Statistical methods for the analysis of complex omics data

Tissier, R.

Citation

Tissier, R. (2018, December 4). *Statistical methods for the analysis of complex omics data*. Retrieved from https://hdl.handle.net/1887/67092

Version:	Not Applicable (or Unknown)
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/67092

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

Leiden University
Repository

The following handle holds various files of this Leiden University dissertation:
http://hdl.handle.net/1887/67092

Author: Tissier, R.
Title: Statistical methods for the analysis of complex omics data
Issue Date: 2018-12-04

Statistical methods for the analysis of complex omics data

Renaud Laurent Michel Tissier

Cover design: IPSKAMP Printing
Printing: IPSKAMP Printing

Copyright ©2018 Renaud Tissier, Amsterdam, the Netherlands.
All rights reserved. No part of this publication may be reproduced without prior permission of the author.

ISBN: 978-94-028-1239-8

This thesis was funded by the EU under the FP7 project MIMOmics (Methods for Integrated Analysis of Multiple Omics Datasets), (Grant/Award Number: FP7/Health/F5/2012/305280)

Statistical methods for the analysis of complex omics data

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 4 December 2018
klokke 10.00 uur

door

Renaud Laurent Michel Tissier

geboren te Louviers, Frankrijk in 1987

Promotor:	Prof. Dr. J. J. Houwing-Duistermaat
Co-Promotors:	Dr. M. Rodríguez-Girondo Dr. S. Tsonaka
Leden promotiecommissie:	Prof. Dr. P. E. Slagboom Prof. Dr. J.H. Barret · *University of Leeds, Leeds* Dr. W.N. van Wieringen · *Vrije Universiteit, Amsterdam*

Table of Contents

1 Introduction **1**
- 1.1 Introduction 1
- 1.2 Measure of dependence: Pearson correlation coefficients 3
- 1.3 Family studies 5
  - 1.3.1 Modelling between subject correlation in family studies 5
  - 1.3.2 Modelling ascertainment in family studies 6
  - 1.3.3 Examples of family studies 7
- 1.4 Secondary phenotypes 8
- 1.5 Correlated predictors 10
- 1.6 Outline of the thesis 12

2 Secondary phenotype analysis in ascertained family designs **15**
- 2.1 Introduction 16
- 2.2 Methods 18
  - 2.2.1 Retrospective likelihood approach 18
  - 2.2.2 Mixed-effects models for the analysis of family data 19
  - 2.2.3 Genotype probability 21
  - 2.2.4 Estimation and statistical testing 22
  - 2.2.5 Continuous polygenic score 22
  - 2.2.6 Inclusion of covariates in the model 23
- 2.3 Simulation Study 23
  - 2.3.1 Simulation results for a SNP 24
  - 2.3.2 Simulation results for a polygenic score 27
- 2.4 Application: Analysis of the Leiden Longevity Study 29
  - 2.4.1 Triglyceride levels analysis 29
  - 2.4.2 Glucose levels analysis 30
- 2.5 Discussion 31

3 Secondary phenotype analysis in family proband designs **33**
- 3.1 Introduction 34
- 3.2 Methods 37
  - 3.2.1 Naive approach: ignoring sampling 37
  - 3.2.2 Joint modeling under retrospective likelihood 37

3.2.3 Conditioning on probands 39
3.3 Simulation study 40
3.3.1 Simulation Setup 40
3.3.2 Simulation Results 42
3.4 The social anxiety disorder study 48
3.5 Discussion 51

4 Gene co-expression network analysis for family studies 53
4.1 Background 54
4.2 Methods 54
4.2.1 Study sample 54
4.2.2 Single probe analysis 55
4.2.3 Network constructions 55
4.2.4 Phenotype analysis 55
4.3 Results 56
4.3.1 Results obtained with all probes 56
4.3.2 Analysis of Q1 57
4.3.3 Results obtained with the 25% most heritable probes 58
4.4 Discussion 59

5 Improving stability of prediction models based on correlated omics 61
5.1 Introduction 62
5.2 Methods 64
5.2.1 Step 1: Network construction 65
5.2.2 Step 2: Hierarchical clustering 67
5.2.3 Step 3: Outcome prediction 67
5.2.4 Software implementation 69
5.3 Simulation Study 69
5.3.1 Simulation setup 69
5.3.2 Simulation results 71
5.4 Real data analysis 79
5.4.1 DILGOM: metabolites 79
5.4.2 DILGOM: Transcriptomics 80
5.4.3 Breast cancer cell lines 86
5.5 Discussion 88

6 Integration of several omic sources in prediction models 91
6.1 Introduction 92
6.2 Network-based group-penalized prediction 93
6.3 Group inference 97
6.3.1 Network construction: WGCNA 97
6.3.2 Hierarchical clustering 97
6.4 Outcome prediction 98

- 6.5 Simulations . . . 99
  - 6.5.1 Simulation setup . . . 99
  - 6.5.2 Simulation study . . . 100
- 6.6 Real data analysis . . . 106
  - 6.6.1 DILGOM . . . 106
  - 6.6.2 Breast cancer cell lines . . . 107
- 6.7 Discussion . . . 109

Bibliography **120**

List of Publications **121**

Summary **125**

Samenvatting **127**

Dankwoord **131**

Curriculum Vitae **133**

1
Introduction

1.1 Introduction

In the last decades, technical developments in biomolecular research have made it possible to collect various omics measurements such as, gene expression, transcriptomics, proteomics, metabolomics, and glycomics. All these measurements, have improved our knowledge of the biological functions in the human body and the mechanisms which get activated in complex diseases. Prediction of disease phenotypes using such omics data in addition to classical environmental factors has also been made possible, opening thereby new research directions for personalized medicine.

Despite the broad availability of omics measurements, the statistical analysis does not always match the complexity of the data generating process. In particular, data collected in complex study designs such as family studies are often analysed without properly taking into account the sampling mechanism and the relatedness of family members. Moreover, incorporating biological or empirically derived information is not always exploited minimizing thereby the potential of state-of-the-art prediction approaches. Furthermore, prediction models are typically based on a unique omic source, thereby, neglecting the potential gain by combining multiple sources of omic predictors. This fact limits the accuracy of personalized prediction and work has to be done on the integration of multiple omic sources in prediction models as there is, actually, no state-of-the-art approach for this type of prediction problem. The development of advanced statistical methods to address the aforementioned complexities is the topic of this thesis. In particular, we present: (i) methods for modelling associations between phenotypes and omics data while cor-

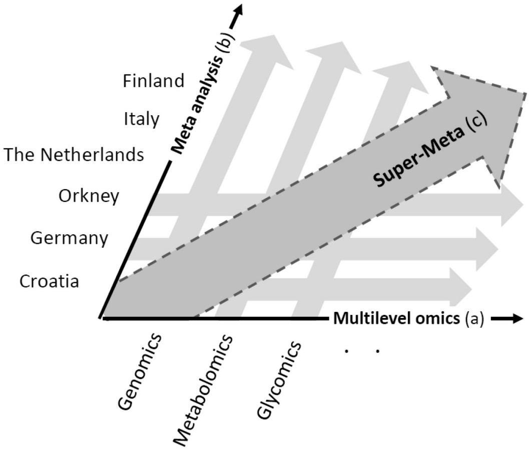


Figure 1.1: Diagram of the super meta-analysis combining several datasets as well as integrating multiple omic sources available in the MIMomics consortiums.

recting for the sampling mechanism in family studies, (ii) methods to build networks of omic features which are collected in family data, (iii) methods to improve prediction by adding information of the correlation structures in group penalization models using only one omic source, and (iv) the extension to several omic sources prediction models.

The research conducted in this thesis was part of the European collaborative project: Methods for Integrated analysis of Multiple Omics datasets (MIMOmics). The goals of the project were: (1) the development of a statistical framework of methods for all analysis steps needed for identifying and interpreting omics-based biomarkers, and (2) to integrate such data derived from multiple omics platforms within studies and across studies and populations. The second goal of the project, namely the development of a super meta-analysis framework, is visualized in Figure 1.1. To establish this super meta-analysis framework the development of robust statistical methodologies which are able to take into account the dependence between omic features, relatedness of individuals in the studies, high dimensionality of datasets and the sampling process were needed.

The methods developed in this thesis can be applied on the (a) and (b) axis of Figure 1.1. In particular, the methods that will be presented in the next chapters focus on the proper analysis of separate omics data under complex study designs and the integration of the various omics in predicting disease outcomes. The proper analysis of omics data under complex study designs allows the integration of the results via meta-analyses (axis (b)) and the analysis of multiple phenotypes simultaneously (axis(a)), while integrating various omics sources in a single prediction model grants the possibility to combine several measurements from one study (axis (a)). The development of such methods was necessary to achieve super meta-analyses.

The rest of this introductory chapter is organized as follows. We will first discuss the key ingredient linking all thesis' chapters, namely the dependence between random variables and present the measures used in the coming chapters to quantify it. Next, we will present the modelling of this dependence in three settings: (i) between individuals i.e. when analysing data from family studies, (ii) between omic features, i.e. when building networks and prediction models and (iii) between outcomes measured on the same subjects. Finally, we will close with a short presentation of the chapters included in this thesis.

1.2 Measure of dependence: Pearson correlation coefficients

One of the most common measures of dependence is the correlation which captures a particular type of dependence, namely linear dependence. Let X and Y two random variables with finite variances σ_x^2 and σ_y^2, the correlation of X and Y which is denoted by ρ_{xy} is given by:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x^2 \sigma_y^2}, \tag{1.1}$$

where σ_{xy} is the covariance between X and Y.

In the case where multiple random variables are recorded, e.g. multiple omic features available in a dataset M, the correlation coefficient ρ_{xy} can be applied on all possible pairs of features leading thereby to the correlation matrix $\mathbf{R}_{\mathbf{M}}$. Modelling such linear associations between omic features in the context of multivariate regression models or network methods is our main concern in Chapters 2-6.

In particular, in Chapters 2 and 3 the correlation matrix of the error terms of multivariate regression models is used to model the dependencies between multiple features measured on the same members of the same family. In Chapters 4-6 the correlation matrix is the input to construct weighted networks. Weighted networks, in general are defined as an adjacency matrix $\mathbf{A} = [a_{ij}]$, where each coefficient a_{ij} represent how close features i and j are. Each non zero coefficient in the matrix $\mathbf{A}$ represent an existing edge between two nodes in the network. One straightforward approach to compute a network of features is to compute their correlation matrix. More sophisticated methods have been developed to obtain more relevant or more interpretable networks. Specifically, in this thesis, we use the weighted gene coexpression network analysis (WGCNA , Zhang and Horvath (2005)) which uses a soft thresholding approach to make the adjacency sparser. The thresholding used in WGCNA is designed to produce network following a free scale topological criteria. This criteria, explained in Chapters 4-6, allows network to follow a hub model. Hubs models are believed to be representative of certain biological mechanisms, and are especially relevant in gene expression (Zhang and Horvath, 2005), helping investigators to identify groups of related features with a meaningful biological interpretation of these groups.

A limitation of the correlation coefficient (Eq 1.1) is that it cannot distinguish direct from indirect linear dependencies. For instance, let X, Y and Z three random variables of interest. Figure 1.2 illustrates one example of possible relationships between them, where the presence of a link between the nodes implies the presence of a direct linear dependence. In this case, the fact that X and Y are both linearly related to Z will lead to ρ_{xy} different from 0 indicating the existence of correlation between them. This dependence is indirect. Making the distinction between direct and indirect dependencies is necessary when trying to identify groups of biologically related features. In this case, the use of partial correlation (Fisher, 1924) is preferred to avoid confounding effects.

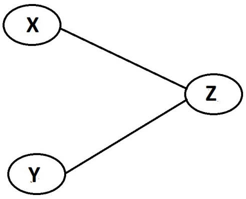


Figure 1.2

The partial correlation, is the correlation between X and Y given other variables, i.e. the conditional dependence between them. In particular, for the triplet of random variables (X, Y, Z) the partial correlation ρ^*_{xy} of X, Y given Z is written as:

$$\rho_{xy} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{1-\rho^2_{xz}}\sqrt{1-\rho^2_{yz}}}$$

In the case of a high number of omics features in a dataset M, the partial correlation for all pairs of variables can be derived in terms of the correlation matrix $\mathbf{R}_{\mathrm{M}}$ by $\mathbf{R}^*_{\mathrm{M}} = \mathrm{scale}(\mathbf{R}^{-\mathbf{1}}_{\mathrm{M}})$, where for a matrix $\mathbf{A}$, $\mathrm{scale}(\mathbf{A}) = \mathrm{diag}(\mathbf{A})^{-1/2}\mathbf{A}\mathrm{diag}(\mathbf{A})^{-1/2}$, with $\mathrm{diag}(\mathbf{A})$ the vector of the diagonal elements of $\mathbf{A}$. Note that when the number of variables exceeds the number of samples in the dataset, the correlation matrix is not invertible. To compute the partial correlation matrix in this case and to make it sparser regularized regression methods can be used, as described in Chapter 5. In particular, penalty functions are introduced to shrink correlation coefficients either separately or as a group in order to allow for inversion of the covariance matrix while retaining only relevant coefficients, forcing other partial correlation coefficients to be equal to 0. Furthermore, the use of group penalization grants the possibility to include a priori knowledge on existing relationships between variables. The partial correlation matrix is a key ingredient of the Gaussian graphical models (Lauritzen, 1996). These models are used in Chapter 5 and can be applied in Chapter 6. Graphical models are used to provide a representation of the existing direct interdependence between several variables allowing investigators to identify groups of features. Compared to WGCNA this approach does not force the network to follow a hub structure. The groups of identified features with graphical models are often smaller since they only contain features having direct interdependences.

Apart from the linear dependence between omic features, another potential source of correlation in genomic data is the relationships between members of the same family. Methods for modelling such dependence is the topic of the next section.

1.3 Family studies

Family studies are often used in genetic research to understand the role of genetics and shared environment in the etiology of disease. In the last years, in addition to genetic markers, several omic measurements are being collected for existing family studies in order to further improve our understanding of human diseases. In family studies, a commonly used design oversamples families enriched with the disease under study, i.e. we only recruit families with at least a certain number of cases. This is the so called multiple-cases family study design. Statistical inference under such a design is known to be robust to population stratification and efficient for detecting rare genetic variants as they tend to aggregate within families. Despite the strengths of family studies, we should acknowledge that recruiting disease-enriched families is harder than the sampling in case-control studies. Moreover, the statistical analysis of family data requires sophisticated approaches (de Andrade and Amos, 2000; Kraft and Thomas, 2000) which explicitly model the familial relationships and deal with the biased sampling design.

1.3.1 Modelling between subject correlation in family studies

Regarding the within families correlations, mixed-effects models are typically used. Let Y_i be a quantitative phenotype for family $i = 1, \ldots, n$ and X_i the $n_i \times (p+1)$ design matrix of p omic features. The linear mixed effects model to study the association between X_i and Y_i in a family i is written as:

$$Y_i = \beta X_i + \sigma_G g_i + \sigma_E e_i + \sigma_\epsilon \epsilon_i \tag{1.2}$$

with β the fixed effects parameter vector, g_i the $n_i \times 1$ random effects vector which models the familial genetic correlation, e_i the shared environment $n_i \times 1$ random effects vector and $\epsilon_i \sim N(0; \sigma_\epsilon^2 I_i)$ the $n_i \times 1$ residual error terms vector. The genetic and environmental variances are σ_G^2 and σ_E^2, respectively. The random variables, g_i, e_i and ϵ_i are assumed to be independent and to follow the multivariate normal distribution with $g_i \sim N_{n_i}(0; K_i)$ and $e_i \sim N_{n_i}(0; E_i)$, where E_i is the environmental correlation matrix often defined as a unit matrix, as all family members share the same environment. K_i is the relationship matrix representing the genetic relatedness between family members.

Let a and b be two members of family i, then the coefficient of the relationsgip matrix between a and b is $k_{iab} = 2^{-d(a,b)}$, where $d(a, b)$ is the genetic distance between family members a and b. The coefficient of relatedness k_{iab} represents the probability that a random allele is shared identical by descent (IBD, Thompson (2008)) by a and b, i.e. the probability that the allele is inherited from a common ancestor. From equation 1.2 it follows $\mathbf{Y_i} \sim (\beta\mathbf{X}_i, \sigma_G^2\mathbf{K}_i + \sigma_E^2\mathbf{E}_i + \sigma_\epsilon^2 I_i)$. For a sample of randomly selected families, the parameters β, σ_G, σ^2, and σ_ϵ can be estimated by maximizing the likelihood function L:

$$\mathbf{L} = \prod_i P(\mathbf{Y}_i|\mathbf{X}_i)$$

For binary phenotypes Y_i, generalized linear mixed model such as probit mixed models or logistic models are used. Such models are used in chapters 2-3 of this thesis.

1.3.2 Modelling ascertainment in family studies

The analysis of family studies is complicated by the oversampling of disease enriched families also known as ascertainment. To derive unbiased estimates of the omics effects on disease phenotypes and heritability related parameters in this case, we need to correct for the chosen sampling scheme. In the literature several approaches have been proposed to address this issue. Namely, the prospective, retrospective and joint likelihood approach (Kraft and Thomas, 2000). Let E be an exposure (catagorical variable), Y the case-control status (binary variable), and S the ascertainment process.

Under the prospective likelihood approach, we condition on the sampling process as shown in equation below:

$$P(\mathbf{Y} \mid \mathbf{E}, \mathbf{S}) = \frac{P(\mathbf{E}, \mathbf{Y}, \mathbf{S})}{P(\mathbf{E}, \mathbf{S})} = \frac{P(\mathbf{S} \mid \mathbf{Y}, \mathbf{E})P(\mathbf{Y} \mid \mathbf{E})}{P(\mathbf{S} \mid \mathbf{E})},$$

which can be further simplified by assuming complete ascertainment, i.e. for all individuals included in the sample $P(S \mid Y) = 1$. We obtain:

$$P(\mathbf{Y} \mid \mathbf{E}, \mathbf{S}) = \frac{P(\mathbf{Y} \mid \mathbf{E})}{P(\mathbf{S} \mid \mathbf{E})}$$

For multiple-cases family studies the denominator $P(\mathbf{S} \mid \mathbf{E})$ can be easily modelled. However, for more complex sampling design, modelling the ascertainment process can be challenging. In such cases, the retrospective likelihood is preferred as this approach corrects implicity for the ascertainment if the ascertainment process depends only on the case-control status.

The retrospective likelihood is based on modelling the distribution of covariates conditional on the outcome and the ascertainment and can be expressed as follows:

$$P(\mathbf{E} \mid \mathbf{Y}, \mathbf{S}) = \frac{P(\mathbf{S} \mid \mathbf{Y}, \mathbf{E})P(\mathbf{E} \mid \mathbf{Y})}{P(\mathbf{S} \mid \mathbf{Y})} = P(\mathbf{E} \mid \mathbf{Y}) \tag{1.3}$$

By application of Bayes'rule, in equation 1.3, The probability of $\mathbf{Y}$ given $\mathbf{E}$ becomes:

$$P(\mathbf{E} \mid \mathbf{Y}) = \frac{P(\mathbf{Y} \mid \mathbf{E})}{P(\mathbf{Y})} = \frac{P(\mathbf{Y} \mid \mathbf{E})}{\sum_E P(\mathbf{Y} \mid \mathbf{E}) P(\mathbf{E})}$$

As previously stated, the main advantage of the retrospective likelihood is the fact that the ascertainment does not need to be modelled. However, this approach does need to model the distribution of the exposure E within families. Therefore, specific assumptions have to be made which provide biased parameter estimates in case of model misspecification. Another drawback of this approach is the loss efficiency by possibly over-conditioning on the phenotype Y of interest and the ascertainment event (Kraft and Thomas, 2000).

The last approach, the joint likelihood, is based on modelling the joint distribution of the exposure and phenotype given the sampling process and is given as follows:

$$P(\mathbf{Y}, \mathbf{E} \mid \mathbf{S}) = \frac{P(\mathbf{E}, \mathbf{Y}, \mathbf{S})}{P(\mathbf{S})} = \frac{P(\mathbf{S} \mid \mathbf{Y}, \mathbf{E}) P(\mathbf{Y} \mid \mathbf{E}) P(\mathbf{E})}{\sum_E P(\mathbf{S} \mid \mathbf{E}) P(\mathbf{E})}$$

This approach combines both disadvantages of the prospective and retrospective likelihood as both ascertainment process and the distribution of the exposure within the family have to be modelled, but is the most efficient as it needs the weakest conditioning (Kraft and Thomas, 2000). Indeed, this approach relies only on the conditioning on the ascertainment process. In the specific case of a family study following a multiple cases design and the exposure of interest is a single nucleotide polymorphism (SNP), both the ascertainment and distribution of the SNP within the family can be modelled.

1.3.3 Examples of family studies

In this thesis data from two family studies are analysed. The Leiden Longevity Study (LLS, Schoenmaker et al. (2006); Houwing-Duistermaat et al. (2009)) is a family-based study set up to identify mechanisms that contribute to healthy ageing and longevity. The inclusion criteria of the study are sibships with at least two alive nonagenarian siblings. Several secondary phenotypes and GWAS data were measured for the offspring of these siblings and their partners. Since the offspring have at least one nonagenarian parent, they are also likely to become long-lived. Therefore, the set of offspring and their partners corresponds to a multiple cases design with related subjects where the offspring are considered as cases and their partners as controls. 421 families with 1671 offspring (cases) and 744 partners (controls) have been included in the study. In Chapter 2, we study the relationships between SNPs and metabolites measured in LLS. Namely, triglyceride levels and glucose levels.

The Leiden Family Lab study on Social Anxiety Disorder (LFLSAD, Bas-Hoogendam et al. (2018)) is a two generation multiplex family study aiming to identify endophenotypes linked to the social anxiety disorder (SAD). Families were considered eligible for inclusion when they contained at least one adult with a primary diagnosis of SAD and whom had at least one child living at home with the proband, showing SAD symptoms.

In addition to these probands other family members were included in the study leading to 9 families with a total number of samples of 132. In Chapter 3, we aim to identify endophenotypes, i.e. heritable phenotypes associated with a primary phenotype of interest, using electroencephalography (EEG) measurements.

1.4 Secondary phenotypes

In genetic studies, apart from the genetic variants and primary disease phenotype e.g. case-control status, a number of omic and non-omic phenotypes are collected as well. These additional phenotypes are known as secondary phenotypes. For instance, in the LLS in addition to case-control status and GWAS, metabolites, classical environmental factors, etc., are measured. Similarly, in the Leiden Family Lab Study omics and fMRI data are available.

In these studies, one of the main research questions is to identify genetic variants associated with these additional secondary phenotypes. In the context of LLS this would help us investigate the presence of pleiotropy, namely the existence of genes associated with multiple phenotypes. The study of pleiotropic effects is important to understand the underlying biological mechanisms of complex diseases. Identifying pleiotropic effects can improve personalized medicine as well. Since specific genetic variants may show strong associations with multiple traits but in opposite directions (Solovieff et al., 2013), identifying pleiotropic effects will help to better prevent and identify possible side effects after gene therapy or genome editing treatments (Solovieff et al., 2013; Gratten and Visscher, 2016). In the LFLSAD, one of the primary objectives is to identify endophenotypes for social anxiety disorder and the genetic variants associated with them. A trait is declared as endophenotype of a specific disease if it is associated with the disease status, if it manifests whether illness is active or in remission (state-independent), and when the trait and the disease status co-segregate within a family. The search for endophenotypes is important as psychologic diseases are complex to diagnose and diagnosis can be subjective. Therefore, identification of genetic variants or biomarkers associated with the disease is difficult. Studying instead the association between genetic variants and highly heritable disease-related phenotypes is needed to understand the relationship between psychological disorders and the genome.

In both studies testing for genetic variants associated with secondary phenotypes is complicated by the sampling mechanism. In both designs, as explained in Section 1.3.2, there is over-representation of cases and we may obtain biased estimates of the association between secondary phenotypes and genetic variants or biomarkers if this is ignored. In the literature several ad-hoc solutions have been initially proposed: testing for association only on cases, testing for association only on controls, or simply adjusting for the case-control status in the regression model used. However, none of these methods properly corrects the sampling mechanism and the relationships between the primary phenotype, the secondary phenotype, and the genetic variants. Figure 1.3 illustrates 6 scenaria for the possible relationships between a SNP (G), a secondary phenotype (X), the case-control status (Y) and the sampling process (S) (Monsees et al., 2009) .

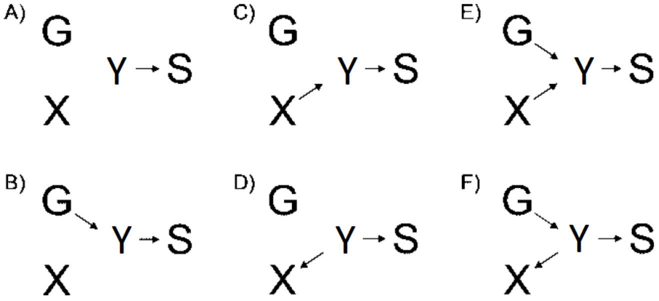


Figure 1.3: Directed acyclic graphs representing the different relationship between a SNP G, a secondary phenotype X, a primary phenotype Y and the sampling process S. Here we assume that the sampling process depends only on the primary phenotype. A: There is no association between G, X, and Y, B: G influences Y, C: X influences Y, D: Y influences X, E: G and X influences Y, F: G influences Y and Y influences X. Bias will occur when estimating the effect of G on X in scenarios B to F. Scenarios D and F induce reverse causality problems and are not considered in this thesis.

In general, the primary and secondary phenotypes are expected to be correlated as they are collected on the same individual. In this case and for the multiple cases studies we consider in this thesis, the sampling distribution of the secondary phenotypes in the study sample is not representative of its distribution in the general population. A naive analysis which ignores this feature will lead to biased estimates of the effect of genetic variants on the secondary phenotype.

In case-control studies, inverse-probability-weighting approaches (Richardson et al., 2007; Monsees et al., 2009) have been proposed to deal with the sampling mechanism on the primary phenotype. Inverse-probability-weighting is an alternative to regression-based adjustment of the outcomes. This approach focus on the idea that individuals have unequal probabilities to be sampled. To correct for bias induced by the sampling mechanisms and obtain proper estimates in the population of interest individuals are weighted by their inverse probability to be included in the study. Therefore, giving a larger weight to individuals having a small probability to be included in the study. This approach is very efficient and simple but can create imbalance if weights are not properly computed. Therefore, proper modelling of the probability of being included in the study is needed. For family studies, the use of inverse-probability-weighting methods is challenging (Rodríguez-Girondo et al., 2018) because we need to compute the probability that a family is recruited in the study which is not available for our studies. Alternatively, the retrospective likelihood approach can be used. The retrospective likelihood, as explained in Section corrects 1.3.2, implicitly for the ascertainment as follows:

$$P(\mathbf{X}, \mathbf{G} \mid \mathbf{Y}, \mathbf{S}) = \frac{P(\mathbf{S} \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}) P(\mathbf{G}, \mathbf{X} \mid \mathbf{Y})}{P(\mathbf{S} \mid \mathbf{Y})} = P(\mathbf{X}, \mathbf{G} \mid \mathbf{Y})$$

Thus under the assumption that the ascertainment depends only on the case-control status, we require modelling: $P(\mathbf{X}, \mathbf{G} \mid \mathbf{Y})$ which using the Bayes rule is further rewritten as:

$$P(\mathbf{X}, \mathbf{G} \mid \mathbf{Y}) = \frac{P(\mathbf{X}, \mathbf{Y} \mid \mathbf{G})P(\mathbf{G})}{P(\mathbf{Y})}$$

To study the effect of a genetic variant on secondary phenotype we need to explicitly model the correlation between X and Y. In the case where X and Y are both continuous then we can assume multivariate normal.

In our case, we have a mix of outcomes and thus we build the joint distributions using latent variables. This model is presented in detail in Chapter 2.

1.5 Correlated predictors

An important complication in the discovery of biomarkers associated with a phenotype and in the prediction of disease phenotypes is the complex correlation structure of features and the fact that most phenotypes are associated with a combination of biomarkers from various omic sources. As a matter of fact, only few diseases are single gene disorders and most of the disease are due to a complex combination of biological and environmental factors.

In this case, separate analysis of omic features using univariate regression models is not advisable. The non-independence of the separate statistical tests and the strong multiple testing correction penalty, due to the high number of omic features, limit the ability of univariate models to discover new associations. A simple solution is then to include all variables of interest as covariates in a multiple regression model. Let Y be the quantitative phenotype of interest and $X = (X_1; \ldots; X_K)$ a set of biomarkers. The linear model is written as:

$$\mathbf{Y} = \alpha + \sum_{k=1}^{K} \beta_k \mathbf{X}_k + \epsilon$$

with α the intercept of the model, β_k the effect size of the k^{th} biomarker, and ϵ the vector of residuals. Even though this method is rather simple, due to the high dimensionality and presence of collinearity in omic features, this model, often, cannot include the whole set of predictors. This is only possible in combination with regularized regression such as the ridge regression (Hoerl and Kennard, 1970) or lasso regularization (Tibshirani, 1996) or variable selection techniques. The lasso regularization is widely used as this approach also forces numerous effect sizes, depending on the size of the penalty, to be equal to zero leading to simpler models. However this approach is also sensitive to the correlation structure between omic features. It will randomly select one variable from a set of strongly correlated features leading to models hard to reproduce and to understand the underlying biological mechanisms. Ridge is appropriate for correlated features, but does not shrink coefficients in the model towards zero leading to complex models which are hard to interpret. Finally, the presence of strong correlations increases the possibility of confounder effects and therefore the quantity of false positive associations.

A solution to overcome these issues is to incorporate the correlation structure in the model. In particular, a two-step procedure may be followed. In the first step, the goal is to identify groups of closely related variables. In the second step, this grouping information is used in the statistical analysis. For the first step, there are two possibilities: either use a Biology- or a data-driven approach. In the biology-driven approach the idea is to incorporate the knowledge about pathways, i.e. groups of single omic features working on a specific cellular function. However, this approach has some limitations: first, the relationship between the variables from the same pathways are not always linear and thus these variables might not be correlated, and second, our knowledge about pathways is still incomplete and therefore we incorporate only a partial picture of the data structure in the model. For the data-driven approach the idea is to empirically derive the correlation structure of the data and to apply clustering algorithms in order to identify clusters of strongly correlated variables. Network construction methods are discussed in Chapters 5 and 6.

Once the groups of omic features have been identified, we can proceed with the statistical analysis of step 2. As far as testing is concerned, one approach is to use the grouping information for dimension reduction and then test for association between omics and phenotypes. In particular, this can be done either by selecting the most "important" variable of the cluster based on a specific criterion or by using a summary measures such as the mean or the first principal component (Pearson, 1901). Association between phenotypes and the summary measures can then be tested in order to detect the group of variables related to these phenotypes. Advantages of such an approach are that it is straightforward to summarize clusters of features in one variable and that we considerably reduce the number of tests to be performed. A downside of this approach is that the use of summary measures makes it hard to reproduce the original results. In prediction models, the grouping information can be incorporated in the statistical analysis, via group penalization methods (Yuan and Lin, 2006; Jacob et al., 2009; Simon et al., 2013; van de Wiel et al., 2014). Thereby, only the most important groups needed for the prediction will be selected leading to more stable and easier to interpret prediction models. Note that recently methods have been developed which allow features to be in different clusters allowing the data to mimic more closely the reality as many biomarkers are part of several pathways. These methods are presented and discussed further in detail in Chapters 5 and 6.

The approaches discussed above have been investigated and applied for a single omic source. Extensions to the simultaneous analysis of multiple omic sources are not straightforward. The integration of multiple omic sources is challenging due to the existing heterogeneity between the different omic sources, in terms of dimensionality, scale, and possible differences in noise structure. These complications make the identification of relationships between them difficult and no state-of-the-art method for integrating different omic sources is available. In Chapter 6 we discuss possible approaches to incorporate several omic sources in prediction models.

1.6 Outline of the thesis

The rest of this thesis contains 5 chapters. As explained in the previous sections, the analysis of omics data can be complicated by several sources of correlation. Table 1.1 shows the correlation structures handled in the different chapters. All chapters may be read in any preferred order, as they have been published or submitted independently. However, we feel that the order in which this thesis has been organized enhances the understanding of the links between the topics of the different chapters. In particular, even though Chapters 2 and 3 both focus on the ascertainment correction for secondary phenotypes in family study designs, we feel that Chapter 3 should be read after 2 as the methodology applied in Chapter 3 is developed in Chapter 2. Chapter 4 may be regarded as the link between the first part (Chapters 2 and 3) and the second part of the thesis (Chapters 5 and 6). In Chapter 4, we still consider family study designs but we use an alternative approach to test for associations between omics data and disease phenotypes, namely correlation networks. Correlation networks are the key ingredient in Chapters 5 and 6, where novel network-based approaches are presented to perform prediction of outcomes in population-based studies with one and multiple omic sources, respectively.

Dependencies	Individuals	Outcomes	Features
Chapter 2	✓	✓	
Chapter 3	✓	✓	
Chapter 4	✓		✓
Chapter 5			✓
Chapter 6			✓

Table 1.1: Overview of the between units dependencies modelled in the different chapters of this thesis.

In Chapter 2, we present a novel approach for the analysis of secondary phenotypes in multiple-cases family studies, i.e. families selected for having at least a certain number of cases. In particular, we work under the retrospective likelihood approach and explicitly model the dependence of the secondary phenotypes and the case-control status using a latent variable approach. A shared random effect is assumed to model the association between the primary and secondary outcome. For the analysis of the primary and secondary phenotypes properly chosen mixed-effects models are used to address the familial relationships. The performance of this approach is empirically evaluated in terms of bias, type I error and robustness to model misspecification. We use the LLS to illustrate the methods.

Chapter 3 explores the performance of the approach developed in Chapter 2 in another very common design for family studies, the proband design. In this design, families are selected based on the phenotype of specific family members. For the analysis of primary phenotypes in this case, the conditioning on proband approach is typically considered. This approach has been recently applied for the analysis of secondary phenotypes (Greenwood et al., 2007; Turetsky et al., 2015) collected under the proband design. However, the dependency between the primary and secondary phenotypes is not modelled. Therefore, in the context of proband designs we compared our method presented in Chapter 2 with the conditioning on proband approach. Both methods are compared in terms of bias in the estimates of genetic effect on secondary phenotypes and heritability in an extensive simulation study. The relative performance of the two methods has been illustrated on electroencephalography (EEG) data from the LFLSAD.

Chapter 4 presents weighted gene coexpression analysis (WGCNA) with family data using a meta-analysis approach. To take into account between family variation, we proposed to perform the WGCNA on each family separately and to combine the obtained results using a meta-analysis approach. This approach was compared with two ad-hoc applications of WGCNA: (1) ignoring the family structure and (2) decorrelation of the gene expression via use of mixed models. To compare their performance, each method was applied on the simulated dataset provided by the Genetic Analysis Workshop 19 (GAW19).

Chapters 5 and 6 present network-based approaches for the prediction of health outcomes using omic sources. In particular, Chapter 5 investigates the combination of network analysis to identify clusters of correlated variables and the incorporation of this information in group penalization in order to improve stability and prediction ability of prediction model using a single omic source. We have considered several combinations of network analysis methods and group regularization approaches. Specifically, as network construction approaches we have used WGCNA and gaussian graphical modelling and as group regularization approaches we have considered: the group lasso, sparse group lasso, and adaptive group ridge. These combinations are compared with common regularization approaches such as lasso, ridge, and elastic net in terms of prediction ability and variable selection via double cross-validation. All methods have been applied to two different datasets: (1) the Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) study where gene expression and metabolomics at baseline were used to predict BMI after 7 years follow-up, and (2) the publicly available breast cancer cell line pharmacogenomics dataset in which we predict the response to treatment of cell lines using gene expression.

Chapter 6 extends Chapter 5 by allowing for the integration of several omic sources in the prediction model. To combine both datasets in the prediction model several approaches have been investigated. The first approach is to perform the network analysis separately on each omic source and combine them in the group regularization approach to predict the outcome. Even though this approach is robust to heterogeneity between omic sources, it is not able to capture interactions between omic sources. The second approach we considered is to apply the approach of Chapter 5 on the stacked omic sources. Even though this approach, potentially, can identify groups of related features coming from various omic sources, it is highly sensitive to difference in scale and heterogeneity between sources. The last approach, is relatively close to the first approach with one additional step consisting of building a new network of summary measures of the clusters obtained in the first step. Clusters containing related summary measures from different omic sources are obtained, and clusters containing features from both omics sources can then be derived from them. Finally, the group penalization is performed using an overlapping group lasso approach allowing the variables to be in different groups. The performance of these approaches has been assessed using metabolomics and gene expression data from the DILGOM study and CNV and gene expression from the breast cancer cell line pharmacogenomics dataset to predict the same outcomes as in Chapter 5.

R codes of the methods developed in Chapters 2-3, Chapters 4-5, and supplementary materials of the different chapters can be found at the git repository:
https://github.com/RenTissier

2

Secondary Phenotype Analysis in Ascertained Family Designs: Application to the Leiden Longevity Study

Abstract

The case-control design is often used to test associations between the case-control status and genetic variants. In addition to this primary phenotype a number of additional traits, known as secondary phenotypes, are routinely recorded and typically associations between genetic factors and these secondary traits are studied too. Analysing secondary phenotypes in case-control studies may lead to biased genetic effect estimates, especially when the marker tested is associated with the primary phenotype and when the primary and secondary phenotypes tested are correlated. Several methods have been proposed in the literature to overcome the problem but they are limited to case-control studies and not directly applicable to more complex designs, such as the multiple-cases family studies. A proper secondary phenotype analysis, in this case, is complicated by the within families correlations on top of the biased sampling design. We propose a novel approach to

---

This chapter has been published as: Renaud Tissier, Roula Tsonaka, Simon P. Mooijart, P. Eline Slagboom, Jeanine J. Houwing-Duistermaat (2017). Secondary Phenotype Analysis in Ascertained Family Designs: Application to the Leiden Longevity Study. *Statistics in Medicine* 36(14), 2288-2301.

accommodate the ascertainment process while explicitly modelling the familial relationships. Our approach pairs existing methods for mixed-effects models with the retrospective likelihood framework and uses a multivariate probit model to capture the association between the mixed type primary and secondary phenotypes. To examine the efficiency and bias of the estimates we performed simulations under several scenarios for the association between the primary phenotype, secondary phenotype, and genetic markers. We will illustrate the method by analysing the association between triglyceride levels and glucose (secondary phenotypes) and genetic markers from the Leiden Longevity study, a multiple-cases family study that investigates longevity.

2.1 Introduction

In order to understand biological mechanisms underlying disease and health, epidemiological studies measure genetic markers, classical variables, and novel omics datasets and model the relationship between these variables and the phenotype of interest. Here we consider outcome dependent sampling designs with binary outcome variables. In addition to studying these binary (primary) phenotypes, the classical or omics variables are typically also analysed as outcome variables (secondary phenotypes). For example modelling of associations between these traits and genetic factors, such as single-nucleotide polymorphisms (SNPs) or polygenic risk scores (sumscores based on SNPs)(Dubdbridge, 2003). However, an important complication which is often ignored is that a proper analysis of the secondary traits should correct for the sampling mechanism on the primary phenotype (Figure 2.1). Note that we assume that the secondary phenotype has an effect on the primary phenotype. The reverse situation will not be treated due to reverse causality challenges (Monsees et al., 2009). In our motivating case study, the Leiden Longevity study (LLS, Houwing-Duistermaat et al. (2009)) families with at least two long-lived siblings are recruited. Obviously, these families do not represent a random sample from the population and inferences cannot be generalized to the whole population, unless the sampling mechanism is properly modelled. Several datasets are measured in the offspring of the long-lived siblings, namely lipidomics, glycomics, metabolomics, and imaging. These offspring share a part of their genetic variation with the long-lived parent and therefore are expected to represent a healthy subpopulation while the partners represent the population. As data example we will model the effect of genetic factors on the secondary traits glucose and triglyceride levels in the offspring (cases) and their partners (controls). To be able to extrapolate results to the general population, we need to account for the over sampling of long-lived subjects in the families of the LLS. There are several multiple-case family studies. For human longevity, GEHA (Genetics of Healthy aging, Skytthe et al. (2011)) used the same study design as the LLS. Other examples are Genetics in Familial Thrombosis (GIFT with at least two cases with thrombosis) (de Visser et al., 2013; Tsonaka et al., 2013) and the ongoing study from Leiden Family Lab (famlab: https://www.leidenfamilylab.nl) which recruits families with at least two cases with social anxiety disorder. The novel methods presented in this paper will also be essential for modelling secondary phenotypes in these studies.

In the context of case-control studies Monsees et al. (2009) showed that bias can occur when estimating the SNP effect on secondary phenotypes if the primary and secondary phenotypes are associated. This is often the case because both outcomes are measured on the same subjects and secondary phenotypes are typically chosen for their potential associations with the primary phenotype. They also showed that the amount of bias is dependent on the prevalence of the primary phenotype, the strength of the association between the primary and secondary phenotypes, and the association between the tested marker and the primary trait (see Figure 2.1).

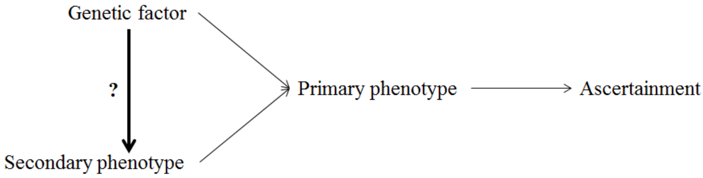


Figure 2.1: Directed acyclic graph representing the case where bias is expected when estimating the association between the genetic marker and the secondary phenotype. Arrows represent existing association between each node of the graph. A secondary phenotype analysis investigates whether there is an association between the genetic factor and the secondary phenotype

To deal with the bias problem, investigators first used ad hoc methods i.e. using controls only, cases only, combined data of cases and controls or joint analysis of cases and controls adjusting for the case-control status. However, several authors showed that these simple approaches can lead to false positive results (Monsees et al., 2009; Lee et al., 1997; Lin and Zeng, 2009). This is due to the sampling design, namely, the secondary phenotype data are not sampled according to the case-control design as the primary phenotype. Several sophisticated methodologies have been developed to correct for the sampling mechanisms and provide unbiased genetic effect estimates: (i) inverse-probability-of-sampling-weighting approaches (Monsees et al., 2009; Richardson et al., 2007; Schifano et al., 2013) which correct for the sampling mechanism by weighting appropriately individuals in case-control studies, (ii) retrospective likelihood-based approaches which indirectly adjust for ascertainment (Lin and Zeng, 2009; He et al., 2011), and (iii) a weighted combination of two estimates obtained with the retrospective likelihood approach in the presence or not of an interaction between SNPs and primary phenotypes (Li and H., 2012).

Even though these approaches can successfully correct for the biased design used to collect the data, they are not directly applicable to more complex designs such as the LLS which motivates this work. In particular, inverse probability weighting approaches require knowledge of the sampling weights for each family. These weights are not available for the LLS because it is unknown what the prevalence of families with at least two nonagenarians is in the population. In addition, the correlations between the family members cannot be ignored and therefore it is evident that statistical methodology for proper secondary phenotypes analysis in this context is needed. To this end, under the retrospective likelihood framework, we develop a multivariate probit regression model inspired by the work of Najita et al. (2009) to model jointly the distribution of the primary and secondary phenotype. This approach allows us to deal with the ascertainment issue while taking into account the individual relatedness and the genetic and environmental variations.

The paper is organised as follows: in Section 2, we present the retrospective likelihood approach to correct for the over sampling of long-lived subjects and the multivariate probit regression model for the joint modelling of the mixed type primary and secondary phenotypes. In Section 3, we evaluate empirically the performance of the method in terms of bias and efficiency and contrast it with the naive approach which ignores the sampling mechanism. Finally, in Section 4 we illustrate the potential of our proposed method in the analysis of triglyceride levels and glucose in the LLS.

2.2 Methods

2.2.1 Retrospective likelihood approach

Let N be the total number of families in the study. For the family i ($i = 1 \dots N$) of size n_i, let Y_i, X_i and G_i be the $n_i \times 1$ vectors for the case-control status, the secondary phenotype and the genotype, respectively. Motivated by the LLS, we will work under the retrospective likelihood approach to correct for the ascertainment of the families. Such an approach is attractive when modelling the ascertainment mechanism is not straightforward, as in the LLS where sampling depends on the previous generation (an example of a pedigree in LLS is shown in Figure 2.2). In fact the retrospective likelihood approach implicitly corrects for the ascertainment mechanism, under the assumption that the ascertainment depends only on the primary phenotype Y. In particular, for the ith family it holds:

$$P\left(X_i, G_i \mid Y_i, Asc\right) = \frac{P\left(Asc \mid Y_i, G_i, X_i\right) P\left(G_i, X_i \mid Y_i\right)}{P\left(Asc \mid Y_i\right)} = P\left(X_i, G_i \mid Y_i\right), \quad (2.1)$$

with Asc the ascertainment process. By applying Bayes rule we obtain:

$$P\left(X_i, G_i \mid Y_i\right) = \frac{P\left(X_i, Y_i \mid G_i\right) P\left(G_i\right)}{P\left(Y_i\right)} = \frac{P\left(X_i, Y_i \mid G_i\right) P\left(G_i\right)}{\sum_{g \in G} P\left(Y_i \mid g\right) P\left(g\right)}. \quad (2.2)$$

To fully specify (2.2) we need to model properly: the conditional joint distribution of the primary and the secondary phenotypes given the genotype P($X_i, Y_i \mid G_i$), the marginal probability of the primary phenotype P($Y_i \mid G_i$), and the genotype probability of the ith family P(G_i). Each one of these elements are described in Sections 2.2.2 and 2.2.3.

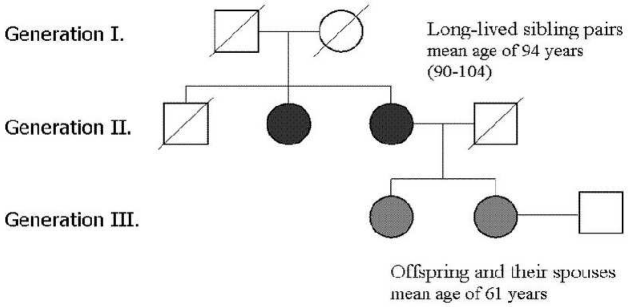


Figure 2.2: Example of a family pedigree from the LLS. Squares and circles represent men and women respectively, crossed symbols represent deceased individuals. In black are the long-lived individuals on whom the ascertainment is based, in grey are the cases of the study (offsprings of long-lived siblings) and in white are the controls.

2.2.2 Mixed-effects models for the analysis of family data

To model the correlation of the phenotypes Y and X within families, a common choice is to use random effects. For the binary primary phenotype we propose to use a multivariate probit model with random effects. The advantage of this model is that it involves only the integrals of the multivariate normal cumulative distribution function for which efficient algorithms have been developed. In contrast, for the more commonly used logistic regression model, the integrals have to be approximated for example by using Gauss-Hermite quadrature which might be computationally intensive for large pedigrees. Let $b_i^Y = \left(b_{i1}^Y, \ldots, b_{in_i}^Y\right)^T$ be a set of family specific random effects designed to handle familial genetic correlation and $G_i = (g_{i1}, \ldots, g_{in_i})^T$ be the vector of genotypes for family i. For the probit model, the observed response Y is viewed as a censored observation from an underlying continuous latent variable Y^* with:

$$Y_{ij} = y_{ij} \Leftrightarrow \gamma_{y_{ij}} < Y_{ij}^* < \gamma_{y_{ij}+1}, Y_{ij} \in \{0,1\}, j = 1, 2, ..., n_i$$

where $-\infty = \gamma_0 < \gamma_1 < \gamma_2 = +\infty$ are suitable threshold parameters. For the underlying latent variable Y^* we assume the mixed-effects regression model

$$Y_i^* = \alpha_0 + \alpha_1 G_i + \sigma_{G_Y} b_i^Y + \sigma \epsilon_i^Y,$$

where $\epsilon_i^Y \sim N_{n_i}(0, I_{n_i})$ is independent of b_i^Y. Here $\alpha = (\alpha_0, \alpha_1)$ denotes the regression coefficient vector with α_0 the intercept and α_1 the parameter representing the effect of the genotype on Y. At the family level we assume $b_i^Y \sim N_{n_i}(0, \mathbf{R}_i)$, with $\mathbf{R}_i$ the coefficient of relationships matrix with elements $r_{lm} = 2^{-d_{lm}}$ with d_{lm} denoting the genetic distance between subjects l and m in the family. The parameter σ_{G_Y} represents the residual additive genetic variation not explained by g_{ij}. Note that σ_{G_Y} models the polygenic inheritance in a family.

For identifiability reasons restrictions are required on both the scale and location of Y^*, namely we set $\sigma^2 = 1$ and $\gamma_1 = 0$. Thus, in the mixed-effects probit regression the disease risk $\pi_{ij} = P(Y_{ij} = 1 \mid b_{ij}^Y, g_{ij})$ conditional on the random-effects b_{ij}^Y and genotypic information g_{ij} is modelled as follows

$$P\left(Y_{ij} = 1 \mid g_{ij}, b_{ij}^Y\right) = \Phi\left(\alpha_0 + \alpha_1 g_{ij} + \sigma_{G_Y} b_{ij}^Y\right), \tag{2.3}$$

with $\Phi\left(z\right)$ the cumulative distribution function of the standard normal distribution. The marginal density under the probit model takes the form:

$$f(y_{ij} \mid g_{ij}; \alpha, \sigma_b) = \int_{b_i^Y} \int_{\gamma_{y_{ij}}}^{\gamma_{y_{ij}}+1} f(y_{ij}^* \mid g_{ij}, b_i^Y; \alpha, \sigma_b) f(b_i^Y) dy_{ij}^* db_i^Y.$$

To model the secondary phenotype X_i we use a linear mixed model:

$$X_i = \beta_0 + \beta_1 G_i + \sigma_{G_X} b_i^X + \sigma_\epsilon \epsilon_i^X, \tag{2.4}$$

where $\beta = (\beta_0, \beta_1)$ denotes the regression coefficient vector with β_0 the intercept and β_1 the parameter representing the effect of the genotype on X, $b_i^X \sim N_{n_i}(0, \mathbf{R}_i)$ is the random parameter used to model the genetic correlation structure within each family for the secondary trait, and σ_ϵ is the residual standard deviation.

To model jointly X and Y using the model specifications (2.3 and 2.4), we introduce a shared random effect $u_{ij} \sim N(0, 1)$ and propose the following model:

$$\begin{aligned} Y_i^* &= \alpha_0 + \alpha_1 G_i + \sigma_{G_Y} b_i^Y + \sigma_u u_i + \epsilon_i^Y, \\ X_i &= \beta_0 + \beta_1 G_i + \sigma_{G_X} b_i^X + \delta\sigma_u u_i + \sigma_\epsilon \epsilon_i^X, \end{aligned} \tag{2.5}$$

where u_i is assumed to be independent of $b_i^Y, b_i^X, \epsilon_i^Y$, and ϵ_i^X. We introduce a coefficient δ in order to have different phenotypic variances for the random effect u_i. In case of small datasets or small family sizes, it can be better to constrain δ to be equal to 1 for a simpler model. Let Σ_{X_i} and $\Sigma_{Y_i^*}$ denote the corresponding variance-covariance matrices of the marginal distributions of X_i and Y_i^* and let $\Sigma_{XY_i^*}$ be their covariance. The joint distribution of Y^* and X is then $(Y_i^*, X_i) \backsim \mathcal{N}_{2n_i}\left(\begin{bmatrix} \alpha_0 + \alpha_1 G_i \\ \beta_0 + \beta_1 G_i \end{bmatrix}, \begin{bmatrix} \Sigma_{Y_i^*} & \Sigma_{XY_i^*} \\ \Sigma_{XY_i^*} & \Sigma_{X_i} \end{bmatrix}\right)$. In the special case for $n_i = 2$, the variance-covariance matrix becomes:

$$\Sigma_i = \begin{pmatrix} \sigma_{G_Y}^2 + \sigma_u^2 + 1 & \sigma_{G_Y}^2 2^{-d(1,2)} & \sigma_{G_X}\sigma_{G_Y} + \delta\sigma_u^2 & \sigma_{G_X}\sigma_{G_Y} 2^{-d(1,2)} \\ \sigma_{G_Y}^2 2^{-d(1,2)} & \sigma_{G_Y} + \sigma_u^2 + 1 & \sigma_{G_X}\sigma_{G_Y} 2^{-d(1,2)} & \sigma_{G_X}\sigma_{G_Y} + \delta\sigma_u^2 \\ \sigma_{G_X}\sigma_{G_Y} + \delta\sigma_u^2 & \sigma_{G_X}\sigma_{G_Y} 2^{-d(1,2)} & \sigma_{G_X}^2 + \delta^2\sigma_u^2 + \sigma_\epsilon^2 & \sigma_{G_X}^2 2^{-d(1,2)} \\ \sigma_{G_X}\sigma_{G_Y} 2^{-d(1,2)} & \sigma_{G_X}\sigma_{G_Y} + \delta\sigma_u^2 & \sigma_{G_X}^2 2^{-d(1,2)} & \sigma_{G_X}^2 + \delta^2\sigma_u^2 + \sigma_\epsilon^2 \end{pmatrix}. \tag{2.6}$$

Using the properties of the multivariate normal distribution, the joint distribution for the observed primary and secondary phenotypes takes the form:

$$\begin{aligned} P\left(Y_i, X_i \mid G_i\right) &= \int P\left(Y_i^*, X_i \mid G_i\right) dy_i^* \\ &= \int P\left(Y_i^* \mid X_i, G_i\right) P\left(X_i \mid G_i\right) dy_i^* \\ &= P\left(X_i \mid G_i\right) \int P\left(Y_i^* \mid X_i, G_i\right) dy_i^*. \end{aligned}$$

Thus by using the probit regression model for the primary trait we have developed an efficient approach to model the correlation between the primary and secondary trait.

From model (2.5) and the variance-covariance matrix (2.6), several marginal correlations between and within family members can be deduced:

$$\begin{aligned} cor\left(X_{ij}, X_{ij'}\right) &= \frac{\sigma_{G_X}^2 2^{-d\left(j,j'\right)}}{\left(\sigma_{G_X}^2 + \delta^2\sigma_u^2 + \sigma_\epsilon^2\right)} = \rho_X \\ cor\left(Y_{ij}^*, Y_{ij'}^*\right) &= \frac{2^{-d\left(j,j'\right)}\sigma_{G_Y}^2}{\left(\sigma_{G_Y}^2 + \sigma_u^2 + 1\right)} = \rho_Y \\ cor\left(X_{ij}, Y_{ij}^*\right) &= \frac{\sigma_{G_X}\sigma_{G_Y} + \delta\sigma_u^2}{\sqrt{\left(\sigma_{G_X}^2 + \delta^2\sigma_u^2 + \sigma_\epsilon^2\right)\left(\sigma_{G_Y}^2 + \sigma_u^2 + 1\right)}} = \rho_{XY} \\ cor\left(X_{ij}, Y_{ij'}^*\right) &= \frac{2^{-d\left(j,j'\right)}\sigma_{G_X}\sigma_{G_Y}}{\sqrt{\left(\sigma_{G_X}^2 + \delta^2\sigma_u^2 + \sigma_\epsilon^2\right)\left(\sigma_{G_Y}^2 + \sigma_u^2 + 1\right)}} = \rho_{XY}', \end{aligned}$$

where ρ_{XY} represents the association between the primary and secondary phenotype. We can also derive the closed form for the heritability estimates of the secondary phenotype which quantifies the percentage of genetic variation in the total variance:

$$H^2 = \frac{\sigma_{G_X}^2}{\left(\sigma_{G_X}^2 + \delta\sigma_u^2 + \sigma_\epsilon^2\right)}. \tag{2.7}$$

Note that when genetic factors are included in the model formula (3.2) gives the residual heritability.

2.2.3 Genotype probability

Finally another key component in the formulation of the retrospective likelihood (2.2) is the computation of the genotype probability for each family i. Let G_{mj} and G_{pj} denote the genotypes of the mother and father of an individual j if this individual is a nonfounder member of family i. Under the assumption of random mating and mendelian inheritance, the genotype probabilities can be written as presented by Thomas (2004):

$$P(G_i) = \prod_{j=1}^{J} \begin{cases} P(g_{ij} \mid G_{mj}, G_{pj}) & \text{if } j \text{ is a nonfounder} \\ P(g_{ij}) & \text{if } j \text{ is a founder} \end{cases} .$$

The probabilities $P(g_{ij} \mid G_{pj}, G_{mj})$ are the transmission probabilities which can be modelled using mendelian inheritance. Finally $P(G_{pi})$, $P(G_{mi})$, and $P(g_{ij})$ can be modelled by assuming Hardy-Weinberg proportions $(1-q)^2$, $2q(1-q)$, q^2 which depend on q, the minor allele frequency. Here we propose to use external information for q or to estimate q from the control sample before maximizing the likelihood. Note that when genotypes of the parents are missing the probability can be obtained by summing over the possible parental genotypes. In case of more complex pedigree a recursive algorithm known as peeling (Elston and Stewart, 2013) can be used . For the LLS where families are sibships the probability is as follows:

$$L(\theta; Y, X) = \prod_i \frac{\{P(X_i \mid G_i) \int P(Y_i^* \mid X_i, G_i)\, dy_i^*\} \sum_{G_p} \sum_{G_m} \prod_j P(G_{ij} \mid G_m, G_p)\, P(G_p)\, P(G_m)}{\sum_g \sum_{G_p} \sum_{G_m} \int P(Y_i^* \mid g)\, P(g \mid G_m, G_p)\, P(G_p)\, P(G_m)}, \tag{2.8}$$

where $\theta = (\alpha_0, \alpha_1, \sigma_{G_Y}, \beta_0, \beta_1, \sigma_{G_X}, \sigma_\epsilon, \delta, \sigma_u)$ is the model parameters vector.

2.2.4 Estimation and statistical testing

To estimate the parameters of the joint model we maximize the logarithm of the likelihood described in (2.8). This involves a combination of numerical optimization and integration. For the evaluation of the integral in the multivariate normal distribution, we use the deterministic algorithm Miwa described in Miwa et al. (2003). For the optimization, we use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm implemented in the function `optim(.)` in R. The BFGS algorithm is a quasi-Newton method, which means that the Hessian matrix does not need to be evaluated directly but is approximated by using specified gradient evaluations. To test for the presence of an effect of the SNPs on the secondary phenotype we use the likelihood ratio test. Note that when the interest of a researcher is solely testing for genetic association a score statistic is an alternative to the likelihood ratio statistic.

2.2.5 Continuous polygenic score

Our approach can also be applied in the case of modelling the association between continuous covariates and secondary phenotypes. For example polygenic scores have been used to summarise genetic effects among an ensemble of SNPs that have been identified in large GWASes (International Schizophrenia Consortium et al., 2009; (IMSGC) et al., 2010; Simonson et al., 2011). Polygenic scores are typically linear combinations of SNPs: $G = \sum_k \delta_k SNP_k$, where $\delta_k = 1$ or δ_k is obtained from previous GWASes. For genetic scores, we need to integrate over the distribution of the polygenic score instead of summing over the genotypes in the denominator of (2.2). For the distribution of the polygenic score we use a multivariate normal distribution $G_i \backsim \mathcal{N}_{n_i}(\mu_g, \sigma_g R_i)$, with μ_g

the mean value of the genetic score, σ_g the standard deviation of the genetic score and R_i the relationship matrix of family i. The likelihood contribution for family i is given by:

$$\frac{P\left(Y_i, X_i \mid G_i\right) P\left(G_i\right)}{P(Y_i)} = \frac{P\left(Y_i, X_i \mid G_i\right) P\left(G_i\right)}{\int_{y_i^*} P(y_i^*) dy_i^*} = \frac{P\left(Y_i, X_i \mid G_i\right) P\left(G_i\right)}{\int_{y^*} \int_{g_i} P(y_i^* \mid g_i) P(g) dy_i^* dg_i}.$$

Computation of the integral $\int_{y^*} \int_g P(y^* \mid g) P(g) dy_* dg$ can be quite intensive and challenging. In order to gain efficiency we write the marginal model of Y^* (2.5) as $Y_i^* = \alpha_0 + b_i^{Y*} + u_i + \epsilon_i^Y$, with $b_i^{Y*} = \sigma_{G_Y} b_i^Y + \alpha_1 G_i$. Now Y_i^* follows the following multivariate normal distribution: $Y_i^* \backsim \mathcal{N}_{n_i}\left(\alpha_0 + \alpha_1 \mu_g, \Sigma_{Y_i^*} + \alpha_1^2 \sigma_g^2 R_i\right)$. Note that when a polygenic risk score is included in the model for the secondary phenotype, the parameter σ_{G_Y} represents the residual polygenic inheritance.

2.2.6 Inclusion of covariates in the model

Often, researchers want to adjust for covariates such as age, sex, treatment etc in the model. Let Z be such a covariate. To estimate the effect Z on the secondary phenotype we propose to maximize the joint likelihood of X and G conditionally on the primary phenotype Y and Z. Thereby we avoid modeling of the distribution of Z within the families. Indeed, under the assumption of independence between genotype and Z we obtain:

$$\begin{aligned} P\left(X_i, G_i \mid Y_i, Z_i\right) = \frac{P\left(X_i, Y_i, Z_i, G_i\right)}{P\left(Y_i, Z_i\right)} &= \frac{P\left(X_i, Y_i \mid G_i, Z_i\right) P\left(G_i\right) P\left(Z_i\right)}{P\left(Y_i \mid Z_i\right) P\left(Z_i\right)} \\ &= \frac{P\left(X_i, Y_i \mid G_i, Z_i\right) P\left(G_i\right)}{P\left(Y_i \mid Z_i\right)}. \end{aligned} \tag{2.9}$$

2.3 Simulation Study

A simulation study was set up to evaluate the performance of our proposed method for the estimation of the association between a genetic factor and the secondary phenotype and the estimation of the heritability of the secondary phenotype. We compare the proposed method with the naive approach which is typically followed in practice, namely analysis of the secondary trait without correcting for the sampling mechanism. In particular, in this case, we fit the standard linear mixed-effects model for the secondary phenotype and explicitly model the familial relationships as described in (2.4). The two methods are compared in terms of bias, Root Mean Square Error (RMSE) and 95% coverage probabilities. We consider SNPs (discrete variables) and polygenic scores (continuous variables). Several settings are considered for the disease prevalence, the strength of the association between the genetic factor and the primary phenotype, the strength of the ascertainment mechanism and the number of sibships. We simulated sibships of size 5.

With respect to the familial relationships, we consider only sibships such that our simulation resembles the LLS design. For the prevalence of the primary phenotype we consider two settings namely a disease prevalence of 1% which corresponds to $\alpha_0 \approx -2.32$ and of 5% which corresponds to $\alpha_0 \approx -1.64$. In addition the variance parameters have been chosen such that they correspond to a heritability of 50%. Specifically we use σ_{G_X}=2, $\sigma_{G_Y} = \sqrt{3}$, $\sigma_{u_X} = \sigma_{u_Y} = \sqrt{2}$ and $\sigma_\epsilon = \sqrt{2}$. This corresponds to a correlation of 0.78 between the primary and the secondary phenotypes. To speed up computations, we assume that $\sigma_{u_X} = \sigma_{u_Y}$ when fitting the models to the simulated datasets. For each scenario, 500 datasets are simulated using model (2.5).

2.3.1 Simulation results for a SNP

The genotypes of the SNPs are simulated assuming a minor allele frequency of 0.3 in the population. For the secondary phenotype model the following fixed effects values are used: $\beta_0 = 3.5$ and $\beta_1 = 0.2$, whereas for the primary phenotype model the effect sizes are $\alpha_1 = 0.1$ or 0.5. Finally, for each of the four scenarios (rare or common disease, and weak and strong SNP effect on the primary phenotype) we consider two ascertainment mechanisms, namely the sampled sibships of size five have at least one affected or at least two affected members.

Figure 3.3 presents the estimates and 95% confidence intervals for the scenario of 400 sibships. Figure 3.3 shows that ignoring the sampling mechanism (naive method) leads to biased estimates of the SNP effect and the size of this bias increases with the strength of the ascertainment mechanism and the association between the SNP and the primary phenotype. Overall we observe that the proposed method gives unbiased estimates of the SNP effect on the secondary phenotype. The coverage probabilities reach the nominal level (see section A of supplementary material). Regarding the prevalence of the primary phenotype, we observe that for the naive method bias increases with lower prevalence, while the proposed method remains robust to the lower amount of information due to the rare primary phenotype. In general, the proposed method leads to smaller RMSE than the naive approach and better coverage probabilities.

In Table 2.1 we present the heritability estimates of the secondary phenotype for a common disease, under the various ascertainment mechanisms and the two values of α_1. It is obvious that the heritability estimates are influenced by the ascertainment mechanisms when using the naive approach. Indeed the naive method tends to underestimate the heritability for each mechanism and this underestimation increases as the ascertainment mechanisms become more stringent. The heritability estimates are 25-27$\%$ for sibships with at least one affected sibling and drop to 13-14$\%$ for sibships with at least 2 affected siblings. On the contrary, the proposed method is robust to the stringency of the ascertainment mechanism.

Next, we study the robustness of our approach to one violation of the model assumptions, namely we simulated under a logit link for the primary phenotype and used the probit link for modelling. Results for the SNP effect and the heritability are presented

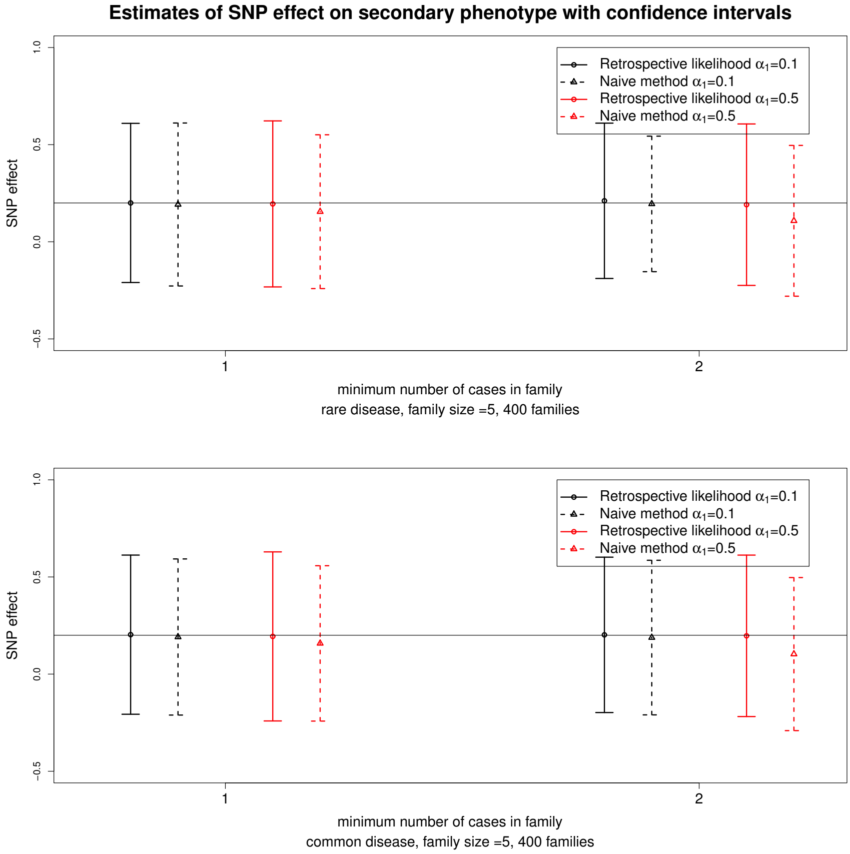


Figure 2.3: Estimates and 95% confidence intervals for the SNP effect on the secondary phenotype for the retrospective likelihood approach and the naive method. Results are obtained from 500 simulated datasets of 400 families for 2 ascertainment schedules. The top and bottom panel correspond to a rare or common primary phenotype with a prevalence around 1% and 5% respectively. In black and red are represented results for small (α_1=0.1) and large (α_1=0.5) effect sizes of the SNP on the primary phenotype, respectively. The horizontal line corresponds to the true SNP effect on the secondary phenotype.

in Table 2.2. These results show that even though our approach gives biased estimates for the primary phenotype model, the parameters estimates for the secondary phenotype model are not affected. All the results are presented in Section A of the Supplementary Material.

Although we focus on parameter estimation, model fitting, and heritability estimation for genetic association with a secondary phenotype, we also investigate the performance of the likelihood ratio test under the null hypothesis of no genetic association with a secondary phenotype at two levels of genetic association with the primary phenotype. In

		SNP model		Polygenic score model	
Ascertainment	α_1	Retrospective	Naive	Retrospective	Naive
1. 2 cases					
	0.10	0.48(0.07)(0.22)	0.13(0.07)(0.37)	0.50(0.03)(0.13)	0.14(0.03)(0.36)
	0.50	0.48(0.07)(0.22)	0.14(0.07)(0.36)	0.52(0.03)0.12)	0.15(0.03)(0.34)
2. 1 case					
	0.10	0.50(0.08)(0.17)	0.25(0.08)(0.25)	0.48(0.04)(0.12)	0.25(0.03)(0.24)
	0.50	0.50(0.08)(0.17)	0.27(0.08)(0.24)	0.50(0.04)(0.10)	0.26(0.04)(0.23)

Table 2.1: Heritability results of the simulation studies for a SNP and a polygenic score: Estimates with standard deviations and RMSE (in brackets) for the heritability of the secondary phenotype for a common disease (prevalence $\approx 5\%$), when families with at least one and at least two cases are sampled and for two values of α_1, i.e. SNP or polygenic score effect on primary phenotype. Datasets consist of 400 families of size 5. Results are based on 500 replicates.

Ascertainment	α_1	β_1	heritability
0.True value		0.200	0.500
1.At least 2 cases			
	0.100	0.199(0.104)(0.104)(0.948)	0.509(0.017)(0.110)
	0.500	0.197(0.106)(0.110)(0.945)	0.516(0.014)(0.108)
2.At least 1 case			
	0.100	0.200(0.104)(0.107)(0.961)	0.510(0.012)(0.096)
	0.500	0.199(0.107)(0.111)(0.960)	0.513(0.010)(0.087)

Table 2.2: Robustness: Estimates of the effect size of the SNP on the secondary phenotype (β_1) and heritability of the secondary phenotype are given for a common disease (prevalence $\approx$ 5%), for the two ascertainment mechanisms and two values of α_1. Into brackets are standard deviations, RMSE and coverage probability (for the effect size only). Datasets consist of 400 families of size 5. Results are based on 500 replicates.

	nominal level (α)	Retrospective likelihood	Naive method
At least 2 cases			
α_1 =0.1			
	0.05	0.0509	0.0580
	0.01	0.0118	0.0152
	0.001	0.0017	0.0025
α_1 =0.5			
	0.05	0.0505	0.0878
	0.01	0.0113	0.0222
	0.001	0.0013	0.0043
At least 1 case			
α_1 =0.1			
	0.05	0.0524	0.0514
	0.01	0.0102	0.0098
	0.001	0.0018	0.0014
α_1 =0.5			
	0.05	0.0522	0.0558
	0.01	0.0098	0.0097
	0.001	0.0009	0.0016

Table 2.3: Type I errors rates for testing for association between a genetic marker and a secondary phenotype for four scenarios. Families with at least one and with at least two cases are considered. Two values for the association between the SNP and the primary phenotype namely α_1 = 0.1 and α_1 = 0.5 are used. Datasets consist of 400 families of size 5. Results are based on 10000 replicates.

each of the four considered scenarios, we simulate 10,000 replicates. In Table 3.2 the emprical type I error rates are given for the rare disease scenario (i.e. prevalence 1%). We observe that while our approach preserves the type I error rate at a nominal level, the naive approach has, systematically, an inflated type I error rate. The type I error rate for the naive method increases with stronger ascertainment and larger SNP effect on the primary phenotype.

2.3.2 Simulation results for a polygenic score

To study the performance of the proposed method for polygenic score, we simulated centered and standardized scores. The parameters of the secondary phenotype model were chosen as for the SNP simulations: $\beta_0 = 3.5$ and $\beta_1 = 0.2$, whereas for the primary phenotype model effect sizes of $\alpha_1 = 0.1$ or 0.5 were used. Figure 3.4 presents the estimates and confidence intervals for datasets with 400 sibships. Our approach provides unbiased estimates of the effect of the polygenic score on the secondary phenotype. In contrast, the naive approach provides biased estimates and the bias increases when the ascertainment process is more stringent or when α_1 is larger.

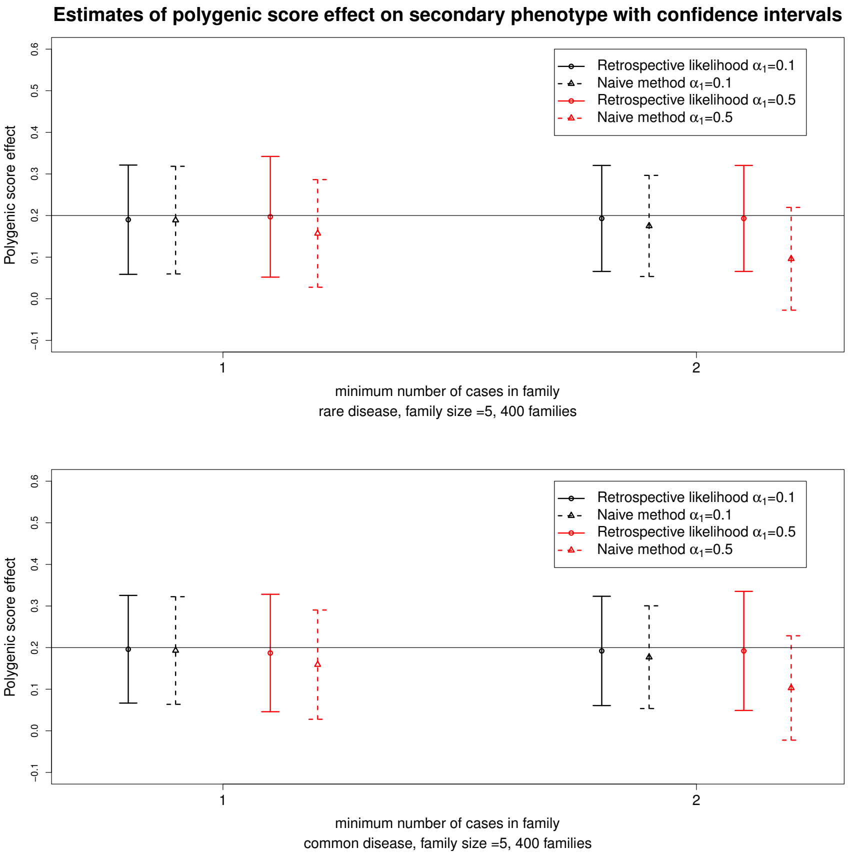


Figure 2.4: Estimates and 95% confidence intervals for the polygenic score effect on the secondary phenotype for the retrospective likelihood approach and the naive method. Results are obtained from 500 simulated datasets of 400 families for 2 ascertainment schedules. The top and bottom panel correspond to a rare or common primary phenotype with a prevalence around 1% and 5% respectively. In black and red are represented results for small (α_1=0.1) and large (α_1=0.5) effect sizes of the polygenic score on the primary phenotype, respectively. The horizontal line corresponds to the true polygenic score effect on the secondary phenotype.

The results of the residual heritability estimates after adjustment for polygenic scores agree with the results obtained when a SNP is included in the model (Table 2.1). The naive approach did not perform well: estimates between 25-26% and 14-15% for an ascertainment process of at least one affected sibling and at least two affected siblings respectively instead of 50%.

2.4 Application: Analysis of the Leiden Longevity Study

In this Section, we will exemplify our proposed method in the analysis of the LLS briefly introduced in Section 1. The LLS is a family-based study set up to identify mechanisms that contribute to healthy ageing and longevity. The inclusion criteria of the study are sibships with at least two nonagenarian siblings, i.e. the selection takes place at Generation II (Figure 2.2). Several secondary phenotypes and GWAS data have been measured for the offspring of these siblings (Generation III in Figure 2.2) and their partners. Since the offspring have at least one nonagenarian parent, they are also likely to become long-lived. Therefore, the set of offspring and their partners corresponds to a case-control design with related subjects where the offspring in Generation III are considered as cases and their partners as controls. Overall 421 families with 1671 offspring (cases) and 744 partners (controls) have been included in the study. Because the families are relatively small we use the model which assumes an equal variance for the shared effect for the two traits.

Here we model the association between genetic factors and the secondary phenotypes triglyceride and glucose levels. For both traits, there is evidence of an association with human longevity and both traits are normally distributed. For the sake of comparison in addition to our proposed method, we will present results using the naive approach i.e. standard linear mixed model. Analyses using the linear mixed model which conditions also on the case-control status will not be presented because the parameters do not have a comparable interpretation between the two approaches. The p-values presented below are obtained using the likelihood ratio test.

2.4.1 Triglyceride levels analysis

Triglyceride levels have been found to be associated with the primary trait longevity (p-value = 0.0005 for women and p-value = 0.04 for men) and the size of association is sex dependent. Therefore a sex-stratified analysis has been considered further. For the purposes of our illustration, we restricted our analysis to seven genes on chromosome 11 which are known to be associated with Triglyceride levels. These genes are *APOA1, APOA4, APOA5, APOC3, ZNF259, BUD13* and *DSCAML1*. The selection of the genes was performed using the NHGRI-EBI GWAS catalog (Welter et al., 2014). For these genes, we have genotypes of 41 SNPs which have no missing values in our datasets. Triglyceride levels were standardized and we included age as a covariate in the analysis.

We ran the analysis with the constrained approach, i.e. $\delta = 1$. We observe that none of the SNPs analysed is significantly associated with Triglyceride levels either in men or in women, hence for most SNPs the estimates of the effect sizes agree between the two approaches. The SNPs showing the largest differences are, in men, SNP 22: β_1^{RA} = 0.047 for our Retrospective Approach (RA) and β_1^{NA} = 0.052 for the Naive Approach (NA) and SNP 26: β_1^{RA} = 0.088 and β_1^{NA} = 0.092. For women more SNPs give different estimates between the two approaches, i.e. SNP 1 (β_1^{RA} = 0.024, β_1^{NA} = 0.020), SNP 2 (β_1^{RA} = 7.2e-06 β_1^{NA}=0.006), SNP 13 (β_1^{RA} = -0.013, β_1^{NA} = -0.009) and SNP 19 (β_1^{RA} = 0.011, β_1^{NA} = 0.007) showed the biggest differences. Results for the SNPs are presented in Section B of the Supplementary Material.

We verified whether the assumption of equal variances for the primary and secondary phenotype for the shared effects is justified. We fitted also the model with non constrained δ. We noticed that for some of the SNPs the model parameters are hard to estimate and the estimates of the variances of the shared and residual random effects in the model for the second phenotype are swapped. Overall the estimates of the effect of the SNP on the secondary phenotype are very similar to the model which assumes equal variances. Results of these analyses are presented in Section B of the Supplementary Material.

2.4.2 Glucose levels analysis

In previous analysis of glucose levels in the offspring and partners of the LSS, Mooijaart et al. (2010) studied the association between glucose and a polygenic score. The genetic score was defined as the total number of risk alleles across 15 SNPs which are known to be associated with Type II diabetes. The Generalized Estimating Equation method was applied to take into account the familial relationships. The paper showed that a higher number of Type II diabetes risk alleles is associated with a higher serum concentration of glucose ($p - value$ = 0.016). A statistically significant association was found between glucose level and case-control status (p-value$<$ 0.001). However, the sampling process was not taken into account in the analysis and thus the results might be biased. We applied the proposed method to estimate the heritability of glucose levels and to test for the presence of an association between the glucose levels and the polygenic score. In addition, we applied the naive approach which did not correct for case-control status. We did not stratify according to sex in these analyses.

For this analysis the polygenic score was standardized. Using the Retrospective approach, the association between the genetic score and the glucose level is estimated by $\beta_1^{RA} = 0.630$ with a standard error of $stE = 0.023$ ($p - value = 0.015$). The naive approach also yields a significant association between the genetic score and glucose levels ($\beta_1^{NA} = 0.622$, $stE = 0.026$, $p - value = 0.020$). By using the Naive Approach (NA) we obtained for the glucose levels a genetic variance of $\sigma^2_{G_X} = 0.302$ and a total variance of $\sigma^2_T = 1.322$, which corresponds to a residual heritability of $h^2_{NA} = 0.228$. Our Retrospective approach (RA) yields a genetic variance of $\sigma^2_{G_X} = 0.384$ and a total variance of $\sigma^2_T = 1.457$ which corresponds to a residual heritability of $h^2_{RA} = 0.263$.

2.5 Discussion

In this paper, we developed a new method for the proper analysis of secondary traits for multiple-cases family designs. A key component in our proposed method is the joint modelling of the primary and secondary phenotypes. We developed a multivariate probit model which can also capture the within families dependencies. A retrospective likelihood approach has been followed to correct for the ascertainment process. Thereby unbiased estimates of the association between genetic factors and secondary traits can be obtained. Simulation results showed that our approach preserves the type I error at nominal level and provides accurate estimates irrespective of the disease prevalence, the strength of the association between the genetic variants and the primary phenotype, and the ascertainment mechanism. Another important empirical finding is that the heritability estimates for the secondary traits can be severely underestimated unless the sampling mechanism is taken into account. With respect to the analysis of the motivating case study, for the SNPs the differences between the effect sizes obtained by our proposed method and the naive approach were small. The small differences obtained between the naive and the retrospective approach are mainly due to the small effect sizes of the genetic markers selected on the primary phenotype. Indeed, the three main factors influencing the magnitude of the bias when using the naive approach are the correlation between the secondary phenotype and the primary phenotype, the strength of the ascertainment, and the strength of the association between the genetic marker and the primary phenotype.

Heritability is one of the properties that a trait needs to possess to be declared an endophenotype for a specific disease. The other criteria are: the trait is associated with the disease status in the population, the trait manifests whether illness is active or in remission (state-independent), and the trait and the disease status co-segregate within a family (Gottesman and Gould, 2003). The Leiden Family Lab (https://www.leidenfamilylab.nl) aims to identify endophenotypes for social anxiety disorder. The study comprises families with at least two cases with social anxiety. The methods presented in this paper will be used for the analyses of this study to identify endophenotypes and are relevant for other family studies, as well.

In this paper, we proposed to include additional covariates in the model by using the likelihood conditional on these covariates. Alternatively the joint likelihood of the secondary phenotype, genotype, and covariate conditionally on the primary phenotype can be used. This alternative approach might be more efficient (Balliu et al., 2015). However this likelihood requires distributional assumptions for the covariates within families which can be complex for related individuals. Moreover maximization of the likelihood might become time consuming. Ghosh *et al* (Ghosh et al., 2013) propose a pseudo-likelihood and a profile approach to include covariates in a secondary phenotype analysis for case-control data. This work needs to be extended to family data. A Monte Carlo approach might be considered to compute the integrals (Tsonaka *et al* (Tsonaka et al., 2015)).

Typically there are missing genotypes. In unrelated individuals, genotypes can be imputed based on the haplotype structure obtained from a reference panel. For family data, the imputation should also take into account the genotypes of other family members. Software exists which can perform such analysis, for example the Genotype Imputation Given Inheritance (GIGI) program (Cheung et al., 2013). However for the computation of the denominator in equation (2.2) these imputed genotype probabilities have to be taken into account.

Due to the computational intensity of the proposed method, it is not yet possible to run full GWAs analyses of secondary phenotypes. However, the proposed method can be used on a set of pre-selected variants e.g. after an initial screening with the naive approach to the primary and secondary phenotypes or when investigating pleiotropic effects. To reduce computation time of the multivariate integrals in the numerator and the denominator, a faster algorithm can be used than the one used in this paper. The randomized Quasi-Monte-Carlo procedure, developed by Genz (1992), is less accurate but faster especially for large pedigrees. Development of less computational intensive methods is one of the topics for future research.

With regard to pleiotropic effects, a criticism of probit random-effects models is that in the presence of high dimensional random effects we cannot move from the subject-specific interpretation for the fixed effects parameters to the population-level interpretation as in the random-intercepts case. When the outcome is binary and families are relatively small, estimation of the intercept and variances terms can be difficult, and consequently coverage probabilities can be poor. Tsonaka et al. (2013) showed efficiency gains by using information on disease prevalence. Their methods need to be adapted to our setting of the analysis of two phenotypes. When the parameters of the primary phenotype model are not of interest and this model is only used to correct for the ascertainment mechanism which is driven by the primary phenotype, we showed that secondary phenotype analyses with the proposed method are robust to using the probit instead of the logit link function.

Future directions in the LSS and Leiden Family Lab Study will address the pending availability of multiple omics and fMRI data, respectively, and joint modelling of several glycans or voxels is of interest. Extending our approach, in this case, is algebraically straightforward, but practical implementation may be challenging due to computational intensity especially with a large number of secondary phenotypes. Use of composite likelihood approaches might be a solution and is our current research topic.

Finally, an attractive alternative approach to properly analyse secondary traits is to apply inverse probability weighting. However, it is crucial to correctly specify the weights. Currently, we do not have sufficient information to be able to estimate these weights for our studies. However with access to electronic records for research, such as information from general practitioners to estimate the weights, it is likely that inverse probability weighting approaches can be developed.

3

Statistical methods for the analysis of secondary phenotypes in family proband designs

Abstract

Numerous epidemiological studies comprise collections of traits in addition to the primary phenotype. Typically these studies use an outcome dependent design in which subjects with extreme values for the primary phenotype are oversampled. These additional traits are secondary phenotypes and straightforward analysis which ignores the study design may yield biased effect estimates. Especially when the covariate of interest is also associated with the primary phenotype and when the primary and secondary phenotypes are correlated, the selection based on the primary phenotype needs to be modelled. Family studies use various types of ascertainment procedures. The most common ones are the proband design, i.e. family members of a specific subject (probands) are recruited, and the multiple cases design, i.e. families with a specific distribution of the primary phenotype are recruited. For example the families should have at least two cases. Recently we proposed an approach for the analysis of secondary phenoytpes in multiple cases family

---

This chapter is in preparation for publication as: Renaud Tissier, Jeanine J. Houwing-Duistermaat. Statistical methods for the analysis of secondary phenotypes in family proband designs.

studies. The approach is based on the retrospective likelihood and joint modelling of the primary and secondary phenotypes. Here, will consider the proband design. We compare via mathematical formula and simulations the performance of our approach and an often used approach which is based on the conditional distribution of the trait values given the trait values of the probands. The last approach is implemented in the SOLAR-eclipse software. We will illustrate the methods by analyzing data from the Social Anxiety Disorder (SAD) family study. We conclude that our approach performes well and yields unbiased estimates for the heritability and the SNP effects on the secondary phenotype for a proband design. However, when the information on the probands is missing, there appears to be a small bias. We showed that conditioning of the trait values of the proband violates the model assumptions and hence leads to biased estimates.

3.1 Introduction

Family studies are an important tool to understand the relationship between genetic, lifestyle and shared environmental factors and complex traits. Typically residual correlation between outcomes of family members exists due to unobserved shared genetic, lifestyle and environmental factors. In addition most of the family studies use outcome dependent sampling. These two issues make the statistical analysis of data from family studies challenging. The correlation between the outcomes has to be modelled and corrections for the ascertainment process are required to obtain unbiased parameter estimates with correct standard errors. Here we consider the two most commonly used selection schemes (Figure 3.1). For the proband design ascertainment is based on the primary phenotypes of probands, i.e., families are selected in the study because one or more specific family members are known to have a extreme value of a trait of interest. One example is the Family Violence study where parent-child agreement on child maltreatment was examined in a multigenerational study (Compier-de Block et al., 2017). A part of the families was recruited via probands selected from an epidemiological study. Although the families in this design are recruited via the probands, not always data for the probands are available. For example, a family can be recruited for having at least two members who deceased due to cardio-vascular diseases (Irvin et al., 2014). Missingness of data on probands is another complication in the proband family design. The second ascertainment process is based on recruiting families with at least a certain number of affected family members. An example is the Leiden Longevity Study where sibships with at least two nonagenarian siblings are included in the family (Houwing-Duistermaat et al., 2009). These two types of ascertainment processes need probably different ascertainment corrections. For the proband design the correction comprises using the conditional distributions of the outcomes on the family members given the outcomes of the probands. For the muliple case family studies corrections need to be based on the probability for a family to have a certain number of affected members.

While a lot of work is available on ascertainment corrections when the aim of the analysis is to model the primary phenotype (de Andrade and Amos, 2000), work on secondary

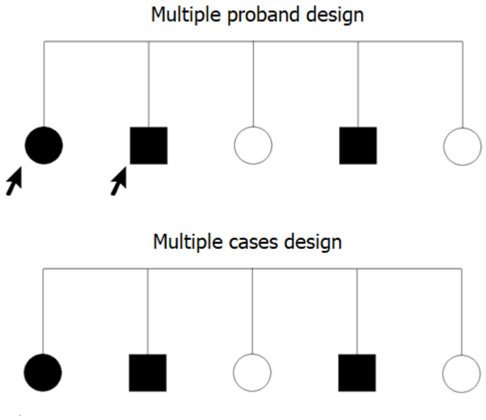


Figure 3.1: Example of pedigree representing the two types of ascertainment process for familiy studies for a family of five siblings. Top, the multiple proband design where the family is selected based on 2 specific family members being affected. Bottom, the multiple cases family design where family are selected for having at least 2 affected members. Squares and circle represent the two genders. Black filled symbols represent affected family members.

continuous phenotypes is scarce. Epidemiological studies comprise however often many secondary phenotypes, for example omics variables, such as metabolomics, proteomics, and glycomics, and Electroencephalography (EEG) and MRI data. In psychology, identification of endophenotypes of illnesses is an important research topic. Endophenotypes are by definition secondary phenotypes. They satisfy four conditions, namely the phenotype is associated with the disease, is heritable, manifests whether the disease is active or in remission (state-independent), and co-segregates with the disease status within families. Since in psychology the definition of diseases might be challenging, these endophenotypes offer an alternative to detect underlying genetic mechanisms (Gottesman and Gould, 2003; Glahn et al., 2007; Miller and Rockstroh, 2013; Iacono et al., 2017). However for proper analysis of endophenotypes (secondary traits) corrections for the sampling mechanism based on the primary phenotype need to be made. For case-control studies this was shown by Monsees et al. (2009) and for multiple case families this was shown by Tissier et al. (2017). Both papers studied the effect of selection on parameter estimates for different scenarios of relationships between genetic markers (G), the secondary phenotypes (X) and the primary phenotype (Y). Four of these scenarios are given as directed acyclic graphs in Figure 2.1. When the secondary and primary phenotype X and Y are correlated, the estimate of the parameter modelling the effect of G on X will be biased under the alternative hypothesis. Since both outcomes are measured on the same subjects and the secondary phenotypes are typically chosen for their potential associations with the primary phenotype, this is often the case. The scenarios C and D in Figure 2.1

correspond to these situations. The authors also showed that the amount of bias depends on the prevalence of the primary phenotype, the strength of the association between the primary and secondary phenotypes, and the association between the tested marker and the primary trait. Note that Monsees presented two more scenarios. However since they involve reverse causality, they were not further considered in Monsees paper and we will not consider them neither.

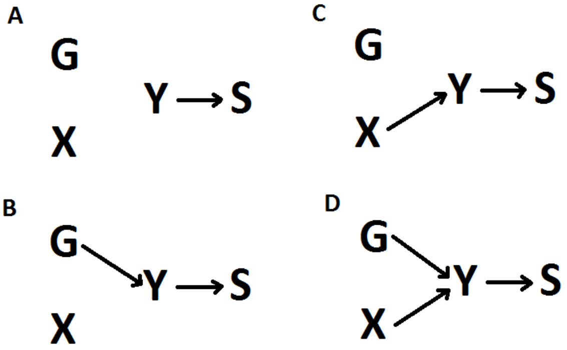


Figure 3.2: Directed acyclic graphs representing possible relationships between the ascertainment (S), the primary phenotype (Y), the secondary phenotype (X) and the genetic marker (G)

Here we will consider is our recently developed method based on a retrospective likelihood approach and joint modeling of primary and secondary phenotype (Tissier et al., 2017) for the proband design. In addition we will consider using the conditional distribution of the secondary phenotype values of the families given the secondary phenotypes of the probands. This approach was recently applied in two papers (Greenwood et al., 2007; Turetsky et al., 2015) for the proband design. They used the method implemented in the software SOLAR (Almasy and Blangero, 1998). However the method in SOLAR was developed for the analysis of primary quantitative phenotypes in families with a proband design. We will show that this approach is not appropriate for secondary phenotypes. Another complication is that when data on the probands are not available, conditioning on the outcome of the probands is not possible. We will compare the performance of these two approaches to the naive approach.

The rest of the paper is organized as follows, in the next section we present the strategies to correct for the ascertainment. We present a simulation study to assess the strength of possible bias of parameter estimes as function of the correlation between primary and secondary phenotype for different scenarios. We finally illustrate the methods by analyzing Electroencephalography (EEG) data from the social anxiety disorder (SAD) family study from the Leiden Family Lab (famlab: https://www.leidenfamilylab.nl) which recruited families based on having at least two members (one parent and one child) of the same family affected by SAD (Bas-Hoogendam et al., 2018).

3.2 Methods

First we introduce some notation. Let N be the total number of families in the study. For family i ($i = 1 \dots N$) of size n_i, let Y_i, X_i and G_i be the $n_i \times 1$ vectors for the case-control status, the secondary phenotype and the genotype, respectively.

3.2.1 Naive approach: ignoring sampling

If we ignore the ascertainment, estimates of the parameters modelling the effect of genetic markers G on a phenotype of interest X can be obtained by maximizing the prospective likelihood L:

$$L = P(X \mid G) = \prod_i P(X_i \mid G_i),$$

with $P(X_i \mid G_i)$ the conditional probability of the phenotype X given the covariate G in family i. This probabilitly might be modeled by a multivariate linear mixed model where $X_i \sim \mathcal{N}_{n_i}(\mu_X, \Sigma_X)$ as follows:

$$X_i = \beta_0 + \beta_1 G_i + \sigma_{G_X} b_i^X + \sigma_\epsilon^X \epsilon_i^X, \tag{3.1}$$

where $\beta = (\beta_0, \beta_1)$ denotes the regression coefficient vector with β_0 the intercept and β_1 the parameter representing the effect of the genotype on X, $b_i^X \sim N_{n_i}(0, \mathbf{R}_i)$ the random effect which models the genetic correlation structure of the secondary phenotype within each family, and σ_ϵ the residual standard deviation. The heritability can be estimated by the proportion of the genetic variance in the phenotypic variance:

$$H^2 = \frac{\sigma^2_{G_X}}{\sigma^2_{G_X} + \sigma^2_\epsilon}. \tag{3.2}$$

3.2.2 Joint modeling under retrospective likelihood

Next we consider to use the retrospective instead of the prosepective likelihood. For a family i the retrospective likelihood can be written as follows:

$$P\left(X_i, G_i \mid Y_i\right) = \frac{P\left(X_i, Y_i \mid G_i\right) P\left(G_i\right)}{P\left(Y_i\right)} = \frac{P\left(X_i, Y_i \mid G_i\right) P\left(G_i\right)}{\sum_{g \in G} P\left(Y_i \mid g\right) P\left(g\right)}.$$

This approach is advantageous because it implictly corrects for the ascertainment process, because in the proband design the selection depends only on the primary phenotype. To estimate the effect of the genetic factor G on X, we need to model the joint conditional probability $P\left(X_i, Y_i \mid G_i\right)$. To deal with the mixture of binary and quantitative outcome variables, a multivariate mixed probit model is proposed. Let $b_i = (b_{i1}, \dots, b_{in_i})^T$ be a set of family specific random effects which model the familial genetic correlation and let $G_i = (g_{i1}, \dots, g_{in_i})^T$ be the vector of genotypes for family i. For the probit model, the observed response Y is modelled as a censored observation from an underlying continuous latent variable Y^*. For Y^* a mixed-effects regression model is used

$$Y_i^* = \alpha_0 + \alpha_1 G_i + \sigma_{G_Y} b_i^Y + \sigma \epsilon_i^Y, \tag{3.3}$$

where $\epsilon_i^Y \sim N_{n_i}(0, I_{n_i})$ is independent of b_i^Y. Here $\alpha = (\alpha_0, \alpha_1)$ denotes the regression coefficient vector with α_0 the intercept and α_1 the parameter representing the effect of the genotype on Y. At the family level we assume $b_i^Y \sim N_{n_i}(0, \mathbf{R}_i)$, with $\mathbf{R}_i$ the coefficient of relationships matrix with elements $r_{lm} = 2^{-d_{lm}}$ with d_{lm} the genetic distance between subjects l and m in the family. The parameter σ_{G_Y} represents the residual additive genetic variatio, i.e. the variation which is not explained by the observed genotype g_{ij}. For identifiability reasons σ is fixed to 1. Now the conditional probability to have the disease $\pi_{ij} = P(Y_{ij} = 1 \mid b_{ij}, g_{ij})$ on g_{ij} is modeled as follows:

$$P\left(Y_{ij} = 1 \mid g_{ij}, b_{ij}^Y\right) = \Phi\left(\alpha_0 + \alpha_1 g_{ij} + \sigma_{G_Y} b_{ij}^Y\right),$$

with $\Phi\left(z\right)$ the cumulative distribution function of the standard normal distribution. The marginal density under the probit model takes the form:

$$f(y_{ij} \mid g_{ij}; \alpha, \sigma_{G_Y}) = \int_{b_i^Y} \int_{\gamma_{y_{ij}}}^{\gamma_{y_{ij}}+1} f(y_{ij}^* \mid g_{ij}, b_i^Y; \alpha, \sigma_{G_Y}) f(b_i^Y) dy_{ij}^* db_i^Y.$$

To model the secondary phenotype X_i we extend Equation 3.1. Now X and Y are jointly modelled using the model specifications (Equation 3.1 and Equation 3.3).

$$\begin{aligned} Y_i^* &= \alpha_0 + \alpha_1 G_i + \sigma_{G_Y} b_i + \epsilon_i, \\ X_i &= \beta_0 + \beta_1 G_i + \sigma_{G_X} b_i + \sigma_\epsilon^X \epsilon_i, \end{aligned} \tag{3.4}$$

Note that additional random effects can be included in this model to take into account unobserved shared environmental factors. In this paper we are focused on the genetic variability and do not model environmental variation. Let Σ_{X_i} and $\Sigma_{Y_i^*}$ denote the corresponding variance-covariance matrices of the marginal distributions of X_i and Y_i^* and let $\Sigma_{XY_i^*}$ be their covariance. The joint distribution of Y^* and X is then $(Y_i^*, X_i) \backsim \mathcal{N}_{2n_i}\left(\begin{bmatrix} \alpha_0 + \alpha_1 G_i \\ \beta_0 + \beta_1 G_i \end{bmatrix}, \begin{bmatrix} \Sigma_{Y_i^*} & \Sigma_{XY_i^*} \\ \Sigma_{XY_i^*} & \Sigma_{X_i} \end{bmatrix}\right)$.

with the marginal correlation between family members j and j' in family i:

$$\begin{aligned}
cor\left(X_{ij}, X_{ij'}\right) &= \frac{\sigma^2_{G_X} 2^{-d(j,j')}}{\left(\sigma^2_{G_X} + \sigma^2_{\epsilon}\right)} \\
cor\left(Y^*_{ij}, Y^*_{ij'}\right) &= \frac{2^{-d(j,j')}\sigma^2_{G_Y}}{\left(\sigma^2_{G_Y} + 1\right)} \\
cor\left(X_{ij}, Y^*_{ij}\right) &= \frac{\rho\sigma_{G_X}\sigma_{G_Y}}{\sqrt{\left(\sigma^2_{G_X} + \sigma^2_{\epsilon}\right)\left(\sigma^2_{G_Y} + 1\right)}} = \rho_{XY} \\
cor\left(X_{ij}, Y^*_{ij'}\right) &= \frac{2^{-d(j,j')}\rho\sigma_{G_X}\sigma_{G_Y}}{\sqrt{\left(\sigma^2_{G_X} + \sigma^2_{\epsilon}\right)\left(\sigma^2_{G_Y} + 1\right)}},
\end{aligned}$$

where ρ represents the genetic correlation between the primary and secondary phenotypes. The heritability estimates of the secondary phenotype which quantifies the proportion of genetic variation in the total variance is obtained by:

$$H^2 = \frac{\sigma^2_{G_X}}{\left(\sigma^2_{G_X} + \sigma^2_{\epsilon}\right)}. \tag{3.5}$$

All the parameters of the models are estimated by maximization of the log likelihood function.

3.2.3 Conditioning on probands

This approach is developed for analysis of quantitative primary phenotypes in a proband design and implemented in the SOLAR software and FISHER. The method is described by Hopper and Mathews (1982), Beaty and Liang (1987) and de Andrade and Amos (2000). Corrections for ascertainment are made by conditioning the likelihood for each pedigree on the trait values of the pedigree probands. Let Asc be the ascertainment process and let Y^P and G^P be the primary phenotype and the genotype values for the probands, respectively. The prospective likelihood L_i for family i can be written as follows:

$$L_i = P\left(Y_i \mid G_i, Asc\right) = \frac{P\left(Y_i, G_i, Asc\right)}{P(G_i, Asc)} = \frac{P\left(Asc|Y_i, G_i\right)P\left(Y_i|G_i\right)}{P(Asc \mid G_i)},$$

By assuming that the ascertainment is complete and depends only on Y_i^P and not on the covariates or Y^{NP}, the values for the non-proband family members, we have $P(Asc|Y_i, G_i) = 1$. Furthermore, $P(Asc \mid G_i) = P(Y_i^P \mid G_i) = P(Y_i^P \mid G_i^P)$. Hence L_i can be written as follows:

$$L_i = \frac{P\left(Y_i|G_i\right)}{P(Y_i^P \mid G_i^P)},$$

The phenotype Y_i is assumed to follow a multivariate normal distribution, $Y_i \sim \mathcal{N}_{n_i}(\mu, \Sigma_Y)$, with:

$$Y_i = \alpha_0 + \alpha_1 G_i + \sigma_{G_X} b_i^Y + \sigma_\epsilon^Y \epsilon,$$

where $\beta = (\beta_0, \beta_1)$ denotes the regression coefficient vector with β_0 the intercept and β_1 the parameter representing the effect of the genotype on Y_i, $b_i^Y \sim \mathcal{N}_{n_i}(0, \mathbf{R}_i)$ is the random effect which models the genetic correlation structure within each family for the secondary trait, and σ_ϵ is the residual standard deviation. Several studies have used the same approach for analysis of secondary phenotypes. Specificaly let X_i^P be the value of the secondary phenotype for the proband(s) of family i, then the following likelihood is used (3.2.3):

$$L_i = \frac{P\left(X_i | G_i\right)}{P(X_i^P \mid G_i^P)},$$

where X_i follows a multivariate normal distribution $X_i \sim \mathcal{N}_{n_i}(\mu, \Sigma_X)$ which is modeled by the model of the naive approach (Equation 3.1). However, Asc is not based on X^P but on Y^P, hence conditioning on X^P might not be an appropriate way to correct for the ascertainment. Therefore this approach might provide biased estimates.

To overcome this issue we propose to model jointly X_i and Y_i in order to take into account the existing correlation between the primary and secondary phenotypes and to use the conditional distribution given the proband(s) values X_i^P and Y_i^P. We have:

$$L_i = P(X_i, Y_i \mid X_i^P, G_i, Y_i^P) = \frac{P\left(Y_i, X_i, X_i^P, G_i, Y_i^P\right)}{P(X_i^P, G_i, Y_i^P)} = \frac{P\left(Y_i, X_i \mid G_i\right) P(G_i)}{P(X_i^P, Y_i^P \mid G_i) P(G_i)}$$

$$= \frac{P\left(Y_i, X_i \mid G_i\right)}{P(X_i^P, Y_i^P \mid G_i^P)},$$

where the joint distribution of X_i and Y_i can be modeled as proposed in Eq 3.4 for the retrospective likelihood approach. Estimates of the parameters are obtained by maximizing the likelihood function:

$$L = \prod_i L_i = \prod_i \frac{P\left(Y_i, X_i \mid G_i\right)}{P(X_i^P, Y_i^P \mid G_i^P)}.$$

3.3 Simulation study

3.3.1 Simulation Setup

A simulation study was set up to evaluate the performance of the described methods. The methods were compared in terms of bias and 95% coverage probabilities of the estimated parameters using data from proband designs. For simplicity we used continuous primary and secondary phenotypes. Probands had a value for the primary phenotype above the 90^{th} quantile in the population. We considered families with five siblings.

Families which were included for analysis had either one or two probands. As genetic biomarker a single nucleotide polymorphism (SNP) with an additive effect on both outcomes was simulated.

The following joint distribution of the primary and secondary phenotypes X and Y for a family was used:

$$(Y_i, X_i \mid G_i) \backsim \mathcal{N}_{2n_i}\left(\begin{bmatrix} \alpha_0 + \alpha_1 G_i \\ \beta_0 + \beta_1 G_i \end{bmatrix}, \begin{bmatrix} \Sigma_{Y_i} & \Sigma_{XY_i} \\ \Sigma_{XY_i} & \Sigma_{X_i} \end{bmatrix}\right)$$

with as covariances between the secondary phenotypes X and the primary phenotypes Y for family members j and j':

$$\begin{aligned} cov(X_{ij}, Y_{ij}) &= \rho\sigma_{G_X}\sigma_{G_Y} \\ cov(X_{ij}, Y_{ij'}) &= 2^{-d(j,j')}\rho\sigma_{G_X}\sigma_{G_Y}. \end{aligned}$$

Here, $d(j, j')$ is the degree of kinship between family members j and j' and ρ is the correlation between the two phenotypes. For simplification of notation let $Z_i = (Y_i, X_i \mid G_i)$. Then the distribution of Z can be written as follows:

$$(Z_i) = \left(Z_i^{NP}, Z_i^{P}\right) \backsim \mathcal{N}_{2n_i}\left(\begin{bmatrix} \mu_{Z_i^{NP}} \\ \mu_{Z_i^{P}} \end{bmatrix}, \begin{bmatrix} \Sigma_{Z_i^{NP}} & \Sigma_{Z_i^{NP}Z_i^{P}} \\ \Sigma_{Z_i^{P}Z_i^{NP}} & \Sigma_{Z_i^{P}} \end{bmatrix}\right)$$

with P and NP representing the proband and other family members, respectively. Now we first simulated the phenotypes of the probands Z_i^P by generating samples from the multivariate normal distribution and keeping only the extreme values of Y_i^P (values larger than the 90^th percentile. Let z_i^P be the realization of Z_i^P, then phenotypes for the family members of the proband(s) were simulated by using the conditional multivariabe distribution of $(Z_i^{NP} \mid Z_i^P)$:

$$\begin{aligned} \mu_{(Z_i^{NP}|Z_i^{P})} &= \mu_{Z_i^{NP}} + \Sigma_{Z_i^{NP}Z_i^{P}}\Sigma^{-1}_{Z_i^{P}}(z_i^P - \mu_{Z_i^{P}}) \\ \Sigma_{(Z_i^{NP}|Z_i^{P})} &= \Sigma_{Z_i^{NP}} - \Sigma_{Z_i^{NP}Z_i^{P}}\Sigma^{-1}_{Z_i^{P}}\Sigma_{Z_i^{P}Z_i^{NP}} \end{aligned}$$

A minor allele frequency of 0.3 was used to simulate the SNP genotypes for the parents and the genotypes for the offspring were simulated under Mendels first Law. For the secondary phenotype, the following fixed effects values were used: $\beta_0 = 2$ and $\beta_1 = 0.5$, whereas for the primary phenotype model the effect sizes were $\alpha_0 = 5$ and $\alpha_1 = 0.5$.. The heritability of both phenotypes was fixed to 50%. Specifically, we used σ_{G_X}=σ_{ϵ_X}=0.5 and $\sigma_{G_Y} = \sigma_{\epsilon_Y}$=1. For the correlation between the phenotypes, we varied ρ from 0 to 1 with steps of 0.1.

For study designs with ascertainment based on one and on two probands and for each ρ, 500 replicates were generated. Each replicate was considered twice namely all family members were used in the analysis and only the non-proband family members were analyzed.

3.3.2 Simulation Results

Figure 3.3 presents the means of the estimates of the SNP effect on the secondary phenotype (top panel) and of the heritability of the secondary phenotype (bottom panel) across the replicates. In the left panel the estimates are based on the datasets with one proband, while the right panel corresponds to datasets with two probands. With regards to the estimates of the effect of the SNP on the secondary phenotype, the retrospective likelihood approach provides unbiased estimates for both types of ascertainments. Conditioning on the value of the secondary phenotype of the proband(s) provides only unbiased estimates when the correlation between the primary and secondary phenotypes is one. When the correlation decreases the estimated effect size of the SNP decreases from $\beta_1 = 0.5$ for a correlation of one to $\beta_1 = 0.091$ and $\beta_1 = 0.079$ for a correlation ρ of zero, and for one and two probands, respectively. Conditioning on the values of both phenotypes of the proband provides unbiased estimates for ρ equal to zero, $\beta_1 = 0.495$ and $\beta_1 = 0.491$ for one and two probands, respectively. However, when ρ increases this approach underestimates the effect sizes namely $\beta_1 = 0.452$ and $\beta_1 = 0.427$ for one and two probands, respectively and $\rho = 1$.

Finally, the naive approach provides constant estimates almost independent of the correlation between X and Y. The estimates are $\beta_1 \approx 0.37$ and $\beta_1 \approx 0.27$ for selection schemes based on one or two probands. For all approaches the bias of the estimates increases with the strength of the ascertainment.

With regards to the heritability estimates, just as for the SNP effects the retrospective likelihood provides unbiased estimates. For the naive approach the bias of the estimates of the heritability appears to depend on the correlation ρ. For ρ between 0 and 0.6 the estimates are almost unbiased but after 0.6 they start to decrease to $h^2 = 0.410$ and $h^2 = 0.356$ for $\rho = 1$ for families which are selected based on one proband and two probands, respectively. This suggests an impact of the sampling process on the variance of X. As the correlation between X and Y increases the variance of the distribution of X in the families decreases. The same phenomenon is observed for the approaches which condition on the outcome(s) of the probands, an unbiased estimate of the heritability is obtained when both phenotypes have a small correlation (between 0 and 0.3) while the estimate decreases to $h^2 = 0.309$ for a one-proband design and to $h^2 = 0.236$ for a two-proband design when conditioning on the proband's secondary phenotypes and $\rho = 1$. For conditioning on both phenotypes of the probands the estimates of the SNP effect and the heritability are unbiased except for values of ρ close to one where both estimators show a small bias.

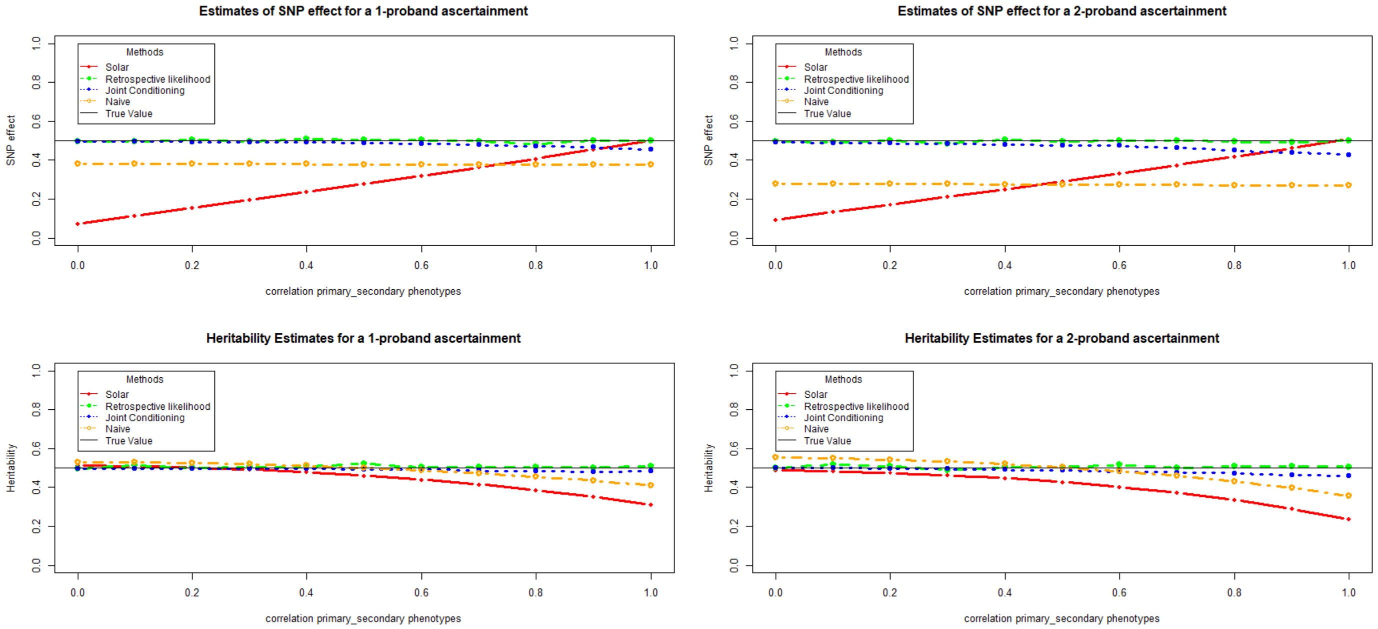


Figure 3.3: SNP effect (top) and Heritability (bottom) estimates for a proband design provided by SOLAR, the retrospective likelihood and the naive approach. Estimates are provided for various correlation between the primary phenotype and the secondary phenotype and for the two different ascertainments, namely 1 proband (left panel) or 2 probands (right panel). The black line represent the true value of the heritability and the SNP effect

Figure 3.4 presents the simulations results for the situation where there is no data available for the probands. Note that for this situation we can only apply the retrospective likelihood and the naive approach. In terms of SNP effects, both approaches provide similar results for both ascertainment schedules. For a one-proband ascertainment the retrospective likelihood provides estimates with a small bias which increases with ρ, namely from $\beta_1 = 0.467$ to $\beta_1 = 0.459$ while the naive approach obtains a SNP effect estimation from $\beta_1 = 0.459$ to $\beta_1 = 0.442$. For a two-proband design the bias is larger, the values reduces to $\beta_1 = 0.430$ and to $\beta_1 = 0.405$ for the retrospective and naive approach, respectively. For the heritability, the retrospective likelihood slightly overestimates the heritability with $h^2 \approx 0.54$ for 2 probands. For the naive approach, we observe a similar behavior as for the families which includes the phenotypes of the probands: the approach provides unbiased estimates for a small correlation between X and Y, while the bias increases when the correlation becomes larger: the heritability estimates reduces to $h^2 = 0.403$ for a one-proband design and to $h^2 = 0.312$ for a two-proband design for $\rho = 1$.

The estimate of the SNP effect on the secondary phenotype by the naive approach is slightly better when the information on the proband is missing. However, the heritability estimates provided by the naive approach are worse than for the scenario where the probands are included in the study. For the retrospective likelihood, the lack of information on the probands leads to bias for both SNP effect and heritability. However, the bias remains relatively small.

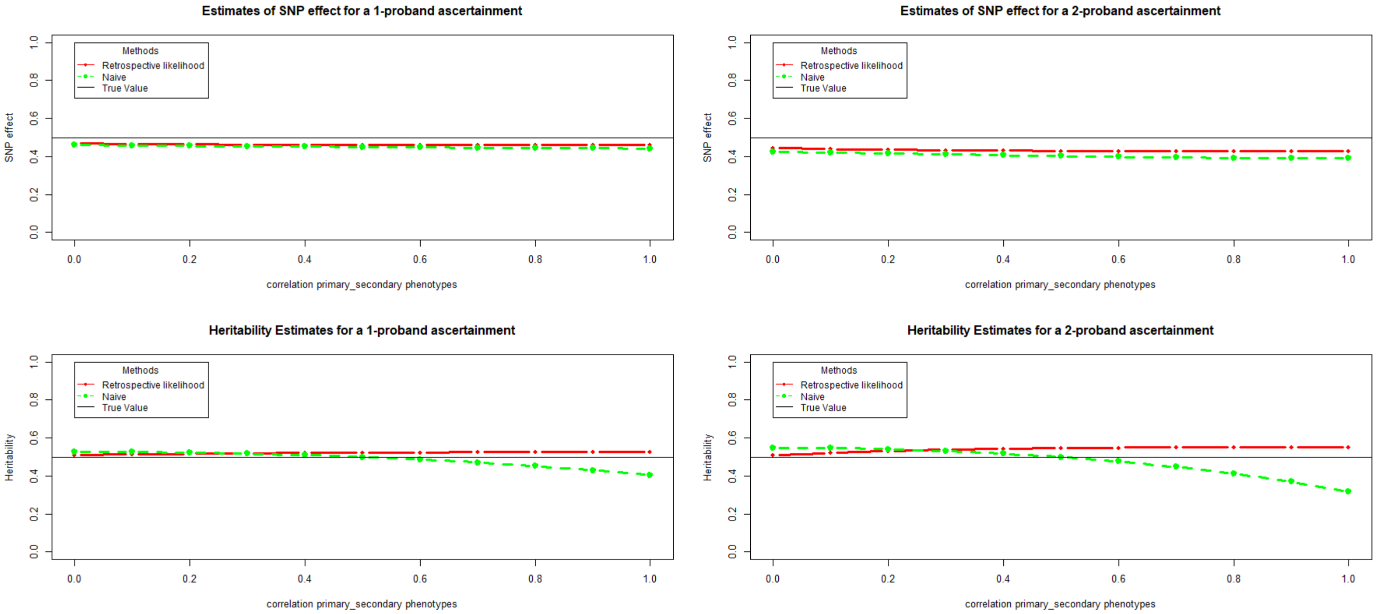


Figure 3.4: SNP effect (top) and heritability (bottom) estimates for a proband design, where probands are not in the study, provided by the retrospective likelihood and the naive approach. Estimates are provided for various correlation between the primary phenotype and the secondary phenotype and for the two different ascertainment, namely 1 proband (left panel) or 2 probands (right panel). The black line represent the true value for the heritability and SNP effect.

As the heritability provides only information about the proportion of genetic variance in the total variance of a phenotype, we investigated the 95% coverage probabilities for the genetic variance of X. Table 3.1 summarizes the results. the retrospective likelihood approach has a stable coverage probability between 90.9% and 96.8% accross all different scenarios. When conditioning on the secondary phenotype only, the coverage probability is $\geq 90\%$ for a correlation ρ between 0.2 and 0.5 for a 1-proband ascertainment and between 0 and 0.3 for 2-proband ascertainment rule. The coverage probability drastically shrinks, when ρ increases, to 3.2% and 0% when families are selected based on one or two probands, respectively. The naive approach provides, overall, better coverage probabilities than conditioning on the phenotypes of the probands. The performance of the naive approach is strongly impacted by the ascertainment strength. Finally, the conditioning on the values of the primary and secondary phenotypes provides coverage probabilities $\geq 90\%$ only for $\rho \leq 0.4$ and $\rho \leq 0.3$ for 1 and 2 probands ascertainment, respectively.

Study design	With Probands				Without Probands	
Correlation (X,Y)	Retrospective likelihood	Solar	Naive	Joint	Retrospective likelihood	Naive
1 proband						
0.0	93.1	87.0	86.9	94.3	93.5	90.6
0.1	91.6	88.4	88.4	93.2	91.5	91.2
0.2	95.4	91.4	89.8	95.4	91.0	92.2
0.3	92.8	94.0	91.1	92.4	89.6	93.4
0.4	95.8	95.2	94.8	91.0	89.8	93.8
0.5	95.8	95.0	94.5	88.4	88.9	94.0
0.6	95.7	89.4	94.3	81.0	87.6	93.6
0.7	93.9	73.2	87.3	76.7	88.8	89.2
0.8	95.7	47.2	79.4	70.4	86.9	81.8
0.9	93.9	17.3	62.4	62.9	87.4	67.6
1.0	93.7	3.2	39.6	56.9	87.8	44.4
2 probands						
0.0	94.7	94.0	71.9	92.8	95.3	85.6
0.1	90.9	94.0	74.5	91.4	93.1	85.2
0.2	96.8	94.2	80.8	91.6	89.2	87.4
0.3	96.4	93.2	87.9	87.3	85.9	90.6
0.4	95.4	89.0	92.4	79.5	84.1	93.4
0.5	92.5	84.0	96.0	74.0	82.5	94.8
0.6	94.0	67.6	93.2	66.8	80.4	92.0
0.7	95.0	43.6	80.8	61.9	78.4	84.0
0.8	92.5	17.0	59.2	59.1	77.9	66.2
0.9	91.1	2.6	23.8	48.1	78.2	33.8
1.0	96.4	0.0	3.8	41.8	79.4	11.2

Table 3.1: 95% Coverage probabilites of the genetic variance of the secondary phenotype, $\sigma^2_{G_X}$, for the various scenarios.

3.4 The social anxiety disorder study

Recently we used the retrospective likelihood approach and SOLAR to analyze data from a proband design where data on probands were missing (Fuady et al., 2018). Here we focus on the analysis of various secondary phenotypes in a family study with the proband design where data for the probands are available. The primary phenotype is social anxiety disorder (SAD), which is a psychiatric disorder which is characterized by extreme anxiety and avoidance in one or more social situations (Association., 2013). SAD is a common and debilitating internalizing disorder (Furmark, 2002). The risk for developing SAD is higher for individuals with a close family member with SAD than for individuals without family members with SAD (Isomura et al., 2015), and estimates of the heritability of SAD vary from 20 % to 56 % (Distel et al., 2008; Kendler et al., 1992; Isomura et al., 2015). The goal of the Leiden Family Lab study on Social Anxiety Disorder (LFLSAD, Bas-Hoogendam et al. (2018)) is to investigate whether behavioral and electrocortical responses to social evaluation are candidate endophenotypes of SAD. Families were selected based on two probands: one adult with SAD (target participant) and his/her child with clinical or subclinical SAD. The final version of the dataset included 138 participants from nine extended families. To measure electroencephalography (EEG) activity during a stressful social situation, participants had to perform a task without knowing it beforehand to avoid anticipatory stress. EEG measures were done at four different time points: before knowing they had to perform the speech (Resting state 1), before doing the task (Anticipation), maximum three minutes after the task (Recuperation), and after thirty minutes after completion of the task (Resting state 2). More details about the procedure can be found in Harrewijn et al. (2016). Here we analyze the EEG measure of the correlation between the delta and beta wave of the brain. This correlation between cerebral waves was measured on three 3 different frequency bands, namely 14-20 Hz (cor delta low beta), 20-30 Hz (cor delta high beta), and 14-30Hz (cor delta beta), leading to three measures of the correlation between delta waves and beta waves of the brain for four time points. We applied the retrospectively likelihood approach and the approach which conditions on the secondary phenotype values of the probands.

In the first columns of Table 3.2 for each secondary phenotype, the correlation between the secondary and the primary phenotype SAD and the estimates and p-values of the parameter modelling the association between the various correlations between beta and delta waves and subclinical SAD are given. The latter estimates were obtained from the following linear mixed effect model:

$$X_i = \beta_0 + \beta_1 SAD + \beta_Z Z + \sigma_{G_X} b_i^X + \sigma_\epsilon \epsilon_i^X,$$

with Z vector of covariates to adjust for sex, age and age^2. The random effects $b_i^X \sim N_{n_i}(0, \mathbf{R}_i)$ model the genetic correlation structure within each family. We used the package *coxme* in R to fit this model. Only the correlation between the beta and delta waves at low frequency in the anticipation period, was found to be associated with SAD, with an effect size of -0.071 for an estimated correlation of 0.394 with SAD.

Table 3.2 presents, also, the estimates for the heritability obtained using the three described methods (retrospective likelihood, conditioning on secondary phenoty of the proband, and conditioning on both phenotypes of the proband). The heritability was computed using equation 3.5, presented in section 3.2. Just as in the simulations, the heritability estimates obtained by the retrospective likelihood are always higher than the heritability estimates obtained by SOLAR or by conditioning on both phenotypes of the probands. As expected the differences between heritability estimates across the approaches is related to the correlation between the primary phenotype SAD and the secondary phenotypes, the EEG variables. Indeed the largest identified difference in heritability estimates corresponds to a secondary phenotype highly correlated with SAD ,while the smallest difference between the heritability estimates coincides with the smallest correlation between the EEG variables and SAD. With the largest difference in heritability estimates between the retrospective likelihood and SOLAR is 0.223 (with $h^2 = 0.580$ and $h^2 = 0.357$, respectively) which corresponds to the only statistically significant associated secondary phenotype. This phenotype also has the strongest correlation with SAD. As expected conditioning on both phenotypes of the probands provides for this variable an estimate of the heritability between the estimates of the retrospective likelihood and SOLAR namelh $h^2 = 0.532$. Finally, we observe that when the heritability estimates obtained by the retrospective likelihood approach and SOLAR are small, the heritability estimates which is obtained by conditioning on both proband's pheonotypes are even smaller.

EEG variable	genetic correlation SAD-EEG	β_1(p-value)	h^2 SOLAR	h^2 Retrospective	h^2 joint conditioning
Resting State 1					
cor delta beta	0.066	.032(.47)	.075	.124	.039
cor delta low beta	0.042	.025(.53)	.086	.087	.084
cor delta high beta	0.102	.039(.30)	.067	.118	.015
Anticipation					
cor delta beta	0.266	-.053(.17)	.289	.391	.374
cor delta low beta	0.394	-.071(.01)	.357	.580	.532
cor delta high beta	0.161	-.028(.48)	.397	.422	.427
Recuperation					
cor delta beta	0.009	-.019(.73)	.165	.187	.098
cor delta low beta	0.004	-.003(.94)	.107	.111	.045
cor delta high beta	0.040	-.027(.60)	.301	.330	.040
Resting State 2					
cor delta beta	0.114	.061(.21)	.303	.356	.305
cor delta low beta	0.186	.057(.19)	.200	.336	.216
cor delta high beta	0.085	.041(.38)	.169	.215	.200

Table 3.2: Results of the data analysis. Estimates of the association between the secondary phenotypes and subclinical SAD, and estimates of the heritability of the secondary phenotypes obtained by the three considered approaches

3.5 Discussion

In this paper, we studied the effect of incorrect ascertainment corrections on estimates of model parameters for secondary phenotypes as function of the correlation between primary and secondary phenotypes. We illustrate the methods with data from the social anxiety disorder study from the Leiden Family Lab.

For a proband design where data for the proband(s) are available and analyzed, simulations results showed that the retrospective likelihood performs well and the estimates of the effect of the SNP on the secondary phenotype and of the heritability of the secondary phenotype are both unbiased. In contrast conditioning on the secondary phenotype values of the probands might yield biased results. This could be expected since the families are selected based on the primary and not on the secondary phenotype. We observed that SNP effects are underestimated when the secondary phenotype has a small correlation with the primary phenotype. When the correlation between the primary and secondary phenotype is large the bias shrinks and finally disappears when the correlation is equal to 1. Note that in our simulations we chose the primary and secondary phenotypes to be continuous, in case of binary primary phenotype it is not possible to achieve a correlation of 1 with a continuous secondary phenotype. Therefore, estimates will always be biased. An explanation for the fact that there is no bias for $\rho = 1$ might be that conditioning on the values of the secondary phenotypes of the proband is equivalent to conditioning on the primary phenotype when the correlation between the primary and the secondary phenotype is large. Furthermore, the bias observed for both estimates is enhanced by the strength of the ascertainment. It is well known that this approach performs less well when the number of probands increases (Boehnke and Greenberg, 2018). While conditioning on the phenotypes of the proband has proved to be an efficient and fast way to study primary phenotypes in a proband design, it is not recommended to use it for the analysis of secondary phenotypes in this design.

Conditioning on the primary and secondary phenotypes of the probands appears to perform better than only conditioning on the secondary phenotype. This approach provides on average less biased estimates than just conditioning on the secondary phenotype. If the averaged estimate over the 500 datasets is close from the true value, the coverage probabilities provided by this approach are small especially when ρ 0.3. Although not always performing well this approach has two advantages: 1) when the primary phenotype is binary, this approach is computationally efficient compared to the retrospective likelihood, especially for large pedigrees. The reason is that the dimension of the integral in the denominator is smaller; 2) for quantitative primary phenotypes, the approach is implemented in the SOLAR software as the analysis of multiple continuous phenotypes can performed. Finally, the naive approach, as expected, was unable to provide proper SNP effects and heritability estimates. The estimates were only unbiased when the correlation between primary and secondary phenotypes was small.

In studies where data on the probands are not available, the retrospective likelihood provided biased results both for the heritability and SNP effect. Although this bias appears to be small. Apparently, information on the phenotypes of the probands is essential to correct for ascertainment via these probands. For this situation the naive approach appears to perform relatively well and estimates of SNP effects are quite similar to the estimates obtained from the retrospective likelihood. The estimates of the heritability appear to be unbiased when the correlation between the phenotypes is not too large. Therefore, the naive approach might be a good alternative for the retrospective likelihood when pedigrees are large and the retrospective approach is computational too demanding.

The results of the estimation of the heritability of EEG phenotypes in families with social anxiety disorder agreed with the results obtained in the simulations. The variable which showed a statistically significant association with subclinical social anxiety disorder has a larger heritability when the retrospective likelihood approach compared to other approaches is used. This variable could be an appropriate candidate endophenotype for the social anxiety disorder. Due to the lack of genetic data in the study, we were unable to investigate the difference between conditioning on the phenotypes of the probands and the retrospective likelihood approach for fixed effects.

We are currently running more simulations to obtain a better understanding of the underestimation of the heritability when the correlation between primary and secondary phenotypes is high when we are only conditioning on the secondary phenotype. Finally, the methods will be implemented in a R package for the conditioning on the joint distribution of both primary and secondary phenotypes. This approach can be computationally less challenging than the retrospective likelihood approach when pedigrees are large while providing mostly unbiased results.

4

Gene co-expression network analysis for family studies based on a meta-analytic approach

Abstract

For a better understanding of the biological mechanisms involved in complex traits or diseases, networks are often useful tools in genetic studies: coexpression networks based on pairwise correlations between genes are commonly used. In case of a family-based design, it can be problematic when there is a large between-family variation in expression levels. We propose here a gene coexpression network analysis for family studies. We build a coexpression network for each family and then combine the results. We applied our approach to data provided for analysis in the Genetic Analysis Workshop 19 and compared it to 2 naive approaches-ignoring correlations among the expressions and decorrelating the gene expression by using the residuals of a mixed model and a single-probe analysis. Our approach seemed to better deal with heterogeneity with regard to the naive approaches. The naive approaches did not provide any significant results, while

---

This chapter has been published as: Renaud Tissier, Hae-Won Uh, Erik van den Akker, Brunilda Balliu, Spiridoula Tsonaka, Jeanine J. Houwing-Dusitermaat (2016). Gene co-expression network analysis for family studies based on a meta-analytic approach. *BMC Proceedings*; 10(Suppl 7): 119-123.

our approach detected genes via indirect effects. It also detected more genes than the single-probe analysis.

4.1 Background

Weighted gene co-expression network is a widely used method for studying biological networks based on pairwise correlations. This method provides more insight in the underlying biological mechanisms and offers a tool for dimension reduction by summarizing identified modules (clusters) of genes (Plaisier et al., 2009; de Jong et al., 2012). How to perform such an analysis for family data is an open question. For family data Kraft et al. (2003) noted that testing association between expression levels and traits without taking into account the family structure can lead to spurious results, especially when the number of families is small and in the presence of large between-family variation. In this paper, we propose a novel strategy for network analyses in a small set of relatively large families. For this family-based approach, we first construct family-specific co-expression networks and test for association between the modules and the traits of interest. Common set of genes for all families were obtained by using the intersection and the union of family specific modules. We compare this family-based approach with two naive approaches: namely, one using the gene expression of the families directly (ignoring correlation) and one that first decorrelates the gene expressions and then applies the standard approach. We also compare our results with single probe analyses.

4.2 Methods

4.2.1 Study sample

The gene-expression dataset is composed of 647 individuals from 17 large families. These samples are from the dataset described in Almasy and Blangero (1998). Here, we focus on the largest 5 families: namely family 2, 5, 6, 8 and 10 with 65, 55, 45, 62 and 49 family members, respectively. The total number of individuals is 276. In total gene expressions of 20634 probes are available. We used the simulated quantitative phenotypes Systolic Blood Pressure (SBP) and the phenotype Q1 at time point 1 as outcome variables. The simulation model of SBP comprises 15 genes and that of Q1 does not contain any of these genes. SBP, Q1 and all probes were corrected for age and sex by regressing out covariates and using residuals.

In order to decorrelate the gene expressions, we fitted for each probe a linear mixed model: $X_{ij} = \ + u_{ij} + v_i + {}_{ij}$, with X_{ij} the value of the probe for the individual j in family i, u_{ij} a normally distributed random genetic effect: $u_ij\ N(0, S)$ where $S = 2 * K * s_g$ with K kinship matrix and $s_g$genetic variance, v_i a normally distributed random effect representing shared environmental effects, and ${}_{ij}$ a normally distributed residual. To obtain the residuals X^*_{ij} of this model we used the function lmekin, which fits linear mixed models with specific structure of the variance-covariance matrix from the package coxme (Therneau, 2018) in R.

4.2.2 Single probe analysis

For the single probe analysis the following mixed model was used:

$$Y_ij = +u_ij + v_i + X_ij + _ij$$

with Y_{ij} the value of SBP or Q1 and X_{ij} the value of the probe for individual j of family i. The random effects u_{ij}, v_i and $_{ij}$ are the genetic effect, the shared environmental effect and residuals respectively. The parameter β represents the effect of the probe on the outcome variable.

4.2.3 Network constructions

Co-expression networks were built on the dataset without correction for family structure based on X_ij (naive approach), the dataset adjusted for family structure based on X_ij^* (naive decorrelated approach), and on the datasets from the five families separately.

We used signed co-expression networks. The adjacency matrix $A = [a_{lk}]$ of each network was computed as follows: $a_{lk} = |0.5 + 0.5cor(x_l, x_k)|^\gamma$, with $cor(x_l, x_k)$ the correlation between x_l the values vector of probe l and x_k the values vector of probe k. The parameter γ is acting as a soft threshold in the adjacency matrix, when we increase the value γ the coefficient of the adjacency matrix will tend to 0 except for values really close to 1. We used the biweight midcorrelation based on the median, which is more robust than the Pearson correlation. The co-expression networks were constructed with the R package WGCNA (Langfelder and Horvath, 2008). For each obtained module, the first principal component (eigengene) was computed.

4.2.4 Phenotype analysis

From all modules and all families, the following models were fitted:

$$Y_j = +u_j + \beta eigengene_j^k + _j,$$

where Y_j is the outcome, u_j the random genetic effect and $eigengene_j^k$ the value of the eigengene of module k of family member j. Let E_{F2}^M to E_{F10}^M be the most significant eigenvalues of the family specific networks (N_F2 to N_F10) and let E_F^M be the most significant eigenvalue of these five eigenvalues and M_F^M be the corresponding module. Identify the modules of the family-specific networks, which have the highest overlap with M_F^M (denoted as M_{F2}^O to M_{F10}^O). Next, two common sets of genes for all families were obtained by taking the intersection ($M_F = M_{F2}^O \cap M_{F5}^O \cap M_{F6}^O \cap M_{F8}^O \cap M_{F10}^O$) and the union ($M_F = M_{F2}^O \cup M_{F5}^O \cup M_{F6}^O \cup M_{F8}^O \cup M_{F10}^O$) of the family specific modules. The first principal components of the two common sets were computed. The principal component that explained most of the variance of the corresponding set of genes was used as the eigengene EF of the family based approach.

The eigengenes of the naive approach (EN), the naive approach after decorelation (END) and the family-based approach (EF) are tested for association with the two phenotypes SBP and Q1. Here, the following mixed model was used:

$$Y_{ij} = \ + u_{ij} + v_i + \beta eigengene^k_{ij} + {}_{ij}$$

with Y_{ij} the phenotype value for individual j of family i and $eigengene^k_{ij}$ the value of eigengene of module k of individual j of family i. And u_{ij}, v_i and ${}_{ij}$ are the genetic effect, the shared environmental effect and residuals respectively. The parameter β represents the effect of the eigengene k on the outcome variable.

Finally since spurious associations are especially expected in the presence of large between family heterogeneity (Kraft et al., 2003) we also performed a network analysis using the subset of 25% most heritable probes when performing the network analysis (n=4911 probes with heritability between 0.33 and 0.88).

To test for significance we used a nominal alpha level of 0.05 and the Bonferroni correction was applied to take into account multiple testing.

4.3 Results

4.3.1 Results obtained with all probes

For per family analysis, the module that showed the highest correlation with the SBP was the magenta module obtained in family 8 (M^M_{F8}) (β=2.52, p=0.0011). M^M_{F8} comprises 710 genes. For each family, the number of genes of the module with the highest overlap is given in Table 4.1. The intersection and the union of these five family modules, comprises 62 and 1746 probes respectively. The first principal component (eigengene) of the probes in the intersection set explained more than 50% of the variance for each family, while for the union set the eigengene explained only between 23% and 31% of the variance of the expression levels. Therefore the eigengene of the intersection set was used as summary for the family approach (E_F). In Table 4.2, for each family the effect of E_{Fi} on SBP (β of model (2)) is given. For families 2 and 8, the eigengenes (E_{F2} and E_{F8}) were significantly associated with SBP.

When analysing all families together none of the approaches provided significant results. The joint analysis of the families using EF as eigengene in model (3) did not provide a significant association SBP (β=-0.13, p=0.49). For the naive approach, the eigengene of the module magenta (E_N) had the smallest p-value (β=-3.21, p=0.01). For the naive approach using the decorrelated dataset, the eigengene of the module grey60 (E_{ND}) had the smallest p-value (β=-3.03, p=0.0061). After multiple-testing correction (between 43 and 50 modules in each network) none of the results were significant. Finally the single

	M^O_{F2}	M^O_{F5}	M^O_{F6}	M^O_{F8}	M^O_{F10}
Module size	446	694	499	710	446
Size of the overlap with M^M_F	187	308	240	710	372

Table 4.1: Module size of M^O_{F2} to M^O_{F10} and overlap size with M^M_F in the all-probes analysis

	All probes		25% most heritable probes	
	SBP	Q1	SBP	Q1
E_{F2}	-0.57(0.2) (0.02)^a	9.90(4.3)(0.02)	0.27(0.1)(0.07)	-1.62(1.0)(0.11)
E_{F5}	0.34(0.2)(0.21)	14.0(4.7)(3.3e-3)	0.18(0.2)(0.41)	-2.13(1.3)(0.11)
E_{F6}	0.08(0.3)(0.78)	8.90(3.7)(0.02)	0.66(0.3)(0.01)^a	2.49(0.9)(9.6e-3)
E_{F8}	-0.62(0.3)(0.04)^a	10.47(4.2)(0.01)	0.07 (0.2)(0.68)	2.47(1.0)(0.02)
E_{F10}	0.14(0.3)(0.67)	7.55(4.5)(0.09)	0.02(0.2)(0.91)	-2.22(1.2)(0.06)
E_F	-0.13(0.2)(0.49)	-	0.21(0.09)(0.02)^a	-
E_N	-3.21(1.3)(0.01)	2.75(0.7)(5.6e-4)^a	1.93(0.8)(0.01)	-0.96(0.5)(0.06)
E_{ND}	-3.03(1.1)(6.1e-3)	-1.41(0.4)(9.4e-4)^a	1.94(0.7)(5e-3)^a	-0.41(0.2)(0.06)

Table 4.2: Parameter estimates of the association between eigengenes and Q1 and SBP. In parentheses are standard errors and p values, respectively. For E_{F2} to E_{F10} model (2) was used; for E_F, E_N and E_{ND} model (3) was used. For Q1 the association results for E^M_{F2} to E^M_{F10} are presented. ^{a} Denotes significant test after multiple testing corrections.

probe analysis preformed in the five families by using model (1) provided one significantly associated probe with SBP (CRIP2; β= -13.68, p= 1.7e-06).

The intersection module of the family based approach did not contain any of the 15 genes used for the simulation. Also the identified gene of the single probe analysis is not among these 15 genes. We hypothesized that correlation might exist between E_{F2}, E_{F8}, and the gene expression of CRIP2 on one hand and the set of 15 genes on the other hand. Indeed E_{F2} showed significant correlation with PSMD5 (p=0.004) and GTF2IRD1 (p=0.007) and E_{F8} showed significant correlation with ZNF443 (p=5e-05), PSMD5 (p=3e-05) and ABTB1 (p=6e-05). When the presence of these 15 genes in the modules was investigated, it appeared that they were in different modules (see Table 4.3). The gene CRIP2 which was significant in the single probe analysis showed significant correlation with the gene KRTAP11-1 (p= 3.1e-03).

4.3.2 Analysis of Q1

The results of the analysis of Q1 are also given in Table 4.2. For the family approach, none of the modules obtained in family-specific network analysis was significantly asso-

	N_N	N_{ND}	N_{F2}	N_{F5}	N_{F6}	N_{F8}	N_{F10}
MAP4	-	-	-	-	-	7	-
NRF1	-	-	-	1	-	-	-
TNN	11	-	19	5	14	-	-
LEPR	-	1	19	-	-	-	-
FLT3	5	-	-	8	4	-	1
GTF2IRD1	-	4	13	3	-	-	-
FLNB	9	-	16	21	13	-	2
ZNF443	8	-	5	1	23	6	1
GSN	2	15	3	-	-	1	-
CABP2	11	-	-	5	14	2	16
LRP8	-	-	6	-	-	12	-
PSMD5	3	1	18	10	28	17	-
GAB2	20	15	1	3	22	-	5
ABTB1	3	-	4	4	1	2	2
KRTAP11-1	4	19	2	1	18	4	1

Table 4.3: List of the top genes involved in the simulation model and their module number in each network. -, Denotes the grey module in which all nonclustered genes are combined. The different colours represent genes in the same module for a specific network

ciated with Q1 and no common set could be defined. In table 4.2 the estimates of strongest associated modules E_F^M for each family are given. For the naive approach, the module red (β=2.75, p=0.00056) was significant and for the naive approach using the decorrelated data the module green (β=-1.41, p=0.00094) was significantly associated with Q1.

4.3.3 Results obtained with the 25% most heritable probes

For the naive and the family approaches, the results of the network based analyses using only the gene expressions of the 25% most heritable probes (n=4911 probes with heritability between 0.33 and 0.88) are also given in Table 4.2. None of the 15 genes used in the simulation model for SBP was among this set of most heritable probes. For Family 6 the E_{F6} was significantly associated with SBP (p=0.01). The association of E_F in the five families with SBP was also significant (p=0.02). For Q1 none of the approaches provided significant results. With regard to the single probe analysis, no other probes than CRIP2 was significantly associated.

4.4 Discussion

In this paper, we have proposed a novel strategy to perform a co-expression network analysis with family data: building a network for each of the large pedigrees, and defining a common module by taking the intersection of family specific modules. We compared our family-based approach with two naive network approaches and a single probe analysis. All analyses were performed in a small set of five relatively large families. None of the 15 genes in the simulation model was identified in this small dataset. However the family-based approach identified significant associations between the eigengene and SBP in two families. This eigengene was significantly correlated with 4 of the 15 genes. When analyzing all families jointly the family eigengene was not significant. Also the naive network approaches did not provide any significant result. The single probe analysis provided one significant gene which was correlated with one of the 15 genes. To study the performance of the methods with regard to false positive findings, we also analyzed the trait Q1 for which no gene expressions were included in the simulation model. The family approach did not provide a significant finding, while both naive approaches identified a significant module for Q1. The result in the naive approach based on gene expression (X_{ij}) is in line with the findings of Kraft et al. (2003). We did not expect to have a false positive finding when using the decorrelated data (X^*_{ij}) as input for our network analysis. Possible explanations for this finding are the fact that the correlation based on the kinship coefficient might not be appropriate for gene expressions, and randomness. In addition to the set of all probes, also networks were built using only the 25% most heritable probes. Especially for these variables that show large between-family variation spurious associations might occur when the family structure is not taken into account. This was not confirmed in our analysis. More research is needed to study the sensitivity of the methods for between-family variation.

We did not know the answers when we developed the family-based approach and analyzed the data. The simulation model used to create the datasets may not be well suited to pick up the 15 genes directly by network analysis. The 15 genes present in the model were in different pathways: they were not correlated. Moreover our approach was able to identify indirect effects: i.e. the eigengenes were correlated with the 15 genes. Thus the significant association of the family based network approach represented the largest number of genes from the simulation model. We expect that especially in the presence of large between-family variation our approach would perform best. A thorough simulation study is required to investigate the performance of our method further.

Network analysis provides a tool to reduce the number of tests by first summarizing the data in sets of genes with correlated gene expressions and summarizing the gene set by the first principle component. Another obvious reduction step is to only consider the heritable probes for the analysis. It appeared that by using the heritable probes the results across the families were less heterogeneous. The family approach as well as the naive approach using decorrelated data provided significant results for SBP. In this paper we combined the family-specific modules by taking the intersection of the modules which showed most overlap. This approach worked well for the relatively small set of five families. When we applied our method to the six largest families, similar results were obtained (data not shown). However intersection might not be the most appropriate approach to combine modules across families, because the intersection set becomes too small. Alternative approaches have to be developed. For example lasso type of methods can be used to select probes from the union sets. Development of a method for constructing a common set from the family specific modules is a topic for future research. Finally more research is needed to evaluate the performance of our method with regard to false positive and false negative findings in relationship to heterogeneity, family size, the number of families and the heritability of gene expressions.

5

Improving stability of prediction models based on correlated omics data by using network approaches

Abstract

Building prediction models based on complex omics datasets such as transcriptomics, proteomics, metabolomics remains a challenge in bioinformatics and biostatistics. Regularized regression techniques are typically used to deal with the high dimensionality of these datasets. However, due to the presence of correlation in the datasets, it is difficult to select the best model and application of these methods yields unstable results. We propose a novel strategy for model selection where the obtained models also perform well in terms of overall predictability. Several three step approaches are considered, where the steps are 1) network construction, 2) clustering to empirically derive modules or pathways, and 3) building a prediction model incorporating the information on the modules. For the first step, we use weighted correlation networks and Gaussian graphical modelling. Identification of groups of features is performed by hierarchical clustering. The grouping

---

This chapter has been published as: Renaud Tissier, Jeanine J. Houwing-Duistermaat, Mar Rodríguez-Girondo (2018). Improving stability of prediction models based on correlated omics data by using network approaches. *PLoS One* 13(2):e0192853.

information is included in the prediction model by using group-based variable selection or group-specific penalization. We compare the performance of our new approaches with standard regularized regression via simulations. Based on these results we provide recommendations for selecting a strategy for building a prediction model given the specific goal of the analysis and the sizes of the datasets. Finally we illustrate the advantages of our approach by application of the methodology to two problems, namely prediction of body mass index in the DIetary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome study (DILGOM) and prediction of response of each breast cancer cell line to treatment with specific drugs using a breast cancer cell lines pharmacogenomics dataset.

5.1 Introduction

The advent of the omic era in biomedical research has led to the availability of an increasing number of omics measurements representing various biological levels. Omics datasets (e.g. genomics, methylomics, proteomics, metabolomics, and glycomics) are measured to provide insight in biological mechanisms. In addition, new predictions models can be built based on omics predictors. Omic data are typically high-dimensional (i.e. $n < p$, n sample size and p the number of variables) and they present unknown dependence structures reflecting various biological pathways, co-regulation, biological similarity or coordinated functions of groups of features. Since traditional regression methods have been developed for low-dimensional settings only, they are too restrictive and hence unable to deal with omic datasets and to determine the actual role of their various components. As a result, an important methodological challenge in omic research is how to incorporate these complex datasets in prediction models for health outcomes of interest. This paper is motivated by the previous work of Rodríguez-Girondo et al. (2018) in which we showed that metabolomics were predictive of future Body Mass index (BMI) using data from the DIetary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome study (DILGOM)(Inouye et al., 2010). However, when we tried to identify the important metabolites, using lasso regression for variable selection in a cross-validation framework, we obtained inconsistent effect sizes and variable selection frequencies. Specifically, metabolites with largest effects were not always selected and highly correlated variables presented different selection frequencies. These results inspired us to develop more stable prediction models by using network methods.

To obtain a good balance between stability and predictive ability, we propose to incorporate information on the structure between features from an omics dataset into predictions models for health outcomes. The incorporation of such a structure in prediction models is a relatively new and expanding strategy in prediction models. For classification problems methods have been developed, such as the partial correlation coefficient matrix (PPCM) method (Rao and Lakshminarayanan, 2007), network-based support vector machines (Zhu et al., 2009), or the selection protein-protein interactions discriminative subnetworks (Chuang et al., 2007). In this paper we focus on the prediction on continuous outcomes. Also several methods have been developed for this type of outcomes.

Zhang and Horvath (2005), and Reis (2013) have proposed to identify clusters of related variables inside the network and to include a summary measure of these clusters, namely principal components and partial least squares. While these approaches provide good results in terms of prediction accuracy, one of their major drawbacks is the chosen summary measures which are hard to interpret and replicate. An alternative approach is network penalization as proposed by Li and Li (2008), using the laplacian matrix of the network matrix to build a lasso-type penalization in order to force the effect sizes of variables related to each other in the network to be similar. However, it is relatively heavy in terms of computations and therefore not able to handle too large datasets. Winter et al. (2012) proposed to first rank variables based on their univariate association with the outcome and their relationships between each other and then use the top ranked variables in a prediction model. While this approach can provide good predictions in some settings, it depends on various tuning parameters and therefore reproducibility is a challenge. Recently, network-based boosting methods (Shim et al., 2017) and combination of network-based boosting and kernel approaches (Friedrichs et al., 2017) have been proposed to improve prediction models for GWAs and gene expression studies. These methods include known relationships between genetic markers and phenotypes of interest in order to detect new genetic-phenotypes relationship and therefore improve prediction models. However, for some omic type of data, such as metabolomics and transcriptomics, our lack of knowledge limits the application of these methods only to certain omic sources such as genomics.

In this paper, we propose a flexible approach allowing investigators to apply several types of network analysis approaches to estimate the structure of the data as well as several possible group-penalizations methods. Namely, our approach consists of three steps (Fig5.1): network analysis (to empirically derive relations within an omic dataset), clustering (to empirically establish groups of omic related features) and predictive modeling using the aforementioned grouping structure (via group-based variable reduction or group-penalization). This strategy allows a lot of flexibility in terms of both network analysis and prediction models, as different type of omics data have different properties and might need different network analysis strategies or prediction models to obtain proper and biologically relevant results. Finally, to avoid overoptimism in absence of an external validation set, a common situation in omic research, cross-validation of the whole three-step procedure is used.

The rest of the paper is organized as follows: we present the various methods involved in our three-step approach. An intensive simulation study is then presented to empirically evaluate the performance of the various studied methods in terms of predictive ability and variable selection properties. Standard regularized regression methods such as lasso, ridge and elastic net are also considered. The methods are applied to two sets of omic sources (metabolomics and transcriptomics) measured at baseline for the prediction of BMI after seven years of follow-up using DILGOM and on gene expression to predict treatment response from the publicly available breast cancer cell line pharmacogenomics dataset (https://genomeinterpretation.org/content/breast-cancer-cell-line-pharmacogenomics-dataset). In the last section, the results are discussed and

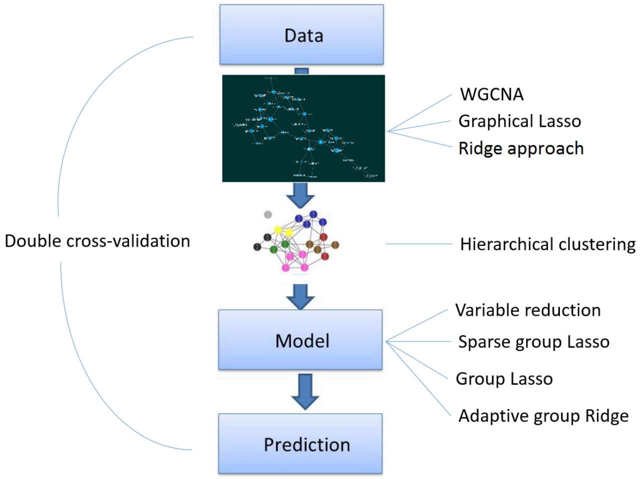


Figure 5.1: Method summary. Step 1: Networks of features are derived from the data. Step 2: Using hierarchical clustering, modules of features are identified. Step 3: Prediction models are derived using grouping information from Step 2.

concluding remarks are provided.

5.2 Methods

A common approach to build prediction models in high-dimensional settings or in presence of strong correlation between features is regularized regression (Hastie et al., 2009), which has shown to have good properties in terms of predictive ability in various omic settings (Ghosh and Chinnaiyan, 2005; Zemmour et al., 2015; Shahabi et al., 2006; Pena et al., 2016). The choice of the shrinkage type imposes certain constrains in the estimated parameters which can lead to unstability or to models which are difficult to interpret. The lasso approach (Tibshirani, 1996) introduces a l_1-norm constrain of the vector β of regression coefficients and shrinks some of the regression coefficients towards zero, introducing sparsity by only selecting ‘the most important variables’ in the model. In the presence of (groups of) correlated features, lasso penalization appears not to perform well in terms of stability since it tends to randomly choose among the strongly correlated features and can select at most n variables before saturation. Alternatively, ridge regression (Hoerl and Kennard, 1970) considers a l_2-norm constrain of the regression coefficients, which does not allow for explicit variable selection but typically handles well strong correlations. Still, these ridge models are difficult to interpret since sparsity is not obtained. Alternative penalizations as elastic net (Zou and Hastie, 2005) have been proposed to overcome limitations of lasso and ridge regression, producing sparse models but also allowing to select more than n correlated variables.

In the rest of this paper, let the observed data be given by $(\mathbf{y}, \mathbf{X})$, where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathbf{T}}$ is the continuous outcome measured in n independent individuals and $\mathbf{X}$ is a matrix of dimension $n \times p$, representing an omic predictor source with p features. We propose a three-step approach (Fig 5.1) to get an interpretable prediction model for $\mathbf{y}$ based on $\mathbf{X}$, where $\mathbf{X}$ is high-dimensional $(p > n)$. In the first step, we estimate the intensity matrix (network) of $\mathbf{X}$, which contains the degree of relation among the features of $\mathbf{X}$. We investigate three different techniques for network estimation: weighted gene co-expression network analysis (WGCNA, Zhang and Horvath (2005)), where the relationship is based on Pearson correlation, and two proposals based on gaussian graphical modeling (Lauritzen, 1996), where the relationship is given by the precision matrix. Here two different penalization methods are considered. Namely, ridge (Ha and Sun, 2014) and lasso (Friedman et al., 2007). In the second step, we identify modules (groups) of features by applying hierarchical clustering to the dissimilarity matrix obtained from the estimated network of Step 1. The grouping information is incorporated in the prediction model. Here we consider two strategies: group-based variable reduction and group-penalization. In the variable reduction approach, 'hubs' in each group are identified, i.e. variables with the strongest connectivity within a module, and then included in a standard regression. Group penalization, such as adaptive group ridge (van de Wiel et al., 2014), group lasso (Yuan and Lin, 2006), and sparse group lasso (Simon et al., 2013), penalizes the features from the same module jointly. Finally, double cross-validation (Rodríguez-Girondo et al., 2018; Mertens et al., 2006, 2011) was applied, over all steps, to obtain proper tuning parameters and summary performance measures in absence of an external validation set.

5.2.1 Step 1: Network construction

A network is, by definition, an adjacency matrix $\mathbf{A} = [a_{ij}]$, where a_{ij} is either an indicator of presence of connection (edge) between two features (nodes) x_i and x_j or a value between 0 and 1 which represents how close the two nodes are. We focus on the latter case because of its continuous nature, and we refer to the resulting networks as weighted networks.

WGCNA

Co-expression networks based on pairwise correlations have been proposed in the context of analyzing gene expression data Zhang and Horvath (2005). Due to the presence of many correlated gene expression data, a parameter β (soft threshold) is introduced in order to shrink "low" pairwise correlation values towards zero. The parameter β might be chosen in such a way that the free-scale topology criterion holds, i.e, the fraction of nodes with k edges should follow the power law $P(k) \approx k^{-\gamma}$, with $P(k)$ the fraction of nodes in the network with k edges and γ a constant with a value comprised between 2 and

3. The rationale behind the free scale topology criterion relies on maximizing the within cluster connectivity while minimizing the between cluster connectivity.

Co-expression networks have been successfully used in the context of transcriptomics (Oldham et al., 2006, 2008; Stuart et al., 2003). A drawback of the approach is that the soft thresholding does not provide a sparse network as none of the correlation coefficients is set to zero. In some omic settings, such as metabolomics and glycomics where correlations are high the network might be too dense to interpret. This limitation has motivated the use of alternative approaches such as Gaussian graphical models based on partial correlations which are, by definition, more sparse.

Gaussian Graphical Modeling

Partial correlation coefficients represent the pairwise correlation between two variables conditional on all other variables. Thus the linear effects of all other variables are removed and association is based on the remaining signals. The use of partial correlations appears to provide sparser and more biologically relevant networks compared to networks based on Pearson correlation (Krumsiek et al., 2011; Schäfer and Strimmer, 2005).

In the low-dimensional setting ($p < n$) the partial correlation matrix is straightforward estimated as $\mathbf{P} = -scale(\mathbf{S}^{-1}) = -diag\,(\mathbf{S})^{-\frac{1}{2}}\,\mathbf{S}diag\,(\mathbf{S})^{-\frac{1}{2}}$, where $\mathbf{S}$ is the sample variance-covariance matrix.

However, note that the calculation of partial correlations relies on the inversion of the sample variance-covariance matrix, which is challenging (or impossible) in case of strong collinearity between variables or in high-dimensional ($p > n$) situations. To overcome this difficulty, several authors have considered penalizing the covariance matrix in order to invert it. In this work, we focus on two methods namely a ridge-type (Ha and Sun, 2014) and a lasso-type penalty (Friedman et al., 2007).

Ridge-penalty approach Ha and Sun (2014) proposed a method to obtain a sparse partial correlation matrix, based on a ridge-type penalty to invert the variance-covariance matrix. Specifically, let $\mathbf{S}$ be the empirical variance-covariance matrix. To deal with singularity of $\mathbf{S}$ due to collinearity or high-dimension a positive constant to the diagonal elements of $\mathbf{S}$ is added, $\mathbf{S}' = \mathbf{S} + \lambda\mathbf{I_p}$. For any $\lambda > 0$, $\mathbf{S}'$ has full rank. The partial correlation matrix $\mathbf{R}$ is estimated as follows:

$$\hat{\mathbf{R}} = -scale\left(\mathbf{S}'^{-1}\right)$$

When the penalty parameter λ goes to infinity, the partial correlation matrix is shrunk towards the identity matrix. To obtain a sparse matrix, it is tested whether each coefficient r_{ij} is significantly different from zero by applying a Fisher's z-transformation (Fisher, 1924) on the partial correlation estimates and assuming that these transformations follow a mixture of null and alternative hypotheses. Efron's central matching method (Efron, 2004) allows to estimate the null distribution of this test statistic by approximating the mixture distribution using polynomial Poisson regression.Thus, p-values can be computed for each estimated partial correlation r_{ij}, and a sparse network (if r_{ij} not significant, r_{ij} is set to zero) is obtained.

Lasso approach An alternative penalization method is to apply a lasso-type penalty when estimating the inverse of the estimated variance-covariance matrix (Friedman et al., 2007). Assume that we have n multivariate normal observations of dimension p, with mean vector μ and variance-covariance matrix $\mathbf{\Sigma}$. To estimate $\mathbf{S}$ the following penalized log-likelihood has to be maximized:

$$L(\Theta) = log\,(det\,(\Theta)) - trace\,(\mathbf{S}\Theta) - \lambda||\Theta||_1$$

with $\Theta = \mathbf{\Sigma}^{-1}$. The optimal tuning parameter λ is determined by minimizing the AIC ($AIC = n \times tr\,(S\Theta) - log\,(det\,(\Theta)) + 2E$) with E the number of non-zero elements in Θ. Note that, especially for small values of the penalty parameter, the resulting partial correlation matrix is not exactly symmetric. Symmetry can be imposed by duplicating one of the estimated triangular matrices (upper or lower).

5.2.2 Step 2: Hierarchical clustering

Hierarchical clustering is used to detect groups of related features from the estimated network which was obtained with the methods introduced in the previous section.

Specifically, we have used the dynamic tree cut algorithm based on the dendogram obtained by hierarchical clustering (Langfelder et al., 2008). This is an adaptive and iterative process of cluster decomposition and combination until the number of clusters becomes stable. This approach, in contrast to a constant height cut-off method, is capable of identifying nested clusters and is implemented in the R package WGCNA. The measure used for the hierarchical clustering aproach was the topological overlap dissimilarity measure. The topological overlap of two nodes quantifies their similarity in terms of the commonality of the nodes they connect (Yip and Horvath, 2007) and is given by:

$$TOM_{ij} = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{min\,(k_i, k_j) + 1 - a_{ij}}$$

with a_{ij} the weight between i and j in the adjacency matrix, and $k_i = \sum_u a_{iu}$. The topological overlap dissimilarity measure is now defined as : $dissTOM_{ij} = 1 - TOM_{ij}$.

5.2.3 Step 3: Outcome prediction

Finally, we incorporate the obtained grouping information in the prediction models. One of the major challenges in prediction using high dimensional data is to avoid overfitting. Overfitting occurs when a model is too complex, i.e when it has too many parameters. We used two of the most standard approaches for parameter reduction which are a priori variable reduction based on variable importance and shrinkage methods. Namely, we consider within-group variable selection and regularized regression models with group penalization. In general, regularized regression models are characterized by the optimization problem $\min_{\beta \in R^p} \left(\|y - \sum X\beta\|_2^2 + R(\beta) \right)$ where $R(\beta)$ is the regularization or penalty term. Examples of commonly used penalization functions are: $R(\beta) = \lambda \sum_j |\beta_j|$ (lasso; Tibshirani (1996)), $R(\beta) = \lambda \sum_j \beta_j^2$ (ridge; Hoerl and Kennard (1970)) and $R(\beta) = \alpha \sum_j \beta_j^2 + (1-\alpha) \sum_j |\beta_j|\ \alpha \in (0,1)$ (elastic net; Zou and Hastie (2005)).

Variable importance

The general idea of this simple approach is to retain the most relevant (according to some pre-defined criterion) variables from each of the estimated groups obtained by hierarchical clustering in step 2. We propose to consider only the most strongly connected variables within its group (‘hubs’), assuming that strong connectivity is indicative of biological importance and hence relevance to predict the outcome of interest. Specifically, for a specific group G:

$$hub_G = \max_i \left(\sum_{j \in G} \mathbf{I}_{a_{ij} \neq 0} \right)$$

with a_{ij} the ij element of the adjacency matrix. If multiple nodes have the same maximum, all these hubs are selected. Ridge regression is used to deal with collinearity in case of several selected hubs.

Group penalization

An alternative to within-cluster variable selection is to consider cluster-based penalties in the context of regularized regression.

Group lasso Group lasso (Yuan and Lin, 2006) selects groups of variables since it simultaneously shrinks all the coefficients belonging to the same group towards zero. The group lasso estimator is given by:

$$\min_{\beta \in R^p} \left(\left\| y - \sum_{l=1}^{L} X_l \beta_l \right\|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l} \, \|\beta_l\|_2 \right)$$

where $l \in (1 \cdots L)$ represents the index of the group of predictors, L is the the total number of clusters, X_l is the matrix of predictors in the group l and $\sqrt{p_l}$ is a penalty to take into account the varying group size. The tuning parameter λ is made by cross-validation based on minimization of the AIC. The group lasso estimator is asymptotically consistent even when model complexity increases. Note that if each group contains just one variable, group lasso is equivalent to the standard lasso (Tibshirani, 1996).

Sparse group lasso Sparse group lasso (Simon et al., 2013) can be applied when one also wish to select variables within a group. Shrinkage is carried out at the group level and at the level of the individual features, resulting in the selection of important groups as well as members of those groups. The sparse group lasso estimator is given by:

$$\min_{\beta \in R^p} \left(\left\| y - \sum_{l=1}^{L} X_l \beta_l \right\|_2^2 + (1-\alpha)\lambda \sum_{l=1}^{L} \sqrt{p_l} \left\| \beta_l \right\|_2 + \alpha\lambda \left\| \beta \right\|_1 \right)$$

where l, X_l, $\sqrt{p_l}$ and are defined as in group lasso. Note that the sparse group lasso is a combination of group lasso and lasso. The parameter α regulates the weight of each approach. For $\alpha = 1$ sparse group lasso equals lasso and for $\alpha = 0$ group lasso.

Adaptive group-regularized ridge regression Finally, the recently proposed adaptive group ridge approach van de Wiel et al. (2014) which extends ridge regularized regression to group penalization is considered. The adaptive group ridge considers group specific penalties λ_l for the L groups. The adaptive group ridge estimator is given by:

$$\min_{\beta \in R^p} \left(\left\| y - \sum_{l=1}^{L} X_l \beta_l \right\|_2^2 + \sum_{l=1}^{L} \lambda_l \sum_{q \in G_l} \beta_q^2 \right)$$

where l and X_l are defined as in group lasso, G_l is the lth group of variables and λ_l is the penalty term for the group G_l. The penalty terms can be expressed as: $\lambda_l = \lambda_l' \lambda$ with λ a unique penalty term and λ_l' as penalty multipliers for each group.

5.2.4 Software implementation

The proposed three-step approach has been implemented in the `R` function PredNet which is available at github (https://github.com/RenTissier/NetPred). The function allows to apply all the possible combinations of the previously presented network analysis and group penalization methods. The function calls the packages `WGCNA` (co-expression based on pairwise correlation), `huge` (gaussian graphical modeling), `GGMridge` (ridge-penalty approach), `grpreg` (group lasso), `SGL` (sparse group lasso), and `GRridge` (adaptive group-regularized ridge regression).

5.3 Simulation Study

5.3.1 Simulation setup

An intensive simulation study was conducted to study the performance of our proposed prediction methods using estimated grouping information and to compare them wit9h existing regularized regression methods (without grouping information), such as lasso, ridge and elastic net ($\alpha = 0.5$). We also included the special case of 'known clustering', in which we assume that the true underlying grouping structure is known, mimicking the situation in which information on biological clustering is available from previous analyses or open source pathway databases. The omic predictor $\mathbf{X}$ is simulated from a zero-mean multivariate normal distribution with correlation matrix $\mathbf{\Sigma}$. Following the recent literature on pathway and network analysis of omics data (Zhang and Horvath, 2005), we generated $\mathbf{\Sigma}$ according to a hub observation model with added realistic noise (Hardin et al., 2013).

The continuous outcome $\mathbf{y}$ is generated by $\mathbf{y} = \mathbf{X}\beta + \epsilon$, where β is the vector of regression coefficient of size p, and $\epsilon \sim N(0,1)$. The singular value decomposition (svd; Jolliffe (2008)) of $\mathbf{X}$, $\mathbf{X} = \mathbf{UDU}^t$ allows to generate $\mathbf{y}$ in terms of the various latent modules present in $\mathbf{X}$ since they represent different independent subspaces of features accounting for different proportions of variation in $\mathbf{X}$. In practice, we first generate β^*, the regression coefficients corresponding to each independent module (given by $\mathbf{U}$), and we then transform it to the predictor space by using $\beta = \mathbf{U^t}\beta^*$.

Within this general framework, we consider three different scenarios: (Scenario a) $\beta_j^* = 0.01, j = 1$; $\beta_j^* = 0, j \neq 1$. $\mathbf{y}$ is then associated to a high variance subspace of $\mathbf{U}$, corresponding to the largest eigenvalue of $\mathbf{X}$. (Scenario b) $\beta_j^* = 0.01$, $j = 4$; $\beta_j^* = 0$, $j \neq 4$. The association with $\mathbf{y}$ relies on a low-variance subspace of $\mathbf{U}$. Hence, we expect lower predictive ability of $\mathbf{X}$ compared to Scenario a. (Scenario c) $\beta_j^* = 0.01$, $j = 1, 4$; $\beta_j^* = 0$, $j \neq 1, 4$. The association with $\mathbf{y}$ relies on several subspaces of $\mathbf{U}$. As a result, Scenario c is less sparse than Scenarios a and b.

For each scenario, we considered two sample sizes ($n = 50$ and $n = 100$), different number of features in $\mathbf{X}$, ($p = 200$ features and $p = 4000$), and different number of underlying modules ($k = 4$ and $k = 8$). Each module presents various within-correlation levels and in all the scenarios, we assumed the presence of one module of uncorrelated variables. Fig 5.2 shows the corresponding heatmaps of Σ for $k = 4$ (left panel) and $k = 8$ (right panel). For each scenario, we generated $M = 500$ replicates and for each trial we consider 10-fold partitions in order to obtain cross-validated summary measures.

We evaluated our methods in terms of obtaining the correct grouping structure, of prediction performance, and variable selection. Grouping is summarized in two ways. On the one hand, we compared the estimated number of groups with the underlying parameter k. On the other hand, for each of the k underlying modules, we calculated the correct and incorrect classification rates (belonging or not belonging to the underlying module taken as reference) of each of the p features. Predictive ability is measured by $Q^2 = \frac{\sum_{i=1}^{n}(p_i - p_{0i})^2}{\sum_{i=1}^{n}(y_i - p_i)^2}$, the cross-validated version of the fraction of variance explained by the prediction model, in which the performance of the model-based is compared to the naive double cross-validated predictions $\mathbf{p_0}$ based on the mean value of the outcome variable $\mathbf{y}$(Rodríguez-Girondo et al., 2018). Variable selection properties are assessed by comparing the simulated β coefficients with the average estimated regression coefficients.

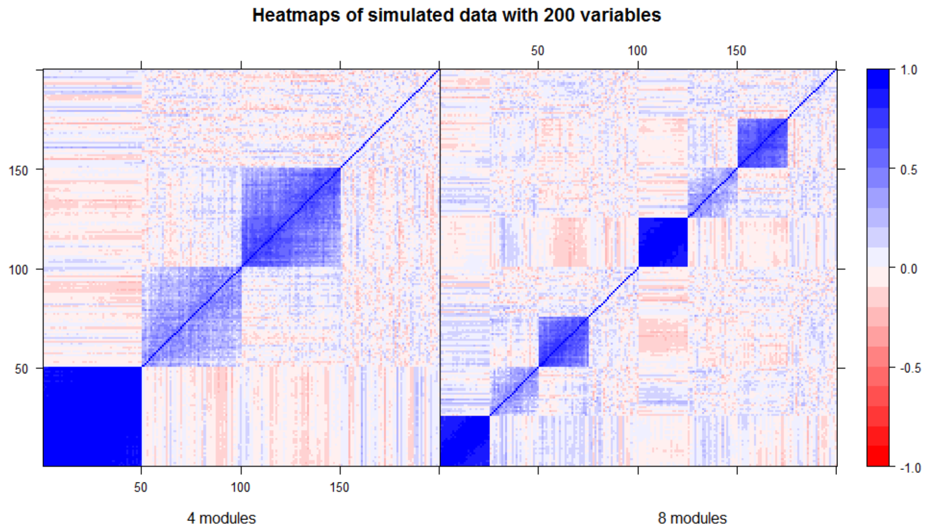


Figure 5.2: Simulation study; correlation matrices. Example of simulated correlation matrices obtained with 200 variables for 4 and 8 modules respectively.

5.3.2 Simulation results

Network analysis and clustering Table 5.1 and 5.2 show the performance of the studied methods for network analysis and hierarchical clustering. WGCNA obtains number of clusters closer to the truth than graphical lasso and the ridge-penalty approach. WGCNA estimates, on average, $\hat{k} = 3$ and $\hat{k} = 5$ for $k = 4$ and $k = 8$ underlying modules, respectively. This slight underestimation of k yields a large number of false positives (see Table 5.2). Focusing on the situation of $k = 4$, and taking the group with highest simulated within correlation as reference, Table 5.2 shows a false positive rate of 38.2% for WGCNA, mainly due to the incorrect assignment of features of the second cluster to the first one. In contrast, graphical lasso overestimates the number of simulated modules.

The number of estimated modules is not affected by the number of underlying modules (for example, $\hat{k} = 14$ for both $k = 4$ and $k = 8$ with $n = 50$), but it increases with the number of p simulated features. This is likely due to the reliance of graphical lasso on partial correlations instead of Pearson correlations. After having a closer look at the estimated modules, we observe that graphical lasso generates $\hat{k}$ groups, which are subsets of the underlying simulated k modules. In other words, graphical lasso does not group together features belonging to different underlying modules (WGCNA does), and the estimated modules can be grouped in such a way that the original k modules are recovered. This translates in a very small false positive rate when taking any of the k simulated modules as reference (see Table 5.2). Finally, the ridge-penalty approach is, in most of the

	200 variables			
	4 modules		**8 modules**	
	n=50	**n=100**	**n=50**	**n=100**
WGCNA	3.1(2-5)	3.0(2-5)	5.0(3-8)	5.0(4-7)
Graphical lasso	14.7(9-21)	17.0(12-23)	14.4(9-21)	17.6(13-25)
Ridge penalty	1.0(1-3)	1.3(1-6)	1.5(1-8)	9.8(1-21)
	1000 variables			
	4 modules		**8 modules**	
	n=50	**n=100**	**n=50**	**n=100**
WGCNA	3.1(2-5)	3.0(2-5)	5.6(4-18)	5.0(4-11)
Graphical lasso	48.3(40-86)	76.5(57-93)	59.6(39-81)	77.5(63-95)
Ridge penalty	10.2(1-71)	52.6(3-72)	13.1(1-69)	61.5(6-81)

Table 5.1: Simulation study. Average number of clusters obtained accross cross-validation by WGCNA, graphical lasso, and ridge penalty. The minimum and maximum number of clusters identified are presented in brackets.

		50 Individuals			100 Individuals		
		TPR	FNR	FPR	TPR	FNR	FPR
module 1	WGCNA	.999	.001	.382	.998	.002	.375
	Graphical lasso	.308	.692	.000	.259	.741	.000
	Ridge penalty	.999	0.001	.997	.962	.038	.951
module 3	WGCNA	.918	.082	.190	.989	.011	.148
	Graphical lasso	.189	.811	.001	.192	.808	.000
	Ridge penalty	.999	.000	.997	.960	.040	.951

Table 5.2: Simulation study. Average (across 10 cross-validation folds and 500 replicates) true positive rate (TPR), false negatives rate (FNR) and false positives rate (FPR) for WGCNA, graphical lasso and ridge penalization. Top part: Scenario a. Reference module: module 1 (corresponding to the first 50 variables in Fig 2 left panel which present the highest level of correlation). Bottom part: Scenario b. Reference module: module 3 (corresponding to the variables 100-150 in Fig 2 left panel).

cases, not able to lead to the identification of any cluster with small number of features and subjects (see $p = 200$ and $n = 50$ in Table 5.1). For larger number of individuals and variables, the number of clusters is overestimated for the same reason as graphical lasso. Namely, the reliance of this method on partial correlations.

Predictive ability

Table 5.3 and Table 5.4 show the results in terms of the predictive accuracy measure Q^2 for $p = 200$ and $n = 50$ and, for $p = 1000$ and $n = 50$ respectively. Table A and Table B in S1 File, show results for $n = 100$. Adaptive group ridge and group lasso present similar performances in most of the studied situations. These two methods outperform the other considered three-step approaches. Also they are the best performing methods when the known grouping was used. Further, these approaches may outperform the commonly used regularized regression methods lasso, ridge and elastic net regression in terms of predictive ability. Specifically, group lasso relying on grouping structure coming from WGCNA and graphical lasso systematically outperforms ridge and lasso and it presents a similar predictive ability than elastic net when $p = 200$. For $p = 1000$ the predictive ability of the standard ridge, lasso and elastic net is lower while the methods based on group lasso and adaptive group ridge present similar behavior than for $p = 200$. Therefore, the gain of these new approaches appears to be larger when the number of predictors increases.

Compared to adaptive group ridge, group lasso was less sensitive to the chosen network method. Namely, all scenarios adaptive group ridge presents bad performance when using the ridge penalty approach Ha and Sun (2014) for network construction. The performance of group lasso is robust with respect to the studied network construction methods in all the studied scenarios, and close to its performance when using the true underlying grouping structure. Sparse group lasso provides proper results in terms of prediction ability when the clustering is known a priori, with Q^2 values only slightly lower than the corresponding values of adaptive group ridge and group lasso. However, when the grouping is estimated, its performance drops. The predictive ability appears to drop to a $Q^2 < 0.1$ for scenario b, which is 8 times lower than the predictive ability obtained with a combination of graphical lasso and group lasso. The variable selection approach based on selecting hubs only provides satisfactory results when using the WGCNA method for network construction in scenario a.

		4 modules			8 modules		
Scenario		a	b	c	a	b	c
A Priori	Sparse group lasso$_{0.5}$	.79(.01)	.51(.06)	.65(0.02)	.75(.02)	.71(.02)	.69(0.03)
	Sparse group lasso$_{0.9}$	.79(.01)	.48(.06)	.59(.03)	.74(.02)	.69(.04)	.65(0.04)
	Sparse group lasso$_{0.1}$	.79(.01)	.53(.06)	.66(.02)	.75(.02)	.72(.02)	.70(0.03)
	Group lasso	.87(.01)	.53(.07)	.77(.02)	.84(.02)	.78(.03)	.81(0.02)
	Group ridge	.94(.01)	.43(.08)	.69(.07)	.90(.02)	.73(.06)	.85(0.03)
WGCNA	Hubs	.81(.03)	.15(.10)	.59(.11)	.81(.05)	.18(.13)	.55(.12)
	Sparse group lasso$_{0.5}$	.72(.12)	.15(.12)	.57(.15)	.41(.21)	.28(.19)	.36(.20)
	Sparse group lasso$_{0.9}$	.73(.13)	.13(.22)	.53(.13)	.41(.22)	.26(.12)	.35(.19)
	Sparse group lasso$_{0.1}$	.69(.12)	.16(.12)	.58(.15)	.39(.20)	.29(.17)	.36(.19)
	Group Lasso	.90(.02)	.58(.07)	.87(.02)	.83(.04)	.76(.06)	.83(.04)
	Group ridge	.78(.03)	.46(.06)	.62(.05)	.69(.07)	.61(.08)	.53(.09)
Graphical lasso	Hubs	.52(.20)	.26(.15)	.51(.18)	.52(.22)	.45(.20)	.51(.22)
	Sparse group lasso$_{0.5}$	.69(.13)	.08(.06)	.45(.16)	.31(.21)	.22(.15)	.27(.18)
	Sparse group lasso$_{0.9}$	.68(.13)	.06(.05)	.42(.16)	.32(.21)	.19(.15)	.26(.17)
	Sparse group lasso$_{0.1}$	.69(.13)	.08(.06)	.46(.16)	.31(.21)	.24(.15)	.28(.18)
	Group lasso	.92(.01)	.54(.08)	.87(.03)	.86(.03)	.76(.06)	.86(.03)
	Group ridge	.93(.02)	.46(.08)	.61(.06)	.85(.08)	.71(.06)	.70(.11)
Ridge penalty	Hubs	.52(.06)	.11(.02)	.47(.06)	.27(.10)	.22(.07)	.27(.09)
	Sparse group lasso$_{0.5}$	.77(.09)	.42(.05)	.67(.02)	.68(.07)	.63(.04)	.67(.04)
	Sparse group lasso$_{0.9}$	.79(.07)	.46(.06)	.61(.03)	.72(.05)	.66(.05)	.65(.05)
	Sparse group lasso$_{0.1}$	.73(.08)	.40(.04)	.68(.02)	.62(.09)	.59(.03)	.63(.04)
	Group lasso	.87(.02)	.48(.06)	.84(.02)	.79(.04)	.71(.05)	.78(.03)
	Group ridge	.67(.05)	.07(.03)	.69(.05)	.47(.06)	.32(.07)	.45(.07)
Common	Lasso	.88(.03)	.52(.10)	.73(.05)	.81(.04)	.74(0.06)	.79(0.05)
	Ridge	.67(.05)	.07(.03)	.59(.06)	.46(.06)	.55(0.04)	.70(0.03)
	Elastic net	.96(.04)	.74(.26)	.79(.20)	.87(.02)	.81(.04)	.89(.02)

Table 5.3: Simulation study. Results obtained in terms of average Q^2 (across 500 replicates) for scenarios a,b,c, p=200 variables, k=4 and k=8 modules, and n=50 individuals. Standard errors are given in brackets. The first column represents the method used to build the network. A Priori represents the situation were the true clustering of the predictors is known and no network analysis is performed.

		4 modules			8 modules		
Scenario		a	b	c	a	b	c
	Sparse group lasso$_{0.5}$	.80(.002)	.64(.02)	.63(.03)	.77(.016)	.69(.036)	.69(.030)
	Sparse group lasso$_{0.9}$	.80(.001)	.56(.036)	.54(.047)	.76(.019)	.62(.047)	.67(.056)
A Priori	Sparse group lasso$_{0.1}$	.80(.002)	.66(.026)	.66(.032)	.77(.016)	.70(.033)	.72(.025)
	Group lasso	.89(.003)	.76(.021)	.71(.046)	.87(.011)	.81(.022)	.84(.016)
	Group ridge	.97(.011)	.65(.076)	.55(.083)	.95(.018)	.87(.033)	.78(.065)
	Hubs	.87(.026)	.48(.12)	.45(.324)	.45(.324)	.13(.127)	.08(.088)
	Sparse group lasso$_{0.5}$	.74(.143)	.61(.098)	.57(.153)	.43(.244)	.36(.206)	.32(.221)
WGCNA	Sparse group lasso$_{0.9}$	.74(.147)	.54(.090)	.53(.138)	.44(.252)	.35(.193)	.29(.223)
	Sparse group lasso$_{0.1}$	.70(.134)	.62(.098)	.58(.155)	.40(.227)	.34(.196)	.32(.205)
	Group lasso	.94(.01)	.85(.031)	.87(.027)	.88(.036)	.79(.043)	.78(.058)
	Group ridge	.80(.037)	.59(.061)	.62(.059)	.70(.067)	.50(.088)	.62(.096)
	Hubs	.52(.054)	.55(.054)	.21(.039)	.42(.059)	.46(.063)	.43(.050)
	Sparse group lasso$_{0.5}$	.79(.032)	.54(.110)	.12(.08)	.46(.251)	.32(.202)	.34(.185)
Graphical lasso	Sparse group lasso$_{0.9}$	.79(.030)	.49(.122)	.09(.075)	.46(.249)	.30(.191)	.30(.195)
	Sparse group lasso$_{0.1}$	.79(.030)	.56(.111)	.13(.083)	.46(.254)	.32(.208)	.37(.180)
	Group lasso	.96(.01)	.81(.039)	.61(.084)	.93(.023)	.83(.044)	.82(.054)
	Group ridge	.96(.02)	.61(.062)	.59(.075)	.81(.127)	.66(.106)	.75(.069)
	Hubs	.02(.052)	.07(.064)	.01(.028)	.04(.069)	.05(.075)	.05(.060)
	Sparse group lasso$_{0.5}$	.59(.245)	.57(.163)	.13(.14)	.69(.137)	.62(.140)	.59(.136)
Ridge penalty	Sparse group lasso$_{0.9}$	.70(.186)	.49(.148)	.13(.149)	.72(.132)	.59(.136)	.60(.147)
	Sparse group lasso$_{0.1}$	.47(.254)	.59(.164)	.13(.127)	.59(.139)	.58(.130)	.53(.116)
	Group lasso	.91(.031)	.79(.029)	.42(.065)	.82(.053)	.75(.042)	.70(.059)
	Group ridge	.75(.07)	.63(.078)	.10(.055)	.53(.097)	.48(.11)	.37(.10)
	Lasso	.91(.016)	.59(.060)	.51(.080)	.87(.035)	.68(.065)	.70(.074)
Common	Ridge	.80(.028)	.73(.037)	.26(.046)	.66(.041)	.63(.044)	.539(.050)
	Elastic net	.92(.015)	.54(.089)	.60(.057)	.87(.032)	.69(.067)	.68(.065)

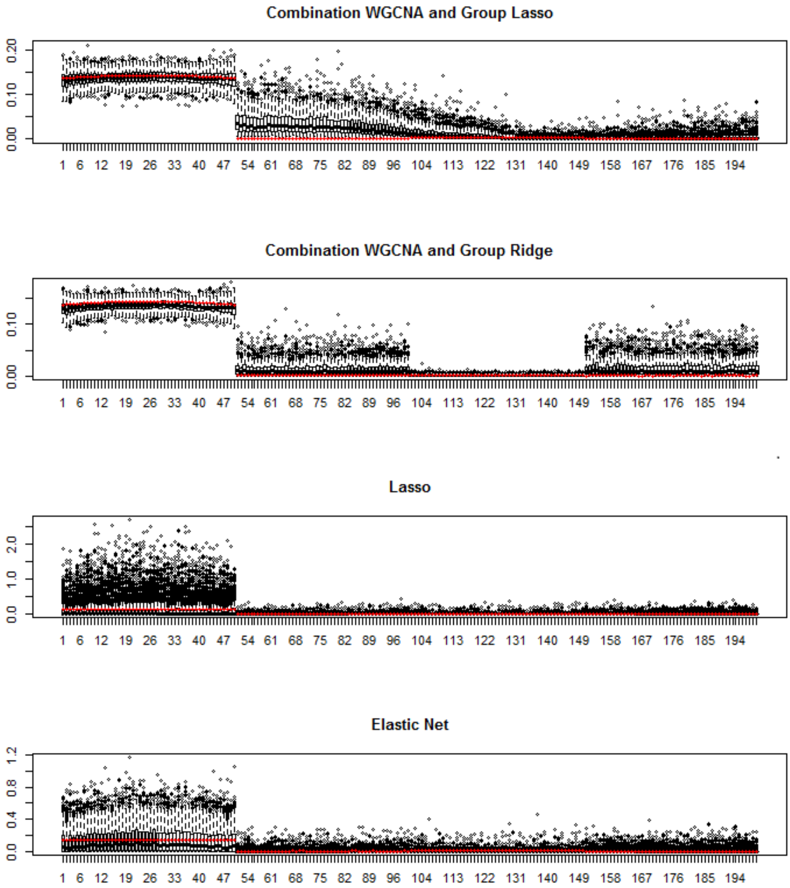
Table 5.4: Simulation study. Results obtained in terms of average Q^2 (across 500 replicates) for scenarios a,b,c, p=1000 variables, k=4 and k=8 modules, and n=50 individuals. Standard errors are given in brackets. The first column represents the method used to build the network. A Priori represents the situation were the true clustering of the predictors is known and no network analysis is performed.

Variable selection

Finally, we investigated the variable selection properties of the best performing (in terms of predictive ability) three-step procedures. Figs 5.3 and 5.4 show for scenario a, $k = 4$, $p = 200$ and $n = 100$ the variable selection properties of adaptive group ridge and group lasso in combination with WGCNA and graphical lasso, respectively. In both cases, the performance of lasso and elastic net is also shown. For each method, each boxplot shows for each of the p variables of $\mathbf{X}$ the distribution of the average estimated regression coefficients over the 10 fold cross-validation folds for each of the $M = 500$ Monte Carlo trials. The true simulated regression coefficients are also shown (red dots). Complete results for all scenarios are presented in the S2 File, Figure A to Figure R.

These results show that our three step approaches perform well in terms of specific regression coefficient estimation and variable selection. The four investigated approaches given by the combination of WGCNA and graphical lasso with adaptive group ridge and group lasso clearly separate informative from non-informative variables. In contrast, lasso regression, especially in scenario a, shows a very poor performance. The mean estimated coefficients by lasso for all p variables are close to zero, while the variability is very high for the features with non-zero effects, reflecting that lasso randomly selects a few of the informative variables and assigns a very large effect to them. To a lesser extent, the same phenomenon is also observed for elastic net. Even if the mean estimate for informative variables is larger and variability is lower than for lasso, the overall performance of elastic net is inferior to our three-step methods based on including grouping information.

Fig 5.3 top panel shows that the combination of WGCNA and group lasso tends to overestimate the effect of the variables belonging to the second cluster of variables. This is due to the underestimation of the number of clusters by WGCNA and the joint penalization of group lasso. Interestingly, adaptive ridge is less affected by this issue. When using graphical lasso as network analysis method, the first informative group of variables is clearly separated from the rest, and the estimation is close to the theoretical one (Fig 5.4).


Combination WGCNA and Group Lasso

Combination WGCNA and Group Ridge

Lasso

Elastic Net

Figure 5.3: Simulation study: Variable selection results with WGCNA. Variable selection results for scenario a, $k = 4$, $p = 200$, and $n = 100$. Box-plots of the absolute values of the estimated parameters for the 200 variables over the 500 simulated datasets are plotted. The red points represent the absolute average true values over the 500 datasets.

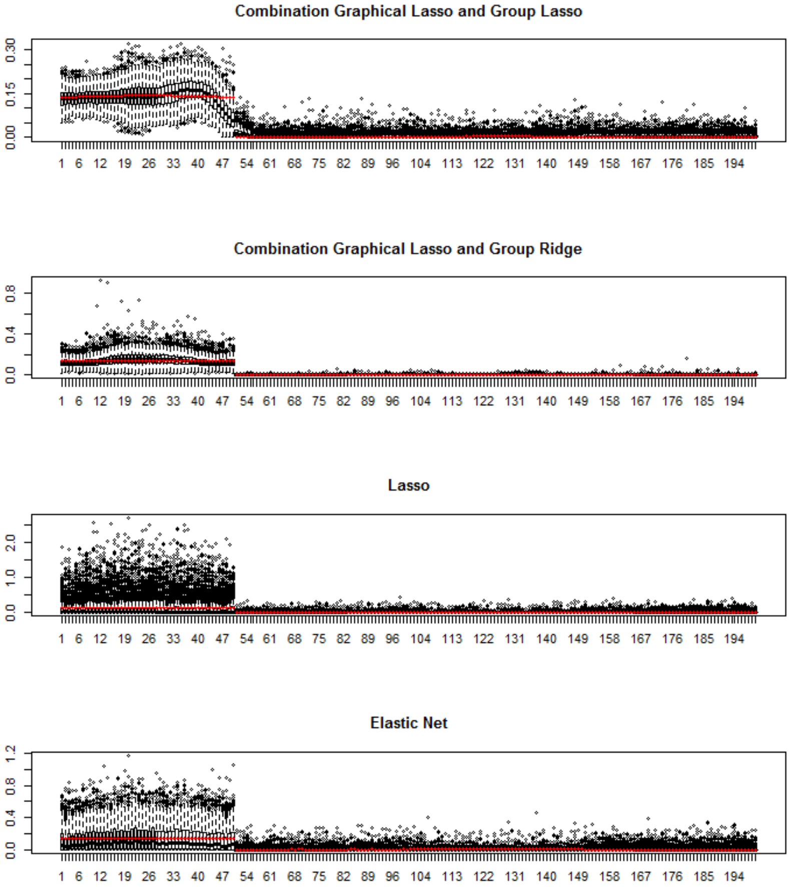


Figure 5.4: Simulation study: Variable selection results with Graphical Lasso. Variable selection results for scenario a, $k = 4$, $p = 200$, and $n = 100$. Box-plots of the absolute values of the estimated parameters for the 200 variables over the 500 datasets simulated are plotted. The red points represent the absolute average true values over the 500 datasets.

5.4 Real data analysis

We analyzed data from the DILGOM study and from the breast cancer cell line pharmacogenomics dataset. In both cases, the aim is to obtain biological insights about the features which drive the prediction of BMI and treatment response.

In the DILGOM study we consider two omics datasets measured at baseline to predict the body mass index (BMI) after seven years of follow-up. Serum nuclear magnetic resonance (NMR) spectroscopy metabolites measures and gene expression profiles were considered. The analysed sample contained n = 258 individuals for which both types of omic measurements and the outcome of interest (log-transformed BMI) were available. In the breast cancer cell lines dataset, we were interested in using gene expression for predicting the response to the Erlotinib drug. Treatment response is measured using the GI50 index, a quantitative measure which measures the growth inhibitory power of the test agent. The analysed sample consisted of 45 breast cancer cell lines.

5.4.1 DILGOM: metabolites

The serum metabolomic data consists of quantitative information on 57 metabolic measure of various types, including lipids, lipoprotein subclasses, amino acids, cholesterol, glycolysis-related metabolites and fatty acids (see S3 File, Table A). Table 5.5 and Table 5.6 show the main results for the prediction of BMI after 7 years of follow-up using serum NMR metabolites as predictors. Table 5.5 shows the performance of each method in terms of predictive ability measured through Q^2. We observe that adaptive group ridge and group lasso provide the best results and that they perform slightly better than ridge, lasso and elastic net. Namely, for adaptive group ridge when using graphical lasso $Q^2 = 0.244$ and for adaptive group ridge in combination with WGCNA $Q^2 = 0.233$, while for ridge $Q^2 = 0.227$ and for lasso $Q^2 = 0.222$. Also, group lasso combined with WGCNA outperforms ridge and lasso (Q^2 of 0.241). Variable selection based on hubs presents a notably lower predictive ability (best performance is reached with graphical lasso, $Q^2 = 0.176$) than methods based on regularization, except for sparse group lasso, which is not competitive at all ($Q^2 < 0.002$ in all cases). Table 5.6 shows the variable selection properties of the two top performing methods; the combination of WGCNA and group lasso and the combination of graphical lasso and adaptive group ridge. The top 12 variables selected by the combination of WGCNA and group lasso approach are shown in the left part of Table 5.6, jointly with their average regression coefficient, selection frequency over the 10 cross-validation folds used in the analysis, and their cluster membership. For each of these top 12 variables, average effect and selection frequencies over the 10 cross-validation folds are also shown for the combination of graphical lasso and adaptive group ridge, lasso, and elastic net. These top 12 variables represent two different families of metabolites. Namely, lipids and fatty acids (XSVLDLL, XLHDLL, SM, SHDLL, FAW6), and amino-acids and glycolysis-related metabolites (ALB, TYR, PHE, GLY, GLOL, GLC). This means that the three-step approach based on WGCNA and group lasso consistently points out these groups of metabolites as those driving the prediction of BMI. Accordingly, these two families of metabolites are well separated in the network analysis plus clustering steps (by both WGCNA and graphical lasso methods), consistently belonging to different clusters (see columns labeled 'Cluster' in Table 6).

Interestingly, our three-step approach based on the combination of WGCNA and group lasso provides similar effect estimates for metabolites XSVLDLL, SM, FAW6 and SHDLL (.038,.034,.031, and .030, respectively), all of them belonging to the same cluster of lipids and fatty acids. The combination of graphical lasso and adaptive group ridge provides similar results in terms of effect size. On the contrary, lasso provides more extreme estimates due to within-group random variable selection, i.e. lasso selects at random oen feature over a set of highly correlated variables. Specifically, lasso assigns quite different effect estimates to the lipids and fatty acids group (XSVLDLL:.036, SM:.018, FAW6:.017, SHDLL:.003). The effect size of SHDLL is particularly counter-intuitive since high density lipids are well established risk factors for obesity (Shamai et al., 2011). Elastic net appears not to solve this issue and provides similar results than lasso.

	WGCNA	Graphical lasso	Ridge penalty
	Q^2	Q^2	Q^2
Hubs + ridge	0.153	0.176	0.153
Group lasso	$\mathbf{0.241}$	0.225	0.221
Sparse group lasso $\alpha = 0.5$	0.013	0.010	0.015
Sparse group lasso $\alpha = 0.9$	0.003	0.012	0.013
Sparse group lasso $\alpha = 0.1$	0.013	0.007	0.016
Group ridge	$\mathbf{0.233}$	$\mathbf{0.244}$	0.225
Lasso	0.227	0.227	0.227
Ridge	0.222	0.222	0.222
Elastic net	0.208	0.208	0.208
Number of Clusters	4	7	4-6

Table 5.5: DILGOM metabolomics. Prediction accuracy of the models obtained for the different approaches on metabolites. In bold are the combinations of network analyses and prediction approaches which perform better than lasso, ridge, and elastic net.

5.4.2 DILGOM: Transcriptomics

Due to the computational intensity of the graphical lasso approach, we considered two sets of gene expression probes for analysis. A set of 2980 probes which was only analysed by WGCNA to perform network analysis and a set of 732 filtered probes (probes with a variance higher than 1) were WGCNA and graphical lasso were used. The main results are presented in Table 5.7 and Table 5.8. Table 5.7 presents the prediction ability results of the used methods. For the set of filtered probes (left part of Table 5.7), the best method with regard to predictive performance is the combination of WGCNA and group lasso ($Q^2 = 0.258$). Adaptive group ridge appears to provide poor results ($Q^2 = 0.158$ in combination with WGCNA and $Q^2 = 0.188$ in combination with graphical lasso) in the transcriptomics context. In contrast to the observed results regarding the NMR metabolites, adaptive ridge is clearly outperformed by lasso ($Q^2 = 0.227$) and elastic net ($Q^2 = 0.253$), but still provide better results than the ridge regression ($Q^2 = 0.071$). Also, we observe that for transcriptomics elastic net provides better results than lasso which was not the case for the metabolites. For the larger set of probes (right part of Table 5.7), the best prediction accuracy is achieved using the combination of WGCNA with group lasso $Q^2 = 0.418$ while lasso and elastic net show similar predictive abilities with $Q^2 = 0.257$ and $Q^2 = 0.265$, respectively. Ridge presented better prediction accuracy with the large set of probes but its performance is still very low (Q2=0.131). In line with the simulation study, the benefits of our three-step proposal is larger when the number of probes increases.

Table 5.8 presents the number of variable selected for the two group lasso approaches (based on WGCNA and graphical lasso), lasso, and elastic net. The left part of Table 5.8 shows the results for the filtered set of probes and the right part shows the results for the large set of probes. For the filtered set of probes, it appears that group lasso retains more variables than lasso and elastic net. WGCNA in combination with group lasso provided 687 variables which were selected at least once during the cross validation process, while the combination of graphical lasso and group lasso provided 485 variables. Lasso and elastic net identified only 78 and 123 variables, respectively. Moreover, the models obtained with group lasso are more stable than those obtained with the standard approaches, lasso and elastic net. Indeed, using WGCNA, 19.9% of the 687 variables are selected in all the 10 cross-validation folds and for graphical lasso 18.9% of the 485 variables are selected. In contrast, for lasso and elastic net only 3.8% and 5.6% of the variables are selected in all the cross-validation folds. For the larger set of probes, the number of variables always selected increased for lasso and elastic net with respectively 13 and 21 variables, this is not the case for the combination of WGCNA with group lasso with 48 variables always selected for the set of 2928 probes while 137 variables were always selected with the smaller set of probes. From the 48 variables obtained, only 5 were also included in the previous set of 137 variables.

To investigate the biological relevance of the selected variables in the prediction models obtained, a gene set enrichment analysis was performed using the Gene Set Enrichment Analysis software (GSEA;Subramaniana et al. (2005); Mootha et al. (2003)) on the variables always selected by each approach during the cross-valiadation process. A gene set enrichment analysis consists of comparing the set of gene identified with a priori known group of genes that have been grouped together by their involvement in the same biological pathway. Table 5.9 presents the results of the enrichment analysis when using the large set of transcriptomics. None of the pathways obtained in the enrichment analysis by the different methods has been previously identified as related to BMI. The enrichment analysis based on the 137 and 92 genes obtained from the filtered set of probes was more insightful. Among the 137 genes selected by the combination WGCNA and group lasso, 33 were associated with cardiovascular disease ($p = 0.019$) and 6 of these 33 genes were associated with obesity ($p = 0.044$). Among the 92 genes obtained with the combination of graphical lasso and group lasso, 3 of them where included in the glucagon signaling pathway ($p = 0.070$) and 3 were in the insulin resistance pathway ($p = 0.080$). These results are not surprising since it is known that increased insulin and decreased glucagon secretion play a role in obesity Schade and Eaton (1974). Due to the small number of variables of lasso and elastic net, 7 and 3 predictors respectively, the enrichment analysis did not provide associated pathways.

	WGCNA + Group lasso			Graphical lasso + adaptive group ridge		
Variable	Average beta	Frequency	Cluster	Average beta	Rank	Cluster
GLOL	.064	10	1	.039	5	6
TYR	.060	10	1	.070	2	1
ALB	-.059	10	1	-.075	1	1
GLY	-.041	10	1	-.039	4	1
PHE	.038	10	1	.046	3	1
XSVLDLL	.038	10	2	.017	16	2
XLHDLL	-.038	10	3	-.034	7	5
HIS	-.036	10	1	-.030	8	1
SM	.034	10	2	.016	17	2
FAW6	.031	10	2	.003	31	3
GLC	.031	10	1	.037	6	1
SHDLL	.030	10	2	.030	9	5
	Lasso			Elastic Net		
	Average beta	Frequency	Rank	Average beta	Frequency	Rank
GLOL	.074	10	4	.063	10	3
TYR	.080	10	3	.068	10	2
ALB	-.086	10	2	-.069	10	1
GLY	-.037	10	6	-.035	10	7
PHE	.038	10	5	.042	10	5
XSVLDLL	.036	10	7	.038	10	6
XLHDLL	-.089	9	1	-.056	10	4
HIS	-.024	9	8	-.020	10	11
SM	.018	8	10	.011	8	17
FAW6	.017	7	12	.011	8	14
GLC	.018	10	11	.022	10	9
SHDLL	.003	3	20	.005	7	20

Table 5.6: DILGOM metabolomics. Top 12 metabolites (in terms of average beta) selected by the combination of WGCNA and group lasso, their selection frequencies and cluster membership. For lasso, graphical lasso + ridge, and elastic net, the rank of the variables according to the absolute values of the average effect size is added.

	Filtered set (p=732)		Larger set (p=2980)
	WGCNA	Graphical lasso	WGCNA
	Q^2	Q^2	Q^2
Group lasso	0.258	0.215	0.418
Group ridge	0.158	0.188	
Lasso	0.227	0.227	0.257
Ridge	0.071	0.071	0.131
Elastic net	0.253	0.253	0.265
Number of clusters	16-17	32-36	40-45

Table 5.7: DILGOM transcriptomics. Prediction accuracy of the models obtained by combination of networks and prediction models as well as lasso, ridge, and elastic net for transcriptomics.

	Filtered set (p=732)			Larger set (p=2980)		
	Always	At least once	Proportion	Always	At least once	Proportion
WGCNA and group lasso	137	687	0.199	48	252	0.190
Graphical lasso and group lasso	92	485	0.189			
Lasso	3	78	0.038	13	134	0.097
Elastic net	7	123	0.056	21	176	0.119

Table 5.8: DILGOM transcriptomics. Number of variables selected during the cross-validation process, at least once, in all croos-validation folds and the proportion of variables selected all in the set of variables selected at least once.

method	Pathway	Number variables	FDR
WGCNA and group lasso	Genes transcriptionally modulated in the blood of multiple sclerosis patients in response to subcutaneous treatment with recombinant IFNB1	10	9.68 e-15
	Genes up-regulated in CD34+ hematopoetic cells by expression of NUP98-HOXA9 fusion off a retroviral vector at 3 days after transduction	10	3.86 e-12
	Genes representing interferon-induced antiviral module in sputum during asthma exacerbations	8	1.27 e-11
Lasso	Genes exclusively down-regulated in B lymphocytes from WM (Waldenstroem's macroblobulinemia) patients but with a similiar expression pattern in the normal cells and in the cells from CLL (chronic lymphocytic leukemia) patients.	2	5.62 e-3
	Genes down-regulated in erythroid progenitor cells from fetal livers of E13.5 embryos with KLF1 knockout compared to those from the wild type embryos	6	5.62 e-3
Elastic net	Genes down-regulated in CD4+ T lymphocytes transduced with FOXP3.	3	1.55 e-3
	Genes up-regulated in MCF7 cells (breast cancer) after stimulation with NRG1	4	1.55 e-3
	Genes down-regulated in normal hematopoietic progenitors by RUNX1-RUNX1T1 fusion	4	1.55 e-3

Table 5.9: DILGOM transcriptomics. Top significant pathways identified by enrichment analysis using the GSEA software for all predictions model using the variables always selected during the cross-validation process of the breast cancer cell lines study on the transcriptomics data. For each method, the number of variables common to the pathway and the set of variables selected at least 5 times and the false discovery rate (FDR) of the enrichment test are presented.

5.4.3 Breast cancer cell lines

The main results of the prediction of the treatment response of breast cancer cell lines to Erlotinib are presented in Table 5.10. The best prediction performance is again the combination of WGCNA and group lasso with $Q^2 = 0.654$. Ridge with $Q^2 = 0.610$ performs better than lasso and elastic net with $Q^2 = 0.571$ and $Q^2 = 0.564$, respectively. For this dataset the combination of WGCNA and group lasso is less stable and is not always able to pick the same variables during the cross-validation process, while lasso and elastic net are able to always pick 2 probes. With regards to variables selected at least 5 times by WGCNA + group lasso, lasso and elastic net, all 3 methods have a similar number of selected variables with respectively 22, 18 and 25. The intersection between 3 identified sets of variables is empty. The enrichment analysis identified genes related to breast cancer for the WGCNA + group lasso and elastic net approaches as presented table 5.11. This was not the case for lasso.

	Q^2	Number of Variables	
		At least 5 times	always
WGCNA and group lasso	0.654	22	0
Lasso	0.571	18	2
Ridge	0.610	5376	5376
Elastic net	0.564	25	2
Total number of variables		5376	5376

Table 5.10: Breast Cancer analysis. Prediction accuracy and numbers of variable selected at least 5 times and always selected in the 10-fold cross-validation process of the different approaches on the whole set of probes for the Breast cancer cell lines.

method	Pathway	Number variables	FDR
WGCNA and group lasso	Candidate genes in genomic amplification regions in hepatocellular carcinoma (HCC) samples	6	5.61 e-10
	Genes within amplicon 17q11-q21 identified in a copy number alterations study of 191 breast tumor samples.	6	6.49 e-8
	Genes up-regulated in DLBCL (diffuse large B-cell lymphoma) cell lines sensitive to stimulation of CD40 relative to the resistant ones	5	4.62 e-5
Lasso	Genes up-regulated in confluent IMR90 cells (fibroblast) after knockdown of RB1 by RNAi	7	5.621 e-6
	Genes up-regulated in the neural crest stem cells (NCS), defined as p75+/HNK1+	5	5.92 e-6
	Genes down-regulated in BEC (blood endothelial cells) compared to LEC (lymphatic endothelial cells)	5	6.66 e-6
Elastic net	Genes down-regulated in TMX2-28 cells (breast cancer) which do not express ESR1 compared to the parental MCF7 cells which do	11	5.41 e-10
	Genes up-regulated in confluent IMR90 cells (fibroblast) after knockdown of RB1 by RNAi.	9	2.52 e-8
	Genes positively correlated with recurrence free survival in patients with hepatitis B-related (HBV) hepatocellular carcinoma (HCC)	5	4.57 e-6

Table 5.11: Breast Cancer analysis. Top significant pathways identified by enrichment analysis using the GSEA software for all predictions model using variables selected at least 5 times during the cross-validation process on the transcriptomics data of the breast cancer cell lines study. For each method, the number of variables common to the pathway and the set of variables selected at least 5 times and the false discovery rate (FDR) of the enrichment test are presented.

5.5 Discussion

In this paper, we presented a new strategy to obtain accurate, stable and interpretable prediction models. The key components of our proposed approach are to capture the correlation structure of the features within an omic dataset, to derive clustering information, and to include it in a group penalization model. Our approach seems to provide interpretable models by capturing underlying biological mechanisms impacting the phenotype of interest.

Our applications showed that the proposed three step approach can outperform the standard regularized regression approaches in terms of prediction ability, stability and biological interpretation in high-dimensional settings or when groups of strongly correlated features are present in the data. Our analyses highlighted the weakness of methods such as lasso and elastic net in terms of stable variable selection in highly correlated datasets. Indeed, for the metabolites, our WGCNA and group lasso combination selected a group of highly correlated metabolites (cluster 2 including XSVLDLL, SM, FAW6, and SHDLL) while lasso selected XSVLDLL all the times in the cross-validation process but SHDLL only 3 out of 10 times. In addition it appeared that for the large transcriptomics dataset the prediction accuracy is also larger for our proposed methods than for the standard regularization methods. The analysis of the breast cancer cell lines study showed some limitations in terms of stability for our network-based approach when the number of samples is relatively small. Probably the networks obtained during the cross-validation steps are less stable for a small number of samples leading to a less stable clustering and prediction model. Further with regard to transcriptomics, the obtained groups of gene expression features identified by our strategies were enriched for known pathways linked to BMI (DILGOM) and breast cancer (breast cancer cell lines). This was only the case when using the filtered transcriptomics dataset. This was not always the case for lasso, ridge, and elastic net. For the unfiltered transcriptomic datasets, the gene sets were not enriched for pathways related to the outcome. Here more research is needed. These results suggest that our proposed approaches can indeed improve the understanding of prediction models while keeping a good prediction accuracy.

The performance of our approaches compared to the standard approaches was in line with the results obtained from the simulation study. Indeed the combination of WGCNA or graphical lasso with group lasso appeared to provide the most stable results, hence probably better interpretable. The prediction accuracy of these approaches was also good and for large omics datasets even better than the prediction accuracy of the standard approaches. Further our simulations showed that several group penalization models (sparse group lasso and adaptive group ridge) are quite sensitive to the used grouping structure. In contrast the group lasso approach proved to be quite robust with respect to the network approach used. Also, we have explored the idea of reducing the omic dataset dimensionality by choosing ‘important’ features by group based on network topology (such as our ‘hubs’ selection). This attractive approach to reduce the prediction complexity only performed well when using WGCNA for predictors which are highly associated to the phenotype of interest. Its performance was very sensitive to the used network method and bad in low-signal situations. Overall the combination of graphical lasso for network construction and group lasso was the best performing method in our simulation study. However, this approach computationally challenging for a large number of features and, therefore, cannot deal with large omics datasets. Moreover in the real data analysis better results were obtained when WGCNA was combined with group lasso. Therefore, for large datasets we recommend the combination of WGCNA and group lasso, while for smaller datasets both network approaches can be applied.

The presented work can be extended in various ways. So far, all our analyses focused on prediction of a continuous outcome, but all the obtained results apply, in principle, to other types of response variables, such as binary outcomes (classification problems) and to time-to-event data. Also, prior knowledge on biological grouping could be included in our three-step approaches if available, even if it is only partial. Our simulation study showed good results if the correct underlying clustering is known. Given that such biological knowledge is only partially known in many omic applications, we have proposed to use network analysis to infer the correlation structure. Including external prior biological in the first step of network construction may lead to an improvement of the clustering obtained and, therefore, of the proposed methods. Another possible extension is to build prediction models with two or more sets of omic predictors. It is known (Rodríguez-Girondo et al., 2018) that using a common penalization (such as lasso or ridge) to the extended dataset containing both omic sets to be combined can lead to worse predictive ability than using only one of these omic sets. Therefore, applying our three-step approach to the stacked dataset of different omic predictors may outperform current methods. Alternatively, more advanced network techniques as multi-layer networks Kivelä et al. (2014), based on obtaining the correlation structure between and within the omic sets may be improve prediction models. These extensions are currently under investigation.

To conclude, we presented a set of methods which provides accurate and stable predictions possibly leading to better interpretation, as is shown in the real data application. In the DILGOM study, a much more stable set of metabolomic predictors for BMI was obtained compared to standard approaches. Moreover, better predictions were obtained with our approach when using a large set of gene expression probes to predict BMI. Regarding the prediction of breast cancer, identified gene modules with our approach appeared to be interpretable since enrichment analyses showed that selected features could be linked with breast cancer tumors. This was not the case when using the standard approaches.

6

Integration of several omic sources in prediction models using network-based approaches

Abstract

In the last decades, biomolecular research developments led to an increasing number of omics measurements. These measurements have been widely used for prediction of numerous phenotypes and diseases. The next step is to combine and use various types of omics data to further improve prediction. However, the combination of heterogeneous datasets, in terms of scale, noise structure, and normalizations, is challenging and there is not yet any state-of-the art approach. In this paper, we propose methods based on network analysis and group penalization to combine several omics sources in one prediction model while taking into account the possible interaction between them. An extensive simulation study has been performed in order to compare this new approach with the common regularization methods lasso, ridge and elastic net. Finally we illustrate the advantages of our approach by application of the methodology to two problems, namely prediction of body

---

This chapter has been submitted for publication as: Renaud Tissier, Mar Rodríguez-Girondo, Jeanine J. Houwing-Duistermaat. Integration of several omic sources in prediction models using network-based approaches.

mass index in the DIetary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome study (DILGOM) and prediction of response of each breast cancer cell line to treatment with specific drugs using a breast cancer cell lines pharmacogenomics dataset. The results show that prediction models based on multiple omics should carefully account for within and across-omic correlations. In that case, predictive performance can be improved and single-omic models can be outperformed.

6.1 Introduction

One of the current main challenges in prediction modeling is the integration of several sources of omic predictors in a single model. Recent developments in biomolecular research, have resulted into the availability of an increasing number of omics measurements such as genomics, methylomics, proteomics, metabolomics, and glycomics for one subject. For predicting health-related traits, these omic sources have shown promising results in some settings using single omic source models (Ibrahim-Verbaas et al., 2014; Bahado-Singh et al., 2014; Lemesle et al., 2015). Indeed, in statistics much work has been done in developing these single omics prediction models (Hastie et al., 2009; Bühlmann and van de Geer, 2011). The next challenge is the integration of multiple sources of omic data which may potentially improve the performance of single-omic prediction models. However, there is currently no state-of-art method for achieving this goal.

In this paper, we aim to investigate the combined predictive ability of heterogeneous omic sources. These sources differ with regard to dimensionality of the datasets, normalization procedures used and presumably their error structures. For example, multiplicative noise is often encountered in imaging analysis (Liu et al., 2013) and therefore commonly met in fluorescence based measures as transcriptomics (Sásik et al., 2002) while others are expected to have an additive noise structure. We encountered this challenge in two motivating studies. Our first data application refers to a question from the Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) study, namely the prediction of body mass index (BMI) after seven years of follow-up based on the combination of baseline metabolomics and transcriptomics. Our second application comprises copy number variants and gene expression for treatment response prediction using the publicly available Breast cancer cell line pharmacogenomics dataset (https://genomeinterpretation.org/content/breast-cancer-cell-line-pharmacogenomics-dataset). In these studies, we have previously shown better predictive ability by including network information for single-omic analysis. In this paper we extend these methods to multiple omics datasets.

The literature on prediction based on multiple omic datasets is scarce. A first straightforward approach is to apply existing techniques for high-dimensional prediction such as regularized regression models to the stacked dataset of omic features. It has been shown that this naive approach can provide poorer predictive ability than using only one of the available omic sources (Rodríguez-Girondo et al., 2018) due to the heterogeneity across omics datasets. Hence more sophisticated strategies are needed.

Another approach is to first perform a dimension reduction to the omics datasets. For example Acharjee et al. (2016), first performs random forest on each omic source to select the most relevant features. These features are then combined in a single prediction model. In order to understand possible interactions between omic sources, these features are then combined in a network. Potentially, such network could be included in a prediction model using the methods developped by Tissier et al. (2018). Alternatively, latent structures within and between datasets can be identified by using O2PLS (Trygg and Wold, 2003; Bouhaddani et al., 2016). These latent structures can be summarized in a few independent components (dimension reduction) which can be included in a prediction model. This approach is interesting as it has been developed to deal with heterogeneous datasets. However, disadvantages of first applying dimension reduction step is the loss of possible relevant information, either by ignoring the joint distribution of two different omics features or the relationship between the omics datasets an the outcome.

In this paper, we will not perform a-priori dimension reduction, instead, we propose group penalization. Therefore, as part of the model building procedure, group inference is performed using network analysis followed by clustering. To this end we extended our approach for a single omic source (Tissier et al., 2018) to multiple omic sources. Specifically, we explore how to include groups containing features from different omic sources and whether this is beneficial in terms of predictive performance. Under this general framework, we will investigate several possible alternatives regarding the inference of groups and the incorporation of this information in regularized regression. It is indeed unclear if, due to the heterogeneity between omic sources as for example metabolites and transcriptomics, combining the datasets before performing the subsequent group inference might lead to poor results. An alternative is to restrict group inference to each of the omic sources which, however, might miss possible across-omic relations due to shared biological pathways.

The rest of the paper is organized as follows: in Section 2, we present the general methodological framework based on grouped regression and several variants regarding group inference and grouped lasso regression. Section 3 contains technical details about the specific method used for group inference. Section 4 focuses on outcome prediction. An intensive simulation study is presented in Section 5 to empirically evaluate the performance of the different studied methods in terms of predictive ability and variable selection properties. The results of the integrated approach are compared with single omic source predictive ability. In Section 6 the methods are applied to two different studies. Main conclusions and a final discussion follow in Section 7

6.2 Network-based group-penalized prediction

Let the observed data be given by $(\mathbf{z}, \mathbf{Y}, \mathbf{X})$, where $\mathbf{z} = (z_1, \ldots, z_n)^{\mathbf{T}}$ is the continuous outcome measured in n independent individuals, and $\mathbf{Y}$ and $\mathbf{X}$ are matrices of dimension $n \times p$ and $n \times q$ respectively, representing two omic predictor sources with p and q features. Let $\mathbf{M}$ be the stacked dataset of $\mathbf{Y}$ and $\mathbf{X}$. Our main goal is to build a predictive model for $\mathbf{z}$ based on $\mathbf{Y}$ and $\mathbf{X}$ with good predictive performance.

The matrices $\mathbf{X}$ and $\mathbf{Y}$ might be high-dimensional ($n < p, q$) and present complex dependence structures, potentially shared due to existing biological pathways, or coordinated functions of groups of features.

We propose a general framework of grouped regularized regression methods including group inference as part of the model building procedure. Group inference relies on first estimating the existing relations among features using network analysis techniques and then deriving groups of features using hierarchical clustering. Based on this general framework, three approaches are proposed, with variable level of complexity in group inference.

The first algorithm named **GLasso0** consists of constructing a separate network for each omic source and to perform subsequent hierarchical clustering on each of the resulting adjacency matrices. Finally, group lasso regression is performed. This approach only allows for omic-specific groups of features, so correlation across omic sources cannot be captured. The second proposed method, **GLasso**, starts by building a unique network from the stacked dataset $\mathbf{M}$. Subsequent hierarchical clustering is performed on the resulting adjacency matrix and group regression is also based on group lasso. This approach is a direct application of the method proposed by Tissier et al. (2018) on the stacked dataset $\mathbf{M}$ and it potentially allows for groups including features from different omic datasets. However, when noise structures of the omic datasets are different stacking the datasets might be problematic for network construction. Finally, the third proposed approach, **OverlapLasso**, is an extension of **GLasso0** and allows for overlapping groups of features. Namely, after obtaining the omic-specific groups of features, an extra network analysis and hierchical clustering is conducted at the group level to try to incorporate extra shared information by the two omic sources.

In all three approaches weighted gene co-expression network analysis (WGCNA) and the dynamic tree cut algorithm for hierarchical clustering were used. Outcome prediction relies on group lasso in the two first procedures (**GLasso0** and **GLasso**) and on a extension to allow the presence of features on multiple groups in the case of **OverlapLasso**. Specific components used in each step are described in more detail in the next section. For each approach, double cross-validation (Mertens et al., 2006, 2011) of the whole process (including group inference) was applied to obtain proper tuning parameters and summary performance measures in absence of an external validation set. The three procedures have been implemented in the R function MultiPredNet which is available at github (https://github.com/RenTissier/MultiPredNet). The function calls the packages `WGCNA` (for network construction and hierarchical clustering), `grpreg` (group lasso) and `grpregOverlap` (overlapping group lasso).

The basic structure of the the three procedures is as follows:

GLasso0 **Network-based group-penalized prediction model based on omic-specific group inference**

Step 1 Network construction

Input $\mathbf{Y}, \mathbf{X}$

Output $A_{\mathbf{Y}}, A_{\mathbf{X}}$ two adjacency matrices

Step 2 Hierarchical clustering

Input $A_{\mathbf{Y}}, A_{\mathbf{X}}$

Output $P_{\mathbf{Y}},\ P_{\mathbf{X}}$ omic-specific clusters

Step 3 Outcome prediction: Group lasso

Input $(\mathbf{M}, P_Y, P_X)$

Output $p+q$ β regression coefficients

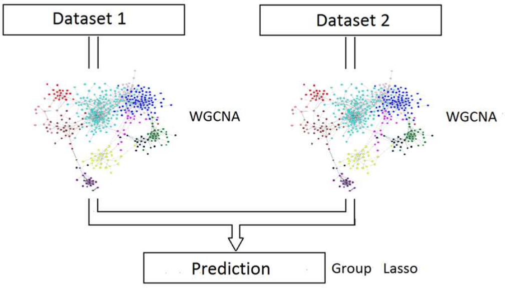


Figure 6.1: GLasso0

GLasso **Network-based group-penalized prediction model on stacked datasets Y and X**

Step 1 Network construction

Input $M = (\mathbf{Y}, \mathbf{X})$

Output A_M adjacency matrix

Step 2 Hierarchical clustering

Input A_M

Output P_M clusters

Step 3 Outcome prediction: Group lasso

Input $(\mathbf{M}, P_M)$

Output $p+q$ β regression coefficients

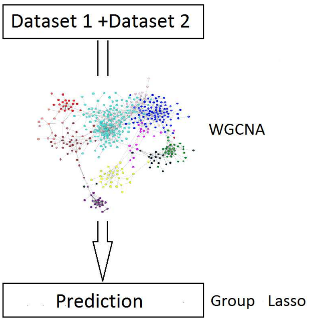


Figure 6.2: GLasso

OverlapLasso **Network-based overlapping group-penalized prediction model based on omic-specific group inference**

Step 1 Network construction

Input $\mathbf{Y}, \mathbf{X}$

Output $A_{\mathbf{Y}}$, $A_{\mathbf{X}}$ two adjacency matrices

Step 2.a. Hierarchical clustering

Input $A_{\mathbf{Y}}, A_{\mathbf{X}}$

Output $P_{\mathbf{Y}}$, $P_{\mathbf{X}}$ omic-specific clusters

Step 2.b. Principal component analysis

Input $P_{\mathbf{Y}}, P_{\mathbf{X}}$

Output $\mathbf{U} = (PC_{P_{\mathbf{Y}}}, PC_{P_{\mathbf{X}}})$ set of two first principal components of each cluster in $P_{\mathbf{Y}}$ and in $P_{\mathbf{X}}$

Step 2.c. Network construction +hierarchical clustering on $\mathbf{U}$

Input $\mathbf{U}$

Output $P_{\mathbf{U}}$ clusters

Step 2.d. Identification of $(\mathbf{Y}, \mathbf{X})$-shared clusters in $P_{\mathbf{U}}$

Input ***Input:*** $P_{\mathbf{U}}$

Output ***Output:*** $P_{\mathbf{U_M}}$ clusters

i. Identify the m clusters obtained in **Step 2.c.** which contain elements from both $P_{\mathbf{X}}$ and $P_{\mathbf{Y}}$

ii. For each $i = 1, \ldots, m$ of the identify clusters in **i.** identify the corresponding variables from $\mathbf{X}$ and $\mathbf{Y}$

iii. Denote by $P_{\mathbf{U_M}}$ the corresponding set of m new clusters.

Step 3 Outcome prediction: Overlapping group lasso

Input $(\mathbf{M}, P_Y, P_X, P_{UM})$

Output $p + q$ β regression coefficients

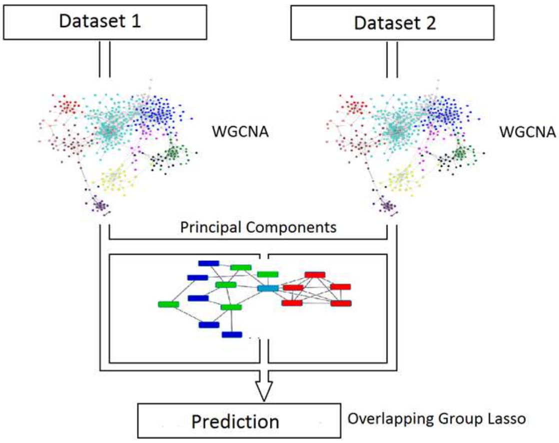


Figure 6.3: OverlapLasso

6.3 Group inference

6.3.1 Network construction: WGCNA

The first step of the group inference in all three proposed approaches involves network analysis. We use weighted networks, in general defined as an adjacency matrix $\mathbf{A} = [a_{ij}]$, where each coefficient a_{ij} represents how close features x_i and x_j are.

In this paper, we have used weighted co-expression networks based on pairwise correlations (Zhang and Horvath, 2005), originally proposed in the context of gene expression data. Due to the presence of high correlations in omic data, a parameter β (soft threshold) is introduced in order to shrink 'low' pairwise correlation values towards zero. The parameter β might be chosen in such a way that the free-scale topology criterion of the resulting network holds, i.e, the fraction of nodes with k edges should follow the power law $P(k) \approx k^{-\gamma}$, with $P(k)$ the fraction of nodes in the network with k edges and γ a constant with a value comprised between 2 and 3. The rationale behind the free scale topology criterion is the existence of hubs.

Co-expression networks have been successfully used in the context of transcriptomics (Stuart et al., 2003; Oldham et al., 2006, 2008). We also have shown good predictive performance when using WGCNA as part of group inference in the metabolomics setting (Tissier et al., 2018). We have extensively compared different network analysis methods for group inference in the context of single-omic prediction and WGCNA has shown a good balance between clustering accuracy and computational efficiency. Note that our proposed methods could be easily adapted to include other possible network analysis techniques.

6.3.2 Hierarchical clustering

Hierarchical clustering is used to detect groups of related features, using the previously estimated network to derive a metric matrix.

Specifically, we used the topological overlap dissimilarity measure as metric for the hierarchical clustering. The topological overlap of two nodes quantifies their similarity in terms of the commonality of the nodes they connect (Yip and Horvath, 2007). The topological overlap between node i and node j is given by:

$$TOM_{ij} = \frac{\sum_u a_{iu}a_{uj} + a_{ij}}{min\left(k_i, k_j\right) + 1 - a_{ij}}$$

with a_{ij} the weight between i and j in the adjacency matrix, and $k_i = \sum_u a_{iu}$. The topological overlap dissimilarity measure is defined as : $dissTOM_{ij} = 1 - TOM_{ij}$.

To obtain the groups of features, we applied the dynamic tree cut algorithm (Langfelder et al., 2008) on the dendogram defined by the dissimilarity measure. This algorithm is an adaptive and iterative process of cluster decomposition and combination until the number of clusters becomes stable. In contrast to a constant height cut-off method, this approach is capable of identifying nested clusters.

6.4 Outcome prediction

We focus on grouped regularization approaches, which are able to deal with high-dimensional data and keep group related features together, improving the interpretability and stability of the resulting models. Here, we consider group and overlapping group lasso.

Group lasso

GLasso and **GLasso** are using Group lasso (Yuan and Lin, 2006) to build the prediction model. This approach, simultaneously shrinks all the coefficients belonging to the same pre-specified group towards zero and hence selects groups of related features. Assume that L groups were obtained after subsequent hierarchical clustering. The final prediction model is given by the group lasso estimator:

$$\min_{\beta \in R^{(p+q)}} \left(\left\| z - \sum_{l=1}^{L} M_l \beta_l \right\|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l} \, \|\beta_l\|_2 \right)$$

where $l \in (1 \cdots L)$ represents the index of the group of predictors. M_l is the matrix of predictors in the group l and $\sqrt{p_l}$ is a penalty to take into account the varying group size. The choice of the tuning parameter λ is made by cross-validation based on minimization of the AIC. The group lasso estimator is asymptotically consistent even when model complexity increases. Note that if each group contains just one variable, group lasso is equivalent to the standard lasso (Noah et al., 2013).

Overlaping group lasso

Overlapping group lasso is a group-based regularized regression method which allows predictors to be part of several clusters. Suppose that the set of $p + q$ features from the stacked dataset **M** are assigned to L possibly overlapping groups ($P_{\mathbf{Y}}, P_{\mathbf{X}}, P_{\mathbf{UM}}$ obtained in **Step 2.d.** of the **OverlapLasso** procedure). Jacob et al. (Jacob et al., 2009) proposed the overlapping group lasso estimator:

$$\min_{\beta \in R^{M}} \left(\|z - M\beta\|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l} \, \|\gamma_l\|_2 \right) \tag{6.1}$$

where $\gamma_1, \ldots, \gamma_L$ are L vectors of dimension $(p+q) \times 1$ called latent coefficient vectors. The resulting $p+q$ regression coefficients of interest, β_m, $m = 1, \ldots, p+q$, are obtained as $\beta_m = \sum_l \gamma_{lm}$, where $\gamma_{lm} = 0$ if the variable m is not in the group l. Equation 6.1 can be rewritten as:

$$\min_{\gamma \in R^{M}} \left(\left\| z - \sum_{l=1}^{L} M \gamma_l \right\|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l} \, \|\gamma_l\|_2 \right)$$

In our case, in the **OverlapLasso** procedure, after deriving omic-specific and not overlapping groups of features ($P_{\mathbf{Y}}$ and $P_{\mathbf{X}}$) in **Step 2.a.**, the matrix $\mathbf{U}$ composed of the two first principal components scores of each omic-specific group is defined. The rationale behind this idea is to get a comparable common scale for the two omics sources $\mathbf{Y}$ and $\mathbf{X}$. Network and hierarchical clustering is then applied to $\mathbf{U}$ in order to obtain shared clusters by $\mathbf{Y}$ and $\mathbf{X}$ (**Step 2.c.**). Namely, if one detected group at this stage combines latent scores belonging to different omic sources, we take the corresponding features in $\mathbf{Y}$ and $\mathbf{X}$ and combine them in a new group. As a result of these extra steps, features can belong to several groups, and groups can contain features from both omic sources.

6.5 Simulations

6.5.1 Simulation setup

An extensive simulation study was conducted to investigate the performance of the three proposed methods and to compare them with the standard strategies of using common regularized approaches, namely, lasso, ridge, and elastic net regression on stacked datasets. We simulated two omic predictors $\mathbf{Y}$ and $\mathbf{X}$ from a zero-mean multivariate normal distributions with correlation matrices $\mathbf{\Sigma}_Y$ and $\mathbf{\Sigma}_X$, respectively. Following the recent literature on pathway and network analysis of omics data (Tissier et al., 2018), we generated $\mathbf{\Sigma}$ according to a hub model with noise based on realistic situations (Hardin et al., 2013). Correlation between $\mathbf{Y}$ and $\mathbf{X}$ was created by calculating the singular decomposition (svd Jolliffe (2008)) of $\mathbf{X}$, $\mathbf{X} = \mathbf{U^X D^X U^X}^t$ and of $\mathbf{Y}$, $\mathbf{Y} = \mathbf{U^Y D^Y U^Y}^t$ and replacing the second column of $\mathbf{U^Y}$ by the second column of $\mathbf{U^X}$. Next the continuous outcome $\mathbf{z}$ was simulated as follows $\mathbf{z} = \mathbf{X}\beta^{\mathbf{X}} + \mathbf{Y}\beta^{\mathbf{Y}} + \epsilon_{\mathbf{z}}$, where β^X and β^Y are vectors of regression coefficients of length q and p respectively, and $\epsilon_z \sim N(0, 1)$. This procedure enables the generation of outcomes variables $\mathbf{z}$ affected by latent modules or biological pathways present in $\mathbf{X}$ and $\mathbf{Y}$.

In practice, we first generate β^{X*} and β^{*Y}, the regression coefficients corresponding to each independent module (given by $\mathbf{U^X}$ and $\mathbf{U^Y}$), and we then transform them to the predictor space by using $\beta^X = \mathbf{U^{Xt}}\beta^{\mathbf{X}*}$ and $\beta^Y = \mathbf{U^{Yt}}\beta^{\mathbf{Y}*}$. Next we added noise to the matrices $\mathbf{X}$ and $\mathbf{Y}$. To test the impact of various noise structures on model performance, we considered additive and multiplicative noise structures. Noise is added to $\mathbf{X}$ (analogously to $\mathbf{Y}$) as follows:

$$\mathbf{X_{noise}} = \mathbf{X} + \epsilon_X \text{ for an additive noise,}$$
$$\mathbf{X_{noise}} = \mathbf{X} \times \epsilon_X \text{ for a multiplicative noise, with} \epsilon_X \sim \mathbf{N}(0, I).$$

We simulated sets of $\mathbf{X}$ and $\mathbf{Y}$ with 100 and 1000 features organized in four and two modules of correlated features respectively. The correlations within the four modules of $\mathbf{X}$ range between 0.98 and 0.9 for the first modules, between 0.8 and 0.7 for the second module, between 0.4 and 0.1 for the third module, and between 0.6 and 0.5 for the fourth module. For $\mathbf{Y}$, the correlation within each module is between 0.9 and 0.85, and between 0.6 and 0.4 for the first and second module, respectively.

We considered three scenarios for the relationship between the phenotype $\mathbf{z}$ and the features in $\mathbf{X}$ and $\mathbf{Y}$: (Scenario a) The phenotype is simulated using only one independent principal component of the smaller dataset $\mathbf{X}$ with $\beta_j^{X*} = 0.01, j = 3$ $\beta_j^{X*} = \beta_j^{Y*} = 0$, $j \neq 3$; (Scenario b) The phenotype is simulated using both an independent principal component and the joint component of $\mathbf{X}$ and $\mathbf{Y}$ $\beta_j^{X*} = \beta_j^{Y*} = 0.01, j = 2$; $\beta_j^{X*} = 0.01$, $j = 3$; $\beta_j^{X*} = 0, j \neq 2, 3$, and $\beta_j^{Y*} = 0, j \neq 2$; (Scenario c) the phenotype is simulated using only 5 specific variables of the smaller dataset $\mathbf{X}$, i.e. $\beta_j^X = 0.1$ if j is one of the selected variables, 0 otherwise.

Finally, to investigate the impact of different noise structure on the proposed approaches, we considered for each of the three scenarios two situations, namely both datasets have an additive noise structure and X is subject to additive noise while Y is subject to multiplicative noise.

The methods are evaluated with regard to prediction performance and variable selection. Predictive ability is measured in terms of the $Q^2 = \frac{\sum_{i=1}^n (p_i - p_{0i})^2}{\sum_{i=1}^n (z_i - p_i)^2}$, the cross-validated version of the fraction of variance explained by the prediction model, in which the performance of the model-based double cross-validated predictions $\mathbf{p}$ is compared to the naive double cross-validated predictions $\mathbf{p_0}$ based on the mean value of the outcome variable $\mathbf{z}$(Rodríguez-Girondo et al., 2018). Performance with regard to variable selection is measured by the average number of truly associated features selected in the model.

6.5.2 Simulation study

Table 6.1, 6.2 and 6.3 present the obtained results for scenario a, b and c. Here, the left panels show results when both omic sources are subject to additive noise and the right panels shows the results when an additive noise structure for $\mathbf{X}$ and a multiplicative noise structure for $\mathbf{Y}$ is used.

Scenario a corresponds to the situation where solely $\mathbf{X}$ has an effect on the outcome $\mathbf{z}$ and $\mathbf{Y}$ is only indirectly associated with z via its correlation with $\mathbf{X}$. Table 6.1 shows that stacking the two datasets led to a worse predicting ability than only using the smaller dataset $\mathbf{X}$. When both datasets have the same noise structure, the largest loss was found for lasso and elastic net (from Q^2=.673 to Q^2=.560 and from Q^2=.669 to Q^2=.559, respectively), while **GLasso0** and **OverlapLasso** have a similar predictive performance as a single-omic prediction (Q^2=.673 for **GLasso0** and Q^2=.670 for **OverlapLasso**). This robustness for adding a non associated dataset was even more evident for the situation where the datasets have different noise structures. The decrease in predictive ability of **GLasso** (from Q^2=.677 to Q^2=.414), lasso (from Q^2=.673 to Q^2=.409), elastic net (from Q^2=.669 to Q^2=.406) and ridge (from Q^2=.494 to Q^2=.323) was larger than for the situation of one noise structure, while **GLasso0** and **OverlapLasso** showed almost similar results than using only dataset $\mathbf{X}$ (Q^2=.653 and Q^2=.652, respectively).

	Q^2					
	Same noise structures			Different noise structures		
Datasets	$\mathbf{X}(\mathbf{p}=\mathbf{100})$	$\mathbf{Y}(\mathbf{p}=\mathbf{1000})$	Combined datasets	$\mathbf{X}(\mathbf{p}=\mathbf{100})$	$\mathbf{Y}(\mathbf{p}=\mathbf{1000})$	Combined datasets
GLasso	.677(.155)	.385(.181)	.609(.194)	.677(.155)	.385(.170)	.414(.187)
GLasso0			.673(.154)			.653(.181)
OverlapLasso			.670(.122)			.652(.205)
Lasso	.673(.049)	.373(.049)	.560(.033)	.673(.049)	.364(.175)	.409(.167)
Elastic net	.669(.048)	.372(.048)	.559(.032)	.669(.048)	.362(.174)	.406 (.167)
Ridge	.494(.031)	.381(.025)	.482(.034)	.494(.031)	.312(.161)	.323(.152)

Table 6.1: Predictive ability performance. Results of simulation study for scenario a. Results are based on 500 replicates. Into brackets are the standard errors

Table 6.2 presents the simulation results for scenario b. Here, the shared principal component of both datasets influences the outcome variable and the relationships are stronger. As a consequence, the predictive ability of all methods was improved. When both datasets have the same noise structure (left panel) the best performing approach was **OverlapLasso** with Q^2=.971 followed by **GLasso0** and ridge with Q^2=.942 and Q^2=.918, respectively. For these three approaches, combining datasets led to an improved prediction accuracy compared to when these methods are applied to just one of the datasets. **OverlapLasso** outperforms the competing methods, especially when both datasets have an additive noise, suggesting the benefits of taking the correlation between the datasets into account when the signal comes from the shared part of $\mathbf{X}$ and $\mathbf{Y}$. Performing group inference in the stacked dataset provided relatively worse results. When $\mathbf{X}$ and $\mathbf{Y}$ have different noise structures, none of the approaches was able to improve prediction accuracy compared to solely using the dataset $\mathbf{X}$. The most strongly impacted methods were **GLasso** with a predictive accuracy reducing from Q^2=.854 when the datasets have the same noise structures to Q^2=.542, and ridge with Q^2=.918 reducing to Q^2=.319. **GLasso0** and **OverlapLasso** appeared to be more robust. The advantage of **OverlapLasso** over **Glasso0** disappeared in this setting, suggesting that our proposed approach to detect correlated groups across datasets using the principal components of omic-specific groups fails in presence of different noise structures.

Q^2						
	Same noise structures			Different noise structures		
Datasets	$\mathbf{X}(\mathbf{p}=\mathbf{100})$	$\mathbf{Y}(\mathbf{p}=\mathbf{1000})$	Combined datasets	$\mathbf{X}(\mathbf{p}=\mathbf{100})$	$\mathbf{Y}(\mathbf{p}=\mathbf{1000})$	Combined datasets
GLasso	.883(.185)	.809(.136)	.854(.144)	.883(.185)	.391.(190)	.542(.182)
GLasso0			.942(.162)			.842.(.178)
OverlapLasso			.971(.160)			.843(.195)
Lasso	.900(.033)	.761(.049)	.838(.049)	.900(.033)	.376(.180)	.763.(.083)
Elastic net	.899(.032)	.761(.048)	.836(.048)	.899(.032)	.367(.173)	.763(.083)
Ridge	.841(.034)	.806(.031)	.918(.025)	.841(.034)	.305(.146)	.319(.061)

Table 6.2: Predictive ability performance. Results of simulation study for scenario b. Results are based on 500 replicates. Into brackets are the standard errors

Table 6.3 presents the results for scenario c. In this setting, the signal comes from only a small set of almost independent variables of dataset **X**. It is, therefore, expected that lasso and elastic net perform the best. Indeed, lasso and elastic net outperformed other approaches when using only **X** (Q^2=.711 and Q^2=.699, respectively). However, lasso and elastic net performed less when applied to the stacked datasets (Q^2=.577 and Q^2=.566, respectively). Regarding the network-based approaches, **GLasso** performed worse for the combined dataset compared to the single-omic prediction based on solely **X** with a loss of prediction ability of approximately 4%. In contrast, the predictive ability remained the same when group inference is omic-specific (Q^2=.655 for **GLasso0**) and it was slightly improved when taking into account the correlation between the datasets (Q^2=.681 for **OverlapLasso** compared to $Q^2 = .654$ for **GLasso** using only **X**). As in previous scenarios, naive combinations performed worse when the datasets have different noise structures compared to the situation when they have a similar error structure. Here, **OverlapLasso** also did not outperform **GLasso0**, suggesting that the correlation structure between the datasets is not well captured. Hence **OverlapLasso** was not able to improve the prediction accuracy of **GLasso0**.

Finally, Table 6.4 presents the average number of times the truly associated variables were selected across the 500 simulations. Indeed, the best results were obtained with **GLasso** using only dataset **X**, with 2.83 of the 5 variables correctly selected in average. This value reduced to 1.62 out of 5 when combining the two datasets with the same noise structure and to 0.11 when combining datasets with different noise structures. This indicates that the Pearson correlation-based network approach was sensitive to the noise structure of the datasets and that having a mixture of noise structure yields an incorrect network structure. Although, **GLasso0** and **OverlapLasso** provided worse results than **GLasso** based solely on **X**, they were the most robust among the studied combination approaches, with respectively 2.27 and 1.58 variables correctly selected when both datasets have the same noise structures, and 1.33 and 1.33 when datasets have different noise structures. Finally, with, in average, less than one correct variable included in the prediction model, the performance of lasso and elastic net were weak, even when considering only **X**. This is probably caused by the fact that these methods tend to randomly select a subset of correlated variables which might not necessary include the associated features (Tissier et al., 2018).

	Q^2					
	Same noise structures			Different noise structures		
Datasets	$\mathbf{X}(\mathbf{p} = \mathbf{100})$	$\mathbf{Y}(\mathbf{p} = \mathbf{1000})$	Combined datasets	$\mathbf{X}(\mathbf{p} = \mathbf{100})$	$\mathbf{Y}(\mathbf{p} = \mathbf{1000})$	Combined datasets
GLasso	.654(.167)	.474(.180)	.611(.164)	.654(.167)	.389(.179)	.578(.180)
GLasso0			.655(.177)			.658(.208)
OverlapLasso			.681 (.181)			.662(.210)
Lasso	.711(.234)	.439 (.219)	.577 (.212)	.711(.234)	.370(.172)	.525(.198)
Elastic net	.699(.240)	.436 (.218)	.566(.214)	.699(.240)	.369(.172)	.521(.196)
Ridge	.640 (.267)	.428 (.198)	.417(.207)	.640(.267)	.312(.152)	.475(.175)

Table 6.3: Predictive ability performance. Results of simulation study for scenario c. Results are based on 500 replicates. Into brackets are the standard errors

	Same noise structures		Different noise structures	
	Only using X	Combined datasets	Only using X	Combined datasets
GLasso	2.83	1.62	2.83	0.11
GLasso0		2.27		1.33
OverlapLasso		1.58		1.33
Lasso	0.89	0.62	0.89	0.80
Elastic net	0.98	0.70	0.98	0.93

Table 6.4: Variable selection. Results from simulation study for scenario c. Average number of true signal carrying variables selected accross the 10-fold cross-validation and the 500 simulations. The phenotype was simulated using only 5 variables from dataset X .

6.6 Real data analysis

To illustrate and compare the performance of the proposed approaches on real data, we analyzed data from the DILGOM study and from the breast cancer cell line pharmacogenomics dataset, introduced in Section 1. For both cases, we aim to build a single prediction model based on two omic datasets. In the DILGOM study we consider NMR metabolites and gene expression profiles, both measured at baseline, to predict BMI after seven years of follow-up. The analyzed sample contained n =258 individuals. In the breast cancer cell lines dataset, we were interested in using gene expression and copy number variants for predicting the treatment response of the Erlotinib drug. Treatment response is measured through the GI50 index, a quantitative measure which quantifies the growth inhibitory power of the test agent. The analysed sample consisted of 45 breast cancer cell lines.

6.6.1 DILGOM

The NMR metabolomic data consists of quantitative information on 57 metabolic measures, mainly composed of measures on different lipid subclasses, but also amino acids, and creatine. The set of gene expression profiles consist of 2980 probes. Tables 6.5 and 6.6 present the main results for the prediction of BMI after 7 years of follow-up. Table 6.5 shows the performance of each method in terms of predictive ability measured through Q^2. All network-based approaches, applied to both datasets simultaneously, perform similarly (Q^2=.414, Q^2=.425, and Q^2=.422 for **Glasso**, **Glasso0** and **OverlapLasso**, respectively) and better than lasso, elastic net and ridge (with $Q^2 = .295$, $Q^2 = .307$, and $Q^2 = .104$ respectively). Comparing combined with single omic approaches, these three group based approaches perform similarly or better.

Table 6.6 presents the main results with regards to variable selection properties of the different approaches. We can see that the network based approaches selected more

		Q^2	
Datasets	Metabolites (p=57)	Gene Expression (p=2980)	Combined datasets
GLasso	.241	.418	.414
GLasso 0			.425
Overlap Lasso			.422
Lasso	.227	.257	.295
ENet	.208	.265	.307
Ridge	.222	.131	.104

Table 6.5: Predicitive ability. Results obtained from analyzing the DILGOM datasets. Predictive ability obtained for each dataset and for the combined dataset for each method.

	GLasso	GLasso0	Overlap Lasso	Lasso	Enet	Ridge
Variables not selected	2639	2565	2562	2937	2923	0
Variables selected at least once*	343(0.11)	417(0.13)	420(0.14)	45(0.01)	59(0.02)	2982(1.00)
Variables always selected**	71(0.20)	210(0.50)	210(0.50)	14(0.31)	20(0.33)	2982(1.00)

Table 6.6: Variable selection. Results obtained from analyzing the DILGOM datasets. Number of variables selected selected or not during the cross-validation process. * into brackets are the proportion of variable selected in the total set of variables available. ** into brackets are the proportion of variable selected in the set of variable selected at least once

variables than lasso and elastic net with a proportion of variables selected at least once between 0.11 and 0.14 during the cross-validation process for the network approaches and between 0.01 and 0.02 for lasso and elastic net, respectively. With regards to stability, from the variables selected at least one time a larger proportion was selected all the time for **GLasso0** and **OverlapLasso** (0.5) compared to elastic net and **GLasso** ($\simeq$ 0.3). Finally, we observe that **OverlapLasso** and **GLasso0** had almost the same prediction ability, moreover, they selected the same variables suggesting that there is almost no correlation between the two omic datasets.

6.6.2 Breast cancer cell lines

The first omic dataset consists of quantitative information on almost 100,000 different copy number variants. After filtering copy number variants present in all the 45 breast cancer cell lines and keeping only one variable per gene, we obtained 637 copy number variants. The second omic set consist of a set of 5375 probes. Tables 6.7 and 6.8 present the main results for the prediction of the treatment response of the Erlotinib drug. Table 6.7 shows the performance of each method in terms of predictive ability measured by Q^2.

	Q^2		
Datasets	CNV	Gene Expression	Combined datasets
GLasso	.476	.651	.504
GLasso 0			.905
Overlap Lasso			.933
Lasso	.934	.571	.576
ENet	.836	.564	.563
Ridge	.454	.610	.614

Table 6.7: Predicitive ability. Results obtained from analyzing the Breast cancer cell lines. Predictive ability obtained for each dataset and for the combined dataset for each method.

	GLasso	GLasso0	Overlap Lasso	Lasso	Enet	Ridge
Variables not selected	5639	5448	5467	5953	5940	0
Variables selected at least once*	373(0.06)	564(0.09)	545(0.09)	59(0.01)	72(0.01)	6012(1.00)
Variables selected at least 5 times**	22(0.04)	34(0.06)	43(0.11)	2(0.03)	2(0.02)	6012(1.00)
Variables always selected**	0(0.00)	0(0.00)	0(0.00)	2(0.02)	2(0.02)	6012(1.00)

Table 6.8: Data analysis: Breast cancer cell lines. Number of variables selected selected or not during the corss-validation process. * into brackets are the proportion of variable selected in the total set of variables available. ** into brackets are the proportion of variable selected in the set of variable selected at least once

The performance of **OverlapLasso** and of **GLasso0** was similar to the best performing method applied to one dataset (CNVs, Lasso with $Q^2 = .934$). Note that **OverlapLasso** showed a little better results than **GLasso0**. This is probably due to the presence of correlation between the datasets. Figure 6.4 present the network of principal components obtained after clustering the datasets separately, highlighting the presence of existing correlation between groups from different omic sources.

Table 6.8 shows the results with regards to variable selection properties of the different approaches. We observe that only lasso and elastic net select variables all the times during the cross-validation process. These two methods selected the same two variables all the times. However, when looking at variables selected at least 5 times in the 10-fold cross-validation process, we observe that the network approaches selected more variables than lasso and ridge. Namely, **GLasso**, **GLasso0** and **OverlapLasso** selected 22, 34 and 43 variables half of the time, respectively. Note that **OverlapLasso** is the most stable approach when looking at the proportion of variable selected at least 5 times among the variables selected (0.11).

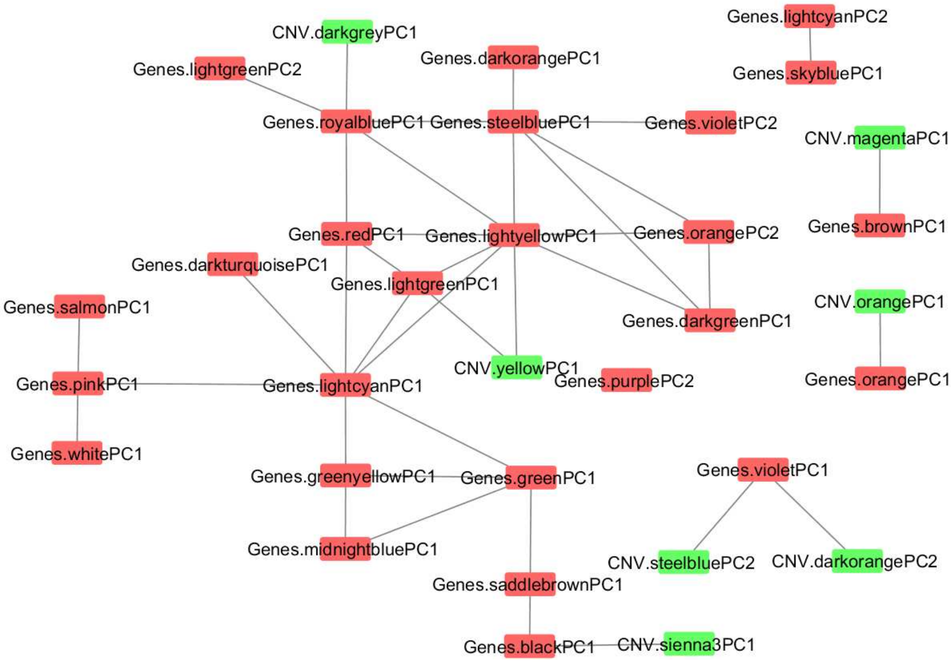


Figure 6.4: Data analysis: Breast cancer cell lines. Network of principal components of the modules identified in each datasets. In red are the principal components of the transcriptomics modules and in green from the CNVs module. Edges are showing existing correlations between principal components. Colors in the names of the different nodes represent different modules of omic variables.

6.7 Discussion

In this paper, we proposed a new strategy to integrate two omics datasets in a prediction model. The key components of our proposed approach are to capture groups of correlated features within and between omic datasets and to include it in a group penalization model. Simulations results showed that naively stacking datasets is usually not a good strategy as it often perform worse than a model based on a single omic datasets. This result confirmed previous research of our group (Rodríguez-Girondo et al., 2018). As seen in the simulations (scenario b), including information about the correlation between the omic datasets, through the inclusion of overlapping groups (**OverlapLasso**), might improve the prediction accuracy. Therefore, when the studied omic datasets have a similar noise structure, **OverlapLasso** is the recommended method to build prediction models. The loss of prediction accuracy of **GLasso** when the two datasets have different noise structures, suggests that correlation-based network approaches might be sensitive to different noise distributions. Building the networks and performing clustering on each omic sets provided the most robust results especially in the presence of different noise structure. In case of different noise structure for the studied datasets, **OverlapLasso** can also be the recommended method if both datasets are considered. However in this situation a model based on only one dataset might still be the best.

The analysis of the datasets of the two applications confirm the simulation results. Indeed, in the case of the the NMR metabolites and gene expression profiles of DILGOM we clearly see that using **GLasso0** and **OverlapLasso** provided the best results. In this particular case, stacking the two sets of omics in a naive way using lasso and elastic net also improves the predictive accuracy, which was not the case in our simulations. This might be explained by the fact that the features within each omic set are less correlated than in our simulations settings and especially lasso performs better in this case. Finally, the fact that **GLasso0** and **OverlapLasso** provided the same predictive ability, is probably due to a weak correlation between the two different types of omics. For the breast cancer cell lines, where the correlation between the omics set is stronger, **OverlapLasso** provides a better predictive accuracy than **GLasso0** which agrees with the results obtained in the simulation scenarios. Overall the network based approaches were more stable when analyzing the two datasets jointly especially for the DILGOM datasets where a large proportion of variables selected at least once were actually allways selected during the cross-validation process.

To conclude, we presented a strategy to integrate two different omic sets of features into a prediction model in order to improve the prediction ability of single-omic based models. Our approach is highly flexible and several types of group penalization methods or network analysis approaches can be used. When the noise structure of the two datasets is similar and the signal comes from the joint component of the two datasets the principal components approach to build clusters containing features of both datasets showed some improvement. More research is needed for identifying the best method to detect possible correlation between groups of omic features from different sources, especially when these sources are subject to different error structures. Another topic of future work is to formally assess added predictive value using network and group penalization (Rodríguez-Girondo et al., 2018).

Bibliography

Acharjee, A., B. Kloosterman, R. G. Visser, and C. Maliepaard (2016). Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC Bioinformatics 17 Suppl 5*.

Almasy, L. and J. Blangero (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *The American Journal of Human Genetics 62(5)*, 1198 – 1211.

Association., A. P. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Arlington, VA: American Psychiatric Publishing.

Bahado-Singh, R., R. Ertl, R. Mandal, T. C. Bjorndahl, A. Syngelaki, B. Han, E. Dong, P. B. Liu, Z. Alpay-Savasan, D. S. Wishart, and K. H. Nicolaides (2014). Metabolomic prediction of fetal congenital heart defect in the first trimester. *American Journal of Obstetrics & Gynecology 211(3)*.

Balliu, B., R. Tsonaka, S. Boehringer, and J. Houwing-Duistermaat (2015). A retrospective likelihood approach for efficient integration of multiple omics factors in case-control association studies. *Genetic Epidemiology 39(3)*, 156 – 165.

Bas-Hoogendam, J. M., A. Harrewijna, R. L. M. Tissier, M. J. W. van der Molena, H. van Steenbergen, I. M., V. Vliet, C. G. Reichart, J. J. Houwing-Duistermaat, E. Slagboom, N. J. A. van der Wee, and M. Westenberg (2018). The Leiden Family Lab study on Social Anxiety Disorder: a multiplex, multigenerational family study on neurocognitive endophenotypes. *International Journal of Methods in Psychiatric Research (In press)*.

Beaty, T. H. and K. Y. Liang (1987). Robust inference for variance components models in families ascertained through probands: I. conditioning on proband's phenotype. *Genetic Epidemiology 4*, 203 – 210.

Boehnke, M. and D. A. Greenberg (2018). The leiden family lab study on social anxiety disorder: a multiplex, multigenerational family study on neurocognitive endophenotypes. *International Journal of Methods in Psychiatric Research In press*.

Bouhaddani, S., P. Houwing-Duistermaat, J. J. andd Salo, M. Perola, G. Jongbloed, and H.-W. Uh (2016). Evaluation of o2pls in omics data integration. *BMC Bioinformatics 17(Suppl 2)*, S11.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Berlin: Springer.

Cheung, C. Y., E. A. Thompson, and E. M. Wijsman (2013). Gigi: an approach to effective imputation of dense genotypes on large pedigrees. *The American Journal of Human Genetics 92(4)*, 504 – 516.

Chuang, H.-Y., E. Lee, Y.-T. Liu, D. Lee, and T. Ideker (2007). Network-based classification of breast cancer metastasis. *Molecular Systems Biology 3:140.*

Compier-de Block, L. H. C. G., L. R. A. Alink, M. Linting, L. J. M. van den Berg, B. M. Elzinga, A. Voorthuis, M. S. Tollenaar, and M. J. Bakermans-Kranenburg (2017). Parent-child agreement on parent-to-child maltreatment. *Journal of Family Violence 32*(2), 207–217.

de Andrade, M. and C. I. Amos (2000). Ascertainment issues in variance components models. *Genetic epidemiology 19*, 333 – 344.

de Jong, S., M. P. M. Boks, T. F. Fuller, E. Strengman, E. Janson, C. G. F. de Kovel, A. P. S. Ori, N. Vi, F. Mulder, J. D. Blom, B. Glenthøj, C. D. Schubart, W. Cahn, R. S. Kahn, S. Horvath, and R. A. Ophoff (2012). A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes. *PLoS One 7*(6), 1–10.

de Visser, M. C., R. van Minkelen, V. van Marion, M. den Heijer, J. Eikenboom, H. L. Vos, P. E. Slagboom, J. J. Houwing-Duistermaat, F. R. Rosendaal, and R. M. Bertina (2013). Genome-wide linkage scan in affected sibling pairs identifies novel susceptibility region for venous thromboembolism: Genetics in familial thrombosis study. *Journal of Thrombosis and Haemostasis 11(8)*, 1474–84.

Distel, M. A., J. M. Vink, G. Willemsen, C. M. Middeldorp, H. Merckelbach, and D. I. Boomsma (2008). Heritability of self-reported phobic fear. *Behavior Genetics 38*, 24 – 33.

Dubdbridge, F. (2003). Power and predictive accuracy of polygenic risk scores. *The American Journal of Psychiatry 160(4)*, 636 – 645.

Efron, B. (2004). Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association 99*, 96–104.

Elston, R. C. and J. Stewart (2013). A general model for the analysis of pedigree data. *Human Heredity 21*, 523–542.

Fisher, R. A. (1924). The distribution of the partial correlation coefficient. *Metron 3*, 329–332.

Friedman, J., T. Hastie, and R. Tibshirani (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9(3)*, 432–441.

Friedrichs, S., J. Manitz, P. Burger, C. I. Amos, A. Risch, and J. e. a. Chang-Claude (2017). Pathway-based kernel boosting for the analysis of genome-wide association studies. *Computational and Mathematical Methods in Medicine*, Article ID 6742763.

Fuady, A. M., R. Tissier, and J. J. Houwing-Duistermaat (2018). Genome-wide analysis in multiple-case families: assessing the relationship between triglyceride and methylation. *BMC Proceedings*.

Furmark, T. (2002). Social phobia: overview of community surveys. *Acta Psychiatrica Scandinavica 105*, 84 – 93.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics 1*, 141 – 150.

Ghosh, A., F. A. Wright, and F. Zou (2013). Unified analysis of secondary traits in case-control association studies. *Journal of the American Statistical Association 108(52)*, 140 – 151.

Ghosh, D. and A. M. Chinnaiyan (2005). Classification and selection of biomarkers in genomic data using lasso. *Journal of Biomedicine and Biotechnology 2005(2)*, 147–154.

Glahn, D. C., P. M. Thompson, and J. Blangero (2007). Neuroimaging endophenotypes: Strategies for finding genes influencing brain structure and function. *Human Brain Mapping 28*, 488 – 501.

Gottesman, I. I. and T. D. Gould (2003). The endophenotype concept in psychiatry: Etymology and strategic intentions. *American Journal of Psychiatry 160*, 636 – 645.

Gratten, J. and P. M. Visscher (2016). Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Medicine 8*, 78.

Greenwood, T. A., D. L. Braff, G. A. Light, K. S. Cadenhead, M. E. Calkins, D. J. Dobie, R. Freedman, M. F. Green, R. E. Gur, R. C. Gur, J. Mintz, K. H. Nuechterlein, A. Olincy, A. D. Radant, L. J. Seidman, L. J. Siever, J. M. Silverman, W. S. Stone, N. R. Swerdlow, D. W. Tsuang, M. T. Tsuang, B. I. Turetsky, and N. J. Schork (2007). Initial heritability analyses of endophenotypic measures for schizophrenia: the consortium on the genetics of schizophrenia. *Archives of General Psychiatry 64(11)*, 1242 – 1250.

Ha, M. J. and W. Sun (2014). Partial correlation matrix estimation using ridge penalty followed by thresholding and re-estimation. *Biometrics 70(3)*, 765–773.

Hardin, J., S. R. Garcia, and D. Golan (2013). A method for generating realistic correlation matrices. *The Annals of Applied Statistics 7(3)*, 1733 – 1762.

Harrewijn, A., M. J. W. van der Molen, and P. M. Westenberg (2016). Putative eeg measures of social anxiety: Comparing frontal alpha asymmetry and deltaâĂŞbeta cross-frequency correlation. *Cognitive, Affective, & Behavioral Neuroscience 6*, 1086–1098.

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

He, J., H. Li, A. C. Edmonson, D. J. Rader, and M. Li (2011). A gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies. *Biostatistics 13(3)*, 497 – 508.

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

Hopper, J. L. and J. D. Mathews (1982). Extensions to multivariate normal models for pedigree analysis. *Annals of Human Genetics 46*, 373 – 383.

Houwing-Duistermaat, J. J., A. Callegaro, M. Beekman, R. G. Westendorp, P. E. Slagboom, and J. C. van Houwelingen (2009). Weighted statistics for aggregation and linkage analysis of human longevity in selected families: the Leiden Longevity Study. *Statistics in Medicine 28(1)*, 140 – 151.

Iacono, W. G., S. M. Malone, and S. I. Vrieze (2017). Endophenotype best practices. *International Journal of Psychophysiology 111*, 115 – 144.

Ibrahim-Verbaas, C. A., M. Fornage, J. C. Bis, S. H. Choi, B. M. Psaty, J. B. Meigs, M. Rao, M. Nalls, J. D. Fontes, and C. J. e. a. OâĂŹDonnell (2014). Predicting stroke through genetic risk functions: the charge risk score project. *Stroke 45 (2014)*, 403–412.

(IMSGC), I. M. S. G. C., W. S. Bush, S. J. Sawcer, P. L. de Jager, J. R. Oksenberg, J. L. McCauley, M. A. Pericak-Vance, and J. L. Haines (2010). Evidence for polygenic susceptibility to multiple sclerosisâĂŞthe shape of things to come. *The American Journal of Human Genetics 86(4)*, 421 – 425.

Inouye, M., J. Kettunen, P. Soininen, K. Silander, S. Ripatti, and L. S. e. a. Kumpula (2010). Metabonomic, transcriptomic, and genomic variation of a population cohort. *Molecular Systems Biology 21(6).*

International Schizophrenia Consortium, ., P. S. M., N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, and P. Sklar (2009, jul). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature 460(7256)*, 748 – 752.

Irvin, M. R., D. Zhi, R. Joehanes, M. Mendelson, S. Aslibekyan, S. A. Claas, K. S. Thibeault, N. Patel, K. Day, L. W. Jones, L. Liang, B. H. Chen, C. Yao, H. K. Tiwari, J. M. Ordovas, D. Levy, D. Absher, and D. K. Arnett (2014). Epigenome-wide association study of fasting blood lipids in the genetics of lipid-lowering drugs and diet network study: Clinical perspective. *Circulation 130*(7), 565–572.

Isomura, K., M. Boman, C. Ruck, E. Serlachius, H. Larsson, P. Lichtenstein, and D. Mataix-Cols (2015). Population-based, multi-generational family clustering study of social anxiety disorder and avoidant personality disorder. *Psychological Medicine 45*, 1581 – 1589.

Jacob, L., G. Obozinski, and J.-P. Vert (2009). Group lasso with overlap and graph lasso. *Proceedings of the International Conference on Machine Learning (ICML) ICML '09.*

Jolliffe, I. T. (2008). *Principal Component Analysis.* New York: Springer-Verlag.

Kendler, K. S., M. C. Neale, R. C. Kessler, A. C. Heath, and L. J. Eaves (1992). The genetic epidemiology of phobias in women - the interrelationship of agoraphobia, social phobia, situational phobia, and simple phobia. *Archives of General Psychiatry 49*, 273 – 281.

Kivelä, M., A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Proter (2014). Multilayer networks. *Journal of complex networks 2(3)*, 203–271.

Kraft, P., E. Schadt, J. Aten, and S. Horvath (2003, May). A family-based test for correlation between gene expression and trait values. *The American Journal of Human Genetics 72*(5), 1323–1330.

Kraft, P. and D. C. Thomas (2000). Bias and efficiency in family-based gene-characterization studies: Conditional, prospective, retrospective, and joint likelihoods. *The American Journal of Human Genetics 66*(3), 1119–1131.

Krumsiek, J., K. Suhre, T. Illig, J. Adamski, and F. J. Theis (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology 5:21.*

Langfelder, P. and S. Horvath (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics 9*(1), 559.

Langfelder, P., B. Zhang, and S. Horvath (2008). Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Journal of the American Statistical Association 99*, 96–104.

Lauritzen, S. L. (1996). *Graphical models.* Oxford, Clarendon Press: Oxford statistical science series.

Lee, A., L. McMurchy, and A. J. Scott (1997). Re-using data from case-control studies. *Statistics in Medicine 16(12)*, 1377 – 1389.

Lemesle, G., F. Maury, O. Beseme, L. Ovart, P. Amouyel, N. Lamblin, P. de Groote, C. Bauters, and F. Pinet (2015). Multimarker proteomic profiling for the prediction of cardiovascular mortality in patients with chronic heart failure. *PLoS One 10(4)*.

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

Li, H. and G. M. H. (2012). Efficient adaptively weighted analysis of secondary phenotypes in case-control genome-wide association studies. *Human Heredity 73*, 159 – 173.

Lin, D. Y. and D. Zeng (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology 33*, 256 – 265.

Liu, J., T.-Z. Huang, Z. Xu, and X.-G. Lv (2013). High-order total variation-based multiplicative noise removal with spatially adapted parameter selection. *Journal of the Optical Society of America A 30*, 1956–1966.

Mertens, B. J. A., M. E. de Noo, R. A. E. M. Tollenaar, and A. M. Deelder (2006). Mass spectrometry proteomic diagnosis: enacting the double crossvalidatory paradigm. *Journal of Computational Biology 13*, 1591–1605.

Mertens, B. J. A., Y. E. M. van der Burgt, B. Velstra, W. E. Mesker, R. A. E. M. Tollenaar, and A. M. Deelder (2011). On the use of double crossvalidation for the combination of proteomic mass spectral data for enhanced diagnosis and prediction. *Statistics and Probability Letters 81*, 759–766.

Miller, G. A. and B. Rockstroh (2013). Endophenotypes in psychopathology research: where do we stand? *Annual Review of Clinical Psychology 9*, 177 – 213.

Miwa, T., A. J. Hayer, and S. Kuriki (2003). The evaluation of general non-centred orthant probabilities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*, 223 – 234.

Monsees, G. M., R. M. Taqmimi, and P. Kraft (2009). Genome-wide association scans for secondary traits using case-control samples. *Genetic Epidemiology 33*, 717 – 728.

Mooijaart, S. P., D. van Heems, R. Noordman, M. P. Rozing, C. A. Wijsman, A. J. M. de Craen, R. G. J. Westendorp, M. Beekman, and E. P. Slagboom (2010). Polymorphisms associated with type 2 diabetes in familial longevity: The Leiden Longevity Study. *Aging 3*, 55 – 62.

Mootha, V. K., C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, and J. e. a. Lehar (2003). Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics 34*, 267–273.

Najita, J. S., Y. Li, and P. J. Catalano (2009). A novel application of a bivariate regression model for binary and continuous outcomes to studies of fetal toxicity. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 58(4)*, 555 – 573.

Noah, S., J. Friedman, T. Hastie, and R. Tibshirani (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics 22:2*, 231–245231–245.

Oldham, M., S. Horvath, and D. Geschwind (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences 103(47)*, 17973 – 17978.

Oldham, M., G. Konopka, K. Iwamoto, P. Langfelder, T. Kato, S. Horvath, and D. Geschwind (2008). Functional organization of the transcriptome in human brain. *Nature Neuroscience 11(11)*, 1271 – 1282.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine 2(11)*, 559 – 572.

Pena, M. J., A. Heinzel, P. Rossing, H. Parving, G. Dallma, and K. e. a. Rossing (2016). Serum metabolites predict response to angiotensin ii receptor blockers in patients with diabetes mellitus. *Journal of Translational Medicine 14*, 203.

Plaisier, C. L., S. Horvath, A. Huertas-Vazquez, I. Cruz-Bautista, M. F. Herrera, T. Tusie-Luna, C. Aguilar-Salinas, and P. Pajukanta (2009). A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia. *PLoS Genetics 5*(9), 1–10.

Rao, K. R. and S. Lakshminarayanan (2007). Partial correlation based variable selection approach for multivariate data classification methods. *Chemometrics and Intelligent Laboratory Systems 86(1)*, 68–81.

Reis, M. S. (2013). Applications of a new empirical modelling framework for balancing model interpretation and prediction accuracy through the incorporation of clusters of functionally related variables. *Chemometrics and Intelligent Laboratory Systems 127*, 7– 16.

Richardson, D. B., P. Rzehak, J. Klenk, and S. K. Weiland (2007). Analyses of case-control data for additional outcomes. *Epidemiology 8(4)*, 441 – 445.

Rodríguez-Girondo, M., J. Deelen, P. E. Slagboom, and J. J. Houwing-Duistermaat (2018). Survival analysis with delayed entry in selected families with application to human longevity. *Statistical Methods in Medical Research 27*(3), 933–954.

Rodríguez-Girondo, M., P. Salo, T. Burzykowsky, M. Perola, J. J. Houwing-Duistermaat, and B. Mertens (2018). Sequential double cross-validation for augmented prediction assessment in high-dimensional omic applications. *Annals of Applied Statistics*.

Sásik, R., E. Calvo, and J. Corbeil (2002). Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. *Bioinformatics 18(12)*, 1633–1640.

Schade, D. S. and P. R. Eaton (1974). Role of insulin and glucagon in obesity. *Diabetes 23(8)*, 657–661.

Schäfer, J. and K. Strimmer (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology 4*, Art. 32.

Schifano, E. D., L. Li, D. C. Christiani, and X. Lin (2013). Genome-wide association analysis for multiple continuous secondary phenotypes. *The American Journal of Human Genetics 92(5).*

Schoenmaker, M., A. J. M. de Craen, P. H. E. M. de Meijer, M. Beekman, G. J. Blauw, P. E. Slagboom, and R. G. J. Westendorp (2006). Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *European Journal Of Human Genetics 14*, 79 EP –.

Shahabi, A., J. P. Lewinger, J. Ren, C. April, A. E. Sherrod, and J. G. e. a. Hacia (2006). Novel gene expression signature predictive of clinical recurrence after radical prostatectomy in early stage prostate cancer patients. *Prostate 76(14)*, 1239–1256.

Shamai, L., E. Lurix, M. Shen, G. M. Novaro, S. Szomstein, and R. e. a. Rosenthal (2011). Association of body mass index and lipid profiles: evaluation of a broad spectrum of body mass index patients including the morbidly obese. *Obesity Surgery 21(1)*, 42–47.

Shim, J. E., C. Bang, S. Yang, T. Lee, S. Hwang, and C. Y. e. a. Kym (2017). Gwab: a web server for the network-based boosting of human genome-wide association data. *Nucleic Acids Research 45(1)*, W154–W161.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics 22(2)*, 231–245.

Simonson, M. A., A. G. Wills, M. C. Keller, and M. B. McQueen (2011). Recent methods for polygenic analysis of genome-wide data implicate an important effect of common variants on cardiovascular disease risk. *BMC Medical Genetics 12*, 146.

Skytthe, A., S. Valensin, B. Jeune, E. Cevenin, F. Balard, M. Beekman, V. Bezrukov, H. Blanch, L. Bolund, K. Broczek, C. Carru, K. Christensen, L. Christiansen, J. C. Collerton, and R. Cotichini (2011). Design, recruitment, logistics, and data management of the GEHA (genetics of healthy ageing) project. *Experimental Gerontology 46(11)*, 934–945.

Solovieff, N., C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics 14*, 483–495.

Stuart, J. M., E. Segal, D. Koller, and S. K. Kim (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science 302(5643)*, 249 – 255.

Subramaniana, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Eberta, and M. A. e. a. Gillette (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences 21(1)*, 15545–15550.

Therneau, T. M. (2018). *coxme: Mixed Effects Cox Models.* R package version 2.2-7.

Thomas, D. C. (2004). *Statistical Methods in Genetic Epidemiology.* Oxford: Oxford University Press.

Thompson, E. A. (2008). The ibd process along four chromosomes. *Theoretical Population Biology 73*(3), 369 – 373.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B 58(1)*, 267–288.

Tissier, R., J. J. Houwing-Duistermaat, and M. Rodríguez-Girondo (2018). Improving stability of prediction models based on correlated omics data by using network approaches. *PLoS One 13(2)*, e0192853.

Tissier, R., R. Tsonaka, S. P. Mooijaart, E. Slagboom, and J. J. Houwing-Duistermaat (2017). Secondary phenotype analysis in ascertained family designs: application to the leiden longevity study. *Statistics in Medicine 36(14)*, 2288 – 2301.

Trygg, J. and S. Wold (2003). O2-pls, a two-block (xâĂŞy) latent variable regression (lvr) method with an integral osc filter. *Journal of Chemometrics 7(1)*, 53–64.

Tsonaka, R., M. C. H. de Visser, and J. J. Houwing-Duistermaat (2013). Estimation of genetic effects in multiple cases family studies using penalized maximum likelihood methodology. *Biostatistics 14(2)*, 220 – 231.

Tsonaka, R., D. van der Woude, and J. J. Houwing-Duistermaat (2015). Marginal genetic effects estimation in family and twin studies using random-effects models. *Biometrics 71(4)*, 1130 – 1138.

Turetsky, B. I., T. A. Greenwood, A. Olincy, A. D. Radant, D. L. Braff, K. S. Cadenhead, D. J. Dobie, R. Freedman, M. F. Green, R. E. Gur, R. C. Gur, G. A. Light, J. Mintz, K. H. Nuechterlein, N. J. Schork, L. J. Seidman, L. J. Siever, J. M. Silverman, W. S. Stone, N. R. Swerdlow, D. W. Tsuang, M. T. Tsuang, and M. E. Calkins (2015). Abnormal auditory n100 amplitude: a heritable endophenotype in first-degree relatives of schizophrenia probands. *Biol Psychiatry 64(12)*, 1051 – 1059.

van de Wiel, M. A., T. G. Lien, W. Verlaat, W. N. van Wieringen, and S. M. Wilting (2014). Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in medicine 35(3).*

Welter, D., J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson (2014). The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic Acids Research 42(Database Issue)*, D1001–D1006.

Winter, C., G. Kristiansen, S. Kersting, J. Roy, D. Aust, and T. e. a. Knösel (2012). Google goes cancer: Improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Computational Biology*.

Yip, A. M. and S. Horvath (2007). The generalized topological overlap matrix for detecting modules in gene networks. *BMC Bioinformatics 8(22)*.

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

Zemmour, C., F. Bertucci, P. Finetti, B. Chetrit, T. Filleron, and J. M. Boher (2015). Prediction of early breast cancer metastasis from dna microarray data using high-dimensional cox regression models. *Cancer Informatics 14(Suppl 2)*, 129–138.

Zhang, B. and S. Horvath (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology 4*, Article17.

Zhu, Y., X. Shen, and W. Pan (2009). Network-based support vector machine for classification of microarray samples. *BMC Bioinformatics 10(Suppl I)*, S21.

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

List of Publications

Tissier R.L.M., Uh H.W., van den Akker E., Balliu B., Tsonaka S., Houwing-Duistermaat J.J. (2016). Gene coexpression network analysis for family studies based on a meta-analytic approach. *BMC Proceedings 10*, (Suppl 7): 119–123.

Tissier R.L.M., Tsonaka S., Mooijaart S.P., Slagboom E., Houwing-Duistermaat J.J. (2017). Secondary phenotype analysis in ascertained family designs: application to the Leiden longevity study. *Statistics in Medicine 36*, (14): 2288–2301.

Tissier R.L.M., Houwing-Duistermaat J.J., Rodrìguez-Girondo M. (2018). Improving stability of prediction models based on correlated omics data by using network approaches. *Plos One 13*, (2): e0192853.

Tissier R.L.M., Rodrìguez-Girondo M., Houwing-Duistermaat J.J. Integration of several omic sources in prediction models using network-based approaches. Submitted for publication.

Tissier R.L.M, Houwing-Duistermaat J.J. Statistical methods for the analysis of secondary phenotypes in family proband designs. Submitted for publication.

Fuady A M, **Tissier R.L.M.**, Houwing-Duistermaat J.J. (2018). Genome-wide analysis in multiple-case families: assessing the relationship between triglyceride and methylation. *BMC Proceedings*, 2(Suppl 9):67–71.

Harrewijn A., van der Molen M.J.W., van Vliet I.M., **Tissier R.L.M.**, Westenberg P.M. (2018). Behavioral and EEG responses to social evaluation: A two-generation family study on social anxiety. *Neuroimage Clin*, (17): 549–562.

Bas-Hoogendam J.M., Harrewijn A., **Tissier R.L.M.**, van der Molen M.J.W., van Steenbergen H., Van Vliet I.M., Reichart C.G., Houwing-Duistermaat J.J., Slagboom E., van der Wee N.J.A., Westenberg P.M. (2018). The Leiden Family Lab study on Social Anxiety Disorder: a multiplex, multigenerational family study on neurocognitive endophenotype. *International Journal of Methods 27*, (2):e1616

Bas-Hoogendam J.M., van Steenbergen H., **Tissier R.L.M.**, Houwing-Duistermaat J.J., Westenberg P.M., van der Wee N.J.A. (2018). Subcortical brain volumes, cortical thickness and cortical surface area in families genetically enriched for social anxiety disorder - A multiplex multigenerational neuroimaging study. *EBioMedicine*, epub ahead of print.

Summary

This dissertation focuses on the development of new statistical methods designed to take into account existing structures inside omic datasets. The major challenge in analysing omic datasets is the strong dependencies which are present. Taking into account and modelling the different dependency structures can lead to further improvements of our knowledge of the biological mechanisms. Therefore, improving our ability to predict diseases.

Chapter 1 provides a general introduction to the existing dependency structures possibly faced when studying omic datasets. First, the most common measure of dependence is described, i.e. the Pearson correlation coefficient. Next, the different dependency structures are described. Namely, dependencies between individuals, between outcome measures and between omic features. For each of these dependency levels the challenges faced and the commonly used methods are described.

Chapters 2 and 3 present methods for the analysis of secondary phenotypes in ascertained family studies. Chapter 2 presents a new approach to analyse secondary phenotype for the multiple case family design. Where families are selected when they have at least a specific number of cases. The proposed method is illustrated by a data example obtained from the Leiden Longevity Study, which is a multiple-cases family study that investigates human longevity (primary phenotype). Here the association between, triglyceride levels and glucose (secondary phenotypes), and genetic markers was estimated. Chapter 3 presents methods used in the literature for secondary phenotype analysis for the proband family design. This design comprises family members of specific probands (often cases with the primary outcome). These methods are then compared with the method previously developed in Chapter 2. The real data analysis presented in this chapter is part of the Social Anxiety Disorder (SAD) family study, and aims to identify possible endophenotypes of SAD.

Chapter 2 develops an approach to obtain unbiased association estimates between secondary phenotypes and biomarkers as well as unbiased heritability estimates of the available secondary phenotypes. This method accommodates the ascertainment process while explicitly modelling the familial relationships. To do so, Our approach uses the retrospective likelihood in order to correct for the ascertainment process with existing methods for mixed-effects models. The retrospective likelihood approach automatically corrects for the ascertainment. A multivariate probit model is used to capture the association between the mixed type primary (binary variable) and secondary phenotypes (continuous variable).

Estimates are then obtained by maximizing the log-likelihood.

An important empirical finding is that the heritability estimates for the secondary traits can be severely underestimated unless the sampling mechanism is taken into account. Extensive simulations show that the presented method preserves the type I error at nominal level and provides accurate estimates irrespective of the disease prevalence, the strength of the association between the genetic variants and the primary phenotype, and the ascertainment mechanism. Currently, a key limitation of this approach is the computational time of multivariate integrals, especially in case of large pedigrees.

Chapter 3 investigates the performances of the previous method, from Chapter 2, for the analysis of proband family study design. Theses performances are compared with methods currently used in the litterature. Namely, ignoring the ascertainment process and modelling the conditional distribution of the secondary phenotype values of the families given the secondary phenotypes of the probands. Furthermore, we propose an extension of the latter approach, by modelling the joint conditional distribution of the primary and secondary phenotype values of the families given the joint distribution of the primary and secondary phenotypes of the probands.

Extensive simulations show that only the retrospective likelihood approach developed in Chapter 2 is able to obtain unbiased heritability estimates of the secondary phenotype as well as association estimates of the secondary phenotypes with genetic markers. Furthermore, conditioning on the secondary phenotype values of the proband can severely underestimate heritability estimates and therefore limiting the identification of candidate endophenotypes of primary phenotypes. Only the retrospective likelihood approach could identify a candidate endophenotypes of SAD in the real data analysis . Another important key point of this chapter is that current methods provide biased estimates when the proband information is missing. Therefore, the use of such study design should not, at this time, be considered.

Chapter 4 considers the problem of conducting gene co-expression network analysis for family studies. A large between-family variation in expression levels could severely bias the network structure obtained if the pedigree structure is not taken into account. To overcome this issue, we propose a meta-analytic approach. We first build the omic network for each pedigree to identify clusters of correlated microarray probes. The eigengene (first principal component) of each cluster of each pedigree are then tested for association with a phenotype of interest. After identification of the most strongly associated cluster, clusters presenting the largest overlap with this cluster in each family are then combined with this one. Finally, the eigengene of the combined cluster is then tested for association with the phenotype. This method was used for analysis of the simulated dataset provided for the Genetic Analysis Workshop 18. This method was compared with methods such as: single probe analysis, ignoring the pedigree structure, and build the network on "decorrelated" omic variables.

Chapter 5 and Chapter 6 presents new methods to incorporate grouping information in prediction models in order to obtain more stable and possibly interpretable models. All the analyses shown in these chapters are using data from the DIetary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome study (DILGOM) and the publicly available breast cancer cell lines pharmacogenomics dataset for illustration.

In Chapter 5, a new strategy for model selection based on three steps is presented : Network construction of omic features, empirical derivation of modules of related feature via clustering, and construction of prediction model incorporating the grouping information. This approach aims to overcome issues caused by the presence of strong correlations. Several methods are considered to performs steps 1 and 3 of the developed approach. We compare the performance of this strategy with standard regularized regression such as lasso, ridge regression, and elastic net via simulations.

Simulation and data application results show that this strategy provide more stable prediction models and can perform, in terms of prediction accuracy, as well as standard regularized regression. Indeed methods such as lasso or elastic net tend to select randomly one variable from group of strongly correlated variable leading to unstable models and, therefore, the results are hard to reproduce. Comparisons in prediction performance of the various combinations of network approaches and prediction models allows us to provide guidelines in which combination of methods to use. The combination of graphical lasso and group lasso is overall the best performing approach. However, in large datasets the use of WGCNA instead of graphical lasso is preferred due to the intensive computations needed for graphical lasso.

Chapter 6 studies how to use different omics datasets simultaneously in prediction models. Combining several omic sources in one prediction model is challenging due the presence of strong heterogeneity between omic sources. Heterogeneity in terms of dimensionality, normalization procedures, and error structures. In this chapter we propose several strategies to integrate two omic sources in one prediction model. Specifically, we propose three strategies: 1) stacking both omic sources together and applying the approach proposed in Chapter 5, 2) performing network construction and clustering on each omic source separately and build the prediction model, 3) performing network construction and clustering on each omic source separately, identifying correlation between clusters and between omic sources, and incorporation of this information in the prediction model. The data examples in this chapters comprise metabolomics and transcriptomics datasets from Dilgom and, and Copy number variants and gene expression from the breast cancer cell lines pharmacogenomics dataset.

The key components of our proposed approach are to capture groups of correlated features within and between omic datasets and to include this information by a group penalization model. Simulations results showed that naively stacking datasets is usually not a good strategy as it often perform worse than a model based on a single omic datasets. Including information about the correlation between the omic datasets might improve the prediction accuracy. When the noise structures from both omic sources are different, performing the network analysis and clustering on each omic sources separately proved to be more robust in terms of predictive accuracy than stacking the datasets together.

Samenvatting

Deze dissertatie richt zich op de ontwikkeling van nieuwe statistische methodes waarbij rekening wordt gehouden met afhankelijkheidsstructuren in omics-datasets en met de modellering van deze structuren. Statistische modellering van deze structuren kan leiden tot verdere verbetering van onze kennis van biologische mechanismen. Door rekening te houden met de structuur zijn wij mogelijk beter in staat om ziekten te voorspellen. In hoofdstuk 1 wordt de meest gebruikelijke mate van afhankelijkheid beschreven, namelijk de correlatiecoëfficiënt van Pearson. Verder wordt een algemene inleiding gegeven over verschillende afhankelijkheidsstructuren waarmee iemand bij het bestuderen van omics-datasets mogelijk te maken krijgt: afhankelijkheden tussen personen, gemeten resultaten en omics-eigenschappen. Voor elk van deze afhankelijkheidsniveaus worden de uitdagingen beschreven die iemand kan tegenkomen en de daarvoor meestal gebruikte methodes.

In hoofdstuk 2 en 3 worden methodes beschreven voor de analyse van secundaire fenotypen in onderzoek naar geverifieerde families. In hoofdstuk 2 wordt ingegaan op een nieuwe aanpak voor het analyseren van het secundaire fenotype voor een opzet voor familieonderzoek met meerdere casussen, waarvoor families worden geselecteerd waarin minimaal een specifiek aantal casussen voorkomt. De voorgestelde methode wordt onderbouwd aan de hand van een voorbeeld op basis van gegevens van het onderzoek Leiden Lang Leven, een familieonderzoek aan de hand van meerdere casussen, waarin de veroudering bij mensen (primair fenotype) wordt onderzocht. Hier werd een inschatting gemaakt van de parameters die zorgen voor de verbanden tussen triglyceridespiegels en glucose (secundaire fenotypen) en genetische markers. In hoofdstuk 3 worden methodes beschreven die binnen de literatuur zijn terug te vinden ten aanzien van secundairefenotypeanalyse voor een onderzoeksopzet op basis van de familie van proefpersonen. Bij deze opzet worden familieleden van specifieke proefpersonen (vaak casussen met de primaire resultaten) meegenomen in het onderzoek. De prestaties van de beschikbare methodes worden vergeleken met onze methode, die beschreven staat in hoofdstuk 2. De analyse van werkelijke gegevens in dit hoofdstuk maakt onderdeel uit van het familieonderzoek naar socialeangststoornissen (Social Anxiety Disorder, SAD) en richt zich op het vaststellen van kandidaat- endofenotypen van SAD.

Hoofdstuk 2 omschrijft een benadering voor het verkrijgen van onvertekende associatieve schattingen tussen secundaire fenotypen en biomarkers en onvertekende erfelijkheidsschattingen voor secundaire fenotypen. Deze methode biedt ruimte voor het vaststellingsproces en zorgt voor expliciete modellering van de familierelaties. Om dit te bereiken maken wij bij onze benadering gebruik van retrospectieve waarschijnlijkheid, ten behoeve van modellen met gemengde effecten. De benadering op basis van retrospectieve waarschijnlijkheid corrigeert automatisch voor de vaststelling. De willekeurige effecten zorgen voor modellering van de familierelaties. Om de associatie tussen de primaire fenotypen (binaire variabele) van het gemengde type en de secundaire fenotypen (continue variabele) te kunnen bepalen wordt gebruikgemaakt van een multivariaat probitmodel. Door maximalisatie van de retrospectieve log-waarschijnlijkheid kunnen er schattingen worden gedaan.

Een belangrijke empirische bevinding is dat de erfelijkheidsschattingen voor de secundaire trekken sterk kunnen worden onderschat, tenzij rekening wordt gehouden met de wijze van monstername. Uit uitgebreide simulaties is gebleken dat de hier gepresenteerde methode de fout van type 1 op een nominaal niveau houdt en zorgt voor nauwkeurige schattingen, ongeacht de prevalentie van de ziekte, de sterkte van de associatie tussen de genetische varianten en het primaire fenotype, en ongeacht het vaststellingsmechanisme. Momenteel is een belangrijke beperking van onze benadering de aanwezigheid van multivariate integralen, waarvan de berekening veel tijd kost, vooral als er sprake is van een grote stamboom.

In hoofdstuk 3 wordt onderzoek gedaan naar de prestaties van onze methode voor de analyse van gegevens van proefpersonenonderzoeken met familiebenadering, zoals die in hoofdstuk 2 beschreven worden. De prestaties worden vergeleken met methodes die vandaag de dag binnen de literatuur gangbaar zijn, namelijk methodes die het vaststellingsproces negeren of die gezien de secundaire fenotypen van de proefpersonen de voorwaardelijke spreiding van de waarden van de secundaire fenotypen van de families modelleren. Verder pleiten wij voor een uitbreiding van deze laatste wijze van aanpak, waarbij bij de gezamenlijke, voorwaardelijke spreiding van de primaire en secundaire fenotypewaarden van de familieleden wordt uitgegaan van de gezamenlijke spreiding van de primaire en secundaire fenotypen van de proefpersonen.

Uit uitgebreide simulaties is gebleken dat alleen de benadering op basis van retrospectieve waarschijnlijkheid die in hoofdstuk 2 werd ontwikkeld ook echt in staat is om onvertekende erfelijkheidsschattingen te krijgen van het secundaire fenotype, evenals onvertekende parameterschattingen voor de associaties tussen de secundaire fenotypen en genetische markers. Bovendien kan conditionering op de secundaire fenotypewaarden van de proefpersoon leiden tot een ernstige onderschatting van de erfelijkheid en dat kan ook de identificatie beperken van kandidaat-endofenotypen van primaire fenotypen. Alleen de benadering op basis van retrospectieve waarschijnlijkheid kon binnen de analyse van werkelijke gegevens een kandidaat-endofenotype van SAD vaststellen. Een ander belangrijk punt binnen dit hoofdstuk is dat uit alle methodes vertekende schattingen voortkomen als de informatie van de proefpersoon ontbreekt. Daarom moet de toepassing van een dergelijke onderzoeksopzet op dit moment niet worden overwogen.

In hoofdstuk 4 wordt gekeken naar het probleem van het uitvoeren van netwerkanalyse voor co-expressie van genen bij familieonderzoeken. Een grote variatie in expressieniveaus tussen families onderling zou een aanzienlijke vertekening vormen voor de verkregen netwerkstructuur als geen rekening zou worden gehouden met de stamboomstructuur. Om dit probleem te voorkomen, stellen wij een meta-analytische benadering voor. Wij bouwen eerst het omics-netwerk op voor iedere stamboom, om zo clusters van correlerende microarray-probes vast te kunnen stellen. De eigengenen (eerste hoofdcomponent) van ieder cluster van elke stamboom worden vervolgens getest op associatie met een belangwekkend fenotype. Na bepaling van het sterkst geassocieerde cluster, worden de clusters die binnen iedere familie hiermee het meest overlappen met dit cluster gecombineerd. Ten slotte wordt het eigengen van het gecombineerde cluster getest op associatie met het fenotype. Deze methode werd gebruikt voor de analyse van de gesimuleerde dataset die beschikbaar werd gesteld voor de Genetic Analysis Workshop 18 en de prestaties ervan werden vergeleken met die van andere methodes, waaronder: enkeleprobeanalyse, waarbij de stamboomstructuur wordt genegeerd, en opbouw van het netwerk op basis van gedecorreleerde omics-variabelen. In hoofdstuk 5 en 6 presenteren wij nieuwe methodes om groeperingsinformatie in te bouwen in voorspellingsmodellen, om zo stabielere en waar mogelijk interpreteerbare modellen te kunnen krijgen. Bij alle analyses in deze hoofdstukken wordt ter illustratie gebruikgemaakt van gegevens uit het DILGOM-onderzoek (DIetary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome) en van de openbaar toegankelijke, farmacogenomische dataset van borstkankercellijnen.

In hoofdstuk 5 wordt een nieuwe aanpak gepresenteerd voor modelselectie op basis van drie stappen: opbouwen van een netwerk van omics-eigenschappen, empirische derivatie van modules met vergelijkbare eigenschappen door middel van clustering en ten slotte het opbouwen van een voorspellingsmodel, waarin de groeperingsinformatie is ingebouwd. Deze aanpak is erop gericht om problemen als gevolg van de aanwezigheid van sterke correlaties tegen te gaan. Er worden verschillende methodes afgewogen voor het uitvoeren van stap 1 en 3 van de ontwikkelde benadering. Wij vergelijken de prestaties van deze strategie door middel van simulaties met de standaard geregulariseerde regressie, zoals LASSO, Ridge-regressie en elastic net.

Uit de resultaten van de simulaties en datatoepassing blijkt dat deze aanpak leidt tot stabielere voorspellingsmodellen en dat deze wat betreft nauwkeurigheid van de voorspellingen even goed werkt als de standaard geregulariseerde regressie. Bij methodes zoals LASSO of elastic net wordt meestal een willekeurige variabele geselecteerd uit een groep sterk correlerende variabelen, wat leidt tot instabiele modellen en daardoor tot problemen met de reproductie van de resultaten. Door voorspellende prestaties van de diverse combinaties van netwerkbenaderingen en voorspellingsmodellen te vergelijken, kunnen we richtlijnen geven voor de te gebruiken combinatie van methodes. De combinatie van grafische LASSO en groeps-LASSO is de aanpak die over het algeheel gesproken de beste prestaties geeft. Bij grote datasets heeft echter WGCNA de voorkeur boven grafische LASSO, aangezien voor grafische LASSO erg intensieve berekeningen nodig zijn.

In hoofdstuk 6 onderzoeken we hoe binnen voorspellingsmodellen verschillende omics-datasets simultaan kunnen worden gebruikt. Het combineren van verschillende omics-bronnen binnen een voorspellingsmodel is een hele uitdaging, gezien de sterke onderlinge heterogeniteit van omics-bronnen. De gegevenssets variëren in termen van dimensionaliteit, normalisatieprocedures en foutstructuren. In dit hoofdstuk stellen wij drie strategieën voor om twee omics-bronnen binnen een voorspellingmodel te integreren. Onze specifieke voorstellen zijn: 1) stapeling van de beide omics-bronnen en toepassing van de benadering die is voorgesteld in hoofdstuk 5, 2) netwerkconstructie en clustering van beide omics-bronnen afzonderlijk en bouwen van het voorspellingsmodel, 3) netwerkconstructie en clustering van beide omics-bronnen afzonderlijk, bepaling van de correlatie tussen de clusters en tussen de omics-bronnen en inbouwen van deze informatie in het voorspellingsmodel. De voorbeelden van de gegevens in dit hoofdstuk omvatten datasets bestaande uit metabolomics en transcriptomics uit het DILGOM-onderzoek en kopienummervarianten en genexpressie van de dataset van farmacogenomische borstkankercellijnen.

De belangrijkste componenten van de door ons voorgestelde benadering zijn het bepalen van groepen van intern of onderling correlerende eigenschappen van de omics-datasets en het integreren van deze informatie door middel van een groepspenalisatiemethode. Uit simulaties blijkt dat het naïef stapelen van datasets meestal geen goede strategie is, aangezien het model meestal slechter presteert dan een model dat is gebaseerd op een enkele omics-dataset. Toevoeging van informatie over de onderlinge correlatie van de omics-datasets kan de voorspellingsnauwkeurigheid mogelijk verbeteren. Als de ruisstructuren van beide omics-bronnen verschillen, blijkt dat in termen van voorspellingsnauwkeurigheid het uitvoeren van de netwerkanalyse en clustering voor iedere omics-bron afzonderlijk robuuster is dan de stapeling van datasets.

Dankwoord

The work in this thesis could not have been done without a lot of important person that deserve to be acknowledged.

First, I would like to thank Prof.dr Jeanine Houwing-Duistermaat and Dr. Roula Tsonaka for selecting me for this PhD. Dr. Roula Tsonaka never stopped to encourage and support me while Prof.dr Jeanine Houwing-Duistermaat never stopped to fight for me and has given invaluable insights on statistics and on the research environment. I would equally like to express my gratitude to Dr. Mar Rodrìguez-Girondo for accepting to join my supervision during the PhD, leading to fruitful collaborations and friendship.

Secondly, I would like to thank the reading committee: Prof.dr Eline Slagboom, Prof.dr. Jennifer Barrett and Dr. Wessel van Wieringen for their time, their feedback and their interest in this work.

I would like to thank all the colleagues from the LUMC who contributed directly and indirectly to my personal and professional development during my time time at LUMC. Particularly my office mates Brunilda Balliu, Hae-Won Uh, Ivonne Martin and Angga Fuady with whom I had constructive discussions, pure moment of friendship and they have always been there during the tough times. Special thanks are due to all the PhD students and postdoctoral fellow: Alexia Kakourou, Georgios Bartzis, Markus de Jong, Saïd El Bouhaddani, Rosa Meijer, Mia Klinten Grand, Carlo Lancia, Laudia Sala, Sonia Amodio, Theodore Balan, Dimitris Ziagkos, Jesse Hemerik, Irene Yi Sum Man, Kate Xu, Eleni Panagiotou, Xinpei Gao, Roberta Rovito, Zhenia Aizenberg. They have been an important part of my life during my PhD and they continue to be. I would also like to thank the rest of my colleagues from the Medical Statistics and Bioinformatics from whom I learned a lot: Stefan Boehringer, Bart Mertens, Jelle Goeman, Hein Putter, Szymon Kielbase, Ramin Monajemi, Ron Wolterbeek, Theo Stijnen and Saskia le Cessie.

I would like to thanks all the member of MIMomics with whom I learned to work in a multi-disciplinary environment, allowing me to develop communication skills and increase my knowledge on numerous fields: Gastone Castellani, Ettore Mosca, Daniele Remondini, Karli Reiding, Elisa Benedetti, Trishanta Padayachee, Ivo Ugrina, Lucija Klaric, Tomasz Burzykowski, Felix Agakov, Yuri Aulchenko, Lennart Karlsen, Gordan Lauc, Pietro Lio, Carlo Berzuini.

I would also like to acknowledge the people from LIDA for being so welcoming and having shared with me great moments. Thank you so much: Alison, Amira, Hara.

I am thankful to Prof.dr Michel Westenberg for giving me the opportunity to work in his group and to continue my work on family studies. Many special thanks to my office mates at the social sciences institute: Anita Harrewijn, Janna Marie Bas-Hoogendam, Sara Jakobsson Månsson, Jiemiao Chen and Simone Vogelaar.

A really important person I would like to thank is Anna for being the light of the end of my PhD. I am so happy to share my life with you and hope to be able to provide as much support to you as you did to me.

To conclude, I would like to give all my gratitude to my family and especially to my grand parents Roger, Denise, Bernadette, my parents Laurent and Isabelle and my brother Bertrand. Merci pour tout, vos encouragements, votre soutien dans les moments difficile et plus particulièrement votre patience envers moi.

Curriculum Vitae

Renaud Tissier was born on the 16th of September 1987, in Louviers, France. He finished his secondary education in 2005 at the Lycée Georges Clémenceau in Nantes, France. After three years studying in a preparatory class for high scientific school at the Lycée Saint-Joseph in la Roche-sur-Yon, France, he successfully integrated the National School for Statistics and Information Analysis, ENSAI, in Bruz, France. Where he obtained his engineering diploma (Msc equivalent) in statistics with specialization in biostatistics in 2012.

In 2013, he started his PhD at the Department of Medical Statistics and Bioinformatics, Leiden University Medical center, under the supervision of Prof.dr. Jeanine Houwing-Duistermaat, dr. Roula Tsonaka and Dr Mar Rodrìguez-Girondo. His work was founded by the FP7 grant MIMomics and focused on the development of novel statistical methodology for the analysis of complex omics data. The results of this research are presented in this thesis. Chapter 2 of this thesis has been awarded with the Best Student Presentation Award at the 43rd European Mathematical Genetics Meeting (2015). Renaud also spent one year and three months as a visiting searcher at the Leeds Institute of Data Analysis, Leeds, United Kingdom.

In 2017, he joined Prof.dr. Michel Westenberg research group in the social sciences institute of the Leiden University, where he worked as a postdoctoral fellow providing statistical support and supervision for the Leiden Family Lab study on Social Anxiety Disorder.

In September 2018, he joined, as a postdoctoral fellow, dr. Renee de Menezes in the Big statistics group of the department of Epidemiology and Biostatistics of the Vrij University Medical centre, where he works on the development of statistical methodology for the analysis of CRISPR data and the data integration of CRISPR and RNA sequencing datasets.