabolites	0.105 (0.001)	0.884 (0.002)	0.906 (0.002)
BATMAN: lipids	0.323 (0.002)	0.611 (0.003)	0.738 (0.003)
BATMAN: all metabolites and lipids	0.111 (0.001)	0.874 (0.002)	0.902 (0.002)
900 MHz (manually pre-processed data)			
Binning: top 45 integration regions	0.170 (0.002)	0.813 (0.003)	0.846 (0.002)
Lasso			
400 MHz (manually pre-processed data)			
Binning: top 45 integration regions	0.136 (0.002)	0.825 (0.003)	0.901 (0.002)
BATMAN: all metabolites	0.333 (0.002)	0.632 (0.003)	0.700 (0.003)
BATMAN: lipids	0.261 (0.002)	0.704 (0.003)	0.771 (0.003)
BATMAN: all metabolites and lipids	0.197 (0.002)	0.768 (0.003)	0.834 (0.002)
900 MHz (`PepsNMR` automatically pre-processed data)			
Binning: top 90 integration regions	0.206 (0.002)	0.764 (0.003)	0.821 (0.003)
BATMAN: all metabolites	0.112 (0.001)	0.877 (0.002)	0.899 (0.002)
BATMAN: lipids	0.324 (0.002)	0.612 (0.003)	0.737 (0.003)
BATMAN: all metabolites and lipids	0.122 (0.001)	0.862 (0.002)	0.892 (0.002)
900 MHz (manually pre-processed data)			
Binning: top 45 integration regions	0.170 (0.002)	0.799 (0.003)	0.859 (0.002)
Orthogonal Partial Least Squares - Discriminant Analysis			
400 MHz (manually pre-processed data)			
Binning: top 60 integration regions	0.141 (0.002)	0.768 (0.003)	0.943 (0.002)

Features	Misclassification	Sensitivity	Specificity
BATMAN: all metabolites	0.307 (0.002)	0.650 (0.003)	0.732 (0.003)
BATMAN: lipids	0.225 (0.002)	0.697 (0.003)	0.848 (0.002)
BATMAN: all metabolites and lipids	0.217 (0.002)	0.725 (0.003)	0.838 (0.003)
900 MHz (`PepsNMR` automatically pre-processed data)			
Binning: top 90 integration regions	0.193 (0.002)	0.746 (0.003)	0.863 (0.002)
BATMAN: all metabolites	0.109 (0.001)	0.883 (0.002)	0.898 (0.002)
BATMAN: lipids	0.253 (0.002)	0.714 (0.003)	0.778 (0.002)
BATMAN: all metabolites and lipids	0.103 (0.001)	0.891 (0.002)	0.903 (0.002)
900 MHz (manually pre-processed data)			
Binning: top 60 integration regions	0.199 (0.002)	0.697 (0.003)	0.897 (0.002)
Random forest			
400 MHz (manually pre-processed data)			
Binning: top 60 integration regions	0.160 (0.002)	0.805 (0.003)	0.872 (0.002)
BATMAN: all metabolites	0.362 (0.002)	0.584 (0.003)	0.689 (0.003)
BATMAN: lipids	0.279 (0.002)	0.689 (0.003)	0.750 (0.003)
BATMAN: all metabolites and lipids	0.265 (0.002)	0.703 (0.003)	0.765 (0.003)
900 MHz (`PepsNMR` automatically pre-processed data)			
Binning: top 80 integration regions	0.222 (0.002)	0.746 (0.003)	0.807 (0.003)
BATMAN: all metabolites	0.124 (0.001)	0.858 (0.002)	0.893 (0.002)
BATMAN: lipids	0.284 (0.002)	0.691 (0.003)	0.739 (0.003)
BATMAN: all metabolites and lipids	0.131 (0.002)	0.855 (0.002)	0.881 (0.002)
900 MHz (manually pre-processed data)			
Binning: top 45 integration regions	0.175 (0.002)	0.773 (0.003)	0.874 (0.002)
Support vector machines			
400 MHz (manually pre-processed data)			
Binning: top 90 integration regions	0.137 (0.001)	0.838 (0.003)	0.887 (0.002)
BATMAN: all metabolites	0.353 (0.002)	0.548 (0.005)	0.739 (0.004)
BATMAN: lipids	0.225 (0.002)	0.713 (0.003)	0.833 (0.002)
BATMAN: all metabolites and lipids	0.226 (0.002)	0.735 (0.003)	0.811 (0.003)
900 MHz (`PepsNMR` automatically pre-processed data)			
Binning: top 80 integration regions	0.213 (0.002)	0.759 (0.003)	0.813 (0.003)
BATMAN: all metabolites	0.142 (0.001)	0.850 (0.002)	0.866 (0.002)
BATMAN: lipids	0.246 (0.002)	0.699 (0.003)	0.806 (0.003)
BATMAN: all metabolites and lipids	0.117 (0.001)	0.879 (0.002)	0.887 (0.002)
900 MHz (manually pre-processed data)			
Binning: top 80 integration regions	0.177 (0.002)	0.785 (0.003)	0.858 (0.003)

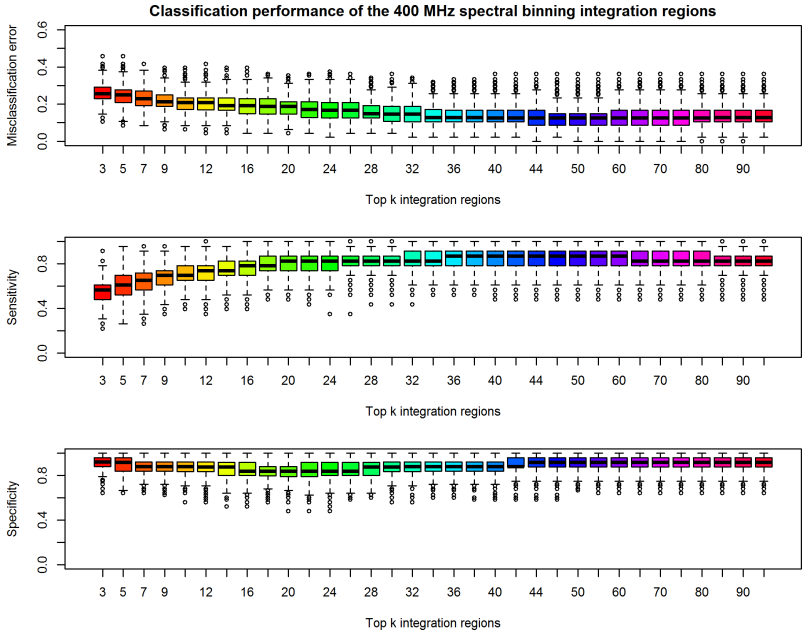

Figure D.1.1: Classification performance of the elastic net models utilizing the top k 400 MHz spectral binning integration regions.

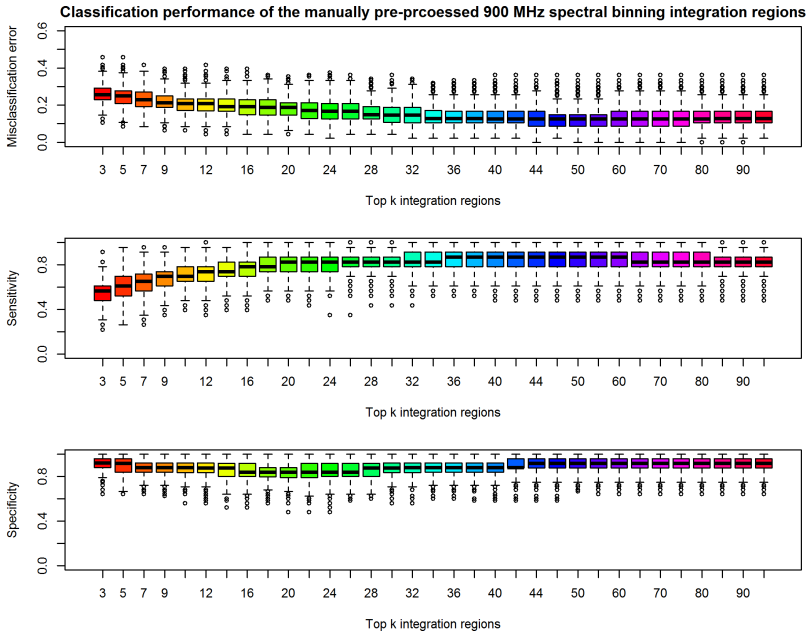

Figure D.1.2: Classification performance of the elastic net models utilizing the top k PepsNMR pre-processed 900 MHz spectral binning integration regions.

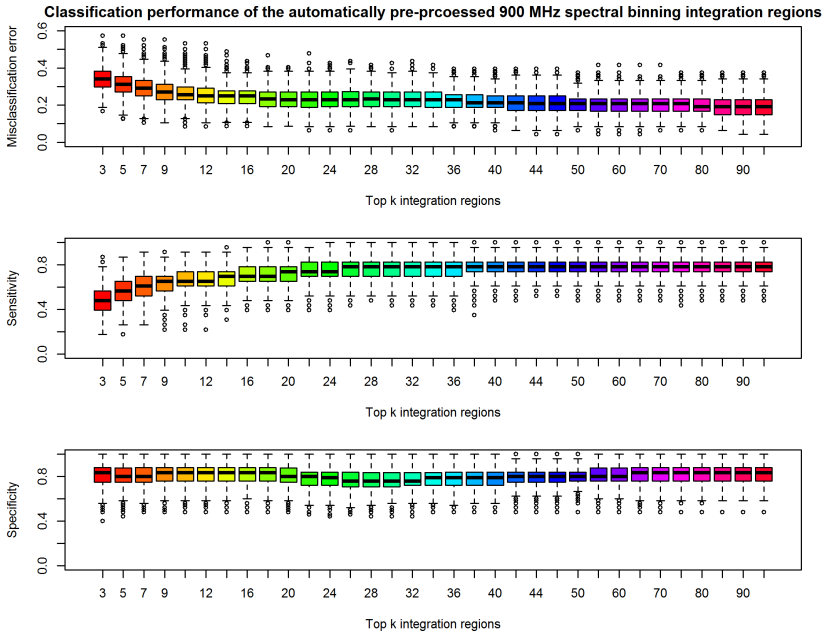

Figure D.1.3: Classification performance of the elastic net models utilizing the top k manually pre-processed 900 MHz integration regions.

Summary

Omics technologies have rapidly advanced giving rise to an extensive amount of omics data with widespread availability. The analysis of omics data can lead to the identification of molecular profiles that are associated with disease status, susceptibility, or progression, or it may provide insight into biological pathways or processes that differ in diseased and control patients. Biological processes are, however, extremely intricate and obtaining biologically meaningful information from this mass of data is a non-trivial task. To capture the complexity of biological processes, research is now centering on the integrative analysis of omics data. However, methodological development in this area is lacking. As a result, complex data is analysed in rather simple ways that fail to capture the complexity of the biological problem. The research presented in Part I of this dissertation aims to improve on currently implemented methods for the integrative analysis of omics datasets.

A way to enhance our understanding of the development and progression of complex diseases is to investigate the influence of cellular environments on gene co-expression (i.e., gene-pair correlations). Investigating whether metabolites regulate the co-expression of a predefined gene module (a set of co-expressed (correlated) genes belonging to the same biological pathway) is one of the relevant questions posed in the integrative analysis of metabolomic and transcriptomic data (Inouye et al., 2010a). In Part I of this dissertation, three statistical models are described for investigating the association between gene-module co-expression and metabolite concentrations. The suitability and versatility of the proposed models are investigated through simulation studies and an application to real-life data. Specifically, a subset of the DILGOM (**DI**etary, **L**ifestyle, and **G**enetic determinants of **O**besity and **M**etabolic syndrome) study data (Inouye et al., 2010a) is analysed.

Part I of the dissertation begins with a description of a simple linear regression (SLR) approach that has been previously implemented for the investigation of conditional co-expression (Inouye et al., 2010a). Attention is drawn to several limitations of the approach. As an alternative, a multivariate linear model for studying the dependence between categorised metabolite concentrations and gene-module co-expression is proposed in Chapter 4. The approach addresses the limitations of the linear-regression-based analysis. Through a simulation study it is shown that the SLR approach suffers from a highly inflated type I error probability and that the proposed multivariate model is less prone to the detection of spurious conditional correlations.

Often, changes in gene co-expression are investigated across two or more biological conditions defined by categorising a continuous covariate. However, the selection of arbitrary cut-off points may have an influence on the results of an analysis. To

address this issue, in Chapter 5, a multivariate linear model for investigating the association between gene-module co-expression and a continuous covariate is proposed.

Fitting a multivariate model that fully captures the dependence structure of several variables can become increasingly challenging as the number of parameters and the size of the variance-covariance matrix increases. Chapter 7 provides a more computationally feasible solution for investigating the conditional co-expression of a gene-module. In particular, a copula-based pseudo-likelihood approach is proposed. The multivariate density described in Chapter 5 is replaced by a pseudo-likelihood function formed by the product of all pairwise densities over the set of all possible gene pairs within the gene module. Furthermore, bivariate densities are modeled using Gaussian, Gumbel-Hougaard, and Clayton copulas that specify the gene-pair correlations as a function of the metabolite concentrations. In addition to reducing the computation burden, this approach facilitates the estimation of other non-parametric measures of association such as Kendall's tau and Spearman's rho.

High-throughput techniques enable the measurement of the chemical composition of cells, tissues, or, biofluids. The reproducibility, precision, and inherent noise of the measurements vary between techniques. In some instances, the biological signal may constitute only a small portion of the collected measurements. Efficient extraction of the biological signal is required before the data can be analysed. A variety of approaches exist to extract biological signal. The adopted approach can have an impact on downstream analyses. In Part II of this dissertation, the impact of the method for extracting metabolic signal from proton nuclear magnetic resonance (^{1}H-NMR) data on the classification of lung cancer samples is studied.

Extracting metabolic information from NMR spectra is complex due to the fact that an immense amount of detail on the chemical composition of a biological sample is expressed through a single spectrum. The simplest approach to quantify the signal is through spectral binning which involves subdividing the spectra into regions along the chemical shift axis and integrating the peaks within each region (Louis et al., 2015). However, due to overlapping resonance signals, the integration values do not always correspond to the concentrations of specific metabolites. An alternate, more advanced statistical approach is spectral deconvolution. BATMAN (**B**ayesian **A**u**T**omated **M**etabolite **A**nalyser for **N**MR data) (Astle et al., 2012; Hao et al., 2014) performs spectral deconvolution using prior information on the spectral signatures of metabolites. In this way, BATMAN estimates relative metabolic concentrations. Both spectral binning and spectral deconvolution using BATMAN were applied to 400 MHz and 900 MHz NMR spectra of blood plasma samples from lung cancer patients and control subjects (Chapter 11). The relative concentrations estimated by BATMAN were compared with the binning integration values in terms of their ability to discriminate between lung cancer patients and controls (Chapter 12). For the 400 MHz

data, the spectral binning approach provided greater discriminatory power. However, for the 900 MHz data, the relative metabolic concentrations obtained by using BATMAN provided greater predictive power. While spectral binning is computationally advantageous and less laborious, BATMAN estimated features correspond directly with specific metabolites and therefore have a simpler interpretation.

Samenvatting

'Omics' technologie is bezig aan een sterke opmars, waardoor er een grote stijging in hoeveelheid en beschikbaarheid van deze omics data is. Het analyseren van deze 'omics' data kan leiden tot de identificatie van moleculaire profielen die geassocieerd worden met eigenschapen van ziektes zoals status, ontvankelijkheid, vooruitgang, maar het kan ook inzicht verschaffen in de biologische methodes en processen die verschillen tussen de zieke en controle patiënten. Maar biologische processen zijn extreem ingewikkeld en het verkrijgen van zinvolle biologische informatie uit deze grote hoeveelheid data is geen sinecure. Om de complexiteit van deze biologische processen te vatten, richt onderzoek zich momenteel op het gezamenlijk analyzeren van verschillende bronnen van omics data. Desondanks blijft de methodologische ontwikkeling hier achter. Hierdoor wordt complexe data geanalyseerd op een relatief simplistische wijze die de complexiteit van het biologisch probleem niet vatten. Het onderzoek voorgesteld in deel I van dit proefschrift richt zich op het verbeteren van de huidige geïmplementeerde methodes voor de integratieve analyse van omics gegevens.

Een manier om ons begrip omtrent de ontwikkeling en de vooruitgang van complexe ziektes te vergroten, is om de invloed te onderzoeken die een cellulaire omgeving heeft op de gene co-expressies (b.v.; gene-paar correlaties). Eén van de belangrijke vragen die gesteld werden in de gezamenlijke analyse van metaboloom en transcriptoom data (Inouye et al., 2010a), is het onderzoek naar het regulerend effect van metabolieten op de co-expressie van vooraf bepaalde gen modules (een paar genen die een co-expressie (correlatie) hebben en behoren tot het zelfde biologisch mechanisme). In het eerste deel van deze thesis worden er drie statistische modellen beschreven voor het onderzoek naar de associatie tussen gene-module co-expressie en metaboliet concentraties. De geschiktheid en de veelzijdigheid van de voorgestelde modellen zijn onderzocht met behulp van gesimuleerde data en reële data. Meer specifiek is er een gedeelte van de DILGOM (**DI**eet, **L**evensstijl en **G**enetische determinant van **O**besity en **M**etabolisch syndroom) studie data (Inouye et al., 2010a) gebruikt.

Deel I van de verhandeling begint met de beschrijving van een eenvoudige lineaire regressie (ELR), die voorheen gebruikt werd bij het onderzoek naar voorwaardelijke co-expressie (Inouye et al., 2010a). Hier wordt de aandacht gevestigd op de vele beperkingen van deze methode. Als alternatief wordt, in Hoofdstuk 4, een multivariaat lineair model voor de studie van afhankelijkheid tussen metaboliet concentraties (verdeeld in categorien) en gen-module co-expressie voorgesteld. Deze aanpak geeft een antwoord op de beperkingen van de lineair regressie analyse. Door middel van een simulatie studie is er aangetoond dat de ELR een serieus verhoogde kans op type I fouten heeft, en dat het multivariate model minder gevoelig is voor het waarnemen

van onechte voorwaardelijke correlaties.

Vaak worden veranderingen in gen co-expressies onderzocht over twee of meerdere biologische condities, die gedefinieerd worden door het verdelen van een continue variabele. Echter de keuze van deze arbitraire categorieën, en meer bepaald het begin en eindpunt, kan een invloed hebben op het resultaat van de analyse. Om dit probleem aan te kaarten wordt, in Hoofdstuk 5, een multivariaat lineair model voorgesteld voor het onderzoek naar het verband tussen gene-module co-expressie en continue covariaat.

Het schatten van een multivariaat model dat de afhankelijkheid van verschillende variabelen beschrijft, kan zeer complex worden wanneer het aantal parameters en de grote van de variatie-covariatie matrix toeneemt. Hoofdstuk 7 introduceert een computationeel meer haalbare oplossing voor het nagaan van voorwaardelijke co-expressie van een gene-module. Meer specifiek wordt er een pseudo-likelihood aanpak gebaseerd op copulas voorgesteld. De multivariate dichtheidsfunctie, beschreven in Hoofdstuk 5, is vervangen door een pseudo-likelihood functie die het product is van alle paar-gewijze dichtheden over de set van alle mogelijke gene paren binnen een gene module. Verder zijn de bivariate dichtheidsfuncties, gemodeleerd met behulp van Gaussian, Gumbel-Hougaard en Clayton copulas die de correlatie tussen twee genen verduidelijken als een functie van de metaboliet concentraties. Bovenop het verminderen van de computationele belasting, vergemakkelijkt deze methode de schatting van niet-parametrische associatie-maten zoals Kendall's tau en Spearman's rho.

High-throughput technieken maken het meten van de chemische samenstelling van cellen, weefsels of vloeistoffen mogelijk. De herhaalbaarheid, nauwkeurigheid en de inherente ruis van de metingen verschillen van techniek tot techniek. In bepaalde omstandigheden, bestaat het biologisch signaal slechts uit een klein deel van de verzamelde data. Een efficiënte extractie van het biologisch signaal is een noodzaak vooraleer de data geanalyseerd kan worden. Verscheidene benaderingen bestaan om het biologisch signaal eruit te halen. De gekozen benadering kan een invloed hebben op de verdere analyse. In deel II van deze thesis wordt de invloed van de methode voor het extraheren van het metabolische signaal van proton nucleair magnetisch resonantie (H-NMR) data op de classificatie van longkanker stalen bestudeerd.

Extractie van metabolische informatie van het NMR spectrum is complex doordat een gigantische hoeveelheid aan details over de chemische samenstelling van een biologisch monster in één enkel spectrum aanwezig is. De meest eenvoudige aanpak om het biologische signaal te bepalen is *spectral binning* wat inhoud dat het spectrum onderverdeeld wordt in zones langs de chemische verschuiving as, en het integreren van de pieken binnen elke zone (Louis et al., 2015). Echter door overlapping van de resonantie-signalen, komen de geïntegreerde waarden niet altijd overeen met de concentraties van een specifiek metaboliet. Een alternatieve, meer geavanceerde statistis-

che methode is *spectral deconvolution*. BATMAN (**B**ayesian **A**u**T**omated **M**etabolite **A**nalyser for **N**MR data) (Astle et al., 2012; Hao et al., 2014) voert spectrale deconvolutie uit met behulp van gekende informatie over spectrale kenmerken van metabolieten. Op deze wijze schat BATMAN relatieve metabolische concentraties. Zowel spectral binning en spectral deconvolution door BATMAN werden toegepast op 400 MHz en 900 MHz NMR spectrum van bloedplasma stalen afkomstig van patiënten met longkanker en van een controlegroep (Hoofdstuk 11). De relatieve concentraties, die door BATMAN geschat zijn, werden vergeleken met de binning integratie waardes, wat de mogelijkheid geeft om een onderscheid te maken tussen longkanker patiënten en de controle groep (Hoofdstuk 12). Voor de 400 MHz data, heeft de spectral binning aanpak een groter onderscheidend vermogen. Hiertegenover staat dat voor de 900 MHz data, de relatieve metabolische concentraties verkregen door BATMAN een groter voorspellend vermogen hadden. Ondanks het feit dat spectral binning minder rekenkracht vereist, komen de kenmerken geschat door BATMAN overeen met de specifieke metabolieten waardoor de interpretatie eenvoudiger wordt.