Statistical Methods for the Reconstruction of Metabolite Networks in Humans and Plants

Georgios Bartzis

Propositions

1. Conditioning on variation from specific covariables allows us to better disentangle different metabolic aspects.
   (this thesis)

2. A single network containing information from multiple omics levels is more informative than multiple networks from single omic sources.
   (this thesis)

3. The fourth scientific paradigm of data-driven discovery cannot do without classic statistics.

4. Interpretation is crucial for imbuing numbers with actionable insights.

5. Progress in a complex world still hinges on the bold questioning pioneered by the ancient Greeks.

6. The value of any network, whether in society or in science, lies in the connections we choose to explore, without losing sight of the broader whole they create.

Propositions belonging to the thesis, entitled

Statistical Methods for the Reconstruction of Metabolite Networks in Humans and Plants

Georgios Bartzis
Wageningen, 16 May 2025

Statistical Methods for the Reconstruction of Metabolite Networks in Humans and Plants

Georgios Bartzis

Thesis committee

Promotors
Prof. Dr F.A. van Eeuwijk
Professor of Applied Statistics
Wageningen University & Research

Prof. Dr J.J. Houwing-Duistermaat
Professor of Mathematics
Radboud University, Nijmegen

Co-promotor

Dr C.F.W. Peeters
Associate Professor of Applied Statistics
Wageningen University & Research

Other members
Prof. Dr I. Athanasiadis, Wageningen University & Research
Prof. Dr P.E. Slagboom, Leiden University Medical Center
Prof. Dr M. Suarez Diez, Wageningen University & Research
Dr J.J.J. van der Hooft, Wageningen University & Research

This research was conducted under the auspices of the C.T. de Wit Graduate School
of Production Ecology & Resource Conservation (PE&RC)

Statistical Methods for the Reconstruction of Metabolite Networks in Humans and Plants

Georgios Bartzis

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rectus Magnificus,
Prof. Dr C. Kroeze,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 16 May 2025
at 3:30 pm in the Omnia Auditorium

Georgios Bartzis
Statistical Methods for the Reconstruction of Metabolite Networks in Humans and Plants,
125 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2025)
With references, with summary in English

DOI https://doi.org/10.18174/670261

To my parents, Ioannis & Evangelitsa
And to my sisters, Eleftheria & Dimitra-Paraskevi

Without data, you're just
another person with an opinion
-*William Edwards Deming*

Life isn't a series of
simple questions and answers!
-*Kaido*

Contents

1

Introduction

1.1 Preface

1.1.1 Background

Over the past decades, rapid advances in high-throughput technologies enabled the characterization of different biological levels of an organism. The term “ome” is commonly used to refer to the complete set of molecules of a particular type in an organism. For example, the genome refers to the complete set of genes, and the metabolome to the complete set of metabolites. The term “omics” is typically used to refer to studying the entire set of molecules of a particular type i.e. genomics is the study of the genome, proteomics is the study of the proteome, and metabolomics is the study of the metabolome. The quantification and introduction of those data into biomedical research was proclaimed to broaden our understanding in the structure and functions of living organisms (Bilello, 2005) establishing the field of systems biology. In systems biology, relationships between and within different biological levels describe characteristics and functions of an organism. Hence, in order to understand the organization of those interactions, the era of omics has gained a lot of attention.

1.1.2 Rationale and challenges of integrative analysis

As both time and cost of use have been significantly reduced, high-throughput technologies are routinely used (Lightbody et al., 2018) to generate massive datasets. These datasets often contain multiple omic data types and may also contain confounding variables, which are typically present in conventional studies. As these multi-omic datasets are collected from the same set of samples (possibly over time), it highlights the importance of the integration of data across different omic layers to gain a more comprehensive understanding of biological systems.

Particularly, to fully utilize this multi-omic information, integrative methodologies are increasingly popular for better understanding the interplay between different biological functional levels. By utilizing information from different biological levels, we gain knowledge that would not have been evident by investigating the components independently. These integrative methods combine biology with mathematical modeling and have promised to improve the insight in biological processes underlying complex traits and elucidate biological pathways (Sun & Hu, 2016; C. Clark, Rabl, Dayon, & Popp, 2022).

While multi-omic modeling can potentially provide a more complete understanding of biological systems than single-omic approaches, it also presents several challenges. One of the biggest challenges is addressing differences in data nature. Consequently, deciding which type of model is better suited for addressing the aforementioned complications can be difficult. The challenges of model selection are two-fold, i.e. prediction or interpretation. Selecting the appropriate model for each purpose is not a straightforward task. With regard to prediction, a robust model that produces highly accurate results is essential, whereas for interpretation, the selected model's outcomes must be reasonably supported by the biological literature. It is not uncommon for the model chosen for prediction to differ from the one chosen for ease of interpretation. Consequently, selecting a model that satisfies both requirements can be difficult, as it requires a careful evaluation of available methods with their strengths and limitations.

The metabolome is the level of most interest for us and thus we will be working around that level. For integrating information from other omic sources, we will also be working with genomics and phenotypes.

1.2 Metabolome

1.2.1 Definition

Metabolomics is the scientific study of small ($<$ 1500 daltons) biologically active molecules (Barchet, 2013). These molecules are the intermediate or end products of metabolic processes (Nielsen & Jewett, 2007; Medina-Cleghorn & Nomura, 2013). The entirety of metabolic processes and reactions, in which the end product works as input for another reaction is called metabolism. Finally, the complete collection of metabolites together with their interactions found in a biological sample (e.g. cells, organs, tissues, biofluids, etc) is called the metabolome. The metabolome comprises of reactions and associations that are used to determine the physiological and biochemical activities of a cell (Medina-Cleghorn & Nomura, 2013). The metabolome is relatively close to the phenotypic level and its interdependent nature has been suggested to be responsible for phenotypic diversity in plants (Keurentjes, 2009). As a result, metabolites are typically studied in conjunction to phenotypic traits.

1.2.2 Importance

Measuring the total collection of chemical reactions that continuously take place as results of internal and external perturbations enable taking a snapshot of the organism at a given time point. Considering that metabolites capture information from all functional levels of a cell they have been given an important place in the study of biological systems (Nielsen, 2003). By analyzing the metabolome, researchers can gain insights into the biochemical pathways and processes that underlie the phenotype of an organism, such as its physical and physiological characteristics (A. Zhang et al., 2014). This can help bridge the gap between the genotype (the genetic makeup of an organism) and the phenotype (the observable traits of an organism), as the metabolome provides a link between the genetic information and the actual metabolic processes that give rise to the phenotype. As it has been characterized as intermediate phenotype, the metabolome has vastly been used for drug discovery and biomarker detection. (Tebani, Abily-Donval, Afonso, Marret, & Bekri, 2016).

Another advantage of studying the metabolome is its reduced size. While the genome and proteome typically consist of tens of thousands of components, targeted metabolites are generally comprised of a few hundred metabolites. Compared to the metabolites, the analysis of the enormous genomic and proteomic data sets pose considerable computational and interpretative challenges.

The amplification effect observed in metabolites is an added advantage of using metabolites. This effect refers to the magnified change in downstream metabolite concentrations resulting from upstream effects, such as simple genomic, proteomic or transcriptomic changes (Bory, Boulieu, Chantin, & Mathieu, 1990; Gowda et al., 2008; Fiehn, 2002; Rinschen, Ivanisevic, Giera, & Siuzdak, 2019). The larger signal makes the separation of signal from noise easier in general for metabolites (as compared to upstream omics features). Hence, the amplification effect highlights the importance of metabolites in identifying potential biomarkers and therapeutic targets for various diseases.

In systems biology, the metabolome is expected to be vastly beneficial considering it displays a strong interdependent nature (Bartel, Krumsiek, & Theis, 2013; Rinschen et al., 2019). In other words, metabolites often function in metabolic pathways where the activity of one metabolite can impact the abundance or activity of others. This implies that alterations in a single metabolite could potentially have consequences on other metabolites. As metabolites form connections that convey useful information, multivariate statistical tools such as network analysis have widely been used by recovering the underlying metabolic reactions (Amara et al., 2022) and representing them as graphical models. Metabolites are represented as nodes in a graph and their realized relationships as edges connecting them (Ursem, Tikunov, Bovy, Van Berloo, & Van Eeuwijk, 2008; DiLeo, Strahan, den Bakker, & Hoekenga, 2011; H. Wang et al., 2015).

1.3 Graph theory and network analysis

1.3.1 Graph and network definitions

Graph theory is the study of graphs which are mathematical structures modeling relationships between nodes (representing objects/characteristics). The pairwise relationships are usually represented by edges connecting the nodes. Network analysis is part of graph theory in which the nodes and edges have attributes. While, in graph theory, the interest is in questions mainly regarding the topology and properties of graphs, network analysis tools answer questions concerning what those networks represent (Barnes & Harary, 1983).

To put it concisely, in the context of systems biology, network analysis is a way of studying the joint data distribution of all elements of an omics source of data. In the metabolomic case, the metabolites are seen as a set of random variables with an underlying joint distribution. The associations between metabolites are inscribed in, or are analogous to the variables' variance-covariance matrix. Consequently, by using data measurements on the concentration levels for each metabolite over different samples, the joint distribution is estimated. To ease interpretation, the recovered associations are typically visually represented in a graph, with metabolites being nodes and their pairwise associations as edges connecting them. As network analysis is a data-driven approach estimating an underlying model, it is possible to identify novel metabolic relationships that might be missed by other approaches. Extracted patterns are then used as a basis to point out biological mechanisms underlying the traits of interest (Toubiana, Fernie, Nikoloski, & Fait, 2013).

As there is a very close link between graph theory and network analysis, (i) the terms graph theory and network analysis have been used interchangeably, and (ii) many concepts from graph theory have been used in network analysis (Barnes & Harary, 1983) in order to better understand the nature of the network under investigation. For representing a graph or a network, a matrix encoding all pairwise associations between nodes is used, collected in a so-called intensity matrix (denoted as $\boldsymbol{W}$). As previously stated, this usually is analogous to the (possibly sparsified) empirical variance-covariance matrix of the data. Due to its data driven nature, the intensity matrix estimation is typically sensitive to data pre-processing. For removing spurious associations due to noise in the data, the intensity matrix can be transformed into a Boolean matrix, i.e adjacency matrix (denoted $\boldsymbol{A}$) by applying a hard threshold. The existence of association between two nodes is represented by a non-zero element in the corresponding cell (0 for absence, and 1 for presence).

In Table 1.1 we present a toy example of an adjacency table that describes the connections between four nodes in a network. As discussed, the table is represented as a Boolean matrix, where the presence of an edge between two nodes is denoted by 1, and absence of an edge is denoted by 0. The network representation of the adjacency matrix can be seen in Figure 1.1, where each node is represented as a circle and each edge is represented as a line connecting the nodes. When analyzing small networks, as in our toy example, it may be easy to infer pairwise associations from an adjacency matrix, making network representation less advantageous. On the other hand, when dealing with a larger number of nodes/variables, network representation can offer more meaningful insights. As can be observed from the adjacency matrix, 0 entries between nodes signify the absence of edges in the network representation, while 1 entries indicate the presence of edges. This simple

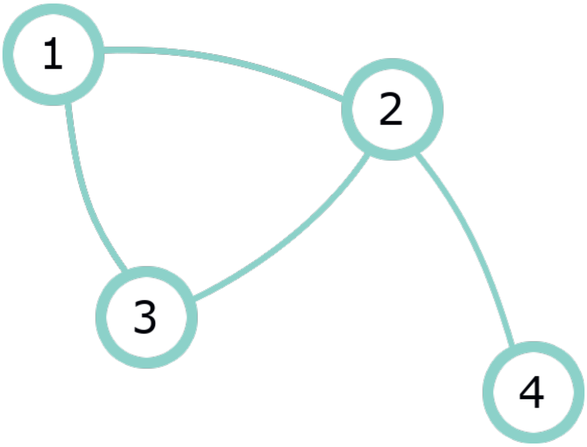


Fig. 1.1: Network representation of the adjacency matrix in Table 1.1, where each node is depicted as a circle and each edge is depicted as a line connecting the nodes

example serves as a useful tool for understanding how adjacency tables can be used to represent network connections.

	Node 1	Node 2	Node 3	Node 4
Node 1	0	1	1	0
Node 2	1	0	1	1
Node 3	1	1	0	0
Node 4	0	1	0	0

Table 1.1: Adjacency matrix of a 4-node network represented as a Boolean matrix, where 1 denotes the presence of an edge and 0 denotes the absence of an edge. The matrix displays the connections between nodes.

With the massive public availability of high throughput data, the development and use of network analysis gained considerable recognition. Part of the popularity can also be attributed to the intuitive representation of the interactions between the components involved. Complex data can easier be understood, described, and analyzed using network analysis.

1.3.2 Network models and tools

In this thesis, our focus is on the analysis of the metabolome by using undirected networks, meaning that we are not inferring causality and the edges between metabolites do not have a direction. In other words, the relationship between any two metabolites is symmetric. That allows us to explore the correlation structure of the data without the need to specify a priori which metabolites are drivers or targets. We will explore techniques for integrating other omics sources where each omics source can belong to a different network layer. We will analyze the network structure using network-based methods based on correlation patterns to gain insights into the underlying biological mechanisms. We therefore, base our estimation on Pearson's correlation (*Weighted Gene Co-expression Analysis*; *WGCNA*) and partial correlation (*Graphical LASSO*; *GL*). The former is based on observed correlations and may contain spurious associations as they are indicative of direct and indirect associations as well as association by confounding. Contrarily, Graphical LASSO is based on partial correlations, meaning that a pair of nodes does not share an edge if they are conditionally independent given all other variables. In other words, Graphical LASSO eliminates spurious relationships. For evaluating our network reconstruction methods, a module identification method can typically be used. The identification of modules is relevant for understanding the modular organization of the network, as well as the functional relationships between the nodes. The ability to identify functionally related nodes or modules based on literature can provide insights into the underlying biological mechanisms, facilitate the identification of potential biomarkers or drug targets, and predict new interactions or pathways (Barabási, Gulbahce, & Loscalzo, 2011).

1.4 Network analysis and the metabolome

Network analysis has become a powerful tool for the analysis of metabolomics data in recent years. These methods enable complex data visualization and interpretation, allowing the identification of meaningful patterns and relationships (Toubiana et al., 2013; Perez De Souza, Alseekh, Brotman, & Fernie, 2020). Overall, the integration of network analysis into metabolomics research provides critical insights into metabolic function and disease mechanisms, leading to new diagnostic and therapeutic strategies. Likewise, network analysis in plants provides a comprehensive view of the metabolic pathways and networks involved in plant growth, development, and responses to environmental stress. Consequently, it can be used to identify key metabolic pathways associated with plant traits making it a valuable tool for plant improvement. While humans and plants have distinct metabolic pathways and processes, there are some similarities that can be observed. For instance, the same basic classes of biomolecules, including carbohydrates, lipids, proteins, and nucleic acids are utilized in both humans and plants. Despite their potential, network-based approaches have limitations that need to be considered. Data preprocessing is crucial, including noise removal and data normalization, to ensure that the network analyses are accurate and reliable. Edges in metabolite networks are susceptible to covariables related to the study designs and other sources of variation (Rinschen et al., 2019). Therefore, as much information as possible should be used for identifying meaningful patterns. Another limitation is the risk of false positives arising from spurious connections due to noise or other factors. Expert knowledge in both metabolomics and network analysis is essential for the interpretation of network-based results. While network-based approaches are promising for metabolomics data analysis and interpretation, careful consideration of the limitations and challenges involved is necessary to ensure their effective implementation. These challenges should be taken into account when designing and performing network-based analyses to ensure that they provide reliable and interpretable results. Methods for modeling the joint metabolite distribution while at the same time adjusting for variability originating from nuisance factors.

1.5 Objectives and outline of thesis

In this thesis we develop a framework for reconstructing metabolite networks for humans and plants while accounting for various sources of biological (environmental conditions) and technical variation (study design), while addressing nuisance variation. In **Chapter 2**, we study how information on the study design can be incorporated to estimate metabolite networks. While metabolites concentrations and their extracted correlation patterns are susceptible to covariables related to the study design, often the research interest needs to be focused on variation coming from a certain omic source. We work in the regression framework in conjunction with network analysis methods for demonstrating how to extract sets of metabolites sharing a similar relation to a covariable of interest (Bartzis et al., 2017). For each metabolite, the metabolic variation is decomposed by using a linear regression model; the part related to the covariable of interest is retained and further used for network analysis. The proposed method is demonstrated using metabolite data coming from (i) an Arabidopsis Thaliana desiccation tolerance experiment (Maia, Dekkers, Provart, Ligterink, & Hilhorst, 2011) and (ii) an epidemiological study in humans (DILGOM study) (Inouye et al., 2010; Kettunen et al., 2012).

In **Chapter 3**, we extend our network reconstruction pipeline by working with repeated measurements. In a longitudinal setting the within sample dependence should be modeled when the data are analyzed. We use human metabolite data that are measured at two time points (2007 and 2014). Additionally, SNP data together with a self reported food frequency questionnaires were available. We use a linear mixed effects model to study the metabolite concentrations. By using longitudinal measurements, time effects and subject specific random effects can be estimated. In this application, the subject specific effects represent lifestyle by being the remaining unmeasured shared sources of metabolic variation. Using lifestyle, dietary patterns and time, we identify metabolites sharing similar relationships to diet and lifestyle.

In **Chapter 4**, we propose a guided network reconstruction approach for studying metabolite networks subject to another omic source having a network organization of its own, e.g. SNPs or genes. A key question in systems biology is how to model omics data at a systems level (integrative analysis), instead of each layer separately. We investigate if reconstructing networks of a particular omics source benefits when information from the underlying network organization of another omics source is used. For demonstration, we use multi-omics information collected from a recombinant inbred line (RIL) population of two natural Arabidopsis variants (Joosen et al., 2013; Joosen, 2013). We incorporate information from *guiding* omics data (SNPs or genes) into the analysis of the *target* dataset (metabolites) by utilizing a regularized linear model explicitly accounting for the network organization of the guiding dataset. When the guiding dataset consists of SNP data, this method can provide results for QTL detection.

In **Chapter 5**, we revisit the question of associating a response with an omic source having a network organization of its own, as in Chapter 4, but instead we work in the genomic selection domain (Bartzis, Peeters, & Eeuwijk, 2022). A training population that has been both genotyped and phenotyped is used to describe a marker-trait relationship. The prediction model is then applied to unphenotyped samples to produce genomic estimated breeding values measuring the samples' genetic merit. The most common approach to do so is by using the *genomic best linear unbiased predictor* (GBLUP), where the marker effects are estimated as random in a linear mixed model. We propose a regularized linear model (namely proximity smoothed BLUP; psBLUP) penalizing the L2-norm of the coefficients (like GBLUP) while additionally encouraging smoothness on neighboring marker effects, since it is anticipated that some of them might be correlated due to their spatial proximity within the chromosomes. We compare psBLUP to GBLUP for 64 phenotypes from a Arabidopsis RIL population (metabolites in this case; (Joosen et al., 2013; Joosen, 2013)) concluding that that psBLUP yields superior accuracy by explicitly accounting for the dependence between marker effects.

Chapter 6 provides a overview of the thesis, followed by a short summary of the research findings. We discuss a general view of the proposed framework using a linear model unifying all aspect of metabolic variation that have been discussed. We conclude the chapter by discussing the relevance of the proposed framework in terms of applicability, generalizability and transferability.

2

Estimation of Metabolite Networks with Regard to a Specific Covariable: Applications to Plant and Human Data

Bartzis, G., Deelen, J., Maia, J., Ligterink, W., Hilhorst, H.W., Houwing-Duistermaat, J.J., Van Eeuwijk, F., & Uh, H.W. (2017) *Metabolomics*, *13*, 129

2.1 Introduction

In recent years, network analysis of biological datasets has become an increasingly popular tool for studying the relationships between large numbers of variables that occur in omics research on transcripts, metabolites, proteins and others. In networks, variables are represented by nodes and their relationships, direct and indirect interactions (physical or functional), are represented by edges or links. One is often interested in the joint distribution of a set of variables conditional on a particular covariable. For example, one may want to study the relations between a set of metabolites with regard to Body Mass Index (BMI). As an illustration, in Figure 2.1 we show the concentrations of eight Very Large Density Lipoprotein (VLDL) particles that are associated with BMI and gender in a similar way, i.e. males have higher VLDL concentrations than women; low VLDL concentrations are associated to low and medium BMI categories, and high VLDL concentrations to high BMI. The aim of the current paper is to detect these groups of metabolites with similar relationships by using network analysis.

For network estimation, networks with different features are typically used for representing networks, i.e. undirected or directed networks. In undirected networks, edges connecting two nodes do not have a direction indicating a symmetric relationship between them. In biology, undirected network estimation methods based on correlation are often preferred. These methods perform well with large numbers of samples and variables. However, little is known about their performance in a small sample setting. Therefore, estimating, describing, visualizing and comparing networks for relatively small samples is an ongoing topic of research (Kolaczyk & Krivitsky, 2015). We will consider two methods of estimating undirected networks. The first is Weighted Gene Co-expression Network Analysis (WGCNA) based on correlation (B. Zhang & Horvath, 2005; Langfelder & Horvath, 2008; W. Zhao et

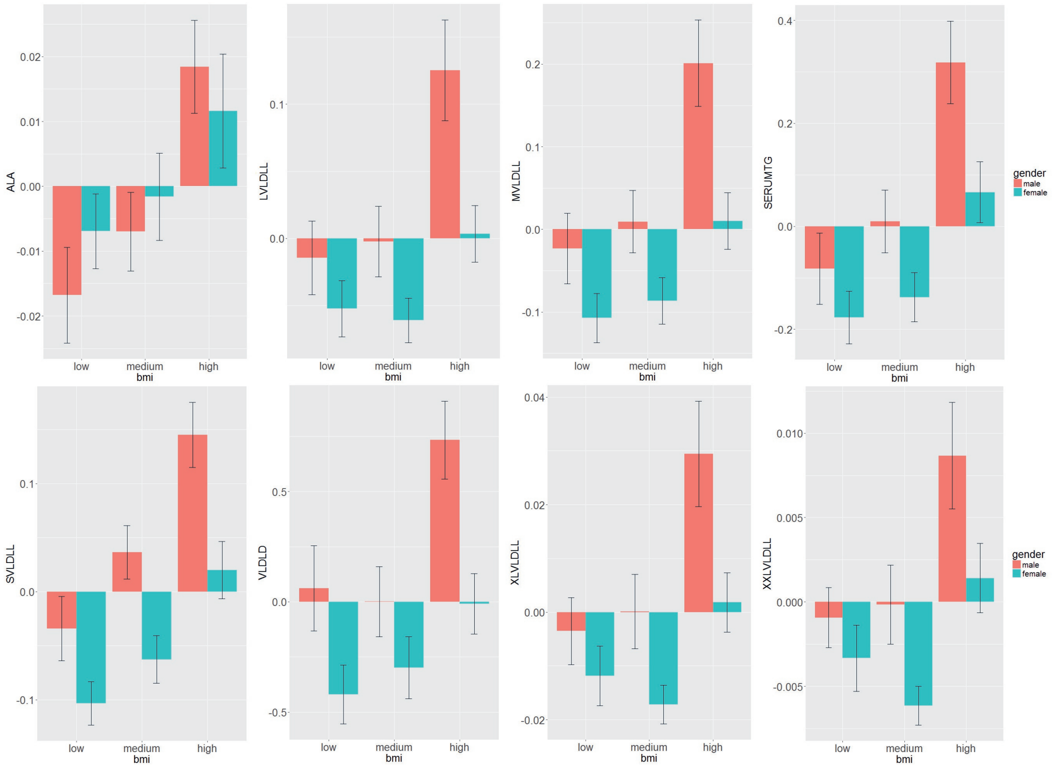


Fig. 2.1: Barplots representing the metabolite concentrations in humans by BMI class and sex for metabolites belonging in the VLDL module

al., 2010). While WGCNA was mainly developed for analysis of gene-expression data, applications on metabolite data have been reported as well (DiLeo et al., 2011; G. Zhang et al., 2013). WGCNA is based on the concept of scale free networks implying the existence of a few highly connected nodes (hubs) participating in a very large number of metabolic reactions (W. Zhao et al., 2010). In WGCNA, the strength of the connection between nodes is typically dictated by a similarity measure (W. Zhao et al., 2010). The second method is the Graphical LASSO (GL) (Friedman, Hastie, & Tibshirani, 2001, 2008) based on partial correlations. For this Gaussian Graphical Model (GGM) based method, for two nodes not sharing a direct edge in a network the implication is that they are conditionally independent given all other variables. To obtain a sparse network an L1 penalty can be used. The penalty can be determined using a stability selection algorithm (StARS) (H. Liu, Roeder, & Wasserman, 2010) to select a stable set of edges.

For describing and comparing the estimated networks using the two methods above, we first characterize the topology of the networks by measuring three types of network concepts (Dong & Horvath, 2007; Horvath & Dong, 2008), i.e. density, centralization, and heterogeneity.

We will consider two different data applications for estimating metabolite networks. In the first application, the metabolites are coming from an experiment with seeds in which the desiccation tolerance of these seeds was investigated as a function of genotype and a managed environment condition, or treatment. Seeds from two genotypes of the well-known plant *Arabidopsis thaliana*, the genotype Columbia-0 (Col-0) and the abscisic acid deficient mutant 2-1 (*aba2-1*) were selected. Germinated seeds at radicle protrusion were selected and either frozen in liquid nitrogen and stored at -80°C directly or after 3d of incubation in -2.5 MPa polyethylene glycol (PEG), 5µM ABA (ABA) or a combination of -2.5MPa PEG + 1µM ABA (PEG + ABA). Therefore, four treatments for metabolic profiling have been considered here: i) no treatment (control), ii) PEG, iii) ABA or iv) both PEG and ABA. In this paper, we focus on estimating metabolite networks based on genotype-related information.

The second application concerns serum metabolites of unrelated individuals coming from the capital region of Finland. In this observational study our specific interest lies in estimating networks of metabolites, conditioned on the individuals' BMI status which might interact with individuals' Sex and Age.

The rest of the paper is organized as follows. In Section 2, we propose our method for selecting information relevant to a certain covariable prior to network estimation. Additionally we review some existing network estimation methods. In Section 3, we demonstrate our network estimation approaches on metabolite data coming from plants and humans, and we conclude with a Discussion in Section 4.

2.2 Methods

A network consists of a set of nodes (or vertices) connected by a set of edges (or links). In this paper we consider undirected networks of metabolites, where the nodes correspond to the metabolites, and the edges between metabolites represent their relationship. For a network of P nodes, the network structure can be represented by the $P \times P$ adjacency matrix $\mathbf{A}$. For undirected networks, the elements a_{ij} of the adjacency matrix are defined as follows:

$$a_{ij} = \begin{cases} 1, & \text{if there is an edge between node } j \text{ and node } i \\ 0, & \text{otherwise.} \end{cases}$$

Note that the adjacency matrix $\mathbf{A}$ is symmetric and has zeros on the diagonal. The degree or connectivity k_i of a node i is defined as $k_i = \sum_{j \neq i} a_{ij}$: i.e. k_i equals the number of neighbors of node i in the network. In addition to the adjacency matrix we typically consider a $P \times P$ intensity matrix $\mathbf{W}$ where the elements w_{ij} represent the strength of the relationship between node i and j. If nodes i and j are not linked, the weight w_{ij} is equal to zero. Popular choices for the weights w_{ij} are

Pearson's correlation coefficient, Mutual Information, Euclidean distance, partial correlation and Topological Overlap.

We will consider absolute Pearson's correlation coefficient and partial correlations with no self-edges ($w_{ii} = 0$). Analogously to the degree of a node i (k_i), the strength or weighted degree is defined as $s_i = \sum_{j \neq i} w_{ij}$. The strength of a node takes into account both the connectivity as well as the weights of the edges.

2.2.1 Identifying the specific parts of metabolic concentrations relevant to a covariable

The weights represent the relationships between the metabolite concentrations. In addition one might be interested in the relationships between specific parts of the metabolite concentrations: for example the part which is related to a covariable of interest. The idea is that metabolites with similar relationships with this covariable will tend to be close to each other in the network (See for example Figure 2.1). The parts of the metabolite concentration concerning this specific covariable can be obtained by fitting linear regression models (or ANOVA) to the metabolic variables. Let $\mathbf{Y} = (\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(P)})$ be the concentrations of P metabolites and let $\mathbf{Y}^{(p)}$ be the vector of concentrations for the pth metabolite. Assume that $\mathbf{Y}$ follows a multivariate Gaussian distribution $\mathbf{Y} \sim N_P(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. From m covariables let $\mathbf{X}_m$ be the categorical covariable of main interest. The remaining $m-1$ covariables are denoted as $\mathbf{X}^{(-m)} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{m-1}\}$.

Now we propose the following regression model:

$$\mathbf{Y}^{(p)} = \beta_0^{(p)} + \boldsymbol{\beta}_1^{(p)}\mathbf{X}_m + \sum_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} \gamma_{\boldsymbol{\delta}}^{(p)} \prod_{j=1}^{m-1} \mathbf{X}_j^{\boldsymbol{\delta}_j} + \sum_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} \boldsymbol{\eta}_{\boldsymbol{\delta}}^{(p)} \mathbf{X}_m \circ \prod_{j=1}^{m-1} \mathbf{X}_j^{\boldsymbol{\delta}_j} + \varepsilon^{(p)}, \quad (2.1)$$

where $\boldsymbol{\varepsilon}^p \sim N(0, \sigma^2_{(p)})$ represents the random noise, $\boldsymbol{\Delta}$ is the space with elements all vectors of length $m-1$ with all combinations of zeros and ones, except all zeros, i.e. $\boldsymbol{\Delta} = \{(1,0,\ldots,0), (0,1,\ldots,0), \ldots, (1,1,\ldots,1)\}$, and $\circ$ is the Hadamard product. The term $\sum_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} \boldsymbol{\gamma}_{\boldsymbol{\delta}} \prod_{j=1}^{m-1} \mathbf{X}_j^{\boldsymbol{\delta}_j}$ models all main effects and second and higher order interaction terms of covariables in $\mathbf{X}^{(-m)}$. For example, for $m-1=2$, $\sum_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} \boldsymbol{\gamma}_{\boldsymbol{\delta}} \prod_{j=1}^{2} \mathbf{X}_j^{\boldsymbol{\delta}_j} = \boldsymbol{\gamma}_{10}\mathbf{X}_1 + \boldsymbol{\gamma}_{01}\mathbf{X}_2 + \boldsymbol{\gamma}_{11}\mathbf{X}_1 \circ \mathbf{X}_2$. Similarly, the term $\sum_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} \boldsymbol{\eta}_{\boldsymbol{\delta}} \mathbf{X}_m \circ \prod_{j=1}^{m-1} \mathbf{X}_j^{\boldsymbol{\delta}_j}$ models all terms interacting with $\mathbf{X}_m$. To ensure identifiability of all parameters a level of the covariable is used as reference category.

The relevant information for the pth metabolite with regard to the categorical covariable $\mathbf{X}_m$ is given by:

$$\hat{\mathbf{Y}}^{(p)} = \hat{\boldsymbol{\beta}}_1^{(p)}\mathbf{X}_m + \sum_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} \hat{\boldsymbol{\eta}}_{\boldsymbol{\delta}}^{(p)} \mathbf{X}_m \circ \prod_{j=1}^{m-1} \mathbf{X}_j^{\boldsymbol{\delta}_j}. \quad (2.2)$$

Now the edge between node p and q in a network can be based on correlation between $\hat{\mathbf{Y}}^{(p)}$ and $\hat{\mathbf{Y}}^{(q)}$.

2.2.2 Network Estimation Methods

We consider two network approaches, namely WGCNA based on pairwise correlations between metabolites and GL based on partial correlations.

Weighted Gene Co-expression Network Analysis (WGCNA)

WGCNA (Horvath & Dong, 2008; B. Zhang & Horvath, 2005; Langfelder & Horvath, 2008) has been developed to efficiently analyze the correlation patterns among genes using gene expression data from microarray experiments. Network construction with WGCNA is typically followed by identifying clusters (modules) of highly correlated genes.

Estimation of networks - modeling the intensity matrix:

Complex networks may display non-trivial topological features, such as a heavy tail in the empirical distribution of the degree of the nodes (the number of edges connected to a node). In biology, often networks with many low degree nodes and a few high degree nodes are of interest. Such networks are called scale free. To determine whether a network is scale free the log of the degree frequency ($\log(P(k))$) is plotted against the logarithm of the degree ($\log(k)$). A linear relationship indicates that the network is scale-free. A scale-free degree distribution can be expressed as $P(k) \propto k^{\gamma}$ and in a weighted case, i.e. in the context of WGCNA, the intensity w_{ij} can be written on a logarithmic scale, $\log w_{ij} = \gamma \log |\text{Cor}(y_i, y_j)|$ for $\gamma > 1$. The threshold parameter γ might be chosen in a way such that the network approximately satisfies the scale-free topology criterion.

In our experience, however, not all biological datasets yield scale free networks. If the network is not scale free γ will be determined by the amount of noise in the dataset. For two Gaussian random variables, the magnitude of random noise for correlation coefficients from N samples is $1/\sqrt{N}$. In order to sufficiently suppress low correlations due to noise, we take the smallest value for γ in such a way that the total noise is smaller than one (personal communication with Peter Langfelder):

$$\frac{P(P-1)}{2(\sqrt{N})^{\gamma}} < 1. \tag{2.3}$$

Module identification:

A network might consist of a set of modules of closely interconnected metabolites. Average linkage hierarchical clustering based on a dissimilarity measure is a popular method to define a dendrogram of the network. The modules are obtained by cutting this tree (the two-step dynamic hybrid algorithm) (Langfelder, Zhang, & Horvath, 2008). Here, we will use the following dissimilarity measure:

$$Diss(w_{ij}) = 1 - |w_{ij}|. \tag{2.4}$$

Graphical LASSO (GL)

GL is another popular approach to obtain a network for a set of variables. Assume that the metabolite concentrations (N by P data matrix $\mathbf{Y}$) follow a multivariate Gaussian distribution with mean vector μ and variance-covariance matrix $\mathbf{\Sigma}$. For simplicity we can assume that the data are centered, i.e. $\mu = \mathbf{0}$. The inverse covariance matrix $\mathbf{\Sigma}^{-1}$ is the precision matrix. When its elements are equal to zero, the pair of metabolites is conditionally independent given the other metabolites. A Gaussian graphical model (Lauritzen, 1996) is a network based on these conditional independence relationships.

To estimate $\mathbf{\Sigma}^{-1}$ a penalized log-likelihood approach can be used. Define the precision matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$. Under the Gaussian model, the log-likelihood function is given by (up to a constant):

$$\ell(\mathbf{\Theta}) \sim \log|\mathbf{\Theta}| - \mathrm{tr}(\mathbf{S}\mathbf{\Theta}), \tag{2.5}$$

where $\mathbf{S} = \mathbf{Y}^{\top}\mathbf{Y}/N$ is the sample covariance matrix. Maximizing expression 2.5 with respect to $\mathbf{\Theta}$ leads to the maximum likelihood estimate $\hat{\mathbf{\Theta}}$. Note that the elements of $\hat{\mathbf{\Theta}}$ are in general not exactly equal to zero. Further in the high-dimensional setting ($P > N$), $\mathbf{S}$ is singular and cannot be inverted to obtain $\hat{\mathbf{\Theta}}$.

Therefore a penalized version of the log likelihood is typically maximized (Friedman et al., 2008; Hastie, Tibshirani, & Friedman, 2009; Friedman et al., 2001; Rothman, Bickel, Levina, Zhu, et al., 2008; Yuan & Lin, 2006). For the Lasso penalty the log likelihood function is as follows:

$$\ell_{\lambda}(\mathbf{\Theta}) \sim \log|\mathbf{\Theta}| - \mathrm{tr}(\mathbf{S}\mathbf{\Theta}) - \lambda||\mathbf{\Theta}||_1, \tag{2.6}$$

where λ is a non-negative tuning parameter. For $\lambda = 0$, the resulting network will be fully connected. While λ increases, sparsity is induced to the estimated $\hat{\mathbf{\Theta}}$ and the network starts to lose edges to the point that no more edges are left. Consequently, elements of the resulting estimated precision matrix will be exactly equal to zero.

The tuning parameter λ can be chosen so that the number of edges are biologically relevant and straightforward to interpret. A statistical approach for choosing λ can be based on Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), or K-fold cross-validation. To obtain a stable edge set with a low false discovery rate, StARS (Stability Approach for Regularization Selection) (H. Liu et al., 2010) is an attractive approach. It provides a penalty corresponding to the least amount of regularization that simultaneously makes a network sparse and is replicable under random sampling. GL with StARS is implemented in the *huge* R package (T. Zhao, Liu, Roeder, Lafferty, & Wasserman, 2012).

2.2.3 Network Evaluation

We now consider several measures to describe a specific network or a subset of a network and to compare networks, namely density, centralization and heterogeneity

(Dong & Horvath, 2007). Let $\mathbf{M}$ be the weights of the edges of the nodes in a network, i.e. $\mathbf{M}$ is the $P \times P$ matrix $\mathbf{W}$ for WGCNA after suppressing low correlations due to noise or the stability matrix estimated by StARS for GL.

The network density of $\mathbf{M}$ is the mean of these weights and is estimated by:

$$Density(\mathbf{M}) = \sum_i \sum_{j \neq i} \frac{w_{ij}}{P(P-1)} = \frac{\bar{s}}{(P-1)}, \tag{2.7}$$

where $\bar{s}$ is the mean of s. A value close to one indicates high interconnectedness.

The centralization is the difference between the strength of the most connected node in the network with respect to the average network and is given by

$$\begin{aligned} Centralization(\mathbf{M}) &= \frac{P}{P-2}\left(\frac{\max(s)}{P-1} - Density\right) \\ &= \frac{P}{(P-1)(P-2)}\left(\max(s) - \bar{s}\right) \\ &\approx \frac{1}{P}\left(\max(s) - \bar{s}\right). \end{aligned} \tag{2.8}$$

This measure is large when the network is a star, i.e. the network contains one highly connected node.

Finally the variation of the strength of the nodes might be of interest. Heterogeneity equals the coefficient of variation of the strength distribution:

$$Heterogeneity(\mathbf{M}) = \frac{\sqrt{\mathrm{var}(s)}}{\bar{s}}. \tag{2.9}$$

These measures can be computed for a total network and for subnetworks or modules.

2.3 Application to data

We will now analyze, describe and visualize the correlation structure of two metabolites datasets by using the described network approaches. The datasets are from an experiment aimed to study desiccation tolerance in germinated Arabidopsis seeds (Maia, Dekkers, Dolle, Ligterink, & Hilhorst, 2014) and from an epidemiological cohort which studies the relationship between dietary, lifestyle, and genetic determinants and obesity and metabolic syndrome (DILGOM), which is a subset of the Finrisk 2007 survey (Inouye et al., 2010; Kettunen et al., 2012). The studies differ in the types of subjects (plants versus human), the study designs (experimental design (completely randomized) versus random sample of the population) and in sizes (27 versus 419). For both studies about the same number of metabolites are available, namely 56 and 58 metabolites for the experiment and epidemiological studies, respectively.

2.3.1 Experimental Design

Desiccation tolerance (DT) is the ability of certain organisms to lose most of their cellular water content and become extremely dry and re-hydrate without the accumulation of lethal damage. DT is common in seeds of land plants. Such seeds acquire DT during development and become sensitive again to extreme dehydration around the point of visible germination. Yet, if confronted with suboptimal conditions, such as osmotic stress, germinated desiccation sensitive seeds are able to activate global changes in gene expression and metabolite composition to re-establish DT (Maia et al., 2011). Here, we are interested in the network structure of the metabolic phenotype of two Arabidopsis genotypes, a Wild type, Col-0, and an abscisic acid-insensitive (*aba2-1* mutant). Germinated seeds were subjected to a set of treatments including the application of osmotic stress by polyethylene glycol (PEG) and abscisic acid (ABA) to re-establish DT. In addition to the network structure of the observed metabolite concentrations also the network structure with regard to the relationship between the metabolites and the genotype (wild type and *aba2-1* mutant) will be studied.

In total 56 metabolites were measured for 15 samples of Arabidosis wildtype seeds and 12 samples of *aba2-1* mutant seeds. In Table 2.1, the number of samples for each combination of genotypes (Col-0, *aba2-1*) and treatment (control or no treatment), -2.5MPa polyethylene glycol (PEG), 5µM Abscisic acid (ABA) or both (PEG+ABA)) are given.

Table 2.1: Experimental design for plant data. Cell counts denote the number of samples obtained per combination of treatment and Genotype

	No ABA		Yes ABA	
	No PEG	Yes PEG	No PEG	Yes PEG
Col-0	3	6	3	3
aba2-1	3	3	3	3

To obtain the genotypic part of the metabolites we fitted Eq. 2.2 to the data where $\mathbf{X}_m = \mathbf{G}$ represents the genotype; the four treatment levels were represented by two dummy variables denoting the administration of PEG and ABA. Thus, $\mathbf{X}^{(-m)} = (\mathbf{PEG}, \mathbf{ABA})$. Specifically for the p^{th} metabolite, the concentration associated with the seed's genotype is given by:

$$\tilde{\mathbf{Y}}^{(p)} = \hat{\boldsymbol{\beta}}_1^{(p)}\mathbf{G} + \hat{\boldsymbol{\eta}}_{10}^{(p)}\mathbf{G} \circ \mathbf{PEG} + \hat{\boldsymbol{\eta}}_{01}^{(p)}\mathbf{G} \circ \mathbf{ABA} + \hat{\boldsymbol{\eta}}_{11}^{(p)}\mathbf{G} \circ \mathbf{PEG} \circ \mathbf{ABA}. \quad (2.10)$$

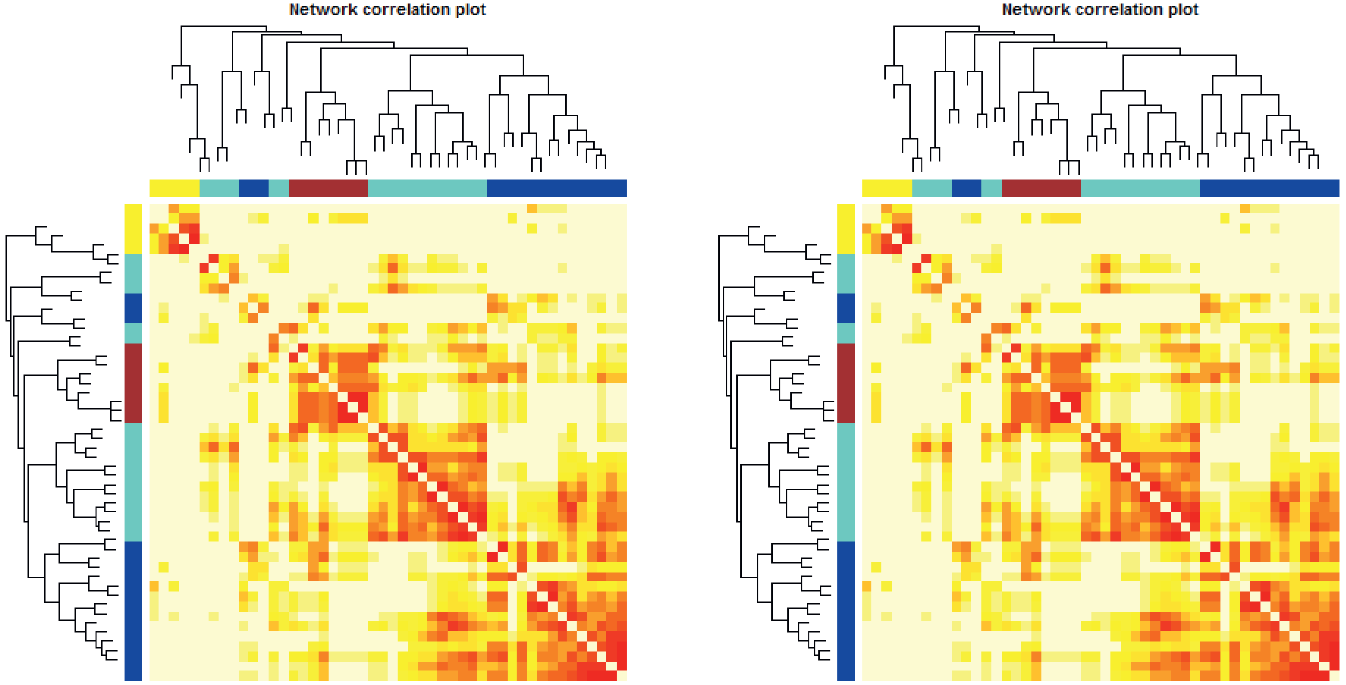
Only significant variables were included in the models. Here we applied a hierarchical approach: if the highest order interaction was significant, all terms were included in the model. If the highest order interaction was not significant it was removed from the model and the significance of the consecutive highest order interaction terms was checked. Finally if the genetic effect was not significant, the metabolite was discarded from further analysis. 48 metabolites were used for the network analysis of genotype related metabolic variation.

Hereafter we will denote networks of metabolites containing the original metabolite values ($\mathbf{Y}^{(p)}$) as $\mathcal{N}_O$ and networks containing the genotype related values ($\tilde{\mathbf{Y}}^{(p)}$) as $\mathcal{N}_G$.

Networks estimated by WGCNA

For estimation of the intensity matrix $\mathbf{W}$ we first tried to find the soft thresholding parameter γ corresponding to a scale free topology. However it appeared that for reasonable powers (γ being less than 12) both networks ($\mathcal{N}_O$ and $\mathcal{N}_G$) were not scale-free. Therefore based on the sample size ($N = 27$) and the number of metabolites ($P = 54$ for $\mathcal{N}_O$ and $P = 48$ for $\mathcal{N}_G$), $\gamma = 5$ was chosen for suppressing low correlations due to noise (see Section 2.2.2).

In Figure 2.2, the intensity matrices $\mathbf{W}_{\mathcal{N}_O}$ and $\mathbf{W}_{\mathcal{N}_G}$ are depicted together with their corresponding dendrograms obtained by average linkage hierarchical clustering. The visualization of the heatmaps reveals higher correlations (deeper red colors) between the genotype related metabolite values in ($\mathcal{N}_G$) compared to the original metabolite values ($\mathcal{N}_O$).



(a)

(b)

Fig. 2.2: Heatmap plots for plant metabolites. Heatmaps were estimated for (a) $\mathcal{N}_O$ when the original metabolite values were used, and (b) $\mathcal{N}_G$ when using the metabolite information related to the Genotype. Deep red colors denote high values of absolute correlation between pairs of metabolites, while lighter colors correspond to weaker correlations. Dendrograms were obtained using hierarchical clustering while modules correspond to square blocks along the diagonal. Interconnected modules are color coded by using the color bands beneath the displayed dendrograms.

Table 2.2: Network statistics when using WGCNA (Density, Centralization, Heterogeneity) for modules and network as whole in plant data

		$\mathcal{N}_O$[a]			
Module	Density	Centralization	Heterogeneity	Nr Nodes	Color Coded
Module 1[b]	0.33	0.20	0.33	6	Green
Module 2[c]	0.30	0.18	0.35	11	Turquoise
Module 3[d]	0.27	0.14	0.41	9	Brown
Module 4[e]	0.23	0.22	0.56	7	Yellow
Module 5[f]	0.20	0.12	0.31	11	Blue
Module 6[g]	0.19	0.13	0.41	5	Black
Module 7[h]	0.18	0.17	0.30	5	Red
Complete	0.06	0.04	0.45	54	

[a] Network using the original metabolite values

[b] Module 1: Fructose, Fructose-6-phosphate, Glucose, Glucose-6-phosphate, Glyceric-acid, Xylose

[c] Module 2: 2-Aminoadipic-acid, Glutamine, Isoleucine, Lysine, Nicotinic-acid, Phenylalanine, Proline, Pyroglutamic-acid, Threonine, Tyrosine, Valine

[d] Module 3: 5-Aminocarboxy-4,6-dihydroxypyrimidine, Ascorbic-acid, Glutamate, Glycerol, Hexonic-acid, Monothylphosphate, Phosphoric-acid, Threonate, Urea

[e] Module 4: Aspartate, Beta-alanine, Citrate, Ethanolamine, Maltose, Serine, Tryptophan

[f] Module 5: Alfa-Ketoglutaric acid, Allantoine, Asparagine, Fucose, Galactinol, Glycine, Leucine, Malate, Raffinose , Suberyl-glycine, Succinic acid

[g] Module 6: Alanine, Methionine, Myo-inositol, Pentonic acid, Sucrose

[h] Module 7: Anhydroglucose, Benzoic acid, Fumarate, Trans-Sinapinic acid, Xylofuranose

		$\mathcal{N}_G$[i]			
Module	Density	Centralization	Heterogeneity	Nr Nodes	Color Coded
Module 8[j]	0.62	0.15	0.17	8	Brown
Module 9[k]	0.53	0.26	0.30	5	Yellow
Module 10[l]	0.42	0.17	0.31	17	Blue
Module 11[m]	0.38	0.19	0.33	18	Turquoise
Complete	0.21	0.16	0.44	48	

[i] Network using Genotypic-related information

[j] Module 8: Alanine, Fructose, Fructose-6-phosphate, Fumarate, Glucose-6-phosphate, Threonate, Trans-Sinapinic acid, Tyrosine

[k] Module 9: 2-Aminoadipic-acid, Anhydroglucose, Benzoic-acid, Maltose, Xylofuranose

[l] Module 10: Ascorbic-acid, Aspartate, Glucose, Glutamine, Glyceric-acid, Isoleucine, Leucine, Lysine, Raffinose, Methionine, Monomethylphosphate, Phenylalanine, Serine, Threonine, Tryptophan, Valine, Xylose

[m] Module 11: 5-Aminocarboxy-4,6-dihydroxypyrimidine, Allantoine, Asparagine, Citrate, Galactinol, Glycerol, Glycine, Hexonic acid, Malate, Myo-inosytol, Pentonic acid, Phosphoric acid, Proline, Pyroglutamic acid, Suberyl-glycine, Succinic acid, Sucrose, Urea

Table 2.3: Characterization of modules and network when estimating networks by GL and by using Density, Centralization and Heterogeneity (edges' stability has been used as weight) in plant data

		$\mathcal{N}_O$[a]			
Module	Density	Centralization	Heterogeneity	Nr Nodes	Color Coded
Module 1[b]	0.51	0.33	0.37	7	Brown
Module 2[c]	0.39	0.25	0.40	6	Yellow
Module 3[d]	0.23	0.19	0.48	10	Blue
Module 4[e]	0.22	0.19	0.46	10	Turquoise
Complete	0.05	0.07	0.71	54	

[a] Network using the original metabolite values

[b] Module 1: Glutamine, Isoleucine, Lysine, Pyroglutamic acid, Threonine, Tyrosine, Valine

[c] Module 2: Allantoine, Galactinol, Glycine, Leucine, Raffinose , Succinic acid

[d] Module 3: 2-Aminoadipic acid, Alanine, Aspartate, Beta-alanine, Citrate, Methionine, Phenylalanine, Proline, Serine, Tryptophan

[e] Module 4: 5-Aminocarboxy-4,6-dihydroxypyrimidine, Ascorbic acid, Fructose, Fructose-6-phospate, Glucose-6-phosphate, Glyceric acid, Glycerol, Monomethylphosphate, Pentonic acid, Phosphoric acid, Sucrose, Threonate, Urea

		$\mathcal{N}_G$[f]			
Module	Density	Centralization	Heterogeneity	Nr Nodes	Color Coded
Module 5[g]	0.49	0.33	0.41	7	Yellow
Module 6[h]	0.37	0.23	0.31	9	Turquoise
Module 7[i]	0.29	0.33	0.48	8	Brown
Module 8[j]	0.25	0.21	0.44	9	Blue
Complete	0.05	0.07	0.77	48	

[f] Network using Genotypic-related information

[g] Module 5: Glutamine, Glyceric acid, Isoleucine, Leucine, Monomethylphosphate, Valine, Xylose

[h] Module 6: Ascorbic-acid, Aspartate, Glycine, Phenylalanine, Proline, Pyroglutamic acid, Raffinose, Serine, Succinic acid

[i] Module 7: Alanine, Fructose-6-phospate, Fumarate, Glucose-6-phosphate, Lysine, Threonate, Threonine, Tyrosine

[j] Module 8: Allantoine, Asparagine, Galactinol, Glycerol, Pentonic acid, Phosphoric acid, Suberyl-glycine, Sucrose, Urea

As seen in Table 2.2, the density and centralization of the complete networks are relatively low. The network $\mathcal{N}_G$ has a density of 0.21 and centralization of 0.16. The network $\mathcal{N}_O$ has a density of 0.06 and a centralization of 0.04. With regard to heterogeneity the two networks are similar ($\mathcal{N}_G$ 0.44 and $\mathcal{N}_O$ 0.45). The two most dense modules of $\mathcal{N}_G$ have a moderate density of 0.62 and 0.53, but still small centralizations (0.15 and 0.26). The modules of $\mathcal{N}_O$ have small densities, namely smaller than 0.33.

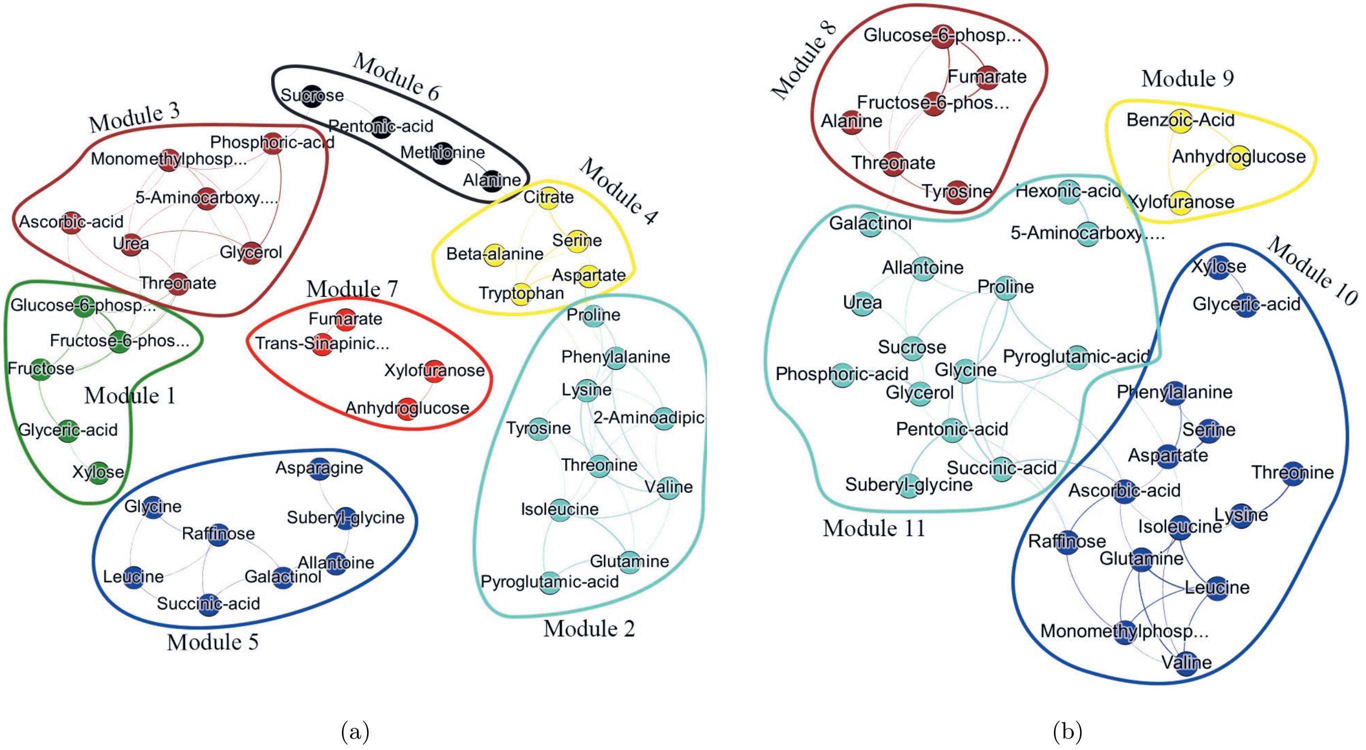


Fig. 2.3: Sparse plant metabolite networks estimated by WGCNA with (a) the original metabolite values and (b) the Genotype related metabolite values. Modules have been color coded as indicated by column "Color Coded" in Table 2.2 in their corresponding network ($\mathcal{N}_O$ or $\mathcal{N}_G$). Colors have been selected by the two-step dynamic hybrid algorithm implemented in the R-package WGCNA

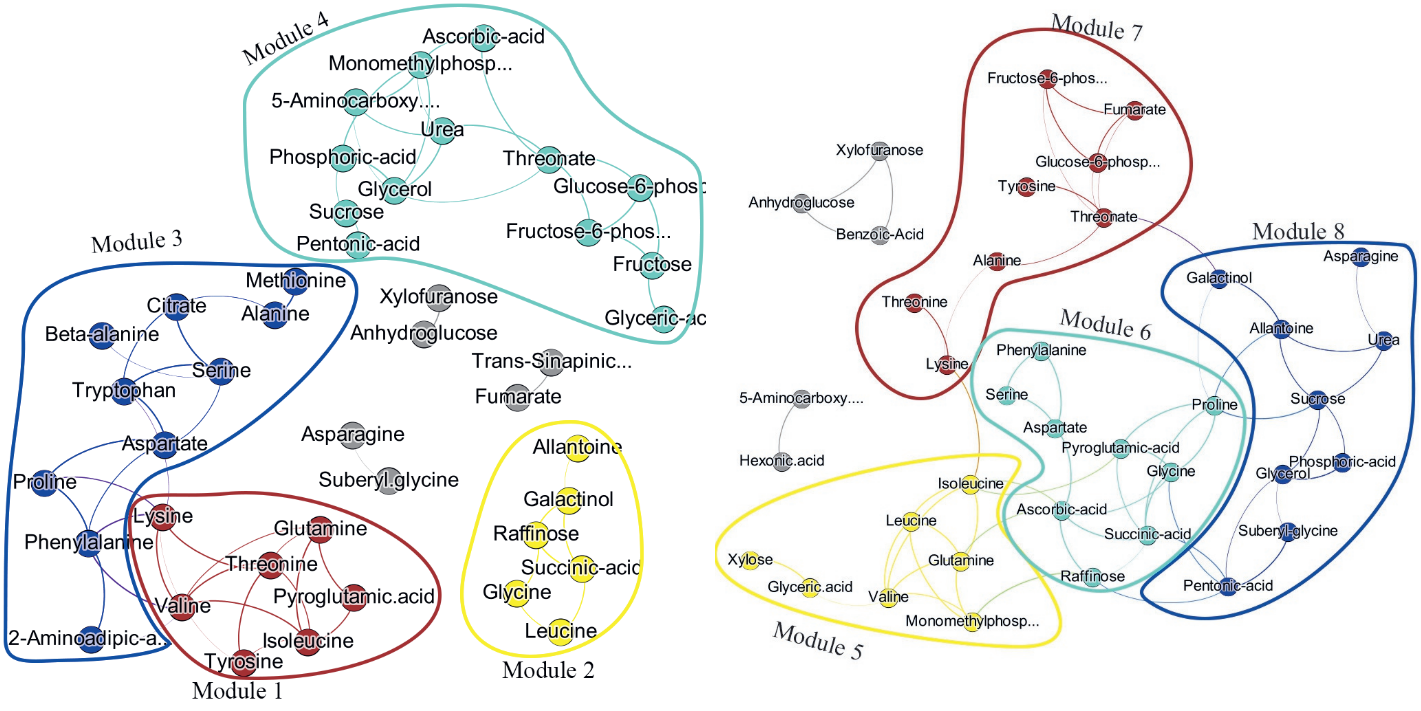
In Figure 2.3, the top 5% of the strongest edges are visualized for the network $\mathcal{N}_O$ and $\mathcal{N}_G$. Here, a threshold of 0.36 for $\mathcal{N}_O$ and of 0.73 for $\mathcal{N}_G$ was used for keeping the top 5% of the edges.

Networks estimated by GL

Next we consider the GL approach for estimation of networks. The regularization parameter λ controlling the network's sparsity was selected by using StARS (H. Liu et al., 2010). We randomly draw 100 subsamples of size 22 for estimating λ. A disagreement allowance of 5%, gave a $\lambda_{\mathcal{N}_O}$ of 0.82 and a $\lambda_{\mathcal{N}_G}$ of 0.94. In Figure 2.4, the results are depicted. Here, the edges' thickness and transparency denotes the edge's stability i.e. frequency of edge occurrence in the 100 datasets. With regard to density, centralization and heterogeneity (Table 2.3) the two networks gave similar values. However, there was not much overlap between the identified modules of $\mathcal{N}_G$ and $\mathcal{N}_O$ obtained by the GL approach. Additionally, there does not seem to be much overlap between the modules obtained by different network estimation methods (WGCNA vs GL).

Top connected plant metabolites:

In Table 2.4, the top connected plant metabolites for the $\mathcal{N}_O$ and $\mathcal{N}_G$ networks which were estimated by WGCNA or GL are given. In general, the GL approach yielded higher degrees of the metabolites for the $\mathcal{N}_G$ than for the $\mathcal{N}_O$ network. This was not the case when the network was estimated by using WGCNA. Apart from small differences, the two methods (WGCNA and GL) appeared to give similar lists of top connected metabolites but the order was different. Additionally, the top



(a) (b)

Fig. 2.4: Estimated metabolite networks for plant data based on GL. In (a), the estimated network is based on the original metabolite values, whereas in (b) Genotype related metabolite values have been used. Different modules have been color coded by the colors pinpointed in column "Color Coded" of Table 2.3. The colors selection was guided by the two-step dynamic hybrid algorithm implemented in the R-package WGCNA

connected metabolites of $\mathcal{N}_G$ were different from the ones of $\mathcal{N}_O$ in both WGCNA and GL cases. These results were also observed in Figures 2.3 and 2.4.

Table 2.4: List of top connected plant metabolites for network estimation using WGCNA and GL. Results are displayed for the $\mathcal{N}_O$ and $\mathcal{N}_G$ networks. Green color denotes that the metabolite appears in both $\mathcal{N}_O$ and $\mathcal{N}_G$ networks. Blue is for metabolites that appear only in the list of $\mathcal{N}_O$ network and violet for metabolites that appear only in the list of the $\mathcal{N}_G$ network

WGCNA			
$\mathcal{N}_O$[a]		$\mathcal{N}_G$[b]	
Metabolite	Degree	Metabolite	Degree
Lysine	8	Threonate	6
Threonine	7	Succinic-acid	6
Threonate	7	Isoleucine	6
Valine	6	Ascorbic-acid	6
Phenylalanine	6	Sucrose	5
Isoleucine	6	Proline	5
Urea	5	Pentonic-acid	5
Monomethylophosphate	5	Glycine	5
Glycerol	5	Glutamine	5
Fructose-6-phosphate	5	Valine	4
5-Aminocarboxy-4,6-dihydroxypyrimidine	5	Pyroglutamic-acid	4
Tryptophan	4	Monomethylophosphate	4
Serine	4	Raffinose	4
Phosphoric-acid	4	Leucine	4
Raffinose	4	Allantoine	4

GL			
$\mathcal{N}_O$[a]		$\mathcal{N}_G$[b]	
Metabolite	Degree	Metabolite	Degree
Valine	6	Threonate	6
Threonine	5	Succinic-acid	6
Threonate	5	Proline	6
Phenylalanine	5	Isoleucine	6
Monomethylophosphate	5	Ascorbic-acid	6
Lysine	5	Valine	5
Isoleucine	5	Sucrose	5
Glycerol	5	Pentonic-acid	5
Urea	4	Glycine	5
Tryptophan	4	Glutamine	5
Serine	4	Pyroglutamic-acid	4
Glutamine	4	Monomethylophosphate	4
Aspartate	4	Raffinose	4
Succinic-acid	3	Leucine	4
Proline	3	Glycerol	4

[a] $\mathcal{N}_O$ network using the original metabolite values

[b] $\mathcal{N}_G$ network using Genotypic-related information

2.3.2 Epidemiological Study (DILGOM)

The other metabolomics dataset used has been measured in the epidemiological cohort DILGOM. A detailed description of the study and of the metabolomic dataset can be found in (Inouye et al., 2010) and (Kettunen et al., 2012). We excluded subjects who were diagnosed with diabetes, who received cholesterol medication, or who had outlying values for fasting glucose levels (more than 10 mmol/l). In addition only subjects with complete data were considered. After excluding these samples, we had 419 subjects (202 males and 217 females) aged between 25 and 74 years (median 53). The metabolomic data were measured by nuclear magnetic resonance (^{1}H NMR) and comprise absolute quantitative measurements on 137 serum metabolites. Because of high correlation we removed 78 lipid particle subfractions and only the total lipid concentrations per particle size were used. Additionally, one more metabolite (FALEN) was excluded since its measurements were not completely trusted. Our final dataset consisted of $P = 58$ metabolites: 25 Lipoproteins, 13 Lipids and Fatty Acids, 9 Amino acids and 11 other small metabolites, e.g. involved in glycolysis. We adjusted all the metabolites for Diastolic Blood Pressure (DBP) and Blood Pressure Medication (BPM) (binary) by linear regression. In the rest of this section we will denote these adjusted metabolites as metabolites. We are interested in the network of metabolite concentrations and in the part of the metabolite concentrations related to BMI.

Since the metabolites and BMI depend on age and sex, these variables were included in model 2.2 as $(\mathbf{X}_1, \mathbf{X}_2)$. The continuous BMI values were categorized into three equally sized classes (1^{st} thirtile = 24.38, 2^{nd} thirtile = 27.56). Thus in Eq. 2.2, the total number of covariables is $m = 3$ and the variable $\mathbf{X}_3$ represents the indicator variables for the three BMI categories. The model 2.2 includes the main effects, first and second order interactions. Specifically for the p^{th} metabolite, the following equation was used,

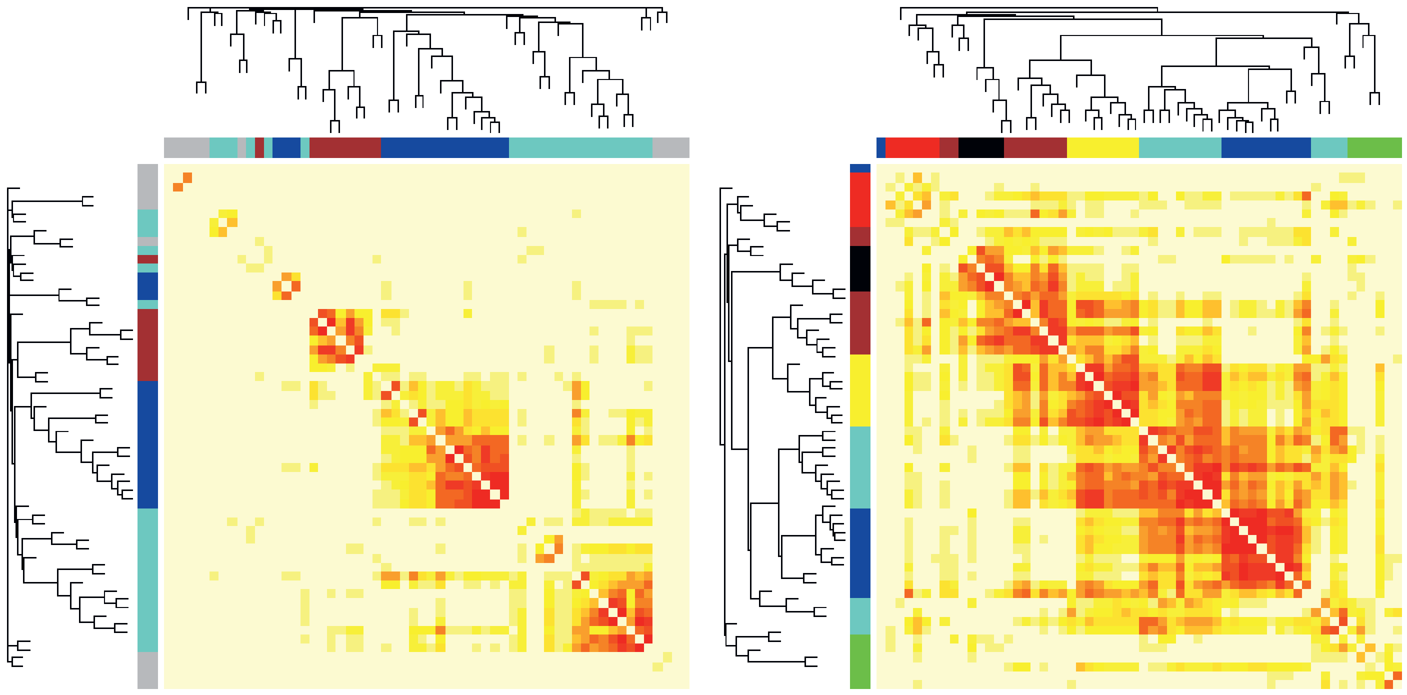
$$\tilde{\mathbf{Y}}^{(p)} = \hat{\boldsymbol{\beta}}_1^{(p)}\mathbf{BMI}+\hat{\boldsymbol{\eta}}_{10}^{(p)}\mathbf{BMI}\circ\mathbf{Age}+\hat{\boldsymbol{\eta}}_{01}^{(p)}\mathbf{BMI}\circ\mathbf{Sex}+\hat{\boldsymbol{\eta}}_{11}^{(p)}\mathbf{BMI}\circ\mathbf{Age}\circ\mathbf{Sex}. \quad (2.11)$$

Analogously to the plant application, networks of metabolites containing the original metabolite values $(\mathbf{Y}^{(p)})$ are denoted by $\mathcal{N}_O$, and networks of metabolites containing the relevant information on BMI $(\tilde{\mathbf{Y}}^{(p)})$ by $\mathcal{N}_B$.

Networks estimated by WGCNA

For estimation of the intensity matrix, we first tried to determine a soft-thresholding parameter γ for which the network had a scale free topology. For both networks (the original and BMI related metabolite values), the scale-free topology did not hold for reasonable powers, i.e. γ less than 12. Therefore, $\gamma = 3$ was chosen based on the sample size ($n = 419$) and the number of metabolites ($P = 58$) as described in Section 2.2.2. The absolute values of Pearson's correlation coefficients are visualized in Figure 2.5; darker red colors represent strong correlations, and lighter colors weaker correlations. The heatmap of $\mathcal{N}_B$ shows several larger blocks of highly correlated metabolites while the heatmap of $\mathcal{N}_O$ shows several small and distinct clusters.

For visualizing and depicting edges in both networks we used the following thresholds: 0.22 for $\mathcal{N}_O$ and 0.66 for $\mathcal{N}_B$. These thresholds correspond to the top 10% of the edges (Figure 2.6). For identifying interconnected modules, we applied average linkage hierarchical clustering and obtained dendograms. Modules were defined as branches of the dendogram and were identified using the two-step dynamic hybrid algorithm. The descriptives of the two networks are given in Table 2.5. The complete network $\mathcal{N}_B$ has a larger density (0.23 versus 0.07) and centralization (0.17 versus 0.11) and a lower heterogeneity value (0.50 versus 0.76) than the complete network $\mathcal{N}_O$. The first three modules of $\mathcal{N}_B$ are more dense than the modules of N_O (0.63 to 0.77 versus 0.15 to 0.32). The centralization and the heterogeneity of the modules of N_G are smaller than of the modules of N_O. In both cases,



(a)

(b)

Fig. 2.5: Correlation matrix plot. The plots were generated for (a) $\mathcal{N}_O$ using the original metabolite values, and (b) $\mathcal{N}_B$ using BMI information. Dendrograms were obtained by ordering metabolites using hierarchical clustering. Modules of interconnected metabolites correspond to square blocks along the diagonal, while deep red colors denote strong correlations (on absolute value). Metabolites belonging in the same module are color coded by the colors (coming from the two-step dynamic hybrid algorithm) that are indicated by the color band below each dendrogram

centralization exhibits relatively low values denoting that there is not a dominant metabolite in each of the modules.

Table 2.5: Characterization of modules and network in humans with WGCNA using Density, Centralization and Heterogeneity

		$\mathcal{N}_O$[a]			
Module	Density	Centralization	Heterogeneity	Nr Nodes	Color Coded
FA/LDL[b]	0.32	0.19	0.45	17	Blue
HDL[c]	0.31	0.23	0.55	9	Brown
VLDL/AA[d]	0.15	0.19	0.72	22	Turquoise
Complete	0.07	0.11	0.76	58	

[a] Network using the original metabolite values

[b] FA/LDL: APOB, DHA, FAW3, FAW3FA, FAW6,IDLC, IDLL, LA,LDLC, LLDLL,MLDLL, PC,SERUMC,SLDLL, SM, TOTPG, XSVLDLL

[c] HDL: ALB, APOA1, HDL2C, HDLC, LHDLL, MHDLL, SHDLL, XLHDLL

[d] VLDL/AA: ALA, FAW6FA, GLC, GP, HDL3C, ILE, LAC, LDLD, LEU, LVLDLL, MUFA, MVLDLL, PHE, PYR, SERUMTG, SVLDLL, TOTFA, TYR, VAL, VLDLD, XLVLDLL, XXLVLDLL

		$\mathcal{N}_B$[b]			
Module	Density	Centralization	Heterogeneity	Nr Nodes	Color Coded
VLDL[e]	0.77	0.11	0.11	8	Yellow
HDL[g]	0.67	0.15	0.15	5	Black
LDL/IDL[h]	0.63	0.18	0.35	11	Blue
FA/Lipids[i]	0.59	0.16	0.19	13	Turquoise
FA/Others[j]	0.54	0.15	0.27	9	Brown
FA[k]	0.20	0.20	0.42	6	Green
AA/Lipoproteins[l]	0.19	0.16	0.38	6	Red
Complete	0.23	0.17	0.50	58	

[e] Network using BMI-related information

[f] VLDL: ALA, LVLDLL, MVLDLL, SERUMTG, SVLDLL, VLDLD, XLVLDLL, XXLVLDLL

[g] HDL: ACACE,APOA1,BOHBUT,HDL2C,HDLC

[h] IDL/LDL: ACE, IDLC, IDLL, LDLC, LEU, LLDLL, MLDLL, SERUMC, SLDLL, SM, XSVLDLL

[i] FA/lipids: APOB, FAW3FA, FAW6, GLOL, LA, LAC, LDLD, MUFA, PC, PYR, SHDLL, TOTFA, TOTPG

[j] FA/Others: GLC, GP, HDLD, ILE, LHDLL, PHE, TYR, VAL, XLHDLL

[k] FA: CIT, DHA, FAW3, FAW6FA, GLN, MHDLL

[l] AA/Lipoproteins: ALB, CREA, GLY, HDL3C, HIS, UREA

Table 2.6: Characterization of modules and network in GL using Density, Centralization and Heterogeneity (edges stability has been used as weight) in humans

		$\mathcal{N}_O$[a]			
Module	Density	Centralization	Heterogeneity	Nr Nodes	Color Coded
LDL/IDL[b]	0.71	0.31	0.34	9	Blue
FA[c]	0.65	0.45	0.28	6	Yellow
BCAA/VLDL[d]	0.56	0.23	0.38	9	Turquoise
HDL [e]	0.45	0.27	0.45	6	Brown
Complete	0.07	0.17	0.98	58	

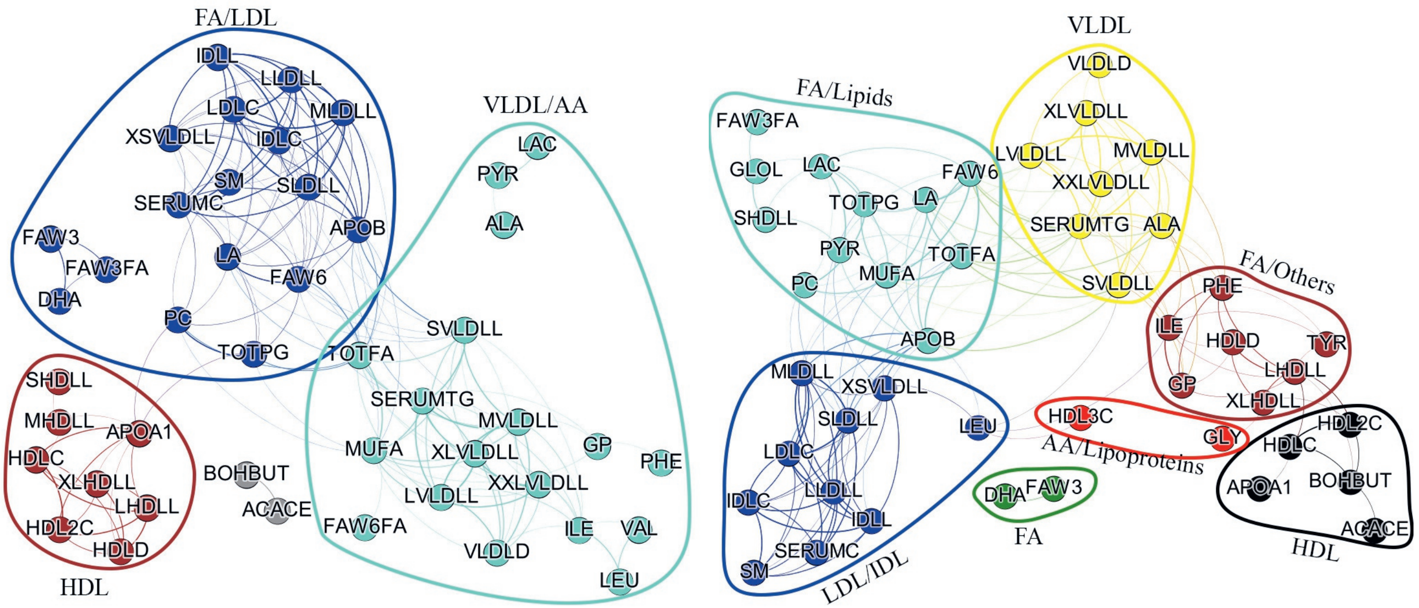
[a] Network using the original metabolite values
[b] LDL/IDL: IDLC, IDLL, LDLC, LLDLL, MLDLL, SERUMC, SLDLL, SM, XSVLDLL
[c] FA: APOB, FAW6, LA, MUFA, SVLDLL, TOTFA
[d] BCAA/VLDL: ILE, LEU, LVLDLL, MVLDLL, SERUMTG, VAL, VLDLD, XLVLDLL, XXLVLDLL
[e] HDL: APOA1, HDL2C, HDLC, HDLD, LHDLL, MHDLL, PC, TOTPG, XLHDLL

		$\mathcal{N}_B$[f]			
Module	Density	Centralization	Heterogeneity	Nr Nodes	Color Coded
LDL/IDL[g]	0.79	0.27	0.24	9	Blue
VLDL[h]	0.61	0.30	0.35	8	Yellow
FA[i]	0.44	0.56	0.45	9	Brown
Lipids/HDL[j]	0.30	0.31	0.42	10	Turquoise
Complete	0.05	0.11	0.98	58	

[f] Network using BMI-related information
[g] LDL/IDL: IDLC, IDLL, LDLC, LLDLL, MLDLL, SERUMC, SLDLL, SM, XSVLDLL
[h] VLDL: ALA, LVLDLL, MVLDLL, SERUMTG, SVLDLL, VLDLD, XLVLDLL, XXLVLDLL
[i] FA: APOB, FAW6, LA, LAC, MUFA, PC, PYR, TOTFA, TOTPG
[j] Lipids/HDL: ACACE, BOHBUT, GP, HDL2C, HDLC, HDLD, ILE, LHDLL, PHE, XL-HDLL

Networks estimated by GL

Next we considered the GL approach. The penalty parameter λ was selected by using StARS (H. Liu et al., 2010). We used a size for the subsamples of 205 (almost 50% of the total sample size) and a disagreement allowance of 2% across the networks. After subsampling 100 times, we obtained a $\lambda_{\mathcal{N}_O}$ of 0.68 and $\lambda_{\mathcal{N}_B}$ of 0.91. Figure 2.7 depicts the results. Here the edge thickness represent the stability of the estimated edges. The evaluation measures are also calculated based on stability (Table 2.6). The density, centralization, and heterogeneity was very similar for both networks. The first modules of each network contained exactly the same metabolites and had also very similar properties. Interestingly the second dense



(a)

(b)

Fig. 2.6: Estimated sparse metabolite networks in humans based on WGCNA. In (a), the network is based on the original metabolite values. In (b) the BMI related metabolite network is displayed. Metabolites have been colored by the color of the module they belong to. Information on the color of each module can be found in Table 2.5. Colors were selected based on the two-step dynamic hybrid algorithm implemented in the R-package WGCNA

module of $\mathcal{N}_G$ (VLDL) and third of the $\mathcal{N}_O$ (BCAA/VLDL) had a large overlap with regard to VLDL particles. Note that VLDL particles were also identified by the WGCNA approach as a module. In Figure 2.1 metabolite concentrations stratified by BMI and Sex are depicted for the VLDL module. It appeared that all these eight metabolites had a high value in obese men. For females the values were much lower for these metabolites except for Alanine. Note that Alanine is also on the border of the cluster (Figures 2.6b and 2.7b) and shares connections with metabolites from other modules. Finally, the third densest module in $\mathcal{N}_B$ (FA) identified by GL exhibits a relatively high centralization. This denotes that some metabolite(s) had higher than average degree (hub metabolite(s)). The metabolite with the highest strength (6.99) and degree (7) as measured by the stability in GL within the module is MUFA (Monounsaturated Fatty Acids). MUFA being connected to all metabolites in the FA module (only Pyruvate is not a direct connection) indicates that it might have the most representative metabolite profile among the rest. LLDLL (Total lipids of large LDL particles), with 8 connections and strength 7.98 is the metabolite with the most connections within the Lipoproteins module. In the VLDL module, LVLDLL has the most connections (6 connections) with strength of 5.98.

Top connected metabolites:

In Table 2.7 the top 15 metabolites are given for combinations of the two approaches (WGCNA and GL) and two types of correlation structures (metabolites and BMI specific metabolites). The degree k of the metabolites was higher for WGCNA than for GL. The same conclusion holds when comparing Figures 2.6b and 2.7b.

2.4 Discussion

In this paper we studied and visualized the correlation structure of two datasets of metabolomics coming from two different studies. The studies differ in types of subjects, study designs and sizes. We considered two methodological approaches to estimate the networks: namely WGCNA based on correlation and GL based on partial correlation. The methods were applied to the metabolite values and to parts of the metabolite values which incorporate the effect of a covariable of interest, e.g. genotype or BMI. We compared the obtained networks in terms of density, centralization and heterogeneity of the relationships between the nodes.

The density of the networks based on the covariable specific part of the metabolites had larger or similar values than the density of the networks based on the metabolites itself. Interestingly, for the metabolites measured in the epidemiological study, a number of lipoproteins (i.e. IDLC, IDLL, LDLC, LLDLL, SERUMC, SLDLL, SM, XSVLDLL) was clustered together when keeping the original or the BMI-specific metabolite values in both network estimation methods. In addition, for the BMI specific part both approaches identified exactly the same module consisting mainly of very large density lipoprotein (VLDL) particles with a density

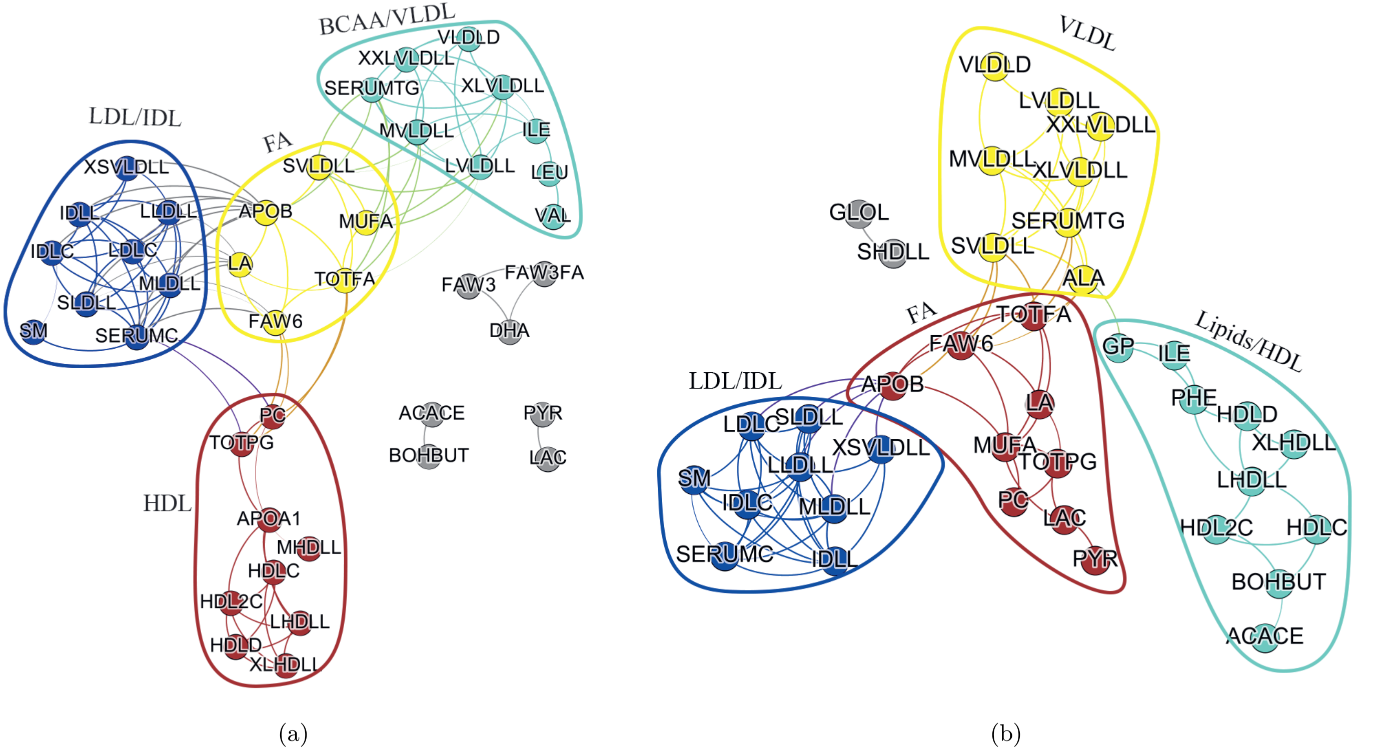


Fig. 2.7: Sparse metabolite networks estimated by GL. The networks were constructed (a) using the original metabolite values, and (b) using the metabolite values driven by variation originating from BMI. Metabolites were colored based on the module they belong to. Information on the membership of each metabolite and on the color they have been colored can be found in Table 2.6

Table 2.7: List of top connected plant metabolites for network estimation using WGCNA and GL . Results are displayed for the $\mathcal{N}_O$ and $\mathcal{N}_G$ networks. Green color denotes that the metabolite appears in both $\mathcal{N}_O$ and $\mathcal{N}_G$ networks. Blue is for metabolites that appear only in the list of $\mathcal{N}_O$ network and violet for metabolites that appear only in the list of the $\mathcal{N}_G$ network

WGCNA				GL			
$\mathcal{N}_O$[a]		$\mathcal{N}_B$[b]		$\mathcal{N}_O$[a]		$\mathcal{N}_B$[b]	
Metabolite	Degree	Metabolite	Degree	Metabolite	Degree	Metabolite	Degree
TOTFA	16	TOTFA	17	APOB	13	LLDLL	9
FAW6	16	SVLDLL	15	SERUMC	12	MLDLL	9
APOB	15	APOB	15	SERUMTG	10	APOB	9
SVLDLL	14	SERUMTG	14	TOTFA	10	SVLDLL	8
SERUMC	14	SLDLL	14	LVLDLL	9	IDLL	8
LA	14	FAW6	14	MVLDLL	9	LDLC	8
MUFA	14	MLDLL	12	LLDLL	9	SERUMTG	8
XSVLDLL	12	MUFA	12	MLDLL	9	SLDLL	7
MLDLL	12	ALA	11	SLDLL	9	IDLC	7
SLDLL	12	LVLDLL	10	IDLC	9	SERUMC	7
SERUMTG	12	MVLDLL	10	XLVLDLL	8	FAW6	7
MVLDLL	11	XSVLDLL	10	IDLL	8	TOTFA	7
IDLL	11	XLVLDLL	9	LDLC	8	MUFA	7
LLDLL	11	IDLL	9	FAW6	8	LVLDLL	6
IDLC	11	ILE	9	LA	8	MVLDLL	5

[a] Network using the original metabolite values

[b] Network using BMI-related information

of 0.77 using Pearson correlation, and a density of 0.61 using stability in the GL approach. This module is characterized by high values for obese men. The relationship between BMI and VLDL concentrations is known. In several studies it has been found that obese people have elevated VLDL concentrations (almost by 50% ;(Magkos, Mohammed, & Mittendorfer, 2008)) compared to lean individuals (Mittendorfer, Patterson, & Klein, 2003; Chan, Barrett, & Watts, 2004; Goff, D'Agostino, Haffner, & Otvos, 2005; Magkos et al., 2008; Magkos & Mittendorfer, 2009). This can be attributed to the hepatic overproduction of VLDL particles (Chan et al., 2004; Ooi et al., 2005) which characterizes obesity. Hepatic overproduction of VLDL particles is also stimulated by atherogenic dislipidemia (which is commonly present in obese people) and promoted by increased liver fat (Grundy, 2004; Klop, Elte, & Cabezas, 2013), also common in obese people. Finally, abdominal fat (known as visceral adipose tissue) and BMI have been found to be positively associated to VLDL particle concentrations and size suggesting again the association between obesity and high levels of VLDL (Sam et al., 2008).

The results for the data from the Arabidopsis desiccation tolerance experiments were harder to interpret. In general, the densities and centralizations of the networks were smaller. This might be the result of the small sample size. Indeed the γ parameter to shrink small values in the WGCNA approach is larger for the experimental than for the epidemiological design (5 and 3 respectively). However for the GL approach there was less difference in the shrinkage, namely the shrinkage parameter was 0.68 and 0.91 for N_O and N_B in the epidemiology study and 0.82 and 0.94 for N_O and N_G in the experimental design. While in the epidemiological study the network specific to BMI provided interesting results, this was less the case for the desiccation tolerance data. One reason may be that the main effect of the genotype on the metabolite variation was smaller than the main effect of the treatment. We presented a treatment corrected metabolic network reconstrution in which the wild type and mutant type genotypes are compared across treatments. This is a sensible analysis in plant genetics (van Eeuwijk, Bink, Chenu, & Chapman, 2010). However, from a seed physiological perspective an analysis correcting for genotype and comparing the treatments across the two genotypes may be more interesting. We will present the results of such an analysis elsewhere.

WGCNA and GL are complementary approaches for metabolite inference since the former recovers meaningful modules and the latter recovers meaningful edges, e.g. for the DILGOM study MUFA is the dominant metabolite in the FA module when GL is used. WGCNA is based on the correlation structure, and the obtained results are therefore straightforward to interpret. To reduce the noise in the data, a soft threshold is often applied to sufficiently shrink small correlations to zero. WGCNA allows detection of modules with high density. GL is based on Gaussian graphical models, in which the conditional independence is inferred by the zero entries in the precision matrix. In order to induce sparsity to the precision matrix and to estimate stable sets of edges with a low false discovery rate, a stability selection approach, StARS, was applied. This can lead to detection of modules with high centralization. For both approaches modules are selected by constructing a dendogram based on average linkage and cutting the branches.

In both cases, WGCNA and GL, the user has to choose specific tuning parameters: the soft threshold in WGCNA and the disagreement allowance and the number of the subsamples for GL with StARS. For a large dataset, StARS might not be needed and the penalty parameter might be chosen based on cross validation making the procedure to be data driven. The impact of the tuning parameters in WGCNA and GL is high; so data driven methods for selecting them would be ideal.

For clear visualization of the constructed networks using WGCNA, arbitrary thresholds were applied to depict the top 5% of the edges in the experimental design and 10% in the human data. These numbers were selected together with plant biologists and epidemiologists so that a set of meaningful edges was recovered by the data. The disagreement allowance parameter in GL was set to 5% in the experimental design and 2% in the epidemiological data for the same reason.

Apart from the two considered methods for estimating networks of metabolites, other well established methods have been used in practice based on conditional independence using regression techniques, ridge regularization and partial correlation. First, a simple approach for estimating sparse networks using regression techniques is by estimating the edge set for each variable by fitting a Lasso model to each variable using the remaining variables as predictors (Meinshausen & Bühlmann, 2006). The non-zero Lasso coefficients identify the adjacent nodes to which each variable is connected to. The shortcoming of this method compared to GL which estimates the network structure simultaneously, is that an edge between two nodes (p and q) might be estimated from p to q but not vice versa. For overcoming this, an AND or an OR rule can be applied. This method asymptotically estimates the set of non zero elements of the precision matrix, but it does not yield the maximum likelihood estimator (Banerjee, Ghaoui, & d'Aspremont, 2008). In contrast to Lasso penalty in GL which estimates entries in the precision matrix (and subsequently in the adjacency matrix) exactly equal to zero, a ridge penalty can also be used in the penalized log likelihood (Ha & Sun, 2014). The elements in the precision matrix will shrink, but will not be exactly zero unless a threshold is chosen for having exactly zeros (Efron, 2012). In this case, GL seems to be advantageous due to the sparsity that is preferred for interpreting the results. Also, by applying a threshold in the ridge-based method, a second parameter should be estimated on top of the ridge penalty λ_{ridge}. This usually involves a two-dimensional grid search approach that can be time-consuming and the optimization problem might not be convex as well. Finally, the edges of the network can be chosen by estimating Pearson's partial correlation coefficient for any given pair of variables directly. Partial correlation eliminates edges that appear from indirect effects which is a desired characteristic (Krumsiek, Suhre, Illig, Adamski, & Theis, 2011). The shortcoming of the method is that it is not applicable in the high dimensional setting. In this case, partial correlation cannot be estimated by either using the linear regression, or the matrix inversion methods. Using a recursive formula is also not possible and is computationally expensive. Therefore, here we used WGCNA which is a commonly used approach and is based on observed correlations which are easy to interpret. GL is also used since it has all desirable features (sparsity, computational speed, can be used in high-dimensional setting) that the other methods fall short of.

For identifying and selecting modules in this paper, the two-step dynamic hybrid algorithm was used with $1 - w_{ij}$ as a dissimilarity measure in WGCNA and one minus stability of the edges (number of edges occurrences from the subsampling scheme) in GL. Another popular clustering method is the modularity optimization. Dividing a network in modules so that the modularity is optimal, will result in many edges within the modules and few edges between modules (Newman & Girvan, 2004). Module identification by using modularity optimization is a known data-driven approach, which is used without specifying any arbitrary parameters. Here the dynamic hybrid algorithm was used instead, because while it takes the modular structure into account (through the dendrogram), it additionally does not have a constant cut-off height, so is able to identify nested clusters by using the cluster shape.

Some issues still need to be addressed. For example network estimation methods should thoroughly be investigated on the sample size sensitivity (large vs small sample sizes) and in changes on tuning parameters (soft-threshold in WGCNA, disagreement allowance in GL). Further work for repeated measurements of the metabolites over time should be considered.

The regression framework to study specific parts of the metabolite concentrations worked well for the epidemiological dataset. We recovered sets of metabolites that are associated to a categorical covariable of interest in the same way. By using a network approach coupled with a module identification method, sets of metabolites which regulate the covariable of interest in the same way can be detected.

3

Estimating Metabolite Networks Subject to Dietary Preferences and Lifestyle

Bartzis, G., Peeters, C.F.W., Uh, H.W., Houwing-Duistermaat, J.J., & Van Eeuwijk, F. (currently under revision for *Metabolomics*)

3.1 Introduction

The metabolome is the complete collection of metabolites, which are intermediates or end products of metabolic pathways associated with cells, tissues or organs (Nielsen & Jewett, 2007). The metabolome captures information from all functional levels of a cell (Nielsen, 2003). It has been used as a tool for biomarker detection, drug discovery and safety, diet strategies and genetic disease testing (Tebani et al., 2016) because it reflects the underlying biochemical activity. Since it dynamically interacts with other molecules and the environment (Beisken, Eiden, & Salek, 2015), it occupies a unique place in systems biology, where an organism is viewed as a complex web of interacting molecular entities (Nielsen & Jewett, 2007). Additionally, metabolites themselves, by being sensitive to variation coming from genetics, time, and environmental stimuli, are widely used for assessing any type of systematic change in biochemical activity (Tebani et al., 2016). The variability induced by these sources of variation, produces fluctuations in the metabolite concentrations that spread through enzymatic reactions and create correlation patterns (Morgenthal, Weckwerth, & Steuer, 2006). Metabolic network analysis tools that recover these correlation patterns have been described in the literature by representing metabolites as nodes in a graph and their relationships as edges connecting the nodes (Morgenthal et al., 2006; Ursem et al., 2008; Weng et al., 2019; Watson, MacNeil, Arda, Zhu, & Walhout, 2013; Floegel et al., 2014; Bartzis et al., 2017).

3.1.1 Objective and Motivation

In this study, our interest is in recovering meaningful metabolite patterns using network analysis, while metabolite measurements are taken repeatedly on the same individuals over time. Previously, we incorporated information on the study design when metabolite networks were estimated (Bartzis et al., 2017). Extending this approach, here we work with repeated metabolite measurements. In this setting, metabolite concentrations of a subject are dependent (due to time) and this dependence should be taken into account when the data are analyzed. We use a linear mixed effects model for the metabolite concentrations, allowing us to estimate time, and subject-specific random effects. When time, diet, and genetics are included in the model, the subject-specific effects represent all remaining unmeasured shared sources of metabolic variation, i.e., lifestyle. Lifestyle can be defined as the collection of smaller environmental effects, like physical activity, sleeping patterns, interests, etc. Since it is thought of in an abstract way, it is usually not quantifiable and common approaches ignore it as a part of random variation.

Here, by having a repeated measures design together with information on diet and genetics, time and lifestyle effects can be estimated. By working in the network framework, our interest is in recovering meaningful metabolite patterns associated with specific dietary preferences and lifestyle. Additionally, working with the estimated time effects allows the recovery of metabolite sets that change similarly over time.

This work is motivated by the DILGOM (DIetary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome) study. More information can be found in (Inouye et al., 2010; Kettunen et al., 2012). This Finnish population study investigates how nutrition, diet, lifestyle, psychosocial factors, environment, and genetics are linked to obesity and the metabolic syndrome.

3.1.2 Contribution

By using dietary preferences we have a better understanding of the extent to which the metabolic patterns are influenced by diet (Pallister et al., 2015). To date, several studies have assessed the interplay between diet and metabolism in the context of nutritional epidemiology (Guertin et al., 2014; Floegel et al., 2014; Pallister et al., 2015; Schmidt et al., 2015; Xu et al., 2010). A standard technique for quantifying dietary patterns is by making use of Food Frequency Questionnaires (FFQ) with Exploratory Factor Analysis (EFA) (Hu, 2002; Newby & Tucker, 2004). By exploring correlation patterns among food items, common underlying dietary factors are identified and a score summarizing dietary patterns is typically determined (Hu, 2002). Here, we use these scores for summarizing dietary information and studying the interplay between metabolites regarding dietary and lifestyle choices, as well as the interrelationship among the metabolites with regard to time.

Despite the extended use of network analysis in modeling metabolic pathways, to our knowledge, only a few other studies have investigated the link between diet and serum metabolites using network analysis (Watson et al., 2013; Floegel et al., 2014; D. D. Wang et al., 2018; Weng et al., 2019). This allows the determination of how habitual factors associate to metabolite classes (lipoproteins, amino acids, etc). In this work, we use lifestyle information and take into account the genetic contribution to metabolite concentrations. While studying the metabolome using network analysis, we address genetic variation by using Polygenic Risk Scores (PRS) (Dudbridge, 2013) for summarizing genetic information.

For network estimation, we use undirected networks, where the relationship between two connected nodes is symmetric. The estimation of metabolite connection patterns in this paper is based on the graphical LASSO (glasso) (Hastie, Tibshirani, Friedman, & Friedman, 2009; Friedman et al., 2008). Compared to methods that are based on the observed correlation structure (and therefore recovering edges based on indirect associations), glasso is based on partial correlation and recovers edges while avoiding spurious associations.

3.1.3 Overview

The rest of the paper is organized as follows. In Section 3.2 the motivating dataset is described. In Section 3.3, we propose an extension to the method of (Bartzis et al., 2017) in the repeated measures setting for selecting information relevant to certain sources prior to network estimation. Additionally, we review existing methods for summarizing genetic and environmental variation. In Section 3.4, we demonstrate how to select specific variation parts in metabolite data, which deviates from standard approaches in nutritional epidemiology by addressing simultaneously metabolite variation induced by time, genetics, lifestyle, and diet. We conclude the article with a discussion in Section 3.5.

3.2 Data

In this Section, we will describe the data used in this study coming from an epidemiological cohort, namely DILGOM, a subset of the FINRISK study (Inouye et al., 2010; Kettunen et al., 2012). In this study, metabolite data measured at two time points were available (2007 and 2014) together with information on food preferences (described by a Food Frequency Questionnaire; FFQ), genetic information (single nucleotide polymorphisms; SNPs), as well as age and biological gender (hereafter indicated as sex). Since only two time points (2007 and 2014) are included in our data, time was considered a binary factor. Similarly, sex was also a binary factor. Finally, we use the genetic information (SNP information) to calculate a genetic quantitative score per subject to capture the different genetic background of individuals. Our interest here is in studying metabolite patterns with regard to dietary and lifestyle choices. As part of the data cleaning process, we excluded subjects who were diagnosed with diabetes and had outlying fasting glucose levels (over 10mmol/l). In addition only subjects with complete information on age, sex and food preferences were further considered. After applying the exclusion criteria, 364 subjects (171 males and 193 females) aged between 25 and 74 years (median 51) at the first time point (2007) and 211 subjects (104 males and 107 females) aged between 32.11 and 81.23 years (median 59.1) at the second time point (2014) were considered for further analysis. For each time point (2007 and 2014), the continuous age values were transformed into a binary factor (1 for individuals younger than 50 years old and 2 for individuals equal to or older than 50 years old). We opted for the age of 50 due to its alignment with the: i) recommended age for certain health screenings and check-ups for monitoring age-related health issues (Chen, Stock, Hoffmeister, & Brenner, 2019), ii) average onset of menopause among women (Bromberger et al., 1997).

3.2.1 Food Frequency Questionnaire

A FFQ was given to the participants of the study in 2007 to record their eating and drinking patterns. The FFQ contained the eating frequency of 40 food items (e.g. pizza, meat, chocolate, etc.) in a scale from 1 (rarely) to 8 (more than 4 times per day) and the daily drinking frequency of 15 beverages under a typical serving (e.g. cups of coffee, glasses of milk, etc.). The FFQ will be subjected to Exploratory Factor Analysis (EFA) to extract factors that correspond to interpretable diets.

3.2.2 SNP data

For computing the individual genetic effect on the subject-specific metabolite profiles, information on approximately 38 million genotyped and imputed SNPs was available for every individual. Since the metabolite variation explained by each SNP separately is rather small, PRS is used as described in Section 3.3.2.

3.2.3 Metabolite data

Metabolite data in both time points were measured by nuclear magnetic resonance and comprise absolute quantitative measurements on 228 serum metabolites (groups of lipoproteins, lipids, amino-acids, fatty acids and others). Metabolites that were expressed as percentages (78 metabolites) were removed and their concentration levels were retained. Additionally, we removed 83 lipid particle subfractions (due to high correlation) and only the total lipid concentrations per particle size were used. Furthermore, 5 more metabolites were removed since they were expressed either as fractions or they were highly correlated with retained metabolites. Finally, 7 metabolites were eliminated for not having information on any possible association with any SNP (Kettunen et al., 2012); hence the data that were considered for analysis consisted of 55 metabolites.

3.3 Methods

Often, one might be interested in metabolite variation from a specific source, such as, for example, diet (Bartzis et al., 2017). By estimating networks based on this source of variation, metabolites that associate to that source of variation in the same way, will share an edge. For estimating networks that contain information on parts relevant to this source of variation, we take two steps: 1) we identify an appropriate model for the responses (metabolites here) and 2) we select the part of the responses that we are interested in to extract a network. A network consists of a set of p nodes (metabolites) connected by a set of edges (relationships between metabolites) and is represented by a symmetric $p \times p$ matrix $\mathbf{A}$ (adjacency matrix) of 1s and 0s depending on whether the corresponding metabolites are connected. In addition, an intensity matrix $\mathbf{W}$ can be considered where the elements represent the intensity of the connection between the nodes (essentially a weighted version of A). In this paper, we consider as intensity the stability of the estimated edges (i.e., the probability of an edge being true, as calculated by the network estimation method in 3.3.3). The number of neighbors of a node i which is the sum of row or column i of matrix $\mathbf{A}$, is called degree. By taking into account both the degree and the intensity of the edges, the strength of node i (s_i) can be calculated as the sum of row or column i of matrix $\mathbf{W}$. Following the estimation of the intensity and adjacency matrices, groups of closely interconnected metabolites are usually identified using a clustering algorithm where the similarity measure is based on $\mathbf{W}$ (as described in Section 3.3.3).

3.3.1 Estimating subject-specific metabolite effects

At the first step, since we work in the repeated measures framework, the correlation between the measurements is modeled using linear mixed effect models with random intercepts, representing the shared unobserved factors.

Let $\mathbf{Y}^{(p)}$ be the vector of concentrations of the pth metabolite over two time-points. Further, assume that $\mathbf{T}$ is the covariable denoting the discrete point in time where the metabolite concentrations were measured for each individual. Finally, we can have m other covariables, e.g. genetics, dietary preferences, age, and sex.

For the pth metabolite we model the within subject correlation by using subject-specific effects from a random-intercepts linear mixed model. For identifying the part of the metabolite concentrations associated to dietary and lifestyle choices we then fit the following model:

$$\begin{aligned}\mathbf{Y}^{(p)} = \beta_0^{(p)} + &\beta_1^{(p)}\mathbf{Age} + \beta_2^{(p)}\mathbf{Sex} + \beta_3^{(p)}\mathbf{T} + \beta_4^{(p)}\mathbf{F} + \beta_5^{(p)}\mathbf{G} + \beta_6^{(p)}\mathbf{Age} \circ \mathbf{Sex} + \\ &\beta_7^{(p)}\mathbf{Age} \circ \mathbf{T} + \beta_8^{(p)}\mathbf{Age} \circ \mathbf{F} + \beta_9^{(p)}\mathbf{Age} \circ \mathbf{G} + \beta_{10}^{(p)}\mathbf{Sex} \circ \mathbf{T} + \\ &\beta_{11}^{(p)}\mathbf{Sex} \circ \mathbf{F} + \beta_{12}^{(p)}\mathbf{Sex} \circ \mathbf{G} + \beta_{13}^{(p)}\mathbf{T} \circ \mathbf{F} + \beta_{14}^{(p)}\mathbf{T} \circ \mathbf{G} + \\ &\beta_{15}^{(p)}\mathbf{F} \circ \mathbf{G} + \mathbf{u}^{(p)} + \boldsymbol{\varepsilon}^{(p)}, \end{aligned} \tag{3.1}$$

where $\boldsymbol{\varepsilon}^{(p)}$ is the random noise, and $\circ$ is the Hadamard product. In model 3.1, $\mathbf{u}^{(p)}$ are the subject-specific effects representing all unmeasured shared factors. $\mathbf{G}$, represents the genetics, and $\mathbf{F}$ the dietary patterns. $\mathbf{Age}$ and $\mathbf{Sex}$ are the vectors containing the age and sex of the subjects, respectively. Note that the subject-specific effects ($\mathbf{u}^{(p)}$) are conditioned on multiple terms, i.e., age, sex, time, genetics, diet, and their interactions. Therefore, this source of metabolic variation is not associated with them, thus is a variable accounting for individual metabolic differences, hence lifestyle. In principle, lifestyle is hard to estimate since it depends on many factors that are not available to us. Here, by having measurements over time we were able to estimate it as the random intercepts of the linear mixed effects model conditioned on all other sources of variation. Finally, the time interval between the repeated measures is the same for all subjects.

The relevant metabolite part related only to dietary and lifestyle choices in the linear mixed model 3.1 that will be used for estimating metabolite networks is given by:

$$\tilde{\mathbf{Y}}_L^{(p)} = \hat{\beta}_4^{(p)}\mathbf{F} + \hat{\beta}_8^{(p)}\mathbf{Age} \circ \mathbf{F} + \hat{\beta}_{11}^{(p)}\mathbf{Sex} \circ \mathbf{F} + \hat{\beta}_{13}^{(p)}\mathbf{T} \circ \mathbf{F} + \hat{\beta}_{15}^{(p)}\mathbf{F} \circ \mathbf{G} + \hat{\mathbf{u}}^{(p)}. \quad (3.2)$$

The quantification of the dietary (F) and genetic (G) parts for inclusion in model 1, using EFA and PRS, is described in Section 3.3.2.

3.3.2 Identifying diets with Exploratory Factor Analysis

Different diet patterns strongly influence disease risks and have an effect on health. Many studies have examined the associations between intakes of individual foods (Hu et al., 1999) and health or lifestyle.

However the intake of one food or nutrient is often correlated with the intake of another (Randall, Marshall, Brasure, & Graham, 1992; Hu et al., 1999). Therefore, dietary patterns can be identified by using the correlation among the foods, typically by using EFA (Slattery, Boucher, Caan, Potter, & Ma, 1998; Hu et al., 1999, 2000; Williams et al., 2000).

EFA is a latent variable model attempting to explain complex relationships between observed variables by using an unobserved structure (Rencher, 2003). The dimension of the latent vector is lower than the dimension of the observable variables. In EFA, we have a set of observed variables (e.g. food preferences) generated by a number of unobserved latent variables (e.g. diets). The idea is to identify and summarize the unobserved variables that explain the dependence between the observed variables.

For p observed variables, and m unobserved factors, let $\mathbf{o}$ be the observed centered eating frequencies, i.e., $\mathbf{o} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_p)^\top$. For notational simplicity, we leave out the notation for observations. Also let $\mathbf{f} = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m)^\top$ be the unobserved diets. Factor analysis is expressing each food frequency as a linear combination of the diets, i.e.,

$$\mathbf{o} = \mathbf{Lf} + \boldsymbol{\epsilon}, \quad (3.3)$$

where $\mathbf{L}$ is the $p \times m$ loadings matrix measuring the dependence of observed variables on factors, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_p)^\top$ is the random error distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ with $\boldsymbol{\Psi}$ being diagonal, $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi})$, and $\mathrm{Cor}(\mathbf{f}, \boldsymbol{\epsilon}) = \mathbf{0}$. Factor analysis is expressing the covariance of the observed variables ($\mathrm{Cov}(\mathbf{o}) = \boldsymbol{\Sigma}$) in terms of $\mathbf{L}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$, i.e., $\boldsymbol{\Sigma} = \mathbf{L}\boldsymbol{\Phi}\mathbf{L}^\top + \boldsymbol{\Psi}$. The number of latent factors can be chosen based on the factor interpretability. In order to do so, we need to carefully examine the loading matrix for different selections of number of factors in combination with usage of the scree plot (Cattell, 1966). A typical practice for making the results more interpretable is by employing factor rotation, where the estimated loading matrix is generally transformed by multiplying it by an orthogonal or non orthogonal matrix. Here, we use the “oblimin” transformation method that allows the factors to be correlated.

The latent diets $\mathbf{F}$ can be quantified by using the “ten Berge” factor scores (Ten Berge, Krijnen, Wansbeek, & Shapiro, 1999). In that way, the correlation of the dietary patterns is preserved when the sample factor score correlations are computed.

Polygenic Risk Scores (PRS)

A popular practice for uncovering the genetic variants that influence metabolite concentrations are the metabolite-based genome-wide association studies, i.e., mGWAS (Raffler et al., 2015). It has been shown that in GWAS, common single nucleotide polymorphisms (SNPs) exhibit significant roles in determining phenotypic variation (Chatterjee, Shi, & García-Closas, 2016). Although separate SNPs typically explain only a moderate proportion, a combination of them can explain a substantial part of the phenotypic variation (Chatterjee et al., 2016; Dudbridge, 2013). Therefore, polygenic risk scores (PRS) are widely used for summarizing genetic effects $\mathbf{G}$ from a set of markers associated to a phenotype of interest.

Typically, for estimating a PRS (denoted as $\mathbf{G}$ here) for a phenotype $\mathbf{Y}$ (e.g. the concentration levels of a metabolite), we have a set of l SNPs ($\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_l)^\top \in \{0, 1, 2\}^l$). For each of the markers, the effect size (η) is determined, e.g. by the estimated coefficients from a linear regression of $\mathbf{Y}$ on each of the l SNPs.

3

For computing the PRS ($\mathbf{G}$), a subset of the top $\tilde{l}$ associated SNPs is used (Euesden, Lewis, & O’reilly, 2015). Then, the linear combination of the top SNPs weighted by their corresponding effect sizes is calculated. For subject i the PRS is computed as:

$$\mathbf{G}_i = \sum_{j=1}^{\tilde{l}} \eta_j \mathbf{S}_{ij}. \tag{3.4}$$

Note that $\mathbf{G}$ contains genetic information related to $\mathbf{Y}$ and can be used for further analysis.

3.3.3 Network estimation using glasso

In this paper, for network estimation we use glasso (Hastie, Tibshirani, Friedman, & Friedman, 2009; Friedman et al., 2008), which is based on partial-correlations using conditional independence and recovers conditional associations between the nodes.

In glasso, it is assumed that the metabolite concentrations follow a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. The network is estimated as the non-zero entries of the precision matrix ($\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$). For estimating the precision matrix, typically a penalized log-likelihood approach is used which produces a sparse estimate of Theta. The penalized version of the log-likelihood that is maximized (Hastie, Tibshirani, Friedman, & Friedman, 2009; Friedman et al., 2008) uses a LASSO penalty as follows:

$$\ell_\lambda(\boldsymbol{\Theta}) \propto \log|\boldsymbol{\Theta}| - \mathrm{tr}(\boldsymbol{K}\boldsymbol{\Theta}) - \lambda||\boldsymbol{\Theta}||_1, \tag{3.5}$$

where $\mathbf{K}$ is the sample covariance matrix of $\tilde{\mathbf{Y}}_L$ and λ is a non-negative tuning parameter controlling the sparsity of the estimated precision matrix $\hat{\mathbf{\Theta}}$. Here we use the stability approach for regularization selection (StARS) (H. Liu et al., 2010) for obtaining the regularization parameter making the network sparse and replicable under random sampling. In StARS multiple overlapping subsamples of the data are selected and sparse networks are estimated for λ values in a grid. For the optimal λ, the stability of an edge can be calculated as the observed relative frequency of it being estimated over the subsamples.

Module identification

A network usually consists of a set of modules that have closely interconnected metabolites. A typical way of identifying them is by using the two-step dynamic hybrid algorithm (Langfelder et al., 2008) on the metabolite dendrogram resulting by using $\mathbf{W}$ as the similarity matrix. An alternative way for module identification is by using the Girvan-Newman algorithm (Newman & Girvan, 2004) based on edge-betweenness implemented in the rags2ridges R-package (Peeters, Bilgrau, & van Wieringen, 2022). Using different module identification algorithms allows us to get a better understanding of the network's modular structure and possibly identify sets of nodes that consistently emerge. In this study, we base $\mathbf{W}$ on StARS. We first recover the adjacency matrix $\mathbf{A}$ for the estimated optimal λ resulting to a stable network under random subsampling. The symmetric $\mathbf{A}$ matrix contains 1s and 0s depending on whether the corresponding nodes are connected for the optimal λ value. The non-zero entries (edges) are then weighted by the relative estimated frequency of the edge being present over all subsamples used. Subsequently, the matrix containing the relative estimated frequency of the stable estimated edges is $\mathbf{W}$ and can be used for module identification.

3.3.4 Network Characterization

We now consider three measures to describe a network or a module, namely density, centralization and heterogeneity (Dong & Horvath, 2007). For a square symmetric $p \times p$ matrix $\mathbf{M}$, let s be the strength of a node (row sum of $\mathbf{W}$) and $\bar{s}$ the mean of all s. The quantities in Table 3.1 are then computed for network or module characterization. For density, a value close to one indicates high interconnectedness between the nodes in the network/module. High values in Centralization, denote a star shaped network, i.e., the network contains one highly connected node. Finally, the heterogeneity indicates the amount of edge diversity in the network/module.

Table 3.1: Measures that can be used for describing a network or part of a network

Quantity	Formula	High values denote that...	Range
Density	$\sum_i \sum_j \frac{M_{ij}}{p(p-1)}$	$\mathbf{M}$ is highly interconnected	$[-1, 1]$
Centralization	$\frac{1}{p}\,(\max(s) - \bar{s})$	$\mathbf{M}$ contains hub node(s)	$[-1, 1]$
Heterogeneity	$\frac{\sqrt{\mathrm{var}(s)}}{\bar{s}}$	the values in $\mathbf{M}$ are diverse	$[0, \infty)$

3.4 Application to Data

In this section, we will use the methods of section 3.3 for analyzing, visualizing, and evaluating the conditional correlation structure of the metabolite data subject to: i) dietary and lifestyle variation, ii) time variation, while addressing for individuals genetic background. Prior to that, we need to estimate dietary ($\mathbf{F}$) and genetic information ($\mathbf{G}$) for adding them together with age, sex, and time in model 3.1, while lifestyle is estimated by the empirical Bayes estimates of the random intercepts of this model.

The FFQ data were used for identifying latent dietary patterns emerging from the complex relationships between the 55 observed eating/drinking items. To estimate dietary information we used EFA. The loading matrix $\mathbf{L}$ was estimated and can be seen in Table 3.2. The visual inspection of the scree plot (Figure 3.1) revealed that the number of possible latent diets that could be recovered from the data were 5 to 6. Here, we chose 6 diets. For the rest of the paper we refer to these 6 diets as: F_1, F_2, F_3, F_4, F_5, F_6.

The dietary patterns were determined after closely examining the factor loadings: a) fast food (F1), b) vegetarian (F2), c) high caloric (F3), d) fish (F4), e) juice (F5), and f) balanced (F6) respectively. Finally, we assumed that individuals do not change dietary patterns in a 7-year period time. That means the dietary scores are the same in the two time points.

For estimating the individuals genetic background, we first selected 48 SNPs based on a mGWAS study (Kettunen et al., 2012) on 8,330 Finnish individuals (DILGOM was part of the study) and Linkage Disequilibrium pruning for dealing with correlation between SNPs. In Equation 3.4, the PRS ($\mathbf{G}_i$) was computed for every available metabolite with η obtained from the supplementary material of (Kettunen et al., 2012).

3.4.1 Metabolite networks for separate timepoints

In order to analyze the data, we first estimated metabolite networks with regard to the different time points, i.e., 2007 and 2014. The estimated metabolite networks can be seen in Figure 3.2. For those networks, the metabolites were corrected for age and sex differences by keeping the residuals of a linear model with each metabolite as response and Age, Sex, and their interaction as predictor variables.

In the case of the network concerning the metabolite measurements in 2007, three modules have been identified using the Girvan-Newman algorithm. The first mainly consisted of VLDLs, the second of Lipoproteins, while the last one had mainly high density lipoproteins. For the metabolite network when using the 2014 measurements, four modules were identified using the Girvan-Newman algorithm. Those modules had high overlap with the ones from the 2007 network.

As the data concern metabolite measurements of the same individuals over different time points, the intra-subject correlation is not utilized when networks are estimated separately. Additionally, repeated measures allow us to estimate time effects and study interactions between time, genetics, and dietary patterns. By utilizing a detailed mixed model which suitably models the repeated measure data, random intercepts are estimated and can be interpreted as the residual metabolic variation not attributed to dietary, demographic or genetic information, i.e., lifestyle. The estimated lifestyle information can further be used together with dietary information to estimate metabolite networks subject to those two sources of metabolic variation.

3.4.2 Metabolite networks subject to dietary patterns and lifestyle

For estimating networks of metabolites with respect to dietary and lifestyle information, model 3.1 was first fitted on the data. Then, for the pth metabolite, the values $\tilde{\mathbf{Y}}_L^{(p)}$ (Model 3.2) were used for network estimation. To estimate a metabolite network using glasso, we first selected the tuning parameter λ controlling the network sparsity in glasso ($\lambda = 0.533$).

Using the two-step dynamic hybrid algorithm, 12 modules were identified (VLDL1, VLDL2, lipid metabolism, lipoproteins, ω-3 FA, carbohydrate metabolism, glycogenesis, AA, ketone bodies, BCAAs, HDL; Figure 3.3a). In Table 3.3, the clusters of interconnected metabolites were characterized by using our descriptive measures (density, centralization and heterogeneity) for clusters that contain five or more metabolites. The complete network displays a small value in terms of density (0.10) and high value for heterogeneity (0.64) compared to the identified modules (high densities and low heterogeneity). This implies a good module separation (high density within modules compared to low for nodes in different modules).

Using the Girvan-Newman algorithm, 8 modules were identified (VLDL, lipoproteins, ω-3 FA, HDL/ketone bodies/lipid metabolism, AA; Figure 3.3b). In Table 3.4, the clusters of interconnected metabolites were again characterized by density, centralization and heterogeneity. As in the two-step dynamic hybrid algorithm, here the complete network displays again a small value in terms of density (0.10) and high value in heterogeneity (0.64). Contrarily to the networks for the separate time points, the identified modules had high densities and low heterogeneity pointing again to good module separation.

3.4.3 Comparison between networks for separate time points and networks subject to dietary patterns and lifestyle

Compared to the case of separate networks per time point, here we estimate different parts of metabolic variation before we reconstruct metabolic networks. This resulted in better separated networks, i.e., higher number of modules. The estimated modules can be better identified as the different metabolic classes, i.e., amino acids, VLDLs, HDLs, ω-3 fatty acids, etc.

For the estimated modules using the two-step hybrid algorithm, the HDL module (Cit, HDL.C, HDL.D, HDL2.C, L.HDL.L, LDL.D, XL.HDL.L), the lipoproteins module (IDL.C, IDL.L, L.LDL.L, LDL.C, S.LDL.L, SM), and one of the VLDL modules (L.VLDL.L, M.VLDL.L, Serum.TG, VLDL.D, XL.VLDL.L, XXL.VLDL.L) had the highest amount of metabolites (7, 6, and 6 respectively). By inspecting the metabolite profiles (Figure 3.4), it can be observed that the metabolic profiles for the HDL module were not as homogeneous as for the other modules (density was estimated at 0.46). Conversely the profiles for the lipoproteins and VLDL modules were much more homogeneous, seen also by their density (0.96 and 0.86, respectively).

Interestingly, using the Girvan-Newman algorithm for module identification, the lipid metabolism module was clustered together with the ketone bodies and the high-density lipoproteins resulting in 16 metabolites within the cluster. The lipoproteins module in this case contained 11 metabolites and had again high density (0.76).

The VLDL and HDL modules appeared to have on average opposite associations to every diet. The negative association might stem from HDL transporting very-low-density lipoprotein to the liver, where they are broken down. Mainly, by following a fast-food, a vegetarian, or a high-caloric diet, a negative association to HDL was observed in our data (Figures 3.4 and 3.5).

3.5 Discussion

In this work, our interest was on recovering metabolite networks under a repeated measures setting. By having information on various sources of variation (age, sex, time, genetic background, dietary preferences), lifestyle was able to be estimated as the random effects of a linear mixed effects model with the metabolite concentrations as response variable. By estimating time effects and quantifying lifestyle, metabolite networks were estimated with regard to lifestyle and dietary preferences while addressing for individual's genetic differences. For network estimation, we considered the glasso method which is based on conditional independence. The network estimation method was applied to human metabolite data and interconnected modules having the same relationship to diets and lifestyle were identified using two methods: i) the two-step dynamic hybrid algorithm, and ii) the Girvan-Newman algorithm. Obtained networks and modules were described in terms of density, centralization, and heterogeneity.

As the data were collected at two time-points (2007 and 2014), we estimated metabolite networks for the separate time-points, as a benchmark, for comparing to the more elaborate model which utilizes information on time, genetics, and dietary patterns. We observed that the separate networks had fewer, more dense modules. Contrarily, a network accounting for different sources of variation resulted in more modules that were also more homogenous in terms of constituent metabolites. When information related to dietary preferences and lifestyle was retained in metabolite networks, several groups of biologically associated metabolites were clustered together.

By working on the repeated measures setting, networks subject to time variation can also be estimated. Identified modules will contain metabolites that change similarly over time. In order to perform such analysis, the metabolite part related to time variation that can be used for network estimation is given by:

$$\tilde{\mathbf{Y}}_T^{(p)} = \hat{\beta}_3^{(p)}\mathbf{T} + \hat{\beta}_7^{(p)}\mathbf{Age} \circ \mathbf{T} + \hat{\beta}_{10}^{(p)}\mathbf{Sex} \circ \mathbf{T} + \hat{\beta}_{13}^{(p)}\mathbf{T} \circ \mathbf{F} + \hat{\beta}_{14}^{(p)}\mathbf{T} \circ \mathbf{G}. \tag{3.6}$$

Although, several studies have examined the interplay between diet and metabolism, to our knowledge, this is the first attempt studying metabolite patterns in the network framework while simultaneously modeling diet ($\mathbf{F}$), lifestyle ($\mathbf{u}$), and genetics ($\mathbf{G}$), when concentrations are measured over time. Under this design, metabolite measurements are dependent and this dependence should be taken into account when the data are analyzed. By using linear mixed effects models, we were able to decompose and select the part of metabolic variation relevant to specific covariables, i.e., $\mathbf{F}$ and $\mathbf{u}$. Established relationships were identified and metabolites were separated by their different biochemical classes.

However, some limitation should be noted. First, we assumed the individual dietary scores, as estimated by the EFA, to be the same between the two time points. Overall dietary patterns tend to be relatively stable over time, while it is suggested that in repeated measures studies, dietary information to be reevaluated after at least seven years (Weismayer, Anderson, & Wolk, 2006). Second the FFQ contains self-reported data, which may have limitations such as recall bias and social desirability bias. Despite the limitations, FFQs are still valuable tools for assessing dietary intake in large-scale epidemiological studies.

Zooming into variation of specific covariables allows us interpreting and inferring different metabolite aspects. Using this framework while working on the metabolome, more can be done. For example, in metabolite identification and characterization, when an unidentified metabolite is included in the network, its properties regarding different aspects can be deduced by carefully examining the edges connected to the metabolite, with respect to different variation sources. Working in the same framework for the reconstruction of metabolite networks in humans and plants (Bartzis et al., 2017), in the future we plan to use information from the graphical structure of lower leveled omic sources (gene or marker level) besides accounting for the study design. This will allow us to use an extra level of information for reconstructing metabolite networks.

Table 3.2: Loadings matrix. Measures the dependence of observed variables on factors. Loadings with absolute value above 0.25 have been indicated.

Food item	Fast Food (F1)	Vegetarian (F2)	High Caloric (F3)	Fish (F4)	Juice (F5)	Balanced (F6)
Chocolate	**0.60**					
Other candies	**0.55**					
Sweet biscuits	**0.52**	**0.29**				
Other sweet pastry	**0.45**					
Salty snacks	**0.44**	**-0.30**				
Ice cream puddings	**0.43**					
Cereals or muesli	**0.37**		**-0.29**			
Store boughtready meal	**0.37**					
Pizza	**0.32**	**-0.28**				
Flavoured yoghurt	**0.31**					
Yeast bread.graham bread						
Cola light/day						
Chocolate milk/day						
Fruits		**0.55**				
Fresh or frosen berries		**0.53**				
Porridge		**0.49**				
Sweet coffeebread or pies	**0.31**	**0.43**				
Low fat cheese		**0.36**				**0.28**
Burgers	**0.29**	**-0.34**				
Rye bread & rye crisp		**0.31**				
Non-flavoured yoghurt		**0.31**				
Cooked vegetables		**0.30**				
Sour milk/day		**0.28**				
Vegetarian food		**0.26**				
Energy drink/day						
Sausages			**0.62**			
Cutlets			**0.59**			
Meat			**0.43**			**0.33**
Cooked or smashed potatoes		**0.40**	**0.42**			
Roasted potatoes or french fries			**0.42**			
Salty pies and pastry	**0.30**		**0.36**			
White bread			**0.29**			
Eggs			**0.29**			
Coffee/day			**0.26**			
Milk/day						
Other cheese						
Coffee/day			**0.26**			
Tap water/day						
Soft drink/day						
Low alcohol /day						
Fish and other fishfood combined				**0.91**		
Salmon & rainbow trout				**0.64**		
Other fish				**0.58**		
Herring				**0.48**		
Tea/day						
Bottled water/day						
Fruit and berry juices					**0.76**	
Poultry meat						**0.60**
Cold cuts			**0.30**			**0.44**
Pasta or rice	**0.26**					**0.36**
Fresh salad. Fresh vegetables		**0.29**				**0.32**
Salad dressing or oil						
Low calory soft drink/day						
Well water/day						

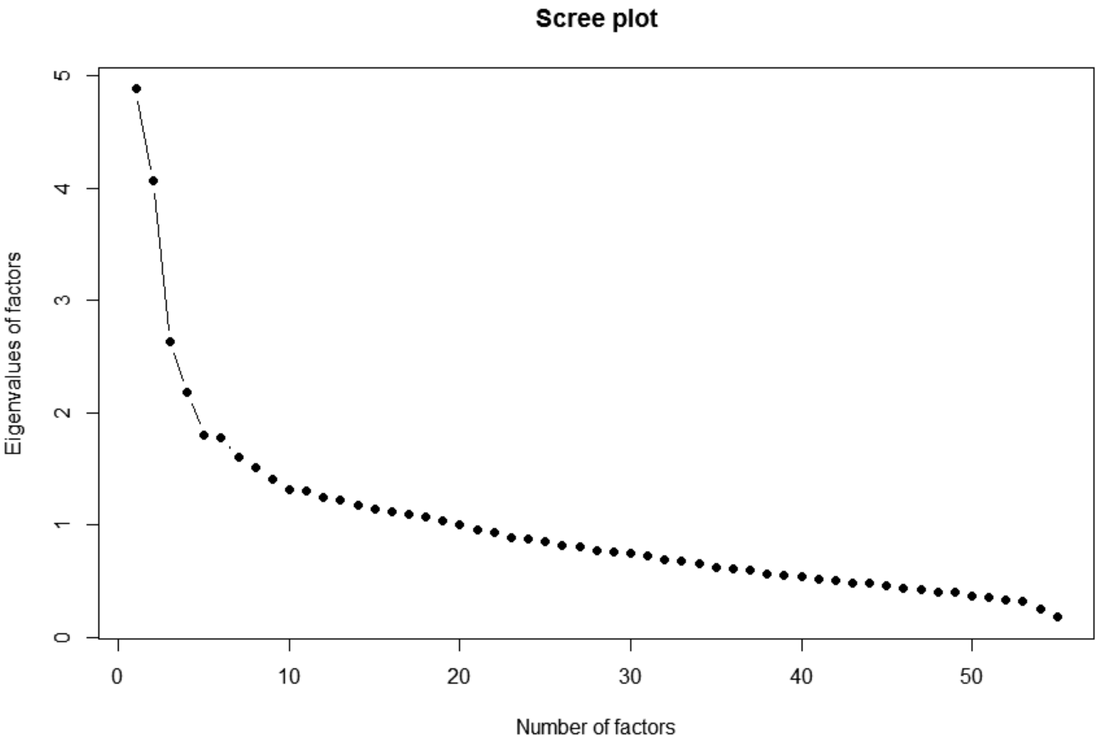
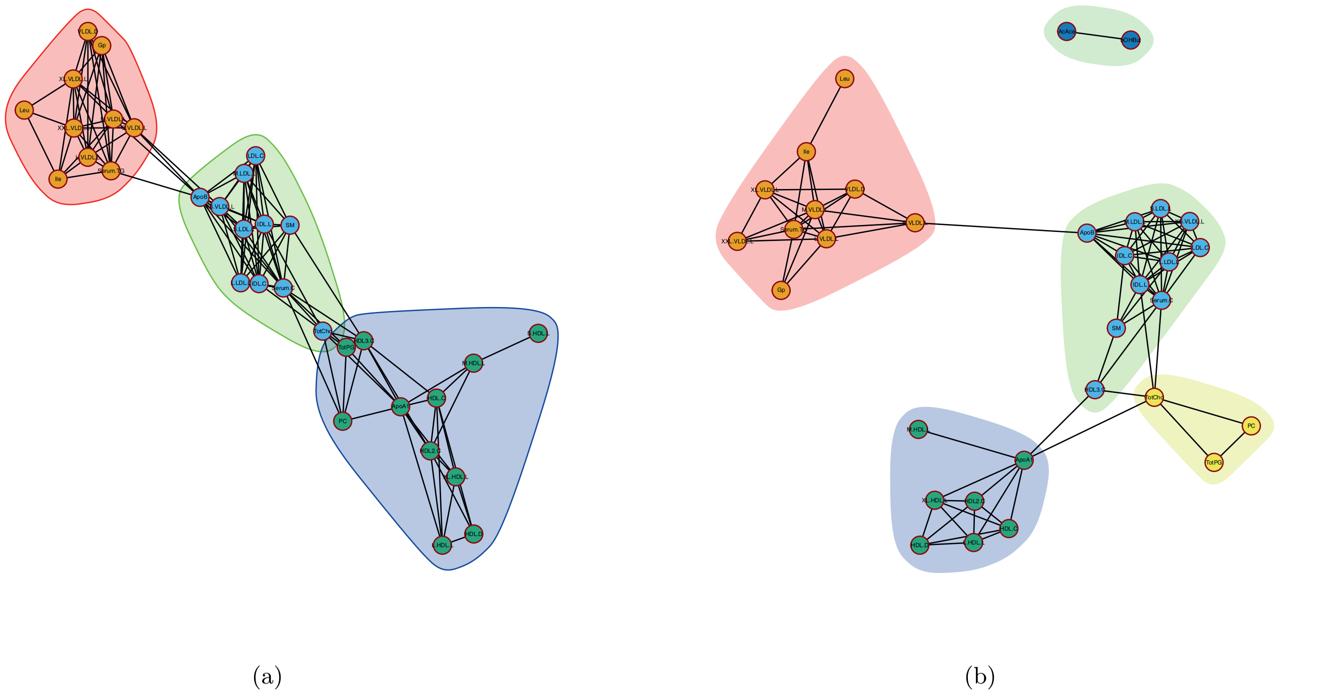
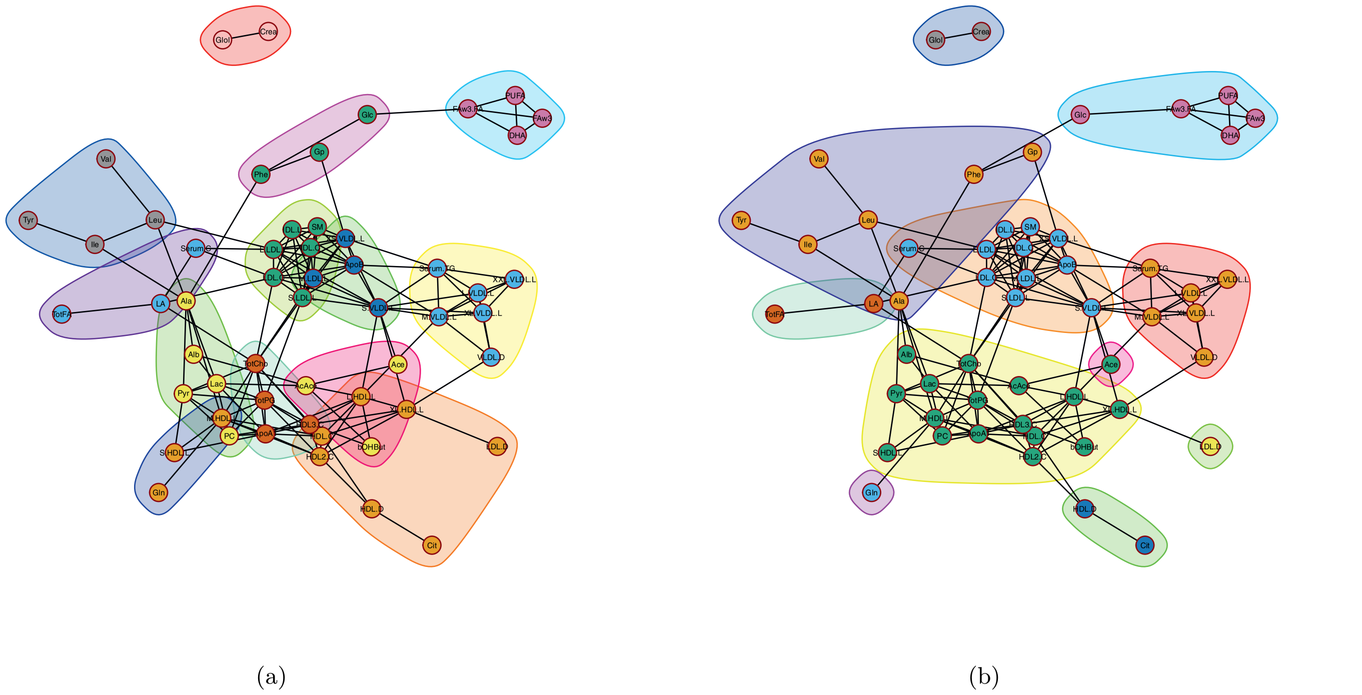


Fig. 3.1: Scree plot for selecting number of dietary profiles



(a) (b)

Fig. 3.2: Estimated metabolite networks and cluster identification using the Girvan-Newman algorithm when Age and Sex have been accounted for. The metabolite networks concern different time points (a) 2007 (b) 2014.



(a)

(b)

Fig. 3.3: Estimated metabolite networks with respect to dietary and lifestyle variation when cluster identification is performed using the (a) two-step dynamic hybrid algorithm (b) Girvan-Newman algorithm.

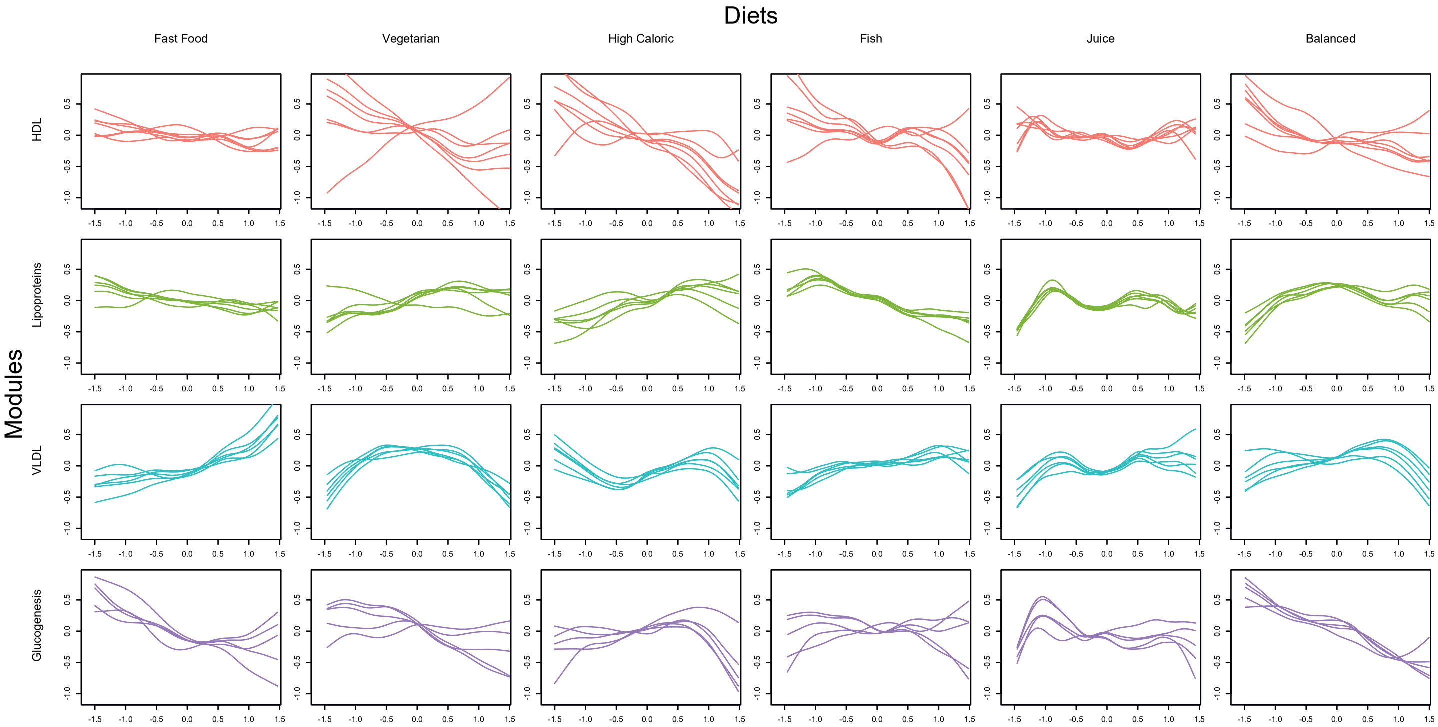


Fig. 3.4: Metabolite profile plots when the metabolite clusters have been identified using the dynamic cut tree algorithm. The relationship of the concentration levels of every metabolite (y-axis) to each diet (factor scores on the x-axis) is depicted using a spline. All metabolites belonging in the same module have the similar relationships to the diets.

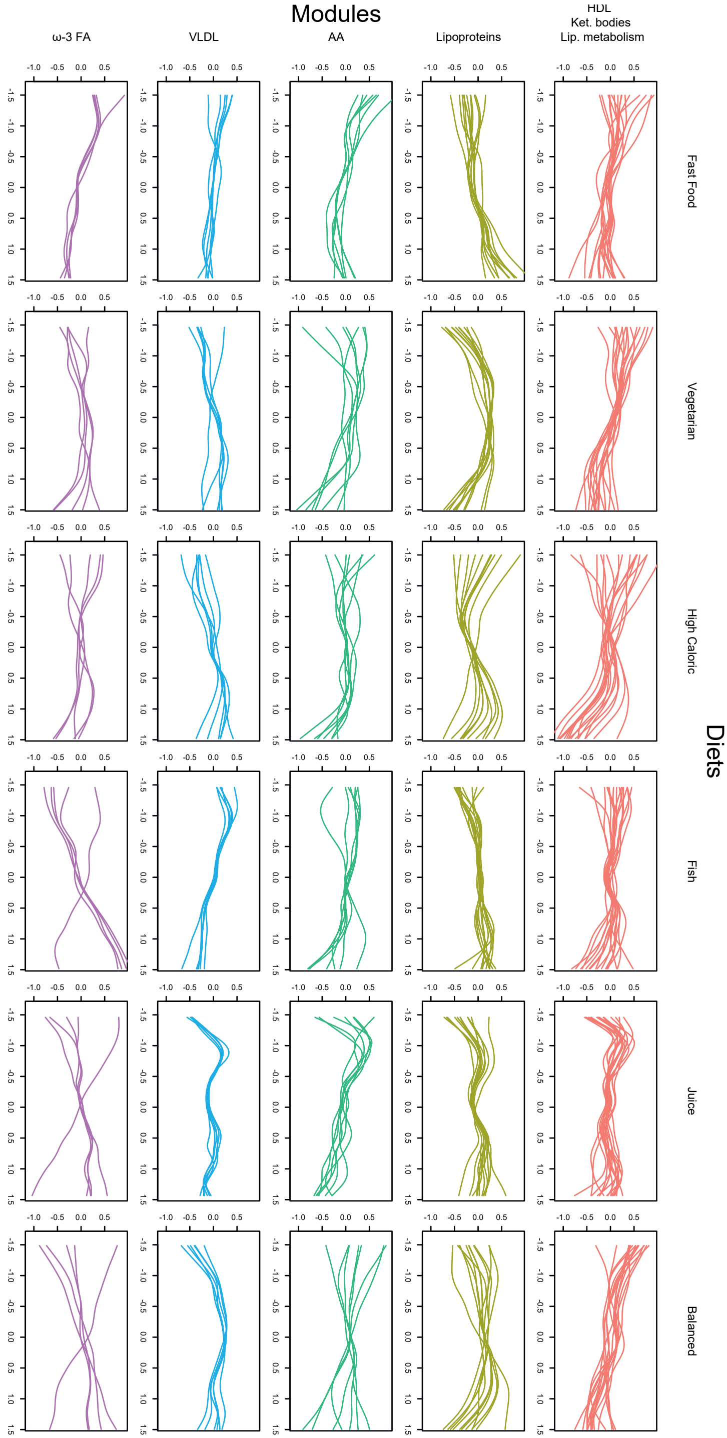


Fig. 3.5: Metabolite profile plots when the metabolite clusters have been identified using the Girvan-Newman algorithm. The relationship of the concentration levels of every metabolite (y-axis) to each diet (factor scores on the x-axis) is depicted using a spline. All metabolites belonging in the same module have the similar relationships to the diets.

Table 3.3: Characterization of the modules when the Girvan-Newman algorithm is used. Only modules with five or more metabolites have been reported.

Module	Density	Centralization	Heterogeneity	Nr of Nodes
HDL [a]	0.46	0.18	0.50	7
Lipoproteins[b]	0.96	0.03	0.04	6
VLDL2[c]	0.86	0.12	0.19	6
gluconeogenesis [d]	0.63	0.28	0.37	5
Complete Network	0.10	0.11	0.64	54

[a] Cit, HDL.C, HDL.D, HDL2.C, L.HDL.L, LDL.D, XL.HDL.L
[b] IDL.C, IDL.L, L.LDL.L, LDL.C, S.LDL.L, SM
[c] L.VLDL.L, M.VLDL.L, Serum.TG, VLDL.D, XL.VLDL.L, XXL.VLDL.L
[d] Ala, Alb, Lac, PC, Pyr

Table 3.4: Characterization of networks and modules subject to dietary patterns and lifestyle, when the Girvan-Newman algorithm is used.Only modules with five or more metabolites have been reported.

Module	Density	Centralization	Heterogeneity	Nr of Nodes
HDL/Ket. bodies/Lip. metabolism [a]	0.37	0.23	0.43	16
Lipoproteins[b]	0.76	0.20	0.29	11
AA [c]	0.30	0.16	0.48	7
VLDL [d]	0.86	0.12	0.19	6
ω-3 FA [e]	0.67	0.23	0.39	5
Complete Network	0.10	0.11	0.64	54

[a] AcAce, Alb, ApoA1, bOHBut, HDL.C, HDL2.C, HDL3.C, L.HDL.L, Lac, M.HDL.L, PC, Pyr, S.HDL.L, TotCho, TotPG, XL.HDL.L
[b] ApoB, IDL.C, IDL.L, L.LDL.L, LDL.C, M.LDL.L, S.LDL.L, S.VLDL.L, Serum.C, SM, XS.VLDL.L
[c] Ala, Gp, Ile, Leu, Phe, Tyr, Val
[d] L.VLDL.L, M.VLDL.L, Serum.TG, VLDL.D, XL.VLDL.L, XXL.VLDL.L
[e] DHA, FAw3, FAw3.FA, Glc, PUFA

4

A Guided Network Estimation Approach Using Multi-Omic Information

Bartzis, G., Peeters, C.F.W., Ligterink, W., & Van Eeuwijk, F. (2024) *BMC Bioinformatics*, *25(1)*, 202

4.1 Introduction

Advances in high-throughput technology have enabled the massive collective quantification of molecular entities, such as messenger RNA, proteins, and metabolites. This age of omics has revolutionalized the field of systems biology, enabling biological systems to be studied using mathematical and computational modeling on high-dimensional omics data. In systems biology, an organism is viewed as a complex web of interacting molecular entities (Nielsen & Jewett, 2007) studied in order to outline how cells, organs, and tissues behave at a molecular level (Raja et al., 2017).

A commonly used tool for analyzing omics data is network analysis. In network analysis, each omics level is assumed to have a network representation where complex associations are visualized by graphical structures. SNPs, genes, metabolites, and/or traits are typically represented by nodes in a graph and their associations (physical, genetic, or functional) by edges connecting them. Extracted patterns are then used to help elucidate biological mechanisms underlying traits of interest.

4.1.1 Methods for omic data integration

A key question in systems biology is how to model omics data at a systems level (integrative analysis), instead of each omics source separately (Kitano, 2002). Several approaches have been developed in the context of integrated analysis, see (Joyce & Palsson, 2006; Fabres, Collins, Cavagnaro, & Rodríguez López, 2017). One such approach for two sets of omics data is canonical correlation analysis (CCA) (Jendoubi & Strimmer, 2019). In order to solve CCA, the inverse of two covariance matrices needs to be computed which is problematic when the number of variables exceeds the number of samples, therefore penalization techniques can be implemented. Similarly, penalized partial least squares (PLS) regression (Lê Cao & Le Gall, 2011) variants (sPLS; sparse Partial Least Squares) have been proposed in order to remove noisy variables resulting in variable selection for both sets of omics data (Lê Cao, González, & Déjean, 2009).

An extension of sPLS is the sparse multi block partial least squares regression (sMBPLS) (W. Li, Zhang, Liu, & Zhou, 2012) in which several genomic data are measured on the same samples. One dataset is considered the response data, while the rest acts as guiding sets. In an application using a dataset containing gene expression (response data), copy number variation, DNA methylation, and micro RNA expression, Li and et al. (W. Li et al., 2012) identified combinations of multiple types of genomic markers that jointly impacted the expression of a set of genes. The covariance between the data blocks and the response block is maximized so that multidimensional modules are discovered associating the guiding with the response data.

Finally, network-based integration methods have also been proposed. The integration may be vertical (across omic-levels) or horizontal (one omic platform through time). The vertical approaches aim to provide a mechanistic understanding of molecular (de)regulation across the omic cascade. An overview of such methods can be found in (Agamah et al., 2022). In this work, we propose an integrative network reconstruction method where the network topology of one type of omics data is conditioned on the network topology of another omics source that is upstream in the omics cascade (Dettmer, Aronov, & Hammock, 2007).

4.1.2 Aim

The question answered in this work is how to integrate information across multiple omics levels. To answer and better understand relationships between different biological functional levels, we need to combine a systems view (requiring network modeling) and a multimodal view (requiring data integration).

In this work, we study whether network reconstruction of a particular omics source can benefit from information from the network organization of another omics source. For example, is metabolite network reconstruction helped by using DNA information? Or does information on a gene expression network aid recovery of the metabolites' network organization? Under our setting, for N samples, there are two sets of omics data; the P-dimensional target dataset (denoted by $\boldsymbol{Y}_{N\times P}$ from hereon) for which the underlying network organization needs to be recovered by using a Q-dimensional guiding dataset (denoted by $\boldsymbol{X}_{N\times Q}$ from hereon) and information on its network structure which is represented by a $Q \times Q$ matrix.

For estimating the network organization of the target data using the *guiding* data set and its network organization, a guided network estimation approach is considered. First, the network organization of the guiding data is estimated. We then regress the target on the guiding data and keep the fitted values on which we estimate a network structure. Alternatively, a guiding network can be used that is available already.

4.1.3 Overview

The rest of the paper is organized as follows. In Section 2, we review some basic network concepts, and propose a guided approach for estimating the network organization of an omics source using information from another omic dataset. In Section 3, we demonstrate our approach on metabolite data coming from the *Arabidopsis thaliana* population.

In the first application, the metabolic network estimation is guided by utilizing SNP information. SNP data and their spatial organization are used as input (DNA structure can be seen as a linear network, with edge intensity analogous to the distance between the markers on the chromosome) (Bartzis et al., 2022). We then identify and retain the part of metabolic variation related to SNP information and its structure and use it for estimating networks of metabolites. In the second data example, we guide the estimation of the metabolite networks by using information coming from gene-expression data and their network organization. Pairs of metabolites will share edges if they are associated to similar gene sets. Here, the data come from the Wageningen Seed Lab and contain SNP, transcriptomic, and metabolic information (Joosen et al., 2013). We consider this to be a standard dataset and demonstrate our integrative network approach. Our aim is to understand the metabolites from a SNP and gene level. Using network analysis we detect groups of metabolites having similar genetic or transcriptomic basis. We conclude the article with some discussion in Section 4.

4.2 Methods

Network analysis is a multivariate type of analysis aimed at recovering the underlying network structure of the data. We consider networks a representation of the pairwise (conditional) (in)dependencies between random variables. The nodes then represent metabolites or other molecular features and the edges represent pairwise dependency. An undirected network is typically encoded into a symmetric matrix $\boldsymbol{W}$ (intensity matrix). The element w_{ij} can be any type of association measure, e.g., the (absolute) partial correlation or (absolute) marginal correlation coefficient. The row- or column-sum of $\boldsymbol{W}$ is called strength and measures the total intensity of the connections of node i: $s(\boldsymbol{X})_i = \sum_j \boldsymbol{W}(\boldsymbol{X})_{ij}$.

4.2.1 Graphical LASSO

A popular approach for obtaining the underlying structure of the data from a set of P correlated variables (measured in N samples) is the Graphical LASSO (GL). In GL, the observational vectors of the data $\boldsymbol{Z}_{N \times P}$, where $\boldsymbol{Z}$ denotes a general dataset (either guiding or target data), are assumed to follow a P-dimensional multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. GL is based on the conditional independence of pairwise relationships, meaning that the precision matrix ($\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$) is estimated. When the element θ_{ij} is equal to zero, variables i and j are conditionally independent given all other variables. The penalized log likelihood using a LASSO penalty (Hastie, Tibshirani, Friedman, & Friedman, 2009; Friedman et al., 2008) is:

$$\ell_\lambda(\boldsymbol{\Theta}) \propto \log|\boldsymbol{\Theta}| - \mathrm{tr}(\boldsymbol{S\Theta}) - \lambda||\boldsymbol{\Theta}||_1, \tag{4.1}$$

where $||\bullet||_1$ is the $L1$-norm and λ is a non-negative tuning parameter governing the sparsity of the estimated precision matrix $\hat{\boldsymbol{\Theta}}$. The tuning parameter λ can be chosen based on subsampling. Here, we use the Stability Approach to Regularization Selection (StARS) (H. Liu et al., 2010) to estimate a set of stable edges. When using StARS, sparse networks are estimated based on multiple overlapping subsamples of the data, for different λ values on a grid. For an optimal λ (resulting in a sparse and stable network under random subsampling) selected by StARS, the absolute estimated precision matrix (similar to (Weber, Striaukas, Schumacher, & Binder, 2023)) $|\hat{\boldsymbol{\Theta}}|$ will be used here as the intensity matrix $\hat{\boldsymbol{W}}(\boldsymbol{Z})$.

4.2.2 Visual representation

To visually represent the sparse precision matrix as a network, the P variables are represented as a set of P nodes/vertices, which are connected by a set of edges, dictated by the non-zero entries of $\boldsymbol{W}(\boldsymbol{Z})$. The intensity of the connections between variables can be visualized by edge thickness with wider edges representing stronger connections. By taking the optimal λ selected by StARS as fixed, the intensity matrix $\hat{\boldsymbol{W}}(\boldsymbol{Z})_t = |\hat{\boldsymbol{\Theta}}_t|$ can be computed based on different subsamples $t = \{1, \ldots, T\}$. The edge-wise standard deviation computed over all $\hat{\boldsymbol{W}}(\boldsymbol{Z})_t$ can be an indicator of the edge's uncertainty. Since a network is a visual representation of an intensity matrix, we will be using both terms interchangeably and denote a network by its estimated intensity matrix $\hat{\boldsymbol{W}}(\boldsymbol{Z})$.

4.2.3 From Guiding to Target Data

Let, $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_P\}$ be the $N \times P$ target omics data matrix. Further, assume that $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_Q\}$ is the $N \times Q$ guiding omics data matrix. If $\boldsymbol{Y}$ contains the concentration levels of P metabolites on N samples, $\boldsymbol{X}$ could contain, for those same samples, data from Q SNPs or the expression levels of Q genes.

To incorporate information from the guiding omics data into the analysis of the target data we work in a regression framework. Prior to any type of multivariate analysis, e.g. network analysis, each of the P variables of $\boldsymbol{Y}$ is regressed on all Q variables of $\boldsymbol{X}$. Subsequently, the fitted values $\hat{\boldsymbol{Y}}(\boldsymbol{X})$ are obtained, e.g. using penalized regression (Tibshirani, 1996). Note that the OLS coefficient estimates cannot be obtained in the high dimensional case $Q > N$. LASSO regression has some attractive properties by performing variable selection, i.e. leading to zero coefficients for some of the variables. On the other hand, no information on the dependencies between $\boldsymbol{X}$ variables ($\hat{\boldsymbol{W}}(\boldsymbol{X})$) is used.

This drawback can be alleviated by using network-constrained regularization (NCR), as proposed by (C. Li & Li, 2008), where the underlying network organization of $\boldsymbol{X}$ is explicitly modeled when regressing each of the P variables of $\boldsymbol{Y}$ on $\boldsymbol{X}$ (Bartzis et al., 2022).

4.2.4 Network Constrained Regularization

We first assume that the data $\boldsymbol{X}$ have an underlying estimable network organization $\hat{\boldsymbol{W}}(\boldsymbol{X})$. For the p^{th} response ($\boldsymbol{y}_p$), the estimated regression coefficients $\hat{\boldsymbol{\beta}}_p \in \boldsymbol{R}^{Q\times 1}$ are obtained as:

$$\hat{\boldsymbol{\beta}}_p = \underset{\boldsymbol{\beta}_p}{\arg\min} \Bigg\{ (\boldsymbol{y}_p - \boldsymbol{X}\boldsymbol{\beta}_p)^\top (\boldsymbol{y}_p - \boldsymbol{X}\boldsymbol{\beta}_p) + \lambda_1 ||\boldsymbol{\beta}_p||_1 \\ + \lambda_2 \sum_{i\sim j} \left(\frac{\beta_{p_i}}{\sqrt{s(\boldsymbol{X})_i}} - \frac{\beta_{p_j}}{\sqrt{s(\boldsymbol{X})_j}} \right)^2 \hat{\boldsymbol{W}}(\boldsymbol{X})_{ij} \Bigg\}, \tag{4.2}$$

where $\sum_{i\sim j}$ denotes the sum over all adjacent unordered ij pairs, $s(\boldsymbol{X})_i$, $s(\boldsymbol{X})_j$ are the strengths of nodes i and j, and the term $\lambda_1||\bullet||_1$ is a LASSO-type penalty inducing a sparse solution in which not all Q predictors enter the model. For selecting the penalty parameters, cross-validation (CV) can be used for estimating the prediction error for set values of λ_1 and λ_2. The chosen penalties are the ones giving the lowest CV error (in our applications we used 5-fold CV). Note that (4.2) can be seen as a generalization of the elastic net (C. Li & Li, 2008).

The part accounting for the network structure of $\boldsymbol{X}$ when estimating $\hat{\boldsymbol{\beta}}_p$ in (4.2) is:

$$\lambda_2 \sum_{i\sim j} \left(\frac{\beta_{p_i}}{\sqrt{s(X)_i}} - \frac{\beta_{p_j}}{\sqrt{s(\boldsymbol{X})_j}} \right)^2 \hat{\boldsymbol{W}}(\boldsymbol{X})_{ij}. \tag{4.3}$$

The regression coefficients $\boldsymbol{\beta}_p$ are smoothed by penalizing the sum of weighted squares of the differences between β_{p_i} and β_{p_j}. Therefore, when nodes i and j share an edge with some weight ($w(\boldsymbol{X})_{ij} \neq 0$) in the network of $\boldsymbol{X}$, they will tend to have similar association to $\boldsymbol{y}_p$. This can be biologically justified since it is expected that connected nodes (in the case of SNPs/genes/metabolites) have similar function (C. Li & Li, 2008) and subsequently their coefficients should be shrunken towards each other. In expression (4.3) it can be seen that the regression coefficients are scaled, since it is expected that nodes with higher strength are more important and therefore have larger coefficients.

The linear model using the NCR criterion, unlike LASSO, preserves the grouping property, meaning that groups of connected variables (predictors linked in $\hat{\boldsymbol{W}}(\boldsymbol{X})$) will enter the model together. This result is shown in (C. Li & Li, 2008).

We then fit values of the target responses on the guiding predictors $\boldsymbol{X}$:

$$\hat{\boldsymbol{y}}_p(\boldsymbol{X}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}_p, \tag{4.4}$$

for each p. These are used for network reconstruction on $\hat{\boldsymbol{Y}}$.

4.2.5 Three-step approach for network reconstruction

For recovering the network structure of the target omics data, i.e. $\boldsymbol{Y}$ using a guiding omics source $\boldsymbol{X}$, we thus have a general 3-step approach:

1. Represent the guiding structure with an estimated or a priori known intensity matrix, i.e. $\hat{\boldsymbol{W}}(\boldsymbol{X})$ using GL;
2. Evaluate expression (4.2) with $\boldsymbol{Y}$, $\boldsymbol{X}$, and $\hat{\boldsymbol{W}}(\boldsymbol{X})_{ij}$ and retain the fitted data matrix $\hat{\boldsymbol{Y}}(\boldsymbol{X})$;
3. Use $\hat{\boldsymbol{Y}}(\boldsymbol{X})$ to reconstruct the target intensity matrix $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}(\boldsymbol{X}))$ using GL.

By using the proposed multi-step approach, the two omics sources are no longer treated independently. The resulting estimated network of the target data is conditioned on the network organization of the guiding data.

4.3 Application to data

We now use the proposed methods for estimating metabolite networks while using information from other omics sources that have a network organization of their own.

We use a Recombinant Inbred Line (RIL) population of a cross between two Arabidopsis accessions, i.e. Bayreuth (Bay-0) and Shahdara (Sha). In this population we want to study the metabolite similarities subject to variation coming from lower leveled omics sources (SNPs or Genes). In our first example we utilize SNP data and their spatial relationship to estimate metabolite networks. Metabolites will be connected if they have the same genetic basis (similar QTLs). In the second example, we use gene expression data and their underlying network organization information when we estimate metabolite networks. Therefore, we identify metabolites with similar transcriptomic basis.

4.3.1 Data

Seeds from 164 lines of the Arabidopsis Bay-0×Sha RIL population were divided into four sub-populations (41 lines each) representing four important developmental stages of seed germination; (i) freshly harvested primary dormant dry seeds (PD), (ii) after-ripened non-dormant dry seeds, (iii) seeds imbibed for 6 hours (6H), and (iv) seeds at radical protrusion (RP).

For determining the metabolite concentrations, all 164 lines were subjected to gas chromatography time of flight mass spectrometry giving 7537 peaks, representing 161 metabolites based on retention time and correlation structure (Tikunov, Laptenok, Hall, Bovy, & De Vos, 2012; Joosen et al., 2013). In total, $P = 64$ metabolites were annotated and were further used in our analysis. Gene expression analysis was performed using the Affymetrix AtSNPtile microarray on the same sub-populations and developmental stages as the metabolites, where the expression levels of 29304 genes were extracted. The top 10% most varying genes

($Q1 = 2931$ genes) were retained for further analysis. Concentration levels of the metabolites and gene expression levels were log transformed and adjusted for the four developmental seed stages by subtracting the mean levels from each group. Finally, information on $Q2 = 1059$ markers (5 chromosomes) was available. More information on the study design and data can be found in (Joosen et al., 2013) and (Joosen, 2013). For the rest of the paper, since metabolites will be the target dataset, we will denote their $N \times P$ data matrix as $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_P\}$. The $N \times Q1$ gene expression data and the $N \times Q2$ SNP data matrix will be used as guiding dataset and will be denoted as $\boldsymbol{X}^G = \{\boldsymbol{x}_1^G, \ldots, \boldsymbol{x}_{Q1}^G\}$ and $\boldsymbol{X}^S = \{\boldsymbol{x}_1^S, \ldots, \boldsymbol{x}_{Q2}^S\}$, respectively.

4.3.2 From SNPs to metabolites

Step 1: The SNP network representation

By having map information known, we represent the SNP data as the simplest type of network, i.e. a one-dimensional linear 'network'. We represent with $\alpha_{(1)}, \ldots, \alpha_{(p)}$ the ordered (in ascending order) genetic/physical position of the markers on the chromosome. The intensity of the connections between neighboring nodes is the relative (genetic/physical) marker proximity is calculated as:

$$\hat{\boldsymbol{W}}(\boldsymbol{X}^S)_{ij} = \hat{\boldsymbol{W}}(\boldsymbol{X}^S)_{ji} = 1 - \frac{\alpha_{(j)} - \alpha_{(i)}}{\alpha_{(p)} - \alpha_{(1)}}, \text{ where } i = 1, \ldots, p-1 \text{ and } j = i+1, \tag{4.5}$$

where $\hat{\boldsymbol{W}}(\boldsymbol{X}^S)_{ij} = 0$ for all other cases and for markers belonging to different chromosomes.

4

Step 2: Estimating the metabolite part related to genetic variation

In order to use $\boldsymbol{X}^S$, and $\boldsymbol{W}(\boldsymbol{X}^S)$ for estimating $\boldsymbol{Y}^M(\boldsymbol{X}^S)$, we work with the NCR as described in Section 4.2.3. Sets of SNPs that relate to each metabolite are identified. For metabolite p, the vector of coefficients $\boldsymbol{\beta}_p^S$ is estimated and used for obtaining the metabolite fitted values as:

$$\hat{\boldsymbol{y}}_p^M(\boldsymbol{X}^S) = \boldsymbol{X}^S \hat{\boldsymbol{\beta}}_p^S.$$

Step 3: Estimating metabolite network related to genetic variation

By using GL coupled with StARS on $\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S)$, the metabolite network using SNP information $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$ was estimated and is visualized in Figure 4.5. The optimal regularization parameter λ^S equalled 0.651, resulting in 98 edges between the metabolites. In the same figure, the network using the original metabolite values ($\boldsymbol{Y}^M$), i.e. $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$ is depicted. In order to compare the two networks, we controlled the sparsity of $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$: select the regularization parameter giving the same number of edges (98 out of 2016 possible edges resulting in sparsity of 0.049). Therefore, the tuning parameter governing the network sparsity in $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$ was selected to be 0.554.

Results and comparison

By examining Figure 4.5, we first see that the uncertainty of the edges is lower in $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$ compared to $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$. The top connected (hub-nodes) metabolites in $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$ are *Proline, Valine, Threonine, Xylose*, and *Serine* with 16, 13, 12, 12, and 10 edges respectively. On the other hand, when we see the network of metabolites with respect to SNP variation, the top connected metabolites are *Serine, (2-Hydroxyethyl)-methanamine, Isoleucine*, and *Proline* with 10, 9, 9, and 7 edges, respectively.

Here, we highlight the major differences between the networks by comparing them. Differences between the networks are visualized in Figure 4.6. Edges are colored with green if they only appear in $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$, red if they only appear in $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$, and grey if they appear in both.

Interestingly, the metabolite losing the most edges by conditioning on SNP information (12) is *Xylose*, showing that the similarity with other metabolites was due to the non-genetic variation. Other metabolites losing multiple edges when we use SNP information are: *Proline* (11), *Valine* (11), *Asparagine* (7), and *Glucose* (7).

On the other hand, the metabolites that gained multiple edges by conditioning on SNP information are *Glutarate* (with 7) and *2-Oxoglutarate, Benzoate, Digalactosylglycerol, D-Xylofuranose, Fumarate, Glucuronate, Phosphoric acid*, and *Tyrosine* (with 5 edges each), showing that their genetic similarity with other metabolites was stronger but it was concealed by their non-genetic variation part.

Finally, the top metabolites retaining many edges are: *Isoleucine* (8), *Serine* (6), *Threonine* (6), *(2-Hydroxyethyl)-methanamine* (5), and *Proline* (5) showing that their genetic similarity with other metabolites was stronger than the non-genetic.

Connection between QTLs and metabolite network

The vector of estimated SNP coefficients can also be used to detect QTLs. Regions where we find SNPs with non-zero coefficients should be highlighted as possible QTL regions. In Figures 4.1-4.4 we provide some results of the correspondence between *Composite Interval Mapping* (CIM; *qtl* R-package) (Z.-B. Zeng, 1994) and QTL detection using NCR while in the supplementary material we present all metabolites. By closely inspecting Figures 4.1-4.4, it is evident that by using NCR we find positions on the chromosome with high CIM test statistic and subsequently possible QTL regions.

The GABA/Maltose pair (Figure 4.1) clearly had similar QTLs and thus share an edge. The Serine/Aspartate (Figure 4.2) and the GABA/Glucose-6-phosphate

(Figures 4.1, 4.3) pairs had an overlap in their QTL profiles but share no edge, which might indicate that the non common potentially identified QTLs are responsible for a big part of the metabolic variation. Finally, the Fructose/Glucose-6-phosphate (Figure 4.3) and GABA-Glycolate (Figures 4.1, 4.4) pairs do not have overlap in QTLs justifying why there are no edges in $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$.

Summarizing, when two metabolites are connected, we usually observe a similarity in QTLs. On the other hand, when metabolites do not share an edge, this is generally due to dissimilar QTLs. Still, there can be situations where metabolites with similar QTLs are not connected, because of either measurement noise, or because non-overlapping QTLs account for a big part of the metabolic variation.

Multigraph representation:

An informative representation can be obtained by visualizing a network that combines all data used here. In Figure 4.7, $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$ and all markers have been visualized; the nodes have been colored so that visual inspection is easier. The 5 chromosomes have been depicted as circular with the start and end being at the topmost point (one moves clockwise from start to end). Edges between a metabolite p and SNPs denote the non-zero estimated coefficients $\hat{\boldsymbol{\beta}}_p^S$.

By looking at Figure 4.6 and Figure 4.7 we identify three interesting metabolite groups. The first consists of: *Nicotinate, GABA, Benzoate, Glutarate, Tyrosine, Digalactosyglycerol, Valine, Trehalose*, and *Maltose*. In Figure 4.6, the edges between the metabolites are green meaning that they are grouped together after using SNP information. Some of them are connected to the biosynthesis of alkaloids derived from shikimate pathway. In Figure 4.7 they have been represented as the dark green colored cluster sharing many edges with chromosome 5.

The second interesting metabolite group consists of: *Allantoin, Gluconate, Fructose, Glucuronate, Mannose, Glucopyranose*, and *Glucose-6-phosphate* and has been colored as ciel blue in Figure 4.7. These metabolites did not share any connections with any metabolites in $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$ but formed a cluster when SNP information was used ($\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$). Those metabolites are involved in sucrose metabolism, glycolysis and are either sugars or closely related to sugars.

Finally, the most interesting metabolite group contains: *Phenylalanine, Proline, Isoleucine, Aspartate, N-Acetylglutamate, Glutamate, (2-Hydroxyethyl)-methanamine, Glycine, Serine,* and *Threonine*. This metabolite group is the one with most grey edges in Figure 4.6, showing that it retained most of its edges when we include non-genetic SNP variation. All metabolites in this group are contained in the biosynthesis of amino-acids. They have been colored red in Figure 4.7 and show strong association with chromosomes 1, 4, and 5.

4.3.3 From genes to metabolites

In the first example we used SNP data, where the network structure was simple. Nevertheless, in many applications, e.g. gene expression data, the underlying network structure is far more complicated than a linear distance-based network and not known *a priori*. In this second example we recover metabolite networks by utilizing gene information.

Step 1: Reconstruction of the gene expression network

In order to reconstruct the gene expression network, we use GL coupled with StARS on $\boldsymbol{X}^G$. The selected regularization parameter based on subsampling was equal to 0.82, resulting in 3347 edges (sparsity of 0.00078). Our strict selection was based on the intention to minimize edges between metabolites due to false positives in gene expression data. The sparse gene expression intensity matrix $\hat{\boldsymbol{W}}(\boldsymbol{X}^G)$ contains the absolute values of the resulting inverse covariance matrix.

Step 2: Estimating the metabolite part related to transcriptional variation

We use expression (4.2) with $\boldsymbol{Y}^M$ as response and the gene expression data $\boldsymbol{X}^G$ as predictors, having an estimated network structure $\hat{\boldsymbol{W}}(\boldsymbol{X}^G)$. The p-th metabolite is regressed on all genes. The vector of estimated coefficients $\hat{\boldsymbol{\beta}}_p^G$ related to the $Q1$ genes is used for recovering the fitted metabolite values related to transcriptional variation ($\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^G)$) as:

$$\hat{\boldsymbol{y}}_p^M(\boldsymbol{X}^G) = \boldsymbol{X}^G \hat{\boldsymbol{\beta}}_p^G \tag{4.6}$$

Step 3: Metabolite networks related to gene variation

To estimate metabolite networks, we use GL on the fitted metabolite values related to transcriptional variation, i.e. $\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^G)$. For comparing with $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$, the regularization parameter λ^G was selected equal to 0.69, resulting in 98 edges for the metabolite network related to gene variation ($\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^G))$). The resulting network has been visualized in Figure 4.8 together with $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$. The edges' width in both figures, denotes the intensity of the connection between the metabolites. The opacity represents the uncertainty for the edge intensity and has been computed based on resampling as in example 1. In Figure 4.8, we see that the uncertainty of the edges is lower in $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^G))$ compared to $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$. By examining $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^G))$, we see that the top connected metabolites are *Arabinose, Xylose, Glucose, Raffinose, Fructose-6-phosphate*, and *Monomethylphosphate* with 13, 13, 12, 12, 11, and 11 edges, respectively.

Metabolites are mainly connected because they are associated to similar (or connected) genes. On the other hand, metabolites that are not connected are usually associated with different sets of genes.

Network of differences between $\boldsymbol{W}(\hat{\boldsymbol{Y}}^{\boldsymbol{M}}(\boldsymbol{X}^{\boldsymbol{G}}))$ and $\hat{\boldsymbol{W}}(\boldsymbol{Y}^{\boldsymbol{M}})$

To highlight the major differences between the networks we visualize their differences in Figure 4.9. Edges are colored with green if they only appear in $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^{M}(\boldsymbol{X}^{G}))$, red if they only appear in $\hat{\boldsymbol{W}}(\boldsymbol{Y}^{M})$, and grey if they appear in both. By examining the differences between the networks, we make the following observations.

The metabolites losing most of their edges are *Proline* (13) and *Valine* (12) showing that their similarity with other metabolites is not due to the transcriptional part of variation. Those metabolites lost many edges in the SNP example as well, showing that their correlation with other metabolites is driven by other sources of variation. Other metabolites losing multiple edges when we use only gene information are: *Aspartate* (7), *Threonine*(7), and *Glutamate* (6).

On the other hand, the metabolites gaining edges when gene variation is used are *Monomethylphosphate* (11), *Arabinose* (10), *Glutarate* (10), *Raffinose* (10), and *Galactinol* (8). Finally, the metabolites keeping most of their edges are: *Xylose* (9), *Serine* (6), *Fructose-6-phosphate* (5), *Glucose* (5), and *Threonine* (5).

Another finding standing out when looking at Figure 4.9 is the two metabolite clusters. One consists of the following amino-acids: Proline, Phenylalanine, threonine, Isoleucine, Valine, Glycine and Serine. Lastly, the cluster containing many green edges is composed of several metabolites that are related to abiotic stress responses in plants like those related to the Raffinose family of oligosaccharides (Sengupta, Mukherjee, Basak, & Majumder, 2015).

4.4 Discussion

In this work, we studied whether estimating the network structure of a particular omics level can be supported by using information coming from the network organization of another omics level. We proposed a three-step approach (Section 4.2.5) based on regularized regression that was demonstrated in two applications. Using this approach, in both applications the recovered networks contained edges with lower uncertainty compared to the original data.

For addressing missingness within guiding and target datasets, an expectation-maximization (EM) algorithm adapted for penalized network estimation offers a theoretical solution, but may escalate computational complexity. Alternatively, matrix completion techniques provide a pragmatic preprocessing step to impute missing data (e.g., (Mazumder, Hastie, & Tibshirani, 2010)), thereby preparing the dataset for our network-based approach.

A natural extension of our three-step method is by using more than two datasets, e.g. SNPs, genes, and metabolites. To estimate such networks, we work sequentially from one omics source to the next. We start from SNP data and their linear structure and work our way to estimate gene expression data subject to SNP variation. Then we use the fitted gene expression values and their estimated network organization to estimate metabolite networks. Even though the rationale of such application is intuitive (propagate information from one omics level to the next), the interpretation is challenging.

By taking a step forward, since metabolites determine many quality traits (nutritional value, drought tolerance, etc) (H. Wang et al., 2015) and are closely related to the phenotypes (Beisken et al., 2015), we could also study phenotypic associations using network analysis. By using our three-step approach for modelling phenotypic associations, we would be able to identify metabolites, genes and DNA regions responsible for these traits. Using this approach, in plant genetics, plant breeders and physiologists can improve adaptation to environmental stress, food quality, and crop yield (Okazaki & Saito, 2012).

Finally, an interesting point of discussion is the choice of NCR in step 2 over other candidate methods, e.g., the LASSO or elastic net. In (C. Li & Li, 2008), these three methods have been compared in different scenarios with respect to sensitivity (true positives), specificity (true negatives) and prediction mean squared error (PMSE). The NCR procedure resulted in better PMSE making it a principal candidate for our multi-step approach. Another alternative candidate method to relate the guiding and the target datasets would be to use L2 regularization instead of L1 in (4.2) making it a Ridge-NCR procedure. The solution of the Ridge-NCR problem with application in genomic prediction may be more interesting, as the L1 regularization tends to drop collinear variables from the model that can potentially carry relevant information. We have presented the results of a Ridge-NCR analysis elsewhere (Bartzis et al., 2022). Lastly, a hybrid between L1 and L2 penalties, aka elastic net-NCR can also be considered. Similar to Lasso, this alternative can produce reduced models by estimating zero-valued coefficients. In addition, not all collinear variables are eliminated, potentially retaining relevant information (similar to Ridge).

Supplementary Information

The supplementary material (`https://static-content.springer.com/esm/art%3A10.1186%2Fs12859-024-05778-7/MediaObjects/12859_2024_5778_MOESM1_ESM.pdf`) contains the correspondence between *Composite Interval Mapping* (CIM) and QTL detection using NCR for all metabolites used in this article. It also includes a list of SNPs associated to each metabolite used.

Figures

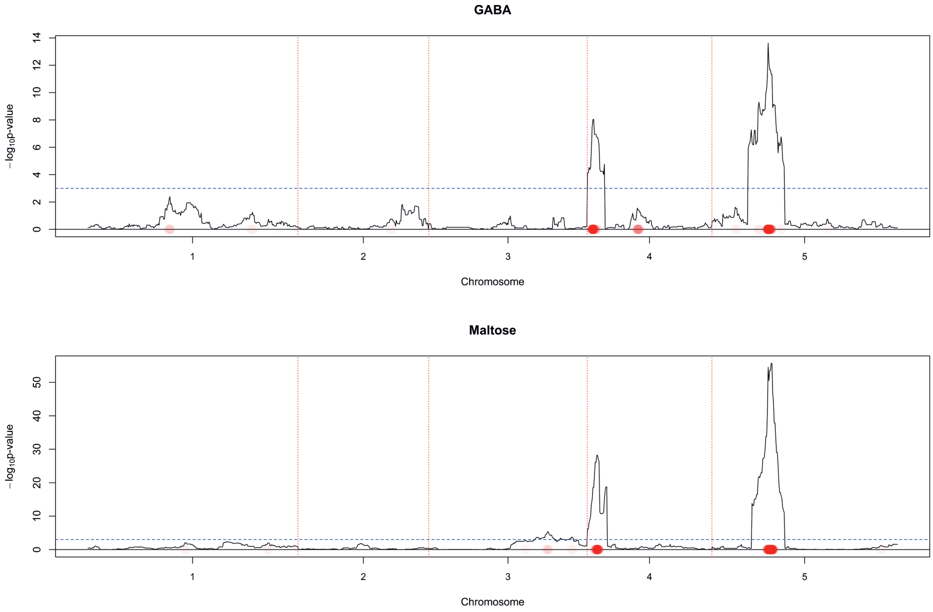


Fig. 4.1: QTL detection for GABA (a) and Maltose (b) using CIM and NCR when the guiding dataset is SNP data. GABA and Maltose share an edge in both $\boldsymbol{W}(\hat{\boldsymbol{Y}}^{M}(\boldsymbol{X}^{S}))$ and $\hat{\boldsymbol{W}}(\boldsymbol{Y}^{M})$ which can be justified by their QTL profile. The $-\log_{10}$ p-value score for every marker is plotted when CIM is used. The red dotted vertical lines are plotted as a visual separation between the 5 chromosomes and the chromosome number is indicated on the x-axis below each segment. The dotted horizontal blue line marks the $-\log_{10}$ p-value score of 3. Red dots on the x-axis are placed on marker positions for which NCR estimated non-zero coefficients. The color transparency indicates the magnitude of the regularized estimated coefficient. The correspondence between CIM and NCR can be seen by noticing that red-dots on the x-axis are in most areas where the $-\log_{10}$ p-value score has high values.

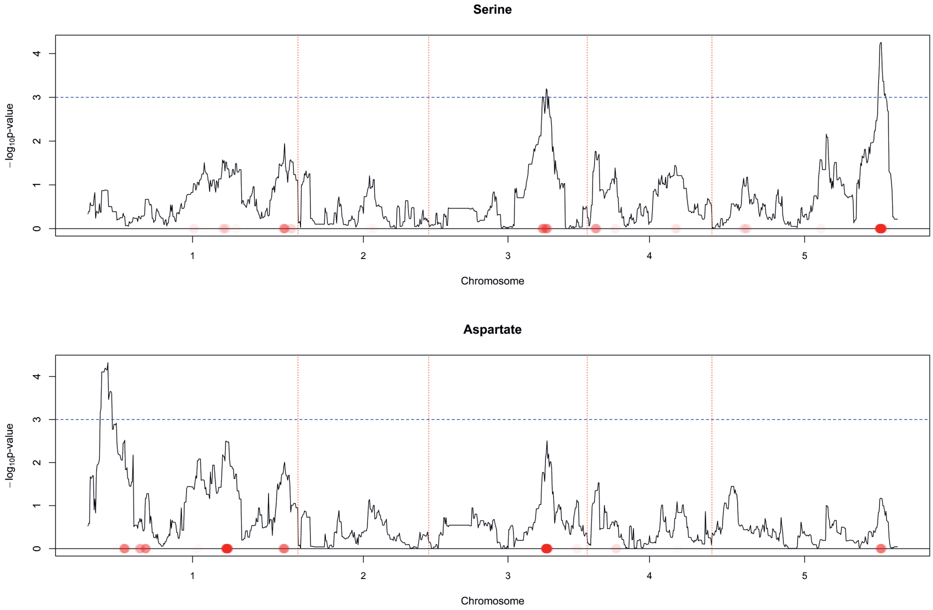


Fig. 4.2: QTL detection for Serine (a) and Aspartate (b) using CIM and NCR when the guiding dataset is SNP data. They only share an edge in $\boldsymbol{W}(\boldsymbol{Y}^M)$, but do not share an edge in $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$, which can indicate that the unique QTLs (for a pair of metabolites) can neutralize correlation induced by common QTLs. In this case unique QTLs are responsible for a bigger part of the metabolic variation. The $-\log_{10}$ p-value score for every marker is plotted when CIM is used. The red dotted vertical lines are plotted as a visual separation between the 5 chromosomes and the chromosome number is indicated on the x-axis below each segment. The dotted horizontal blue line marks the $-\log_{10}$ p-value score of 3. Red dots on the x-axis are placed on marker positions for which NCR estimated non-zero coefficients. The color transparency indicates the magnitude of the regularized estimated coefficient. The correspondence between CIM and NCR can be seen by noticing that red-dots on the x-axis are in most areas where the $-\log_{10}$ p-value score has high values.

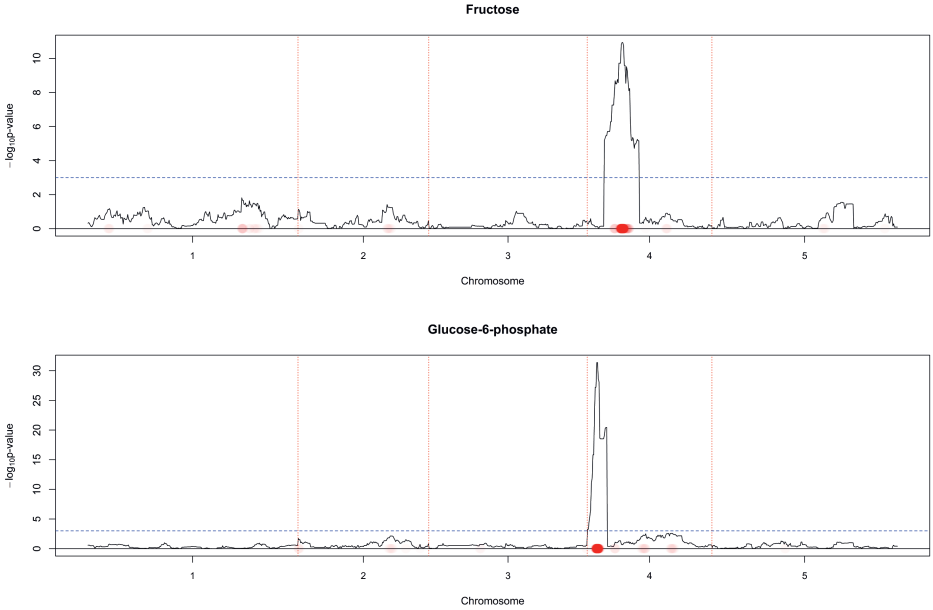


Fig. 4.3: Fructose (a) and Glucose-6-phosphate (b) do not have similar QTLs and therefore do not share an edge in $\boldsymbol{W}(\hat{\boldsymbol{Y}}^{M}(\boldsymbol{X}^{S}))$. The $-\log_{10}$ p-value score for every marker is plotted when CIM is used. The red dotted vertical lines are plotted as a visual separation between the 5 chromosomes and the chromosome number is indicated on the x-axis below each segment. The dotted horizontal blue line marks the $-\log_{10}$ p-value score of 3. Red dots on the x-axis are placed on marker positions for which NCR estimated non-zero coefficients. The color transparency indicates the magnitude of the regularized estimated coefficient. The correspondence between CIM and NCR can be seen by noticing that red-dots on the x-axis are in most areas where the $-\log_{10}$ p-value score has high values.

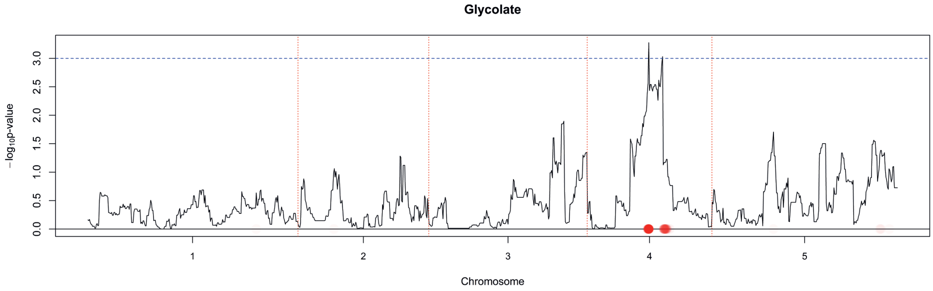


Fig. 4.4: Glycolate's QTL profile using CIM and NCR. The $-\log_{10}$ p-value score for every marker is plotted when CIM is used. The red dotted vertical lines are plotted as a visual separation between the 5 chromosomes and the chromosome number is indicated on the x-axis below each segment. The dotted horizontal blue line marks the $-\log_{10}$ p-value score of 3. Red dots on the x-axis are placed on marker positions for which NCR estimated non-zero coefficients. The color transparency indicates the magnitude of the regularized estimated coefficient. The correspondence between CIM and NCR can be seen by noticing that red-dots on the x-axis are in most areas where the $-\log_{10}$ p-value score has high values.

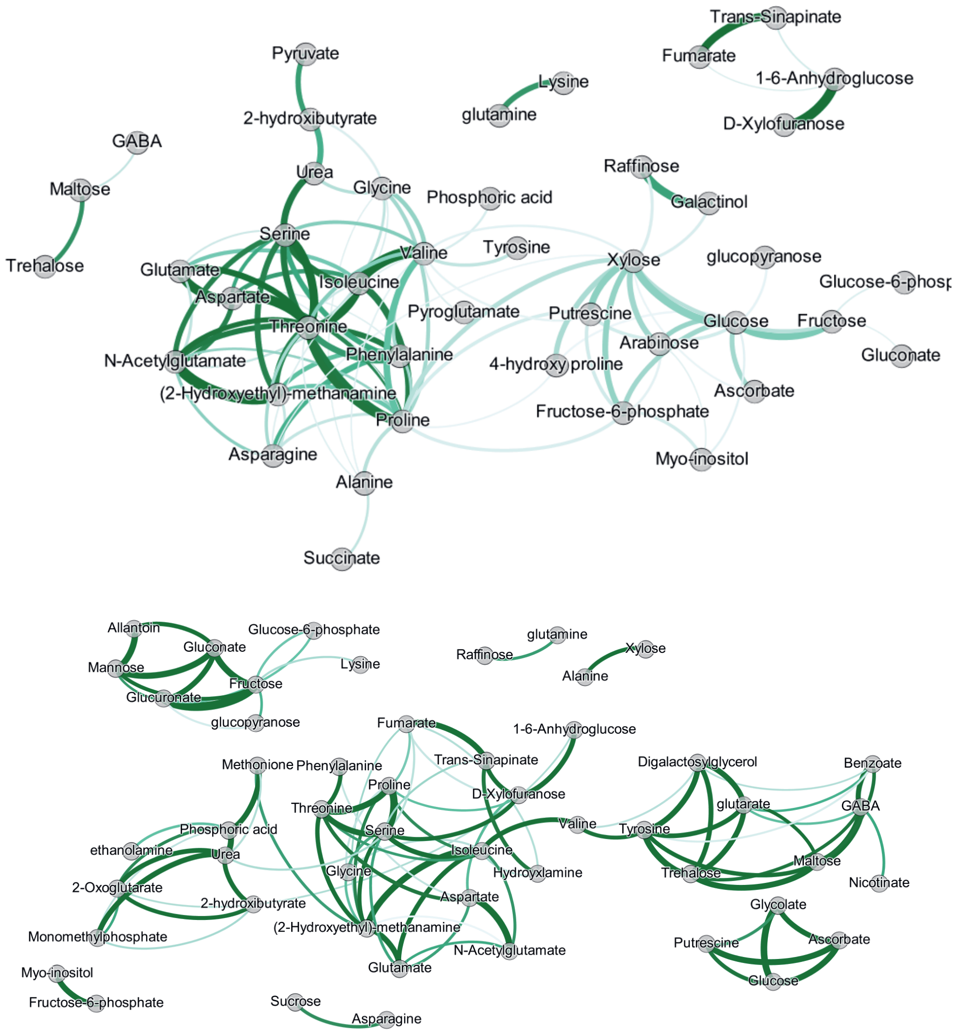


Fig. 4.5: Estimated metabolite networks when: (a) using the original metabolite data ($\boldsymbol{W}(\boldsymbol{Y}^M)$), and (b) using information on SNPs and their network structure ($\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$). Edges' width denotes the intensity of the connection between two nodes, while edges' opacity indicates the uncertainty as measured by the edges standard deviation.

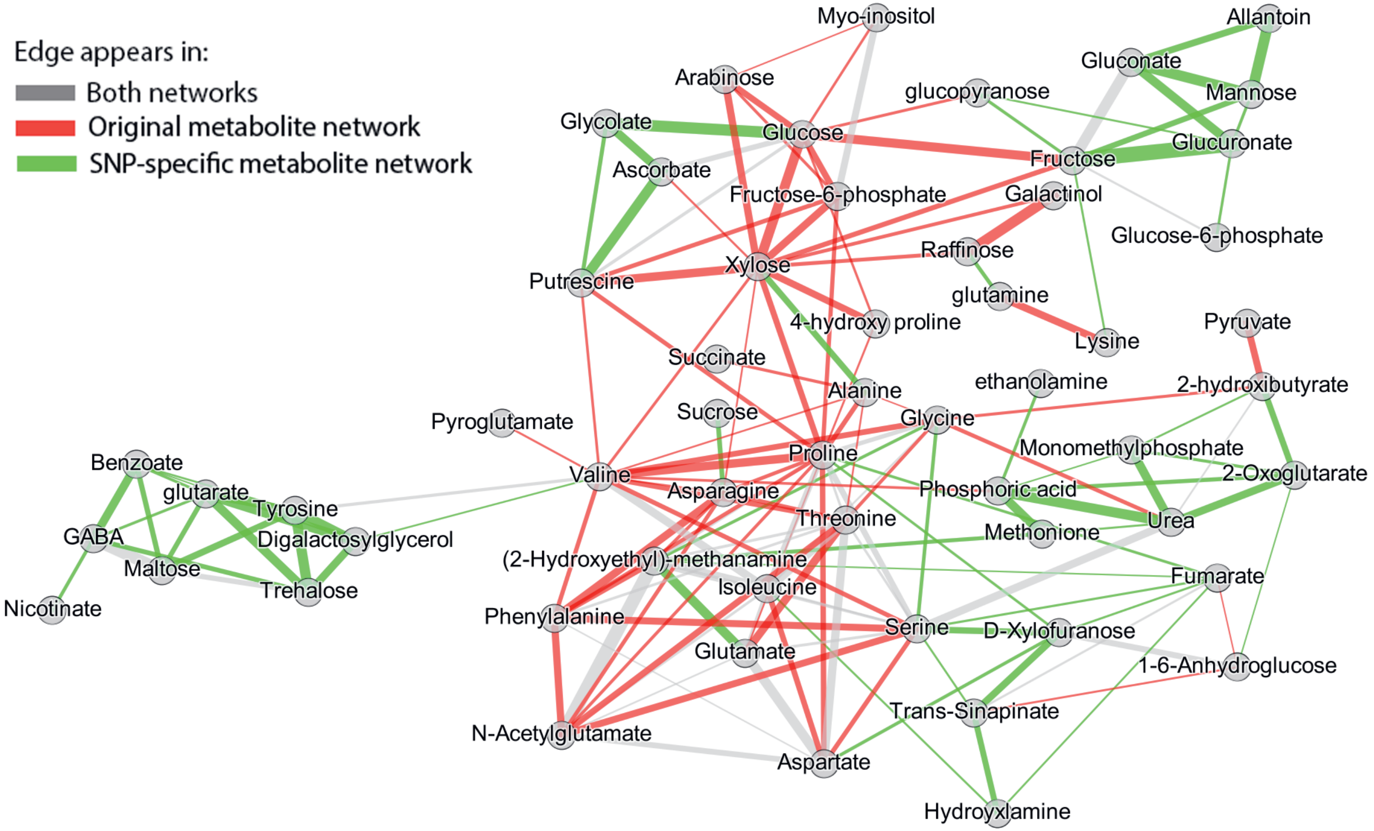


Fig. 4.6: Difference between network based on the original metabolite values ($\boldsymbol{W}(\boldsymbol{Y}^M)$) and network reconstructed when SNP information is used ($\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$). Green edges denote the unique edges that appear in $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$. Red denote the unique edges appearing in $\boldsymbol{W}(\boldsymbol{Y}^M)$. Grey edges are the common edges between $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$ and $\boldsymbol{W}(\boldsymbol{Y}^M)$. The width of the edges denotes the difference between the connections' intensity of $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$ and $\boldsymbol{W}(\boldsymbol{Y}^M)$.

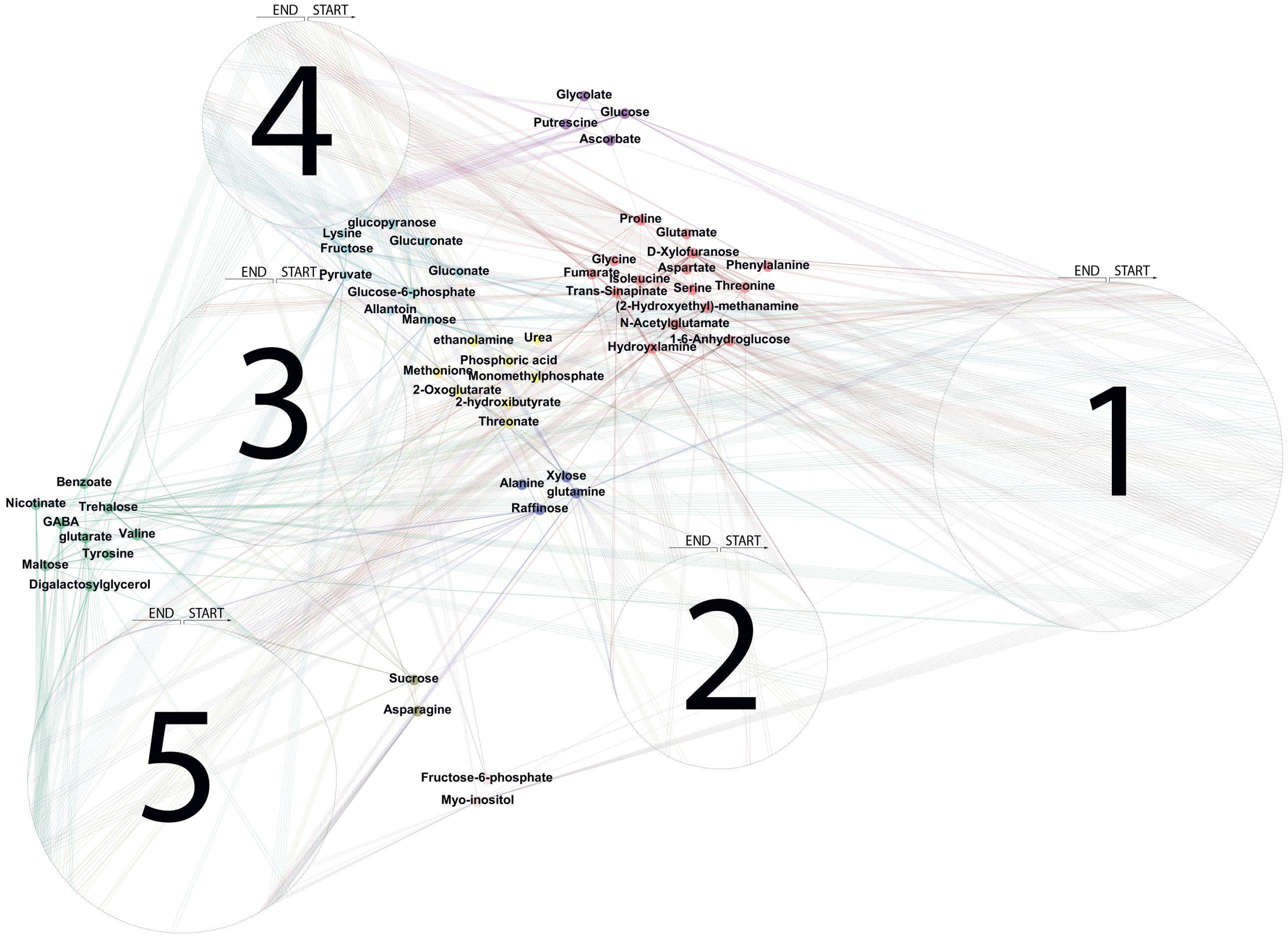


Fig. 4.7: Combined network of metabolites and SNPs. $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$ is visualized together with the five chromosomes which have been folded to be represented by five circular structures. The start and end of each chromosome is at the topmost part (moving clockwise for proceeding from start to end). Non-zero SNP coefficients for every individual model have been visualized as edges connecting metabolites and SNPs. Metabolites have been colored to ease visual inspection.

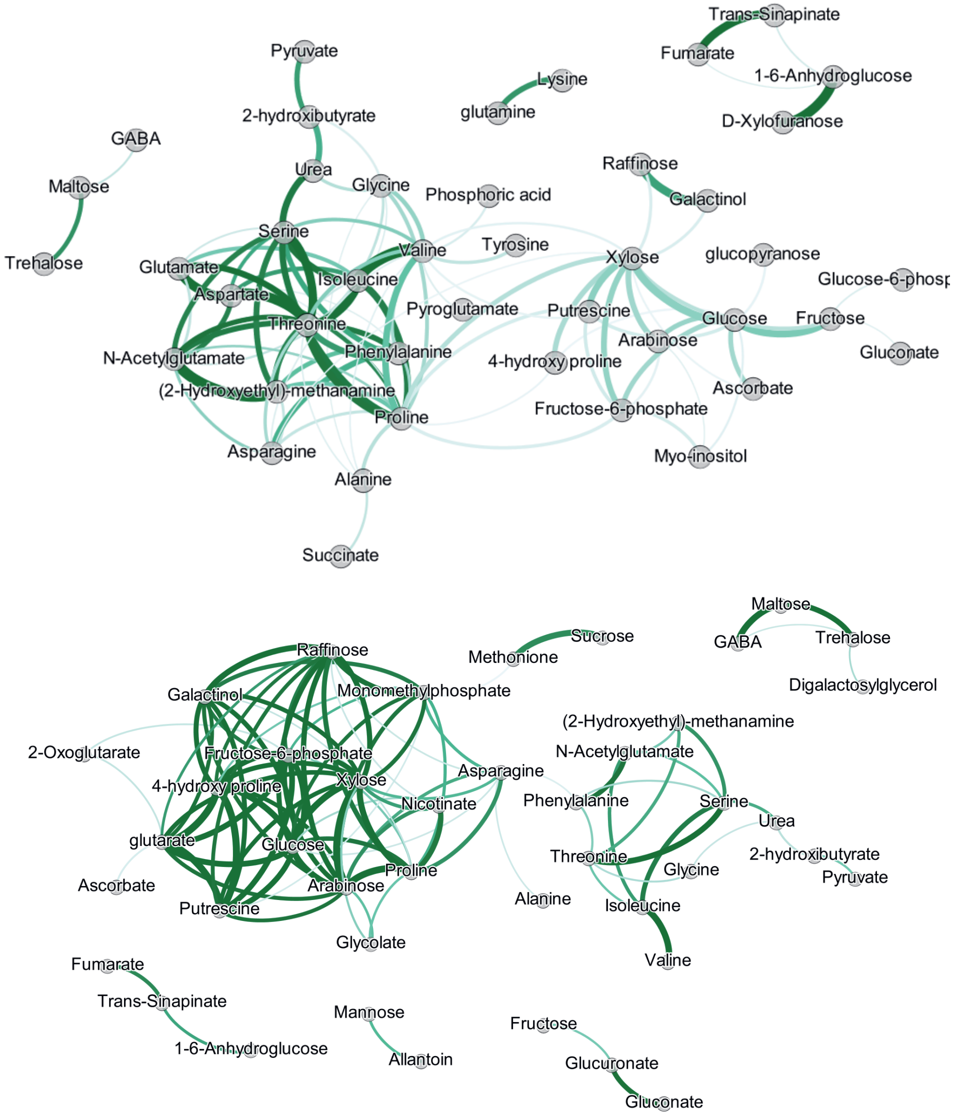


Fig. 4.8: Estimated metabolite networks when: (a) using the original metabolite data ($\boldsymbol{W}(\boldsymbol{Y}^M)$), and (b) using information on SNPs and their network structure ($\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(X^G))$). Edges' width denotes the intensity of the association between two nodes, while edges' opacity indicates the uncertainty as measured by the edges standard deviation.

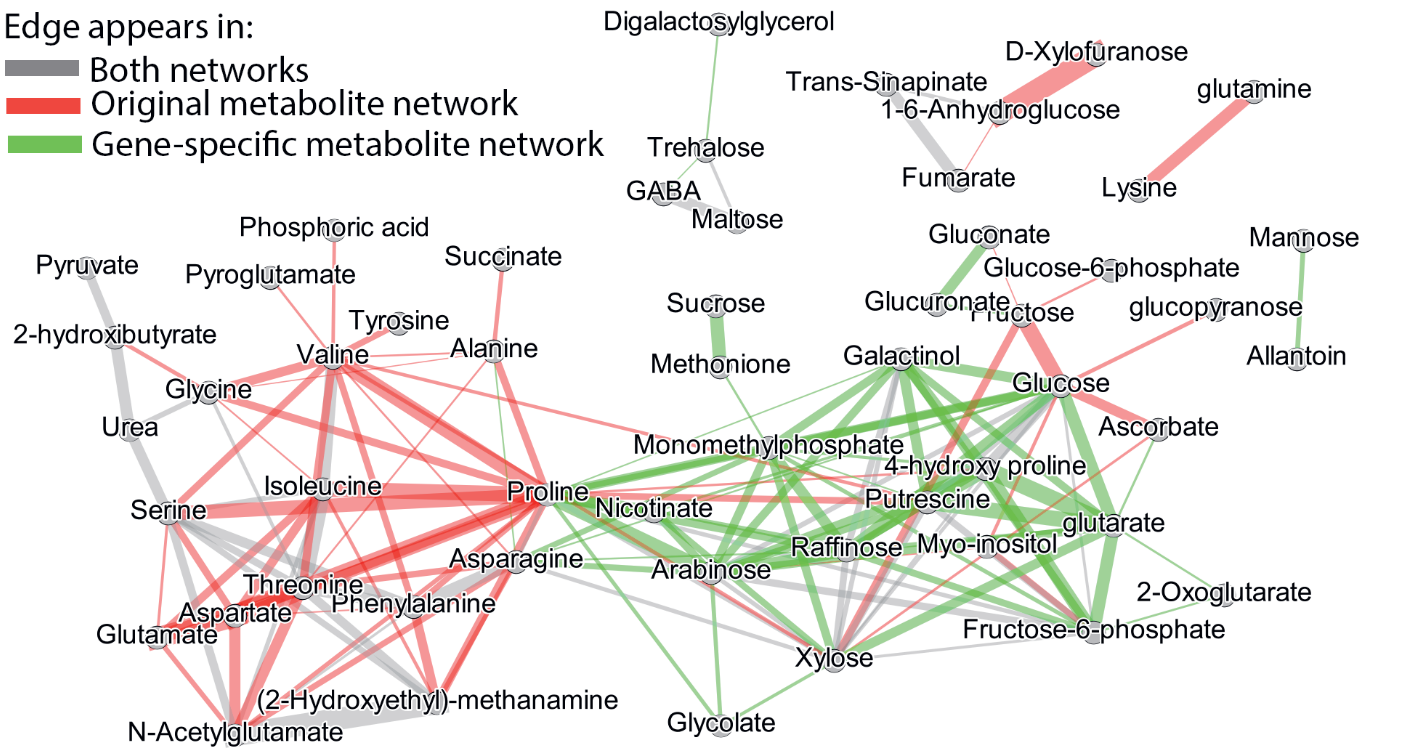


Fig. 4.9: Difference between network based on the original metabolite values ($\boldsymbol{W}(\boldsymbol{Y}^M)$) and network reconstructed when gene expression is used ($\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(X^G))$). Green edges denote the unique edges that appear in $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(X^G))$. Red denote the unique edges appearing in $\boldsymbol{W}(\boldsymbol{Y}^M)$. Grey edges are the common edges between $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(X^G))$ and $\boldsymbol{W}(\boldsymbol{Y}^M)$. The width of the edges denotes the difference between the connections' intensity of $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(X^G))$ and $\boldsymbol{W}(\boldsymbol{Y}^M)$.

5

psBLUP: Incorporating Marker Proximity for Improving Genomic Prediction Accuracy

Bartzis, G., Peeters, C.F.W., & Van Eeuwijk, F. (2022) *Euphytica*, *218(5)*, 54

5.1 Introduction

Genomic selection is a tool applied in animal and plant sciences for improving quantitative traits (Heffner, Sorrells, & Jannink, 2009; B. J. Hayes, Bowman, Chamberlain, & Goddard, 2009; Jannink, Lorenz, & Iwata, 2010; Goddard, Hayes, & Meuwissen, 2010; Van Binsbergen et al., 2015). Genomic values of line performance measuring the genetic merit of lines are calculated using markers (e.g., single nucleoteid polymorphisms; *SNPs*) covering the whole genome (B. J. Hayes et al., 2009). By using high density SNP panels, it is expected that SNPs in linkage disequilibrium (LD) with quantitative trait loci ($QTLs$) contributing to the phenotypic variation (B. J. Hayes et al., 2009; J. Zeng, Garrick, Dekkers, & Fernando, 2018) are included.

A training panel that has been both genotyped and phenotyped is used to build a prediction model describing a marker-trait relationship. A common approach to do so is by regressing phenotypes on all available markers using a linear model (de Los Campos, Hickey, Pong-Wong, Daetwyler, & Calus, 2013). With the prediction model, phenotypic values for non-phenotyped plant genotypes are predicted, which are subsequently used for selection (Hunt, van Eeuwijk, Mace, Hayes, & Jordan, 2018).

The first attempts to incorporate and simultaneously estimate SNP effects to predict phenotypic values were made by (Bernardo, 1994, 1996). These have been popularized by (Whittaker, Thompson, & Denham, 2000) and (Meuwissen, Hayes, & Goddard, 2001) and have been repeatedly used in plant and animal breeding (Bernardo, 2008; VanRaden, 2008; Crossa et al., 2010). However, the availability of high-density SNP panels, where the number of markers (p) exceeds the sample size (n), implies that regularization methods are required in order to estimate all effects.

5.1.1 Common regularization approaches

The most common approach is by using the genomic best linear unbiased predictor (*GBLUP*) method, where a mixed model is fitted to the data with the marker effects as random (normally and independently distributed effects with a common variance) (VanRaden, 2008; de Los Campos et al., 2009). GBLUP has also been alternatively parameterized as a ridge regression (Hoerl & Kennard, 1970) model (referred to as *RRBLUP*) for genomic prediction (Piepho et al., 2012). Therefore, the level of SNP effect shrinkage can be determined with either a grid search over the regularization parameter for RRBLUP, or by using the ratio of variance components in GBLUP (Heslot, Yang, Sorrells, & Jannink, 2012). Finally, RRBLUP can also be parameterized in a Bayesian setting with a Gaussian prior for the marker effects (de Los Campos et al., 2013). We will use RRBLUP and GBLUP interchangeably in this work.

RRBLUP assumes that all SNP effects have equal variance, an assumption that has often been criticized, since both causal and non-causal SNPs receive the same amount of regularization. Contrarily, most of the SNPs in the genome are assumed to contribute little to the phenotype and therefore should be penalized more (Shen, Alam, Fikse, & Rönnegård, 2013). By assuming that SNP effects have different distributions, additional flexibility is added to the BLUP model. One such approach is *MultiBLUP* and *Adaptive MultiBLUP* (Speed & Balding, 2014) assigning different distributions to the effects, based on prior information or data-driven approaches. In these approaches, markers are assigned to groups with different variances expressing whether the markers have large or zero to small contribution to the phenotypic variance. Each group of markers forms a separate genomic relationship matrix.

Another encompassing approach to regularization is by assigning certain prior densities to the marker effects in the Bayesian setting. Using a t-density (which puts more mass at zero and has thicker tails relative to the Gaussian density), for example, implies that small effects receive stronger shrinkage towards zero than strong effects. This approach is colloquially known as *BayesA* (Meuwissen et al., 2001). *BayesB* (Meuwissen et al., 2001) and *BayesC* (Habier, Fernando, Kizilkaya, & Garrick, 2011) are obtained by assuming that SNP effects are a mixture of a point-mass at zero and a (diffuse) distribution on some finite interval. BayesB uses a t-density as the slab, while BayesC uses a normal density. Both induce a combination of variable selection and shrinkage (de Los Campos et al., 2013). Empirical studies show only small differences between GBLUP, BayesA, BayesB, and BayesC, with variable selection methods having better performance in scenarios with large-effect QTLs. When the number of SNPs is small, no difference in performance is observed (de Los Campos et al., 2013).

All aforementioned methods are based on the assumption of independence between SNP effects. Nonetheless, it is anticipated that SNPs will be correlated due to spatial proximity within the chromosomes (Gianola, Perez-Enciso, & Toro, 2003). For modeling the correlation between the effects *ante-BayesA*, *ante-BayesB*, and *BayesN* have been proposed (Yang & Tempelman, 2012; J. Zeng et al., 2018).

In these approaches the effect of a SNP is estimated with respect to the relative physical distance of its preceding neighbour, i.e., they have a distance-specific *ante-dependence* parameter (Núñez-Antón & Zimmerman, 2009). While these are very interesting Bayesian approaches dealing with the spatial proximity of the SNPs, they involve Markov Chain Monte Carlo methods, which become computationally prohibitive for models involving many variables. We offer a simpler alternative method based on penalized regression to account for the spatial proximity.

5.1.2 Contribution

In this article we propose, motivated by the network constrained regularization and variable selection (C. Li & Li, 2008), a regularized linear model: the proximity smoothed BLUP (psBLUP). (C. Li & Li, 2008) use a combination of L_1 (Lasso) and L_2 (ridge) penalties. The former is used for variable selection, the latter for encouraging smoothness on neighboring marker effects. psBLUP uses an L_2 instead of an L_1-norm on the coefficients (like RRBLUP), while similarly to (C. Li & Li, 2008) it imposes a second L_2-norm to encourage smoothness on neighboring effects. psBLUP explicitly accounts for the dependence between marker effects due to the SNPs' relative spatial proximity within chromosomes. A smooth solution on the differences between adjacent marker effects is employed, since it is expected that neighboring markers are in LD with the same QTLs. One feature of the method is that we do not require a strict definition of the markers' proximity, which can be estimated from the data. For example, the correlation coefficient between markers can be used as a measure of LD (Zaykin, Pudovkin, & Weir, 2008). In our applications, we use the squared correlation coefficient for those SNP pairs being equal or less than 10 centimorgan (cM) apart as a measure of proximity and observe that it is sufficient to outperform RRBLUP in terms of accuracy. A big advantage of the method is that its implementation can be done by standard software able to fit linear mixed models (e.g., *R, Python*, etc.), making it easily accessible.

5.1.3 Overview

The remainder is organized as follows. In Section 5.2, we review RRBLUP and propose the psBLUP as a way of incorporating information on the SNPs proximity in genomic prediction. This section also introduces the data with which the two methods (RRBLUP vs psBLUP) are compared in terms of predictive ability: Arabidopsis thaliana data coming from the Seed Lab of Wageningen University and Research, and Barley data from the North American Barley Genome Mapping Project (NABGMP). In Section 5.3 we demonstrate our approach on these two applications and show that psBLUP can lead to a gain in accuracy. We conclude in Section 5.4 by discussing possible extensions for computational efficiency and the advantages of the method in settings with limited sample sizes or low heritability phenotypes.

5.2 Materials and methods

5.2.1 Phenotyped and genotyped datasets

Population 1: Arabidopsis thaliana data from Wageningen

The first population is a Recombinant Inbred Line (RIL) population created from a cross between two natural Arabidopsis accessions, i.e., Bayreuth (*Bay-0*) and Shahdara (*Sha*). The data come from the Seed Lab of Wageningen University and Research (Netherlands). Seeds of 164 RILs were divided into four sub-populations (41 lines each) representing four important developmental stages of seed germination. The concentration levels of 161 metabolites were determined for all 164 lines. Finally, 64 metabolites were retained to be used for further analysis as phenotypes. Concentration levels of the metabolites were log-transformed and adjusted for the four developmental seed stages by subtracting the mean levels from each group. Finally, information on $p = 1059$ markers (5 chromosomes) was available. More information on the study design and data can be found in (Joosen, 2013) and (Joosen et al., 2013).

Population 2: Barley data from NABGMP

The second population concerns the well-known *Steptoe* × *Morex* doubled haploid (DH) population developed by the NABGMP (`https://wheat.pw.usda.gov/ggpages/SxM/`). This DH population was developed between 1991 and 1992 at several locations in North America. It consists of $n = 150$ DH lines of Barley that were evaluated in different environments. We retained five traits for further analysis, i.e., yield (measured in 16 environments), percentage of grain protein (measured in 9 environments), percentage of malt extract (measured in 9 environments), line's height (measured in 16 environments), and the degree of α-amylase activity (measured in 9 environments). A total of 148 lines were genetically characterized by $p = 794$ markers covering the seven barley chromosomes. More information on the study design and data can be found in (P. Hayes et al., 1993) and (Malosetti, Voltas, Romagosa, Ullrich, & Van Eeuwijk, 2004).

5.2.2 Methods for genomic prediction

Let, for n samples, $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$ be a $n \times 1$ centered response vector representing a phenotype of interest ($\sum_i y_i = 0$). Also, let $\boldsymbol{X}$ be a $n \times p$ matrix containing scaled SNPs ($\sum_i x_{ij} = 0$, $\sum_i x_{ij}^2 = n$ for all $j = 1, \ldots, p$). In order to build a genomic prediction model and establish a genotype-phenotype relationship, a vector of SNP effects needs to be estimated. We first present the standard RRBLUP model, before extending to psBLUP.

RRBLUP

In RRBLUP the vector of SNP effects is obtained by minimizing the penalized least squares with respect to $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{RR} := \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{X\beta})^\top (\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda_1 \boldsymbol{\beta}^\top \boldsymbol{I}_p \boldsymbol{\beta} \right\}, \tag{5.1}$$

where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix and where $\lambda_1 \geq 0$ represents the shrinkage parameter controlling the amount of regularization. Since $\hat{\boldsymbol{\beta}}$ depends on λ_1, a cross-validation criterion is typically used to select λ_1 from a grid of possible values.

Another way to select λ_1 is by estimating the variance components of a mixed model with SNP effects as random, since the two models are equivalent (Habier, Fernando, & Dekkers, 2007; Piepho et al., 2012; de Los Campos et al., 2013; de Vlaming & Groenen, 2015). The linear mixed model can be written as:

$$\boldsymbol{y} = \boldsymbol{Xu} + \boldsymbol{\varepsilon}, \tag{5.2}$$

where $\boldsymbol{\varepsilon}$ are the residuals distributed as $N(0, \sigma^2_{\boldsymbol{\varepsilon}} \boldsymbol{I}_n)$ and $\boldsymbol{u}$ are the random effects distributed as $N(0, \sigma^2_{\boldsymbol{u}} \boldsymbol{I}_p)$. The ridge regression model with $\lambda_1 = \sigma^2_{\boldsymbol{\varepsilon}}/\sigma^2_{\boldsymbol{u}}$ gives the same estimated SNP effects as (5.2) (i.e., $\hat{\boldsymbol{\beta}}_{RR} = \hat{\boldsymbol{u}}$). Selecting λ_1 and calculating the SNP effects based on the mixed model is often preferred due to its computational efficiency (S. A. Clark & van der Werf, 2013).

SNP proximity matrix

Before presenting the penalized least squares for obtaining psBLUPs, we briefly introduce the proximity between the SNPs, represented as a matrix. Let $\boldsymbol{W}$ be a matrix containing information on the spatial relationship between SNPs. For example, the matrix element $w_{jj'}$ could contain the LD between the jth and j'th SNPs or the relative (physical/genetic) distance between them. Here, $\boldsymbol{W}$ is calculated using the square of markers' pairwise Pearson correlation coefficient (VanLiere & Rosenberg, 2008) if they are close. We deem markers whose genetic distance is equal or less than 10cM to be close. A genetic distance of 10cM concurs with a recombination rate of at most .1 (Hartl, 2011) which translates to a Pearson correlation of at least .6 (Warrens, 2008). Let j and j' be two SNP indices, let g_j and $g_{j'}$ be the physical/genetic position of the two corresponding SNPs on the chromosome, and let $\boldsymbol{x}_j$ and $\boldsymbol{x}_{j'}$ be two vectors containing genetic information on n samples for those SNPs. The matrix element $w_{jj'}$ is then defined as:

$$w_{jj'} = w_{j'j} = \begin{cases} \rho(\boldsymbol{x}_j, \boldsymbol{x}_{j'})^2, & \text{if } |g_j - g_{j'}| \leq 10\text{cM}, \\ 0, & \text{otherwise}, \end{cases} \tag{5.3}$$

where $\rho(\boldsymbol{x}_j, \boldsymbol{x}_{j'})$ is the Pearson correlation between SNPs j and j'. By that definition, each SNP can be viewed as the center of a local network of SNPs, and is connected to SNPs up to 10cM away. Essentially, for these connections, the squared correlation coefficient is calculated.

Figure 5.1 contains a toy example illustrating how chromosomal spatial information is translated to network information that is explicitly used in psBLUP. On the top panel (chromosomal representation), six SNPs are marked on a segment of a chromosome. The distances between SNPs equal or less than 10cM have been shown with dashed lines. On the center panel, the same SNPs are represented as nodes in a network where an edge is connecting a pair of SNPs if their distance is less than or equal to 10cM. The width of the edges is analogous to the proximity between two SNPs. Finally, the network is represented as a matrix (bottom panel), where the similarity between connected SNP pairs is coded in grey-colored circles. A darker color indicates a stronger similarity. Empty cells imply that the distance between two SNPs is larger than 10cM and they do not share a connection in the network representation.

To estimate the SNP effects using psBLUP we need to calculate the normalized Laplacian matrix $\boldsymbol{L}$ (Chung & Graham, 1997) of $\boldsymbol{W}$ with elements:

$$l_{jj'} = l_{j'j} = \begin{cases} 1 - w_{jj'}/s_j, & \text{if } j = j' \text{ and } s_j \neq 0, \\ -w_{jj'}/\sqrt{s_j s_{j'}}, & \text{if } j \neq j' \text{ and } w_{jj'} \neq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{5.4}$$

where $s_j = \sum_{j'} w_{jj'}$ is the weighted total connectivity of SNP j.

psBLUP

The SNP effects are obtained by minimizing the proximity-penalized least squares with respect to $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{ps} := \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{X\beta})^\top (\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda_1 \boldsymbol{\beta}^\top \boldsymbol{I}_p \boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}^\top \boldsymbol{L\beta} \right\}, \tag{5.5}$$

where $\boldsymbol{L}$ is the normalized Laplacian matrix obtained with expression 5.4 and $\lambda_2 \geq 0$ is the parameter inducing shrinkage on the differences between SNP effects analogous to their proximity. Finally, as in expression 5.1, the term $\boldsymbol{\beta}^\top \boldsymbol{I}_p \boldsymbol{\beta}$ is the L_2-norm shrinking the SNP coefficients.

The term $\boldsymbol{\beta}^\top \boldsymbol{L\beta}$ can also be written as (C. Li & Li, 2008):

$$\boldsymbol{\beta}^\top \boldsymbol{L\beta} = \sum_{j=1}^{p} \sum_{j'=1}^{p} \left(\frac{\beta_j}{\sqrt{s_j}} - \frac{\beta_{j'}}{\sqrt{s_{j'}}} \right)^2 w_{jj'}. \tag{5.6}$$

This implies that the psBLUPs are smoothed by penalizing the sum of weighted squares of the differences between them. Therefore, when SNPs j and j' are close on the chromosome, they are expected to have almost equivalent association to $\boldsymbol{y}$ and thus similar effects, translating in a small difference in coefficients.

Solving psBLUP

Following (Zou & Hastie, 2005) and (C. Li & Li, 2008), we reduce the problem in (5.5) to a ridge regression using the augmented data solution. Let, $\boldsymbol{Q\Lambda Q}^\top$ be

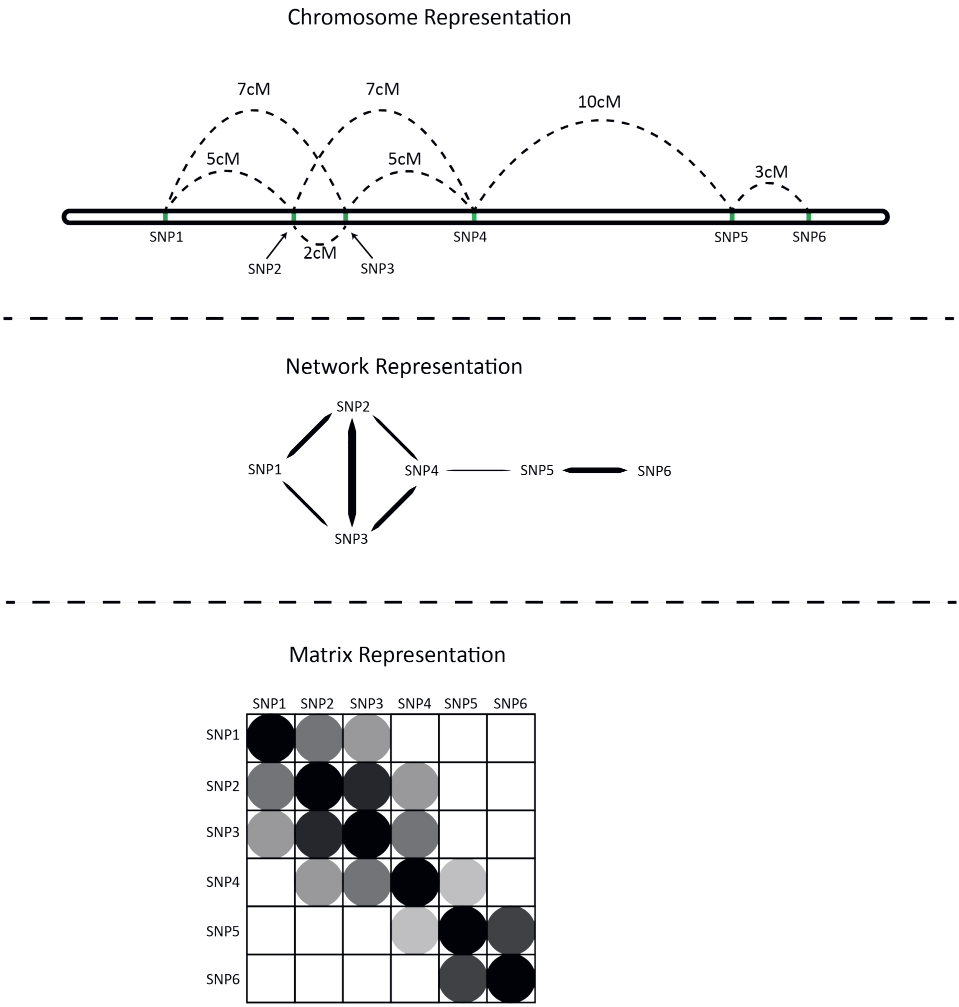


Fig. 5.1: **Chromosomal representation**: six SNPs are marked on a part of a chromosome. Dashed lines indicate the distance between pairs of SNPs. **Network representation**: the six SNPs are represented as nodes in a network with edges connecting only SNPs with distance equal or less than 10cM. SNPs proximity is encoded as edges' width (SNPs with low distance have wider edges). **Matrix representation**: the similarity of all pairs of SNPs is coded using grey-colored circles. Higher similarity is encoded with darker grey. Empty cells indicate that a pair of SNPs does not share an edge in the network representation.

the eigendecomposition of the $p \times p$ normalized Laplacian matrix $\boldsymbol{L}$, with $\boldsymbol{Q}$ the $p \times p$ matrix of eigenvectors and $\boldsymbol{\Lambda}$ the diagonal matrix with the eigenvalues. Define $\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}$, $\gamma = \lambda_1/\sqrt{1+\lambda_2}$, and $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$. The new $(n+p)$-dimensional vector of responses $\boldsymbol{y}^*_{(n+p)}$ and $(n+p) \times p$ matrix of predictors $\boldsymbol{X}^*_{(n+p)\times p}$ are then defined as:

$$\boldsymbol{y}^* = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix}, \qquad \boldsymbol{X}^* = \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2}\boldsymbol{T}^\top \end{pmatrix}.$$

Using $\boldsymbol{y}^*$ and $\boldsymbol{X}^*$, expression (5.5) is rewritten as:

$$\hat{\boldsymbol{\beta}}^*_{ps} := \arg\min_{\boldsymbol{\beta}^*} \left\{ (\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*)^\top (\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*) + \gamma \boldsymbol{\beta}^{*\top} \boldsymbol{I}_p \boldsymbol{\beta}^* \right\}, \tag{5.7}$$

which is a conventional ridge regression model in the augmented data $\boldsymbol{y}^*$ and $\boldsymbol{X}^*$.

Fitting a mixed model is less computationally demanding than the search for an optimal penalty-value for ridge regression. We select the psBLUPs and the regularization parameter γ using the following model:

$$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{u}^* + \boldsymbol{\varepsilon}^* \tag{5.8}$$

where $\boldsymbol{\varepsilon}^*$ is the vector of residuals distributed as $N(0, \sigma^2_{\boldsymbol{\varepsilon}^*}\boldsymbol{I}_{(p+n)})$ and $\boldsymbol{u}^*$ is distributed as $N(0, \sigma^2_{\boldsymbol{u}^*}\boldsymbol{I}_p)$. As the accuracy in terms of correlation is not sensitive to its value, λ_2 was assessed along a crude grid of equidistant values (ranging from 1 to 75). Finally, $\gamma = \sigma^2_{\boldsymbol{\varepsilon}^*}/\sigma^2_{\boldsymbol{u}^*}$ and therefore, $\lambda_1 = (\sqrt{1+\lambda_2})\sigma^2_{\boldsymbol{\varepsilon}^*}/\sigma^2_{\boldsymbol{u}^*}$. Fitting a ridge regression model was done by using the augmented design matrix as input to the *rrBLUP* R-package (Endelman, 2011). The solution to (5.5) is then obtained as $\hat{\boldsymbol{\beta}}_{ps} = (1+\lambda_2)^{-1/2}\hat{\boldsymbol{\beta}}^*_{ps}$.

5.2.3 Evaluation

We evaluate RRBLUP and psBLUP using the following approach. We split the data in training and test sets based on three scenarios:

1. use 25% of the data for training and 75% for testing,
2. use 50% of the data for training and 50% for testing,
3. use 75% of the data for training and 25% for testing.

For each case, RRBLUPs and psBLUPs are estimated. The correlation between the fitted and observed values is used to assess the accuracy of each method. We repeat the process 100 times for computing a mean gain/loss of psBLUP compared to RRBLUP. For each iteration, we calculate the difference in accuracy between psBLUP and RRBLUP. Then, the mean accuracy gain/loss is calculated as the average of the accuracy difference, over the 100 runs.

The selection of scenarios is justified as follows: by using 25-75 training-test split, we investigate how good the model performs when there is little information for estimating SNP-phenotypic relationships, and how in such cases having proximity information can help improve accuracy when generalizing to a much larger population. Inversely, selecting a 75-25 training-test split can show two things: (i) that when having more power and most SNP-phenotypic relationship is explained, spatial information may not add information; (ii) nevertheless, if the sample size is still not an important aspect because studying low heritability traits, spatial information on SNPs can still improve accuracy. Finally, the 50-50 training-test split uses the same number of samples for training and testing.

5.3 Results

5.3.1 Application 1: Wageningen Arabidopsis thaliana data

Here, we want to assess the gain in predictive accuracy when using information on the spatial proximity of the markers, by comparing psBLUP to RRBLUP for 64 metabolites. The markers' proximity was measured using expression (5.3).

The mean accuracy, for each of the three (sample size) scenarios and for each of the two models, was determined as the mean correlation coefficient across all 100 realizations between the predicted genotypic values and observed phenotypes of the test data. A summary of the results is presented in Figure 5.2 for the scenario using 50% of the data for training and the rest for testing (the results for all scenarios can be found in the Supplementary Material). It can be seen that on average, psBLUP gives higher accuracy than RRBLUP, since the gain in accuracy is positive. The mean difference between psBLUP and RRBLUP was 3.3%. The results have also been summarized in Table 5.1.

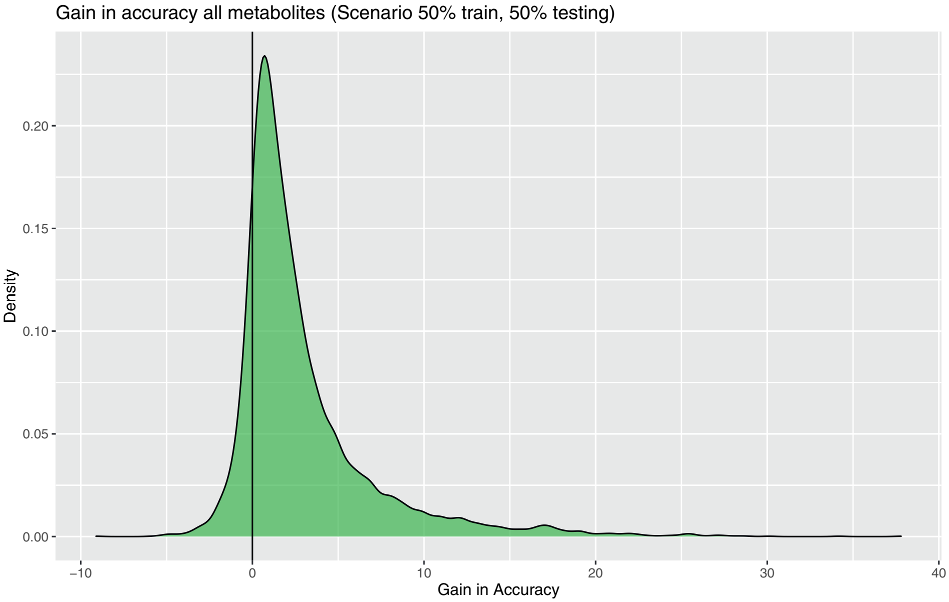


Fig. 5.2: The gain in accuracy when using psBLUP vs RRBLUP for 64 metabolites. The x-axis is expressed in percentages. For every metabolite, psBLUP and RRBLUP was fitted 100 times by randomly sub-sampling 50% of the samples to be used for training the models and 50% for testing.

In Figure 5.2 we observe that the differences in predictive ability between psBLUP and RRBLUP are consistent. Results indicate that phenotypic information

Training set	RRBLUP accuracy	psBLUP accuracy	Gain in accuracy	% of times psBLUP > RRBLUP
25%	20.99% (3.06, 57.84)	24.45% (6.56, 60.77)	3.46% (1.50, 6.85)	**86.6%**
50%	26.49% (4.61, 65.20)	29.40% (9.08, 66.39)	2.91% (0.77, 6.98)	**86.3%**
75%	29.55% (5.00, 69.47)	33.09% (12.61, 70.21)	3.54% (1.17, 8.41)	**86.9%**
Mean	25.68% (4.16, 66.10)	28.98% (7.93, 67.43)	3.30% (1.87, 7.17)	**86.6%**

Table 5.1: The predictive ability of RRBLUP vs psBLUP together with their observed difference using the Arabidopsis metabolite data from Wageningen University Seed Lab. psBLUP and RRBLUP were fitted 100 times under random subsampling for different scenarios: (i) 25% of the samples used for training and 75% for testing, (ii) 50% of the samples used for training and 50% for testing, and (iii) 75% of the samples used for training and 25% for testing. The accuracy is calculated over all iterations of the process. The parentheses contain the $5th$ and $95th$ percentile of the point estimate.

is contained within markers' correlation structure, since using information on the proximity between them yields improved accuracy. In Table 5.1, the accuracy using RRBLUP and psBLUP has been summarized together with the estimated gain (the $5th$ and $95th$ percentile is displayed in the parentheses). In both cases (RRBLUP and psBLUP), the accuracy increases with larger training sample sizes, as expected. The gain in accuracy when using psBLUP ranges for 2.91% to 3.54% in all training set scenarios. In the last column of Table 5.1 we see that psBLUP yields superior accuracy from RRBLUP in more than 86% of the cases for any scenario.

Interestingly, when the predictive accuracy using RRBLUP is high, the gain using psBLUP is small. Inversely, the gain using marker proximity is higher when the genomic prediction model is not so informative. This result has been visualized in Figure 5.3. Each dot represents the mean accuracy using RRBLUP and mean gain in accuracy when psBLUP is used, over 100 runs. For metabolites with high predictive accuracy using RRBLUP, the gain in psBLUP is small, while the highest gains using psBLUP have been observed for metabolites with very low predictive accuracy using RRBLUP.

5.3.2 Application 2: NABGMP barley data

In this application we assess the gain in predictive accuracy when using information on the spatial proximity of the markers, by comparing psBLUP to RRBLUP for 59 trait-environmental combinations (Barley data from NABGMP). The markers proximity was measured using expression (5.3).

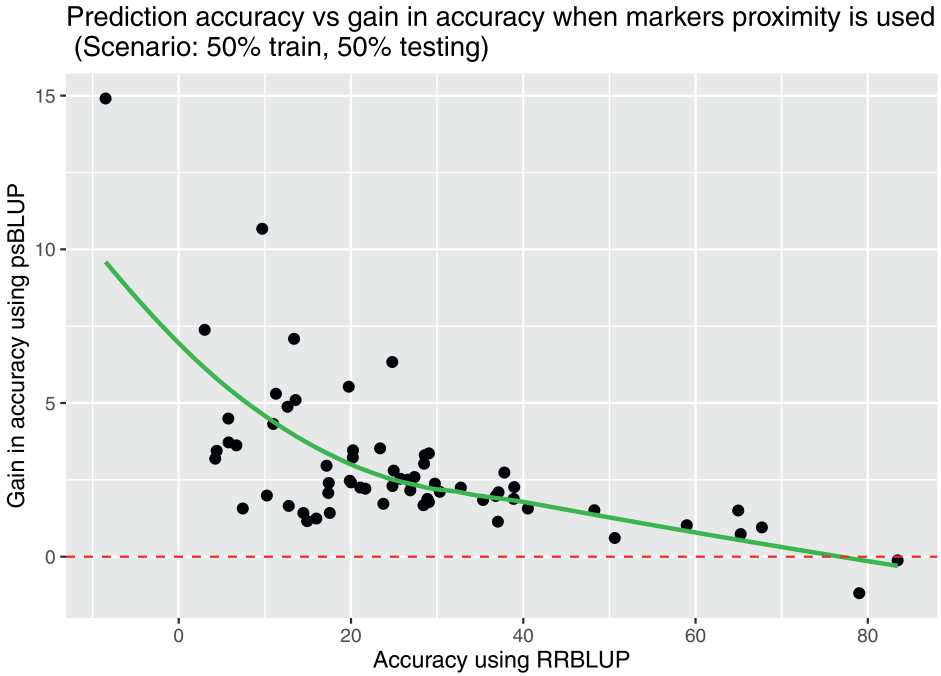


Fig. 5.3: Prediction accuracy vs gain in accuracy for the 64 metabolites used (points in the plot) when markers proximity is used. The y and x-axes are expressed in percentages. The y-axis shows the percentage gain in prediction accuracy when psBLUP is used instead of RRBLUP. The x-axis shows the percentage accuracy for a metabolite. Each dot represents the mean accuracy using RRBLUP and mean gain in accuracy when psBLUP is used, over 100 runs.

As in the first application, the mean accuracy of the models was determined using the mean correlation coefficient between the predicted and observed phenotypes of the test data for each of the three (sample size) scenarios over 100 runs. A summary of the results is presented in Figure 5.4 for the scenario with half the samples used for training and the rest for testing. The results have also been summarized in Table 5.2.

In Figure 5.4 we see that the mean difference in predictive ability between psBLUP and RRBLUP is positive in some cases. In Table 5.2 the results have also been summarized. Across all traits, the accuracy increases for larger sample sizes using either genomic prediction method (RRBLUP or psBLUP). The 5th and 95th percentiles are displayed in the parentheses for each trait-subsampling scenario. In the last column of Table 5.2 the percentage of times psBLUP yields greater accuracy than RRBLUP is shown.

As in the metabolite data application, the gain in predictive accuracy is greater when the accuracy using RRBLUP is lower. The scenario with 50% of the data used

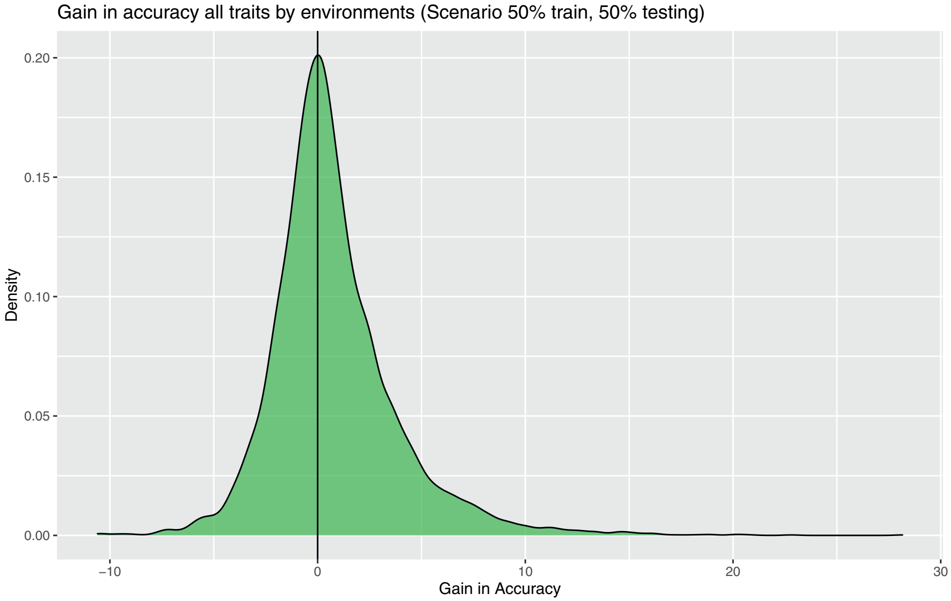
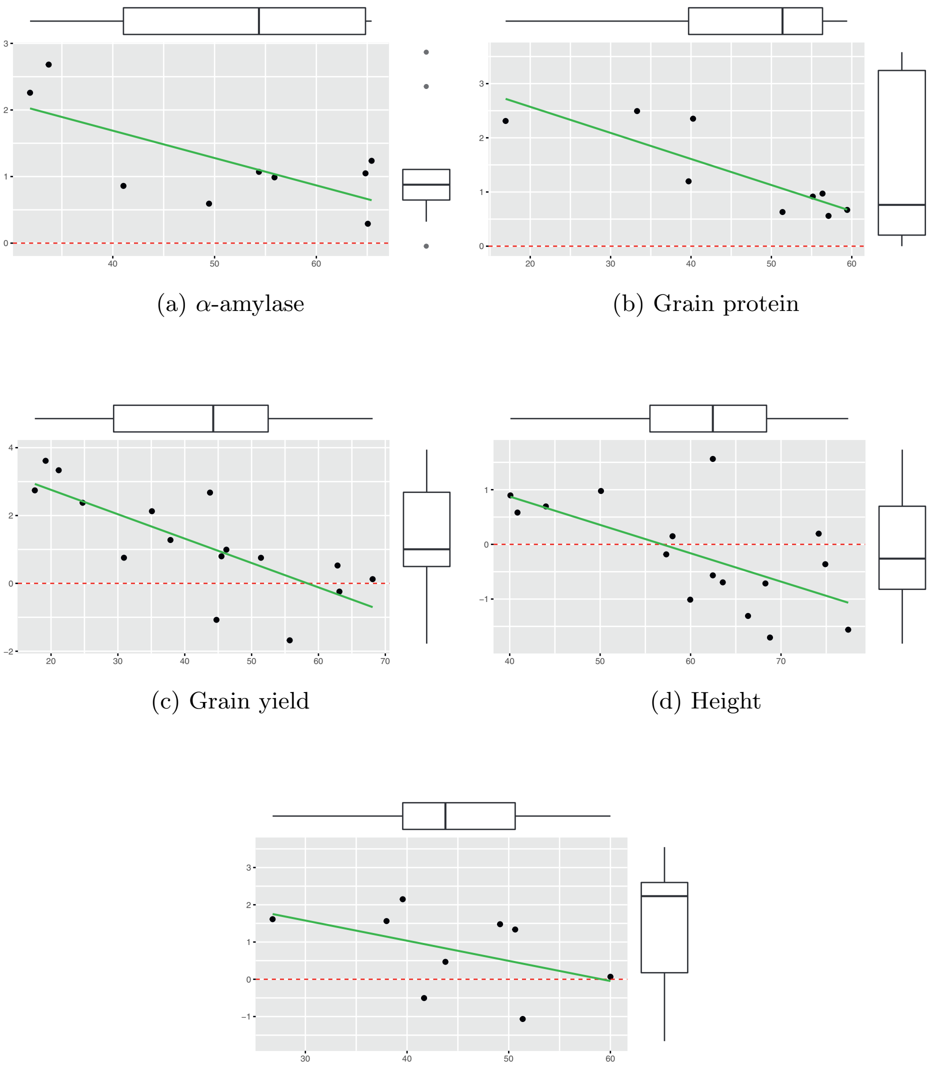


Fig. 5.4: For every combination of the 59 trait-environments of the NABGMP dataset, 100 psBLUP and RRBLUP models were fitted when using 50% of the data for training and the rest for testing. The x-axis is expressed in percentages and shows the absolute difference in accuracies between the best selected psBLUP model and RRBLUP.

as training and the rest as testing (for all five phenotypes) has been visualized in Figure 5.5 were a downward trend can be seen. Each dot shows the mean RRBLUP accuracy and gain in accuracy when using psBLUP over 100 runs. With regard to the traits, we see that plant height has overall the highest accuracy using RRBLUP and subsequently the lowest gain when using psBLUP. The scenarios using a 25-75 and 75-25 split for training and testing can be found in the Supplementary Material.

5.4 Discussion

In this work, we developed a regularized regression model that uses information on the proximity of the explanatory variables in order to increase prediction accuracy. Our model (psBLUP) was used in the context of genomic prediction as an extension of RRBLUP: the spatial proximity between the SNPs was used to improve the predictive ability of RRBLUP. When no penalty is used to account for the dependence between SNP effects, the two methods should be identical by definition.



(a) α-amylase

(b) Grain protein

(c) Grain yield

(d) Height

(e) Malt extract

Fig. 5.5: Percentage gain in prediction accuracy (y-axis) when using psBLUP vs percentage prediction accuracy per trait when using RRBLUP (x-axis). Each dot represents a trait-environmental combination showing the mean RRBLUP accuracy and gain in accuracy when using psBLUP over 100 runs. y- and x- marginal boxplots show the psBLUP gain in accuracy and RRBLUP accuracy, respectively. For ease of comparison, fitted regression lines have been added. The x-axis of each plot represents the RRBLUP accuracy measured in percentages, while the y-axis is the psBLUP gain in accuracy measured in percentages.

For demonstrating the proposed approach two applications were considered. In both applications we utilized SNP information in order to build a prediction model for the responses, using psBLUP and RRBLUP. The two methods were compared with regard to their prediction accuracy. For 100 replications, the models were trained on the 25%, 50%, and 75% of the samples, while the rest were used for evaluating the results.

In the first application, the data were part of a RIL population of 164 lines with 1059 SNPs, and 64 metabolites. The results implied clear superiority of psBLUP over RRBLUP in most of the metabolites and training sample choices. The mean gain in accuracy was 3.3% when using psBLUP instead of RRBLUP. Interestingly, the greatest gain in accuracy was observed when the RRBLUP model was less accurate. On the other hand, metabolites with high RRBLUP accuracy had the smallest gain when psBLUP was used.

In the second application, the data were part of the Steptoe × Morex DH barley population having 148 lines characterized by 794 SNPs. Similarly to the first application, an inverse relationship between accuracy using RRBLUP and accuracy gain when using psBLUP was observed. In the Steptoe × Morex DH barley population, several trait-environmental combinations with accuracy loss were observed. This was mainly observed for height and was combined with high RRBLUP accuracy.

A few things can be noted for the inverse relationship between accuracy gain and training sample size, i.e., greater gain for smaller training sample sizes. In cases were the training sample size is small, the accuracy of the RRBLUP model is expected to be low. Therefore, the variation margin that can be explained by the SNPs' spatial proximity (psBLUP) is high. Modeling the spatial proximity/accounting for correlation between SNP effects is therefore more important for low heritability and smaller training sets.

We note that in some cases (e.g., association panel) neighboring markers can have effects with opposite signs. Then they will wrongly tend to cancel out, leading to smaller overall accuracy. In that case, all predictors can be recoded to be positively associated with the response prior to model fitting. Alternatively, the squared scaled absolute differences between the SNP coefficients could be penalized in expression (5.6).

An advantage of the psBLUP approach is the broad applicability, since it can be used for any continuous outcome and type of predictor variables. Additionally, it can be implemented using standard statistical software that can fit a mixed model, making it easily accessible. Moreover, there is no strict definition for the markers spatial proximity, which can be estimated by the data or by using prior information making the data analysis more flexible.

Some issues still need to be addressed. We utilized the mixed model equivalence to ridge regression for reducing the model tuning to the evaluation of parameters that can be obtained with a single optimization. Even though the speed is greatly improved by solving the mixed model equations on the augmented data, the efficiency needs to be further improved for incorporating high density SNP panels. For estimating the penalized coefficients of the model, the proximity matrix needs to be stored and decomposed. When the number of SNPs is high, the memory needed to store such matrix is sizable. Such problem can partially be solved by encoding the matrices in sparse format. Still, the matrix needs to be decomposed to its eigenvectors and eigenvalues which becomes intensive for big p.

For computational efficiency, when the number of variables far exceeds the number of samples, an alternative parameterization can be used by writing model (5.8) as a single trait mixed model with subject-specific random effects. Let, $\boldsymbol{G} = \boldsymbol{X}^*\boldsymbol{X}^{*\top}$ be the realized additive relationship matrix indicating the relatedness between individuals. By ignoring any fixed effects, the mixed model with subject-specific random effects is written as:

$$\boldsymbol{y}^* = \boldsymbol{\alpha}^* + \boldsymbol{\varepsilon}^* \tag{5.9}$$

where $\boldsymbol{\alpha}^* \sim N(0, \boldsymbol{G}\sigma^2_{\boldsymbol{\alpha}^*})$. The information connecting subject-specific effects $\hat{\boldsymbol{\alpha}}^*$ to SNP effects $\hat{\boldsymbol{u}}^*$ is contained in $\boldsymbol{X}^*$ (Shen et al., 2013). After $\hat{\boldsymbol{\alpha}}^*$ is obtained, the SNP effects can be acquired as:

$$\hat{\boldsymbol{u}}^* = \boldsymbol{X}^{*\top}\boldsymbol{G}^{-1}\hat{\boldsymbol{\alpha}}^*. \tag{5.10}$$

Even though the search grid for the tuning parameter in psBLUP is reduced to one dimension since the mixed model solution is used, the computational time can be demanding for high p and high n by working with the augmented data solution i.e., the predictor data set is a $(n+p)\times p$ matrix. One approach to making the solution more efficient is by estimating the SNP coefficients per chromosome. Since SNPs are considered independent between chromosomes, multiple regularized linear models can be fit. Such approach could potentially yield superior accuracy by estimating chromosome specific regularization parameters and thus making the fit more flexible (by working with much smaller matrices). In addition, a shared λ_1 can also be estimated for each chromosome while λ_2 can vary per chromosome allowing for a better spatial flexibility per chromosome. In that case, the mixed model solution cannot be employed anymore.

Alternatives to psBLUP are the ante-dependence models (Yang & Tempelman, 2012; J. Zeng et al., 2018). These Bayesian models are based on the idea that SNP coefficients are dependent. A typical shortcoming of Bayesian methods is the computational time needed for estimating all coefficients using MCMC methods. For p SNPs, when only the first neighbor is considered (first order dependence), $2p-1$ coefficients need to be estimated, making it burdensome for higher order dependencies and more dense SNP panels. Naturally, for every new SNP incorporated to the model, at least two more coefficients need to be estimated, resulting in additional computational time. We feel that psBLUP offers an alternative perspective to the same problem using a simpler set-up. Finally, the choice of connected neighbors in the ante-dependence models is fixed for all SNPs, while psBLUP allows for different number of neighbors per SNP, making it more flexible.

5

Important future research needs to be done. First, assessing how sensitive the results are to the selection of the proximity matrices. In this paper, we restricted the range within which SNPs were allowed to contribute information to 10cM, which for segregating populations like RILs and DHs is equivalent to a correlation between markers of .6. One could play around with this number to see whether the performance of psBLUP improves. For our choice of 10cM psBLUP often outperformed RRBLUP. Second, a more detailed evaluation of the sample size effect on the estimated accuracy needs to be done. Here, we used 25, 50, and 75% of the data samples as tests. A random subsample (as small as 25% of the original data) can initially be used in any study, to determine what is the maximum potential gain from psBLUP and what are some possible values for the smoothing parameter λ_2.

Finally, the sensitivity to the number of SNP needs to be studied. We would expect that the accuracy gain will be larger when using smaller number of SNPs. Using a big number of SNPs will naturally result in higher RRBLUP accuracy, thus smaller gain.

	Training set	RRBLUP accuracy	psBLUP accuracy	Gain in accuracy	% of times psBLUP > RRBLUP
Grain Yield	25%	34.47%	36.93%	2.46%	**69.9%**
		(11.5, 59.57)	(16.51, 61.6)	(0.03, 5.55)	
	50%	41.75%	42.94%	1.19%	**57.4%**
		(18.79, 64.38)	(22.19, 64.59)	(-1.23, 3.4)	
	75%	45.38%	46.54%	1.16%	**54.2%**
		(21.56, 69.35)	(26.67, 69.98)	(-1.93, 5.38)	
Mean		40.53%	42.14%	1.61%	
		(15.68, 66.87)	(19.34, 67.22)	(-1.47, 5.06)	
Grain Protein	25%	38.17%	40.66%	2.49%	**82.2%**
		(18.28, 51.15)	(21.06, 53.29)	(1.98, 2.92)	
	50%	45.51%	46.86%	1.35%	**65.8%**
		(23.48, 58.5)	(25.86, 59.13)	(0.59, 2.44)	
	75%	49.54%	51.04%	1.50%	**60.8%**
		(27.81, 65.86)	(30.4, 66.41)	(0.33, 2.89)	
Mean		44.41%	46.19%	1.78%	
		(18.27, 62.8)	(20.67, 63.11)	(0.57, 2.94)	
Malt extract	25%	36.77%	38.90%	2.13%	**73.4%**
		(24.76, 48.13)	(26.94, 50.06)	(1.22, 2.85)	
	50%	44.55%	45.34%	0.79%	**60.2%**
		(31.26, 56.55)	(32.85, 56.83)	(-0.84, 1.93)	
	75%	47.53%	49.12%	1.59%	**61.7%**
		(34.06, 61.56)	(36.45, 62.22)	(-0.15, 3.01)	
Mean		42.95%	44.45%	1.50%	
		(27.63, 58.88)	(29.53, 59.49)	(-0.41, 2.99)	
Height	25%	51.81%	52.98%	1.17%	**58.2%**
		(32.91, 68.09)	(35.91, 68.38)	(-0.49, 3.07)	
	50%	60.54%	60.35%	-0.19%	**39.5%**
		(40.66, 75.53)	(41.32, 74.87)	(-1.6, 1.12)	
	75%	64.59%	64.66%	0.06%	**42.0%**
		(43.83, 79.66)	(44.58, 79.02)	(-0.97, 1.07)	
Mean		58.98%	59.33%	0.35%	
		(36.99, 77.25)	(38.41, 76.45)	(-1.21, 2.48)	
α-Amylase	25%	44.74%	47.24%	2.50%	**79.8%**
		(25.81, 60.77)	(29.59, 62.8)	(1.79, 3.78)	
	50%	51.29%	52.52%	1.23%	**66.4%**
		(32.61, 65.28)	(35.03, 66.35)	(0.41, 2.51)	
	75%	53.29%	54.63%	1.34%	**61.1%**
		(34.38, 68.57)	(37.28, 69.47)	(0.21, 2.91)	
Mean		49.77%	51.46%	1.69%	
		(28.55, 67.21)	(31.32, 68.23)	(0.33, 3.15)	

Table 5.2: The predictive ability of RRBLUP vs psBLUP together with their observed difference when using the DH barley data from NABGMP. psBLUP and RRBLUP were fitted 100 times under random subsampling for 3 scenarios: (i) 25% of the samples used for training and 75% for testing, (ii) 50% of the samples used for training and 50% for testing, and (iii) 75% of the samples used for training and 25% for testing. The accuracy is calculated over all iterations of the process. The parentheses contain the 5*th* and 95*th* percentile of the point estimate.

6

Discussion

This work has developed a comprehensive framework for the reconstruction of metabolite networks while meticulously addressing a wide array of biological and technical variations inherent in the study design. In our research setup, we have a specific set of metabolites that we are keen to explore in terms of their network structure. To facilitate the reconstruction of biologically significant associations, we also incorporate additional covariates linked to environmental conditions or aspects of the study design that pique our interest, or that we seek to incorporate into our analysis. Through this approach, we have endeavored to detect (conditional) pairwise associations among metabolites, driven by specific covariates, while simultaneously correcting for any undesired confounding effects.

In our pursuit of estimating metabolite networks within the context of the study design, we adopt a regression framework. In this framework, we conduct regression analyses on the metabolites, employing a design matrix that encompasses the covariates and study design variables that we aim to account for or find interesting. Additionally, we explore the incorporation of other omics data sources within the design matrix to enhance the interpretability of our findings. The fitted values, corrected for uninteresting sources of variation, are subsequently employed to construct the metabolite networks.

Our proposed methodologies have been put into practice with plant and human data, showcasing their versatility across both biological domains. While our primary emphasis was on the metabolome due to its sensitivity to both internal and external variations, and its pivotal role in bridging genetic information to phenotypic traits, it is worth noting that these methods can be readily adapted to other types of omics data as well.

The remainder of this chapter is organized as follows. First, we summarize the main contributions and findings (Section 1). Next, we discuss some limitations of the underlying work (Section 2). These limitations provide inroads for further research (Section 3). We then take perspective by discussing the possible impact of the proposed methodologies (Section 4). We close with some final reflections (Section 5).

6.1 Summary

In Chapter 2, we introduced a novel method (Bartzis et al., 2017) for the identification of sets of metabolites that exhibit similar associations with a specific covariate of interest. To achieve this, we employed a regression framework that enabled us to decompose the overall metabolic variation into distinct components related to the study design and random noise. We then selectively retained and harnessed the portion that was pertinent to the covariate of interest. We used this method in conjunction with a module identification method to effectively recover groups of metabolites that exhibited associations with the covariate of interest. When we applied our proposed method to plant data, our primary aim was to uncover metabolite associations that were dependent on genotypic information, all while accounting for potential nuisance effects. Similarly, in the context of human data, we successfully reconstructed metabolite networks driven by BMI information, while simultaneously controlling for biological variations like age and sex. This approach served to filter out random noise and variation originating from other sources (age, sex in the human data, and osmotic stress by polyethylene glycol (PEG) and abscisic acid (ABA) in the case of plant data), ultimately leading to the identification of biologically relevant clusters.

Building on these methods, Chapter 3 expands the proposed framework to encompass repeated measurements (Bartzis, Peeters, Uh, Houwing Duistermaat, & Eeuwijk, 2023). Our objective was to pinpoint sets of metabolites in situations where measurements were taken from the same individuals over time. Furthermore, we sought to estimate metabolite networks while factoring in genetic variations derived from an additional omics source, namely SNPs. We employed linear mixed models to address the correlations observed in these repeated measurements. In the pursuit of estimating metabolite networks, we incorporated the model's subject-specific effects, which represented lifestyle in this context, in conjunction with dietary preferences. This fusion allowed us to cluster together groups of metabolites that were biologically associated. In our analyses, we also considered SNP information and demographic characteristics as sources of variation that we needed to account for. Consequently, we permitted the edges in the estimated metabolite networks to be driven by variations we specifically targeted, such as lifestyle and dietary information. As a result, we were able to recover estimated modules that exhibited a high degree of homogeneity in terms of their constituent metabolites.

Chapter 4 (Bartzis, Peeters, Ligterink, & Van Eeuwijk, 2024) introduced a method for leveraging information from one type of omics data (referred to as "guiding data") to understand the network organization of metabolites (referred to as "target data"), showcasing its potential utility in various scientific applications, including QTL detection (when the guiding data consist of SNPs). This integrative network reconstruction method recovers associations between variables in the target data, while explicitly accounting for the graphical structure of the guiding data. The associations between guiding and target data are recovered by using network-aware regularized linear models, and target networks are then constructed based on the fitted values of the target variables regressed on the guiding variables. As a result, metabolites in the target data share connections if they display similar associations with variables in the guiding data. As demonstrated in Chapter 4, we successfully estimated metabolite networks driven by SNPs and gene expression, separating signal and noise variation. In this scenario, both omics sources represent vital biological information, serving as the driving force behind the construction of metabolite networks. Given that these two omics data types inherently possess their own network organization, we explicitly model this structure to enhance our ability to interpret the results. Notably, when we incorporated information from other omics sources, the recovered networks exhibit edges with a higher degree of certainty compared to networks constructed from metabolite data alone.

In Chapter 5, we extended the concept of utilizing an omics source and its correlation structure in linear models. We introduced a novel genomic prediction method called proximity smoothed BLUP (psBLUP) with the goal of enhancing predictive accuracy (Bartzis et al., 2022). This method leverages the spatial proximity of genetic markers on chromosomes. Nearby markers tend to exhibit associations with one another due to their physical closeness. To harness this valuable information, we employ a regularized linear model that explicitly takes into account this interdependence between predictor effects. We also discuss various approaches to implementing psBLUP by noticing equivalencies with the ridge solution and mixed models. Much like in Chapter 4, we regard the SNP data as the biological information on which the response must be conditioned. When we incorporate the spatial proximity of genetic markers into our analysis we can observe an improvement in predictive accuracy, especially when compared to simpler methods like genomic BLUP.

6.2 Limitations

Some limitations for the studies in this thesis should be considered, highlighting the complexity of this research. First, data quality and preprocessing is crucial for ensuring that network reconstruction methods are accurate and reliable. In Chapter 3, we utilized information from food frequency questionnaires (FFQ) for examining dietary patterns. While FFQ can provide valuable information on eating habits, they contain self-reported data which are often less reliable due to, for example, issues with bias, memory recollection, and social desirability. While, in our work data preprocessing steps were carefully selected by consulting biologists, it is worth acknowledging that, in many contexts, data processing lacks standardization or guidelines.

A second limitation in network analysis is that network reconstruction methods are fundamentally exploratory tools, geared towards hypothesis generation. The methods reveal underlying correlations within metabolite networks, but their findings should be interpreted with caution due to the inherent variability and exploratory nature of the techniques. Therefore, consulting specialists such as biologists and epidemiologists becomes imperative as their insights can help support the validity of the study's findings.

A third limitation of the present work is that although many different study setups/designs have been considered, we did not investigate network reconstruction methods while dealing with variation due to population or family structures. In such studies, samples often display between-sample correlations primarily due to shared genetic background. Edges in metabolic networks are therefore susceptible to the between-samples variation induced, hence data-driven sample-decorrelation approaches prior to network reconstruction are advisable.

Finally, during this work, we did not develop methods for hybrid study designs. For example, we separately presented methods conditioning on a single study design: (i) measurements over time and (ii) incorporating other omics sources. Separately studying metabolite networks with regard to single sources of variation offers valuable insights into how specific factors influence these networks, but a combination of those can also be interesting as will be discussed below.

6.3 Further research

The framework for extracting metabolite networks as developed in this dissertation provides considerable flexibility. However we also see, stemming from the mentioned limitations, opportunities for further research. We will formulate these opportunities below.

6.3.1 Longitudinal designs

In metabolomics research, capturing the temporal dynamics of metabolite networks is critical. Our investigations primarily utilized static data, which, while informative, offer only a snapshot of metabolic interactions. In Chapter 3 we considered repeated measures. These methods are also of interest in more intensive longitudinal designs, where potentially many time-points are collected. This allows for the observation of metabolic changes over time, potentially revealing transient associations and more accurately reflecting biological processes. Such designs allow tracking metabolite trajectories, offering valuable insights into the sequential and network-level alterations (Buchweitz et al., 2020) that occur in response to various reactions. This approach can potentially elucidate patterns of metabolic regulation and dysregulation that would otherwise remain hidden in cross-sectional studies (Bordbar et al., 2017).

6.3.2 Correlated samples

Metabolic networks are inherently sensitive to various sources of biological and technical variation, including between-sample dependencies arising from shared genetics and environmental exposures. Neglecting the dependence between samples can lead to confounded metabolite associations, particularly when dealing with population and familial structures. A simple way to mitigate this issue is by using a linear whitening transformation, which involves pre-multiplying the metabolite data by the inverse square root of the SNPs variance/covariance matrix across subjects (Kessy, Lewin, & Strimmer, 2018). Alternatively, samples can be decorrelated by regressing out the principal components on the sample space of the SNP data (Lin & Zeng, 2011; L. Liu, Zhang, Liu, & Arendt, 2013; Hellwege et al., 2017) prior to network reconstruction. Finally, the between-sample dependencies can be accounted for when using a linear mixed model (Zhou & Stephens, 2012; Hoffman, 2013; Onifade, Roy-Gagnon, Parent, & Burkett, 2022) with a prespecified variance structure for the random effects. These approaches can help ensure that the recovered metabolite associations are reflective of true biological connections rather than confounded by underlying sample relationships.

6.3.3 Hybrid designs

The developed methodology isolates a single source of variation for network reconstruction. However, biological systems are influenced by a multitude of factors, often occurring simultaneously. Realistically, one would like to integrate longitudinal metabolic data with SNP or transcriptomic information (Suhre et al., 2016). In network analysis, it would not only reflect the dynamics of metabolic networks over time, but also their genetic or transcriptomic underpinnings. By developing methods that can accommodate these hybrid sources of variation, we could construct metabolite networks that more accurately mirror complex biological interactions. Enriching and reshaping the proposed methods can offer opportunities for more precise biomarker identification and deeper metabolic pathway understanding, ultimately leading to more holistic system understanding (Bodein, Scott-Boyer, Perin, Lê Cao, & Droit, 2022).

6.4 Impact

Notwithstanding the limitations and inroads for further research, the current methodology provides a basis for practical impact in metabolomic research. Essentially, this work can assist in real-world biological applications by effectively addressing known interpretative challenges in the field.

6.4.1 Tracking metabolic similarity

A key characteristic of the present work is that it allows us to delve into the intricate connections between metabolites and discern how these associations vary across different biological sources (Inouye et al., 2010; Menni et al., 2015). By reconstructing separate networks for distinct sources of variation like study variables, SNPs or gene information, we can pinpoint the origins of metabolic similarities. This approach allows us to understand whether a pair of metabolites shares a connection due to specific genetic factors, environmental influences, or other aspects of metabolic variation, thus offering a clearer understanding of their interrelations. Additionally, this method illuminates the dynamic nature of these connections, showing how they may change or persist under different biological conditions, thereby providing deeper insights into the multifaceted nature of metabolic networks.

6.4.2 Mapping unknown metabolites

A fundamental feature of our methodology is its potential ability to elucidate the properties of unknown metabolites whose chemical identities are not fully understood. Metabolite identification aided by statistical methodology (Aittokallio & Schwikowski, 2006; Barupal et al., 2012; Heinzmann, Waldenberger, Peters, & Schmitt-Kopplin, 2022) is crucial, especially considering the role of unknown metabolites as potential biomarkers for various diseases (Krumsiek et al., 2012; Pirhaji et al., 2016). By demonstrating, in the preceding chapters, that our approach effectively recovers biologically relevant edges, we have established a foundation for assuming that including unknown metabolites in our analysis would similarly yield insightful and relevant results. Using this framework we are able to tentatively deduce the properties of any unknown metabolite by estimating networks for different facets of metabolic variation and analyzing its connections with metabolites within identified modules, in each network. By revealing their associations and characteristics in relation to specific sources of variation, our framework offers a useful tool for advancing our understanding of unknown yet potentially important molecules in metabolic research.

6.4.3 Guiding research

The exploratory nature of network analysis extends beyond simple visualizations (Aittokallio & Schwikowski, 2006), as they can also serve as tools that provide valuable insights and guide more focused, confirmatory research. By concentrating on specific parts of metabolic variation, we effectively separate signal from noise, thus creating networks with reduced uncertainty and increased relevance. This approach enables researchers and practitioners to not only observe but also hypothesize about biological processes, especially in areas where our understanding is still evolving. Targeted experimental research can then be directed towards specific metabolites or groups of metabolites (Bi et al., 2022) that emerge as potentially important in these refined networks. The exploratory networks produced are thus not just endpoints of analysis: they serve as foundation for targeted investigations, leading to deeper insights and potential discoveries (Amara et al., 2022) in metabolic functions and their broader implications.

6.5 Concluding remarks

As we conclude this thesis, it is important to reflect on the journey of developing a metabolite network estimation framework and its potential across various biological domains. The general principles established here provide a robust foundation for future research, indicating their broad applicability and potential scalability well beyond the initial datasets and biological contexts.

We now delve into topics that indicate how these methodologies are not confined to the specific examples presented but can be adapted to other organisms and study designs. By discussing the potential for generalizability and transferability to aid our understanding of complex biological systems and facilitate the exploration of systems biology and precision medicine, we set the stage for future research directions on the analysis of complex biological systems.

6.5.1 Generalizability

In this thesis, we primarily used linear models prior to metabolite network reconstruction as our data did not display strong indications of non-linear relationships. However, it is crucial to recognize that not all biological data will conform to linear dynamics (Kuznetsov, Knott, & Bonner, 2002; Ritchie, Holzinger, Li, Pendergrass, & Kim, 2015; Yan, Risacher, Shen, & Saykin, 2018). When non-linear effects are evident, we could relax the linearity assumption in various ways. In network reconstruction, pairwise associations can be estimated by employing similarity measures like mutual information (Margolin et al., 2006) or Spearman's rank coefficient (Camacho, De La Fuente, & Mendes, 2005; Bartel et al., 2013), thus enhancing the framework's adaptability. Additionally, the use of copulas (H. Liu, Lafferty, & Wasserman, 2009; H. Liu, Han, Yuan, Lafferty, & Wasserman, 2012) provides a robust framework for capturing complex dependencies that are not adequately described by traditional correlation measures. By separating the modeling of marginal distributions from the dependency structure allows for more precise modeling of metabolic interactions under non-normal conditions. This approach allows a broader coverage of associations while remaining applicable across various biological contexts.

Additionally, our framework, while focused on continuous outcomes and categorical covariables, is versatile enough to handle other data types. For categorical response variables, such as binary disease status, logistic regression can be aptly utilized prior to network reconstruction (L. Li & Liu, 2022).

We also showcased the application of our methods using high-dimensional omics sources to predict continuous outcomes, explicitly modeling the data's inherent network structure to improve prediction accuracy. Although our examples predominantly involved omics data, the principles and methods we developed are applicable to any dataset involving non-independent predictors with inherent network structures (Noorie & Afsari, 2020). This underscores the potential for generalization of our framework beyond the specific cases discussed in this thesis.

6.5.2 Transferability

This thesis has developed a network reconstruction framework from plant and human metabolomic data, demonstrating its efficacy and broad utility. The methodologies can also be applied to the study of other organisms, such as animals.

The study designs (experimental for plants and observational for humans) present distinct preprocessing challenges, from log transformation in plant metabolomics to using raw concentrations in humans, and vary in the level of control over experimental variables. Experimental plant designs allow precise control over environmental conditions, reducing data noise, whereas observational studies often have greater variability due to less control over subjects (Freedman, 2006). Nonetheless, the developed framework is robust and flexible enough for other designs like cohort, case-control, and treatment designs in human studies, or association panels in plant research. Despite their differences in noise levels and preprocessing needs, our framework adapts to these variations, demonstrating versatility across different data characteristics.

Moreover, the principles derived are applicable beyond metabolomics, extending to genomics, proteomics, and transcriptomics, which may face similar preprocessing challenges (Cavill et al., 2009).

In essence, the methodologies developed here offer extensive flexibility, making them applicable across various biological data types and study designs.

References

Agamah, F. E., Bayjanov, J. R., Niehues, A., Njoku, K. F., Skelton, M., Mazandu, G. K., … t Hoen, P. A. (2022). Computational approaches for network-based integrative multi-omics analysis. *Frontiers in Molecular Biosciences*, *9*, 1214.

Aittokallio, T., & Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Briefings in bioinformatics*, *7*(3), 243–255.

Amara, A., Frainay, C., Jourdan, F., Naake, T., Neumann, S., Novoa-Del-Toro, E. M., … Witting, M. (2022). Networks and graphs discovery in metabolomics data analysis and interpretation. *Frontiers in Molecular Biosciences*, 223.

Banerjee, O., Ghaoui, L. E., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, *9*(Mar), 485–516.

Barabási, A.-L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature reviews genetics*, *12*(1), 56–68.

Barchet, G. (2013). A brief overview of metabolomics: What it means, how it is measured, and its utilization. *The Science Creative Quarterly*, *8*.

Barnes, J. A., & Harary, F. (1983). Graph theory in network analysis. *Social networks*, *5*(2), 235–244.

Bartel, J., Krumsiek, J., & Theis, F. J. (2013). Statistical methods for the analysis of high-throughput metabolomics data. *Computational and structural biotechnology journal*, *4*(5), e201301009.

Bartzis, G., Deelen, J., Maia, J., Ligterink, W., Hilhorst, H. W., Houwing-Duistermaat, J.-J., … Uh, H.-W. (2017). Estimation of metabolite networks with regard to a specific covariable: applications to plant and human data. *Metabolomics*, *13*(11), 129.

Bartzis, G., Peeters, C. F. W., & Eeuwijk, F. v. (2022). psblup: incorporating marker proximity for improving genomic prediction accuracy. *Euphytica*, *218*(5), 54.

Bartzis, G., Peeters, C. F. W., Ligterink, W., & Van Eeuwijk, F. (2024). A guided network estimation approach using multi-omic information. *BMC* *bioinformatics*, *25*(1), 202.

Bartzis, G., Peeters, C. F. W., Uh, H. W., Houwing Duistermaat, J. J., & Eeuwijk, F. v. (2023). Estimating metabolite networks subject to dietary preferences and lifestyle. *Under submission.*

Barupal, D. K., Haldiya, P. K., Wohlgemuth, G., Kind, T., Kothari, S. L., Pinkerton, K. E., & Fiehn, O. (2012). Metamapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. *BMC bioinformatics*, *13*, 1–15.

Beisken, S., Eiden, M., & Salek, R. M. (2015). Getting the right answers: understanding metabolomics challenges. *Expert review of molecular diagnostics*, *15*(1), 97–109.

Bernardo, R. (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Science*, *34*(1), 20–25.

Bernardo, R. (1996). Best linear unbiased prediction of maize single-cross performance. *Crop Science*, *36*(1), 50–56.

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop science*, *48*(5), 1649–1664.

Bi, X., Liu, Y., Li, J., Du, G., Lv, X., & Liu, L. (2022). Construction of multiscale genome-scale metabolic models: frameworks and challenges. *Biomolecules*, *12*(5), 721.

Bilello, J. A. (2005). The agony and ecstasy of "omic" technologies in drug development. *Current molecular medicine*, *5*(1), 39–52.

Bodein, A., Scott-Boyer, M.-P., Perin, O., Lê Cao, K.-A., & Droit, A. (2022). Interpretation of network-based integration from multi-omics longitudinal data. *Nucleic acids research*, *50*(5), e27–e27.

Bordbar, A., Yurkovich, J. T., Paglia, G., Rolfsson, O., Sigurjónsson, Ó. E., & Palsson, B. O. (2017). Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics. *Scientific reports*, *7*(1), 46249.

Bory, C., Boulieu, R., Chantin, C., & Mathieu, M. (1990). Diagnosis of alcaptonuria: rapid analysis of homogentisic acid by hplc. *Clinica Chimica Acta*, *189*(1), 7–11.

Bromberger, J. T., Matthews, K. A., Kuller, L. H., Wing, R. R., Meilahn, E. N., & Plantinga, P. (1997). Prospective study of the determinants of age at menopause. *American journal of epidemiology*, *145*(2), 124–133.

Buchweitz, L. F., Yurkovich, J. T., Blessing, C., Kohler, V., Schwarzkopf, F., King, Z. A., ... others (2020). Visualizing metabolic network dynamics through time-series metabolomic data. *BMC bioinformatics*, *21*(1), 1–10.

Camacho, D., De La Fuente, A., & Mendes, P. (2005). The origin of correlations in metabolomics data. *Metabolomics*, *1*, 53–63.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, *1*(2), 245–276.

Cavill, R., Keun, H. C., Holmes, E., Lindon, J. C., Nicholson, J. K., & Ebbels, T. M. (2009). Genetic algorithms for simultaneous variable and sample selection in metabonomics. *Bioinformatics*, *25*(1), 112–118.

Chan, D. C., Barrett, H. P., & Watts, G. F. (2004). Dyslipidemia in visceral obesity. *American Journal of Cardiovascular Drugs*, *4*(4), 227–246.

Chatterjee, N., Shi, J., & García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, *17*(7), 392–406.

Chen, C., Stock, C., Hoffmeister, M., & Brenner, H. (2019). Optimal age for screening colonoscopy: a modeling study. *Gastrointestinal endoscopy*, *89*(5), 1017–1025.

Chung, F. R., & Graham, F. C. (1997). *Spectral graph theory* (No. 92). American Mathematical Society.

Clark, C., Rabl, M., Dayon, L., & Popp, J. (2022). The promise of multi-omics approaches to discover biological alterations with clinical relevance in alzheimer's disease. *Frontiers in Aging Neuroscience*, *14*, 1065904.

Clark, S. A., & van der Werf, J. (2013). Genomic best linear unbiased prediction (gblup) for the estimation of genomic breeding values. In *Genome-wide association studies and genomic prediction* (pp. 321–330). Springer.

Crossa, J., de Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., ... others (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*.

de Los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, *193*(2), 327–345.

de Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., ... Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigrees. *Genetics*.

Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass spectrometry reviews*, *26*(1), 51–78.

de Vlaming, R., & Groenen, P. J. (2015). The current and future use of ridge regression for prediction in quantitative genetics. *BioMed research international*, *2015*.

DiLeo, M. V., Strahan, G. D., den Bakker, M., & Hoekenga, O. A. (2011). Weighted correlation network analysis (wgcna) applied to the tomato fruit metabolome. *PLoS One*, *6*(10), e26683.

Dong, J., & Horvath, S. (2007). Understanding network concepts in modules. *BMC systems biology*, *1*(1), 24.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genet*, *9*(3), e1003348.

Efron, B. (2012). Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with r package rrBLUP. *The Plant Genome*, *4*(3), 250–255.

Euesden, J., Lewis, C. M., & O'reilly, P. F. (2015). Prsice: polygenic risk score software. *Bioinformatics*, *31*(9), 1466–1468.

Fabres, P. J., Collins, C., Cavagnaro, T. R., & Rodríguez López, C. M. (2017). A concise review on multi-omics data integration for terroir analysis in vitis vinifera. *Frontiers in plant science*, *8*, 1065.

Fiehn, O. (2002). Metabolomics—the link between genotypes and phenotypes. *Functional genomics*, 155–171.

Floegel, A., Wientzek, A., Bachlechner, U., Jacobs, S., Drogan, D., Prehn, C., ... others (2014). Linking diet, physical activity, cardiorespiratory fitness and obesity to serum metabolite networks: findings from a population-based study. *International Journal of Obesity*, *38*(11), 1388–1396.

Freedman, D. A. (2006). Statistical models for causation: what inferential leverage do they provide? *Evaluation review*, *30*(6), 691–713.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics Springer, Berlin.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Gianola, D., Perez-Enciso, M., & Toro, M. A. (2003). On marker-assisted prediction of genetic value: beyond the ridge. *Genetics*, *163*(1), 347–365.

Goddard, M. E., Hayes, B. J., & Meuwissen, T. H. (2010). Genomic selection in livestock populations. *Genetics research*, *92*(5-6), 413–421.

Goff, D. C., D'Agostino, R. B., Haffner, S. M., & Otvos, J. D. (2005). Insulin resistance and adiposity influence lipoprotein size and subclass concentrations. results from the insulin resistance atherosclerosis study. *Metabolism*, *54*(2), 264–270.

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

Grundy, S. M. (2004). Obesity, metabolic syndrome, and cardiovascular disease. *The Journal of Clinical Endocrinology & Metabolism*, *89*(6), 2595–2600.

Guertin, K. A., Moore, S. C., Sampson, J. N., Huang, W.-Y., Xiao, Q., Stolzenberg-Solomon, R. Z., ... Cross, A. J. (2014). Metabolomics in nutritional epidemiology: identifying metabolites associated with diet and quantifying their potential to uncover diet-disease relations in populations. *The American journal of clinical nutrition*, ajcn–078758.

Ha, M., & Sun, W. (2014). Partial correlation matrix estimation using ridge penalty followed by thresholding and re-estimation. *Biometrics*, *70*(3), 762–770.

Habier, D., Fernando, R., & Dekkers, J. C. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, *177*(4), 2389–2397.

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC bioinformatics*, *12*(1), 186.

Hartl, D. (2011). *Essential Genetics: A Genomics Perspective* (5th ed.). Sudbury, MA: Jones and Bartlett.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Unsupervised learning*. Springer.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). New York: Springer.

Hayes, B. J., Bowman, P. J., Chamberlain, A., & Goddard, M. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science*, *92*(2), 433–443.

Hayes, P., Liu, B., Knapp, S., Chen, F., Jones, B., Blake, T., … others (1993). Quantitative trait locus effects and environmental interaction in a sample of north american barley germ plasm. *Theoretical and Applied Genetics*, *87*(3), 392–401.

Heffner, E. L., Sorrells, M. E., & Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Science*, *49*(1), 1–12.

Heinzmann, S. S., Waldenberger, M., Peters, A., & Schmitt-Kopplin, P. (2022). Cluster analysis statistical spectroscopy for the identification of metabolites in 1h nmr metabolomics. *Metabolites*, *12*(10), 992.

Hellwege, J. N., Keaton, J. M., Giri, A., Gao, X., Velez Edwards, D. R., & Edwards, T. L. (2017). Population stratification in genetic association studies. *Current protocols in human genetics*, *95*(1), 1–22.

Heslot, N., Yang, H.-P., Sorrells, M. E., & Jannink, J.-L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Science*, *52*(1), 146–160.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hoffman, G. E. (2013). Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PloS one*, *8*(10), e75707.

Horvath, S., & Dong, J. (2008). Geometric interpretation of gene coexpression network analysis. *PLoS Comput Biol*.

Hu, F. B. (2002). Dietary pattern analysis: a new direction in nutritional epidemiology. *Current opinion in lipidology*, *13*(1), 3–9.

Hu, F. B., Rimm, E., Smith-Warner, S. A., Feskanich, D., Stampfer, M. J., Ascherio, A., … Willett, W. C. (1999). Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire. *The American journal of clinical nutrition*, *69*(2), 243–249.

Hu, F. B., Rimm, E. B., Stampfer, M. J., Ascherio, A., Spiegelman, D., & Willett, W. C. (2000). Prospective study of major dietary patterns and risk of coronary heart disease in men. *The American journal of clinical nutrition*, *72*(4), 912–921.

Hunt, C. H., van Eeuwijk, F. A., Mace, E. S., Hayes, B. J., & Jordan, D. R. (2018). Development of genomic prediction in sorghum. *Crop Science*.

Inouye, M., Kettunen, J., Soininen, P., Silander, K., Ripatti, S., Kumpula, L. S., … others (2010). Metabonomic, transcriptomic, and genomic variation of a population cohort. *Molecular systems biology*, *6*(1), 441.

Jannink, J.-L., Lorenz, A. J., & Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics*, *9*(2), 166–177.

Jendoubi, T., & Strimmer, K. (2019). A whitening approach to probabilistic canonical correlation analysis for omics data integration. *BMC bioinformatics*, *20*(1), 1–13.

Joosen, R. V. L. (2013). *Imaging genetics of seed performance.* Wageningen: Wageningen University and Research.

Joosen, R. V. L., Arends, D., Li, Y., Willems, L. A., Keurentjes, J. J., Ligterink, W., … Hilhorst, H. W. (2013). Identifying genotype-by-environment interactions in the metabolism of germinating arabidopsis seeds using generalized genetical genomics. *Plant physiology*, *162*(2), 553–566.

Joyce, A. R., & Palsson, B. Ø. (2006). The model organism as a system: integrating'omics' data sets. *Nature reviews Molecular cell biology*, *7*(3), 198–210.

Kessy, A., Lewin, A., & Strimmer, K. (2018). Optimal whitening and decorrelation. *The American Statistician*, *72*(4), 309–314.

Kettunen, J., Tukiainen, T., Sarin, A.-P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L.-P., … others (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature genetics*, *44*(3), 269–276.

Keurentjes, J. J. (2009). Genetical metabolomics: closing in on phenotypes. *Current opinion in plant biology*, *12*(2), 223–230.

Kitano, H. (2002). Systems biology: a brief overview. *Science*, *295*(5560), 1662–1664.

Klop, B., Elte, J. W. F., & Cabezas, M. C. (2013). Dyslipidemia in obesity: mechanisms and potential targets. *Nutrients*, *5*(4), 1218–1240.

Kolaczyk, E. D., & Krivitsky, P. N. (2015). On the question of effective sample size in network modeling: An asymptotic inquiry. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *30*(2), 184.

Krumsiek, J., Suhre, K., Evans, A. M., Mitchell, M. W., Mohney, R. P., Milburn, M. V., … others (2012). Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. *PLoS genetics*, *8*(10), e1003005.

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., & Theis, F. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology*, *5*(1), 21.

Kuznetsov, V., Knott, G., & Bonner, R. (2002). General statistics of stochastic process of gene expression in eukaryotic cells. *Genetics*, *161*(3), 1321–1332.

Langfelder, P., & Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, *9*(1), 559.

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, *24*(5), 719–720.

Lauritzen, S. L. (1996). *Graphical models.* Clarendon Press.

Lê Cao, K.-A., González, I., & Déjean, S. (2009). integromics: an r package to unravel relationships between two omics datasets. *Bioinformatics*, *25*(21), 2855–2856.

Lê Cao, K.-A., & Le Gall, C. (2011). Integration and variable selection of 'omics' data sets with pls: a survey. *Journal de la Société Française de Statistique*, *152*(2), 77–96.

Li, C., & Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, *24*(9), 1175–1182.

Li, L., & Liu, Z.-P. (2022). A connected network-regularized logistic regression model for feature selection. *Applied Intelligence*, *52*(10), 11672–11702.

Li, W., Zhang, S., Liu, C.-C., & Zhou, X. J. (2012). Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, *28*(19), 2458–2466.

Lightbody, G., Haberland, V., Browne, F., Taggart, L., Zheng, H., Parkes, E., & Blayney, J. K. (2018). Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. *Briefings in bioinformatics*, *1*, 17.

Lin, D., & Zeng, D. (2011). Correcting for population stratification in genomewide association studies. *Journal of the American Statistical Association*, *106*(495), 997–1008.

Liu, H., Han, F., Yuan, M., Lafferty, J., & Wasserman, L. (2012). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, *40*(4), 2293–2326.

Liu, H., Lafferty, J., & Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, *10*(10), 2295–2328.

Liu, H., Roeder, K., & Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in neural information processing systems* (pp. 1432–1440).

Liu, L., Zhang, D., Liu, H., & Arendt, C. (2013). Robust methods for population stratification in genome wide association studies. *BMC bioinformatics*, *14*(1), 1–12.

Magkos, F., & Mittendorfer, B. (2009). Gender differences in lipid metabolism and the effect of obesity. *Obstetrics and gynecology clinics of North America*, *36*(2), 245–265.

Magkos, F., Mohammed, B. S., & Mittendorfer, B. (2008). Effect of obesity on the plasma lipoprotein subclass profile in normoglycemic and normolipidemic men and women. *International journal of obesity*, *32*(11), 1655–1664.

Maia, J., Dekkers, B. J., Dolle, M. J., Ligterink, W., & Hilhorst, H. W. (2014). Abscisic acid (aba) sensitivity regulates desiccation tolerance in germinated arabidopsis seeds. *New phytologist*, *203*(1), 81–93.

Maia, J., Dekkers, B. J., Provart, N. J., Ligterink, W., & Hilhorst, H. W. (2011). The re-establishment of desiccation tolerance in germinated arabidopsis thaliana seeds and its associated transcriptome. *PloS one*, *6*(12), e29123.

Malosetti, M., Voltas, J., Romagosa, I., Ullrich, S., & Van Eeuwijk, F. (2004). Mixed models including environmental covariables for studying QTL by environment interaction. *Euphytica*, *137*(1), 139–145.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., & Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *Bmc bioinformatics* (Vol. 7, pp. 1–15).

Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, *11*, 2287–2322.

Medina-Cleghorn, D., & Nomura, D. K. (2013). Chemical approaches to study metabolic networks. *Pflügers Archiv-European Journal of Physiology*, *465*(3), 427–440.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 1436–1462.

Menni, C., Graham, D., Kastenmüller, G., Alharbi, N. H., Alsanosi, S. M., McBride, M., ... others (2015). Metabolomic identification of a novel pathway of blood pressure regulation involving hexadecanedioate. *Hypertension*, *66*(2), 422–429.

Meuwissen, T., Hayes, B., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

Mittendorfer, B., Patterson, B. W., & Klein, S. (2003). Effect of sex and obesity on basal vldl-triacylglycerol kinetics. *The American Journal of Clinical Nutrition*, *77*(3), 573–579.

Morgenthal, K., Weckwerth, W., & Steuer, R. (2006). Metabolomic networks in plants: Transitions from pattern recognition to biological interpretation. *Biosystems*, *83*(2), 108–117.

Newby, P., & Tucker, K. L. (2004). Empirically derived eating patterns using factor or cluster analysis: a review. *Nutrition reviews*, *62*(5), 177–203.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), 026113.

Nielsen, J. (2003). It is all about metabolic fluxes. *Journal of Bacteriology*, *185*(24), 7031–7035.

Nielsen, J., & Jewett, M. C. (2007). The role of metabolomics in systems biology. In *Metabolomics* (pp. 1–10). Springer.

Noorie, Z., & Afsari, F. (2020). Sparse feature selection: relevance, redundancy and locality structure preserving guided by pairwise constraints. *Applied Soft Computing*, *87*, 105956.

Núñez-Antón, V. A., & Zimmerman, D. L. (2009). *Antedependence models for longitudinal data.* Chapman and Hall/CRC.

Okazaki, Y., & Saito, K. (2012). Recent advances of metabolomics in plant biotechnology. *Plant biotechnology reports*, *6*(1), 1–15.

Onifade, M., Roy-Gagnon, M.-H., Parent, M.-É., & Burkett, K. M. (2022). Comparison of mixed model based approaches for correcting for population substructure with application to extreme phenotype sampling. *BMC genomics*, *23*(1), 98.

Ooi, E. M., Watts, G. F., Farvid, M. S., Chan, D. C., Allen, M. C., Zilko, S. R., & Barrett, P. H. R. (2005). High-density lipoprotein apolipoprotein a-i kinetics in obesity. *Obesity research*, *13*(6), 1008–1016.

Pallister, T., Sharafi, M., Lachance, G., Pirastu, N., Mohney, R. P., MacGregor, A., ... Menni, C. (2015). Food preference patterns in a uk twin cohort. *Twin Research and Human Genetics*, *18*(06), 793–805.

Peeters, C. F. W., Bilgrau, A. E., & van Wieringen, W. N. (2022). rags2ridges: A one-stop-l2-shop for graphical modeling of high-dimensional precision matrices. *Journal of Statistical Software*, *102*, 1–32.

Perez De Souza, L., Alseekh, S., Brotman, Y., & Fernie, A. R. (2020). Network-based strategies in metabolomics data analysis and interpretation: From molecular networking to biological interpretation. *Expert Review of Proteomics*, *17*(4), 243–255.

Piepho, H., Ogutu, J., Schulz-Streeck, T., Estaghvirou, B., Gordillo, A., & Technow, F. (2012). Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. *Crop science*, *52*(3), 1093–1104.

Pirhaji, L., Milani, P., Leidl, M., Curran, T., Avila-Pacheco, J., Clish, C. B., ... Fraenkel, E. (2016). Revealing disease-associated pathways by network integration of untargeted metabolomics. *Nature methods*, *13*(9), 770–776.

Raffler, J., Friedrich, N., Arnold, M., Kacprowski, T., Rueedi, R., Altmaier, E., ... others (2015). Genome-wide association study with targeted and non-targeted nmr metabolomics identifies 15 novel loci of urinary human metabolic individuality. *PLoS Genet*, *11*(9), e1005487.

Raja, K., Patrick, M., Gao, Y., Madu, D., Yang, Y., & Tsoi, L. C. (2017). A review of recent advancement in integrating omics data with literature mining towards biomedical discoveries. *International journal of genomics*, *2017*.

Randall, E., Marshall, J. R., Brasure, J., & Graham, S. (1992). Dietary patterns and colon cancer in western new york. *Nutrition and Cancer*, *18*(3), 265—276.

Rencher, A. C. (2003). *Methods of multivariate analysis*. John Wiley & Sons.

Rinschen, M. M., Ivanisevic, J., Giera, M., & Siuzdak, G. (2019). Identification of bioactive metabolites using activity metabolomics. *Nature reviews Molecular cell biology*, *20*(6), 353–367.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, *16*(2), 85–97.

Rothman, A. J., Bickel, P. J., Levina, E., Zhu, J., et al. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, *2*, 494–515.

Sam, S., Haffner, S., Davidson, M. H., D'Agostino, R. B., Feinstein, S., Kondos, G., ... Mazzone, T. (2008). Relationship of abdominal visceral and subcutaneous adipose tissue with lipoprotein particle number and size in type 2 diabetes. *Diabetes*, *57*(8), 2022–2027.

Schmidt, J. A., Rinaldi, S., Ferrari, P., Carayol, M., Achaintre, D., Scalbert, A., ... others (2015). Metabolic profiles of male meat eaters, fish eaters, vegetarians, and vegans from the epic-oxford cohort. *The American journal of clinical nutrition*, *102*(6), 1518–1526.

Sengupta, S., Mukherjee, S., Basak, P., & Majumder, A. L. (2015). Significance of galactinol and raffinose family oligosaccharide synthesis in plants. *Frontiers in plant science*, *6*, 656.

Shen, X., Alam, M., Fikse, F., & Rönnegård, L. (2013). A novel generalized ridge regression method for quantitative genetics. *Genetics*, genetics–112.

Slattery, M. L., Boucher, K. M., Caan, B. J., Potter, J. D., & Ma, K.-N. (1998). Eating patterns and risk of colon cancer. *American Journal of epidemiology*, *148*(1), 4–16.

Speed, D., & Balding, D. J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. *Genome research*, gr–169375.

Suhre, K., Schwartz, J. E., Sharma, V. K., Chen, Q., Lee, J. R., Muthukumar, T., ... others (2016). Urine metabolite profiles predictive of human kidney allograft status. *Journal of the American Society of Nephrology*, *27*(2), 626–636.

Sun, Y. V., & Hu, Y.-J. (2016). Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Advances in genetics*, *93*, 147–190.

Tebani, A., Abily-Donval, L., Afonso, C., Marret, S., & Bekri, S. (2016). Clinical metabolomics: The new metabolic window for inborn errors of metabolism investigations in the post-genomic era. *International Journal of Molecular Sciences*, *17*(7), 1167.

Ten Berge, J. M., Krijnen, W. P., Wansbeek, T., & Shapiro, A. (1999). Some new results on correlation-preserving factor scores prediction methods. *Linear Algebra and its Applications*, *289*(1-3), 311–318.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tikunov, Y., Laptenok, S., Hall, R., Bovy, A., & De Vos, R. (2012). Msclust: a tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. *Metabolomics*, *8*(4), 714–718.

Toubiana, D., Fernie, A. R., Nikoloski, Z., & Fait, A. (2013). Network analysis: tackling complex data to study plant metabolism. *Trends in biotechnology*, *31*(1), 29–36.

Ursem, R., Tikunov, Y., Bovy, A., Van Berloo, R., & Van Eeuwijk, F. (2008). A correlation network approach to metabolic data analysis for tomato fruits. *Euphytica*, *161*(1-2), 181–193.

Van Binsbergen, R., Calus, M. P., Bink, M. C., Eeuwijk, F. A., Schrooten, C., & Veerkamp, R. F. (2015). Genomic prediction using imputed whole-genome sequence data in holstein friesian cattle. *Genetics Selection Evolution*, *47*(1), 71.

van Eeuwijk, F. A., Bink, M. C., Chenu, K., & Chapman, S. C. (2010). Detection and use of qtl for complex traits in multiple environments. *Current opinion in plant biology*, *13*(2), 193–205.

VanLiere, J. M., & Rosenberg, N. A. (2008). Mathematical properties of the r^2 measure of linkage disequilibrium. *Theoretical population biology*, *74*(1), 130–137.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science*, *91*(11), 4414–4423.

Wang, D. D., Zheng, Y., Toledo, E., Razquin, C., Ruiz-Canela, M., Guasch-Ferré, M., ... Hu, F. B. (2018, nov). Lipid metabolic networks, mediterranean diet and cardiovascular disease in the PREDIMED trial. *International Journal of Epidemiology*, *47*. doi: 10.1093/ije/dyy198

Wang, H., Paulo, J., Kruijer, W., Boer, M., Jansen, H., Tikunov, Y., ... Van Eeuwijk, F. (2015). Genotype–phenotype modeling considering intermediate level of biological variation: a case study involving sensory traits, metabolites and qtls in ripe tomatoes. *Molecular BioSystems*, *11*(11), 3101–3110.

Warrens, M. (2008). On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, *73*, 777–789.

Watson, E., MacNeil, L. T., Arda, H. E., Zhu, L. J., & Walhout, A. J. (2013, mar). Integration of metabolic and gene regulatory networks modulates the c. elegans dietary response. *Cell*, *153*(1), 253–266. doi: 10.1016/j.cell.2013.02.050

Weber, M., Striaukas, J., Schumacher, M., & Binder, H. (2023). Regularized regression when covariates are linked on a network: the 3cose algorithm. *Journal of Applied Statistics*, *50*(3), 535–554.

Weismayer, C., Anderson, J. G., & Wolk, A. (2006). Changes in the stability of dietary patterns in a study of middle-aged swedish women. *The Journal of nutrition*, *136*(6), 1582–1587.

Weng, Y. J., Gan, H. Y., Li, X., Huang, Y., Li, Z. C., Deng, H. M., ... Zhi, F. C. (2019, aug). Correlation of diet, microbiota and metabolite networks in inflammatory bowel disease. *Journal of Digestive Diseases*(9), 447–459. doi: 10.1111/1751-2980.12795

Whittaker, J. C., Thompson, R., & Denham, M. C. (2000). Marker-assisted selection using ridge regression. *Genetics Research*, *75*(2), 249–252.

Williams, D. E., Prevost, A. T., Whichelow, M. J., Cox, B. D., Day, N. E., & Wareham, N. J. (2000). A cross-sectional study of dietary patterns with glucose intolerance and other features of the metabolic syndrome. *British Journal of Nutrition*, *83*(03), 257–266.

Xu, J., Yang, S., Cai, S., Dong, J., Li, X., & Chen, Z. (2010). Identification of biochemical changes in lactovegetarian urine using 1h nmr spectroscopy and pattern recognition. *Analytical and bioanalytical chemistry*, *396*(4), 1451–1463.

Yan, J., Risacher, S. L., Shen, L., & Saykin, A. J. (2018). Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in bioinformatics*, *19*(6), 1370–1381.

Yang, W., & Tempelman, R. J. (2012). A bayesian antedependence model for whole genome prediction. *Genetics*, *190*(4), 1491–1501.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.

Zaykin, D. V., Pudovkin, A., & Weir, B. S. (2008). Correlation-based inference for linkage disequilibrium with multiple alleles. *Genetics*, *180*(1), 533–545.

Zeng, J., Garrick, D., Dekkers, J., & Fernando, R. (2018). A nested mixture model for genomic prediction using whole-genome SNP genotypes. *PloS one*, *13*(3), e0194683.

Zeng, Z.-B. (1994). A composite interval mapping method for locating multiple qtls. In *Proceedings, 5th world congress on genetics applied to livestock production, university of guelph, guelph, ontario, canada* (Vol. 7).

Zhang, A., Sun, H., Yan, G., Wang, P., Han, Y., & Wang, X. (2014). Metabolomics in diagnosis and biomarker discovery of colorectal cancer. *Cancer letters*, *345*(1), 17–20.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, *4*(1).

Zhang, G., He, P., Tan, H., Budhu, A., Gaedcke, J., Ghadimi, B. M., ... others (2013). Integration of metabolomics and transcriptomics revealed a fatty acid network exerting growth inhibitory effects in human pancreatic cancer. *Clinical cancer research*, *19*(18), 4983–4993.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in r. *The Journal of Machine Learning Research*, *13*(1), 1059–1062.

Zhao, W., Langfelder, P., Fuller, T., Dong, J., Li, A., & Hovarth, S. (2010). Weighted gene coexpression network analysis: state of the art. *Journal of biopharmaceutical statistics*, *20*(2), 281–300.

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, *44*(7), 821–824.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

Summary

This thesis focuses on developing a framework for estimating metabolite networks when sources of noise variation are addressed using the study design. Metabolite data hold a unique position in systems biology as intermediates between DNA variation and clinical phenotypes. Variations in metabolic concentrations propagate via enzymatic reactions, forming correlation patterns highlighting the importance of a joint analysis due to their informative interconnections. Network-based approaches are particularly effective for studying the metabolome, as they are inherently multivariate methods that also provide detailed visualizations aiding in the identification of interesting patterns within the data.

Metabolic network analysis, however, is not straightforward as estimated edges are susceptible to variation induced by various sources of biological and technical conditions. A first complication arises from the inherent noise and variation in the metabolome, hindering efforts to isolate and analyze key variables. A second complication is dealing with nuisance variation while modeling temporal correlations in metabolite concentrations from repeated samples. A third complication is estimating metabolite networks while integrating information from other omics levels, requiring the concurrent modeling of multiple omics sources. In this thesis, a statistical framework has been developed to effectively address different study designs by accounting for various sources of biological and technical variation.

In **Chapter 2**, we developed a network analysis approach for recovering a set of metabolite relationships, by finding metabolites sharing a similar relation to a certain source of variation which might also interact with other covariables. By using a linear regression model, the total metabolic information was decomposed and the part relevant to the covariable of interest was used for network estimation. When using the parts related to the specific covariable of interest, the resulting estimated networks displayed higher interconnectedness, and biologically relevant groups of associated metabolites were identified. This work demonstrated how information on the study design can be incorporated to estimate metabolite networks and identifying interconnected metabolites clustered together with respect to their relation to a covariable of interest.

In **Chapter 3** we extended our network reconstruction approaches by working with metabolite concentrations measured at multiple time points, while also incorporating information on other sources of metabolic variation (diet, genetics). Using network analysis we were interested in identifying metabolites sharing similar relationships to diet and lifestyle. Lifestyle was defined as the unmeasured, subject-specific part of the metabolic variation that was not due to diet, genetics, or time in a random intercepts linear mixed model. The metabolite values relevant to diet and lifestyle, instead of the original values, were used as inputs for network estimation methods. This work demonstrated how correcting for several sources of metabolic variation, allows us to look for residual variation and build networks with meaningful metabolite groups sharing similar association to diet and lifestyle.

In **Chapter 4** we proposed an integrative network reconstruction method in which the network organization for a particular type of omics data is guided by the network structure of a related type of omics data upstream in the omic cascade. The structure of these guiding data can be either already known or estimated from the guiding data themselves. Three steps were proposed. First, a network structure for the guiding data was obtained. Second, responses in the target set were regressed on the full set of predictors in the guiding data with a Lasso penalty to reduce the number of predictors and an L2 penalty on the differences between coefficients for predictors that share edges in the network of the guiding data. Finally, a network was reconstructed on the fitted target responses as functions of the predictors in the guiding data. This way we conditioned the target network on the network of the guiding data. By illustrating our approach on two examples, the method detected groups of metabolites that have a similar genetic or transcriptomic basis.

Chapter 5 expanded upon the integrative network reconstruction method of Chapter 4 where a response is expressed as function of an omics source with a network structure. The developed method was applied in genomic selection, where the estimation of phenotypic traits of interest for plants without phenotype is based on the association between single-nucleotide polymorphisms (SNPs) and phenotypic traits for plants with phenotype. In such cases, the number of SNPs far exceeds the number of samples (high-dimensionality), and, therefore, usage of regularization methods is common. The most common approach to estimate marker-trait associations uses the genomic best linear unbiased predictor (GBLUP) method, where a mixed model is fitted to the data. GBLUP is based on the assumption of independence between predictor variables, while it is to be expected that variables will be associated due to their genetic proximity. We proposed a regularized linear model (namely psBLUP: proximity smoothed BLUP) explicitly modeling the dependence between predictor effects. The analysis of Arabidopsis thaliana data and Barley data showed that psBLUP can improve accuracy compared to standard methods.

Chapter 6 summarized the contributions of the proposed framework utilizing a linear model addressing various sources of biological and technical variation, and discusses the framework's relevance concerning its applicability, generalizability, and transferability.

Acknowledgements

Completing a PhD is a journey filled with challenges, perseverance, and moments of doubt, but it is also one of growth, discovery, and deep gratitude. This dissertation is not just the outcome of years of research but also the result of the support, collaboration, and encouragement of many people along the way. It is with deep gratitude that I take this opportunity to express my appreciation to those who have shaped this journey.

I would like to express my deepest gratitude to my promotor, **Fred A. van Eeuwijk**, whose guidance, dedication, and support have been instrumental in shaping my PhD journey. Your expertise and encouragement provided the clarity and direction that were essential for the progress of my research. Your thoughtful approach allowed me to develop as an independent researcher while always offering insightful feedback and support when needed. You were consistently generous with your time, creating opportunities for discussion and exchange of ideas. I am truly grateful for your support and the impact you have had on my academic and professional development.

I would also like to thank my co-promotor, **Jeanine J. Houwing-Duistermaat**. Our regular discussions and your insightful feedback helped shape the early direction of this work. Your high standards and structured approach challenged me to refine my ideas and improve myself. As our collaboration deepened, I came to appreciate not only your expertise but also kindness and trust in my abilities. You respected my ideas and encouraged my growth, providing me with opportunities to develop both as a researcher and as a professional. One such opportunity was organizing *RMSS2017*, an experience that I greatly valued and learned a lot from. I am grateful for the time and effort you dedicated to me and the project.

Finally, I would like to express my deep gratitude to **Carel F.W. Peeters** for stepping in as my daily supervisor at Wageningen during the final stages of my PhD. At a time when progress had stalled, your willingness to engage, read through all the material, and provide constructive feedback made a world of difference. Your dedication to both the research and my success was evident in every discussion we had. Beyond academic guidance, I am especially grateful for your support on a personal level. You did not only help refine my work but also provided the encouragement and understanding I needed to push through a difficult period. Your patience, insights, and genuine care made the journey to completion far more manageable, and for that, I am truly thankful.

Beyond my supervisors, I am deeply grateful to my collaborators and co-authors whose insights and expertise have enriched my research. I sincerely thank **Martin P. Boer**, **Joris Deelen**, **Henk W.M. Hilhorst**, **Willem T. Kruijer**, **Wilco Ligterink**, **Julio Maia**, and **Hae-Won Uh** for their valuable contributions, discussions, and critical feedback. Working together has been an incredibly rewarding experience, and I truly appreciate the effort and perspectives each of you brought to our collaborations.

A significant part of my PhD journey was shaped by office mates, colleagues, and fellow PhD students from both Biometris and LUMC, who made this experience enjoyable. I am deeply grateful to **Alexia**, **Bart**, **Bert**, **Chaozhi**, **Daniela**, **Dennis**, **Dinie**, **Dominique**, **Eleni**, **Emilie**, **Erik**, **George**, **Hein**, **Henry**, **Irene**, **Ivonne**, **Jakub**, **João**, **Julio**, **Katerina N.**, **Katerina P.**, **Kevin**, **Kristina**, **Manya**, **Mar**, **Maria**, **Markus**, **Marta**, **Mia**, **Nadia**, **Ningning**, **Pariya**, **Pietro**, **Renaud**, **Rosa**, **Roula**, **Said**, **Sabine**, **Sonia**, **Stephan B.**, **Stephan W.**, **Szymon**, **Teddy**, and **Vincent** for their support, discussions, and camaraderie. Through your presence and our shared experiences, the workplace became more enjoyable. To **Angga** (and **Suci**), **Floor**, **Paul**, and **Yolande**, whose constant warm smiles made every day a little brighter. No matter the challenges you were facing, your uplifting energy and friendly presence created an atmosphere of positivity. It was always a pleasure to share some space with you, and I truly appreciate the joy and optimism you brought to our daily lives.

Beyond research and work, some of the most memorable moments of my PhD came from the friendships I built along the way. **Dimitris**, a constant presence in my life, has always been an inspiration with his dedication to his goals and his athletic mindset. Our connection remains strong, and I deeply appreciate the friendship we have built. **Roberta** redefined what true friendship means (offering support, celebrating successes, and standing by me during difficult moments). No matter the distance, our bond remains unshakable (and the teddy-bears will always serve as a reminder). I was fortunate to have **Bader**, **Wenhao**, and **Yutaka**, whose company turned even the busiest days into something to look forward to. From lunch breaks filled with laughter to our game nights, endless teasing, and movie discussions, these moments became traditions that I will always want to return to. I am incredibly lucky to have had you all by my side, turning everyday moments into some of my best memories.

A special thank you goes to my paranymphs, who will stand by my side during the defense, but have also been there for me in many ways throughout this journey. **Carlo**, a legend in every sense of the word, was not just an office mate but one of the most caring and supportive people I had the privilege to know. Regardless of the circumstances, I know that I can always turn to you. Your kindness, generosity, and unwavering presence made every day better, and I truly salute your legendariness! **Maikel**, your meticulous attention to detail, structured approach, and readiness to assist with any technical challenge were invaluable. When you accepted being my paranymph with pleasure, it was a reflection of the kindness you has always shown. I could not be more grateful to have you by my side at this moment.

Beyond my academic journey, I am incredibly fortunate to have a group of lifelong friends who have been a constant presence in my life. **Daniella**, **Dimos**, **Ioannis D.**, **Ioannis K.**, **Ioannis P.**, **Kostas K.**, **Kostas P.**, **Sokratis**, **Thodoras**, and **Vasilis**: I have been with you for more than 25 years, and in that time, you have been more than just friends to me; you have been family. Through every challenge, every problem, and every difficult moment, you have always been there. The way we understand each other, the way we lift each other up, is something truly rare, and I am deeply grateful for it. Growing up together has given us a shared way of thinking, a connection that goes beyond words, and a bond that is meant to last for a lifetime. **Manolis** and I first connected over a party game, where we outclassed and outplayed everyone, despite never having spoken before that moment. That instant trust and understanding turned into a great friendship, one that continued throughout our studies! I have always appreciated your sharp thinking, life view, talent, and I am grateful to have had you as both a study partner and, more importantly, as a friend. **Peli** offered me an immense amount of support, especially during the final stages of my PhD. She was always patient, kind, and selfless. During a time of transition and uncertainty, her presence brought stability and comfort, helping me regain my footing. Her understanding, patience, and golden heart stood out even in the most difficult moments, and for that, I will always be grateful. **Vanessa** and I first met during our MSc, and from the start, we clicked. We are always excited to see each other, and our conversations can last for hours. What started as lighthearted fun quickly turned into a true friendship. We are on the same wavelength, whether it's playful joking or catching up after time apart; we share a connection that goes beyond words. She became a great friend, and for that, I am truly grateful. Finally, **Alexandra**, **Chris**, **Dimitris**, **Giorgos**, **Ioanna**, and **Mary** have a steady and comforting presence in my daily life all these years. Exchanging a few words daily, sharing moments, and staying connected in simple ways always makes things a little brighter. I truly appreciate this constant touch and I thank them for the warmth and positivity it brings. Mrs **Mary T.** and Mrs **Vasiliki P.** thank you for always treating me as family, showing me kindness, care, and love. Your constant support, from asking about my well-being to genuinely caring about my work and life, have meant more to me than words can express.

Not all friendships are built in person. When the world stood still (2019), I found a community of incredible friends and their support meant more to me than they probably realize. **Avi**, **Hanna**, and **Ric**, you were always there for me. Every conversation with you has always felt heartfelt and meaningful, and your appreciation always felt sincere. The support and kindness you have shown will stay with me (an *infinity* of gratitude for three truly amazing friends). **Harry**, you and I clicked instantly, through endless banter and teasing, turning every interaction into something fun. Your humor and energy always made my days better. You have been a brother and I am not sorry for the "the stream is back"! **Jordan** (don't forget 11/5 Backstab), you are an incredible friend! Being next to you is giving me the impostor syndrome! The deep discussions we had, meant a lot to both of us, and I feel lucky to have had you by my side during all this time. **Kai** (you're a plague)! Beyond skill, your positivity and company helped me through some very tough moments, reminding me that even in difficult times, there was always something to look forward to. I'm grateful for the countless moments of fun and motivation you brought into my life. **Maya**, when everything was on the line, you had a way of turning the odds in my favor, even when they seemed against me. Your fighting spirit and resilience taught me the value of focus, patience, and determination when it mattered most. **Paul**, you *shirley* have shown me kindness, encouragement, and genuine care! Though we may not have met in person, the impact of your friendship was just as real as any other. You have helped me in ways you may never fully realize, and for that, I will always be grateful.

As I was completing the final stages of my PhD, I had the opportunity to work alongside **Alex**, **Aliki**, **Andreas B.**, **Andreas E.**, **Antonis**, **Manolis**, **Marco**, **Maria**, **Nikos**, **Stavroula**, **Thanos**, and **Vasilis**, who made this period even more enjoyable. While our work was separate from my research, I have truly appreciated the collaboration, discussions, and professional environment we shared. It has been and continues to be a pleasure working with you all!

I would also like to express my heartfelt gratitude to three teachers who left a lasting impact on me during all these years. **Sotirios B.** has played a pivotal role in my academic journey. As my MSc supervisor, he provided endless guidance, support, and encouragement. From attending my first conference together to co-authoring research papers, our collaboration has been truly enriching. His mentorship, patience, trust, and kindness have shaped both my academic development and personal growth. I will forever be grateful for everything! Mrs **Katerina V.**, my high school literature teacher, and Mrs **Mary K.**, my very first English teacher, both created environments where learning felt engaging and approachable. Their encouragement, patience, and dedication helped me build confidence and curiosity, qualities that have stayed with me over the years. I truly appreciate their support and the positive influence they had during my education.

No journey can be completed without the unwavering support of family. To my parents, **Evangelitsa** and **Ioannis**, thank you for everything. Your love, sacrifices, and constant encouragement have shaped me into the person I am today. You have always been my greatest supporters, standing by my side in every step of this journey, which took me far away from home. To my sisters, **Eleftheria** and **Dimitra-Paraskevi**, who have been my lifelong companions, thank you for your endless support, patience, and belief in me; it has meant the world to me. No words will ever be enough to express my gratitude for everything all of you have given me. **Stelios**, and **Iasonas**, thank you for bringing joy into my life whenever I visited. **Kostas**, **Xenia**, **Alexis**, and of course **Vagia**: thank you for your kindness, support, and for always making me feel that no matter where I am, I have a home to return to. This PhD may have been my journey, but it was never walked alone. I carry the love, and support of my family with me, and for that, I am forever grateful.

Biography

Georgios Bartzis was born on the 11$^{\text{th}}$ of November 1988 in Athens, Greece. In 2006 he completed his secondary studies at General High School of Velo, Corinth.

In 2012, he earned a Bachelor's degree in Statistics and Insurance Science from the University of Piraeus, Piraeus. Afterwards he continued his postgraduate studies at University of Piraeus within the Applied Statistics programme. In 2014, he graduated with an MSc degree and was awarded a scholarship in recognition of his performance. His MSc thesis, supervised by Prof. Dr. Sotirios Bersimis, investigated and compared multivariate control charts for monitoring the dispersion of normally distributed processes within the context of statistical process monitoring. He started his PhD in October 2014 as part of a joint project between the Department of Medical Statistics and Bioinformatics at Leiden University Medical Center and Biometris at Wageningen University and Research. The results of his PhD are described in this thesis. During this time, he received the Best Student Paper Award at the 9th IBS-EMR 2017 and organized the Research MIMOmics Summer School, including the Workshop in Omic Studies (2017) in Cambridge, England.

Currently, he works as a data scientist in a Research and Development department, specializing in statistics and machine learning. His role involves conducting academic-oriented research and developing solution-based applications tailored for FMCG companies.

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 30 ECTS (= 20 weeks of activities)

Review/project proposal (4.5 ECTS)
- Metabolite data analysis based on Weighted Networks

Post-graduate courses (15 ECTS)
- Statistical genetics course, University of Helsinki (2015)
- Genetic Genomic Course, LUMC (2014)
- Splines course, IBS in Nijmegen (2015)
- Mixed and Longitudinal Modelling, statistical science for life and behavioral sciences MSc programme, Leiden University (2015)
- Statistical Genetics, statistical science for life and behavioral sciences MSc programme, Leiden University (2015)

Invited review of journal manuscripts (3 ECTS)
- Metabolomics, Profiling of secondary metabolites (2022)
- Metabolomics, Metabolomic analysis of organic acids (2020)
- Metabolomics, Investigation of Species and Environmental Effects on Metabolome (2018)

Competence, skills and career-oriented activities (2.69 ECTS)
- Writing academic English, Leiden University (2016)
- Reviewing a scientific paper, Leiden University (2017)
- Supervising MSc and BSc students, PE&RC (2018)

Scientific Integrity/Ethics in science activities (0.3 ECTS)
- Research Ethics Seminar, Leiden University (2017)

PE&RC Annual meetings, seminars and PE&RC weekend/retreat (0.3 ECTS)
- PE&RC Last year's retreat (2023)

National scientific meetings, local seminars, and discussion groups (11.5 ECTS)
- MIMOmics meetings (2015, 2016, 2017)
- BMS-ANed PhD day (2017,2018)
- LUMC department and group meetings (2014, 2015)
- Biometris department and group meetings (2016, 2017)

International symposia, workshops and conferences (10.4 ECTS)
- 34th LASR Workshop with the EU MIMOmics, Leeds (2017)
- Eucarpia, Ghent (2018)
- EMR-IBS and Italian Region conference, Thessaloniki (2018)
- Panhellenic Statistics Conference, Lamia (2018)
- International Biometric Conference, Barcelona (2018)

Societally relevant exposure (2.1 ECTS)
- Organization of RMSS in Cambridge (2017)
- Guest digital lecture at University of Piraeus (2021)
- Guest lecture at ASOEE (2018)
- Lecture for "Introduction to Statistics for the Life and Behavioural Sciences" at Wageningen (2017)
- Guest lecture at University of Piraeus (2015)

Lecturing/supervision of practicals/tutorials (4.2 ECTS)
- Design and analysis of Biomedical studies (2015)
- Computer practicals in Clinical Technology (2015)
- Computer practicals in Clinical Technology (2016)
- Practicals supervision for Statistics I at WUR (2018)

Nothing happened

The research described in this thesis was financially supported by Biometris and by the European Union's Seventh Framework Program (FP7-Health-F5-2012) under Grant Agreement Number 305280 (MIMOmics).

Cover artwork by Elias Kasselas (livethestorm@yahoo.gr)
Untitled, ink on paper, 29.7 x 21cm
Cover design, digital processing, and typography by Puffin Design Studio
(designstudiopuffin@gmail.com) https://puffindesignstudio.gr/

Printed by ProefschriftMaken