
THE UNIVERSITY
*of* EDINBURGH

# Statistical methods for the testing and estimation of linear dependence structures on paired high-dimensional data: application to genomic data

Adria Caballe Mestres


THE UNIVERSITY *of* EDINBURGH

**Thesis submitted for the degree of**

**Doctor of Philosophy**

**University of Edinburgh**

**School of Mathematics**

**Year of Submission 2017**

# Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Adria Caballe Mestres*)

# Abstract

This thesis provides novel methodology for statistical analysis of paired high-dimensional genomic data, with the aim to identify gene interactions specific to each group of samples as well as the gene connections that change between the two classes of observations. An example of such groups can be patients under two medical conditions, in which the estimation of gene interaction networks is relevant to biologists as part of discerning gene regulatory mechanisms that control a disease process like, for instance, cancer. We construct these interaction networks from data by considering the non-zero structure of correlation matrices, which measure linear dependence between random variables, and their inverse matrices, which are commonly known as precision matrices and determine linear conditional dependence instead. In this regard, we study three statistical problems related to the testing, single estimation and joint estimation of (conditional) dependence structures.

Firstly, we develop hypothesis testing methods to assess the equality of two correlation matrices, and also two correlation sub-matrices, corresponding to two classes of samples, and hence the equality of the underlying gene interaction networks. We consider statistics based on the average of squares, maximum and sum of exceedances of sample correlations, which are suitable for both independent and paired observations. We derive the limiting distributions for the test statistics where possible and, for practical needs, we present a permuted samples based approach to find their corresponding non-parametric distributions.

Cases where such hypothesis testing presents enough evidence against the null hypothesis of equality of two correlation matrices give rise to the problem of estimating two correlation (or precision) matrices. However, before that we address the statistical problem of estimating conditional dependence between random variables in a single class of samples when data are high-dimensional, which is the second topic of the thesis. We study the graphical lasso method which employs an $L_1$ penalized likelihood expression to estimate the precision matrix and its underlying non-zero graph structure. The lasso penalization term is given by the $L_1$ norm of the precision matrix elements scaled by a regularization parameter, which determines the trade-off between sparsity of the graph and fit to the data, and its selection is our main focus of investigation. We propose several procedures to select the regularization parameter in the graphical lasso optimization problem that rely on network characteristics such as clustering or connectivity of the graph.

Thirdly, we address the more general problem of estimating two precision matrices that are

expected to be similar, when datasets are dependent, focusing on the particular case of paired observations. We propose a new method to estimate these precision matrices simultaneously, a weighted fused graphical lasso estimator. The analogous joint estimation method concerning two regression coefficient matrices, which we call weighted fused regression lasso, is also developed in this thesis under the same paired and high-dimensional setting. The two joint estimators maximize penalized marginal log likelihood functions, which encourage both sparsity and similarity in the estimated matrices, and that are solved using an alternating direction method of multipliers (ADMM) algorithm. Sparsity and similarity of the matrices are determined by two tuning parameters and we propose to choose them by controlling the corresponding average error rates related to the expected number of false positive edges in the estimated conditional dependence networks.

These testing and estimation methods are implemented within the R package ldstatsHD, and are applied to a comprehensive range of simulated data sets as well as to high-dimensional real case studies of genomic data. We employ testing approaches with the purpose of discovering pathway lists of genes that present significantly different correlation matrices on healthy and unhealthy (e.g., tumor) samples. Besides, we use hypothesis testing problems on correlation sub-matrices to reduce the number of genes for estimation. The proposed joint estimation methods are then considered to find gene interactions that are common between medical conditions as well as interactions that vary in the presence of unhealthy tissues.

# Lay summary

The term high-dimensional data is used in the statistics community to refer to cases where the number of parameters that have to be estimated is larger than the number of observations. This is a common situation when analyzing omics data, which can be originated from genomics, metabolomics or proteomics, as for example, the number of genes that are identified in organisms such as humans or mice are of order of thousands, whereas the number of subjects that are involved in the studies tend to be one or two order of magnitudes smaller.

Classical statistical inference methods, though, are developed under the assumption of datasets with more observations than covariates. Hence, common statistical inference topics such as hypothesis testing or statistical modeling have to be reconsidered under this challenging high-dimensional paradigm.

The genome activity in an organism depends on the way the genes are interconnected among each other, and might be altered on the presence of illness processes such as cancer. Finding accurate estimations of gene interaction networks from data is important for biologists to understand the gene regulatory mechanisms that control the disease.

This thesis presents statistical methodology related to the estimation and hypothesis testing of gene interaction networks with the purpose to infer common/unique gene-to-gene conditional dependence structures of two classes of samples from the same subject, that as an example, could be determined by the location of their tissues, one being healthy and the other containing a tumor.

# Acknowledgments

Undertaking this PhD has been a life-changing experience for me and I have been able to do it thanks to the guidance and help received from many people.

First and foremost, I would like to express my deepest gratitude to my two supervisors, Dr Natalia Bochkina and Dr Claus-Dieter Mayer, for their continuous advice, dedication and enthusiasm throughout my PhD. Both have given of their time and huge expertise so I could build up my knowledge in the field and complete this thesis. It has been a real pleasure and a learning experience working with them all these years.

I want to thank Dr Ioannis Papastathopoulos for his enormous contribution in one of the projects of my PhD, which corresponds to Chapter 4 in this thesis. Our frequent discussion sessions during the last two years of my studies have been truly inspiring. I am also grateful to my second supervisor, Professor Colin Aitken, for his advice and encouragement given in our regular meetings.

Besides my supervisors and collaborators, I would like to acknowledge the kindness and help offered by many members of staff and students from both the maths department of the University of Edinburgh and Biomathematics & Statistics Scotland. Finally, I want to thank my family and friends for being with me during the whole process. From them, my most special thanks is to Marta, for her patience and huge support at any stage of the project. She has shared with me every single step of a journey that I will never forget.

# Contents

# Chapter 1

# Introduction

## 1.1 Introduction and motivation

The discovery of high-throughput technology has revolutionized the way to collect genomic data differing from old techniques for its capacity to capture the information of a huge number of genes in a single sample under a particular condition. Experiments employing this machinery, e.g., gene expression microarrays, which are reasonably fast and cheap to perform, have been widely used in the last two decades to measure genome profiles of individuals with illness processes such as cancer. As part of the general interest to fight such diseases, many of the platforms that undertake these experiments make the data publicly available for their analysis.

Organisms are made of cells which contain a large number of genes (and also methylation sites, proteins, metabolites, etc.), even though the estimate of this number for humans is still subject to debate; in a recent publication, Ezkurdia et al. (2014) argue that there are about 19,000 protein-coding genes in the whole human genome. One of the main challenges for scientists is to discern the functions of the genes in a biological process and how these interact between each others in a cell. The dependence structure between genes may vary according to the characteristics and conditions of the populations. For instance, a state of illness such as cancer in an organism may modify the way genes are expressed and their relationships in the cell. In that regard, the collection and analysis of genomic data are essential for both discovery and verification of specific genes that have important functions in cancer cells.

The number of samples (e.g., organisms as humans, mice or plants) that are subjected to these type of experiments is often much smaller than the number of genes that are measured. This is referred to as “high-dimensionality" where the number of unknown parameters of interest is much larger than the sample size. The analysis of high-dimensional data using classic likelihood-based statistical methods tends to be either not appropriate or not useful. Finding suitable tools to accommodate data with large dimensions has posed new challenges for the scientific community. Statisticians and mathematicians have studied and proposed different inference procedures that take into account

high-dimensionality issues in the past two decades (Sánchez and Villa, 2008; Buhlmann and van de Geer, 2011). Besides, operational researchers and bioinformaticians have developed computationally efficient methods that process datasets which can be "big" (Marx, 2013; Greene et al., 2014).

The main motivating data for this thesis are presented in Hinoue et al. (2012) and contain the gene expression and methylation presence profile of 25 patients with colon cancer. For every patient there are measures of gene expression in more than 20,000 genes as well as for methylation presence in more than 27,000 sites, for tissues under two medical conditions: a tumor and an adjacent normal colon tissues. The objectives in the analysis of these data are (1) find out which gene associations are or are not common between the two medical conditions, (2) relate the changes to groups of genes that are known to act together in biological functions, (3) measure the connections between methylation presence and gene expression, and use the two datasets together for a joint analysis.

We consider the following four main methodological topics:

A) Hypothesis testing problems involving the comparison of correlation matrices.

B) Selection of the regularization parameters in graphical models.

C) Joint estimation of two precision matrices.

D) Joint estimation of two regression coefficient matrices.

Hypothesis testing problems in A are applied to assess whether the linear dependence structure of a group of genes is equal or not in samples under two medical conditions. Besides, estimation problems in B, C and D are used for finding associations between genes using high throughput genomic data as well as for linking different types of genomic data.

The first topic A is addressed in the literature (Schott, 2007; Li and Chen, 2012; Cai et al., 2013) for testing the equality of two correlation matrices when the two datasets are high-dimensional, and the observations underlying the matrices are independent. Besides, for topic B, sparse precision matrix estimators are developed in Meinshausen and Bühlmann (2006) or Friedman et al. (2007) by maximizing a lasso-penalized likelihood expression. A natural extension of graphical lasso is applied to jointly estimate multiple precision matrices, which is part of our aim C. For instance, Guo et al. (2011) use a group-lasso penalization to control the differences between the non-zero structure of several precision matrices or Danaher et al. (2014) incorporate a fused-lasso penalization option to constrain the absolute value of the precision matrices elementwise differences. In a similar context, following topic D, a penalized least squares estimator, known as regression lasso (Tibshirani, 1996), is employed to find sparse vectors of regression coefficients when the number of covariates is large. The joint estimation of regression coeficients in multiple classes is also studied in the literature. For instance, Zhang and Wang (2012) use a fused-lasso estimator to find the regression coefficients linking high-dimensional explanatory variables and a single response variable in two conditions, or Lam et al. (2016) propose an $L_2$-fused lasso estimator when both explanatory and response variables are high-dimensional.

Most of the testing and estimation methods seen in the literature assume that the multiple datasets consist of independent groups of samples. In this thesis, motivated by colon cancer data introduced in Hinoue et al. (2012), we present novel statistical techniques for the testing and estimation of gene interactions with the aim to investigate changes in the dependence structure between healthy and unhealthy samples when such independence between groups of samples cannot be assumed. The ultimate goal of this thesis is to provide suitable methodology to fully analyze and integrate multiple types of paired high-dimensional datasets corresponding to samples under two medical condition.

Since both healthy and unhealthy (e.g., tumor) samples are observed for every individual, two precision matrices can be jointly estimated to infer the conditional dependence structure of gene expression in healthy samples and tumor samples, as well as its difference matrix. As a pre-estimation step, we consider the simpler problem of testing whether the two precision matrices are exactly equal, in which case the differential precision matrix does not need to be estimated since it can be taken to be a zero matrix. We reformulate this problem to the equivalent hypothesis testing problem of equality of two correlation matrices so we exploit statistics based on the average of squares and maximum of sample correlations differences similarly to the the approaches found in the literature for independent datasets. Moreover, we present a novel test statistic that takes the sum of sample correlation differences that exceed a given threshold. Other relevant hypothesis testing problems involving correlation sub-matrices are also proposed in the same paired high-dimensional data framework.

Graphical lasso (Friedman et al., 2007) adds a penalty term in the likelihood which is affected by a tuning parameter whose choice represents the trade-off between close fit to the data and sparsity of the estimated precision matrix. The selection problem of this sparsity tuning parameter has been given relatively attention in the literature so far, where generally likelihood based methods were used, which may fail for large dimensions (Liu et al., 2010). Alternatively, we propose several procedures to select the regularization parameter in the estimation of graphical models that concentrate on reliably recovering a desired network characteristic (e.g., clustering or graph connectivity) in biological systems.

Gadaleta and Bessonov (2015) integrate gene expression and methylation presence for a dataset with 215 individuals affected with glioblastoma cancer. The authors use lasso penalized maximum likelihood to estimate two networks: the non-zero structure of the regression coefficients using gene expression as response variables and methylation presence as explanatory variables; and the non-zero structure of the precision matrix (using only gene expression data). We develop weighted fused-lasso methods to perform a similar analysis on the colon cancer data by using both healthy and tumor samples and by accounting for paired observations. We jointly estimate marginal precision matrices, by considering a weighted fused graphical lasso approach (WFGL), and the regression coefficients, by a weighted fused regression lasso approach (WFRL). For the tuning parameters of the penalty terms in either WFGL or WFRL, we introduce a novel strategy to select the expected number of false positive edges, which is applied to our paired data setting but could also be used in other lasso/fused

penalized estimators.

Even though the initial motivating data are given by the colon cancer gene expression and methylation presence datasets, throughout the thesis we are also motivated by other related experiments. For instance, we use a dataset that contains the microarray gene expression information of 154 samples for patients with colon tumor and about 18,000 genes, which is publicly available at the Cancer Genome Atlas (TCGA) repository (`https://cancergenome.nih.gov/`). A second dataset provides the gene expression profile of 82 patients with paired samples: the gene expression in a psoriasis vulgaris lesional tissue and the gene expression in its adjacent non-lesional tissue. We also analyze a third dataset that contains the gene expression measurements of 60 patients with lung cancer for a paired tumor and healthy tissues. Both psoriasis and lung cancer data are publicly available in the Gene Expression Omnibus (GEO) database (Edgar et al., 2002) and consist of more than 19,000 genes for each sample.

All proposed methods on correlation matrices testing, regularization parameter selection procedures, or joint estimation of both precision matrices and regression coefficients are implemented within the R package ldstatsHD (Caballe, 2017), which is available in the CRAN repository.

## 1.2 Chapters of the thesis

Chapter 2 is an introductory chapter in which we define and denote some important concepts for the development of the thesis. We discuss the connection between linear dependence structures (correlation and covariance matrices) and conditional linear dependence structures (precision matrices and regression coefficient matrices). We present some theoretical models, which are often employed to characterize biological networks, and that we will consider throughout the thesis to generate the graphical structure of conditional dependence matrices for simulated data analyses. Finally, we introduce several models that are suitable for data generation of biological experiments in paired observations.

In Chapter 3 we review some of the methods that have broken through the statistical literature for the testing and estimation of dependence structures in high-dimensional data. We mainly cover the topics A-D described in Section 1.1, and then we present other major statistical techniques in the multivariate data analysis literature that have been used to summarize dependence between random variables in high-dimensional data. We finish the chapter by drawing attention to the impact that some of the reviewed methods have had in the application to biological data.

The following three chapters, which represent the main methodological contributions of this thesis, are based on several scientific articles and are meant to work as stand alone pieces of text. Chapter 4 is concerned with topic A. We mainly study the hypothesis testing problem of equality of two correlation matrices using two dependent high-dimensional datasets. Nevertheless, other similar testing problems using correlation matrices are considered. These include testing if a row in a correlation matrix is equal to the same row of another correlation matrix, testing if a correlation

matrix is the identity or testing if a row in a correlation matrix is equal to the same row in the identity matrix. We consider test statistics based on the average of squares, maximum and sum of exceedances of sample correlations. For the first problem of equality of correlations, we derive the limiting distribution of the test statistics and we study the behavior of the null distribution p-values using asymptotic and permutation-based distributions. Theoretical results on the power of the tests under different alternatives are presented and backed up by a range of simulation experiments. We apply testing approaches to high-dimensional real case studies of psoriasis lesional and lung tumor gene expression data with the aim of discovering pathway lists of genes that present significantly different correlation matrices on healthy and unhealthy samples.

In Chapter 5 we describe several risk functions which encourage relevant network characteristics such as clustering or graph connectivity, and that are employed to select the regularization parameter of lasso precision matrix estimators (topic B). We conduct an extensive simulation study to show that the proposed methods produce useful results for different network topologies. The approaches are also applied in a high-dimensional real case study of gene expression data with the aim to discover the genes relevant to colon cancer.

The focus of Chapter 6 is the two joint estimation problems corresponding to topics C and D: the joint estimation of two similar sparse precision matrices and their corresponding marginal conditional dependence graphs; and the joint estimation of two regression coefficient matrices and their underlying graph structure. Both estimators are especially useful in the situation of high dimensional data where observations of the two matrices are dependent, as they come from paired observations. We propose novel methods to estimate these conditional dependence matrices simultaneously, a weighted fused graphical lasso estimator (WFGL) and a weighted fused regression lasso (WFRL) which monitor both sparsity and similarity in the estimated matrices. The tuning parameters of sparsity of the matrices are selected by controlling the estimated expected number of false positive edges, and the penalty term controlling similarity of the matrices is weighted for every pair of variables to account for linear dependence between datasets. We observe overestimation of triangular motifs in the estimated precision matrices, so we incorporate an additional step to remove such edges. We conduct a simulation study to show that the proposed methodology successfully recovers the true conditional dependence graphs for different combinations of sample size and dimension. Besides, the proposed approaches are applied to high-dimensional case studies of paired gene expression data with samples in two medical conditions, non-lesional and psoriasis lesional tissues (first dataset) as well as healthy and lung cancer tissues (second dataset), to estimate common networks of genes as well as the differentially connected genes that interact differently in the two types of tissues.

In Chapter 7 we introduce the R package ldstatsHD, linear dependence statistics for high-dimensional data, (Caballe, 2017). This consists of functions that implement the methodology proposed in previous chapters and that can be grouped in three modules: data simulators, testing methods and estimation methods. In this chapter we describe the main functions in each module and then we illustrate the functionality of the implemented code using several examples.

In the final Chapter 8 we will then employ the testing and estimation methodology presented in the thesis to exhaustively analyze a high-dimensional case study of paired gene expression and methylation presence data where samples consider tissues under two medical conditions: healthy and colon cancer. We estimate two types of joint networks, a site-to-gene directed network and a gene-to-gene undirected network. The first is determined by an estimated joint regression coefficient matrix mapping methylation presence in a site to gene expression, whereas the second is found by a joint precision matrix that is only applied to gene expression data. In both cases, we distinguish between a common network of genes/sites as well as a differential network where genes/sites interact differently in tumor and healthy samples. Our findings confirm that methylation sites tend to be negatively related to genes that are nearby. We observe a hub-based structure where the methylation presence of few methylation sites explain the variability (in gene expression) of many different genes. In both gene-to-gene network and site-to-gene network, graph structures for healthy samples tend to be denser than the ones for tumor samples. Finally, the two type of networks are estimated for some important gene sets with known biological interactions. We find several list of genes, that have been related to the disease of interest, such as Tgf-beta, Gaba or EGFR in which site-to-gene interactions change significantly in the two classes of observations.

# Chapter 2

# Types of dependence structures

## 2.1 Link between regression and Gaussian graphical modeling

Consider $n$ independent and identically distributed (i.i.d.) $p$-dimensional random vectors $X = (X_1, \ldots, X_p) \sim N_p(0, \Sigma_X)$, assuming, without a loss of generality, that the mean is zero. The covariance matrix $\Sigma_X$ and its scaled matrix $R_X = [r^x_{ij}] = \text{diag}(\Sigma_X)^{-1/2}\Sigma_X\text{diag}(\Sigma_X)^{-1/2}$, the correlation matrix, measure linear relationship between pairs of variables and they are the key of many multivariate techniques in the statistical literature (Mardia et al., 1979).

The inverse of the covariance matrix $\Sigma_X$ (or sometimes preferable $R_X$), commonly known as precision matrix, and denoted by $\Omega_X = [\Omega^x_{ij}]$ differs from the correlation matrix since it measures linear relationship between pairs of variables accounting for the linear dependence in the rest of the variables. Two variables $X_i$ and $X_j$ are said to be conditionally independent given the rest of the variables if the coefficient $\Omega^x_{ij}$ is zero. The non-zero structure of $\Omega_X$ is characterized by an undirected graph $G(V, E)$ in which nodes $V$ represent the random variables and edges $E$ connect variables whose elements in the precision matrix are non-zero, i.e.,

$$(i, j)\,\&\,(j, i) \in E \iff X_i \not\perp\!\!\!\perp X_j | X_{V\setminus\{i,j\}} \iff \Omega^x_{ij} \neq 0.$$

The graph structure is often represented by a $p \times p$ symmetric matrix called adjacency matrix and denoted by $A_G = [A^G_{ij}]$. The off-diagonal elements of $A_G$ are determined by the precision matrix ($A^G_{ij} = 0$ if $\Omega^x_{ij} = 0$ and $A^G_{ij} = 1$ otherwise) and the diagonal elements are set to zero.

Take $y = X_j$, being the $j$th variable in $X$, update $X_{-j} = X_{V\setminus j}$, and consider the regression model

$$y \sim N(X_{-j}\beta, \sigma^2_e), \tag{2.1}$$

where $\beta$ is the $(p-1) \times 1$ vector of regression coefficients and $\sigma^2_e = \text{Var}(y - X_{-j}\beta)$ is a positive constant.

Then

$$\beta_h = -\Omega^x_{jh}/\Omega^x_{jj},\ \forall\, h \in V \setminus j, \tag{2.2}$$

links precision matrix and regression coefficients.

## 2.2 Data in two conditions and cross-correlation matrix

Consider $n$ independent and identically distributed (i.i.d.) $p$-dimensional random vectors $Y^{(1)}_k = (Y^{(1)}_{k1}, \ldots, Y^{(1)}_{kp})$ and $Y^{(2)}_k = (Y^{(2)}_{k1}, \ldots, Y^{(2)}_{kp})$ associated with condition I (e.g., healthy genes) and condition II (e.g., tumor genes) respectively that jointly follow a multivariate standard normal distribution with correlation R, i.e.,

$$(Y^{(1)}_k, Y^{(2)}_k) \sim N_{2p}(0, R), \quad R = [r_{ij}] = \begin{bmatrix} R_1 & R_{12} \\ R_{12}^\top & R_2 \end{bmatrix}, \tag{2.3}$$

where $R_1$ and $R_2$ are the covariance matrices that correspond to healthy genes and tumor genes, respectively. Assume, without loss of generality, zero mean vector and unit variances for simplicity so covariance matrices coincide with correlation matrices.

Let $\Omega_1 = R_1^{-1}$, $\Omega_2 = R_2^{-1}$, $\Omega_1^J = (R_1 - R_{12}\Omega_2 R_{21})^{-1}$ and $\Omega_2^J = (R_2 - R_{21}\Omega_1 R_{12})^{-1}$. The matrix $\Omega_1$ characterizes marginal dependence of $Y^{(1)}$ whereas $\Omega_1^J$ measures dependence of $Y^{(1)}$ conditionally on $Y^{(2)}$, and similarly for $\Omega_2$ and $\Omega_2^J$. The joint precision matrix $\Omega^J$ is given by the inverse of the joint correlation matrix $R$, with

$$\Omega^J = R^{-1} = \begin{bmatrix} \Omega_1^J & \Omega_{12}^J \\ \Omega_{21}^J & \Omega_2^J \end{bmatrix},$$

In the situations we will consider, the group of observations in $Y^{(1)}$ and $Y^{(2)}$ will not be independent in general but will have a non-trivial cross correlation matrix $R_{12}$. We will start with the simpler independence model though where both $R_{12} = 0$ and $\Omega_{12}^J = 0$, which lead to $\Omega_1 = \Omega_1^J$ and $\Omega_2 = \Omega_2^J$. We further present alternative models which account for dependence, $R_{12} \neq 0$, that seem to be more realistic to justify the paired data setting given in our motivating data, where $\Omega_1$ & $\Omega_2$ may differ slightly from $\Omega_1^J$ & $\Omega_2^J$, respectively.

*Independence model*

This model assumes subject independence in $Y^{(1)}$ and $Y^{(2)}$,

$$Y^{(1)} = Y^{(1)*}; \qquad Y^{(2)} = Y^{(2)*}, \tag{2.4}$$

with $Y^{(1)*} \perp Y^{(2)*}$, thus $R_{12} = 0$. Take model (2.3), the correlation, which is denoted by $R_{ind}$, and its inverse $\Omega_{ind}$, the precision matrix, are specified by

$$R_{ind} = \begin{pmatrix} R_1 & 0 \\ 0 & R_2 \end{pmatrix} \quad \text{and} \quad \Omega_{ind} = \begin{pmatrix} \Omega_1 & 0 \\ 0 & \Omega_2 \end{pmatrix}, \tag{2.5}$$

with $\Omega_1 = R_1^{-1}$ and $\Omega_2 = R_2^{-1}$.

*Additive model*

Assume that data in first condition $Y^{(1)}$ correspond to samples in a "normal" state (e.g., healthy samples) and data in the second condition $Y^{(2)}$ are the samples in the "changing" state (e.g., tumor samples). In a matrix form, this can be specified by

$$Y^{(1)} = Z\Delta^{1/2} + H\bar{\Delta}^{1/2}; \qquad Y^{(2)} = Z\Delta^{1/2} + T\bar{\Delta}^{1/2}. \tag{2.6}$$

Take $\Delta$ to be a $p \times p$ diagonal matrix with the class-correlation magnitudes. Moreover, $\bar{\Delta}$ is also a diagonal matrix with $\bar{\Delta}_{ii} = 1 - \Delta_{ii}$ for any $i \in \{1, \ldots, p\}$. Here $Z$ is the common expression in the two classes, and $H$ and $T$ represent the differential expressions due to changing conditions. We assume that $Z \sim N(0, R_Z)$, $H \sim N(0, R_H)$ and $T \sim N(0, R_T)$ are independent between each other so $R_1 = \Delta^{1/2} R_Z \Delta^{1/2} + \bar{\Delta}^{1/2} R_H \bar{\Delta}^{1/2}$ and $R_2 = \Delta^{1/2} R_Z \Delta^{1/2} + \bar{\Delta}^{1/2} R_T \bar{\Delta}^{1/2}$ define the covariance matrices in $Y^{(1)}$ and $Y^{(2)}$, respectively. Take model (2.3), the correlation, which is denoted by $R_{add}$, and its inverse, $\Omega_{add}$, are specified by

$$R_{add} \doteq \begin{pmatrix} R_1 & \Delta^{1/2} R_Z \Delta^{1/2} \\ \Delta^{1/2} R_Z \Delta^{1/2} & R_2 \end{pmatrix} \quad \text{and} \quad \Omega_{add} \doteq \begin{pmatrix} \Omega_1^J & \Omega_{12}^J \\ \Omega_{21}^J & \Omega_2^J \end{pmatrix}. \tag{2.7}$$

where $\Omega_1^J = (R_1 - \Delta^{1/2} R_Z \Delta^{1/2} R_2^{-1} \Delta^{1/2} R_Z \Delta^{1/2})^{-1}$, $\Omega_{12}^J = -\Omega_1^J \Delta^{1/2} R_Z \Delta^{1/2} R_2^{-1}$, $\Omega_{21}^J = -R_2^{-1} \Delta^{1/2} R_Z \Delta^{1/2} \Omega_1^J$ and $\Omega_2^J = (R_2 - \Delta^{1/2} R_Z \Delta^{1/2} R_1^{-1} \Delta^{1/2} R_Z \Delta^{1/2})^{-1}$. Since $Y^{(1)}$ is considered to be the "normal" state, we could further assume that $R_Z$ is proportional to $R_1$.

*Multiplicative model*

We consider a model with a linear transformation from class $Y^{(1)}$, the "normal" state, to class $Y^{(2)}$, the "changing" state, which is defined by

$$Y^{(1)} = Z\Delta^{1/2} + H\bar{\Delta}^{1/2}; \qquad Y^{(2)} = ZQ\Delta^{1/2} + T\bar{\Delta}^{1/2}. \tag{2.8}$$

The class-correlation matrices $\Delta$ and $\bar{\Delta}$ have the same interpretation as for the additive model. Besides, $Z \sim N(0, R_Z)$, $H \sim N(0, R_H)$ and $T \sim N(0, R_T)$ are independent between each other, and are equivalent to the definition in expression (2.6). Here, we further assume that $R_1$ and $R_2$ are proportional to $R_Z$ and $R_T$, respectively, with $Q = R_2^{1/2} R_1^{-1/2}$ being the transformation matrix. In terms of the model introduced in (2.3), the correlation, which is denoted by$R_{mult}$, and its inverse, $\Omega_{mult}$, are specified by

$$R_{mult} \doteq \begin{pmatrix} R_1 & \Delta^{1/2} R_1^{1/2} R_2^{1/2} \Delta^{1/2} \\ \Delta^{1/2} R_2^{1/2} R_1^{1/2} \Delta^{1/2} & R_2 \end{pmatrix} \quad \text{and} \quad \Omega_{mult} \doteq \begin{pmatrix} \Omega_1^J & \Omega_{12}^J \\ \Omega_{21}^J & \Omega_2^J \end{pmatrix}. \tag{2.9}$$

where $\Omega_1^J = (R_1 - \Delta^{1/2} R_1^{1/2} R_2^{1/2} \Delta^{1/2} R_2^{-1} \Delta^{1/2} R_2^{1/2} R_1^{1/2} \Delta^{1/2})^{-1}$, $\Omega_{12}^J = -\Omega_1^J \Delta^{1/2} R_1^{1/2} R_2^{1/2} \Delta^{1/2} R_2^{-1}$, $\Omega_{21}^J =$

$\Omega_{12}^{'J}$ and $\Omega_2^J = (R_2 - \Delta^{1/2} R_2^{1/2} R_1^{1/2} \Delta^{1/2} R_1^{-1} \Delta^{1/2} R_1^{1/2} R_2^{1/2} \Delta^{1/2})^{-1}$.

*Direct effect model*

This model assumes that a variable in the first condition $Y_{g_1}^{(1)}$ is conditionally independent to a variable in the second condition $Y_{g_2}^{(2)}$, for any $g_1, g_2 \in V$, if $g_1 \neq g_2$ and $V = \{1, \ldots, p\}$. Besides, for some $g_1 \in V$, $Y_{g_1}^{(1)}$ can be conditionally dependent to $Y_{g_1}^{(2)}$ given the rest of the variables in $Y^{(1)}$ and $Y^{(2)}$ (see graphical representation at Figure 2.1). Take model (2.3), the precision matrix $\Omega_{diag}$ is determined by $\Omega_1^J, \Omega_{12}^J, \Omega_{21}^J, \Omega_2^J$, where $\Omega_{12}^J$ and $\Omega_{21}^J$ are diagonal matrices.

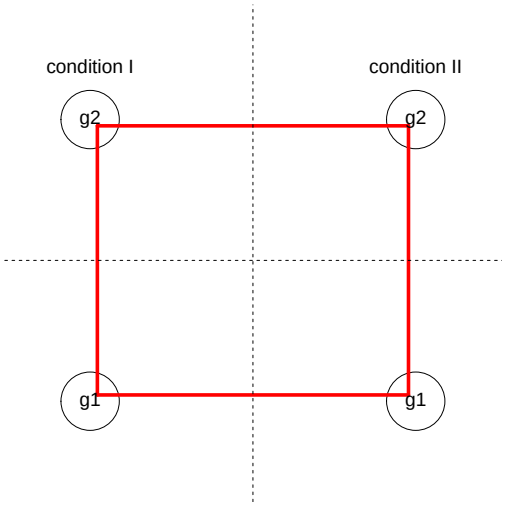



**Figure 2.1.** Square-type conditional graph dependence structure. Gene $g1$ and gene $g2$ are directly connected whithin the same condition, gene $g1$ in the first condition is directly linked with gene $g1$ in the second condition (and similarly for gene $g2$). No direct connections are present between genes $g1$ and $g2$ relating the two conditions.

*Interpretation of proposed models*

The first model, the independence model, is only suitable under the hypothesis that $Y^{(1)}$ and $Y^{(2)}$ come from independent group of observations. In our motivating data, see Section 1.1, there is the gene expression information for classes healthy and tumor in the same individuals. Hence, $R_{12}$ is expected to contain non-zero coefficients. Nevertheless, we keep this model in the list as many testing and estimation methods we will review from the literature in the following Chapter 3 assume independence between random vectors $Y^{(1)}$ and $Y^{(2)}$.

The additive model seems a very reasonable structure for the nature of our data. It considers a clear distinction between a normal state $Y^{(1)}$ and a tumor state $Y^{(2)}$ and it assumes that the expression in tumor samples is equal to the expression in healthy samples plus an additional differential expression term which is independent from the initial state. The multiplicative model reproduces the differences between normal and tumor conditions as follows. The cancer state is given by a transformation of the normal state, which differs from the additive model by the fact that it assumes dependence between initial state and cancer effect. The transformation matrix $Q$ indicates linear dependence between healthy expression and cancer expression. In both models, $Z$ can be interpreted as a source of

systemic variation in gene expression: variation present in all measured tissues of the same individual. Besides, $H$ and $T$ are viewed as category-specific variation.

Finally, the direct effect model is a simplification of the additive and multiplicative models which assumes that $\Omega_{12}^{J}$ is a diagonal matrix where $(\Omega_{12}^{J})_{gg}$ gives the conditional relationship between gene $g$ in a tumor tissue and gene $g$ in a healthy tissue. This diagonal structure considers that the only variables needed to link the gene expression of a gene in a specific state, say gene $g$ in a normal state, are the other genes $V \setminus g$ in the same state (normal) as well as the same gene $g$ in the alternative state (cancer). Hence, it considers conditional independence between gene $g$ in the normal state and all other genes $V \setminus g$ in the alternative state.

# Chapter 3

# Literature overview

## 3.1 Hypothesis testing problems on correlation matrices

Consider $n_1$ independent and identically distributed (i.i.d.) $p$-dimensional random vectors $Y^{(1)} = (Y_1^{(1)}, \ldots, Y_p^{(1)})$ and $n_2$ i.i.d. random vectors $Y^{(2)} = (Y_1^{(2)}, \ldots, Y_p^{(2)})$, which represent measures of the same variables in two different classes, with $Y^{(1)} \sim N(0, R_1)$ and $Y^{(2)} \sim N(0, R_2)$, with $R_1 = [r_{ij}^{(1)}]$ and $R_2 = [r_{ij}^{(2)}]$, assuming unit variances for all variables in the two conditions.

In this section we review some of the methods in the literature that test the hypothesis

$$H_0 : R_1 = R_2 \quad \text{against} \quad H_1 : R_1 \neq R_2, \tag{3.1}$$

when observations in $Y^{(1)}$ and $Y^{(2)}$ are independent. Moreover, we report some other related hypothesis testing methods that only involve a single correlation matrix or that consider sub-matrices of the original $R_1$ and $R_2$.

### 3.1.1 Tests statistics for equality of correlation matrices

*Classical tests*

Random matrix theory ascertains that the sample correlation matrix from normally distributed random variables follows a Wishart distribution. The most powerful test for equality of two correlation matrices is given by the likelihood ratio, which under Gaussianity, it is a function of the determinant of the two sample matrices. The expression of the test statistic is derived in Kullback (1967),

$$T_{Kul} \propto \frac{|\hat{R}_1|^{(1/2)n_1} |\hat{R}_2|^{(1/2)n_2}}{|\hat{R}_1 + \hat{R}_2|^{(1/2)(n_2+n_1)}}, \tag{3.2}$$

and it is only well defined when $\min(n_1, n_2) > p$. Jennrich (1970) suggests another similar proposal

that pursuits a good approximation for

$$T_{Jen} = (LT(\hat{R}_1) - LT(\hat{R}_2))' \Psi^{-1} (LT(\hat{R}_1) - LT(\hat{R}_2)), \tag{3.3}$$

where $LT$ stands for lower triangular matrix and $\Psi$ is the $p(p-1)/2 \times p(p-1)/2$ covariance matrix of the difference coefficients. If $\Psi$ is known then the test is asymptotically chi-squared distributed. Moreover, the author finds an estimator of $T_{Jen}$ in a lower dimension $p$, instead of $p(p-1)/2$, which involves computing the inverse of the average correlation matrix $\bar{\hat{R}} = (n_2\hat{R}_2 + n_1\hat{R}_1)/(n_2+n_1)$, with $n_2 + n_1 > p$ being a necessary condition.

Datasets that arise from biological experiments are frequently high-dimensional, with $n_2 + n_1 \ll p$, and standard statistics, as defined by eq. (3.2) and eq. (3.3), are not suitable. There are two main directions that address this hypothesis testing problem for high-dimensional data in the literature. The first is based on sum of squares statistics, e.g., Schott (2007) and Li and Chen (2012) use the Frobenius norm as a distance measure to compare the two sample correlation matrices. The second is based on extreme value statistics, e.g., Larntz and Perlman (1985) use the maximum of Fisher transform sample correlation coefficient differences in absolute value, Cai et al. (2013) propose an asymptotic test based on the maximum of the squared sample correlation coefficient differences or, similarly, Zhou et al. (2015) apply an extreme value test on Kendall's tau sample correlations. The sum of squares test of Li and Chen (2012) and the extreme value test of Cai et al. (2013) are described below.

*Sum of squares test*

Li and Chen (2012) propose a method to test the equality of covariance matrices, which can be applied to correlation matrices after an appropriate transformation. The authors study the form of the Frobenius norm of the matrix with the sample correlation differences: $\text{tr}\{(\hat{R}_2 - \hat{R}_1)^2\}$. This is decomposed in three terms, $\text{tr}\{(\hat{R}_2 - \hat{R}_1)^2\} = \text{tr}(\hat{R}_1^2) + \text{tr}(\hat{R}_2^2) - 2\text{tr}(\hat{R}_1\hat{R}_2)$, which are estimated using unbiased statistics. Define $\gamma_{2_{ij}}^{(m,s)} = Y_i^{(m)'} Y_j^{(s)}$, $\gamma_{3_{ijk}}^{(m,s)} = Y_i^{(m)'} Y_j^{(s)} Y_i^{(m)'} Y_k^{(s)}$ and $\gamma_{4_{ijkl}}^{(m,s)} = Y_i^{(m)'} Y_j^{(s)} Y_k^{(m)'} Y_l^{(s)}$ with $\bar{\gamma}_2^{(m,s)}$, $\bar{\gamma}_3^{(m,s)}$ and $\bar{\gamma}_4^{(m,s)}$ being the averages of $[\gamma_{2_{ij}}^{(m,s)}]$, $[\gamma_{3_{ijk}}^{(m,s)}]$ and $[\gamma_{4_{ijkl}}^{(m,s)}]$, with $m, s \in \{1,2\}$, $i \neq j$, $j \neq k$, $k \neq l$, respectively. The test statistic is given by

$$T_{Liu} = A_{n_1} + A_{n_2} - 2C_{n_{12}}, \tag{3.4}$$

where $A_{n_m} = \bar{\gamma}_2^{(m,m)} - \bar{\gamma}_3^{(m,m)} + \bar{\gamma}_4^{(m,m)}$ and $C_{n_{12}} = \bar{\gamma}_2^{(1,2)} - \bar{\gamma}_3^{(1,2)} + \bar{\gamma}_4^{(1,2)}$.

Under the null hypothesis ($R_1 = R_2$) and some mild conditions in terms of sample sizes, dimension and dependence, then $T_{Liu}$ tends in distribution to a normal distribution with expected value zero and variance $\sigma_0^2 = 4(n_1^{-1} + n_2^{-1})\text{tr}^2(R^2)$. The variance can be estimated by $\hat{\sigma}_0^2 = 2n_1^{-1}A_{n_1} + 2n_2^{-1}A_{n_2}$ as it is proven to be a ratio-consistent estimator of $\sigma_0^2$.

*Extreme values test*

Cai et al. (2013) consider the maximum of standardized element-wise sample correlation differences

to test the hypothesis of equality between the two matrices. Let

$$D_{ij} = \frac{(\hat{r}_{ij}^{(2)} - \hat{r}_{ij}^{(1)})^2}{\hat{\theta}_{ij}^{(1)}/n_1 + \hat{\theta}_{ij}^{(2)}/n_2},$$

where $\hat{\theta}_{ij}^{(1)}$ and $\hat{\theta}_{ij}^{(2)}$ are the estimators of $\text{var}(\hat{r}_{ij}^{(1)})$ and $\text{var}(\hat{r}_{ij}^{(2)})$, respectively, which can be found by

$$\hat{\theta}_{ij}^{(1)} = n_1^{-1} \sum_{k=1}^{n_1} (Y_{ki}^{(1)} Y_{kj}^{(1)} - \hat{r}_{ij}^{(1)})^2, \quad \hat{\theta}_{ij}^{(2)} = n_2^{-1} \sum_{k=1}^{n_2} (Y_{ki}^{(2)} Y_{kj}^{(2)} - \hat{r}_{ij}^{(2)})^2.$$

The test statistic is given by the maximum of elements in the lower triangular matrix of $D$,

$$T_{Cai} = \max_{i<j} D_{ij} = \max_{i<j} \frac{(\hat{r}_{ij}^{(2)} - \hat{r}_{ij}^{(1)})^2}{\hat{\theta}_{ij}^{(1)}/n_1 + \hat{\theta}_{ij}^{(2)}/n_2}. \tag{3.5}$$

Under the hypothesis of equal correlation matrices and similar mild conditions as for Li and Chen (2012), then $D_{ij}$ are weakly dependent random variables that converge in distribution to a chi-square. The maximum of chi-squared distributed random variables tends in distribution to a Gumbel, i.e.,

$$\Pr(T_{Cai} - 4\log p + \log\log p \le t) \to \exp\{-(8\pi)^{-1/2}\exp(-t/2)\},$$

can be used to assess the evidence of equal correlations.

### 3.1.2 Other tests involving correlation sub-matrices

A useful transformation for the correlation coefficients is given by the Fisher transformation (Fisher, 1924) which can be defined by

$$g : (-1,1) \to \mathbb{R},\ g(z) = \log\{(1+z)/(1-z)\}/2,$$

and it is found to stabilize the variance of sample correlation coefficients. For sufficiently large sample size $n_h$, $h \in \{1,2\}$, the Fisher transformation of a sample correlation estimator $\hat{r}_{ij}^{(h)}$ approximately follows a normal distribution, i.e., $\hat{u}_{ij}^{(h)} = g(\hat{r}_{ij}^{(h)})\sqrt{n_1 - 3} \sim N(g(r_{ij}^{(h)})\sqrt{n_h - 3}, 1)$. The equality of correlation coefficients in different classes is tested in Dunn and Clark (1969) or Steiger (1980) locally for all pairs of variables $(i,j) \in \{1,\ldots,p\}$, $i<j$, by comparing $c(\hat{u}_{ij}^{(1)} - \hat{u}_{ij}^{(2)})$ where $c = 2 - 2\psi_{ij}^{(12)}$ is an estimator for the variance of the difference $\hat{u}_{ij}^{(1)} - \hat{u}_{ij}^{(2)}$. Similarly, Fukushima (2013) recovers a network of tested correlation coefficients using as test statistic $\sqrt{c}\{g(\hat{r}_{ij}^{(1)}) - g(\hat{r}_{ij}^{(2)})\}$ with $c = \{(n_2-3)^{-1} + (n_1-3)^{-1}\}^{-1/2}$. The observed p-values are adjusted by multiple testing by controlling the false discovery rate (Benjamini and Hochberg, 1995).

Li and Chen (2012) also consider the problem of testing whether two correlation sub-matrices are equal or not. The authors propose a Frobenius norm based test statistic on the correlation sub-matrices similar to the statistic in eq. (3.4), which recall was applied to the whole matrices instead.

Raghunathan (2003) tests the equality of two correlation coefficients as well as the equality of two correlation sub-matrices by employing the square of the difference between Fisher-transform sample correlation coefficients and by considering chi-squared null distributions. Finally, Srivastava et al. (2014), among others, propose a method to test whether a single correlation matrix is the identity matrix ($H_0 : R_1 = I$ vs $H_1 : R_1 \neq I$) using related ideas to the hypothesis testing approach presented in Li and Chen (2012), which is discussed in section 3.1.1, for the equality of two correlation matrices. Some of these approaches are implemented within the R package **cocor** (Diedenhofen and Musch, 2015).

## 3.2 Linear regression and Gaussian graphical models

The first problem considered in this section is the linear regression model with Gaussian errors $y_k \sim N(\beta X_k, \sigma_e^2),\ k = 1,\ldots,n$, where $\sigma_e^2$ is a positive constant and $\beta$ represents the linear regression coefficients that relate explanatory variables $X$ with response variable $y$, and it is typically estimated by least squares (LSE),

$$\hat{\beta}_{LSE} = \arg\min_{\beta}\left(\frac{1}{2n}||y - X\beta||_2^2\right), \tag{3.6}$$

with exact solution given by $\hat{\beta} = (X'X)^{-1}X'y$ if $(X'X)^{-1}$ exists ($p < n$ is a necessary condition).

The second problem refers to the estimation of the precision matrix $\Omega_X = R_X^{-1}$, see definition in Section 2.1, which can be estimated by maximum likelihood (MLE) by

$$(\hat{\Omega}_X)_{MLE} = \arg\max_{\Omega_X \succ 0}\ \log\det\Omega_X - \mathrm{tr}(S_X\Omega_X), \tag{3.7}$$

where $S_X = n^{-1}\sum_{k=1}^{n} X_k'X_k$ is the sample covariance matrix. Taking the derivative with respect to $\Omega_X$ by using (i) $\frac{\partial}{\partial B}\log|B| = (B^{-1})'$ and (ii) $\frac{\partial}{\partial B}\mathrm{tr}[BC] = C'$, it is immediate to demonstrate that the maximum likelihood is reached when $\hat{\Omega}_X = S_X^{-1}$ and therefore $\hat{\Sigma}_X = S_X$. However, $p < n$ is a necessary condition for $S_X^{-1}$ to exist.

The study of conditional dependence structures such as $\beta$ and $\Omega_X$ in a high-dimensional data, where the number of unknown parameters to be estimated is larger than the number of observations, and both $\hat{\beta}_{LSE}$ and $(\hat{\Omega}_X)_{MLE}$ are not uniquely defined, is fairly recent. Tibshirani (1996) with the introduction of lasso regression and also Lauritzen (1996) with his book on graphical models opened a door of investigation that has motivated researchers ever since. In this section we review some of the main estimation methods for conditional dependence structures that have been relevant in the statistics literature and that have motivated the work presented in Chapters 5 and 6.

### 3.2.1 Regression models in high-dimensional data

For $p > n$, the least squares estimator defined in eq. (3.6) is not unique and some type of regularization is needed to obtain a tractable problem. Tibshirani (1996) considers an $L_1$ constraint for the regression

coefficients by solving the following optimization problem

$$\begin{aligned}\hat{\beta}^{\tau}_{lasso} = \quad & \underset{\beta}{\arg\min}\ \left(\tfrac{1}{2n}||y-\beta X||_2^2\right) \\ & \text{s.t. } \textstyle\sum_{i=1}^{p}|\beta_i| \le \tau.\end{aligned}$$

Using the Lagrangian multipliers, the constraint can be rewritten as a penalty term in the objective function

$$\hat{\beta}^{\lambda}_{lasso} = \underset{\beta}{\arg\min}\left[\frac{1}{2n}||y-\beta X||_2^2 + P_\lambda(\beta)\right], \quad P_\lambda(\beta) = \lambda||\beta||_1, \tag{3.8}$$

where $||\beta||_1 = \sum_{i=1}^{p}|\beta_i|$ and $\lambda$, which has one-to-one correspondence to $\tau$, represents the trade-off between close fit to the data and sparsity of $\beta$. This method to estimate the regression coefficients in high-dimensional data is commonly known as least absolute shrinkage and selector operator (lasso).

The analogous interpretation of lasso estimates is given in a Bayesian framework (Yuan and Lin, 2005; Park and Casella, 2008; Hans, 2009; Kyung et al., 2010) by finding the mode of the regression coefficients posterior distribution of the model $f(y_k|\beta,\sigma) \sim \mathrm{N}(y_k \mid X_k\beta, \sigma_e^2)$, $k = 1,\ldots,n$, when using independent and identical Laplace priors on $\beta$

$$\pi(\beta|\sigma_e) = \prod_{j=1}^{p}\frac{\lambda}{2\sigma_e}e^{-\lambda|\beta_j|/\sigma_e}.$$

and an inverse gamma prior on $\sigma_e^2$. Gibbs sampling algorithms are employed to approximate the posterior distribution for the regression coefficients.

Other regularization penalties can be used instead of the $L_1$ norm to overcome high-dimensionality problems in solving eq. (3.6). For instance, the ridge regression (Hoerl and Kennard, 1970) constrains the regression coefficients using an $L_2$ norm penalization term by $P_\lambda(\beta) = \lambda||\beta||_2^2$. The comparison between the two penalties is seen in Figure 1 (graphical representation taken from book by Buhlmann and van de Geer (2011)). The $L_2$ norm shrinks regression coefficients towards zero but does not encourage the exact zero values of lasso. This intrinsic variable selection component of the lasso estimates, due to the squared area suggested in the left hand side figure, has made such penalty so appealing in comparison to ridge.

Both $L_1$ and $L_2$ norm constraints can be used together, and its underlying estimator is commonly known as elastic-net (Zou and Hastie, 2005). The incorporation of the $L_2$ norm in the lasso estimation problem can be beneficial in cases where covariates are highly correlated since it acts as grouping effect where correlated variables are either all in or all out of the model.

It is well known for the problem of estimating sparse vectors in high dimensions with the lasso penalty (and also elastic-net), that the variable selection part, with an appropriate $\lambda$, is consistent, however, the estimation of the non-zero values usually has some bias (Wasserman and Roeder, 2009). This is due to the convex relaxation of the desired $L_0$ penalty to the computationally efficient $L_1$

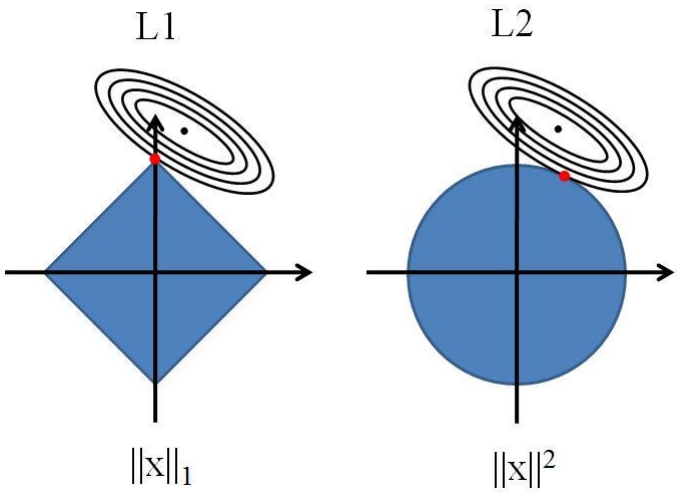



**Figure 3.1.** lasso constraint solution (left) versus ridge constraint solution (right).

penalty. The adaptive lasso (Zou, 2006), the relaxed lasso (Meinshausen, 2007) or the smoothly clipped absolute deviation (SCAD) penalty (Fan Jianqing, 2001; Kim et al., 2008) are procedures that intend to provide unbiased estimators. Besides, the SCAD penalty can be combined with an $L_2$ norm penalty (Zeng and Xie, 2012) to reduce the bias of the estimator and achieve desired grouping properties when covariates are correlated.

Other sparse estimators for the regression coefficients include the Dantzig selector (Candes and Tao, 2007), which estimates $\beta$ by solving an $L_1$ minimization problem that forces the correlation between residuals and any variable entering in the model to be smaller than a value within noise level. Another relevant approach is the least angle regression (LARS) (Efron et al., 2004), which finds a sparse solution without employing a penalization on the least squares function. It is similar to classic forward selection since starts by setting all regression coefficients equal to zero, and a predictor is included in the model once at a time. However, LARS updates the regression estimates so all predictors included in the model are equally correlated (equiangular) to the current residuals.

Linear regression assuming a broad type of exponential family distributions for the errors (GLM) in high-dimensional data is also studied in the literature. Van De Geer (2008) uses a lasso penalization on generalized linear models, James and Radchenko (2009) suggest to employ a generalized Dantzig selector criterion, which is the extension of the Dantzig selector for non Gaussian errors, or Augugliaro et al. (2013) propose a differential geometric least angle regression method based on the LARS algorithm for generalized model selection in high-dimensional data.

Sparse regression estimators are implemented within the free software R in the package **dglars** (Augugliaro et al., 2014), which contains the algorithmic procedures to estimate the regression coefficients for both Gaussian and non-Gaussian errors using lasso, elastic-net and ridge penalizations. Least angle regression and lasso regression are also available within the package **lars** (Hastie and Efron, 2013).

### 3.2.2 Graphical modeling in high-dimensional data

Three lines of approaches are studied in this section to overcome problems in estimating $\Omega_X = [\Omega^x_{ij}]$ when data $X$ are high-dimensional: shrinkage, thresholding and penalization methods. Shrinkage and thresholding operate on the covariance (or correlation) matrix whereas penalization methods are used directly on the precision matrix elements and will be the main focus of attention of this review. Moreover, thresholding and penalization approaches differ from shrinkage approaches by assuming sparsity in the covariance and precision matrix, respectively, i.e., most of the elements in the matrix are assumed to be exactly zero.

When $n/p$ is small, the condition number of the sample covariance matrix is high, meaning that the largest sample eigenvalue is biased upwards and the smallest sample eigenvalue is biased downwards (Pourahmadi, 2007). Shrinkage procedures intend to concentrate the eigenvalues to a common value. Ledoit and Wolf (2004) present a shrinkage estimator of the covariance matrix $\Sigma_X$ by using a linear combination of two models

$$\hat{\Sigma}_X = \lambda T + (1-\lambda)U, \tag{3.9}$$

where $U$ is an unrestricted high-dimensional model for the parameters of interest, $T$ matches such parameters in a lower dimension, and $0 \leq \lambda \leq 1$ is the shrinkage intensity that weights the importance of the two models and allows positive definiteness in the resulting matrix. A common strategy is to use $U = S_X$ (sample covariance) and $T = \text{diag}(S_X)$ (diagonal of sample covariance). Then, off-diagonal elements are shrunk towards zero as $\lambda$ increases. The selection of an optimal shrinkage intensity that balance variance (mostly due to $U$) and bias (mostly due to $T$) in the estimator is proposed in Schäfer and Strimmer (2005). The ridge constraint for matrix inversion (Hoerl and Kennard, 1970) is a sub-case of the shrinkage procedure in eq. (3.9) when $U = I$.

In a Bayesian context, Daniels and Kass (1999, 2001) present several shrinkage alternatives to (3.9) that use Bayesian hierarchical models that go further than placing a Wishart prior distribution (the conjugate prior) on the sample covariance matrix. For instance, the authors describe a Markov chain Monte Carlo (MCMC) algorithm that uses a normal prior distribution centered at zero on the Fisher transformation of the off-diagonal elements of the correlation matrix, or also a similar prior to the Givens angles. In both cases, the eigenvalues of the mode of the posterior distribution of $\hat{\Sigma}_X$ are shrunk towards a constant, positive definiteness is achievable and the inverse of $\hat{\Sigma}_X$ determines the estimated conditional dependence structure.

Thresholding approaches (Bickel and Levina, 2008) first estimate the sample covariance matrix $S_X$ and then set the elements of $S_X$ to zero by a thresholding function. For instance, soft thresholding uses a lasso type penalization by

$$ST(z,\lambda) = sign(z)(|z|-\lambda)_+, \tag{3.10}$$

which fixes to zero the coefficients with lower magnitude than $\lambda$. Other penalizations such as SCAD or adaptive lasso are an extension of the simple soft thresholding and are compared for several data settings in Rothman et al. (2009). This method does not ensure non-singularity in the estimated covariance matrix making the estimation of $\hat{\Omega}_X$ a non-trivial problem. The main advantage of using thresholding in comparison to other approaches is that it is computationally fast and thus is easily applied to real life high-dimensional studies.

Lastly, penalization approaches demand more computational efforts than thresholding since the sparsity is directly assumed in the conditional dependence structure $\Omega_X$. These methods optimize an expression that combines the log-likelihood minus a penalization term

$$(\hat{\Omega}_X^{\lambda})_{PML} = \arg\max_{\Omega_X > 0} \; [\log\det\Omega_X - \mathrm{tr}(S_X\Omega_X) - P_{\lambda}(\Omega_X)], \tag{3.11}$$

where $PML$ stands for penalized maximum likelihood. One of the most famous penalization functions is the lasso or $L_1$ norm (Banerjee et al., 2008; Friedman et al., 2007), and it is defined by

$$P_{GL}^{\lambda}(\Omega_X) = \lambda||\Omega_X||_1 = \lambda \sum_{i=1}^{p}\sum_{j=1}^{p} |\Omega_{ij}^{x}|, \tag{3.12}$$

where $GL$ stands for graphical lasso, and $\lambda$ represents the trade-off between close fit to the data and sparsity of $\Omega_X$. Even though the $L_1$ penalty in (3.12) is applied to all elements of $\Omega_X$, some authors have proposed the same penalty applied to only the off-diagonal elements (Yuan and Lin, 2007). An extension of the graphical lasso is given by the adaptive, or weighted, graphical lasso (Zhou, 2006), which incorporates a weight $V = [v_{ij}]$ for each pair of variables on the penalization by $P_{WGL}^{\lambda,V}(\Omega_X) = \lambda \sum_{i=1}^{p}\sum_{j=1}^{p} v_{ij}|\Omega_{ij}^{x}|$.

The lasso penalization approach has a Bayesian interpretation (Wang, 2012), i.e., the estimator by GL finds similar values to the mode of the posterior distribution of the model $f(X_k|\Omega) \sim \mathrm{N}(X_k \mid 0, \Omega_X^{-1})$, $k = 1,\ldots,n$, assuming a double-exponential prior distribution on the off-diagonal elements of $\Omega_X$ and an exponential distribution on the diagonal ones

$$P(\Omega|\lambda) \propto \prod_{i<j}\{\mathrm{DE}(\Omega_{ij}^{x}|\lambda)\}\prod_{i=1}^{p}\{\mathrm{Exp}(\Omega_{ii}^{x}|\lambda/2)\},$$

where DE represents the double exponential function with density $f(x) = \lambda/2\exp(-\lambda|x|)$ and Exp is the exponential function with density $f(x) = \lambda\exp(-\lambda x)$. Sampling from the posterior distribution is usually done by a MCMC procedure that turns out to be computationally intensive. The reason is that there are as many as $2^{(p(p-1)/2)}$ possible models, that for large $p$, make MCMC visits quite unreliable (Banerjee and Ghosal, 2014). To make the problem tractable, Wong et al. (2003) use the Cholesky decomposition on $\Omega_X$ and set a non-uniform prior distribution for a variable that quantifies the number of non-zero elements in the matrix (reducing the $2^{(p(p-1)/2)}$ possible models). Then, MCMC samples are generated from the posterior distribution given a Metropolis Hasting algorithm. Besides,

Mohammadi and Wit (2015a) use a birth/death MCMC method to reduce the number of operations to $kp^2$, where $k$ is the number of iterations to achieve convergence. These methods are proven to be competitive to lasso estimates. Nevertheless, the computational burden continues to be high when the dimension is of the order of thousands.

Gaussian graphical lasso works fine in the variable selection part under certain conditions (Zhao and Yu, 2006) but generates a biased estimator. In case there is an interest in finding a good approximation of the magnitude of the coefficients, a SCAD penalty (Fan et al., 2009) is typically used instead. Another proposal is the constrained $L_1$-minimization for inverse matrix estimation -CLIME- (Cai et al., 2011). The optimization problem is given by the minimization of the $L_1$ norm of $\Omega_X$ constraining $|\Sigma_X\Omega_X - I|_\infty \leq \lambda$. This method presents some interesting convergence and computational characteristics. For instance, the optimization problem can be separated in $p$ different problems so that parallel computations can be performed. A similar idea is used in Yuan (2010) by fitting $p$ regression models using the so called Dantzig selector estimator (Candes and Tao, 2007).

Several contributions have also been proposed to relax the Gaussian assumption. Among others, Lafferty et al. (2012) present various non-parametric methods to estimate sparse conditional dependence structures, Liu et al. (2012) introduce a semi paremetric copula procedure that employs robust correlation estimators such as Spearman's rho and Kendall's tau, or Abegaz and Wit (2015) describe a Gaussian copula graphical model approach to infer conditional dependence among variables that can be both discrete and continuous.

Some of the reviewed methods can be used in the free statistical software R. The package **GeneNet** (Schäfer et al., 2006) contains shrinkage estimators in the form of eq. (3.9). The package **huge** (Zhao et al., 2012) consists of functions that solve the graphical lasso minimization problem presented in eq. (3.11) and (3.12). Similarly, the package **Camel** (Liu and Wang, 2012) implements the so called tuning-insensitive graph estimation and regression (tiger) approach which can be used to estimate sparse precision matrices. The R package **FastGGM** (Wang et al., 2016) uses a graphical lasso algorithm that is designed to estimate huge biological networks. The Bayesian graphical lasso is also implemented in R by Mohammadi and Wit (2015b) within the package **BDgraph**.

Most of the approaches seen in this section require the selection of a regularization parameter $\lambda$ which controls the sparsity of the estimated regression coefficients / precision matrix elements. Some of the ways that are used in the literature to chose this tuning parameter are reviewed in Section 3.5.

## 3.3 Joint estimation of multiple precision matrices

Consider the problem of estimating two precision matrices corresponding to the conditional dependence structure of two i.i.d. Gaussian p-dimensional vectors $Y^{(1)} : \{Y_1^{(1)}, \ldots, Y_{n_1}^{(1)}\}$ and $Y^{(2)} : \{Y_1^{(2)}, \ldots, Y_{n_2}^{(2)}\}$, where $p \gg n_1$ and $p \gg n_2$. The estimation of precision matrices $\Omega_1 = [\Omega_{ij}^{(1)}]$ (for $Y^{(1)}$ samples) and $\Omega_2 = [\Omega_{ij}^{(2)}]$ (for $Y^{(2)}$ samples) separately in a high-dimensional setting has been well studied in the past few years (see Section 3.2.2) but a potential joint structure (or commonality) of the two condi-

tional dependence structures tends to be ignored in these articles. This particularity is exploited in a few recent contributions by using some type of penalization that encourages similarity between the precision matrices or their underlying graph structures. In Section 3.3.1 we review methods to estimate the two precision matrices together and in Section 3.3.2 we focus on the available alternatives to estimate directly the difference matrix $\Omega_d = \Omega_2 - \Omega_1$.

### 3.3.1 Joint graphical lasso

Define the joint graphical lasso estimation problem (JGL) by

$$\{\hat{\Omega}_1^{\lambda}, \hat{\Omega}_2^{\lambda}\}_{JGL} = \arg\max_{\Omega \succ 0} \sum_{k=\{1,2\}} [\log\det\Omega_k - tr(S_k\Omega_k)] - P_{\lambda_1,\lambda_2}(\Omega_1, \Omega_2), \tag{3.13}$$

which is the sum of log-likelihood functions for the two datasets minus a penalty term. The first important proposal to estimate multiple sparse precision matrices simultaneously was described in Guo et al. (2011). The authors suggest to use a group lasso maximum likelihood estimator (GGL) determined by the optimization problem in eq. (3.13) with penalty

$$P_{\lambda_2}^{GGL}(\Omega) = \lambda_2 \sum_{i \neq j} \left( \sum_{k=\{1,2\}} |\Omega_{ij}^{(k)}| \right)^{1/2}, \tag{3.14}$$

where $\lambda_2$ is a tuning parameter that controls similarity between the graph structures in the two classes, thus ignoring the sign of non-zero values. An algorithm used to solve eq. (3.13) with penalty (3.14) is based on solving two graphical lasso problems iteratively (one for $Y^{(1)}$ and one for $Y^{(2)}$) until convergence by keeping the estimate of the other precision matrix fixed. The maximization problem can be immediately extended to account for datasets with more than two classes.

The two main problems of the precision matrix estimator determined by the penalty in eq. (3.14) are its non-convexity and the control of sparsity (graph structure similarity and sparsity are affected together by $\lambda_2$). Danaher et al. (2014) address these two issues by proposing to use the penalty

$$P_{\lambda_1,\lambda_2}^{GGL}(\Omega) = \lambda_1 \sum_{i \neq j} \sum_{k=\{1,2\}} |\Omega_{ij}^{(k)}| + \lambda_2 \sum_{i \neq j} \left( \sum_{k=\{1,2\}} (\Omega_{ij}^{(k)})^2 \right)^{1/2}. \tag{3.15}$$

In this case, $\lambda_1$ controls the sparsity of the precision matrices and $\lambda_2$ controls their common or not common graph structure. In the same article, Danaher et al. propose to use a fused lasso penalization approach that differs from the group lasso since it encourages similarity between the values of the precision matrix elements rather than their underlying non-zero structure. Fused penalization was previosly used in a time series context in Witten et al. (2009) and Kolar et al. (2012) to smooth consecutive regression coefficients (and it is reviewed later in Section 3.4). This concept is applied to

the precision matrix elements for two classes so

$$P^{FGL}_{\lambda_1,\lambda_2}(\Omega_2,\Omega_1) = \lambda_1||\Omega_2||_1 + \lambda_1||\Omega_1||_1 + \lambda_2 \sum_{i=1}^{p}\sum_{j=1}^{p} |\Omega^{(2)}_{ij} - \Omega^{(1)}_{ij}|. \tag{3.16}$$

Both optimization problems proposed in Danaher et al. (2014) can be solved by an alternating directions method of multipliers (ADMM) algorithm (Boyd, 2010) and are implemented within the R package **JGL** (Danaher, 2013).

A more general method to jointly estimate sparse-similar precision matrices is given in Mohan et al. (2014). The authors consider a node-based approach that focus on the differential patterns between multiple classes, but they do it in the variables space rather than in the edges space. They introduce the following penalty

$$P^{RCON}_{\lambda_1,\lambda_2}(\Omega_2,\Omega_1) = \lambda_1||\Omega_2||_1 + \lambda_1||\Omega_1||_1 + \lambda_2 \sum_{i=1}^{p}\sum_{j=1}^{p} G_q(\Omega^{(2)}_{ij} - \Omega^{(1)}_{ij}), \tag{3.17}$$

where $G_q$ defines the row-column overlap norm (RCON) with $L_1/L_q$ norm and $1 \leq q \leq \infty$ such that

$$G_q(\Omega) = \min_{V} \sum_{g=1}^{p} ||V_g||_q, \quad \text{s.t. } \Omega = V + V^t.$$

By using a penalization on the (possible) non-symmetric matrix $V$, it encourages structures of interest on the columns and rows of the differential matrix $\Omega_d$. For instance, for $q = 1$, the penalty coincides with FGL, and for $q = 2$ or $q = \infty$, the non-zero structure in the differential coefficients is shared in the whole row and column of $\Omega_d$. The assumption of these latter cases is that once a variable is deferentially connected to another variable, then it must be differently connected to all other variables (except for cases where both $\Omega^{(2)}_{ij} = \Omega^{(1)}_{ij} = 0$). An ADMM algorithm is also used to solve the optimization problem. In a similar framework, Cai et al. (2016) provides an $L_\infty/L_1$ optimization problem to jointly estimate the two matrices.

Other proposals include Lee (2015), who defines a CLIME-type optimization problem (Cai et al., 2011) that jointly estimates multiple precision matrices, and that is applicable to Gaussian and non-Gaussian family distributions, or Wit and Abbruzzo (2015), who estimate a joint precision matrix by considering several graph structure designs prior to estimation. Recently, in Xie et al. (2016), the joint estimation is made when the two datasets corresponding to two different classes are dependent. The authors assume an additive model (see Section 2.2): $Y^{(1)}_k = Z_k + H_k$ and $Y^{(2)}_k = Z_k + T_k$, for any $k \in \{1,\ldots,n\}$, where $Z_k$ is the common measure, and $\{H_k, T_k\}$ are the unique structures of the two classes. They use the cross-covariance $\text{cov}(Y^{(1)}_k, Y^{(2)}_k)$ to represent the common structure and describe an expectation maximization (EM) algorithm to infer the common and unique conditional dependence structures.

Bayesian inference for these type of joint models is also available in the literature. For instance, in Peterson et al. (2015) the similarity between related precision matrices is supported by taking a

Markov random field prior to the graph structures. This is done by considering a reference network as well as the two networks corresponding to the two classes. The authors assume that the prior probability of existing an specific edge linking two variables in either of the two unique networks is positively related to the presence of the same edge in the reference network (giving higher prior probabilities). A G-Wishart prior (Roverato, 2002) is placed on the two precision matrices.

Most of the joint estimation procedures seen in this section require the selection of two regularization parameters $\lambda_1$ and $\lambda_2$ that control sparsity and similarity of the estimated precision matrix elements in the two classes, respectively. Section 3.5 presents some of the available alternatives in the literature to estimate the parameters.

### 3.3.2 Direct estimation of differential network

By estimating $\Omega_2$ and $\Omega_1$, it is immediate to obtain the difference matrix $\Omega_d = \Omega_2 - \Omega_1$, or the network structures defined by set of edges $E_1 = \{(ij) : \Omega^{(2)}_{ij} = 0,\ \Omega^{(1)}_{ij} \neq 0\}$ and $E_2 = \{(ij) : \Omega^{(2)}_{ij} \neq 0,\ \Omega^{(1)}_{ij} = 0\}$. However, if the interest is only in $\Omega_d$, its estimation can be done directly, i.e., estimating the marginals $\Omega_2$ and $\Omega_1$ is not required.

In Zhao et al. (2014), the difference matrix $\Omega_d$ is estimated directly by using the fact that, in theory, $R_2\,\Omega_d\,R_1 - (R_1 - R_2) = 0$. The authors suggest to assume sparsity in only $\Omega_d$, allowing $\Omega_2$ and $\Omega_1$ to be dense. The proposed optimization problem follows a CLIME-type constraint,

$$\begin{aligned} \hat{\Omega}_{d_{CLIME}} = \arg\min_{\Omega_d} & |\Omega_d| \\ \text{s.t. } & |(\hat{R}_1\Omega_d\hat{R}_2) - (\hat{R}_2 - \hat{R}_1)|_\infty \leq \lambda_n. \end{aligned} \tag{3.18}$$

Here $\hat{R}_2 = n_2^{-1}\sum_{k=1}^{n_2} Y_k^{(2)'} Y_k^{(2)}$ and $\hat{R}_1 = n_1^{-1}\sum_{k=1}^{n_1} Y_k^{(1)'} Y_k^{(1)}$. Zhao et al. (2014) consider the equivalent linear program where the minimization of $|\Omega_d|$ is subject to $|(\hat{R}_1 \otimes \hat{R}_2)\,\mathrm{vec}(\Omega_d) - \mathrm{vec}(\hat{R}_2 - \hat{R}_1)|_\infty \leq \lambda_n$ with $\otimes$ indicating the Kronecker products operator. This requires the computation of a $p^2 \times p^2$ matrix in the constraint $(\hat{R}_1 \otimes \hat{R}_2)$. The authors use the symmetry property of $\Omega_d$, to further solve the problem in eq. (3.18) by computing a $p(p-1)/2 \times p(p-1)/2$ matrix instead of the $p^2 \times p^2$ matrix given in $\hat{R}_1 \otimes \hat{R}_2$, but stronger theoretical conditions are implied. An ADMM-type recursive algorithm is employed to find an estimator for the differential precision matrix.

A similar problem is tackled in Mitra et al. (2016), who estimate differential networks using a Bayesian formulation. The authors assume a uniform prior for the edges in the first graph and a Bernoulli trial for the equality of edges between the first and second graph. This totally specifies the prior for the graph in the second class. A MCMC approach is used to infer the posterior distribution of the differential network.

## 3.4 Joint estimation of multiple linear regression models

In this section we study the problem of estimating regression coefficient matrices in a high-dimensional framework. We first review an approach for estimating a single sparse regression coefficients matrix and then we discuss the methods available in the literature that jointly estimate sparse regression coefficients in more than one class of observations.

### 3.4.1 Sparse multivariate linear regression

Consider the multivariate Gaussian linear regression model that links pairs of observations $\{X_k, Y_k\}_{k=1}^n$, where $X_k$ are $p$-dimensional covariates and $Y_k$ contain the $q$-dimensional response variables,

$$Y_k \sim N(\beta X_k, \Sigma_e), \text{ for any } k = 1, \ldots, n. \tag{3.19}$$

The covariance matrix $\Sigma_e = \text{cov}(Y_k - \beta X_k)$ describes the residuals linear dependence structure, and the linear regression coefficients $\beta$ relate explanatory variables $X$ to the response variables $Y$ and are typically estimated by least squares:

$$\hat{\beta} = \arg\min_{\beta} \left( \frac{1}{2n} ||Y - X\beta||_2^2 \right), \tag{3.20}$$

with solution $\hat{\beta} = (X'X)^{-1}X'Y$.

A multivariate linear regression method to estimate $\beta$ for high-dimensional data, where $(X'X)$ is not invertible, is proposed in Rothman et al. (2010). The authors jointly estimate the regression coefficients in $\beta$ as well as the precision matrix $\Omega_e = \Sigma_e^{-1}$, that describes the error's conditional dependence, by solving the following minimization problem:

$$(\hat{\beta}, \hat{\Omega}) = \arg\min_{B, \Omega_e} \left\{ \text{tr}\{\frac{1}{n}(Y - X\beta)^t (Y - X\beta)\Omega_e\} - \log|\Omega_e| + \lambda_1 \sum_{i \neq j} |\Omega_{ij}^{(e)}| + \lambda_2 \sum_{j=1}^{p} \sum_{i=1}^{q} |\beta_{ij}| \right\} \tag{3.21}$$

where $\lambda_1$ and $\lambda_2$ are penalization parameters. The estimator in (3.21) is "bi-convex", i.e., it is a convex function once assuming that either $\beta$ or $\Omega_e$ is known. Hence, the proposed algorithm to find a solution of eq. (3.21) uses an iterative process that combines a cyclical coordinate descend algorithm (Friedman et al., 2007) to find $\hat{\beta}_X$ (keeping $\Omega_e$ fix) and a glasso algorithm (Friedman et al., 2007) to find $\hat{\Omega}_e$ (keeping $\beta_X$ fix). The multidimensional regression approach is implemented within the R package **MRCE** (Rothman, 2013).

### 3.4.2 Joint estimation of regression lasso

Consider the extension of the multivariate linear model described in Section 3.4.1 when two samples corresponding to two different classes are observed for both covariates and responses: $\{X_k^{(1)}, Y_k^{(1)}\}_{k=1}^{n_1}$

and $\{X_k^{(2)}, Y_k^{(2)}\}_{k=1}^{n_2}$. Assume predictors and responses are associated by a Gaussian linear model

$$(Y_k^{(1)}, Y_k^{(2)}) \sim N_{2p}\left(\begin{bmatrix}\beta^{(1)} X_k^{(1)}\\ \beta^{(2)} X_k^{(2)}\end{bmatrix}', \Sigma_e\right),\ \Sigma_e = \begin{bmatrix}\Sigma_e^{(11)} & \Sigma_e^{(12)}\\ \Sigma_e^{(12)'} & \Sigma_e^{(22)}\end{bmatrix}, \tag{3.22}$$

where $\Sigma_e$ is the residuals covariance matrix. All methods presented in this section assume independence between datasets, with both $\Sigma_e^{(12)} = 0$ and $\Sigma_X^{(12)} = 0$ and allowing $n_1 \neq n_2$.

Fused regression lasso is initially proposed by Tibshirani et al. (2005) to address the problem of estimating regression coefficients in high-dimensional data when the covariates are ordered (e.g. time ordering). The authors smooth the changes for consecutive estimated coefficients by considering an additional penalty term in the regression lasso optimization problem

$$\hat{\beta}_{OFRL}^{\lambda} = \arg\min_{\beta}\left[\frac{1}{2n}||y - \beta X||_2^2 + P_{\lambda_1,\lambda_2}(\beta)\right], \tag{3.23}$$

with

$$P_{\lambda_1,\lambda_2}^{OFRL}(\beta) = \lambda_1||\beta||_1 + \lambda_2\sum_{i=2}^{p}|\beta_i - \beta_{i-1}|_1,$$

where $OFRL$ stands for ordered fused regression lasso.

This idea is used in Zhang and Wang (2012) to encourage similarity of regression coefficients from different classes in a join linear regression model (JLR):

$$\hat{\beta}_{JLR}^{\lambda} = \arg\min_{\beta}\left[\frac{1}{2n}(||y^{(1)} - \beta^{(1)}X^{(1)}||_2^2 + ||y^{(2)} - \beta^{(2)}X^{(2)}||_2^2) + P_{\lambda_1,\lambda_2}(\beta)\right], \tag{3.24}$$

with

$$P_{\lambda_1,\lambda_2}(\beta)^{FRL} = \lambda_1\sum_{M=\{1,2\}}||\beta^{(M)}||_1 + \lambda_2\sum_{i=1}^{p}\sum_{j=1}^{q}|\beta_{ij}^{(2)} - \beta_{ij}^{(1)}|,$$

where $FRL$ stands for fused regression lasso and the response is a single variable. Let $V = \{1,\ldots,p\}$ be the set of variables, in Zhang and Wang (2012), the optimization problem in eq. (3.24) is solved using a block coordinate descent algorithm, such that pair of parameters $(\beta_i^{(1)}, \beta_i^{(2)})$, for $i \in V$, is updated once at a time considering the rest $(\beta_j^{(1)}, \beta_j^{(2)})$, $j \in V \setminus i$, fixed.

In the technical report by Lam et al. (2016), a related fused penalty proposal is used when the response is also a high-dimensional dataset. The authors present an $L_2$-fused estimator (FRL2) that is the solution of the minimization problem in (3.24) with penalty defined by

$$P_{\lambda_1,\lambda_2}(\beta)^{FRL2} = \lambda_1\sum_{M=\{1,2\}}||\beta^{(M)}||_2 + \lambda_2\sum_{i=1}^{p}\sum_{j=1}^{q}(\beta_{ij}^{(2)} - \beta_{ij}^{(1)})^2.$$

Since all components are in $L_2$ norm, the problem is convex and can be solved through linear regression with an augmented design matrix.

As for the joint estimation of precision matrices, optimization problems that jointly estimate regression coefficients in two classes of observations require the selection of two regularization

parameters $\{\lambda_1, \lambda_2\}$. Section 3.5 describes some of the methods that have been used in the literature.

## 3.5 Selection of tuning parameters

In this section we review some of the methods in the literature to select the tuning parameters in both regression and graphical lasso problems. Firstly we describe standard model selection methods such as Cross Validation (CV), Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), and we give some justification why these methods might not be useful for selecting tuning parameters in our high-dimensional data setting. Finally, we consider three other approaches that have been proposed to account for situations when the sample size is smaller than the dimension.

*Regularization parameter selection for sparsity parameter*

Cross Validation is a widely used technique for variable selection which aims to minimize the predictive error at fixed value of $\lambda$. It is based in splitting the data randomly in two blocks, one for training and one for testing, and finding the predictive error in the testing data using the training to fit the model. The best tuning parameter by cross-validation is the one with the lowest average error over several (or all possible) instances of the splitting process. For instance, the $\lambda$ selection by leave-one-out CV (where testing data only contains one observation) is determined by

$$\lambda_{CV} = \arg\min_{\lambda} \sum_{j=1}^{n} (y_j - X_j \hat{\beta}_{\lambda}^{-j})^2, \tag{3.25}$$

where $\hat{\beta}_{\lambda}^{-j}$ is the lasso (or any other penalization presented in Section 3.2.1) solution using all data except to the pair $(X_j, y_j)$. CV works fine for high-dimensional scenarios, the only consideration is in the objective of the method which falls in the prediction rather than the recovery of the non-zero structure of the regression coefficients. Wasserman and Roeder (2009) show that CV overfits the graphical structure of $\beta$ and propose to perform an addition variable selection stage on the CV-optimal model where some covariates are eliminated by hypothesis testing. When the aim is to estimate a precision matrix in the case $p > n$, when its sample version is not unique, a method based on CV is not as straightforward as there is no "observable" equivalent of the precision matrix. If we consider the link between regression and precision matrix described in Section 2.1, a CV adaptation for graphical models is directly approachable by

$$\lambda_{CV} = \arg\min_{\lambda} \sum_{g=1}^{p} \sum_{j=1}^{n} (X_{jg} - X_{j,-g} \hat{\beta}_{\lambda}^{(g,-j)})^2, \tag{3.26}$$

where $\hat{\beta}_{\lambda}^{(g,-j)}$ is a vector of regression coefficients linking $X_{-j,g}$ with all the other variables $X_{-j,-g}$, which is determined from the estimated precision matrix following eq. (2.1).

Other likelihood (or least squares) based risk functions to select $\lambda$ such as AIC and BIC are useful

to find a compromise between goodness of fit to the data and model over-fitting

$$\lambda_{AIC} = \arg\min_{\lambda} L(y, X, \lambda) + s(\theta_\lambda), \quad \lambda_{BIC} = \arg\min_{\lambda} L(y, X, \lambda) + s(\theta_\lambda)\log(n)/2, \tag{3.27}$$

where $L(y, X, \lambda) = ||y - \hat{\beta}_\lambda X||_2^2$ for linear regression and $L(y, X, \lambda) = -\log\det\hat{\Omega}_X^\lambda + \text{tr}(S_Y\hat{\Omega}_X^\lambda)$ for graphical modeling. The last term $s(\theta_\lambda)$ defines the effective number of free parameters and tends to be approximated by the number of non-zero coefficients. The AIC penalty has its origins from information theory, it is found by minimizing the expected Kullback-Leibler divergence between estimated and "true" models. The BIC comes from a Bayesian background instead, and the obtained risk function is based on the Laplace approximation of the log likelihood of the model assuming constant priors for all possible models. For sufficiently large $n$, the BIC selection finds consistently sparser estimators than AIC. The main conceptual problem of AIC and BIC for high-dimensional problems is that these are asymptotic methods by definition, which assume fixed dimension $p$ for increasing $n$, but with $p > n$ this justification is not appropriate. This justification is supported by Liu et al. (2010) in a simulated data analysis, where AIC and BIC are found to overestimate the graphical structure of $\Omega_X$ even for cases where $n$ is slightly larger than $p$.

An extended version of BIC, called eBIC, is given in Chen and Chen (2008). The eBIC reconsiders the constant priors assumption for the models of BIC by encouraging models with extreme sparsity levels in both ends (highly sparse and dense matrices) as follows

$$\lambda_{eBIC} = \arg\min_{\lambda} L(y, X, \lambda) + K\log(n)/2 + 2\phi\log(\tau(\theta_\lambda)), \ \ \tau(\theta_\lambda) = \begin{pmatrix} K \\ s(\theta_\lambda) \end{pmatrix}, \tag{3.28}$$

with $\theta_\lambda = \hat{\beta}_\lambda$ for regression or $\theta_\lambda = \hat{\Omega}_X^\lambda$ for precision matrix estimation. The hyper-parameter $\phi$ is defined between 0 and 1, so when $\phi = 0$, eBIC coincides with BIC, and as $\phi$ increases, it penalizes sparsity models (in terms of degree distribution) that are more likely to happen just by chance. Another proposal is given in Zhang and Shen (2010) with the introduction of the lasso regression with $\text{RIC}_c$ penalty:

$$\lambda_{RIC_c} = \arg\min_{\lambda} L(y, X, \lambda) + \frac{4n\sigma^2 s(\hat{\beta}_\lambda)(\log p + \log\log p)}{\lambda}, \tag{3.29}$$

where $s(\hat{\beta}_\lambda)$ is the number of non-zero elements in the lasso estimate.

A common consideration for the methods described above is that they use the estimated values for $\beta$ or $\Omega_X$ which make algorithms like neighbourhood selection (Meinshausen and Bühlmann, 2006), see Section 5.2.3, not applicable as only estimate the graph structure of $\Omega_X$. Liu et al. (2010) propose a method that contrasts with the usual variable selection statistics since it only considers the estimated conditional dependence graph structure.The authors consider the stability approach to regularization selection (StARS) to chose $\lambda$ by controlling the desirable approximated variability in the estimated graphs. The variability is estimated for each $\lambda$ using a subsampling approach. The motivation of this method resides in the fact that the selection of $\lambda$ problem, which is difficult to explain by itself,

is transformed to the selection of the desired amount of variability in the graph, which is easier to interpret. Another stability approach is discussed in Meinshausen and Bühlman (2010), who control the expected graph edges false discovery rate. The authors estimate $\Omega_X$ by an average subsampling graphical lasso method such that the effect of the choice of $\lambda$ is very low.

*Regularization parameter selection for joint estimation methods*

The joint estimation problems described in Section 3.3 and Section 3.4 require the selection of two regularization parameters: $\lambda_1$ (sparsity) and $\lambda_2$ (similarity), and the combination of the two characterizes the estimated network sizes (both common network and differential network). Out of a grid of values for both $\lambda_1$ and $\lambda_2$, Danaher et al. (2014) use an AIC criterion that combines the AIC's of the two estimated precision matrices, Guo et al. (2011) consider both BIC and CV to obtain the best $\lambda$, Lee (2015) employ a $K$-fold cross-validation approach, or Xie et al. (2016) find instead the best estimated precision matrices that minimize the eBIC criterion.

## 3.6 Other multivariate methods for high-dimensional data

Other multivariate methods such principal component analysis (PCA), partial least squares (PLS) or canonical correlation analysis (CCA) are used in order to understand the relationship between variables, and to group samples in different clusters. In this section we review some of theses approaches as well as their extensions to encourage sparse solutions.

*Principal component analysis and independent component analysis*

Both PCA (Hotelling, 1933) and independent components analysis -ICA- (Comon, 1994) are multivariate techniques which intend to project a data matrix in a lower dimension by keeping as much information as possible of such original matrix. PCA relies on the second moment of the data (i.e., it finds linear combinations with data that achieve maximum variance) and hence it assumes Gaussian features whereas ICA exploits higher order moments (e.g., minimizes the kurtosis) which are not demanded in a Gaussian context. A sparse variation of the methods is used in a high-dimensional data setting by regularizing the values of the loadings vectors, which describe the relationship between the original variables and the unit-scaled components. Jolliffe et al. (2003) introduce SCoTLASS (Simplified Component Technique- LASSO), which is a procedure that finds the sparse loadings of PCA by directly constraining the $L_1$ norm of the coefficients. Later, Zou et al. (2006) consider the regression reformulation of the PCA problem and include an elastic-net penalization for the loadings in a two stages based algorithm, or Shen and Huang (2008) propose a regularized singular value decomposition (SVD) with lasso/SCAD penalty for the loadings. Similarly, Yao et al. (2012) use soft thresholding on the independent components to obtain the sparse coefficients. These techniques are implemented within the R package **mixOmics** (Le Cao et al., 2016) and functions `spca` and `sipca`. An analogous method for PCA when the input is a contingency table is also available and it is well known as correspondence analysis (Yelland, 2010).

*Sparse partial least squares, canonical correlation analysis and co-inertia analysis*

Partial least squares -PLS- (Wold, 1966) is a multidimensional technique which aims to find a projection of two data sets $X$ and $Y$ such that the covariance between $X$ and $Y$ is maximized. Similarly, canonical correlation analysis -CCA- considers a projection of two data sets $X$ and $Y$ such that the correlation between a linear combination of $X$ and $Y$ is maximized. Lê Cao et al. (2008) for PLS and Lê Cao et al. (2009) for CCA present the extension of these methods for a high-dimensional data setting. The authors enforce sparsity in the projections using an $L_1$ penalization on the loadings for $X$ and an $L_2$ penalization on the loadings for $Y$. The solutions of the underlying optimization problems are found by recursive algorithms which are implemented within the R package **mixOmics** and the functions `spls` and `rcc`. Alternatively, co-inertia analysis -CIA- (Doledec and Chessel, 1994) is used as a general approach to connect two datasets that can be of any type (either continuous or categorical). The CIA method is available in R within the **ADE-4** package (Thioulouse et al., 1997).

*Visualization techniques*

The sparse partial least squares and reduced canonical correlation methods project two datasets (say $X$ and $Y$) in a common space, however they do not find directly the associations between the features on X and Y. González et al. (2012) present a graphical tool, called correlation circle plots, in order to gather similar characteristics among the variables in the new projected space. In particular, it measures the correlation between each of the original variables and the projection of the same in the new space. For instance, considering the representation in the plane, points are within a circle of radius 1 in which similarity between variables far away from the origin can be directly interpreted but more dimensions are needed to explain all the other points.

*Global measure of dependence between two datasets*

Methods reviewed so far in this section like CIA, regularized CCA or sparse PLS are computationally intensive for high-dimensional data. A conceptually simple statistic can be considered to measure global similarity between two data matrices, say $X$ and $Y$. This might be useful to discern which are the most important pair of datasets for a complete analysis (e.g., by CIA, CCA or PLS) when many datasets are available. This statistic is introduced in Escoufier (1973) and it is widely known as RV coefficient

$$\mathrm{RV}(X,Y) = \frac{\mathrm{tr}(\hat{R}_{XY}\hat{R}'_{XY})}{\sqrt{\mathrm{tr}(\hat{R}^2_{XX})\mathrm{tr}(\hat{R}^2_{YY})}}.$$

For high-dimensional data, this measure is highly biased under independence between $X$ and $Y$. Mayer et al. (2011) propose a related statistic $RV_{adj}(X,Y)$ based on adjusted r-squares coefficients $\hat{r}^2_{adj}(x_i,y_j) = 1-(n-1)/(n-2)(1-(\hat{r}^{xy}_{ij})^2)$ by

$$\mathrm{RV}_{adj}(X,Y) = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q}\hat{r}^2_{adj}(x_i,y_j)}{\sqrt{\sum_{i,j=1}^{p}\hat{r}^2_{adj}(x_i,x_j)\sum_{i,j=1}^{q}\hat{r}^2_{adj}(y_i,y_j)}},$$

so that its expected value under independence is equal to zero.

## 3.7 Application to omics datasets

There is an endless number of contributions in the biological literature that apply statistical methods to the analysis of omics (e.g., genomics, proteotomics, metabolomics) datasets. This section highlights a short list of these articles that employ multivariate data analysis techniques for high-dimensional data, especially lasso-based approaches. For an extended list of methods and applications that have broken through in the omics data analysis and integration literature in the past two decades, see Joyce and Palsson (2006) and more recently Bebek et al. (2012) and Wanichthanarak et al. (2015).

*Lasso-based regression*

Yang et al. (2013) identify some of the protein markers associated to progression-free survival of patients with ovarian cancer by applying a lasso regression model. Also employing lasso regression, Simeonov and Himmelstein (2015) relate demographic and cancer risks to several tumor type incidences in order to detect important charactaristics in lung cancer. Timpe et al. (2015) use lasso and elastic-net regression to model the sensitivity of 90 drugs in breast cancer with respect to messenger RNA (mRNA) expression for 160 glycoproteins and two other sets of protein data. Hughey and Butte (2015) also contemplate the elastic-net penalization to classify four lung cancer subtypes using as input the gene expression profile of 629 samples.

*Lasso-based conditional dependence networks*

Chun et al. (2013) apply the group lasso penalized maximum likelihood approaches of Guo et al. (2011) and Danaher et al. (2014) to estimate four conditional graphical models that correspond to the dependence structure of gene expression for four different tissues. They also have the information of another dataset with the genetic markers, so as novelty they consider the estimation problem of the four conditional dependence structures once accounting for the genetic marker profiles. The integration and analysis of methylation with gene expression data is studied in Gadaleta and Bessonov (2015), who integrate gene expression and methylation presence for a dataset with patients with glioblastoma cancer. The authors employ lasso estimates for the regression coefficients linking gene expression (response variables) and methylation presence (explanatory variables) as well as for the precision matrix that considers conditional dependence among genes in only the gene expression data.

*Other multivariate data methods*

In Fagan et al. (2007), co-inertia multivariate technique (CIA) is used to relate two datasets containing the information of gene and proetomic expression for the same individuals. GO annotation terms describe the functions of specific genes according to Gene Ontology (Ashburner et al., 2000) and are superimposed on the CIA projections with the intention to detect the roles of the genes and proteins that are highly expressed. A similar idea is presented in Meng et al. (2014) by employing gene expression of several tumors types as well as a second dataset from ovarian cancer patients profiled from two microarray platforms. Sheng et al. (2011) use ICA to integrate gene expression and copy

number data and then find subsets of the genes with coherent expression patterns and large variation across samples (process commonly known as shaving).

*Data integration for more than two datasets*

Other proposals that study the data integration of multiple omics datsets include Kuznetsov et al. (2009), who use 4 different type of datasets to describe gene connections. These are KEGG pathways, protein-protein networks, expression correlation matrices corresponding to normal human tissues and 6 disease state tissues, and transcription factor binding sites (TFBS). The strength of similarity between datasets is evaluated using a score, and its significance is assessed by comparing it to the analogous scores under random associations. Kamburov et al. (2011) introduce the web tool IMPaLA for joint pathway analysis of transcriptomics or proteomics and metabolomics data. It performs enrichment analysis (it finds terms that are over-represented in a predefined pathway) with user-specified lists of metabolites and genes using over 3000 pre-annotated pathways from 11 databases. Gosline et al. (2012) develop SAMNet (Simultaneous Analysis of Multiple Networks), which uses a constrained optimization approach to analyze signaling and transcriptomic data from multiple experiments and relate estimated graphs to protein-protein interaction networks.

## 3.8 The novelty of the present work

The main aim in this thesis is to develop methodology to fully analyze and integrate multiple high-dimensional datasets that come from the application to genomic data. We focus on testing and estimation problems for linear dependence structures such as correlation matrices, precision matrices and regression coefficient matrices. As reviewed in this chapter, these are topics of great concern in the statistical literature which have been extensively studied in the last 20 years. Nevertheless, some methodological gaps are still present and we intend to explore them in the following chapters.

As an initial topic we consider global statistical testing whether two dependence structures corresponding to samples distinguishing two classes are equal or not, which is formulated as an hypothesis testing problem for equality of correlation matrices. In the statistical literature, several methods are proposed to solve such hypothesis testing issue (see Section 3.1) but typically assume that the observations in the two classes are independent. The studied methods either contemplate sum of squares based test statistics (where all correlation differences influence the test statistic) or maximum test statistics (where only the largest correlation difference is used in the test). As novelty, we propose similar methods to account for cases where observations are dependent, which frequently occur in biological data when, for the same individual, it is obtained the information in more than one sample (e.g., different time points, treatments or tissues). We also consider a test statistic that lies in between sum of squares and maximum test statistics as given by the sum of exceedances above a threshold. This is close to the sum of squares for thresholds near zero and finds more similar powers to the maximum test as the threshold increases. Besides, for a wise selection of the threshold, this procedure

can dominate the power of the test over the other two methods.

Secondly, when reviewing the literature on estimating conditional dependence structures in high-dimensional data, e.g., graphical lasso (see Section 3.2), we realized that most of the methods needed the selection of a tuning parameter $\lambda$, which affects the sparsity levels of estimated precision matrices, denoted by $\hat{\Omega}$ (as well as regression coefficient matrices $\hat{\beta}$). Even though this parameter is crucial for interpreting the graph structure of the estimated conditional dependence structures, researchers have proposed an uncountable number of estimators for the precision/regression coefficient matrices but have overlooked the regularization parameter selection issue in most of the occasions. Standard methods use expressions based on the likelihood function to optimize a certain risk function, e.g., cross-validation, AIC, BIC or RIC (Chen and Chen, 2008; Zhang and Shen, 2010) but ignore the graph structure of the underlying estimated matrices $\hat{\Omega}$ or $\hat{\beta}$. This is tackled in Liu et al. (2010), who control the variability of estimated graphs without employing any likelihood-based expression. In a similar vein we propose several risk functions that only focus on the graph structure of estimated precision matrices $\hat{\Omega}$ that can have an interest for interpreting biological data. For instance, we consider novel selection approaches that monitor network characteristics as clustering structure, graph connectivity or graph vulnerability.

Finally, a natural extension of graphical lasso is studied in Zhang and Wang (2012), Danaher et al. (2014) or Tibshirani et al. (2005), among others, to estimate conditional dependence structures in multiple classes of observations (see Section 3.3 and Section 3.4). These assume both sparsity in the precision matrices (or regression coefficient matrices) and elementwise similarity between such matrices. As seen for the testing procedures, most of the approaches found in the literature assume that observations in different classes are independent. Recently, Xie et al. (2016) accounts for dependence between datasets by assuming an additive model to estimate several precision matrices. Similarly, we propose a general method to estimate joint precision matrices, which is an extension of the fused graphical lasso approach introduced in Danaher et al. (2014), that accounts for linear dependence between datasets. Our method differs from Xie et al. (2016) since it can be used for any type of linear dependence structure between paired observations (see Section 2.2). Inspired by the work of Danaher et al. (2014), we also develop a novel weighted fused regression lasso algorithm that jointly estimates two regression coefficient matrices, and which can be used for both independent and paired observations. In the two proposed joint estimation problems, precision matrices and regression coefficient matrices, the selection of tuning parameters is a major issue, as two parameters controlling sparsity and similarity have to be selected. Standard methods as AIC, BIC or CV have been used in the literature but present similar problems in their usage as for graphical lasso. For practical needs, we provide a new approach that monitors error rates related to the probability of falsely estimating edges in both individual and difference matrices.

# Chapter 4

# Hypothesis testing problems involving correlation matrices

## 4.1 Introduction and motivation

In recent years, the improvements in technology have made it possible to collect and store reliable information for a large number of genes, metabolomics or proteins, among others, on an organism in a single sample. This typically generates datasets where the number of variables $p$ is much larger than the number of observations $n$. Statistical techniques that deal with this type of data, commonly known as high-dimensional data, with the purpose of answering biological questions, are well studied in the literature (Buhlmann and van de Geer, 2011; Sánchez and Villa, 2008). One of the main challenges relates to understanding how the genes function in a biological process and how they interact between each others in a cell. In this regard, measuring and assessing variations of gene interactions on the presence of an illness process such as cancer is important to biologists as part of discerning the gene regulatory mechanisms that control the disease.

A statistical technique that is widely used to measure interaction between pairs of genes from data is given by the Pearson correlation, which quantifies the strength of the linear dependence between two random variables. The main hypothesis testing (HT) problem we study in this chapter assesses the evidence of equality of two correlation matrices $R_1 = [r_{ij}^{(1)}]$ and $R_2 = [r_{ij}^{(2)}]$ that correspond to genomic data $Y^{(1)}$ and $Y^{(2)}$ measured in two different conditions (e.g, healthy and tumor tissues),

$$\mathrm{H}_0 : R_1 = R_2 \;\text{ vs }\; \mathrm{H}_1 : R_1 \neq R_2. \qquad \text{(eq. corr. mat. test)}$$

As part of the literature review, see Section 3.1.1, we found two main directions that address this hypothesis testing problem for high-dimensional data. The first is based on sum of squares statistics (Schott, 2007; Li and Chen, 2012), and the second is based on extreme value statistics (Cai et al., 2013; Zhou et al., 2015). To the best of our knowledge, the tests considered so far in the literature are

applicable when the random vectors $Y^{(1)}$ and $Y^{(2)}$ are independent. Here we study the implications of using the sample correlation matrices when the two datasets are dependent, particularly when they come from paired observations, in which case the cross-correlation is not zero. We propose three different tests which apply to paired data, and that are based on the average, maximum and threshold exceedances of the elementwise correlation differences.

Three other related HT approaches involving correlation matrices are also contemplated in this chapter: (a) we consider the simpler problem of testing if a correlation matrix is the identity matrix with hypothesis

$$\mathrm{H}_0 : R_1 = I \ \text{ vs } \ \mathrm{H}_1 : R_1 \neq I; \qquad \text{(id. corr. mat. test)}$$

(b) we test whether the same $g$th row in two correlation matrices is equal or not with hypothesis

$$\mathrm{H}_0 : \sum_{i \neq g} |r_{gi}^{(1)} - r_{gi}^{(2)}| = 0 \ \text{ vs } \ \mathrm{H}_0 : \sum_{i \neq g} |r_{gi}^{(1)} - r_{gi}^{(2)}| \neq 0; \qquad \text{(eq. corr. row. test)}$$

(c) we assess whether the $g$th variable is linear independent to all the other $p-1$ variables in the data by testing the hypothesis

$$\mathrm{H}_0 : \sum_{i \neq g} |r_{gi}^{(1)}| = 0 \ \text{ vs } \ \mathrm{H}_0 : \sum_{i \neq g} |r_{gi}^{(1)}| \neq 0. \qquad \text{(id. corr. row. test)}$$

The motivation for studying HT problems (b) and (c) lies in the pre-processing stage of omics datasets where the number of variables $p$ (i.g., genes, proteins, methyl sites, etc) is very large, say order of thousands. The statistical analysis of the whole data can involve dealing with $p \times p$ matrices which supposes a challenge for both number of operations and memory space. For instance, conditional dependence structures defined by the inverse of the covariance (or correlation) matrices are widely used in genomic data to find important gene associations in a biological process but the number of genes is usually reduced by some filtering process to speed up the estimation process. In this regard, the proposed correlation sub-matrices based tests could be employed to select only highly correlated or highly differentially correlated genes.

The methodology we develop in this chapter is motivated by genomic data sets that contain, for the same patient, the gene expression information in two different samples corresponding to two different medical conditions. For instance, we use a first dataset that contains the gene expression information of 82 patients with two samples (tissues) for each gene/patient: the expression in a psoriasis vulgaris lesional tissue and the expression in its adjacent non-lesional tissue. We also analyze a second dataset that measures the gene expression of 60 patients with lung cancer for a paired tumor and healthy tissues. In total, there are more than $p = 19{,}000$ genes for each dataset. They are both publicly available in the Gene Expression Omnibus (GEO) database (Edgar et al., 2002) with accession numbers GSE30999 (psoriasis) and GSE19804 (lung cancer).

Even though the complete $p \times p$ correlation matrix is expected to change considerably between

the two classes of observations, testing the equality of correlation for subgroups of genes (of the 19,000) that are known to have functions in a biological process is highly important. We test if the genes interact similarly in the two conditions for 1,320 pathways which describe genes that are known to interact in the same biological process. Using the same gene sets, we further perform the HT of identity correlation matrix on tumor (or lesional) samples to screen the pathways whose genes highly interact between each other. We finally use HT approaches on the correlation matrix rows to test if each of the 19,000 measured genes is related to all the rest of the genes similarly in the two conditions, as well as to find the most correlated genes in tumor (or lesional) samples.

The chapter is structured as follows. In Section 4.2 we explore the hypothesis testing problem of equality of correlation matrices and in Section 4.3 we derive approximate null distributions for the proposed test statistics. Section 4.4 is concerned with other HT problems including identity correlation matrix testing and correlation matrix rows testing. We only provide expressions for the asymptotic power of the tests in the equality of correlation matrices problem. Nevertheless, the analogous expressions for the other described HT problems could then be deduced. In Section 4.5 we use the methodology in simulations in order to assess the accuracy of the proposed tests under the null hypothesis and to compare their power for different characteristics under the alternative hypothesis. Finally, in Section 4.6 we present real data applications where the proposed methodology is used to answer questions that arise from a biological process. All testing methods discussed in this chapter are implemented within the R package **ldstatsHD** (which is presented in Chapter 7).

## 4.2 Hypothesis testing for equal correlation matrices in paired high-dimensional data

### 4.2.1 Mathematical model and biological setting

Consider $n$ independent and identically distributed (i.i.d.) $2p$-dimensional random vectors $Y_k = (Y_k^{(1)}, Y_k^{(2)})$, $k = 1, \ldots, n$, where $Y^{(1)}$ and $Y^{(2)}$ are associated with population I and population II, respectively, and that follow a standard multivariate normal distribution with correlation $R$, i.e.,

$$(Y_k^{(1)}, Y_k^{(2)}) \overset{iid}{\sim} N_{2p}(0, R), \qquad R = [r_{ij}] = \begin{bmatrix} R_1 & R_{12} \\ R_{12}^\top & R_2 \end{bmatrix}, \tag{4.1}$$

where $R_1$ and $R_2$ are the category-specific correlation matrices and the cross-correlation $R_{12}$ is non-zero if the two random vectors $Y^{(1)}$ and $Y^{(2)}$ are linearly dependent. We assume, without loss of generality, unit variances and zero mean vector. The main goal of this section is to test whether the correlation matrix $R_1$ is equal to the correlation matrix $R_2$ with hypothesis $H_0 : R_1 = R_2$ vs $H_1 : R_1 \neq R_2$.

This paired model is related to the following biological setting: the gene expression is measured for the same subject under two conditions or in two different tissues such as healthy and tumor. Different

specifications and biological interpretations of the dependence structure $R$ (or its inverse matrix $\Omega$) are described in Section 2.2 and they all could be considered to apply the methodology proposed in this section. Under $H_0$, the additive and multiplicative models, see equations 2.7 and 2.9, coincide with $R_{12} = R_1\Delta$ in both such cases. In the direct effect model, the model specification is done in the conditional dependence structure, where the cross-joint precision matrix is assumed to be diagonal. Following notation from Section 2.2, under $H_0$ we consider $\Omega_1^J = \Omega_2^J$ and $\Omega_{12}^J$ being a diagonal matrix, so $R_1 = R_2 = (\Omega_1^J)^{-1}(I - A^2)^{-1}$ with $A = (\Omega_1^J)^{-1}\Omega_{12}^J$ and $R_{12} = R_1 A$.

### 4.2.2 Fisher transformation of sample correlations

We denote the sample correlation matrix by $\hat{R}$, which is determined by $\hat{R}_1 = [\hat{r}_{ij}^{(1)}] = Y^{(1)\top}Y^{(1)}/n$, $\hat{R}_2 = [\hat{r}_{ij}^{(2)}] = Y^{(2)\top}Y^{(2)}/n$ and $\hat{R}_{12} = [\hat{r}_{ij}^{(12)}] = Y^{(1)\top}Y^{(2)}/n$. Given the symmetry in the correlation matrices, we consider their lower triangular matrices instead using the same notation with

$$M = \{(i,j) \in \{1,\ldots,p\} : i < j\}, \quad m = \text{Card}(M) = p(p-1)/2. \tag{4.2}$$

An approximate pivot for the correlation coefficient is given by the Fisher transformation (Fisher, 1921), which is defined by $g : (-1,1) \mapsto \mathbb{R}$, $g(z) = \log\{(1+z)/(1-z)\}/2$, such that the elementwise Fisher transformation of $\hat{R}_K$, $K \in \{1,2\}$, weakly converges to a multivariate normal distribution

$$\hat{U}_K = g(\hat{R}_K)\sqrt{n-3} \sim N(g(R_K)\sqrt{n-3}, \Psi_K), \quad K \in \{1,2\}, \tag{4.3}$$

where $\Psi_K = [\psi_{th}^{(k)}]$ is the $m \times m$ correlation matrix between elements in $\hat{U}_K$ as $\psi_{tt}^{(k)} = 1$ for any $t \in M$ and $K \in \{1,2\}$.

### 4.2.3 Correlation of sample correlation coefficients

We assume here and throughout that $r_t < 1$ for any $t \in M$. The non-zero dependence structure between the two random vectors $Y^{(1)}$ and $Y^{(2)}$ leads to correlation between elements in the estimator $\hat{U} = [\hat{U}_1, \hat{U}_2]$ (Elston, 1975; Steiger, 1980), which is found as in eq. (4.3). Take $s = (h,i)$ and $t = (j,l)$, $s,t \in M$, as defined in eq. (4.2), following derivations from Dunn and Clark (1969), the asymptotic correlation of $\hat{u}_s$ and $\hat{u}_t$, $\psi_{st} = \psi_{hi,jl} = \text{cor}(\hat{u}_s, \hat{u}_t)$, as $n \to \infty$, is expressed by

$$\psi_{st} = \psi_{hi,jl} = (\omega_{hh|l}\,\omega_{jj|l})^{-1}[(\omega_{hj|i}\,\omega_{il|j} + \omega_{hj|l}\,\omega_{il|h}) + (\omega_{hl|i}\,\omega_{ij|l} + \omega_{hl|j}\,\omega_{ij|h})]/2, \tag{4.4}$$

where $\omega_{hi|j} = r_{hi} - r_{hj}r_{ij}$ and $\omega_{hh|l} = 1 - r_{hl}^2$.

The difference of Fisher transformed coefficients also approximately follows a normal distribution $\Delta\hat{U} := (\hat{U}_2 - \hat{U}_1) \sim N(U_2 - U_1, \Psi_1 + \Psi_2 - 2\Psi_{12})$ where $\Psi_{12}$ describes the correlation between coefficients in $\hat{U}_1$ and $\hat{U}_2$. The diagonal elements $(\psi_{tt}^{(12)})$, $t \in M$, are estimated by plugging-in the sample correlation coefficients in eq. (4.4). This yields a consistent estimator of $(\psi_{tt}^{(12)})$ for large $n$ but produces

non-negligible bias in the estimation for small $n$. Let $\hat{d}_t$ be the standardized expression of $\Delta\hat{u}_t$, such that

$$\hat{d}_t = \Delta\hat{u}_t\{2(1-\hat{\psi}_{tt}^{(12)})\}^{-1/2}, \quad t \in M, \quad \hat{D} = (\hat{d}_t). \tag{4.5}$$

Under the null hypothesis of equality in the correlation matrices, $\hat{d}_t$ has zero expected value and variance $(\sigma_t^2)_n$ with $(\sigma_t^2)_n \to 1$, $n \to \infty$ for any $t \in M$. Moreover, if $\psi_{tt}^{(12)}$ is known, then $\text{cov}(\hat{d}_t, \hat{d}_k)$ is proportional to $\psi_{tk}^{(1)} + \psi_{tk}^{(2)} - 2\psi_{tk}^{(12)}$, which is non-zero for some $k \neq t$, unless $R = I$.

### 4.2.4 Proposed test statistics

The three test statistics considered here are based on the elementwise standardized differences between transformed sample correlation coefficients in eq. (4.5). These are average of squares ($T_S$), extreme value ($T_M$) and sum of exceedances ($T_E$) test statistics

$$T_S = m^{-1}\sum_{t\in M}\hat{d}_t^2, \tag{4.6}$$

$$T_M = \max_{t\in M}|\hat{d}_t|, \tag{4.7}$$

$$T_E^w(u) = \sum_{t\in M}(|\hat{d}_t| - uw)^2 I(|\hat{d}_t| > u). \tag{4.8}$$

In the sum of exceedances test, $w$ is either 0 or 1 and it is incorporated to weight the importance of high values over the threshold $u$.

## 4.3 Null distributions and asymptotic power

### 4.3.1 Average of squares test

The following lemma provides expressions for the expected value and variance of the average of squares test statistic $T_S$, which is defined in eq. (4.6).

**Lemma 4.1** (*Expected value and variance of $T_S$*)**.** *Let $\mu_2 = \mathbf{E}(\hat{d}_t^2)$ and $\mu_4 = \mathbf{E}(\hat{d}_t^4)$. Define $\bar{\gamma}_2 = 2(m^2-m)^{-1}\sum_{t<h} cov(\hat{d}_t^2, \hat{d}_h^2)$. The expected value and variance of $T_S$ are expressed by*

$$\mathbf{E}[T_S] = \mu_2; \quad var(T_S) = (\mu_4 - \mu_2^2)/m + (1-1/m)\bar{\gamma}_2. \tag{4.9}$$

**Proof.** *The proof of Lemma 1 can be found in Appendix A.1.*

Under $H_0$, asymptotically with $n \to \infty$, $\hat{d}_t^2 \sim \chi_1^2$, for any $t \in M$. Besides, for sufficiently large $n$, it follows from the properties of $\chi_1^2$ that $\mu_2 \doteq 1$ and $\mu_4 \doteq 3$. Let $v = \sum_{t<h} I[\text{cov}(\hat{d}_t^2, \hat{d}_h^2) \neq 0]$ be an integer ranging in $[0, m(m-1)/2]$. If $\text{cov}(\hat{d}_t^2, \hat{d}_h^2) \leq k$, for any $t < h$, for a finite constant $k$, and $v/m \to 0$ as $m \to \infty$, then it follows that $\text{var}(T_S) = (2/m)(1 + O(v/m))$.

However, for a finite dimension, if the correlation matrices are not highly sparse, $v/m$ is not negligible and the dependence parameter $\bar{\gamma}_2$ must be incorporated to assure uniformity in the p-

values of the test under $H_0$. Moreover, since an estimator for the covariance between Fisher transform sample correlations $\psi_{tt}^{(12)}$ (defined in eq. (4.3)) is used, parameters $\mu_2$ and $\mu_4$ can differ slightly from their limiting values (1 and 3) and should be estimated. For sufficiently large $m$ and $n$, $T_S$ is well approximated by a normal distribution with parameters $\mu = \mu_2$ and $\sigma^2 = (\mu_4 - \mu_2^2)/m + (1 - 1/m)\bar{\gamma}_2$ with $\Pr(T_S \leq x \mid H_0) \doteq \Phi(x; \mu, \sigma^2)$ where $\Phi(\cdot; \mu, \sigma^2)$ is the CDF of normal distribution with parameters $\mu$ and $\sigma^2$. Following the central limit theorem, the Gaussian approximation can be appropriate even when $n$ if parameters $\mu_2$ and $\mu_4$ are well specified (not approximated by their limiting values).

Hence, the null hypothesis is rejected at significance level $\alpha$ if the observed value of $T_S$ is greater than

$$t_{S,\alpha} \doteq \mu_2 + z_\alpha \sqrt{(\mu_4 - \mu_2^2)/m + (1 - 1/m)\bar{\gamma}_2}. \tag{4.10}$$

The following theorem gives a lower bound for the power of the average of squares test.

**Theorem 4.1** (*Power of the average of squares test*)**.** *Let $t_{S,\alpha}$ be the asymptotic $\alpha$-quantile of the distribution for $T_S$ under $H_0$ defined by* (4.10) *with $0 < \alpha < 1/2$. Under the alternative hypothesis, let $\bar{\gamma}_2' = 2(m^2 - m)^{-1} \sum_{t<h} cov(\hat{d}_t^2, \hat{d}_h^2 \mid H_1)$ and $\delta_t = |g(r_t^{(2)}) - g(r_t^{(1)})|$ with $\mathscr{S}_d = \{t \in M : \delta_t \neq 0\}$. Denote $\delta_0^2 = \sum_{t \in \mathscr{S}_d} \delta_t^2$. If condition*

$$\delta_0^2 > z_\alpha \sqrt{2m} \{1 + (m-1)\bar{\gamma}_2/2)\}^{1/2}/(n-3) \tag{4.11}$$

*holds, then, as $n, m \to \infty$,*

$$\Pr(T_S \geq t_{S,\alpha} \mid H_1) \geq 1 - \exp\left[ -\frac{1}{2} \left\{ \frac{\frac{(n-3)}{m}\delta_0^2 - z_\alpha [\frac{2}{m}\{1 + (m-1)\bar{\gamma}_2/2\}]^{1/2}}{(m^{-1/2}\{2 + \frac{4s(n-3)}{m}\delta_0^2 + (m-1)\bar{\gamma}_2'\}^{1/2})} \right\}^2 \right] (1 + o(1)).$$

**Corollary 4.1.** *For $\bar{\gamma}_2 < \nu\, k$ and $\nu/m = o(1)$, condition* (4.11) *becomes $\delta_0^2 \gtrsim \frac{m^{1/2}}{n}$ as $(n, m) \to \infty$. Under condition* (4.11)*, when $nm^{-1/2}\delta_0^2 \to \infty$, $\Pr(T_S \geq t_{S,\alpha} \mid H_1) \to 1$.*

**Proof.** *The proof of Theorem 1 can be found in Appendix A.7.1.*

### 4.3.2 Extreme value test

In this section we provide a heuristic approach to approximating the limiting distribution of $T_M$, defined in eq. (4.7), based on two key assumptions: (*i*) we suppose that the sample size $n$ is sufficiently large so that $(\hat{d}_t : t \in M)$ has a Gaussian distribution with standard $N(0,1)$ margins and (*ii*) we assume

$$\max_{t<s\in M} |\mathrm{cov}(\hat{d}_t, \hat{d}_s)| < 1 \quad \text{and} \quad \nu_t = \sum_{s \in M \setminus t} I\{\mathrm{cov}(\hat{d}_t, \hat{d}_s) \neq 0\} = O(m^{\eta_t}), \tag{4.12}$$

for some $\eta_t \in (0,1)$, $t \in M$. Condition (4.12) implies that no two elements of $(\hat{d}_t)$ are perfectly dependent and that there is sufficiently weak dependence structure in the process. If condition (4.12) holds, then adapted versions of extreme value limits for non-stationary Gaussian processes apply

(Leadbetter et al., 1983), i.e., there exist location and scale functions $\mu(m) \in \mathbb{R}$ and $\sigma(m) > 0$, such that

$$\lim_{m\to\infty} \Pr\left( \frac{T_M - \mu(m)}{\sigma(m)} < x \,\middle|\, H_0 \right) = \exp\left\{-\exp(-x)\right\}, \tag{4.13}$$

describes a Gumbel distribution with $\mu(m) + \sigma(m)\, x \to \infty$, as $m \to \infty$, for all $x$. We note that a similar type of extreme value limits are obtained in Cai et al. (2013) for the less general setting where $(Y_k^{(1)}, Y_k^{(2)})$ in expression (4.1) are independent. Additionally, our empirical findings from simulations confirm that this is a reasonable approximation for the distribution of $T_M$ provided $n$ and $m$ are sufficiently large. To back up this result, we illustrate in Appendix A.3 how condition (4.12) links with Leadbetter et al. (1983) conditions for convergence of the maximum of non-stationary Gaussian processes.

In real applications, where $m$ is finite, limit expression (4.13) may fail to approximate the distribution of $T_M$ in two respects. Firstly, it is known that the rate of convergence to the limit distribution is very slow. Secondly, its form is independent of the dependence structure of the process $(\hat{d}_t : t \in M)$, a result that stems from the joint tail properties of the multivariate Gaussian distribution (Sibuya, 1959; Tiago de Oliveira, 1962).

An improved approximation that does take into account the dependence characteristics can be obtained from a sub-asymptotic correction (Eastoe and Tawn, 2012),

$$\Pr\left( \frac{T_M - \mu(m)}{\sigma(m)} < x \,\middle|\, H_0 \right) \doteq \exp\left\{ -\left(\frac{m_E}{m}\right) \exp(-x) \right\}, \quad \text{for large } m, \tag{4.14}$$

where $m_E = m_E(m, x)$ satisfies $m_E/m \to 1$, as $m \to \infty$, for all $x \in \mathbb{R}$, and describes the effective sample size of independent and identically distributed $N(0,1)$ random variables whose maximum has the same distribution with $T_M$. Note that the distribution of $T_M$ in eq. (4.14) is a Gumbel distribution as in eq. (4.13) but with an updated location parameter, say $\mu_{m_E}(m)$, which depends on $m_E$.

The null hypothesis is rejected at significance level $\alpha$ if the observed value of $T_M$ is greater than

$$\begin{aligned} t_{M,\alpha} &\doteq \mu_{m_E}(m) - \sigma(m)\log(-\log(\alpha)) \\ &\sim \{2\log(2m)\}^{1/2} - [\log\theta_m + \log\{-\log(\alpha)\}]/\{2\log(2m)\}^{1/2}. \end{aligned} \tag{4.15}$$

The following theorem gives a lower bound for the power of the extreme value test

**Theorem 4.2** (*Power of the extreme value test*)**.** *Assume* (4.12) *holds. Let* $t_{M,\alpha}$ *be the asymptotic* $\alpha$*-quantile of the distribution for* $T_M$ *under* $H_0$ *defined by* (4.15) *with* $0 < \alpha < 1/2$*. Under the alternative hypothesis, let* $\delta_t = |g(r_t^{(2)}) - g(r_t^{(1)})|$ *with* $\mathscr{S}_d = \{t \in M : \delta_t \neq 0\}$*. If the following condition holds*

$$\max_{t\in\mathscr{S}_d} \delta_t > \frac{1}{\sqrt{n-3}} \left[ \sqrt{2\log(2m)} - \frac{\log\{-\log(\alpha)\}}{\sqrt{2\log(2m)}} \right], \tag{4.16}$$

*then, as* $n, m \to \infty$,

$$\Pr(T_M \geq t_{M,\alpha} \mid H_1) \geq 1 - \exp\left[ -\frac{(n-3)}{2} \left\{ \max_{t \in \mathscr{S}_d} \delta_t - \sqrt{\frac{2\log(2m)}{(n-3)}} \right\}^2 \right] (1 + o(1)).$$

*If* $s = |\mathscr{S}_d| \to \infty$ *and*

$$\min_{t \in \mathscr{S}_d} \delta_t > \frac{1}{\sqrt{n-3}} \left[ \sqrt{2\log(2m)} - \frac{\log\{-\log(\alpha)\}}{\sqrt{2\log(2m)}} \right], \tag{4.17}$$

*then, as* $n, m \to \infty$,

$$\Pr(T_M \geq t_{M,\alpha} \mid H_1) \geq 1 - \exp\left\{ -e^{-\sqrt{2(n-3)\log(2s)} \left[ \min_{t \in \mathscr{S}_d} \delta_t - \sqrt{\frac{2\log(2m)}{(n-3)}} \right]} \right\} (1 + o(1)).$$

**Corollary 4.2.** *As* $n, m \to \infty$, *condition* (4.16) *becomes* $\max_{t \in \mathscr{S}_d} \delta_t^2 \gtrsim (2\log 2m)/(n-3)$. *Under this condition, if* $n^{1/2}(\max_{t \in \mathscr{S}_d} \delta_t - \sqrt{2\log(2m)/n}) \to \infty$, $\Pr(T_M \geq t_{M,\alpha} \mid H_1) \to 1$.

*Similarly, as* $n, m \to \infty$ *and* $s = |\mathscr{S}_d| \to \infty$, *condition* (4.17) *becomes* $\min_{t \in \mathscr{S}_d} \delta_t^2 \gtrsim (2\log 2m)/(n-3)$. *Under this condition, if* $\sqrt{n \log s}\{\min_{t \in \mathscr{S}_d} \delta_t - \sqrt{2\log(2m)/n}\} \to \infty$, $\Pr(T_M \geq t_{M,\alpha} \mid H_1) \to 1$.

**Proof.** *The proof of Theorem 2 can be found in Appendix A.7.2.*

*Extremal index to measure dependence on the sequence*

Expression (4.14) has similarities with a problem studied in the context of stationary time series. Define the stationary sequence $\{Z_t\}_{t=0}^m$ and let $m_C$ determine the length of independent clusters of exceedances with $m_C = o(m)$. Following eq. (4.13), take $x(m) = \mu(m) + \sigma(m)\, x$, under mild conditions, the quantity

$$\begin{aligned} \theta = \lim_{m \to \infty} \theta_m &= \lim_{m \to \infty} \Pr\left\{ Z_1 < x(m), \ldots, Z_{m_C} < x(m) \mid Z_0 > x(m) \right\} \\ &= \lim_{m \to \infty} (m_E / m), \end{aligned} \tag{4.18}$$

is known as the extremal index (O'Brien, 1987) and describes the reciprocal of the expected cluster size of exceedances above large thresholds. For sub-asymptotic models (Eastoe and Tawn, 2012), $\theta_m$ is interpreted as the exceedance probability of $m_C$ consecutive time points just after an exceedance above a high threshold is observed. In a non-stationary process, a cluster-based structure can still be present, but independent clusters may take different sizes. This is studied in Aldous (1989), who proposes an heuristic approach that considers a compound Poisson process with non time homogeneous intensity to represent the extremum of non stationary processes.

In our non-stationary process though, the elements in $(\hat{d}_t : t \in M)$ are not ordered and the general interpretation for $\theta_m$ does not apply. However, we have found empirical evidence that using eq. (4.14) can still improve the representation of $T_M$ under $H_0$. Besides, in Appendix A.4 we use a similar approach as Aldous (1989) to study the form of $\theta_m$ when the correlation matrices $R_1$ and $R_2$ that

generate the data (see eq. (4.1)) are block diagonal.

### 4.3.3 Sum of exceedances test

Let $\mathscr{S}_u = \{t \in M : |\hat{d}_t| \geq u\}$ be the set of exceedances above some threshold $u \geq 0$, let $N_u = \text{Card}(\mathscr{S}_u)$ be the number of elements in $\mathscr{S}_u$ and recall that $m = p(p-1)/2$. The cumulative distribution function of the test statistic $T_E$ under $H_0$ is

$$\Pr(T_E^w(u) < x \mid H_0) = \sum_{k=1}^{m} \left[\Pr(N_u = k \mid H_0)\ \Pr(T_E^w(u) < x \mid H_0, N_u = k)\right]. \tag{4.19}$$

We define several parameters that are used to determine the limiting distribution of $T_E$:

$$\begin{aligned} \gamma_{u_{tj}}^{(w)} &= \text{cov}((|\hat{d}_t| - uw)^2, (|\hat{d}_j| - uw)^2 \mid \hat{d}_t^2 > u, \hat{d}_j^2 > u, d_t = d_j = 0), \\ \eta_0 &= \Pr(|\hat{d}_t| > u \mid d_t = 0), \\ \phi_{tj} &= \Pr(\hat{d}_t^2 > u^2, \hat{d}_j^2 > u^2 \mid d_t = d_j = 0),\ \bar{\phi} = [m(m-1)]^{-1} \sum_{t \neq j} \phi_{tj}. \end{aligned} \tag{4.20}$$

Let $\varphi$ and $\Phi$ be the density and cumulative distribution function of the standard normal distribution, respectively. For sufficiently large expected number of exceedances, the central limit theorem yields $\Pr(T_E^w(u) < x \mid H_0) \doteq \Phi\{x, \mu(m,w), \sigma^2(m,w)\}$ for any $w = \{0, 1\}$, with

$$\begin{cases} \mu(m,w) = m\,\eta_0\,\mu_w \\ \sigma^2(m,w) = m\,\{\eta_0\sigma_w^2 + \mu_w^2(\eta_0 - \bar{\phi})\} + m^2\mu_w^2(\bar{\phi} - \eta_0^2) + \sum_{t \neq j} \gamma_{u_{tj}}^{(w)} \phi_{tj}, \end{cases} \tag{4.21}$$

where for $w = 0$ $\mu_w$ and $\sigma_w^2$ are defined by

$$\begin{cases} \mu_0 = 1 + u\,\varphi(u)/\{1 - \Phi(u)\} \\ \sigma_0^2 = 3 + (u^3 + 3u)\,\varphi(u)/\{1 - \Phi(u)\} - \mu_0^2, \end{cases} \tag{4.22}$$

whereas for $w = 1$ these are

$$\begin{cases} \mu_1 = u^2 + 1 - u\,\varphi(u)/\{1 - \Phi(u)\} \\ \sigma_1^2 = 3 + u^4 + 6u^2 - (5u + u^3)\,\varphi(u)/\{1 - \Phi(u)\} - \mu_1^2. \end{cases} \tag{4.23}$$

The derivation of equations (4.21), (4.22) and (4.23) can be found in Appendix A.2. Note that if the elements in $\hat{D}$ are near independence, then $\bar{\phi} \approx \eta_0^2$, making the third term in the expression for the variance in eq. (4.21) approximately zero, and the whole expression simplifies to $\sigma^2(m,w) \doteq m\,\eta_0\{(1-\eta_0)\,\mu_w^2 + \sigma_w^2\}$. Furthermore, in Appendix A.5 we propose a saddle point approximation for the null distribution of $T_E^{(w)}$ to relax the Gaussian assumption when $m$ is not sufficiently large.

The null hypothesis is rejected at significance level $\alpha$ if the observed value of $T_E^{(w)}$ is greater than

$$t_{E,\alpha}^{(w)} \doteq \mu(m,w) + z_\alpha \sigma(m,w). \tag{4.24}$$

The following theorem shows a lower bound for the power of the sum of exceedances test.

**Theorem 4.3** (*Power of the sum of exceedances test*)**.** *Let $t_{E,\alpha}^{(w)}$ be the asymptotic $\alpha$-quantile of the distribution for $T_E^{(w)}$ under $H_0$ defined by* (4.24) *with $0 < \alpha < 1/2$ and $w$ being either 0 or 1. Consider $\mu_0$ and $\mu_1$ defined by eq.* (4.22) *and eq.* (4.23)*, $\eta_0$ defined by eq.* (4.20) *and $\sigma^2(m,w)$ defined by eq.* (4.21)*. Under the alternative hypothesis, let $\delta_t = |g(r_t^{(2)}) - g(r_t^{(1)})|$ with $\mathscr{S}_d = \{t \in M : \delta_t \neq 0\}$, $s = |\mathscr{S}_d|$, $\eta_t = \Pr(|\hat{d}_t| > u \mid d_t \neq 0)$ and $\mu_{t_w} = \mathbf{E}((|\hat{d}_t| - wu)^2 \mid |\hat{d}_t| > u, d_t \neq 0)$. If the following condition holds*

$$\sum_{t \in \mathscr{S}_d} \mu_{t_w} \eta_t > s\eta_0 \mu_w - z_\alpha \sigma(m,w), \tag{4.25}$$

*then the lower bound for the asymptotic power of sum of exceedances test, with $w = \{0,1\}$, as $n, m\eta_0 \to \infty$, is*

$$\Pr(T_E^{(w)} \geq t_{E,\alpha}^{(w)} \mid H_1) \geq 1 - \exp\left\{-\frac{1}{2}\left(\frac{\sum_{t\in\mathscr{S}_d} \mu_{t_w}\eta_t - s\eta_0\mu_w - z_\alpha\sigma(m,w)}{\sigma_{H_1}(m,w)}\right)^2\right\}(1+o(1)),$$

*where $\sigma^2_{H_1}(m,w)$ can be found following eq.* (A.4)*.*

*Note: Gaussian approximation represents the asymptotic power well if and only if $m\eta_0$ is sufficiently large, with $u < \sqrt{2\log 2m}$ being a necessary condition.*

**Corollary 4.3.** *Assume $\sigma^2(m,w) \doteq m\eta_0\{(1-\eta_0)\mu_w^2 + \sigma_w^2\}$. Let $u = u(\beta)$ with $\beta = 2(1-\Phi(u))$, and let $\mathscr{S}_{du} = \{t \in M, |d_t| \gg u\}$ with $s_u = |\mathscr{S}_{du}|$. When $(m,n,u) \to \infty$, under condition* (4.25)*, if $s_u = k\max(1, s\eta_0, (2m\eta_0)^{1/2})$ for some integer $k > 0$, and $\delta_t^2(n/u^2) \to \infty$ for some $t \in \mathscr{S}_{du}$, $\Pr(T_E^{(w)} \geq t_{E,\alpha}^{(w)} \mid H_1) \to 1$.*

1. *$u = 0$: recovery conditions coincide with the average of squares test (Section 4.3.1).*
2. *$u = \sqrt{2\log 2m} - o(1)$: recovery conditions are similar to extreme value test (Section 4.3.2).*

**Proof.** *The proof of Theorem 3 can be found in Appendix A.7.3.*

### 4.3.4 Estimation of dependence parameters and non-parametric distributions

Under $H_0$, $Y_1^{(1)}, \ldots, Y_n^{(1)} \sim N(0, R_1)$ and $Y_1^{(2)}, \ldots, Y_n^{(2)} \sim N(0, R_2)$ with $R_1 = R_2$. In case $Y_k^{(1)}$ and $Y_k^{(2)}$ were independent for all $k \in \{1, \ldots, n\}$, the elements in $[Y_1^{(1)}, \ldots, Y_n^{(1)}, Y_1^{(2)}, \ldots, Y_n^{(2)}]$ would be exchangeable (i.e., permutation invariant). For paired datasets, $R_{12} \neq 0$ and standard permutation methods are not suitable. Alternatively, we consider a resampling method which keeps paired observations together: find $[(Z_1^{\pi_1}, \ldots, Z_n^{\pi_n}), (Z_1^{\bar{\pi}_1}, \ldots, Z_n^{\bar{\pi}_n})]$ where $\bar{\pi}_k = 1 - \pi_k$, and $Z_k^{\pi_k} = Y_i^{(1)}$ if $\pi_k = 0$ or $Z_k^{\pi_k} = Y_k^{(2)}$ if $\pi_k = 1$, with $\pi_k \sim \text{Bern}(1/2)$. The permutation process is repeated $B$ times and for each replicate $(i = 1, \ldots, B)$

the difference of Fisher transform correlation matrices, defined in eq. (4.5), is calculated and denoted by $\hat{D}^{(i)}$. Finally, a $B \times m$ matrix $\tilde{D}$ is considered where row $i$ contains the lower triangular matrix of $\hat{D}^{(i)}$.

We denote $\tilde{D}^2$ by the elementwise product of the matrix $\tilde{D}$ and $\tilde{D}^4$ by the elementwise product of the matrix $\tilde{D}^2$. The parameters $\mu_2$, $\mu_4$ and $\bar{\gamma}_2$ for the average of squares test defined in eq. (4.9) are estimated using permuted samples such that

$$\hat{\mu}_2 = \frac{1}{Bm}\sum_{i=1}^{B}\sum_{t=1}^{m}\tilde{D}_{it}^2,\ \hat{\mu}_4 = \frac{1}{Bm}\sum_{i=1}^{B}\sum_{t=1}^{m}\tilde{D}_{it}^4,\ \hat{\bar{\gamma}}_2 = \frac{2}{Bm(m-1)}\sum_{i=1}^{B}\sum_{t<h}\mathrm{cov}(\tilde{D}_{it}^2,\tilde{D}_{ih}^2).$$

Regarding the extreme value test, for each replicate of the permutation process, $i = 1,\ldots,B$, the maximum $\hat{T}_M^{(i)} = \max_{t\in M}|\tilde{D}_{it}|$ is computed so that for sufficiently large sample size $n$, $\hat{T}_M^{(i)}$ can be considered as an independent replicate of a Gumbel distributed random variable with parameters $\mu_{m_E}(m)$ and $\sigma(m)$. The location parameter $\mu_{m_E}(m)$ of the Gumbel distribution is estimated by maximum likelihood. Finally, for the sum of exceedances test, the parameter $\sigma^2(m,w)$ defined in eq. (4.21) is estimated by maximum likelihood using permuted samples such that $\Pr(T_E^w(u) < x \mid H_0) \doteq \Phi\{x,\mu(m,w),\hat{\sigma}^2(m,w)\}$ where the parameter $\mu(m,w)$ is also expressed in eq. (4.21).

A non-parametric null distribution for $T_Q$, $Q \in S,M,E$, based on permuted samples is also considered by recording the value of $B$ test statistics, i.e., $\hat{T}_S^{(i)} = m^{-1}\sum_{t=1}^{m}\tilde{D}_{it}^2$, $T_M^{(i)} = \max_{t\in M}|\tilde{D}_{it}|$ or $\hat{T}_E^{(i)} = \sum_{t\in\mathscr{S}_u}(\tilde{D}_{it} - uw)^2$, for $i = 1,\ldots,B$, with $\Pr(T_Q \le x \mid H_0) \doteq B^{-1}\sum_{i=1}^{B} I(\hat{T}_Q^{(i)} \le x)$.

### 4.3.5 Comparison of the tests

Extreme value test is more powerful when it comes to sparse alternatives whereas the average of squares test is useful when the differential correlation matrix is non-sparse and the magnitude of the coefficients is small. The sum of exceedances test lies in between the other two tests. For threshold $u$ near zero, the test statistic is similar to the average of squares test and for $u \approx \sqrt{2\log m}$ it finds similar power to the extreme value test. In between there are infinitely many possibilities and the optimal value is difficult to find without any prior knowledge. In Appendix A.6 we describe an approach to select the threshold that maximizes the lower bound of the power determined in Theorem 2 by integrating out some of the unknown parameters. Furthermore, the weight $w$ is added to the expression of the sum of exceedances since the underlying test powers are complementary regarding sample sizes and number of non-zero correlation differences. For instance, for $w = 1$ the test is powerful for highly sparse differential correlation matrix and small sample sizes (or small magnitude for the difference coefficients). Otherwise, $w = 0$ achieves the most powerful test of the two. We consider a default value of $w = 0$. The theoretical results obtained in this section are completed empirically using simulated data in Section 4.5.

## 4.4 Other hypothesis testing problems using correlation matrices

### 4.4.1 Testing for equal correlation matrix rows in paired high-dimensional data

Consider the problem setting described in Section 4.2.2 where $n$ i.i.d. $p$-dimensional random vectors $Y_k^{(1)} = (Y_{k1}^{(1)}, \ldots, Y_{kp}^{(1)})$ and $Y_k^{(2)} = (Y_{k1}^{(2)}, \ldots, Y_{kp}^{(2)})$, $k = 1, \ldots, n$, are associated to two different classes and jointly follow a standard multivariate normal distribution with joint correlation matrix $R$ (determined by $R_1$, $R_2$ and $R_{12}$). This section studies the HT problem of equality between the row $g$ in $R_1$ and the same row $g$ in $R_2$ with hypothesis $\mathrm{H}_0 : \sum_{i \neq g} |r_{gi}^{(1)} - r_{gi}^{(2)}| = 0$ vs $\mathrm{H}_1 : \sum_{i \neq g} |r_{gi}^{(1)} - r_{gi}^{(2)}| \neq 0$.

Recall from eq. (4.5) that $\hat{D}$ denotes the matrix of Fisher transform correlation differences. We consider an average of squares and extreme value test statistics

$$T_S(g) = (p-1)^{-1} \sum_{i \neq g} \hat{d}_{gi}^2, \qquad g \in \{1, \ldots, p\}, \tag{4.26}$$

$$T_M(g) = \max_{i \neq g} |\hat{d}_{gi}|, \qquad g \in \{1, \ldots, p\}. \tag{4.27}$$

Non parametric null distributions based on permutations are approximated for both test statistics as described in Algorithm 1.

**Algorithm 1** Null distribution and p-values for the equality of correlation rows test

**procedure** $\mathrm{T}_Q(g)$
  Calculate test statistics $T_Q(g)$ for $Q = \{S, M\}$.
  **for** t in 1:B **do**
    Follow Section 4.3.4 to permute data $Y^{(1)}$ and $Y^{(2)}$ to obtain matrices $Z^{\pi}$ and $Z^{\bar{\pi}}$.
    Find $p$ dimensional vectors $\tilde{R}_1^{(t)}(g) = [n^{-1} \sum_{k=1}^n Z_{kg}^{\pi} Z_k^{\pi}]$ and $\tilde{R}_2^{(t)}(g) = [n^{-1} \sum_{k=1}^n Z_{kg}^{\bar{\pi}} Z_k^{\bar{\pi}}]$.
    Calculate the Fisher transform differences of permuted-data sample correlations by

$$\tilde{d}_{gj}^{(t)} = \{g([\tilde{R}_1^{(t)}(g)]_j) - g([\tilde{R}_2^{(t)}(g)]_j)\}\{(n-3)/(2 - 2\hat{\psi}_{gj}^{(t)})\}^{1/2}, \qquad \text{for all} \quad j \neq g.$$

    Compute the average of squares $\tilde{T}_S^{(t)}(g)$ as defined in eq. (4.26) applied to the elements $\tilde{d}_{gj}^{(t)}$, and compute the extreme value test statistic $\tilde{T}_M^{(t)}(g)$ given in eq. (4.27) using $\tilde{d}_{gj}^{(t)}$.
  For $Q = \{S, M\}$, approximate the p-value of test statistic $T_Q(g)$ by

$$\text{p-val}(g) = \frac{1}{B} \sum_{t=1}^{B} I(T_Q(g) < \tilde{T}_Q^{(t)}(g)). \tag{4.28}$$

### 4.4.2 Testing for identity correlation matrix under a single condition

Consider $n$ i.i.d. $p$-dimensional random vectors $Y_k^{(1)} = (Y_{k1}^{(1)}, \ldots, Y_{kp}^{(1)})$, $k = 1, \ldots, n$, that follow a multivariate normal distribution with correlation $R_1$, i.e., $Y_k^{(1)} \sim N_p(0, R_1)$, assuming, without loss of generality, unit variances for simplicity. This section considers the HT problem that assesses whether the correlation $R_1$ is or is not the identity matrix with hypothesis $\mathrm{H}_0 : R_1 = I$ vs $\mathrm{H}_1 : R_1 \neq I$. Recall from Section 4.2.2 that $\hat{R}_1$ denotes the sample correlation lower-triangular matrix of random vector

$Y^{(1)}$. Besides, denote by $\hat{\zeta}_1$ the $m$-dimensional vector containing the Fisher transformation of sample correlation coefficients vector $\hat{R}_1$ with

$$\hat{\zeta}_1 = g(\hat{R}_1)\sqrt{n-3} \sim N(g(R_1)\sqrt{n-3}, \Psi_1), \quad \hat{\zeta}_1 = [\hat{\zeta}_t^{(1)}]_{t\in M}.$$

Under $H_0$, marginally $\hat{\zeta}_t^{(1)}$, $t \in M$, weakly converges to a standard normal distribution.

The three test statistics proposed are equivalent to the ones for equality of correlation matrices but here are applied to $\hat{\zeta}_1$ instead of the difference coefficients in $\hat{D}$

$$T_S^I = m^{-1} \sum_{t\in M} (\hat{\zeta}_t^{(1)})^2, \tag{4.29}$$

$$T_M^I = \max_{t\in M} |\hat{\zeta}_t^{(1)}|, \tag{4.30}$$

$$T_E^{I,w}(u) = \sum_{t\in M} (|\hat{\zeta}_t^{(1)}| - uw)^2 I(|\hat{\zeta}_t^{(1)}| > u). \tag{4.31}$$

Null distributions and powers can be obtained from results in Section 4.3 replacing $\hat{D} = [\hat{d}_t]$ by $\hat{\zeta}_1 = [\hat{\zeta}_t^{(1)}]$.

### 4.4.3 Testing for identity correlation matrix rows under a single condition

Consider the problem setting described in Section 4.4.2 where $p$-dimensional random vectors $Y_k^{(1)} = (Y_{k1}^{(1)}, \ldots, Y_{kp}^{(1)})$, $k = 1, \ldots, n$, follow a multivariate normal distribution with correlation $R_1$. This section assesses whether a variable is or is not linear independent to all other variables with hypothesis $H_0$ : $\sum_{i\neq g} |r_{gi}^{(1)}| = 0$ vs $H_1 : \sum_{i\neq g} |r_{gi}^{(1)}| \neq 0$. The average of adjusted square sample correlation coefficients (S) and the maximum of the absolute value of sample correlation coefficients (M) are the test statistics employed

$$T_S^I(g) = \left(\frac{n-1}{n-2}\right) \frac{T_{SS}^I(g) - 1}{p-1} - \frac{1}{n-2}, \qquad g \in V = \{1, \ldots, p\}, \tag{4.32}$$

$$T_M^I(g) = \max_{i\neq g} |\hat{r}_{ig}^{(1)}|, \qquad g \in V = \{1, \ldots, p\}, \tag{4.33}$$

where $T_{SS}^I(g) = \sum_{i=1}^p (\hat{r}_{ig}^{(1)})^2$ is the sum of squared sample correlation coefficients of variable $g$.

The sum of squared sample correlations $T_{SS}^I(g)$ is computationally fast to obtain for all $g \in V$ simultaneously when $p \gg n$ using some algebra on the definition of correlation coefficient. Given the standardized matrix $Y^{(1)}$, note that the square sample correlation of $Y^{(1)}$ is proportional to $(Y^{(1)\intercal} Y^{(1)})(Y^{(1)\intercal} Y^{(1)})$ which is the product of two $p \times p$ matrices. The same expression can be found by employing fewer number of operations: (1) find $n \times n$ matrix $A =< Y^{(1)}, Y^{(1)\intercal} >$ (2) find $p \times n$ matrix $B =< Y^{(1)\intercal}, A >$ (c) find $T_{SS}^I(g) = \{p(n-1)^2\}^{-1} B_{g.} Y_{.g}^{(1)}$.

Under the hypothesis of total independence presented in Section 4.4.2, say $H_0 : \{r_{ig}^{(1)} = 0$, for all $i \neq g\}$, for the law of large numbers, $T_S^I(g)$ is well approximated by a normal distribution centered at zero. However, in the HT problem presented in this section, $H_0$ allows cases where $\sum_{i\neq j,(i,j)\in V\setminus g} |r_{ji}^{(1)}| \neq 0$,

so some pairs $\hat{r}_{ig}, \hat{r}_{jg}$, $i \neq j \neq g$, can be correlated. A Monte Carlo based procedure that accounts for this linear dependence structure is proposed to find an empirical null distribution. This is done by replicating (i) and (ii) $B$ times with (i) simulate $n$ i.i.d. observations from a standard normal distribution (which is the marginal distribution of any variable $g \in V$) and (ii) find the sample correlation vector that measures the linear dependence between simulated data and all genes in the original data $Y^{(1)}$. The approximate p-values of the tests are found as described in Algorithm 2. Note that the same null distribution can be used for any $g \in V$. Hence, the Monte Carlo based procedure only needs to be done once to test all variables in the dataset.

**Algorithm 2** Null distribution and p-values for the variables linear independence test

1: **procedure**
2: Calculate test statistics $T_S^I(g)$ and $T_M^I(g)$ for any $g \in V = \{1, \ldots, p\}$.
3: **for** $t$ in 1:$B$ **do**
4: Generate $\{z\}_{k=1}^n$ i.i.d. replicates with $z_k \sim N(0,1)$.
5: Compute $(\tilde{r}_j^1)^{(t)} = n^{-1} \sum_{k=1}^n Y_{kj}^{(1)} z_k$, for all $j \in V$,
6: Find $\tilde{T}_S^{I(t)} = p^{-1}\{1 + \sum_{j=1}^p [(\tilde{r}_j^1)^{(t)}]^2\}$, and apply eq. (4.32) to average of squares $\tilde{T}_S^{I(t)}$ instead of $T_S^I(g)$ to obtain $\tilde{T}_S^{I(t)}$. Similarly, find $\tilde{T}_M^{I(t)} = \max_{j \in V} |(\tilde{r}_j^1)^{(t)}|$.
7: For $Q = \{S, M\}$, approximate the p-value of test statistics by

$$\text{p-val}(g) = \frac{1}{B} \sum_{b=1}^{B} I(T_Q^I(g) < \tilde{T}_Q^{I(t)}). \tag{4.34}$$

## 4.5 Simulation study

We analyze the performance of the proposed methods in simulated data sets. We study different structures for the correlation matrix $R$ directly (Section 4.5.1 and Section 4.5.3) or indirectly by setting different graph structures for the precision matrix $\Omega = R^{-1}$ (Section 4.5.2). In both sections we consider two model specifications to generate the data: (i) under $H_0$ to evaluate the size of the testing methods; (ii) under $H_1$ to compare the power of the testing methods.

### 4.5.1 Independent datasets, dense correlation matrices

We can observe in real data, some groups of highly dependent genes whose underlying correlation matrix is non-sparse. In such a case, we argue that asymptotic independence tests are not reliable under $H_0$ even when the datasets are independent. We show this in simulated data by considering a dense correlation matrix denoted by $\tilde{R}$. This matrix is obtained by the sample correlation matrix of a subset of 50 variables from the real dataset described in Section 4.6. In order to obtain a positive definite matrix, we regularize $\tilde{R}$ by

$$\Sigma = \tilde{R} + I\lambda, \tag{4.35}$$

where $\lambda > 0$. Note that as we increase $\lambda$, off-diagonal elements of the correlation matrix decrease.

Data $Y_k^{(1)} \sim N(0,\Sigma_1)$ and $Y_k^{(2)} \sim N(0,\Sigma_2)$, i.i.d. for all $k = 1,\ldots,n$ are generated using the following specifications for the covariance matrices: (i) under $H_0$, we consider $\Sigma_1 = \Sigma_2 = \Sigma$; (ii) under $H_1$, we consider $\Sigma_1 = \Sigma$ and for $\Sigma_2$, we create a two-block diagonal matrix of sizes 40 and 10 by setting to zero the between-block covariance elements of the matrix $\Sigma$. We refer to this model in the results presented in Sections 4.5.4 and 4.5.5 as model 1, which is applied for $n = 50,100$ and $\lambda = 1/2,1,2,3$.

### 4.5.2 Dependent datasets, sparse correlation matrices

We generate data using joint models following notation introduced in Section 2.2. Sparse correlation matrices are obtained by setting almost-block diagonal precision matrices, where each block has a power-law underlying graph structure (see description in Section 5.5.1) and some extra random connections between blocks. Let $A$ be the adjacency matrix with the non-zeros of the precision matrix, the coefficients of the precision matrix are simulated by

$$\Omega^{(0)} = [\omega_{ij}^{(0)}], \qquad \omega_{ij}^{(0)} = \begin{cases} \text{Unif}(0.5,0.9) & \text{if } A_{ij} = 1 \text{ with probability } 0.5\ ; \\ \text{Unif}(-0.5,-0.9) & \text{if } A_{ij} = 1 \text{ with probability } 0.5\ ; \\ 0 & \text{if } A_{ij} = 0. \end{cases} \tag{4.36}$$

Data $(Y_k^{(1)}, Y_k^{(2)}) \sim N(0,\Omega^{-1})$, i.i.d. for all $k = 1,\ldots,n$ are generated using a direct effect model (see definition in Section 2.2) with the following specifications for the joint precision matrix $\Omega$: (i) under $H_0$, $\Omega$ is determined by $\Omega_1^J = \Omega^{(0)}$, $\Omega_2^J = \Omega^{(0)}$ and $\Omega_{12}^J$ being a diagonal matrix with $(\Omega_{12}^J)_{ii} = 0.6$ for $\lfloor p/2 \rfloor$ diagonal elements and $(\Omega_{12}^J)_{ii} = 0$ for the other $\lceil p/2 \rceil$; under $H_1$, let $D_1$ and $D_2$ be two different precision matrices which are generated with the same model as for $\Omega^{(0)}$. We consider $\Omega_1^J = \text{diag}(\Omega^{(0)}, D_1, I)$, $\Omega_2^J = \text{diag}(\Omega^{(0)}, I, D_2)$ and the same specification for $\Omega_{12}^J$ given under $H_0$. In both setting, to obtain a positive definite matrix, we regularize $\Omega$ by $\Omega = \Omega + \lambda I$, with $\lambda$ such that the condition number of $\Omega$ is less than the number of nodes (Cai and Liu, 2011). We refer to this model in the results presented in Sections 4.5.4 and 4.5.5 as model 2, which is applied for $p = 70,120,210$ and $n = 25,50,100,200.$ .

### 4.5.3 Almost identity correlation matrices

We use a toy example to show the behavior of the linear independence tests (both identity correlation matrix and row) when data are generated by a multivariate normal distribution with zero mean vector and correlation matrix

$$R_1 = \begin{pmatrix} 1 & \rho_{12} & 0 & \cdots & 0 \\ & 1 & \rho_{23} & \ddots & 0 \\ & & 1 & \ddots & \vdots \\ & & & \ddots & \rho_{(p-1)p} \\ & & & & 1 \end{pmatrix}$$

We consider different sample sizes $n = 25, 50, 100, 200$ and dimensions $p = 70, 120, 210$. Besides, the coefficient $\rho_{ij}$, for any $j = i + 1$, is fixed to either 0, under $H_0$, or 0.3 under $H_1$. We refer to this model in the results presented in Section 4.5.5 as model 3.

### 4.5.4 Power and size of the equality of correlation matrices test

We consider the average of squares test -S-, the extreme value test -M- and the sum of exceedances test -E- for both $w = 0$ and $w = 1$ with threshold selected as described in Section 4.3.3. We compute the empirical power of the tests to estimate Pr(Reject $H_0 \mid H_1$ true) as well as the test size to estimate Pr(Reject $H_0 \mid H_0$ true) using significance level of $\alpha = 0.05$. We approximate asymptotic null distributions by assuming linear independence between elements in $\hat{D}$ (denoted by AI) or by estimating the dependence parameters using permuted samples (AD). We further approximate non-parametric null distribution (NP) as described in Section 4.3.4. For $w = 1$ we only show the power of the non-parametric null distribution which is labeled by E(NP)$^{(1)}$. Nevertheless, test sizes when $w = 1$ are seen to be similar to the ones provided when $w = 0$.

In Table 4.1 we present the empirical approximations of power and size for the dense correlation matrices scenario (model 1). Generally, tests show a good trade off between false rejection and true rejection rates. For low regularization $\lambda$, as defined in (4.35), asymptotic linear independence tests are not suitable with empirical sizes being larger than the expected 0.05. The average of squares test is the one that dominates the power in this model for $\lambda \geq 2$ and gives similar powers to the sum of exceedances test (with $w = 0$) for $\lambda < 2$. Sum of exceedances test with $w = 1$ achieves worse powers than the test with $w = 0$ for large $\lambda$.

In Table 4.2 we show a similar analysis for dependent datasets with sparse correlation matrices (model 2). Null distributions accounting for dependence (AD and NP) achieve better estimates of the size than asymptotic linear independence tests. Particularly, in the average of squares and sum of exceedances tests adjusting for dependence is desired to obtain a good representation of the null distribution. The asymptotic linear independence extreme value test yields good estimates for the size. It is slightly conservative for large p-values but these do not affect the evidence interpretation. Hence, for sparse dependence structures, the asymptotic extreme value test could be used to speed up the process. The sum of exceedances test with $w = 1$ produces consistently the highest powers among the three tests. Contrarily of what we observe in Table 4.1, the test with $w = 1$ gives better powers than the one with $w = 0$. Moreover, the extreme value test provides higher powers than the average of squares for large sample sizes.

We also analyze the performance of the tests with respect to the proportion of non-zero correlation differences $\rho_s$. In a global analysis, we compute the average power for small proportions ($\rho_s \leq 0.3$) and large proportions ($\rho_s > 0.3$) using the three test statistics. The sum of exceedances test has average powers 0.426 and 0.543 respectively, the extreme value test obtains 0.373 and 0.465, and the average of squares test produces 0.312 and 0.477. Thus, it is $T_S$ that benefits the most from the increase of the

**Table 4.1.** Size, uniformity and power of the equality of correlation matrices test using model 1 -dense correlation matrices- ($\times 10^3$). Test statistics S (average of squares), M (extreme values) and E (exceedances with $w = 0$ or $w = 1$), and null distributions AI (asymptotic independence), AD (asymptotic dependence) and NP (non-parametric) are compared at $\alpha = 0.05$ level.

| | n=50 | | | | n=100 | | | |
|---|---|---|---|---|---|---|---|---|
| $\lambda$ | 0.5 | 1 | 2 | 3 | 0.5 | 1 | 2 | 3 |
| Empirical size | | | | | | | | |
| S(AD) | 62 | 50 | 58 | 53 | 52 | 59 | 60 | 52 |
| M(AD) | 45 | 43 | 49 | 61 | 42 | 48 | 54 | 50 |
| E(AD)$^{(0)}$ | 49 | 54 | 59 | 48 | 52 | 50 | 48 | 54 |
| S(NP) | 61 | 47 | 54 | 52 | 53 | 54 | 57 | 50 |
| M(NP) | 51 | 44 | 47 | 59 | 50 | 50 | 51 | 48 |
| E(NP)$^{(0)}$ | 54 | 50 | 60 | 55 | 46 | 60 | 46 | 58 |
| S(AI) | 306 | 238 | 192 | 133 | 304 | 254 | 192 | 126 |
| M(AI) | 68 | 58 | 59 | 66 | 62 | 54 | 59 | 61 |
| E(AI)$^{(0)}$ | 103 | 126 | 92 | 86 | 200 | 158 | 121 | 88 |
| ks.test p-value to test for uniformity in the correlation test p-values | | | | | | | | |
| S(AD) | 247 | 23 | 716 | 317 | 72 | 400 | 151 | 79 |
| M(AD) | 865 | 121 | 835 | 426 | 147 | 52 | 245 | 646 |
| E(AD)$^{(0)}$ | 51 | 416 | 779 | 211 | 231 | 123 | 532 | 883 |
| S(NP) | 432 | 15 | 134 | 181 | 62 | 432 | 500 | 148 |
| M(NP) | 936 | 69 | 400 | 969 | 181 | 48 | 288 | 181 |
| E(NP)$^{(0)}$ | 288 | 618 | 241 | 500 | 400 | 723 | 648 | 785 |
| S(AI) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M(AI) | 0 | 0 | 24 | 27 | 0 | 0 | 193 | 150 |
| E(AI)$^{(0)}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Empirical power | | | | | | | | |
| S(AD) | 890 | 690 | 342 | 240 | **998** | **992** | 802 | 542 |
| M(AD) | 667 | 270 | 110 | 109 | 996 | 758 | 250 | 122 |
| E(AI)$^{(0)}$ | **950** | 735 | 374 | 202 | **998** | **992** | 790 | 447 |
| S(NP) | 897 | 684 | **380** | **250** | **998** | **992** | **806** | **574** |
| M(NP) | 652 | 280 | 106 | 105 | 996 | 766 | 254 | 118 |
| E(NP)$^{(0)}$ | 943 | **723** | **380** | 223 | **998** | **992** | 787 | 442 |
| E(NP)$^{(1)}$ | 940 | 692 | 304 | 143 | **998** | **992** | 687 | 413 |
| S(AI) | 908 | 588 | 306 | 236 | 998 | 990 | 802 | 450 |
| M(AI) | 702 | 310 | 126 | 072 | 996 | 796 | 248 | 130 |
| E(AI)$^{(0)}$ | 973 | 692 | 251 | 126 | 998 | 972 | 676 | 346 |
| Estimated $\theta$ | | | | | | | | |
| $\hat{\theta}_m$ | .593 | .843 | .915 | .955 | .574 | .828 | .912 | .953 |

number of differential coefficients.

For model 1 (dense difference correlations matrix), the correlation between p-values for the same test statistic using both non-parametric and asymptotic null distributions is very high (around 0.994 in average) whereas the average correlation between extreme value and average of squares p-values is $[0.61, 0.48, 0.36, 0.30]$ in the four regularization parameters used. The p-values for the sum of exceedances test (for both $w$), seem to be more correlated to the p-values for the other two tests with $[0.91, 0.88, 0.82, 0.75]$ against the average of squares and $[0.75, 0.63, 0.55, 0.52]$ against the extreme value. For model 2 (sparse difference correlation matrix), the correlations are smaller with an average of $[0.19, 0.12, 0.07]$ between average of squares and extreme value p-values for the three dimensions used, $[0.55, 0.39, 0.27]$ between average of squares and exceedances and $[0.49, 0.49, 0.48]$ between extreme value and exceedances.

We estimate the extremal index $\theta_m$, which quantifies the dependence structure over high exceedances, and it is defined in Section 4.3.2. In the sparse model 2, the average estimated $\theta_m$ gets close to 1 as the sample size increases. For large $n$, we could assume that $\theta_m$ is equal to 1 and use the

**Table 4.2.** Size, uniformity and power of the equality of correlation matrices test using model 2 -sparse correlation matrices- ($\times 10^3$). Test statistics S (average of squares), M (extreme value) and E (sum of exceedances with $w = 0$ or $w = 1$), and null distributions AI (asymptotic independence), AD (asymptotic dependence) and NP (non-parametric) are compared at $\alpha = 0.05$ level.

| | p=70 | | | | p=120 | | | | p=210 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | 50 | 100 | 200 | 500 | 50 | 100 | 200 | 500 | 50 | 100 | 200 | 500 |
| Empirical size | | | | | | | | | | | | |
| S(AD) | 50 | 50 | 50 | 52 | 49 | 42 | 56 | 52 | 38 | 46 | 48 | 54 |
| M(AD) | 55 | 46 | 51 | 58 | 48 | 54 | 46 | 48 | 48 | 50 | 56 | 44 |
| E(AD)$^{(0)}$ | 50 | 50 | 52 | 51 | 56 | 54 | 56 | 44 | 50 | 43 | 46 | 53 |
| S(NP) | 58 | 54 | 50 | 52 | 55 | 48 | 58 | 50 | 52 | 50 | 50 | 53 |
| M(NP) | 55 | 44 | 51 | 57 | 48 | 54 | 46 | 47 | 47 | 51 | 54 | 44 |
| E(NP)$^{(0)}$ | 47 | 48 | 49 | 49 | 52 | 53 | 54 | 44 | 48 | 46 | 47 | 52 |
| S(AI) | 32 | 58 | 78 | 78 | 22 | 40 | 62 | 69 | 4 | 26 | 44 | 62 |
| M(AI) | 60 | 41 | 47 | 54 | 56 | 57 | 47 | 47 | 62 | 54 | 58 | 42 |
| E(AI)$^{(0)}$ | 56 | 42 | 38 | 66 | 66 | 47 | 46 | 52 | 64 | 52 | 54 | 46 |
| ks.test p-value to test for uniformity in the correlation test p-values | | | | | | | | | | | | |
| S(AD) | 1 | 376 | 37 | 895 | 0 | 929 | 351 | 31 | 0 | 0 | 886 | 286 |
| M(AD) | 58 | 662 | 528 | 266 | 701 | 836 | 917 | 423 | 5 | 837 | 50 | 498 |
| E(AD)$^{(0)}$ | 888 | 58 | 914 | 374 | 155 | 819 | 725 | 349 | 598 | 191 | 85 | 42 |
| S(NP) | 5 | 536 | 29 | 794 | 0 | 648 | 370 | 48 | 0 | 0 | 500 | 164 |
| M(NP) | 87 | 500 | 466 | 341 | 648 | 859 | 936 | 241 | 3 | 723 | 33 | 341 |
| E(NP)$^{(0)}$ | 43 | 536 | 913 | 263 | 43 | 648 | 794 | 466 | 610 | 988 | 241 | 466 |
| S(AI) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| M(AI) | 173 | 255 | 19 | 798 | 513 | 241 | 298 | 78 | 435 | 701 | 19 | 267 |
| E(AI)$^{(0)}$ | 138 | 360 | 10 | 135 | 28 | 207 | 856 | 39 | 0 | 5 | 100 | 42 |
| Empirical power | | | | | | | | | | | | |
| S(AD) | 60 | 144 | 437 | 730 | 78 | 88 | 178 | 398 | 4 | 78 | 152 | 439 |
| M(AD) | 76 | 220 | 715 | 944 | 68 | 76 | 176 | 722 | 42 | 72 | 180 | 651 |
| E(AD)$^{(0)}$ | 101 | 200 | 631 | 910 | 80 | 82 | 170 | 520 | 70 | 74 | 180 | 550 |
| S(NP) | 62 | 150 | 430 | 720 | 96 | **106** | 182 | 404 | **86** | **94** | 160 | 440 |
| M(NP) | 82 | 228 | 706 | 950 | 60 | 58 | 174 | 710 | 44 | 72 | 174 | 649 |
| E(NP)$^{(0)}$ | **102** | 204 | 615 | 960 | 82 | 80 | 180 | 544 | 72 | 76 | 182 | 534 |
| E(NP)$^{(1)}$ | 94 | **316** | **800** | **984** | **102** | 94 | **272** | **816** | 70 | 84 | **232** | **836** |
| M(AI) | 64 | 180 | 458 | 894 | 76 | 76 | 180 | 714 | 56 | 74 | 168 | 632 |
| S(AI) | 68 | 152 | 331 | 630 | 87 | 111 | 162 | 401 | 82 | 91 | 154 | 391 |
| E(AI)$^{(0)}$ | 78 | 261 | 598 | 954 | 106 | 93 | 265 | 819 | 72 | 83 | 214 | 801 |
| Estimated $\theta$ | | | | | | | | | | | | |
| $\hat{\theta}_m$ | .790 | .871 | .908 | .943 | .788 | .848 | .913 | .945 | .770 | .841 | .903 | .937 |

asymptotic approximation which would speed up the results. However, for dense correlations like model 1, $\theta_m$ can be quite small ($\approx 0.6$ for small regularization $\lambda$) and permutations-based tests should be used instead.

### 4.5.5 Power and size of other tests

The HT problems presented in Section 4.4 are also applied to simulated data. We compare the empirical size and power for all proposed test statistics using significance level of $\alpha = 0.05$. For the equality of correlation rows test, we consider model 1 (see Table 4.3) and model 2 (see Table 4.4) to generate the data. The empirical size is close to the expected 0.05 for all scenarios. Moreover, the average of squares test statistic achieves larger power than the maximum test in model 1, whereas the opposite behavior is shown in model 2. This goes in the same direction to the power found for the equality of correlation matrices tests in Section 4.5.4.

Tests based on linear independence are contrasted using model 2 and model 3 (see Table 4.5 and Table 4.6 for identity correlation matrix and identity correlation rows, respectively). Model 2 generates

**Table 4.3.** Size and power of the equality correlation test by rows at $\alpha = 0.05$ level using model 1 -dense correlation matrices- ($\times 10^3$). Test statistics S (average of squares) and M (extreme values) are compared.

| | n=50 | | | | n=100 | | | |
|---|---|---|---|---|---|---|---|---|
| $\lambda$ | 0.5 | 1 | 2 | 3 | 0.5 | 1 | 2 | 3 |
| | | | | Empirical size | | | | |
| S(NP) | 40 | 45 | 57 | 59 | 48 | 55 | 42 | 54 |
| M(NP) | 50 | 41 | 47 | 53 | 52 | 50 | 48 | 45 |
| | | | | Empirical power | | | | |
| S(NP) | 263 | **156** | **77** | **90** | 418 | **293** | **176** | **130** |
| M(NP) | **266** | 122 | 66 | 60 | **506** | 286 | 148 | 91 |

**Table 4.4.** Size and power of the equality correlation test by rows at $\alpha = 0.05$ level using model 2 -sparse correlation matrices- ($\times 10^3$). Test statistics S (average of squares) and M (extreme values) are compared.

| | p=70 | | | | p=120 | | | | p=210 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 25 | 50 | 100 | 200 | 25 | 50 | 100 | 200 | 25 | 50 | 100 | 200 |
| | | | | | | Empirical size | | | | | | |
| S(NP) | 59 | 56 | 42 | 51 | 47 | 56 | 52 | 51 | 52 | 48 | 55 | 42 |
| M(NP) | 51 | 59 | 48 | 52 | 49 | 43 | 48 | 49 | 44 | 46 | 52 | 59 |
| | | | | | | Empirical power | | | | | | |
| S(NP) | **76** | 112 | 157 | 404 | **64** | **83** | 113 | 214 | **59** | **68** | **88** | 93 |
| M(NP) | 69 | **122** | **230** | **622** | 51 | 81 | **146** | **392** | 54 | 47 | 72 | **130** |

sparse correlation matrices but does not achieve the sparsity levels of model 3. The empirical size for non-parametric and dependence-correction tests in Table 4.5 are near the desired 0.05 for all three test statistics. However, asymptotic distributions, especially for maximum and sum of exceedances test, fail to recover the expected size when $n$ is small. This can be due to approximating the Fisher transform sample correlation by a normal distribution, which seems to have problems in the tail of the distribution. Section 4.5.6 studies this particular problem in detail using more simulations. In terms of the power, the maximum does better than the average of squares for highly sparse correlation matrices defined in model 3, but its over-performed by the average of squares and sum of exceedances in the slightly less sparse correlation matrix defined in model 2.

### 4.5.6 Fisher transformation and estimation of correlation of correlations

The Fisher transformation $g(\hat{r}_{ij})$ of a sample correlation coefficient $\hat{r}_{ij}$, for sufficiently large $n$, is established to be well approximated by a normal distribution with expected value $g(r_{ij})$ and variance approximately equal to $n-3$. We use this assumption to propose the null distribution of the asymptotic test statistics to speed up the process, but here we want to determine if this is a reasonable assumption when $n$ is small using simulations.

We generate data from a multivariate normal distribution with zero mean and correlation matrices as defined in model 2 (see Section 4.5.2) with $p = 70$. Initially, since we only want to analyze the utility of the Fisher transformation, we consider independent datasets and $R_1 = R_2$. Moreover, we use several sample sizes $n = 25, 50, 100, 150, 200$. For a generated data set, we estimate the difference between Fisher transform sample correlation matrices, which is denoted by $\hat{D}$ in eq. (4.5), assuming all $\hat{\psi}_{ij} = 0$

**Table 4.5.** Size and power of the identity correlation matrix test using model 2 -sparse correlation matrices- and model 3 - almost identity matrix. Several dimensions $p$ are considered for model 2 whereas several regularization values $\lambda$ are considered for model 3. Test statistics S (average of squares), M (extreme value) and E (sum of exceedances with $w = 0$ or $w = 1$), and null distributions AI (asymptotic independence), AD (asymptotic dependence) and NP (non-parametric) are compared at $\alpha = 0.05$ level.

| | p=70 ($\lambda = 1$) | | | | p=120 ($\lambda = 3$) | | | | p=210 ($\lambda = 5$) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 25 | 50 | 100 | 200 | 25 | 50 | 100 | 200 | 25 | 50 | 100 | 200 |
| Empirical size | | | | | | | | | | | | |
| S(AD) | 56 | 38 | 54 | 46 | 68 | 51 | 55 | 52 | 56 | 50 | 53 | 50 |
| M(AD) | 51 | 44 | 51 | 49 | 38 | 50 | 62 | 50 | 51 | 50 | 50 | 42 |
| E(AD)$^{(0)}$ | 48 | 39 | 54 | 46 | 61 | 50 | 48 | 46 | 60 | 48 | 54 | 48 |
| S(NP) | 52 | 37 | 53 | 43 | 62 | 48 | 49 | 52 | 51 | 50 | 50 | 51 |
| M(NP) | 51 | 42 | 54 | 47 | 41 | 48 | 58 | 49 | 52 | 50 | 50 | 42 |
| E(NP)$^{(0)}$ | 45 | 44 | 52 | 51 | 58 | 44 | 50 | 47 | 58 | 44 | 50 | 48 |
| S(AI) | 57 | 39 | 55 | 46 | 66 | 50 | 58 | 53 | 54 | 50 | 54 | 52 |
| M(AI) | 108 | 75 | 64 | 54 | 136 | 86 | 80 | 59 | 166 | 100 | 68 | 54 |
| E(AI)$^{(0)}$ | 98 | 67 | 65 | 55 | 167 | 88 | 69 | 54 | 277 | 134 | 93 | 62 |
| Empirical power model 2 | | | | | | | | | | | | |
| S(NP) | **444** | **456** | **848** | **994** | **112** | **256** | **412** | **514** | **60** | **70** | 130 | **252** |
| M(NP) | 78 | 116 | 668 | 928 | 64 | 56 | 112 | 180 | 48 | 52 | 54 | 100 |
| E(NP)$^{(0)}$ | 360 | 372 | 826 | **994** | 86 | 182 | 312 | 432 | 50 | 60 | **132** | 178 |
| Empirical power model 3 | | | | | | | | | | | | |
| S(NP) | **300** | 772 | 998 | **1000** | **300** | 742 | 998 | **1000** | **302** | 774 | **1000** | **1000** |
| M(NP) | 108 | 520 | 990 | **1000** | 104 | 428 | 990 | **1000** | 80 | 378 | **1000** | **1000** |
| E(NP)$^{(0)}$ | 292 | **892** | **1000** | **1000** | 274 | **870** | **1000** | **1000** | 298 | **920** | **1000** | **1000** |

(due to having independent random vectors $Y^{(1)}$ and $Y^{(2)}$). We repeat the process 500 times such that we record a $500 \times m$ matrix with i.i.d. replicates of the lower triangular matrix of $\hat{D}$. Then we consider four statistics: (a) the mean of the average of squares by rows; (b) the variance of the average of squares by rows; (c) the mean of the maximum of absolute values by rows; and (d) the variance of the maximum values by rows. In Figure 4.1 we have their representation using 100 instances of the whole process. For first and second order measures such as the mean and variance of the average of squares, see panels (a) and (b), the sample size does not have a big impact on the values of the test statistics. However, the behavior in the tail of the distribution, given here by the maximum is very much dependent on $n$ with decreasing mean and variance (see panels (c) and (d)). For $n$ larger than 100, the mean/variance of the maximum can be quite well approximated by the maximum of a standard normal distribution, which is the marginal null distribution we assume for elements $\hat{d}_t$ in the asymptotic independence test. Moreover, in the variance of $T_S$ we can see the effect of not accounting for the dependence coefficient $\bar{\gamma}_2$ which results in a much larger variance, constant for all $n$, than the expected under a standard normal distribution.

In Section 4.2.3 we define the asymptotic correlation between Fisher transform sample correlation coefficients $[\psi_t]$, and we employ it to standardize the Fisher transform sample correlation differences when the observations of the two datasets are paired. The parameters $[\psi_t]$ have an asymptotic expression which depends on the true correlation coefficients and are estimated employing sample correlation coefficients instead. Below we show using simulations that employing estimated values for $[\psi_t]$ increases the variance of elements $[\hat{d}_t]$ and in consequence the variance of the test statistics.

We consider simulations by model 2 but now with paired observations. We estimate $[\hat{D}]$ using

**Table 4.6.** Size and power of the identity correlation matrix test by rows at $\alpha = 0.05$ level using model 2 -sparse correlation matrices- and model 3 - almost identity matrix. Several dimensions $p$ are considered for model 2 whereas several regularization values $\lambda$ are considered for model 3. Besides, test statistics S (average of squares) and M (extreme values) are compared.

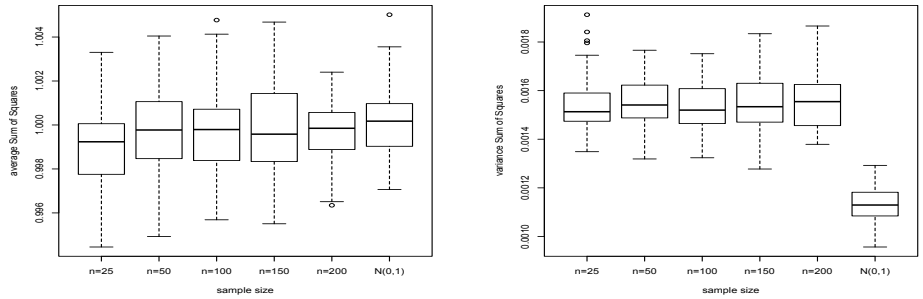
| | p=70 ($\lambda = 1$) | | | | p=120 ($\lambda = 3$) | | | | p=210 ($\lambda = 5$) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 25 | 50 | 100 | 200 | 25 | 50 | 100 | 200 | 25 | 50 | 100 | 200 |
| | | | | | Empirical size model 2 | | | | | | | |
| S(NP) | 58 | 41 | 52 | 51 | 54 | 48 | 51 | 59 | 46 | 45 | 56 | 58 |
| M(NP) | 56 | 39 | 54 | 41 | 47 | 58 | 52 | 48 | 56 | 43 | 48 | 50 |
| | | | | | Empirical size model 3 | | | | | | | |
| S(NP) | 42 | 58 | 49 | 51 | 49 | 52 | 56 | 50 | 58 | 46 | 52 | 57 |
| M(NP) | 41 | 55 | 49 | 46 | 45 | 47 | 47 | 39 | 42 | 41 | 59 | 48 |
| | | | | | Empirical power model 2 | | | | | | | |
| S(NP) | **69** | **97** | **135** | **182** | 42 | **61** | **74** | **94** | 45 | **51** | **62** | **86** |
| M(NP) | 61 | 58 | 117 | 169 | **67** | 46 | 65 | 77 | **46** | 46 | 54 | 68 |
| | | | | | Empirical power model 3 | | | | | | | |
| S(NP) | 84 | 174 | 408 | 836 | **93** | 140 | 307 | 685 | **94** | 115 | 229 | 541 |
| M(NP) | **88** | **249** | **677** | **982** | 80 | **207** | **593** | **975** | 67 | **195** | **524** | **960** |

three expressions for $\hat{\psi}_t$ for all $t \in M$: (a) true values $\hat{\psi}_t = \psi_t$; (b) empirical marginal distribution for $\hat{\psi}_t$ and (c) estimated $\hat{\psi}$. Note that in expression (b) we include bias and variability issues with the fact we are using an estimator but we cancel the dependence structure present between pair of coefficients $(\hat{\psi}_t, \hat{\psi}_h)$ in (c). In Figure 4.2 we show the average of two of our test statistics (average of squares and maximum). In both cases, the effect of using an estimator for $[\psi_t]$ is visible. For instance, using the empirical marginal distribution of $\hat{\psi}_t$, as expected, supposes an increase on the variance of $\hat{d}_t$, and in consequence, the averages of the test statistics go up as well. However, when using the estimator of $[\psi_t]$, since their coefficients are themselves correlated for small sample size, the variance of the $\hat{d}_t$'s diminishes and this is reflected in the average of the two test statistics with a clear decrease.

The last topic we tackle here is deciding which parametric distribution is better to approximate the marginal distribution of $\hat{d}_t$ when random vectors $Y^{(1)}$ and $Y^{(2)}$ come from the same observations, and therefore $[\psi_t]$ coefficients have to be estimated. We compare the goodness of fit for the empirical distribution of all $[\hat{d}_t]$ against two theoretical distributions like standard normal and t-student. To do so we compute the average square difference between estimated values and expected value for the same quantile in the theoretical distribution. The normal approximation seems to get a better fit than the t-student, especially for small sample sizes (see Figure 4.3 (a)). Moreover, the mean square error does not vary much with regards to the sample size (see Figure 4.3 (b)).

## 4.6 Application to psoriasis vulgaris disease and lung cancer gene expression data

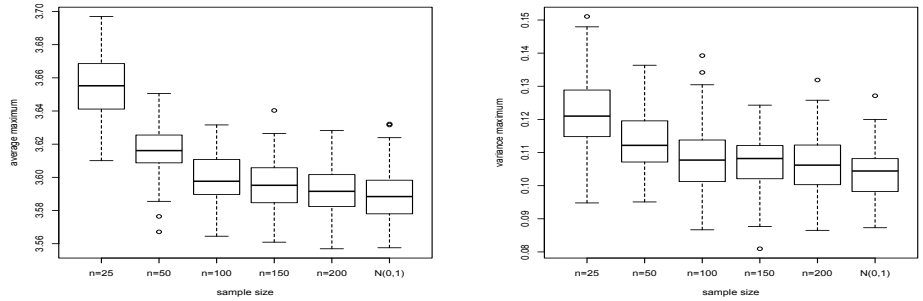
We apply the proposed HT problems to two different real case studies of gene expression data. The first dataset contains the gene expression profiling of 82 patients with psoriasis vulgaris disease in a paired lesional and non-lesional samples (Suárez-Fariñas et al., 2012). The second dataset represents the gene expression in a paired tumor and healthy samples from 60 female non-smoker patients with




**(a)** mean of $T_S$




**(b)** Variance of $T_S$




**(c)** mean of $T_M$




**(d)** Variance of $T_M$

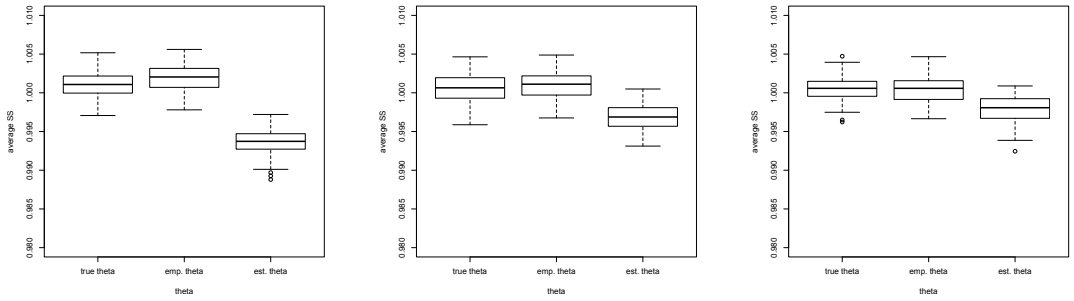
**Figure 4.1.** Boxplots with mean and variance of $T_S$ (average of squares statistic), and mean and variance of $T_M$ (extreme value statistic).

lung cancer (Lu et al., 2010). In both cases, there are 19,507 different genes which have been identified by the biomaRt R package (Durinck et al., 2005).

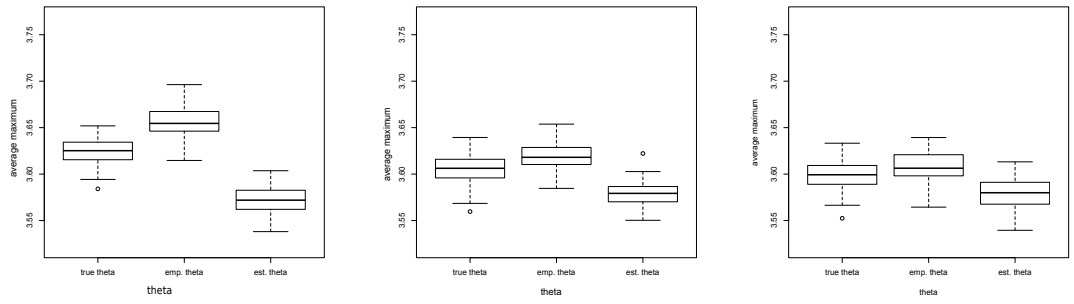
We are particularly interested in knowing how standard gene pathways change in different medical conditions. To assess which biological processes might be linked to changes in the gene connections we download 1,320 gene sets from the MSig database (`http://www.broadinstitute.org/gsea/msigdb/index.jsp`), which represent canonical pathways compiled from two sources: KeGG (`http://www.genome.jp/kegg/pathway.html`) and Reactome (`http://www.reactome.org/`). Then we test the equality of correlation matrices in the two medical conditions by only considering genes in each of the pathways. We also test the null hypothesis of identity correlation matrix in all these pathways lists to highlight the most linearly dependent groups of genes. On gene level we test both the hypothesis of equality and identity in the correlation matrices rows for all genes in the dataset. We use non-parametric null distribution for assessing all HT problems in either of the two datasets (psoriasis and lung cancer).

### 4.6.1 Testing identity and equality of correlation matrices using pathway lists

The hypothesis of identity correlation matrix (see Section 4.4.2) is evaluated for all genes within each of the 1,320 pathway lists, for both lesional (tumor) and healthy samples. Figure 4.4 shows the




**(a)** n=50, mean of $T_S$ **(b)** n=100, mean of $T_S$ **(c)** n=150, mean of $T_S$

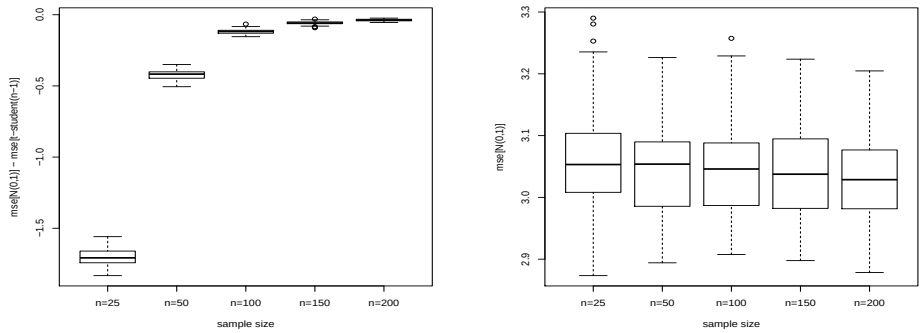



**(d)** n=50, mean of $T_M$ **(e)** n=100, mean of $T_M$ **(f)** n=150, mean of $T_M$

**Figure 4.2.** Boxplots with mean of $T_S$ (average of squares statistic) and mean of $T_M$ (extreme value statistic) using the true value for $\psi_t$ (left), a sampled value from its empirical marginal distribution (center) and the estimate value (right).

confidence interval for the average of squares test statistics. As expected almost all pathway lists are highly significant (indicated by green and red lines in the plots). In the lung cancer data, test statistics tend to be larger for healthy samples, though the largest values correspond to pathways for tumor samples. For psoriasis data, the differences between the two classes are not as big, at least in a general behavior, but pathways in lesional samples tend to have a larger $T_S^I$ than pathways for non-lesional samples. Some of the pathways with largest $T_S^I$ are *"REACTOME GABA A RECEPTOR ACTIVATION"*, *"KEGG MATURITY ONSET DIABETES OF THE YOUNG"*, *"REACTOME OLFACTORY SIGNALING PATHWAY"*, *"REACTOME RECYCLING OF BILE ACIDS AND SALTS"*, *"REACTOME LIGAND GATED ION CHANNEL TRANSPORT"*, and *"REACTOME SEROTONIN RECEPTORS"* for psoriasis data, and *"REACTOME UNWINDING OFDNA"*, *"BIOCARTA TCYTOTOXIC PATHWAY"*, *"BIOCARTA TCAPOPTOSIS PATHWAY"*, *"BIOCARTA THELPER PATHWAY"*, *"BIOCARTA TCRA PATHWAY"*, *"REACTOME ENDOSOMAL VACUOLAR PATHWAY"* for lung cancer data.

We also employ the HT problem for equality of correlation matrices in genes within the 1,320 pathways. In panels (a) and (b) of Figure 4.5 we present the approximated p-values using the three dependence-correction tests: average of squares, maximum and sum of exceedances (see Section 3.1.2) for the psoriasis and lung cancer datasets, respectively. In the sum of exceedances test we give




**(a)** N(0,1) vs t-student **(b)** N(0,1)

**Figure 4.3.** Mean square error differences for the quantile distribution of the empirical distribution of $\hat{d}_t$ against a N(0,1) and a t-student with $n-1$ degrees of freedom.

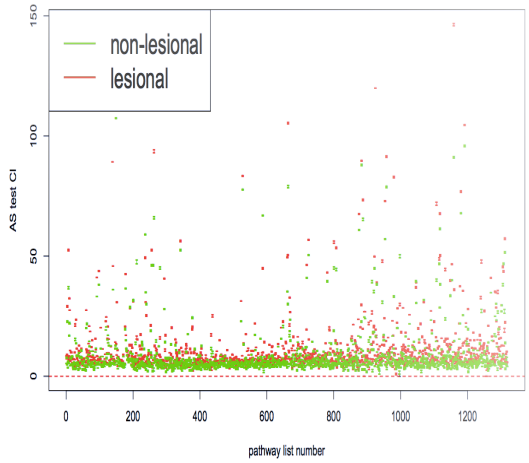
the results for $w = 0$, although they are very similar to the p-values found for $w = 1$.

Firstly for the psoriasis data, 72% of the average of squares test p-values, 34% of the extreme value test p-values and 70% of the sum of exceedances test p-values are smaller than 0.01 and under $H_0$ we were expecting only 1%. About 23% of the lists have the three tests with p-values smaller than 0.01. The correlation between average of squares and sum of exceedances p-values is 0.98, whereas the one between average of squares and maximum is 0.42, and exceedances and maximum is 0.52. Among others, the pathways lists that had the largest average of squares statistic are given in Table 4.7.

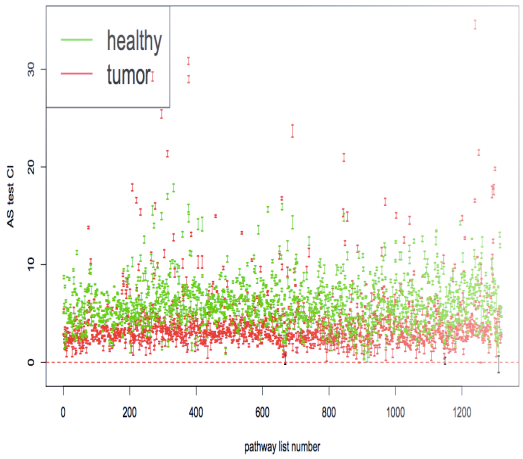
**Table 4.7.** Lists with the largest average of squares test statistic for psoriasis dataset on HT for equality of correlation matrices. Highly overlap label corresponds to pathways lists that contain more than 50% of their genes common to another list.

```
-------------------------------------------------------------------------
 [1] "KEGG_OLFACTORY_TRANSDUCTION"      (highly overlaps with [2])
 [2] "REACTOME_GPCR_DOWNSTREAM_SIGNALING"
 [3] "REACTOME_CLASS_C_3_METABOTROPIC_GLUTAMATE_PHEROMONE_RECEPTORS"
 [4] "REACTOME_PASSIVE_TRANSPORT_BY_AQUAPORINS"
 [5] "REACTOME_GABA_A_RECEPTOR_ACTIVATION"
 [6] "REACTOME_INHIBITION_OF_VOLTAGE_GATED_CA2_CHANNELS_VIA_GBETA_GAMMA_SUBUNITS"
 [7] "BIOCARTA_GABA_PATHWAY"     (highly overlaps with [5])
 [8] "BIOCARTA_ASBCELL_PATHWAY"
 [9] "KEGG_PROXIMAL_TUBULE_BICARBONATE_RECLAMATION"
[10] "REACTOME_UNBLOCKING_OF_NMDA_RECEPTOR_GLUTAMATE_BINDING_AND_ACTIVATION"
 -------------------------------------------------------------------------
```

Secondly for the lung cancer data, 61% of the average of squares test p-values, 35% of the extreme value test p-values and 63% of the sum of exceedances test p-values are smaller than 0.01. The 16% of the lists have the three tests with p-values smaller than 0.01. The correlation between average of squares and sum of exceedances p-values is 0.98, whereas the one between average of squares and maximum is 0.71, and exceedances and maximum is 0.65, which are higher than the ones observed




**(a)** CI psoriasis data




**(b)** CI lung cancer data

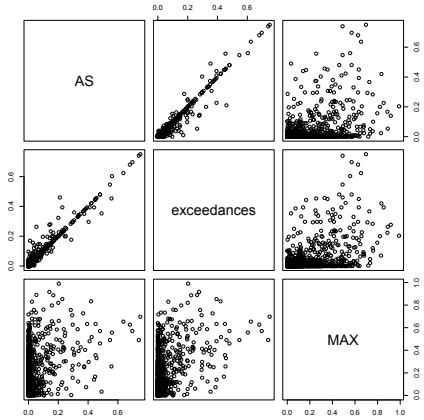
**Figure 4.4.** Hypothesis testing of identity correlation matrix in 1,320 pathway lists. Confidence interval for average of squares test statistic in (a) psoriasis and (b) lung cancer datasets.

for the psoriasis data. The 10 pathways lists that had the largest average of squares statistic are given in Table 4.8.

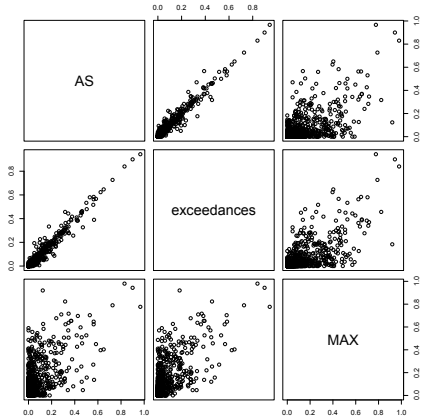
**Table 4.8.** Lists with the largest average of squares test statistic for lung cancer data on HT for equality of correlation matrices. Highly overlap label corresponds to pathways lists that contain more than 50% of their genes common to another list.

```
-----------------------------------------------------------------------
 [1] "REACTOME_ACTIVATION_OF_THE_PRE_REPLICATIVE_COMPLEX"
 [2] "REACTOME_UNWINDING_OF_DNA"  (highly overlaps with [1], [4] and [6])
 [3] "REACTOME_G1_S_SPECIFIC_TRANSCRIPTION"
 [4] "BIOCARTA_MCM_PATHWAY"  (highly overlaps with [1] and [6])
 [5] "REACTOME_ACTIVATION_OF_ATR_IN_RESPONSE_TO_REPLICATION_STRESS"
 [6] "BIOCARTA_LYM_PATHWAY"      (highly overlaps with [1])
 [7] "BIOCARTA_SKP2E2F_PATHWAY"
 [8] "BIOCARTA_IL17_PATHWAY"
 [9] "PID_ATR_PATHWAY"
[10] "BIOCARTA_VITCB_PATHWAY"
-----------------------------------------------------------------------
```

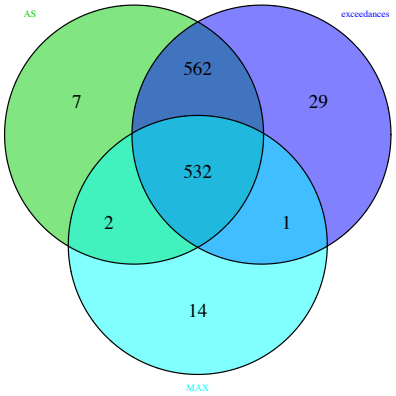
We further adjust the p-values for multiple testing by controlling the false discovery rate, and in Figure 4.5(b) we present a Venn's diagram of the adjusted p-values smaller than 0.05. Comparing the results in the two datasets, the p-values tend to be smaller in the psoriasis dataset. This was expected since the sample size for psoriasis data is fairly larger than the one for the lung cancer data. However, the obtained test statistics are not highly correlated between psoriasis and lung cancer data (p-values correlation of 0.13, 0.13 and 0.05 for $T_S$, $T_E$ and $T_M$) which may indicate that the gene connections are affected differently between the two type of diseases.



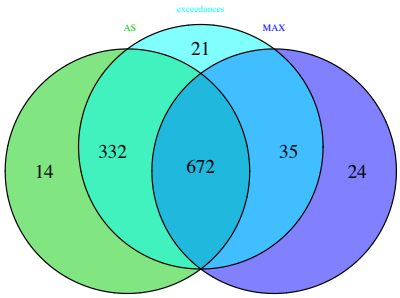

**(a)** p-values psoriasis data




**(b)** p-values lung cancer data




**(c)** Venn's diagram psoriasis data




**(d)** Venn's diagram lung cancer data

**Figure 4.5.** P-values for average of squares, sum of exceedances and maximum test statistics where each point corresponds to a pathway list equality of correlations p-value. Venn's diagram shows the number of rejected lists with an adjusted p-value smaller than 0.05.

### 4.6.2 Testing identity and equality of correlation matrices at gene level

We consider the HT problem of identity correlation matrix rows to find genes that act like hubs in the tumor (lesional) samples, i.e. genes that are highly dependent to many other genes. In the psoriasis dataset, the 93% and 87% of genes have a average of squares p-value smaller than 0.01 for lesional and non-lesional samples, respectively, with the 84% of genes being significant in both conditions (at 0.01 level). For the maximum test, almost all genes (99% for non-lesional and 98% for lesional) achieve p-values smaller than 0.01. The ten genes with the largest average of squares test statistic are VSX1, CALCA, FGB, ITGA4, CFAP65, CDY1, ARPP21, CNGB1, MBD3L1 and VWA3B for non-lesional samples, and AGXT, ADAM30, PEX5L, TRPC5, MUSK, OR2F1, RMST, ATP8B5P, LINC01541 and NEUROG2 for lesional samples. Analogously, in the lung cancer dataset, the 71% and 75% of genes have an average of squares p-value smaller than 0.01 for cancer and healthy samples, respectively, with the 60% of genes being significant in both conditions (at 0.01 level). For the maximum test, 88%, for healthy,

and 86%, for cancer, of the genes achieve p-values smaller than 0.01. The ten genes with the largest average of squares test statistic are DPY19L4, RABAC1, MIGA2, PLA2G4F, ATAD3B, STIP1, TTC31, FBXO3, SMAD2 and UPP1 for healthy samples, and PLIN3, LINC01088, GAST, ZNF839, KCNIP2, CRTC1, MIGA2, RABAC1, RN7SL731P and TAGLN3 for tumor samples.

Moreover, we test whether the genes are equally correlated in healthy and unhealthy samples. Hence, we use the testing procedure of equality of correlation matrix rows described at Section 4.4.1. For the psoriasis dataset, the 52% and the 70% of the genes have p-value smaller than 0.01 for average of squares and maximum tests, respectively. Besides, the 48% have both p-values smaller than 0.01. The genes with largest average of squares statistic are IPO5, HSPA12B, CBARP, GOLGA4, CDK14, VSTM2A, GLRX2, GATS, AQP4-AS1 and TRAV13-2. For the lung cancer dataset, the 32% (average of squares) and 57% (maximum) of genes have p-values smaller than 0.01. The 29% of the genes have both p-values smaller than 0.01. Important genes are FRMD5, P2RX5, PPP2R3C, SPRR1A, PRKAA1, MMP11, GBAS, SLC27A6, TMEM65 and EPS8L3.

## 4.7 Discussion

In this chapter we propose three tests for equality of two correlation matrices: average of squares, extreme value and sum of exceedances tests. These are especially useful for high-dimensional and paired datasets. We further suggest considering dependence-correction or non-parametric tests instead of asymptotic linear independence tests when the correlation matrices are known to be not highly sparse. Asymptotic tests, which assume independence among sample correlation coefficients, are much faster than the other two tests and could be used for highly sparse correlation matrices to speed up the process. For dense correlation matrices though, asymptotic tests can produce a non-negligible bias in the approximated p-values when the null hypothesis is true.

The idea of dependence-correction tests diverges with the methods seen so far in the literature. For instance, the extreme value test proposed in this paper contrasts with the results by Cai et al. (2014) who test the equality of mean vectors by employing the maximum of the square value of element-wise differences. The authors, as we have also done in Appendix A.3, prove that the limiting distribution of the maximum of dependent samples converges to the extreme value distribution of type I under very mild conditions and they examine this limiting distribution to assess the evidence of the test. We estimate the parameters using permuted samples since its known that the convergence of the parameters to the asymptotic ones is slow and we account for bias that arise in paired observations due to estimating correlation of sample correlation coefficients (Olkin and Finn, 1990).

In terms of test power, for a sensible selection of the exceedance threshold, sum of exceedances test is shown to be the most powerful test for sparse alternatives. If the sparsity levels are high, the extreme value test also provides competitive results. In contrast, for dense alternatives and small sample size, the average of squares dominates the asymptotic power.

We use 1,320 pathway lists to test equality of gene dependence's structures between normal and

lung cancer (psoriasis lesional) human samples in groups of genes that are known to interact together in a cell. A large part of the total number of lists has very small p-values. Especially, this happens in the average of squares and sum of exceedances tests. The extreme value test also gives smaller p-values than expected under the null hypothesis but it is more inclined to not reject $H_0$ than the other two tests. This could be an indication, if $H_1$ is true, that we are closer to the dense alternative scenario rather than the sparse scenario. This seems not unlikely as we consider genes in a single pathway so $R_1$ and $R_2$ are probably dense. In contrast, when testing the equality of correlation rows, extreme value test statistics achieve larger power than average of squares in both datasets.

# Chapter 5

# Gaussian graphical lasso and selection of sparsity tuning parameter

## 5.1 Introduction and motivation

In recent years, the study of undirected graphical models (Lauritzen, 1996) has been the focus of attention of many authors. The increasing volume of high-dimensional data in different disciplines makes them a useful tool in order to determine conditional dependence between random variables. For instance, graphical models have been applied to gene expression data sets to find biological associations across genes in Dobra et al. (2004) and Schäfer and Strimmer (2005), as well as in other biological networks (Dokuzoglu and Purtucuoglu, 2017) and in social networks (Goldenberg, 2007). In Gaussian graphical models, which are often used for finding associations between genes using high throughput genomic data, the dependence between the genes is fully characterized by the non-zero elements of the precision matrix $\Omega$ (defined as the inverse of the covariance matrix).

In a high-dimensional framework, where the number of variables $p$ is larger than the number of observations $n$, there is not enough information in the data available to estimate $\Omega$ by standard methods, and hence the underlying conditional dependence (CD) graph. To address this problem, alternative estimators have been proposed in the last two decades using additional information about $\Omega$ such that the estimated covariance matrix and its inverse are of full rank (see Section 3.2.2). In this chapter we consider the graphical lasso penalization method, which adds the penalty $\lambda||\Omega||_1$ with a tuning parameter $\lambda$ in the maximum likelihood to estimate $\Omega$. The penalized maximum likelihood optimization problem is solved using recursive algorithms, for instance we find that three of the most efficient and commonly used ways to solve it are glasso by Friedman et al. (2007), neighborhood selection (MB) by Meinshausen and Bühlmann (2006) and tuning-insensitive graph estimation and regression (tiger) by Liu and Wang (2012).

The choice of the tuning parameter $\lambda$ represents the trade-off between close fit to the data and

sparsity of $\Omega$, and its selection for estimation of the corresponding CD graph structure is the main focus of attention in this chapter. Methods such as Cross Validation (CV), Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) have been widely used to select tuning parameters when $p$ is small. However, they fail once dealing with high-dimensional problems by over-fitting the graph structure of $\Omega$ (Liu et al., 2010; Wasserman and Roeder, 2009). The eBIC criterion introduced by Chen and Chen (2008) extends BIC to account for high-dimensionality problems. Moreover, Liu et al. (2010) propose selecting $\lambda$ by controlling the desirable approximated variability in the estimated graphs using a subsampling approach (StARS). This method contrasts with the usual variable selection statistics since it only considers the estimated CD graph structure. Even though the method is promising and gives an alternative to eBIC, it has a major drawback: another tuning parameter is needed in order to set the maximum variability across samples which can be unknown *a priori* in many applications. Moreover, our simulations show that the default values can lead to overestimation of the network size in certain graph topologies. Meinshausen and Bühlman (2010) present a stability selection approach which controls the graph edges false discovery rate. The authors estimate $\Omega$ by an average subsampling graphical lasso method such that the effect of the choice of $\lambda$ is very low. However, the trade-off between false positive and true positive edges of the selected network by their subsampling approach is worse than the one given by a network with the same number of edges using all the data due to considering smaller effective sample sizes than the original $n$ for estimation.

In the biological literature, the most commonly used approaches to construct gene networks are based on clustering. This is informed by the expected presence of distinct strongly interconnected clusters in biological networks (Eisen and Spellman, 1998; Yi et al., 2007). This gave us the motivation to find $\lambda$ such that the corresponding graph has a clustering structure which can be interpreted by a biologist without restricting it to a block diagonal structure and hence missing potentially important interactions.

Our aim is to select the hyperparameter $\lambda$ such that (a) it produces reliable estimates of the edges of the graph (b) the corresponding CD graph structure is interpretable in terms of network characteristics and (c) works well for networks that arise in biological systems. In this chapter, we propose several such approaches to selecting $\lambda$, in the framework of a general two-step procedure. The main novelty with respect to classical approaches such as AIC or BIC is that we use only the graph structure of the graphical lasso estimator to tune the regularization parameter $\lambda$. The first proposed approach, path connectivity (PC), uses the average geodesic distance of estimated networks to find the graph that corresponds to the biggest change of the number of connections and is associated with splitting of clusters. The second method, augmented mean square error (A-MSE), similarly to the StARS approach, controls the variability of the estimated networks in terms of graph dissimilarity coefficients using either subsampling or a Monte Carlo based approach. The main difference from StARS is the additional bias term to avoid having a tuning parameter. We consider the bias with respect to an initial estimated graph structure which contains a desirable global network characteristic. For instance, we use the AGNES hierarchical clustering coefficient (Kaufman and Rousseeuw, 2009), which

is the third proposed method to choose $\lambda$, to select the graph that presents the highest clustering structure. Although clustering methods exist in the literature, the novelty here is that we use them to select the penalty parameter $\lambda$ in graphical lasso estimation. The last method we employ to select the tuning parameter is called graph vulnerability (VUL) since finds the most vulnerable estimated graph structure, i.e., removing a variable supposes the biggest change in the resulting graph structure.

We compare performance of the proposed approaches as well as of the StARS algorithm and the eBIC criterion on both simulated and real data. The data is a microarray gene expression data set generated by the TCGA Research Network (`http://cancergenome.nih.gov/`). It contains 154 samples for patients with colon tumor and about 18,000 genes. We are particularly interested in finding significant complex gene interactions reliably and relating the observed associations to pathway databases which describe known biochemistry connections between genes. Simulations and real data analysis are performed using the R package **ldstatsHD**, which is fully described at Chapter 7.

The rest of the chapter is organized as follows. In Section 5.2 we review some of the main algorithms to estimate sparse precision matrices as well as their theoretical and computational properties. In Section 5.3 we introduce the tuning parameter selection methodology and in Section 5.4 we give their main algorithmic and computational information. In Section 5.5 we compare the performance of the methods using simulated data and then apply them to a gene expression dataset in Section 5.6.

## 5.2 Gaussian graphical model

### 5.2.1 Problem set up

Consider $n$ independent and identically distributed (i.i.d) observations from a Gaussian model: $Y_k \sim N_p(0,\Sigma)$, $k = 1,\ldots,n$, assuming, without a loss of generality, that the mean is zero. CD (conditional dependence) is totally characterized by the inverse covariance matrix $\Omega = \Sigma^{-1}$, which is widely known as precision matrix. Two Gaussian random variables $Y_i$ and $Y_j$ are said to be conditionally independent given all the remaining variables if the coefficient $\Omega_{ij}$ is zero. Recall from Chapter 2 that CD is often expressed with a graph structure $G(V,E)$ in which each node in $V$ represents a random variable and there is an edge in $E$ connecting two different nodes if the correspondent element in the inverse covariance matrix is non-zero.

The corresponding log likelihood function for $\Omega$ is $\ell(\Omega) = \log\det\Omega - tr(S\Omega)$ where $S = n^{-1}\sum_{k=1}^{n} Y_k^2$. If $S^{-1}$ exists ($p < n$ is a necessary condition), the maximum likelihood estimator (MLE) of $\Omega$ is given by $S^{-1}$. However, in a high-dimensional framework where the number of variables $p$ is larger than the number of observations $n$, the matrix $S$ is singular and so cannot be inverted.

Assume that the CD graph is sparse, and hence that the precision matrix $\Omega$ is sparse. Ideally, we would like to use a penalized likelihood estimator with the penalty proportional to the number of non-zero elements in $\Omega$. However, such optimization problem is non-convex and thus is very computationally intensive. In practice, a likelihood estimator with a convex penalty term proportional

to the $\ell_1$ norm of $\Omega$, a graphical lasso (GL), is commonly used instead:

$$\hat{\Omega}^{\lambda}_{GL} = \arg\max_{\Omega \succ 0} \{\log\det\Omega - tr(S\Omega) - \lambda||\Omega||_1]\}, \tag{5.1}$$

where $||\Omega||_1 = \sum_{i,j=1}^{p} |\Omega_{ij}|$ is the element-wise $\ell_1$ norm of the matrix $\Omega$. For small $\lambda$, the corresponding penalized estimator of $\Omega$ tends to be dense and in the extreme ($\lambda = 0$) it coincides to the initial maximum likelihood problem which may not have unique solution when $p/n$ is large (Pourahmadi, 2011). As $\lambda$ increases, the estimated matrix becomes more and more sparse towards a diagonal matrix. Therefore, the choice of $\lambda$ has a crucial effect on the estimated CD graph structure.

### 5.2.2 Graph notation and distances

We give some basic definitions and properties of networks (Costa and Rodrigues, 2007; Estrada, 2011) which will be used throughout the chapter. The graph structure $G(V,E)$ is often represented by a $p \times p$ matrix, called adjacency matrix and denoted by $A_G$. In the estimation of graphical models, the off-diagonal elements of $A_G$ are determined by the precision matrix (0 if $\Omega_{ij} = 0$ and 1 otherwise) and the diagonal elements are set to zero. Note that graphical models are undirected which means that the correspondent $A_G$ is symmetric.

The distance between a pair of nodes $\{V_i, V_j\} \in G(V,E)$ (also known as the geodesic distance) defines the shortest number of edges connecting node $V_i$ to the node $V_j$ and it is denoted by $g_{ij}$. If there is no path linking the two nodes, then $g_{ij} = \infty$. The correlation coefficient $\rho_{ij}$ between two nodes $\{V_i, V_j\} \in G(V,E)$ and the corresponding dissimilarity measure $d_{ij}$ are given by

$$\rho_{ij} = \eta_{ij}/\sqrt{\kappa_i \kappa_j}, \quad \text{with} \quad d_{ij} = 1 - \rho_{ij}, \; P = [\rho_{ij}], \quad D = J - P \tag{5.2}$$

where $\eta_{ij}$ is the number of neighbors shared by the nodes $V_i$ and $V_j$, $\kappa_i$ is the degree of the node $V_i$ defined as the number of nodes that are directly connected to $V_i$ and $J$ is the matrix of ones.

### 5.2.3 Coordinate descent for regression lasso and Gaussian graphical lasso

In this section, we describe the coordinate descent procedure presented in Friedman et al. (2007) that is used to estimate the lasso regression coefficients and it is also a fundamental step in the Gaussian graphical lasso algorithm. We present the standard glasso method by Friedman et al. (2007) as well as the neighborhood selection strategy by Meinshausen and Bühlmann (2006) and the tiger extension by Liu and Wang (2012).

*Coordinate descent for estimation of regression coefficients*

Let $y$ be a $n$ vector with i.i.d. realizations of a Gaussian random variable and let $X$ be a $n \times p$ matrix

with explanatory variables. The regression lasso optimization problem is defined by

$$\hat{\beta}^{\lambda}_{lasso} = \underset{\beta}{\arg\min} \left\{ \frac{1}{2n} ||y - X\beta||_2^2 + \lambda ||\beta||_1 \right\}, \tag{5.3}$$

where $\hat{\beta}^{\lambda}_{lasso}$ are the estimated regression coefficients with tuning parameter $\lambda$. Note that eq. (5.3) is equivalent to solving

$$\hat{\beta}^{\lambda}_{lasso} = \underset{\beta}{\arg\min} \left\{ \frac{1}{2n} \sum_{k=1}^{n} (y_k - X_{kj}\beta_j - \sum_{i \neq j} X_{ki})^2 \beta_i + \lambda \sum_{i \neq j} |\beta_i| + \lambda |\beta_j| \right\}. \tag{5.4}$$

This problem is solved by a coordinate descent algorithm (Friedman et al., 2007), which is an iterative procedure where regression coefficients are estimated one by one keeping all the other values fixed. For instance, setting $\beta_i = \tilde{\beta}_i$, the parameter $\beta_j$ is estimated (and it is denoted by $\tilde{\beta}^{\lambda}_j$) by minimizing expression (5.4) with respect to $\beta_j$. The solution of the minimization problem is found by

$$\tilde{\beta}^{\lambda}_j \leftarrow ST\left( \frac{\sum_{k=1}^{n} X_{kj}(y_k - \sum_{i \neq j} X_{ki}\tilde{\beta}_i)}{\sum_{k=1}^{n} X_{ki}^2}, \lambda \right), \tag{5.5}$$

where $ST$ is the soft thresholding operator defined by

$$ST(z, \lambda) = sign(z)(|z| - \lambda)_+.$$

Given starting values for $(\tilde{\beta}_j)_{j=1}^{p}$, all the coefficients are updated using eq. (5.5) iteratively until convergence. Friedman et al. (2007) show that the $(\tilde{\beta}^{\lambda}_j)_{j=1}^{p}$ values converge to $(\hat{\beta}^{\lambda}_{lasso_j})_{j=1}^{p}$.

*Glasso*

Banerjee et al. (2008) initially proposed partitioning $\hat{\Sigma}$ (the estimator of the covariance matrix $\Sigma$) and its inverse $\hat{\Omega}$, with $\hat{\Sigma} = \hat{\Omega}^{-1}$, such that the row and column of interest (the variable $i$) are relocated in the last row and column as follows

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{-i,-i} & \hat{\Sigma}_{-i,i} \\ \hat{\Sigma}_{i,-i} & \hat{\Sigma}_{i,i} \end{pmatrix}, \quad \hat{\Omega} = \begin{pmatrix} \hat{\Omega}_{-i,-i} & \hat{\Omega}_{-i,i} \\ \hat{\Omega}_{i,-i} & \hat{\Omega}_{i,i} \end{pmatrix},$$

and identically for the sample covariance matrix,

$$S = \begin{pmatrix} S_{-i,-i} & s_{-i,i} \\ s_{i,-i} & s_{i,i} \end{pmatrix}.$$

Using this scheme, Friedman et al. (2007) show that the graphical lasso maximization problem defined in eq. (5.1) is equivalent to solving $p$ minimization problems

$$\hat{\beta}_i = \underset{\beta_i}{\arg\min} \left\{ \frac{1}{2} ||\hat{\Sigma}^{1/2}_{-i,-i}\beta_i - \hat{\Omega}^{1/2}_{-i,-i} s_{i,-i}|| + \lambda ||\beta||_1 \right\}, \tag{5.6}$$

where regression coefficients in $\hat{\beta}_i$, defining a vector of size $p-1$, are linked to $\hat{\Sigma}$ by $\hat{\Sigma}_{i,-i} = \hat{\Sigma}_{-i,-i}\hat{\beta}_i$. Moreover, the corresponding row of the precision matrix is determined by $\hat{\Omega}_{i,-i} = -\hat{\beta}_i\hat{\Omega}_{i,i}$ and $\hat{\Omega}_{i,i} = 1/(\hat{\Sigma}_{i,i} - \hat{\Sigma}_{i,-i}\hat{\beta}_i)$. The authors present a recursive algorithm to find both $\hat{\Sigma}$ and $\hat{\Omega}$ based on the following steps:

1. Given the tuning parameter $\lambda$, initialize the estimated covariance matrix by $\hat{\Sigma} = S + \lambda I$.

2. Solve the problem in (5.6) for all the nodes permuting $\hat{\Sigma}$ and $\hat{\Omega}$ such that in each case the target node occupies the last row and column of the matrix. The coefficients in $\beta_i$ could be updated by coordinate descent using soft-thresholding under each $\beta_{ij}$, for any $j \neq i$. For instance, let $W = \hat{\Sigma}_{-i,-i}$ and $u = s_{-i,i}$, regression coefficients in the $p-1$ vector $\beta_i$ are updated iteratively by $\hat{\beta}_{ij} = ST(u_j - \sum_{h\neq j} W_{hj}\hat{\beta}_{ih}, \lambda)/W_{jj}$.

3. Continue until convergence in $\hat{\Sigma}$ and $\hat{\Omega}$.

*Neighborhood selection*

An alternative interpretation of the graphical lasso problem is presented in Meinshausen and Bühlmann (2006). Even though the authors do not propose an algorithm to find the precision matrix, they recover in an elegant way the correspondent graph $G(V,E)$, which describes the non-zero structure in the precision matrix. They introduce the concept of neighborhood selection: given the node $i \in V$, find the smallest subset of nodes in $V \setminus \{i\}$, which will form the neighborhood of $i$, denoted by $Y_{\text{ne}_i}$, such that $Y_i$ is perpendicular to all the remaining data $(Y \setminus Y_{\text{ne}_i})$. This problem can be solved by using a lasso type constraint for the number of nonzero elements.

*Tiger*

The minimization problem proposed by Liu and Wang (2012) is of similar fashion as the one given in (5.6). It estimates the precision matrix by solving the next $p$ lasso regressions problems

$$\hat{\beta}_i = \arg\min_{\beta_i} \left\{ \frac{1}{\sqrt{n}} ||Y_i - Y_{-i}\beta_i||_2 + \lambda||\beta_i||_1 \right\}, \quad i \in \{1,\dots,p\}. \tag{5.7}$$

by coordinate descent using a Lagrangian reformulation. The estimator of $\Omega$ is found by computing the next three steps

1. $\hat{\beta}_i = \arg\min_{\beta_i}\{(1 - 2\hat{\Omega}_{-i,i}\beta_i + \beta_i^\intercal\hat{\Omega}_{-i,-i}\beta_i)^{1/2} + \lambda||\beta_i||_1\}$ which can be solved by the coordinate descent algorithm presented above.

2. $\hat{\tau}_i = (1 - 2\hat{\Omega}_{-i,i}\hat{\beta}_i + \hat{\beta}_i^\intercal\hat{\Omega}_{-i,-i}\hat{\beta}_i)^{1/2}$.

3. Given $\hat{\Gamma} = \text{diag}(S)$, $\hat{\Omega}_{ii} = \hat{\tau}_i^{-2}\hat{\Gamma}_{ii}^{-1}$ and $\hat{\Omega}_{-i,i} = \hat{\tau}_i^{-2}\hat{\Gamma}_{ii}^{-1/2}\hat{\Gamma}_{-i,-i}^{-1/2}\hat{\beta}_i$.

### 5.2.4 Theorethical and computational comparison of the methods

The most relevant assumptions of the glasso, neighborhood selection and tiger approaches are the following:

**Allowing high-dimensional cases ($p \gg n$):** for a constant $\gamma > 0$, neighborhood selection assumes that $p = O(n^\gamma)$ and glasso takes $p \leq n^\gamma$. Tiger relaxes the high-dimensionality condition by $\lim_{n\to\infty} \gamma\sqrt{(\log p)/n} = 0$, thus assuming that $\log p$ grows slower than $n$.

**Non-singularity in $\Omega$:** it is shared by all the studied methods. Denote by $\varphi(\Omega)$ the vector with eigenvalues of $\Omega$, the authors bound the condition number of $\Omega$ given a positive constant $c$ by $\varphi_{max}(\Omega)/\varphi_{min}(\Omega) < c$.

**Sparsity in $\Omega$:** neighborhood selection assumes sparsity in the adjacency matrix $A_G$ (defined in Section 5.2.1) such that the sum of non-zero elements in each row is less than the sample size. Tiger also constrains the number of edges so the sum of non-zero elements in each row is less than a constant $\gamma$ with $\gamma^2 \log p = o(n)$.

**Marginal variance of $Y$ and magnitude in the elements of $\Omega$:** tiger assumes that the marginal variance of $Y$ do not diverge fast ($\max_j \Sigma^2_{jj} < \frac{n}{4\log p}$ ) as $n$ grows and neighborhood selection imposes that the non-zero elements of $\Omega$ are bounded away from 0 which makes the recovery of the network more feasible.

If the assumptions above hold, the Frobenius loss function for $\hat{\Omega}$ using the glasso algorithm (Zhou et al., 2010) is given by

$$||Q^\lambda_{gl} - \Omega||_F = O_p\left(2M\sqrt{\frac{(p+s)\log n}{n^{2/3}}}\right),$$

where $p + s = O(n^{2/3}/\log n)$, large constant $M$ and $\lambda_n \asymp \sqrt{\frac{\log n}{n^{2/3}}}$. Similarly, for the tiger estimator, the Frobenius norm error between $\hat{\Omega}$ and $\Omega$ is

$$||Q^\lambda_{tig} - \Omega||_F = O_p\left(k||\Omega||_1\sqrt{\frac{(p+s)\log p}{n}}\right),$$

where $\lambda_n \asymp \zeta\pi\sqrt{\frac{\log p}{2n}}$ with $\zeta \in [\sqrt{2}/\pi, 1]$. The norm error by tiger is lower than glasso if $p = O(n)$. The underlying graph structure of $\Omega$ is recovered using tiger by

$$\inf_{n\to\infty} P(A \subset \hat{A}_{tig}) = 1,$$

which is slightly more conservative than the asymptotic result for the neighborhood selection algorithm, that recovers $A$ by

$$P(\hat{A}^\lambda_{mb} = A) = 1 - O(\exp(-cn^\epsilon)), \qquad n \to \infty \quad \text{and} \quad \epsilon > 0, \ c > 0.$$

Tiger and glasso are asymptotically tuning-parameter free, meaning that the optimal convergence rates defined above hold for any $\lambda$ in its specified interval. The convergence values for the neighborhood selection approach depend on the selection of the optimal tuning parameter $\lambda$ or prediction-oracle solution (Meinshausen and Bühlmann, 2006), i.e.,

$$\lambda_{oracle} = \arg\min_{\lambda} E(\tilde{Y}_i - \sum_{j \in \text{ne}_i} q_{ji}^{\lambda} \tilde{Y}_j), \tag{5.8}$$

where $\text{ne}_i$ defines the neighborhood of conditional dependent variables for the target variable $i$. The problem is that $\tilde{Y}$ is a new unknown matrix for $Y$ and cross validation is normally used to approximate expression (5.8).

In terms of computational time, neighborhood selection is the fastest algorithm of the three, with glasso being slightly faster than tiger. A comprehensive comparison of these methods as well as some other approaches like the PC-algorithm (to find a directed acyclic graph -DAGs-) using simulated data is given in Albieri and Didelez (2014).

## 5.3 Regularization parameter selection

### 5.3.1 General two step procedure to select the tuning parameter

The $\ell_1$ penalized maximum likelihood estimator defined in (5.1) requires the selection of a regularization parameter $\lambda$. If the $\ell_1$ penalization genuinely represented our true prior knowledge about $\Omega$ then one of the standard methods such as the maximum marginal likelihood or cross validation for the elements of $\Omega$ could be used. However, the $\ell_1$ penalty here is used due to its computational convenience, replacing the $\ell_0$ penalty, so these methods are not appropriate. It is well known for the problem of estimating sparse vectors in high-dimensions with the lasso penalty, that the variable selection part, with an appropriate $\lambda$, is consistent, however, the estimation of the non-zero values usually has some bias (Wasserman and Roeder, 2009; Gu et al., 2013). This can be due to the convex relaxation of the desired $\ell_0$ penalty to the computationally efficient $\ell_1$ penalty. Therefore, we suggest to employ methods that use only the variable selection part from the glasso, $\widehat{G}^{\lambda}(V,E)$, for tuning the hyperparameter $\lambda$.

We propose the following two step procedure for estimating $\lambda$:

1. Set $\hat{\Omega}_{GL}^{\lambda}$ as in equation (5.1) for all $\lambda \in \Lambda$, $\Lambda \subset [0, \lambda_{max}]$, $\lambda_{max} > 0$.

2. Choose $\hat{\lambda} = \arg\min_{\lambda} R(\lambda, \widehat{G}^{\lambda}(V,E), \tilde{G}(V,E))$

using risk functions $R$ that are based only on CD graphs $\widehat{G}^{\lambda}(V,E)$ and (possible) initial graph $\tilde{G}(V,E)$. This procedure combines computational efficiency of the lasso algorithm with the choice of $\lambda$ that optimizes relevant characteristics of the CD graph such as connectivity, clustering structure, etc.

### 5.3.2 Proposed risk functions

We propose several risk functions to select $\lambda$ that monitor network characteristics of the conditional dependence graphs that can be applicable to genomic data. It has been observed (Yi et al., 2007) that molecules in a cell work together in groups, with some – usually less strong – interaction between the groups. This motivates our choice of risk functions to encourage a clustering structure in the estimated graphs. We further present the method developed in Liu et al. (2010) to select the tuning parameter by controlling the estimated variability of the graph and the eBIC likelihood-based approach described in Chen and Chen (2008). Both methods are compared to our proposed approaches in simulated data.

*StARS regularization parameter selection*

Liu et al. (2010) propose a resampling approach to select $\lambda$. The procedure is based on subsampling without replacement $T$ samples of size $b$ from the $n \times p$ matrix $Y$. The graph structure $G^{\lambda(t)}(V,E)$ is estimated (e.g., using neighborhood selection, glasso or tiger) for all $t = 1,\ldots,T$. Let $\hat{\theta}^{\lambda}_{ij}$ be the proportion of times that an edge exists connecting two nodes, i.e.,

$$\hat{\theta}^{\lambda}_{ij} = T^{-1} \sum_{t=1}^{N} I(\hat{A}^{\lambda(t)}_{G_{ij}} = 1).$$

Assuming that $\hat{A}^{\lambda(1)}_{G_{ij}},\ldots,\hat{A}^{\lambda(T)}_{G_{ij}}$ are independent, the proportion $\hat{\theta}^{\lambda}_{ij}$ can be viewed as an estimator of the parameter of a binomial distribution, whose variance is given by $\text{var}(\hat{\theta}^{\lambda}_{ij}) \approx \hat{\zeta}^{\lambda}_{ij} = \frac{1}{T}\hat{\theta}^{\lambda}_{ij}(1-\hat{\theta}^{\lambda}_{ij})$. The average of $\hat{\zeta}^{\lambda}$, denoted by $\bar{D}_{\lambda} = \sum_{i<j}\hat{\zeta}^{\lambda}_{ji}/m$ for $m = p(p-1)/2$, can be understood as a measure of stability of all edges for a given graph with regularization parameter $\lambda$. The selection of $\lambda$ by StARS depends on the amount of variability that is allowed in the graph

$$\lambda_{st} = \sup\{\lambda : \bar{D}_{\lambda} \leq \beta\} \tag{5.9}$$

where $\beta$ is a power tuning parameter which controls the magnitude of this variability. Generally, a small $\beta$ corresponds to a large $\lambda$ and a high $\beta$ consequently gives a low $\lambda$. We assume $\beta = 0.05$ for all simulated scenarios presented in the Section 5.5 which is the default value proposed in Liu et al. (2010). The motivation behind this method resides in the fact that the problem of selection of $\lambda$ is transformed to the selection of the maximum amount of variability $\beta$ in the graph, which might be easier to interpret.

*Path connectivity risk function*

To motivate the path connectivity risk function, observe the following obvious property of the graph $\hat{G}^{\lambda}(V,E)$ that corresponds to the penalized estimator $\hat{\Omega}^{\lambda}$ defined by (5.1): for small $\lambda$, the likelihood term dominates and the estimator $\widehat{G}^{\lambda}(V,E)$ is usually a dense graph with $\hat{\Omega}^{\lambda}$ closely fitting the data, and for large $\lambda$, the penalty term dominates and the corresponding estimate is a very sparse graph with $\hat{\Omega}^{\lambda}$ not fitting the data well. Thus, for growing values of $\lambda$, there is a decrease in graph complexity, and the aim here is to capture the value of $\lambda$ that corresponds to the largest change in the complexity

of the graph.

For simplicity, consider a grid of values of $\lambda$, $\Lambda = (\lambda_k)_{k=1}^{M}$ such that $\lambda_k - \lambda_{k-1} = h$, $k = 2, \ldots, M$, and the underlying estimated graphs $\widehat{G}^{\lambda}(V, E)$, for all $\lambda \in \Lambda$. Path connectivity (PC) is a novel approach to find $\lambda$ that finds the biggest change in graph complexity between the graphs $\hat{G}^{\lambda}$ corresponding to two consecutive values of $\lambda \in \Lambda$. In this case the measure of graph complexity is calculated by the *geodesic distance mean* statistic

$$H(\lambda) = \frac{2}{p\,(p-1)} \sum_{i<j} \hat{g}_{ij}(\lambda) I(\hat{g}_{ij}(\lambda) < \infty), \tag{5.10}$$

where $\hat{g}_{ij}(\lambda)$ are the geodesic distances for the graph $\hat{G}^{\lambda}(V, E)$ as defined in Section 5.2.1. To find the largest change in $H(\lambda)$, consider the first order differences of $H(\lambda)$ by $D_h(\lambda) = \Delta_h H(\lambda)$, where $\Delta_h$ refers to the difference operator with bandwidth $h$. The regularization parameter selection by PC is given by the $\lambda$ that produces the most rapid relative descent in the number of graph connections

$$\lambda_{pc} = \arg\min_{\lambda_k \in \Lambda} R_{PC}(\lambda_k) = \arg\min_{\lambda_k \in \Lambda} \left\{ -\left| D_h(\lambda_k) / \bar{D}_h(\lambda_k) \right| \right\}, \tag{5.11}$$

where $\lambda_k$ is the $k$-th ordered element in $\Lambda$ and $\bar{D}_h(\lambda_k)$ is the running average defined as the average of elements $D_h(\lambda)$ with $\lambda \in \{\lambda_1, \ldots, \lambda_k\}$. The difference of the geodesic distance mean is divided by $\bar{D}_h(\lambda_k)$ in eq. (5.11) to favor big jumps for larger $\lambda_k$ (and sparser $\hat{G}^{\lambda}(V, E)$) in comparison to the jumps for smaller $\lambda_k$ which correspond to denser graphs.

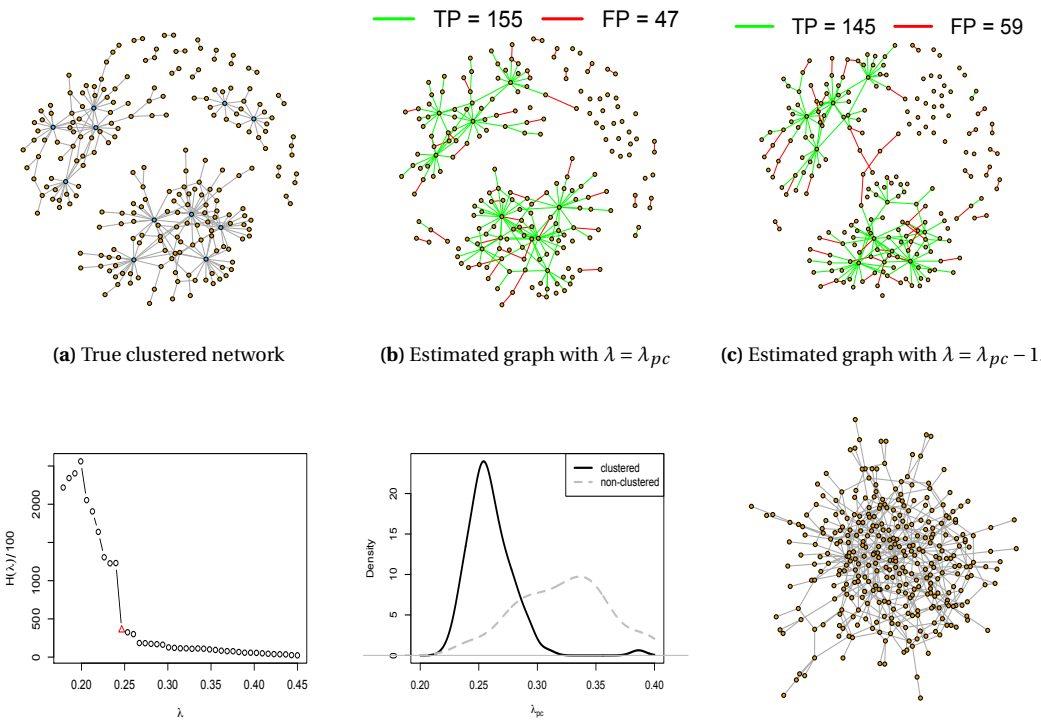
In Figure 5.1 we illustrate the motivation of using the PC selection of $\lambda$ in simulated data (see Section 5.5 for details). The true CD graph structure defined by three non-overlapping clusters is plotted in Figure 5.1(a). We show the geodesic distance mean as function of $\lambda$ for graph estimations in Figure 5.1(d). This presents a few big jumps which are related to the separation of clusters. The last one gives the selected graph by PC and is due to the partition of two clusters (see Figure 5.1(b) for the selected $\lambda_{pc} = \lambda_k$ and Figure 5.1(c) for the previous graph structure defined by $\lambda_{k-1}$). This is a generally observed behaviour in both simulated and real gene expression datasets. In Figure 5.1(e) we show the density estimates of $\lambda_{pc}$ using 100 i.i.d. datasets with $n = 200$, $p = 350$ and two theoretical graph structures: hub-based clustered graph as shown in Figure 5.1(a) and non-clustered/random graph structure as shown in Figure 5.1(f). We can see the clear peak around $\lambda = 0.25$ for the clustered data against a flatter empirical distribution for the non-clustered data.

*A-MSE risk function*

The idea explored in this section is to use a risk function based on network characteristics such as dissimilarities of the graph defined by eq. (5.2). Ideally, we would like to find $\lambda^{\text{oracle}}$ that minimizes

$$R_{MSE}(\lambda) = \mathbf{E}(\sum_{i>j} |d_{ij} - \hat{d}_{ij}(\lambda)|^q), \tag{5.12}$$

for some $q \geq 1$ where $d_{ij}$ are the dissimilarities of the true graph and $\hat{d}_{ij}(\lambda)$ are the dissimilarities of the CD graph estimated by expression (5.2) for a given tuning parameter $\lambda$. For $q = 2$, this risk




**(a)** True clustered network

**(b)** Estimated graph with $\lambda = \lambda_{pc}$

**(c)** Estimated graph with $\lambda = \lambda_{pc} - 1$.

**(d)** Geodesic distance mean

**(e)** Densities of $\lambda_{pc}$

**(f)** True non-clustered network

**Figure 5.1.** Path connectivity regularization parameter selection (PC) using the clustered graph structure in (a) to generate the data. Figure (b) shows the selected network by PC and (c) its previous estimated network. In both networks, true positive edges are in green whereas false positives are in red. The graphical structure in (b) differs from the one in (c) since the two clusters in the bottom are no longer connected by a (false positive) edge. Figure (d) shows the geodesic distance mean statistic over several values for $\lambda$ in which the triangle point is $\lambda_{pc}$. Figure (e) illustrates the empirical distribution of $\lambda_{pc}$ over 100 i.i.d. instances of data with true graph structure in (a), with black solid line, and true graph structure in (f), with grey dashed line. The first concentrates the values to a peak at 0.25 whereas the second is more disperse leading to values of $\lambda_{pc}$ ranging from 0.27 to 0.35.

function can be expressed as a sum of the variance terms and the sum of the squared differences between the initial and the current estimator (the "bias" term),

$$R_{MSE}(\lambda) = \sum_{i>j} [\mathbf{E}(\mathbf{E}[\hat{d}_{ij}(\lambda)] - \hat{d}_{ij}(\lambda))^2 + (\mathbf{E}[\hat{d}_{ij}(\lambda)] - d_{ij})^2]. \tag{5.13}$$

Note that the first term in (5.13), the variance of the estimated distances, gives a stability measure similar to the one proposed in StARS (the latter uses the adjacency matrix instead of the dissimilarities). However, the addition of the bias term for the distance estimator permits avoiding the selection of the power tuning parameter $\beta$ that controls the desired variability in the StARS approach.

The risk function $R_{MSE}(\lambda)$ depends on the unknown true graph structure of $\Omega$; in practice, an unbiased estimator of $R_{MSE}(\lambda)$ is used, commonly obtained by subsampling (bootstrap, cross validation) by comparing estimated values to observations. However, the problem in this setting is that direct observations of $d_{ij}$ are not available. To overcome this issue we propose to use an initial graph

estimate $\tilde{G}(V,E)$ and its dissimilarities coefficients $[\tilde{d}_{ij}]$ in place of observed data. Thus, the choice of $\lambda$ is given by

$$\lambda_{amse} = \arg\min_{\lambda\in\Lambda} \hat{R}_{AMSE}(\lambda) = \arg\min_{\lambda\in\Lambda} \sum_{i>j} \hat{\mathbf{E}}|\tilde{d}_{ij} - \hat{d}_{ij}(\lambda)|^2, \tag{5.14}$$

where $\hat{\mathbf{E}}$ indicates the estimation of the expected value, which is done using either subsampling or Monte Carlo based approaches.

The proposed $R_{AMSE}(\lambda)$ risk can be applied to other network characteristics. By the definition of graph dissimilarities, $d_{ij} = 1$ if nodes $i$ and $j$ are neither directly nor indirectly (share neighbor) connected. Let $h_{ij} = 0$ if $\sigma_{ij} = 1 - d_{ij} = 0$ and $h_{ij} = 1$ if $\sigma_{ij} > 0$. For sparse networks, there are many $(h_{ij} = 0)_{i<j}$ and only few $(h_{ij} = 1)_{i<j}$. Applying the $R_{AMSE}(\lambda)$ to the simplified similarity coefficient $[h_{ij}]$ instead of $[d_{ij}]$, leads to

$$R^h_{AMSE}(\lambda) = \mathbf{E}\sum_{i<j}(h_{ij} - \hat{h}_{ij}(\lambda))^2 = C_h + \mathbf{E}\sum_{(ij)\in\theta(\lambda)}(1 - 2h_{ij}) = C_h + \mathbf{E}[TP(\lambda) - FP(\lambda)],$$

where $\theta(\lambda) = \{(i,j); i < j \,\&\, \hat{h}_{ij}(\lambda) = 1\}$, $\text{FP}(\lambda) = \sum_{i<j} I[h_{ij} = 0, \hat{h}_{ij}(\lambda) = 1]$, $\text{TP}(\lambda) = \sum_{i<j} I[h_{ij} = 1, \hat{h}_{ij}(\lambda) = 1]$ and $C_h$ is independent of $\lambda$. Minimizing $R^h_{AMSE}(\lambda)$ is the same as maximizing the TP and FP differences (also known as Youden index).

Since the true values of $[h_{ij}]$ are unknown, here we assume that an initial graph with "best" global characteristics is available, i.e., exists $\tilde{\lambda}$ such that $\sum_{i<j} h_{ij} \approx \sum_{i<j} \hat{h}_{ij}(\tilde{\lambda})$. An approximation of $\mathbf{E}[\text{FP}(\lambda) - \text{TP}(\lambda)]$ is then found by subsampling or Monte Carlo based approaches with

$$\lambda^h_{amse} = \arg\min_{\lambda} \hat{R}^h_{AMSE}(\lambda) = \arg\min_{\lambda} \sum_{i>j} \hat{\mathbf{E}}|\hat{h}_{ij}(\tilde{\lambda}) - \hat{h}_{ij}(\lambda)|^2. \tag{5.15}$$

In practice, biologists often use clustering algorithms to discover groups of genes. Hence, we propose to use the output of a hierarchical clustering algorithm as an initial estimate of the graph to characterize global structure for the dissimilarities $[d_{ij}]$. We have investigated several clustering algorithms on real and simulated data, and we have not found much difference in the resulting graph estimate. Below we present the AGNES clustering method.

*AGNES risk function*

Clustering of features using a dissimilarity measure has been intensively studied in the literature. Here we focus on the algorithm AGNES (AGglomerative NESting) which is presented in Kaufman and Rousseeuw (2009, chap. 5) and is implemented in the R package **cluster** (Rousseeuw et al., 2013). AGNES finds clusters iteratively joining groups of nodes with the smallest average dissimilarity coefficient. This average is found by considering the dissimilarity coefficients between all possible pairs of nodes coming from two different clusters. Moreover, AGNES provides an agglomerative coefficient (AC) that measures the average distance between a node in the graph and its closest cluster

of nodes. We propose to choose $\lambda$ that maximizes the AC coefficient

$$\lambda_{ac} = \arg\min_{\lambda\in\Lambda} \hat{R}_{AGNES}(\lambda) = \arg\min_{\lambda\in\Lambda} \{-AC(\lambda)\}. \tag{5.16}$$

The details of the AGNES algorithm and the definition of the coefficient AC can be found in Section 5.4. The matrix of dissimilarities $D$ obtained by (5.2) gives a good representation of the complexity of a given graph, so, in addition to being applied as an initial estimate for the A-MSE method described above, AGNES can also be used as a method of choosing $\lambda$.

*Vulnerability risk function*

Another proposed approach to select $\lambda$ corresponds to finding the graph that is most vulnerable from a range of estimated graphs. Vulnerability (VUL) is measured by

$$R_{VUL}(\lambda^t) = -\sum_{i=1}^{p} \frac{E^{\lambda^t} - E_g^{\lambda^t}}{E^{\lambda^t}},$$

where $E^{\lambda}$ is the global efficiency of the original network $\hat{G}^{\lambda}(V,E)$ and $E_g^{\lambda}$ is the global efficiency of the same network once eliminating gene $g$ and their underling connections, which can be expressed by $\hat{G}^{\lambda}(V\setminus g, E\setminus\{g\leftrightarrow \text{ne}(g)\})$. Thus, it measures the effect of removing a node in the estimated network. Global efficiency is defined here by the harmonic mean of the geodesic distance

$$E^{\lambda} = \frac{2}{p(p-1)}\left(\sum_{i<j}\frac{1}{\hat{g}_{ij}(\lambda)}\right)^{-1}. \tag{5.17}$$

We propose to choose $\lambda$ by

$$\lambda_{vul} = \arg\min_{\lambda\in\Lambda} \hat{R}_{VUL}(\lambda) = \arg\min_{\lambda\in\Lambda} \left\{-\sum_{i=1}^{p} \frac{E^{\lambda^t} - E_g^{\lambda^t}}{E^{\lambda^t}}\right\}. \tag{5.18}$$

*eBIC risk function*

The eBIC criterion to select $\lambda$ is presented in Chen and Chen (2008) and provides an extension of BIC for high-dimensional data. As for the standard BIC, it is a likelihood-based expression, so the precision matrix $\Omega$ needs to be estimated. The expression for eBIC risk function is

$$R_{eBIC}(\lambda) = -\log\det\hat{\Omega}_{\lambda} - tr(S\hat{\Omega}_{\lambda}) + K\log(n)/2 + 2\phi\log(\tau(\hat{\Omega}_{\lambda})),\ \ \tau(\hat{\Omega}_{\lambda}) = \begin{pmatrix} K \\ s(\hat{\Omega}_{\lambda}) \end{pmatrix},$$

where $s(\hat{\Omega}_{\lambda})$ is the number of non-zero elements in the precision matrix estimation and $0 \le \phi \le 1$ weights the importance of the sparsity models. For $\phi = 0$, this risk coincides with the $BIC$ criterion. The tuning parameter selection is given by

$$\lambda_{eBIC} = \arg\min_{\lambda} R_{eBIC}(\lambda). \tag{5.19}$$

### 5.3.3 Comparison of the methods

Table 5.1 provides some of the main properties of the 6 risk functions discussed in Section 5.3.2, which are the four proposed methods, as well as StARS (Liu et al., 2010) and eBIC (Chen and Chen, 2008). Likelihood-based risk functions to select $\lambda$ such as AIC, BIC (which are presented in Section 3.5) or eBIC (described above) are useful to compromise between goodness of fit to the data and model overfitting. The additional AIC penalty (given by $p(p-1)$) is smaller than BIC (given by $p(p-1)\log(n)/2$) even for very small $n$. Hence, the selection of $\lambda$ by AIC results in a denser CD graph structure of $\Omega$ than by BIC. Moreover, eBIC, which penalizes the chances of estimating a graph structure with a certain sparsity level, encourages extremer graph sizes, both highly dense and highly sparse graphs, than BIC as the weight $\phi$ grows towards 1. StARS gives a good alternative to select $\lambda$ when only estimating graph structures. It transforms the selection of $\lambda$ problem to the choice of the maximum expected variability allowed in the graph. Even though such a choice is more intuitive than the direct selection of $\lambda$, we find it difficult to use without any prior information; our simulations show that using the default value of the tuning parameter results in high number of false positive edges (see Section 5.5.4).

We provide two computationally fast approaches, AGNES and PC, and the slightly more computationally challenging A-MSE and VUL methods. The AGNES selection tends to find the most clustered graph possible such that different groups of nodes can be interpreted and analyzed. This is found to be a good choice of $\lambda$ to recover global graph structure characteristics when the true precision is block diagonal (see Section 5.5.5 for simulated data analysis). The A-MSE selection uses the AGNES estimator as the initial graph structure with the aim to improve estimations of local network characteristics. The value of $\lambda$ selected by A-MSE is usually smaller than the one given by the initial estimator (AGNES), and it is used to stabilize the trade-off between false positive and true positive edges in the original estimator (AGNES) when $n$ is small (see Section 5.5.4 for simulated data analysis). Moreover, as the sample size increases, the value of $\lambda$ chosen by the A-MSE method tends to the original estimator of $\lambda$ (AGNES). Path connectivity provides an initial good choice of $\lambda$ to find the most sparse graph that is easy to interpret. Starting from the sparsest graph and proceeding to denser graph structures, the PC method monitors the first big change in connectivity of the estimated networks, which is frequently associated with cluster agglomerations. Finally, the graph vulnerability selection approach encourages graph structures that are highly impacted by elimination of variables. This reflects a network characteristic that could also be used individually to each variable in the dataset to measure its importance in the conditional dependence graph.

## 5.4 Algorithms

### 5.4.1 Path connectivity regularization parameter selection

The procedure to select $\lambda$ by path connectivity is detailed in Algorithm 3. It is generally fast and straightforward, i.e., does not require any additional tuning.

**Table 5.1.** Risk functions main characteristics that are separated between statistics that use the likelihood expression (eBIC) and statistics that only use the graphical structure of the estimated precision matrices (PC, A-MSE, AGNES, StARS, VUL).

| method | penalized likelihood | uses network characteristics. | subsampling | fully automatic | fast | highly sparse graph estimates |
|---|---|---|---|---|---|---|
| PC | | ✓ | | ✓ | ✓ | ✓ |
| A-MSE | | ✓ | ✓ | ✓ | | ✓ |
| AGNES | | ✓ | | ✓ | ✓ | |
| VUL | | ✓ | | ✓ | | ✓ |
| StARS | | ✓ | ✓ | | | |
| eBIC | ✓ | | | ✓ | ✓ | ✓ |

**Algorithm 3** Path connectivity algorithm

**procedure** $\mathrm{R}_{PC}(\lambda)$
  Set $\Lambda = (\lambda_k)_{k=1}^{M}$ with $\lambda_k - \lambda_{k-1} = h$, $k = 2,\ldots,M$.
  **for** k in 1 until M **do**:
    Estimate the graph $\hat{G}^{\lambda_k}(V,E)$ using eq. (5.1) and calculate its geodesic distance matrix $[\hat{g}_{ij}]$ as in eq. (5.2).
    Calculate geodesic distance mean $H(\lambda_k) = m^{-1}\sum_{i<j}\hat{g}_{ij}(\lambda_k)I(\hat{g}_{ij}(\lambda_k)<\infty)$ with $m = p(p-1)/2$.
  Calculate $D_h(\lambda_k) = H(\lambda_k) - H(\lambda_{k-1})$ and the running average $\bar{D}_h(\lambda_k) = 1/(M-k-1)\sum_{j=k}^{M} D_h(\lambda_j)$ for $(\lambda_k)_{k=2}^{M}$.
  Return $D_h(\lambda_k)/\bar{D}_h(\lambda_k)$, $k = 2,\ldots,M$.

### 5.4.2 A-MSE regularization parameter selection

The subsampling procedure to select $\lambda_{amse}$ is presented in Algorithm 4. Following Meinshausen and Bühlman (2010), the effective sample size is chosen to be $B = 0.5n$ since the procedure gets the closest to bootstrap. Nevertheless, other effective sizes could be employed, e.g., Liu et al. (2010) suggest to use $B = 10\sqrt{n}$.

**Algorithm 4** Subsampling approach to approximate (5.13)

**procedure** $\mathrm{R}_{AMSE}(\lambda)$
  Set $\Lambda = (\lambda_k)_{k=1}^{M}$ and number of subsampling replicates T.
  **for** t in 1 until T **do**:
    Subsample $B \subset \{1:n\}$ and set $Y_B = (Y_j, j \in B)$.
    Estimate the graphs $\hat{G}^{\lambda_k(t)}(V,E)$ for all $\lambda_k \in \Lambda$ using $Y_B$.
    Find dissimilarities of $\hat{G}^{\lambda_k(t)}(V,E)$ by $\hat{d}_{ij}^{(t)}(\lambda_k) = 1 - \eta_{ij}^{(t)}(\lambda_k)/\sqrt{\kappa_i^{(t)}(\lambda_k)\kappa_j^{(t)}(\lambda_k)}$.
  Set initial graph dissimilarities $\tilde{d}_{ij}(\lambda_k)$ for all $i \le j$.
  Return $T^{-1}\sum_{t=1}^{T}\{\tilde{d}_{ij}(\lambda_k) - \hat{d}_{ij}^{(t)}(\lambda_k)\}^2$ for all $\lambda_k \in \Lambda$.

The algorithm to select $\lambda$ by A-MSE using a Monte Carlo based approach is described in Algorithm 5. It is based on simulating $n$ i.i.d. samples $y_k' \sim N(0, \hat{\Omega}^{\tilde{\lambda}})$, for $k = 1,\ldots,n$ (with same sample size $n$), where $\hat{\Omega}^{\tilde{\lambda}}$ is the estimated precision matrix (using $\tilde{\lambda}$). For the generated new data, graphical lasso estimates using the same tuning parameters sequence $\Lambda = (\lambda_k)_{k=1}^{M}$ are found. Let $\hat{h}_{ij}'(\lambda)$ be the $\lambda$-estimated simplified similarity coefficient obtained using the generated new samples. We conjecture that if (A1) $\sum_{i\neq j}\hat{h}_{ij}(\tilde{\lambda}) \approx \sum_{i\neq j} h_{ij}$, with $h_{ij}$ being the true values, the sum of squared differences between simplified $\hat{h}_{ij}'(\lambda)$ and the initial $\hat{h}_{ij}(\tilde{\lambda})$ is a good approximation of the sum of

squared differences between $\hat{h}_{ij}(\lambda)$, for the original graphical lasso, and $h_{ij}$. Note that under the global characteristic assumption (A1), when $n \to \infty$, $\hat{\Omega}^{\tilde{\lambda}} \to \Omega$ and $\hat{R}^h_{AMSE}(\lambda) \to R^h_{AMSE}(\lambda)$.

**Algorithm 5** Monte Carlo approach to approximate (5.14)

**procedure** R($\lambda$)
  Set $\Lambda = (\lambda_k)_{k=1}^M$ and initial $\tilde{\lambda} \in \Lambda$.
  Estimate $\hat{\beta}$ by using p-regression models: i.e., $Y_i \sim N(Y_{ne_i}\beta_{i,ne_i}, \sigma^2)$ where $ne_i = \{\forall\, j : \hat{A}^{\tilde{\lambda}}_{ij} = 1\}$.
  Find $\hat{\Omega}^{\tilde{\lambda}}$ by symmetrizing the matrix $\beta$ with unit diagonal (i.e., use `forceSymmetric` from R package **Matrix**).
  Find $\hat{R}^{\tilde{\lambda}}$ by inverting $\hat{\Omega}^{\tilde{\lambda}}$ using a quadratic regularization (Danaher et al., 2014).
  **for** t in 1 until T **do**:
    Generate $n$ i.i.d. samples $y_k^{(t)} \sim N(0, \hat{R}^{\tilde{\lambda}})$, for $k = 1, \ldots, n$.
    Estimate the graphs $G^{\lambda_k(t)'}$ for all $\lambda_k \in \Lambda$ with the new sampled data.
    Find dissimilarities $d_{ij}^{(t)'}(\lambda_k) = 1 - \eta_{ij}^{(t)'}(\lambda_k)/\sqrt{\kappa_i^{(t)'}(\lambda_k)\kappa_j^{(t)'}(\lambda_k)}$ and simplified similarities $[\hat{h}_{ij}^{(t)'}(\lambda)]$.
  Return $T^{-1}\sum_{t=1}^T \{\hat{h}_{ij}^{(t)'}(\lambda_k) - \hat{h}_{ij}(\tilde{\lambda}_k)\}^2$.

The Monte Carlo approach does not depend on extra parameters whereas setting a re-sampling sample size is needed for the subsampling approach. However, if $\hat{\Omega}^{\tilde{\lambda}}$ is quite different to the true $\Omega$, which happens for small sample sizes, locally, $h_{ij}$ will be quite different to $\tilde{h}_{ij}$ and the estimator of $\lambda$ will not be reliable.

### 5.4.3 AGNES regularization parameter selection

The AGNES iterative clustering algorithm, including the agglomeration coefficient that is used to select $\lambda$, is detailed in Algorithm 6. The input to the algorithm is a dissimilarity matrix $D = [d_{ij}] = \hat{D}(\lambda)$ based on the graph $\hat{G}^{\lambda}$ corresponding to the estimator $\hat{\Omega}^{\lambda}$ defined by eq. (5.1). AGNES performs hierarchical clustering by iteratively joining groups of nodes with the smallest average dissimilarity coefficient, starting with individual nodes as single clusters and finishing with a single cluster of all $p$ variables. Let $(C_1^{(t)}, \ldots, C_p^{(t)})$ be a partition of $(1:p)$ at iteration $t$, and let $\delta_{k,\ell}^{(t)}$ denote a dissimilarity between clusters $C_k^{(t)}$ and $C_m^{(t)}$. We also record the dissimilarity for each node when it merges with another cluster or node for the first time, denoting it by $\delta_j^\star$, $j = 1, \ldots, p$, and the distance $\delta^\star_{\max}$ between the two clusters merged at the last step into the single cluster.

The coefficient $AC(\lambda)$ measures the average distance between a node in the graph and its closest cluster of nodes. When the dissimilarities within the clusters are small in comparison to the maximum dissimilarity, then $1 - \delta_j^\star/\delta^\star_{\max}$ is large for all $j$ and $AC(\lambda)$ is consequently high.

### 5.4.4 Vulnerability regularization parameter selection

The vulnerability algorithm used to select $\lambda$ is presented in Algorithm 7. This results to a computationally intensive algorithm, i.e., $M \times p$ graphical lasso models need to be computed where $M$ is the size

**Algorithm 6** AGNES clustering algorithm

1: **procedure** $\mathrm{R}_{AGNES}(\lambda)$

2: Initialization: take each node as an individual cluster, i.e. set $C_k^{(0)} = \{k\}$, $k = 1, \ldots, p$, and $\delta_{k,\ell}^{(0)} = d_{k,\ell}$ - dissimilarity between nodes $k$ and $\ell$.

3: At iteration $t \geq 0$:

4: Find pair of clusters $(h,k)$ $(h < k)$ with the smallest dissimilarity, i.e.

$$(h,k) = \arg\min_{i<j} \delta_{i,j}^{(t)},$$

merge them, i.e. set $C_k^{(t+1)} = \{C_k^{(t)}, C_h^{(t)}\}$ and remove cluster $h$: $C_h^{(t+1)} = \emptyset$. Remaining clusters are unchanged: set $C_j^{(t+1)} = C_j^{(t)}$ for $j \neq k, h$.

5: The dissimilarities change to

$$\delta_{j,h}^{(t+1)} = \delta_{h,j}^{(t+1)} = \infty, \quad \delta_{k,j}^{(t+1)} = \delta_{j,k}^{(t+1)} = \frac{1}{2}\left[\delta_{k,j}^{(t)} + \delta_{j,h}^{(t)}\right], \quad \forall\, j \neq k, h.$$

If $|C_k^{(t)}| = 1$, set $\delta_k^\star = \delta_{k,h}^{(t)}$; if $|C_h^{(t)}| = 1$, set $\delta_h^\star = \delta_{k,h}^{(t)}$.

6: If the number of non-empty sets (clusters) in the newly formed partition $(C_j^{(t+1)})$ is more than 1, then set $t = t+1$ and go to step 3; otherwise set $\delta_{\max}^\star = \delta_{k,h}^{(t)}$.

7: Return

$$AC(\lambda) = \frac{1}{p}\sum_{j=1}^{p}\left(1 - \frac{\delta_j^\star}{\delta_{\max}^\star}\right). \tag{5.20}$$

of the grid of $\lambda$. We consider an alternative proposal when $p$ is large that finds the most vulnerable graph with respect to removing groups of variables. Thus, we develop a leave-$K$-out procedure so $K$ variables are removed randomly from the dataset in step 5. This process is repeated $L$ times, $L \ll p$, so $M \times L$ graphical lasso computations are required.

**Algorithm 7** Vulnerability algorithm

1: **procedure** $\mathrm{R}_{VUL}(\lambda)$

2: Set $\Lambda = (\lambda_k)_{k=1}^M$.

3: Estimate the graph $\hat{G}_g^{\lambda_k}(V,E)$ using (5.1).

4: **for** g in 1 until p **do**:

5: Remove $g$th variable from the estimated graph structure $\hat{G}_g^{\lambda_k}(V,E)$.

6: **for** k in 1 until M **do**:

7: Calculate geodesic distance matrix $[\hat{g}_{ij}]$ as in eq. (5.2).

8: Calculate the efficiency $E_g^{\lambda_k} = \frac{1}{m}\left(\sum_{i<j}\frac{1}{\hat{g}_{ij}(\lambda_k)}\right)^{-1}$, with $m = p(p-1)/2$.

9: Return $R_{VUL}(\lambda^t) = -\sum_{i=1}^{p}\frac{E^{\lambda^t} - E_g^{\lambda^t}}{E^{\lambda^t}}$.

## 5.5 Simulated data analysis

In this section we consider simulated data to test the performance of the regularization parameter selection methods using graph structures similar to what can be expected in biological networks. We analyze both the capacity to obtain the true connections and the accuracy in recovering network

characteristics of the true graph.

### 5.5.1 Graph topologies in biological datasets

In "real world" problems that arise from social networks, information networks and biological networks, the graph which defines a kind of level of interaction between nodes (e.g., people in social networks, papers in information networks or genes in biological networks) is unknown but there is typically some knowledge about what sort of network structure can be expected (Newman, 2003; Reinert, 2009; Estrada, 2011).

Biological graph structures which define conditional dependence between nodes by a sparse precision matrix usually present associations in the shape of clusters, meaning that the nodes form groups that are more similar to the nodes within the group than to the nodes of other groups (Eisen and Spellman, 1998). Two distributions that are found to approximate biological networks well are hub-based and power-law networks. Hub-based networks are graphs where only few nodes have a much higher degree (or connectivity) than the rest. This is a common case in biological processes where nodes that behave as hubs may have different biological functions than the other nodes (Lu et al., 2007). The degree of a node $g \in V$ in a graph $G(V,E)$ is defined as the number of edges that connect nodes $V \setminus g$ to $g$. Let $p_b$ be the fraction of nodes in the network that have degree $b$, power-law networks assume that $p_b$ follows a power-law distribution, i.e.,

$$p_b \sim b^{-\alpha} \varsigma(\alpha)^{-1},$$

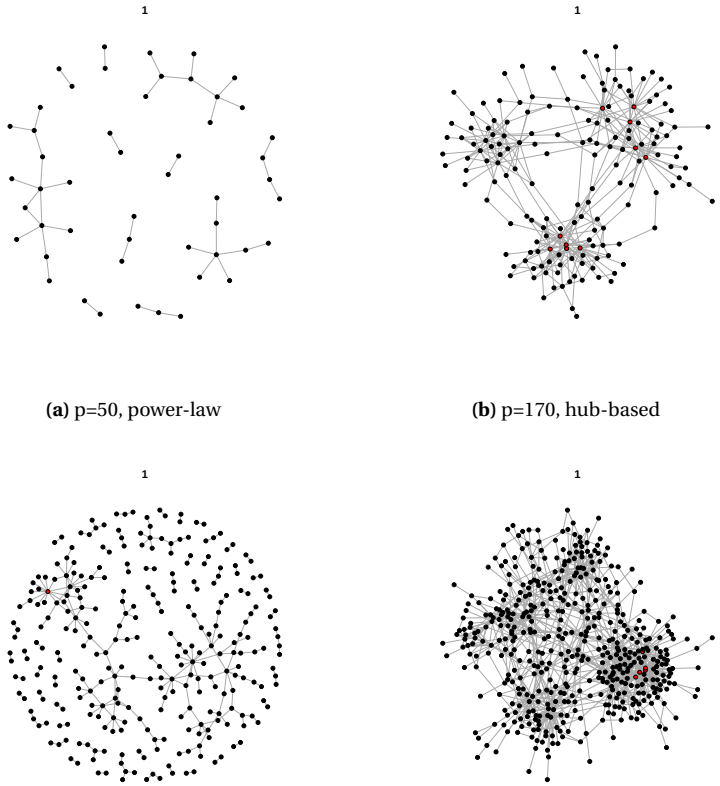
where $b \geq 1$, $\alpha$ is a positive constant and the normalizing function $\varsigma(\alpha)$ is the Riemann zeta function. Following Peng et al. (2009), $\alpha = 2.3$ provides a distribution that is close to what is expected in biological networks.

### 5.5.2 Simulated data

We generate data from multivariate normal distributions with zero mean vector and several almost-block diagonal precision matrices, where each block (or cluster) has a hub-based or power-law underlying graph structure (defined in Section 5.5.1) and there are some extra random connections between blocks. The non-zeros of the precision matrices, which we initially denote by $\Omega^{(0)}$, are obtained following eq. (4.36). These generated matrices may not be positive definite, so we regularized them by $\Omega^{(1)} = \Omega^{(0)} + \delta I$, with $\delta$ such that the condition number of $\Omega^{(1)}$ is less than the number of nodes, so obtaining a positive definite matrix (Cai et al., 2011). Simulated precision matrices are non-singular, sparse and with the non-zero elements bounded away from 0.

We consider precision matrices with $p = 50$, 170, 290 and 500 and sample sizes $n = 50$, 100, 200, 500. The number of clusters (and variables per cluster) for each $p$ setting are: 1 (50), 3 (70, 60, 40), 5 (70, 100, 40, 50, 30), 7 (100, 100, 80, 60, 60, 70, 30). The degree of hub nodes is generated by an Uniform$(5, b)$

where $b$ is one third of the number of variables per cluster. Moreover, the probability for presence of all remaining edges in hub-based models is determined by an Uniform$(0.005, 0.03)$ random variable and the probability for presence of edges in between clusters is given by an Uniform$(0, 0.1)$ random variable. Following Peng et al. (2009), power-law parameter $\alpha$ is set to 2.3 since provides a distribution that is close to what is expected in biological networks. Figure 5.2 shows some of the created networks.



**(a)** p=50, power-law

**(b)** p=170, hub-based

**(c)** p=290, power-law

**(d)** p=500, hub-based

**Figure 5.2.** Graphical represenation of some of the CD structures used to generate simulation data.

We use the R package **huge** (Zhao et al., 2012) to estimate CD graph structures by glasso and neighborhood selection (MB), as well as the R package **camel** (Li et al., 2013) to find the tiger estimates. The glasso and tiger provide the values of the estimated precision matrix whereas MB only give their underlying non-zero structures. In order to compare the proposed methods to the likelihood-based eBIC approach, we only present the results for the glasso procedure. Nevertheless, a performance comparison between the three algorithms to estimate $\Omega$ is presented in Section 5.5.4. We take a sequence of 60 equidistant points for $\lambda$ going from 0.20 to 0.66 for small $n$ and a sequence going from 0.03 to 0.40 for large $n$ (the graphs almost have no change for $\lambda$'s smaller than the lower limit with all nodes connected as well as higher than the upper limit with no edges across nodes). Then we select $\lambda$ by seven different approaches: 1) PC; 2) AGNES; 3) A-MSE (subsampling -sub); 4) A-MSE (Monte Carlo

-mc); 5) VUL; 6) StARS; 7) eBIC. The method StARS (with $\beta = 0.05$) produces the lowest $\lambda$ for almost all the simulated datasets. The eBIC results are strongly dependent on the sample size; the method selects large tuning parameters for small $n$ and low tuning parameters for large $n$ in comparison to A-MSE. The AGNES selections are always larger than A-MSE (sub) but they get close when $n$ increases. The PC $\lambda$ selections do not vary much for different $n$ and $p$ scenarios and produce similar magnitudes to $\lambda$'s selected by A-MSE (mc). The two A-MSE algorithms find similar tuning parameters, with the subsampling approach tending to give slightly larger $\lambda$'s than the Monte Carlo approach.

We assess the performance of the $\lambda$ selection approaches for glasso estimates using two different measures: squared errors in both the partial correlation matrix and the dissimilarity matrix defined in (5.2) and graph recovery with a false positive and true positive analysis. The simulated data analysis is completed by comparing the selected graph structures against the true networks with regards to global network characteristics such as clustering, connectivity and graph topology.

### 5.5.3 Mean square errors for estimated precision and dissimilarity matrices

To measure performance of the methods we use the ranks of the average mean square errors (MSE) of the precision matrix $\Omega$ (Table 5.2) as well as of the dissimilarity matrix $D$ (Table 5.3). This second rate gives a good reference to determine if the estimated graph captures the true local structure. The lowest rank (rank = 1) is assigned to the lowest MSE and the largest rank (rank = 7) is for the largest MSE out of the seven approaches. In the tables, we show the errors for the glasso method.

Even though StARS estimates $\Omega$ well, it produces larger errors than AGNES, A-MSE, PC, VUL and eBIC when minimizing the MSE of the dissimilarity matrix. Particularly, A-MSE (for both subsampling and Monte Carlo approximations) tends to be the best selection for this loss function. We find that eBIC does well for small $n$, contrarily of what is obtained in Liu et al. (2010), but tends to be unreliable for larger sample sizes. AGNES gives good ranks for the power-law scenarios, particularly when $n$ is large. PC and VUL achieve similar levels and are only slightly worse than A-MSE.

### 5.5.4 Graph recovery of graphical modelling approaches

We compare the performance of the three suggested graphical lasso based methods: glasso, neighborhood selection (mb) and tiger. To do so, we present the ROC curve, which corresponds to the graphical representation of the sensitivity (True Positive Rate - TPR) and the complement of the specificity (False Positive Rate - FPR) defined by TPR = TP/P and FPR = FP/N with

$$TP = \sum_{i<j} I(\hat{\Omega}_{ij} \neq 0 \text{ and } \Omega_{ij} \neq 0), \qquad FP = \sum_{i<j} I(\hat{\Omega}_{ij} \neq 0 \text{ and } \Omega_{ij} = 0), \tag{5.21}$$

and $P = \sum_{i<j} I(\Omega_{ij} \neq 0)$, $N = \sum_{i<j} I(\Omega_{ij} = 0)$. Figure 5.3 shows the ROC curves in a unique simulated data set, for $p = 290$ and several $n$ values, which is quite representative of the behavior in the 60 simulations. Each of the three lines corresponds to the FPR-TPR for graph estimation by glasso, MB

**Table 5.2.** Average ranks for the mean square error of the precision matrix using several sample sizes, dimension and network topologies (hub-based and power law). The method StARS finds the best rates (lowest ranks) whereas PC and A-MSE tend to obtain the worst rates (highest ranks).

| | Hub-based | | | | Power law | | | |
|---|---|---|---|---|---|---|---|---|
| n | 50 | 100 | 200 | 500 | 50 | 100 | 200 | 500 |
| | | | | dimension p=50 | | | | |
| PC | 4.34 | 4.84 | 5.55 | 5.66 | 4.05 | 4.96 | 4.80 | 4.53 |
| AGNES | 2.30 | 2.42 | 2.94 | 3.08 | 2.64 | 3.01 | 4.08 | 4.84 |
| A-MSE (sub) | 6.12 | 6.26 | 6.14 | 5.96 | 5.58 | 5.47 | 6.14 | 5.80 |
| A-MSE (mc) | 5.70 | 5.98 | 6.17 | 6.03 | 5.22 | 5.21 | 5.17 | 4.73 |
| VUL | 2.91 | 3.17 | 3.35 | 3.67 | 3.09 | 3.90 | 4.74 | 5.06 |
| StARS | **1.00** | **1.00** | **1.20** | **1.73** | **1.00** | **1.00** | **1.02** | **1.50** |
| eBIC | 5.63 | 4.32 | 2.65 | 1.87 | 6.42 | 4.45 | 2.05 | 1.53 |
| | | | | dimension p=170 | | | | |
| PC | 3.85 | 4.38 | 5.04 | 4.90 | 3.85 | 4.41 | 5.47 | 4.56 |
| AGNES | 2.04 | 2.01 | 2.04 | 2.88 | 2.03 | 2.09 | 2.90 | 4.30 |
| A-MSE (sub) | 6.56 | 6.42 | 6.14 | 5.90 | 6.53 | 5.87 | 6.08 | 6.00 |
| A-MSE (mc) | 5.38 | 5.60 | 5.55 | 5.50 | 4.92 | 4.41 | 4.82 | 4.30 |
| VUL | 3.52 | 4.39 | 5.00 | 5.69 | 3.47 | 4.97 | 5.45 | 5.84 |
| StARS | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| eBIC | 5.66 | 4.20 | 3.23 | 2.13 | 6.21 | 5.26 | 2.27 | 2.00 |
| | | | | dimension p=290 | | | | |
| PC | 3.75 | 3.70 | 4.66 | 5.12 | 3.80 | 4.03 | 5.54 | 5.02 |
| AGNES | 2.00 | 2.01 | 2.01 | 2.63 | 2.00 | 2.01 | 2.48 | 3.92 |
| A-MSE (sub) | 6.60 | 6.51 | 6.22 | 6.02 | 6.87 | 6.49 | 6.27 | 6.38 |
| A-MSE (mc) | 5.27 | 5.51 | 5.49 | 5.51 | 5.02 | 4.23 | 4.71 | 4.47 |
| VUL | 3.86 | 4.88 | 5.38 | 5.35 | 3.48 | 4.66 | 5.19 | 5.21 |
| StARS | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| eBIC | 5.53 | 4.40 | 3.25 | 2.38 | 5.83 | 5.58 | 2.82 | 2.00 |
| | | | | dimension p=500 | | | | |
| PC | 3.70 | 3.78 | 4.58 | 4.62 | 3.68 | 3.88 | 5.67 | 5.47 |
| AGNES | 2.00 | 2.00 | 2.01 | 2.34 | 2.00 | 2.00 | 2.05 | 3.44 |
| A-MSE (sub) | 6.92 | 6.77 | 6.47 | 6.21 | 6.83 | 6.70 | 6.53 | 6.53 |
| A-MSE (mc) | 5.19 | 5.46 | 5.57 | 5.45 | 5.08 | 4.44 | 4.47 | 4.47 |
| VUL | 3.41 | 4.04 | 5.01 | 5.72 | 3.52 | 4.23 | 5.18 | 5.09 |
| StARS | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| eBIC | 5.78 | 4.95 | 3.37 | 2.66 | 5.89 | 5.75 | 3.09 | 2.00 |

and tiger. Glasso usually has lower FPR than the other two methods for large TPR levels. When the TPR is small the rates are similar by the three methods even though MB and tiger result to give a slightly better compromise between true and false edges than glasso. There are not big differences with regards to the two graph topologies, power law and hub-based networks.

In order to quantify how well the tuning parameter selection algorithms recover the non-zero elements in $\Omega$, we compare the true discovery rate (TDR), which can be defined by $TDR = TP/(TP + FP)$ with $TP$ and $FP$ expressions given at eq. (5.21), for each of the estimated networks. In Figure 5.4, we show the average TDR in the 60 simulations for all considered combinations of $n$ and $p$. The TDR turns out to be fairly stable with respect to $n$ for A-MSE, PC and VUL. For AGNES, the TDR increases with $n$ (especially in the power-law scenarios), whereas, for eBIC, this goes down rapidly with $n$. In this analysis we can see the limitations of the eBIC method whose main goal is not the graph recovery of $\Omega$. The eBIC selections go from selecting very sparse graphs with more TP than FP when $n$ is small to selecting much denser graphs with many more FP than TP when $n$ is large.

### 5.5.5 AGNES and A-MSE against oracle tuning parameters

The AGNES regularization parameter selection is considered as initial graph to estimate $\lambda$ by A-MSE in Section 5.4.2. We argue that AGNES produces desired global network characteristics. This is shown here using 60 simulated data sets with $n = 50, 100, 200, 500$, $p = 70, 120, 290$ and graph structure

**Table 5.3.** Average ranks for the mean square error of the dissimilarity matrix using several sample sizes, dimension and network topologies (hub-based and power law). A-MSE tends to be the method with the best rates (lowest ranks). eBIC does well for small sample sizes but fails when the sample size increases.

| | Hub-based | | | | Power law | | | |
|---|---|---|---|---|---|---|---|---|
| n | 50 | 100 | 200 | 500 | 50 | 100 | 200 | 500 |
| dimension p=50 | | | | | | | | |
| PC | 3.56 | 3.22 | 2.37 | 2.40 | 3.83 | 3.32 | 3.32 | 3.68 |
| AGNES | 5.53 | 5.56 | 4.94 | 4.70 | 4.89 | 4.62 | 3.55 | 2.92 |
| A-MSE (sub) | **2.14** | **1.74** | **1.82** | **1.83** | **2.51** | **2.66** | **1.97** | **2.17** |
| A-MSE (mc) | 2.44 | 2.02 | 1.96 | 1.92 | 2.62 | 2.85 | 2.92 | 3.14 |
| VUL | 4.81 | 4.84 | 4.62 | 4.14 | 4.39 | 3.84 | 3.38 | 3.09 |
| StARS | 6.93 | 7.00 | 7.00 | 6.92 | 6.95 | 6.98 | 6.99 | 6.52 |
| eBIC | 2.58 | 3.62 | 5.30 | 6.08 | 2.81 | 3.73 | 5.88 | 6.48 |
| dimension p=170 | | | | | | | | |
| PC | 3.08 | 3.64 | 2.98 | 3.12 | 3.37 | 3.41 | 2.41 | 3.29 |
| AGNES | 5.89 | 5.99 | 5.96 | 5.12 | 5.78 | 5.86 | 4.95 | 3.17 |
| A-MSE (sub) | 3.14 | **1.91** | **1.82** | **2.04** | 3.16 | **2.47** | **2.05** | **1.91** |
| A-MSE (mc) | **1.97** | 2.14 | 2.33 | 2.49 | **2.03** | 2.98 | 2.82 | 3.35 |
| VUL | 4.02 | 3.88 | 3.15 | 2.36 | 3.78 | 3.72 | 3.13 | 3.27 |
| StARS | 7.00 | 7.00 | 7.00 | 7.00 | 7.00 | 7.00 | 7.00 | 7.00 |
| eBIC | 2.90 | 3.43 | 4.76 | 5.87 | 2.88 | 2.58 | 5.64 | 6.00 |
| dimension p=290 | | | | | | | | |
| PC | 3.33 | 3.85 | 3.36 | 2.88 | 2.88 | 3.30 | 2.51 | 2.75 |
| AGNES | 5.87 | 5.92 | 5.99 | 5.37 | 5.97 | 5.88 | 5.42 | 3.76 |
| A-MSE (sub) | 3.39 | 2.46 | **1.70** | **1.92** | 4.06 | 2.91 | **1.67** | **1.64** |
| A-MSE (mc) | **1.77** | **2.08** | 2.33 | 2.42 | **2.00** | 2.82 | 3.04 | 3.24 |
| VUL | 4.19 | 3.60 | 2.89 | 2.78 | 3.21 | 3.89 | 3.51 | 3.61 |
| StARS | 7.00 | 7.00 | 7.00 | 7.00 | 7.00 | 7.00 | 7.00 | 7.00 |
| eBIC | 2.45 | 3.08 | 4.73 | 5.62 | 2.88 | **2.21** | 4.85 | 6.00 |
| dimension p=500 | | | | | | | | |
| PC | 2.58 | 3.67 | 3.40 | 3.34 | 3.40 | 2.79 | 2.30 | 2.43 |
| AGNES | 5.95 | 5.93 | 5.99 | 5.66 | 6.00 | 5.85 | 5.90 | 4.28 |
| A-MSE (sub) | 4.45 | 2.62 | **1.47** | **1.74** | 3.44 | 3.40 | **1.73** | **1.43** |
| A-MSE (mc) | **1.89** | **1.92** | 2.30 | 2.52 | **1.68** | **2.46** | 3.12 | 3.35 |
| VUL | 3.38 | 4.41 | 3.23 | 2.40 | 3.88 | 4.03 | 3.18 | 3.51 |
| StARS | 7.00 | 7.00 | 7.00 | 7.00 | 7.00 | 7.00 | 7.00 | 7.00 |
| eBIC | 2.75 | 2.45 | 4.62 | 5.34 | 2.61 | 2.47 | 4.76 | 6.00 |

generated by a power-law distribution as defined in Section 5.2.2. We compare the oracle $\lambda$ solution that minimizes

$$ADR(\lambda) = \mathbf{E}\{\sum_{i>j}(d_{ij} - \hat{d}_{ij}(\lambda))\}^2. \tag{5.22}$$

against the selected $\lambda$ by AGNES. Figure 5.5(a), Figure 5.5(b) and Figure 5.5(c) present the boxplots with the $\lambda_{ag} - \lambda_{oracle}$ differences for all combinations of $n$ and $p$. The differences are close to zero, especially for $p = 120$ and $p = 290$. Thus, the AGNES estimated graph structure provides a good representation of the global network characteristic in eq. (5.22), at least for this set of simulated data.

Consider the local oracle solution for the regularization parameter that minimizes

$$DR(\lambda) = \mathbf{E}\{\sum_{i>j}(d_{ij} - \hat{d}_{ij}(\lambda))^2\}. \tag{5.23}$$

The A-MSE selections, see Figure 5.5(d), Figure 5.5(e) and Figure 5.5(f), are reasonably close to the oracle $\lambda$, especially for $n > 50$, and in all cases, the oracle value of $\lambda$ is within the 95% confidence interval for the median of $\lambda_{AMSE}$. Although here the expected value of the A-MSE risk function is estimated by Monte Carlo, similar results are found using sub-sampling estimates.

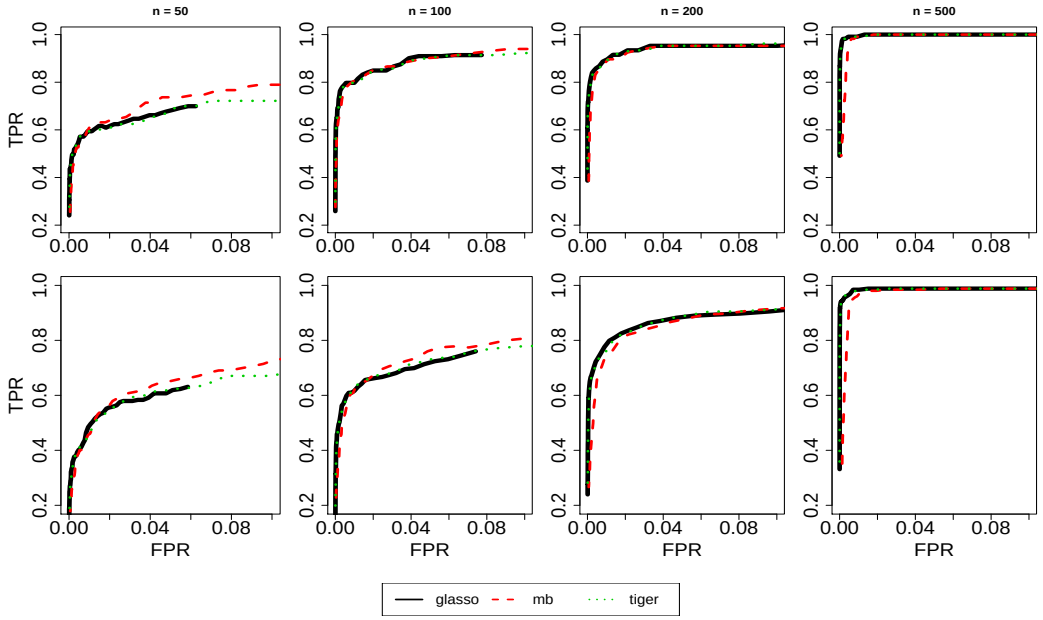



**Figure 5.3.** ROC curves for graph recovery using graphical lasso estimators (glasso, mb and tiger). Hub-based scenarios are on the top figures whereas power-law scenarios are on the bottom figures. The dimension $p = 290$ for all cases.

### 5.5.6 Summary

In our simulations, A-MSE turned out to be the approach with the best estimates of the graph structure dissimilarity matrix as can be seen in Table 5.3. eBIC is also competitive when $n$ is small, but it is not reliable when analyzing larger sample sizes. PC is computationally the fastest method and only does slightly worse than A-MSE in Table 5.3. Moreover, it generally obtains simple graph structures which result in comprehensible connectivity interpretations. The AGNES procedure is usually over-performed by the augmented version A-MSE for small $n$. For large $n$, AGNES and A-MSE have similar $\lambda$ selections with AGNES being significantly faster than A-MSE. StARS (using its default tuning parameters) produces dense graph estimations and achieves the best results when minimizing the mean square error of $\Omega$. Nevertheless, it fails to obtain interpretable network structures due to poor graph recovery.

## 5.6 Application to colon cancer gene expression data

We apply the methods to a case study of genomic data which contain the gene expression profile of 154 colorectal tumor samples and 17,617 genes. The data are generated by the TCGA Research Network: `http://cancergenome.nih.gov/`, and are currently available at the portal `https://gdc-portal.nci.nih.gov/`, under the TCGA cancer program and the Colon Adenocarcinoma disease type.

A reduction on the variable space is applied so that we only keep the most highly correlated genes. We use a filter for the gene's average square correlation with threshold equal to 0.04. Moreover, we add

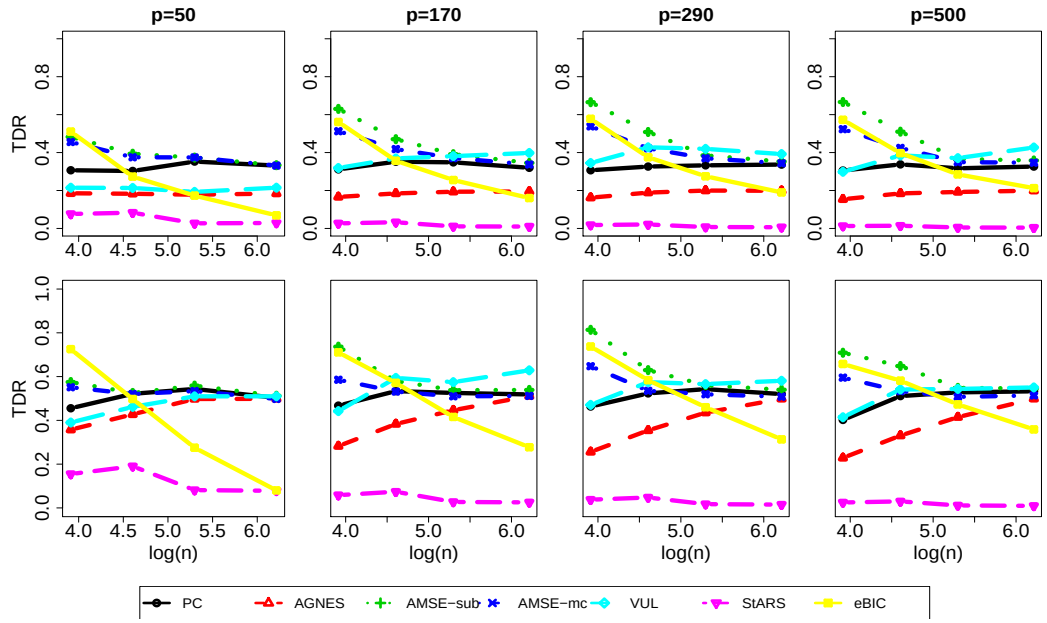
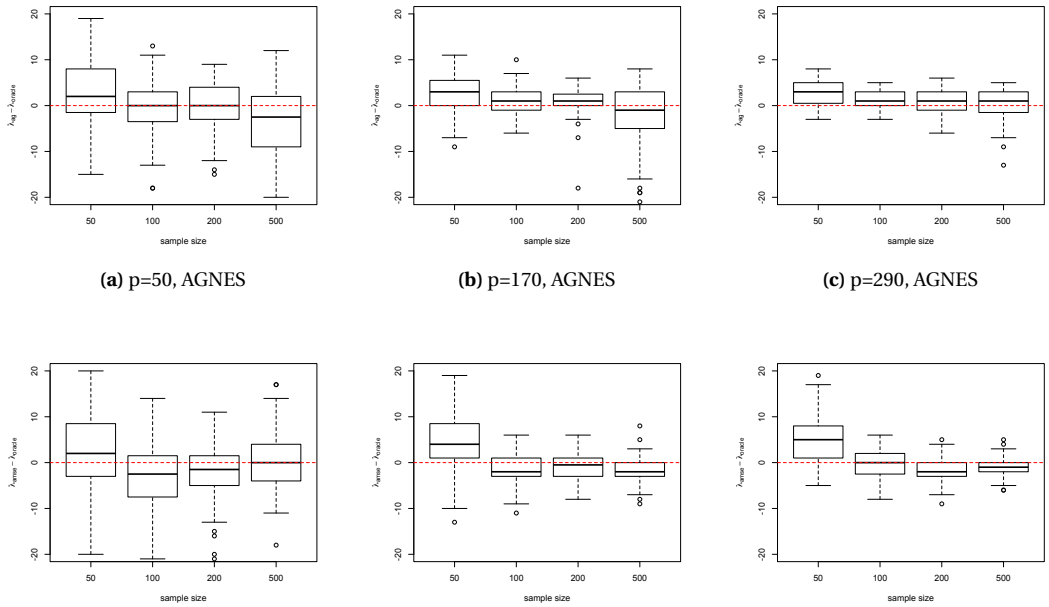



**Figure 5.4.** True discovery rate for all $\lambda$ selection approaches and all combinations of $p$ and $n$. The top figures correspond to hub-based networks and the bottom figures are the power-law networks. The x-axis scale is $n : \log(n)$. eBIC rates decrease with the sample size whereas AGNES, A-MSE, VUL and PC rates slightly increase with the sample size.

the non-filtered genes which have at least one correlation coefficient with the filtered genes larger than 0.5. This means a reduction to the 55% of the genes with a total of 9,723 genes left to analyze. We estimate CD graphs via the Neighborhood selection algorithm of Meinshausen and Bühlmann (2006). We compute 90 different graphs given an equidistant sequence of $\lambda$'s between 0.35 and 0.80. Values of $\lambda$ lower than 0.35 produce almost-fully connected graphs and values above 0.80 produce zero edges in the graph. We use the PC and A-MSE approaches to select one particular graph with $\lambda_{pc} = 0.69$ and $\lambda_{amse} = 0.55$. The graphical representation of the two underlying networks is presented in Figure 5.6. The graph by PC, with 4,819 edges, shows a simpler structure compared to A-MSE, with 19,986 edges.

We separate the graphs in different clusters by applying a Partitioning Around Medoids (Reynolds et al., 2006) on the shortest distance matrix. We choose the number of clusters manually by considering the largest rate of change in the within-subject and between-subject variation such that the PC graph structure contains 15 clusters and the A-MSE contains 18 clusters. To assess which biological processes may be linked to the clusters, we download 1,320 gene sets from the MSig database (Subramanian et al., 2005), which represent canonical pathways compiled from two sources: KeGG (Kanehisa et al., 2016) and Reactome (Milacic et al., 2012). For each pathway we test for a significant over-representation in a cluster by using Fisher's exact test applied to the $2 \times 2$-table defined by pathway and cluster membership with a Bonferroni correction for multiple testing. Note that we use the reduced selection of 9,723 genes here as "background", i.e. the analysis corrects for any over-representation of a pathway in that selection.




**(a)** p=50, AGNES **(b)** p=170, AGNES **(c)** p=290, AGNES

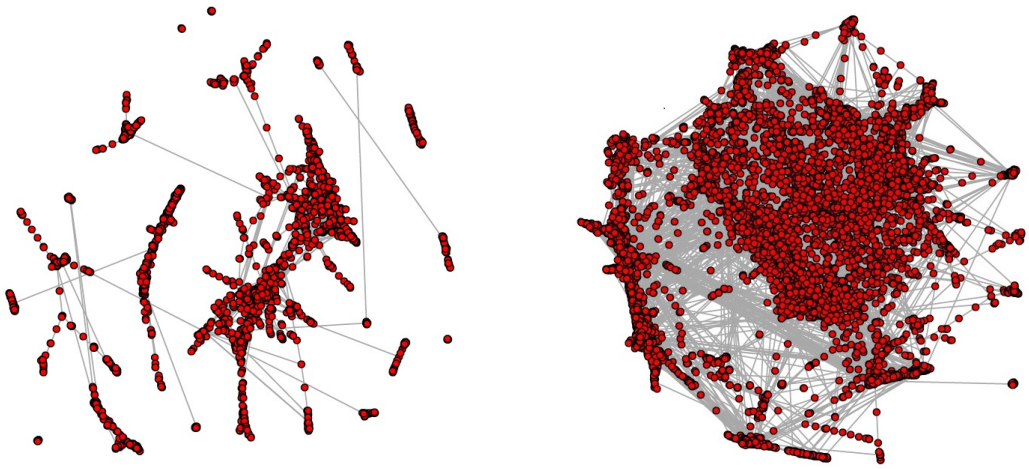
**(d)** p=50, A-MSE **(e)** p=170, A-MSE **(f)** p=290, A-MSE

**Figure 5.5.** AGNES $\lambda$ selections against the oracle best $\lambda$ for the mean square errors of average dissimilarities, see eq (5.22). A-MSE $\lambda$ selections against the oracle best $\lambda$ for mean square errors given local dissimilarities, see eq (5.23).

For the PC and A-MSE selected graphs, respectively, 6 out of 15 clusters of genes, and 7 out of 18 clusters of genes, overlap significantly with at least one pathway gene set (at 0.01 significant level). Besides, a total of 160 and 122 pathway sets (out of 1.320) present significant overlap with clusters of genes defined in the PC and A-MSE graphs. Among the significant lists, PLK1, NFAT, DNA replication or adaptive immune system are pathways associated with tumor cells.

## 5.7 Discussion

This chapter studies the problem of choosing the regularization parameter $\lambda$ for Gaussian graphical models in high-dimensional data assuming we have high level knowledge about the nature of the graph structures, namely strong clustering of gene expression data (e.g., Eisen and Spellman, 1998). The methods we introduce in this chapter take this assumption into account by selecting $\lambda$ so that statistics measuring the degree of clustering (AGNES, A-MSE) or connectivity (PC, VUL) are optimized. We aim to select the sparsest graph such that the real cluster structure is maintained and at the same time it contains a good tradeoff between true and false positive edges. The proposed approaches to select the regularization parameter provide competitive results in a relatively fast computational speed. They present more reliable results than the StARS approach which tends to overestimate the network size. The StARS method accounts for the stability of the estimated graphs and has been



**(a)** PC selected graph. **(b)** A-MSE selected graph.

**Figure 5.6.** Selected graphs by PC and A-MSE to describe conditional gene associations on colon cancer gene expression data. The A-MSE graph is denser than the PC graph but in both cases several clusters of genes are visible.

proven to work well in Liu et al. (2010). It depends, however, on another parameter which controls the maximum amount of variability in the graph. There is no straightforward choice for this parameter and our simulation study shows that using the default value of 0.05 StARS yields uninformative networks with a majority of edges being false positives.

The path connectivity approach introduced here provides a good compromise between structure and false positive edges. The main characteristic of this approach is that it relies on the shortest distance between all pairs of nodes. Interestingly, this quantity tends to show a clear changepoint when studied as a function of $\lambda$, at which the structure of the graph changes radically. It typically produces very informative graphs in all the tested simulated datasets and gives competitive results for the mean square error between dissimilarity matrices as discussed in Section 5.5.2. In the gene expression data set it also provides us with a clearly structured informative graph. PC gives an excellent first choice of $\lambda$ without additional prior information if we want to find an easily interpretable graph.

The A-MSE, with initial graph structure given by the AGNES selected graph, is the best of all the approaches in terms of minimizing the MSE between the true distances and the estimated ones in the simulated data. Also, $\lambda_{amse}$ is generally smaller than $\lambda_{ac}$ leading to less complex graphs than the ones estimated by AGNES. This is a desirable property as we assume only a small proportion of non-zero elements in $\Omega$ and thus with increasing graph density the number of false positive edges grows much faster than the number of true positives and can make the graph become quite inaccurate. However, if the aim is to have fewer false negatives, that is, that as many as possible true edges are included at the expense of a higher number of false positives, then algorithms like AGNES and StARS are more appropriate.

The analysis of the gene expression data underlines some interesting results. The obtained graphs

present a cluster-based structure as we can see in Figure 5.6. Our two new approaches of choosing a regularization parameter, path connectivity and A-MSE, lead to sparse and clustered networks that are easy to interpret. Closer investigation of the results shows that the clusters overlap significantly with a number of pre-defined gene sets and regulatory pathways which indicates that our assumption of a sparse clustered structure rises some biologically meaningful results.

In conclusion, we find that approaches such as PC, A-MSE, AGNES and VUL, which use network characteristics for parameter selection, can be beneficial in estimating sparse partial correlation matrices (and graph structures) for high-dimensional biological data. While maintaining good statistical properties in terms of false discovery rates and mean square error, the results tend to be easier to interpret in terms of network structure and thus are more useful in applications compared to parameter selection methods purely based on mathematical/statistical measures such as AIC or BIC.

# Chapter 6

# Joint estimation of conditional dependence structures

## 6.1 Introduction and motivation

Genomic data produced by high-throughput technology are nowadays easy to collect and store generating many statistical questions. The statistical estimation problem we study in this chapter is motivated by the same type of data we considered in Chapter 4 for hypothesis testing: we want to analyze datasets where genomic profiles are obtained for individuals in two different classes. For instance, we consider two case studies, which were already presented in Chapter 4, that consist of patients with psoriasis vulgaris disease and patients with lung cancer, respectively. In both datasets, there is the genomic profile of more than 19,000 genes for a paired lesional (or tumor) and healthy tissues. The third case study we explore contains the gene expression and methylation profiling of 25 patients with colon cancer in which two samples, one for a colorectal tumor and one for its healthy adjacent colonic tissue, are obtained for every individual. In total, there are more than 24,000 genes and more than 27,000 methyl sites.

The main challenge in the analysis of these data is to understand how genes interact between each other in a cell as well as to detect which groups of gene connections vary from a healthy to a non-healthy state. This can be formulated by an estimation problem of sparse conditional dependence (CD) networks which, under the Gaussian assumption, are fully characterized by their underlying precision matrices. The estimating of precision matrices when data are high-dimensional (dimension is larger than the sample size) represents a challenge as maximum likelihood estimators are no longer suitable (Pourahmadi, 2007). Methods that address this issue to estimate a single precision matrix include sparsity-penalization approaches known as graphical lasso which are extensively investigated in Chapter 5. A natural extension is applied to jointly estimated multiple precision matrices by using an additional penalization term that encourages the similarity between such matrices. For instance,

Guo et al. (2011) use a group-lasso penalization (GGL) or Danaher et al. (2014) incorporate a fused-lasso penalization option (FGL). The FGL method yields better graph recovery rates than estimating the matrices separately when these are expected to be similar. However, it is designed under the assumption of subject independence in the datasets.

Motivated by real data, here we study the probabilistic interpretation of the algorithm proposed by Danaher et al. (2014) when data are paired, with the aim to determine which gene associations are or are not common between two populations (e.g., given by two medical conditions) and relate the changes to cellular biological processes. We end up proposing a new weighted fused-penalty for the estimation of marginal conditional dependence structures (WFGL) that accounts for correlation in the estimators when data are paired. Our analysis shows that the current joint estimation algorithm, for both FGL and WFGL, overestimates triangular motifs structures, so as second contribution, we present a method based on hypothesis testing to correct for this issue.

In a similar framework, we develop a method to estimate joint regression coefficient matrices when data are high-dimensional and possibly paired. This is encouraged by the colon cancer data (Hinoue et al., 2012), where 4 different datasets are available: methylation for healthy and tumor samples, and gene expression for healthy and tumor samples. Gadaleta and Bessonov (2015) previously integrated gene expression and methylation presence for a dataset with 215 individuals affected with glioblastoma cancer. The authors find two networks using lasso-based estimators: the non-zero structure of the regression coefficients using gene expression as response variables and methylation presence as explanatory variables; and the non-zero structure of the precision matrix (using only gene expression data). Here, we take advantage of having both tumor and healthy samples to jointly estimate the regression coefficients as well as the gene expression network using fused lasso penalized marginal likelihood estimators. The analysis of these data is presented separately in Chapter 8.

The chapter is structured as follows. In Section 6.2 we propose a weighted fused graphical lasso algorithm to estimate joint precision matrices. In the following Section 6.3 we present the analogous algorithm to estimate multiple regression coefficient matrices. In Section 6.4 we discuss the issues on overestimating triangular motifs. In Section 6.5 we illustrate the performance of the methods for simulated datasets given different correlation structures, dimension and sample sizes. Finally, in Section 6.6 we estimate CD structures for the motivating applications to gene expression data.

## 6.2 Weighted fused graphical lasso

### 6.2.1 Fused graphical lasso: assumptions and marginal estimator

Consider the problem setting described in Section 2.2 where $n$ i.i.d. $2p$-dimensional random vectors $(Y_k^{(1)}, Y_k^{(2)})$ are observed, with $[Y_k^{(1)}, Y_k^{(2)}] \sim N_{2p}(0, \Omega^{-1})$, $k = 1, \ldots, n$. The matrix $\Omega$ represents the joint

CD structure for $Y^{(1)}$ and $Y^{(2)}$, and it is defined by

$$\Omega = R^{-1} = \begin{bmatrix} R_1 & R_{12} \\ R_{21} & R_2 \end{bmatrix}^{-1} = \begin{bmatrix} \Omega_1^J & \Omega_{12}^J \\ \Omega_{21}^J & \Omega_2^J \end{bmatrix}. \tag{6.1}$$

Danaher et al. (2014) assume independence between observations in the two conditions, see eq. (2.5), where $R_{12} = \Omega_{12}^J = 0$, so $\Omega_1 = R_1^{-1} = \Omega_1^J$ and $\Omega_2 = R_2^{-1} = \Omega_2^J$, and propose the fused graphical lasso (FGL) estimator of $\Omega$

$$\hat{\Omega}_{FGL}^{\lambda} = \arg\max_{\Omega_1,\Omega_2} \left[ \sum_{m=1,2} \log\det\Omega_m - \mathrm{tr}(\Omega_m S_m) - P_{\Lambda_1,\Lambda_2}(\Omega_1,\Omega_2) \right], \tag{6.2}$$

with

$$P_{\Lambda_1,\Lambda_2}(\Omega_1,\Omega_2) = ||\Lambda_1 \circ \Omega_1||_1 + ||\Lambda_1 \circ \Omega_2||_1 + ||\Lambda_2 \circ (\Omega_2 - \Omega_1)||_1, \tag{6.3}$$

where $A \circ B$ is the elementwise product of matrices $A$ and $B$, $\Lambda_1 = [\lambda_{ij}^{(1)}]$ is a $p \times p$ matrix with the sparsity tuning parameters and $\Lambda_2 = [\lambda_{ij}^{(2)}]$ is a $p \times p$ matrix with the similarity tuning parameters. The maximization problem in (6.2) and (6.3) is solved by optimizing its Lagrangian formulation

$$L_\rho = -[\sum_{m=1,2} \log\det\Omega_m - \mathrm{tr}(\Omega_m S_m) + P_{\Lambda_1,\Lambda_2}(A_1,A_2) + \frac{\rho}{2} \sum_{m=1,2} ||\Omega_m - A_m + U_m||_F^2],$$

using the ADMM-type algorithm (Boyd, 2010) described in Algorithm 8. Here, $U_m$ are dual variables, $A_m$ corresponds to $\Omega_m$ and $\rho$ is a positive constant that is used as a regularization parameter with default value equal to 1.

Consider now that the independence assumption does not hold, e.g., paired data setting, and so that $\Omega_{12}^J \neq 0$. The marginal estimators $\hat{\Omega}_1 = [\hat{\Omega}_{ij}^{(1)}]$ and $\hat{\Omega}_2 = [\hat{\Omega}_{ij}^{(2)}]$, being the solution of eq. (6.2) (i.e., step 3 in the ADMM algorithm), are correlated for some pair of variables $(i, j)$ (Steiger, 1980). In the following section we develop the probabilistic interpretation of Algorithm 8, and show that this could be used, even when data are paired, to estimate the marginal conditional dependence structures $\Omega_1 = R_1^{-1}$ and $\Omega_2 = R_2^{-1}$ by considering distinct penalties within matrices $\Lambda_1$ and $\Lambda_2$. We should remark that this method does not find an estimator for the conditional dependence structures $\Omega_1^J$ and $\Omega_2^J$. The precision matrix $\Omega_1^J$ measures linear dependence of $Y^{(1)}$ conditionally on both $Y^{(1)}$ and $Y^{(2)}$ whereas $\Omega_1$ ignores dependence between $Y^{(1)}$ and $Y^{(2)}$ and finds the marginal conditional dependence of $Y^{(1)}$ instead (and similarly for $Y^{(2)}$). We find quite useful to characterize the marginal conditional dependence in our motivating data as the interest is not in understanding gene relationships between tissues, but only the comparison of gene relationships in tumor and healthy populations separately.

Depending on the mathematical model that we assume that generates the data, using marginal estimations may induce some spurious coefficients. From the four models proposed in Section 2.2, independence model, additive model and multiplicative model would not suffer this phenomena too much. For example in the multiplicative model, we assume that the correlation matrices of $Y^{(1)}$

**Algorithm 8** Fused Graphical Lasso

1: Input: $\Lambda_1, \Lambda_2, \rho$

2: Initialization: set iteration $t = 0$, $U_m^{(t)} = 0$ and $\hat{S}_m^{(t)} = S_m$ corresponding to the sample covariance matrix for $m = 1, 2$. Repeat 3-5 until convergence.

3: Find a dense estimator of $\hat{\Omega}_m^{(t)}$ using a quadratic regularized inverse of matrix $\hat{S}_m^{(t)}$ (Witten et al., 2009). Given the eigenvalue decomposition of $\hat{S}_m^{(t)} = V_m^{(t)} D_m^{(t)} V_m^{'(t)}$, the inverse is found by

$$\hat{\Omega}_m^{(t)} = V_m^{(t)} \tilde{D}_m^{(t)} V_m^{'(t)}, \qquad \tilde{D}_{m_{jj}}^{(t)} = \frac{n}{2\rho}\left(-D_{m_{jj}}^{(t)} + \sqrt{(D_{m_{jj}}^{(t)})^2 + 4\rho/n}\right), \tag{6.4}$$

4: Find $[\hat{A}_1^{(t)}, \hat{A}_2^{(t)}]$ by minimizing $\frac{\rho}{2}\sum_{m=1,2} ||A_m - (\hat{\Omega}_m^{(t)} + U_m^{(t)})||_2^F + P_{\Lambda_1,\Lambda_2}(A_1, A_2)$ using a thresholding approach: given $\{\hat{A}_m^{'(t)} = \hat{\Omega}_m^{(t)} + U_m^{(t)}\}_{m=1,2}$, set equal precision matrix elements if the absolute value of the estimated differences are smaller than the corresponding elements of $[\lambda_{ij}^{(2)}/\rho]$:

$$[\hat{A}_{1_{ij}}^{''(t)}, \hat{A}_{2_{ij}}^{''(t)}] = \begin{cases} [.5(\hat{A}_{1_{ij}}^{'(t)} + \hat{A}_{2_{ij}}^{'(t)}), .5(\hat{A}_{1_{ij}}^{'(t)} + \hat{A}_{2_{ij}}^{'(t)})] & \text{if } |\hat{A}_{1_{ij}}^{'(t)} - \hat{A}_{2_{ij}}^{'(t)}| \leq \lambda_{ij}^{(2)}/\rho; \\ [\hat{A}_{1_{ij}}^{'(t)} + \lambda_{ij}^{(2)}(2\rho), \hat{A}_{2_{ij}}^{'(t)} - \lambda_{ij}^{(2)}/(2\rho)] & \text{if } \hat{A}_{1_{ij}}^{'(t)} - \hat{A}_{2_{ij}}^{'(t)} > \lambda_{ij}^{(2)}/\rho; \\ [\hat{A}_{1_{ij}}^{'(t)} - \lambda_{ij}^{(2)}/(2\rho), \hat{A}_{2_{ij}}^{'(t)} + \lambda_{ij}^{(2)}(2\rho)] & \text{if } \hat{A}_{1_{ij}}^{'(t)} - \hat{A}_{2_{ij}}^{'(t)} < -\lambda_{ij}^{(2)}\rho; \end{cases} \tag{6.5}$$

[Notation equivalence $\hat{A}_{k_{ij}}^{'(t)} = (\hat{A}_k^{'(t)})_{ij}$]. Then, set elements in $[\hat{A}_1^{''(t)}, \hat{A}_2^{''(t)}]$ to zero by soft-thresholding with threshold given by $\Lambda_1$: $\hat{A}_{m_{ij}}^{(t)} = \text{sign}(\hat{A}_{m_{ij}}^{''(t)})\left(|\hat{A}_{m_{ij}}^{''(t)}| - \lambda_{ij}^{(1)}\right)_+$, $m = 1, 2$.

5: Set $t = t + 1$. Update $U_m^{(t)} = U_m^{(t-1)} + (\hat{\Omega}_m^{(t-1)} - \hat{A}_m^{(t-1)})$ and $\hat{S}_m^{(t)} = S_m - \frac{\rho}{n}\hat{A}_m^{(t-1)} + \frac{\rho}{n}U_m^{(t)}$ for $m = 1, 2$. Stop if convergence.

6: Output: $\hat{\Omega}_1 = \hat{A}_1^{(t-1)}$, $\hat{\Omega}_2 = \hat{A}_2^{(t-1)}$ and $\hat{\Omega}_d = \hat{A}_2^{(t-1)} - \hat{A}_1^{(t-1)}$.

and $Y^{(2)}$, $\text{cor}(Y^{(1)}) = R_1$ and $\text{cor}(Y^{(2)}) = R_2$ respectively, do not depend on the specification of the paired component $\Delta$, so marginal conditional dependence matrices $\Omega_1$ and $\Omega_2$ would coincide to the scenario where observations in the two populations are independent. In the direct effect model though, for instance $(\Omega_1^J)_{ij} = 0$ does not ensure that the marginal $\Omega_{ij}^{(1)} = 0$ unless either $(\Omega_{12}^J)_{ii} = 0$ or $(\Omega_{12}^J)_{jj} = 0$. Understanding these limitations, in this chapter we only consider the estimation problem of marginal conditional dependence structures, leaving the estimation problem of $\Omega_1^J$ and $\Omega_2^J$ as future work.

### 6.2.2 Monotoring error rates and weighted fused graphical lasso

The joint estimation problems described in Section 6.2 and Section 6.3 require the selection of two regularization parameters: $\lambda_1$ (sparsity) and $\lambda_2$ (similarity), and the combination of the two characterizes the estimated network sizes (both common network and differential network). In Chapter 5 we discuss different ways of choosing sparsity penalization parameters that encourage certain network characteristics, i.e., clustering structure or connectivity of the estimated networks. These could also be applied for the joint estimation algorithm once the parameter $\lambda_2$ is fixed. In this section, an alternative procedure is proposed, though, by choosing $\lambda_1$ and $\lambda_2$ to control the expected proportion of false positive edges (EFPR) at level $\alpha_1$ for both the individual matrices and the difference matrix. This is possible to do directly (without resampling) and fast due to the nature of the ADMM recursive algorithm presented in Section 6.2.1, that, for every iteration, obtain a dense

estimation of the precision matrices before thresholding. By having a dense matrix, the distribution of estimated coefficients whose true values are zero can be approximated. In contrast, other graphical lasso algorithms threshold the coefficients row by row using a regression based approach (Friedman et al., 2007), and the EFPR is commonly controlled using subsampling methods (Meinshausen and Bühlman, 2010), which greatly increase the computational cost.

Define the sets $S_m = \{(i,j),\, i<j : \Omega_{ij}^{(m)} = 0\}$ for $m = 1,2$, $S_0 = \{(i,j),\, i<j : \Omega_{ij}^{(1)} = \Omega_{ij}^{(2)} = 0\}$ and $S_0' = \{(i,j),\, i<j : \Omega_{ij}^{(1)} = \Omega_{ij}^{(2)},\, \Omega_{ij}^{(2)} \neq 0\}$. For a set $S$, denote $|S| = \text{Card}(S)$. Let $d_{ij}^{(1)} = I(\hat{\Omega}_{ij}^{(1)} \neq 0)$, $d_{ij}^{(2)} = I(\hat{\Omega}_{ij}^{(2)} \neq 0)$ and $d_{ij}^{(D)} = I(\hat{\Omega}_{ij}^{(2)} - \hat{\Omega}_{ij}^{(1)} \neq 0)$ determine the estimated graph structures. The objective is to control the error rates

$$\begin{cases} \alpha_1 = |S_0|^{-1} \sum_{(i,j)\in S_0} \sum_{m\in\{1,2\}} \Pr(d_{ij}^{(m)} \neq 0)/2, \\ \alpha_2 = |S_0|^{-1} \sum_{(i,j)\in S_0} \Pr(d_{ij}^{(D)} \neq 0). \end{cases} \tag{6.6}$$

For the difference matrix, ideally we would like to set $\alpha_2 = |S_0 \cup S_0'|^{-1} \sum_{(i,j)\in S_0\cup S_0'} \Pr(d_{ij}^{(D)} \neq 0)$, but since the distribution of the estimators under $S_0'$ depends on the true unknown values of $\Omega_{ij}^{(1)}$ and $\Omega_{ij}^{(2)}$, estimation of $\sum_{(i,j)\in S_0'} \Pr(\hat{\Omega}_{ij}^{(2)} \neq \hat{\Omega}_{ij}^{(1)})$ is challenging. In terms of $\alpha_1$, we would like to distinguish between $\alpha_{11} = |S_1|^{-1} \sum_{(i,j)\in S_1} \Pr(d_{ij}^{(1)} \neq 0)$ and $\alpha_{12} = |S_2|^{-1} \sum_{(i,j)\in S_2} \Pr(d_{ij}^{(2)} \neq 0)$, but these depend on cases where $\Omega_{ij}^{(1)} = 0$ & $\Omega_{ij}^{(2)} \neq 0$ and $\Omega_{ij}^{(1)} \neq 0$ & $\Omega_{ij}^{(2)} = 0$, respectively, and their estimation present similar problems as for $\alpha_2$. Therefore, we will control the simpler rates represented by elements only in $S_0$ instead.

To estimate the error rates defined in eq. (6.6), we will use intermediate steps in Algorithm 8, particularly dense estimators of the precision matrices. In Algorithm 8 at iteration $t$,

- $\hat{\Omega}_{ij}^{(m)} = 0$ if $|\hat{A}_{m_{ij}}^{''(t)}| \leq \lambda_{ij}^{(1)}$, $m \in \{1,2\}$, hence $\Pr(d_{ij}^{(m)} \neq 0) = \Pr(|\hat{A}_{m_{ij}}^{'(t)}| > \lambda_{ij}^{(1)})$;
- $\hat{\Omega}_{ij}^{(2)} - \hat{\Omega}_{ij}^{(1)} = 0$ if either $|\hat{A}_{D_{ij}}^{'(t)}| \leq \lambda_{ij}^{(2)}$ or $\{|\hat{A}_{D_{ij}}^{'(t)}| > \lambda_{ij}^{(2)}$ and $|\hat{A}_{1_{ij}}^{''(t)}| \leq \lambda_{ij}^{(1)}$ and $|\hat{A}_{2_{ij}}^{''(t)}| \leq \lambda_{ij}^{(1)}\}$, hence

$$\Pr(d_{ij}^{(D)} = 0) = \Pr(|\hat{A}_{D_{ij}}^{'(t)}| \leq \lambda_{ij}^{(2)}) + \Pr(|\hat{A}_{D_{ij}}^{'(t)}| > \lambda_{ij}^{(2)} \,\&\, |\hat{A}_{1_{ij}}^{''(t)}| \leq \lambda_{ij}^{(1)} \,\&\, |\hat{A}_{2_{ij}}^{''(t)}| \leq \lambda_{ij}^{(1)})$$

  and $\Pr(d_{ij}^{(D)} \neq 0) = 1 - \Pr(d_{ij}^{(D)} = 0)$, where $\hat{A}_{D_{ij}}^{'(t)} = \hat{A}_{2_{ij}}^{'(t)} - \hat{A}_{1_{ij}}^{'(t)}$

(see step 4 in the algorithm for the definitions of $\hat{A}_{m_{ij}}^{'(t)}$ and $\hat{A}_{m_{ij}}^{''(t)}$). Note that there are two possible ways we can arrive at $\hat{\Omega}_{ij}^{(2)} - \hat{\Omega}_{ij}^{(1)} = 0$, since there are two thresholding steps in the algorithm.

To simplify the notation, we denote $Q_{ij} = \hat{A}_{1_{ij}}^{'(t)}$ and $Z_{ij} = \hat{A}_{2_{ij}}^{'(t)}$. For $(i,j) \in S_0$, we assume that the majority of the pairs $(Q_{ij}, Z_{ij})$ follow a bivariate normal distribution with the following covariance matrix

$$Cov(Q_{ij}, Z_{ij}) = \Sigma_{ls} = \begin{bmatrix} \sigma_{Q_{ij}}^2 & \psi_{ij}\sigma_{Q_{ij}}\sigma_{Z_{ij}} \\ \psi_{ij}\sigma_{Q_{ij}}\sigma_{Z_{ij}} & \sigma_{Z_{ij}}^2 \end{bmatrix}, \tag{6.7}$$

where the correlation between $Q_{ij}$ and $Z_{ij}$ is denoted by $\psi_{ij}$. The assumption of normality is checked for the real data application in Appendix (B.4). To approximate the rates in eq. (6.6), below we assume

that $\sigma^2_{Q_{ij}} = \sigma^2_{Z_{ij}} = \sigma^2$. However, if $\sigma^2_{Q_{ij}} \neq \sigma^2_{Z_{ij}}$, then, replacing $\sigma$ by $\{(\sigma^2_{Q_{ij}} + \sigma^2_{Z_{ij}})/2\}^{1/2}$ and $\psi_{ij}$ by $2\psi_{ij}\sigma_{Q_{ij}}\sigma_{Z_{ij}}/(\sigma^2_{Q_{ij}} + \sigma^2_{Z_{ij}})$ leads to similar expressions.

Probabilities necessary to work out $\Pr(d^{(m)}_{ij} \neq 0)$ and $\Pr(d^{(D)}_{ij} \neq 0)$ under assumption (6.7) are stated in the following lemma.

**Lemma 6.1.** *For $(Q_{ij}, Z_{ij})$ following a bivariate normal distribution with 0 means, variances $\sigma^2_{Q_{ij}} = \sigma^2_{Z_{ij}} = \sigma^2$ and correlation $\psi_{ij}$,*

$$\Pr(Q_{ij} - Z_{ij} > \lambda^{(2)}_{ij}) = P(Q_{ij} - Z_{ij} < -\lambda^{(2)}_{ij}) = 1 - \Phi(\lambda^{(2)}_{ij}/(\sqrt{2}\sigma(1-\psi_{ij})^{1/2})),$$
$$\Pr(|0.5(Q_{ij} + Z_{ij})| > \lambda^{(1)}_{ij} \,|\, |Q_{ij} - Z_{ij}| \leq \lambda^{(2)}_{ij}) = 2[1 - \Phi(\sqrt{2}\lambda^{(1)}_{ij}(1+\psi_{ij})^{-1/2}/\sigma)],$$

*and for any $a < b$, $c < d \leq b + \lambda^{(2)}_{ij}$,*

$$\Pr(Q_{ij} \in [c,d] \,\&\, Z_{ij} \in [a,b] \,\&\, Q_{ij} - Z_{ij} > \lambda^{(2)}_{ij}) =$$
$$\int_c^d \sigma^{-1}\varphi(x/\sigma)\left[\Phi\left(\frac{x(1-\psi_{ij}) - \lambda^{(2)}_{ij}}{\sigma\,(1-\psi^2_{ij})^{1/2}}\right) - \Phi\left(\frac{a - x\psi_{ij}}{\sigma\,(1-\psi^2_{ij})^{1/2}}\right)\right] dx,$$

*where $\varphi$ and $\Phi$ are the density and the cumulative distribution function of the standard normal distribution.*

**Proof.** *proof is given in Appendix B.1, as well as the derivation of the formulas below.*

**Corollary 6.1.** *Define the weights $v_{ij} = (1-\psi_{ij})^{1/2}$. Following lemma 6.1, we set $\lambda^{(2)}_{ij} = \lambda_2 v_{ij}$, such that the probability of recovering differential edges is independent of the linear relationship between variables in the two datasets, i.e., the initial rate*

$$\alpha'_2 \quad = \quad P(|Q_{ij} - Z_{ij}| > \lambda_2(1-\psi_{ij})^{1/2} \,|\, (i,j) \in S_0) = 2[1 - \Phi(\lambda_2/(\sqrt{2}\sigma))].$$

*is the same for all pairs $(i,j) \in S_0$.*

The proportion of false rejections of the difference being 0 is

$$\alpha_2 \quad = \quad \alpha'_2 - (|S_0|)^{-1} \sum_{(i,j)\in S_0} \Pr(|Q_{ij} - Z_{ij}| > \lambda_2 v_{ij} \,\&\, |\hat{A}^{''(t)}_{1_{ij}}| \leq \lambda^{(1)}_{ij} \,\&\, |\hat{A}^{''(t)}_{2_{ij}}| \leq \lambda^{(1)}_{ij})$$

Denote

$$I_\sigma(\lambda^{(1)}_{ij}, \psi_{ij}, \lambda_2) \quad = \quad \int_{-\lambda^{(1)}_{ij} - v_{ij}\lambda_2/2}^{\lambda^{(1)}_{ij} - v_{ij}\lambda_2/2} \sigma^{-1}\varphi(x/\sigma)\left[\Phi\left(\frac{x(1-\psi_{ij}) - \lambda_2 v_{ij}}{\sigma\,(1-\psi^2_{ij})^{1/2}}\right)\right. \tag{6.8}$$
$$\left. -\Phi\left(\frac{\lambda_2 v_{ij}/2 - \lambda^{(1)}_{ij} - x\psi_{ij}}{\sigma\,(1-\psi^2_{ij})^{1/2}}\right)\right] dx,$$

then we can write

$$\alpha_2 \quad = \quad \alpha_2'(1-(|S_0|)^{-1}\sum_{(i,j)\in S_0} I_\sigma(\lambda_{ij}^{(1)},\psi_{ij},\lambda_2). \tag{6.9}$$

Define the complementary events $B_0=\{(i,j)\in S_0, |Q_{ij}-Z_{ij}|>\lambda_2 v_{ij}\}$ and $B_1=\{(i,j)\in S_0, |Q_{ij}-Z_{ij}|\leq\lambda_2 v_{ij}\}$, so that

$$\lambda_{ij}^{(1)} \quad = \quad \lambda_{1_\sigma}(\alpha_1^*,B_0,\psi_{ij})\mathbb{1}_{(i,j)\in B_0}+\lambda_{1_\sigma}(\alpha_1^{**},B_1,\psi_{ij})\mathbb{1}_{(i,j)\in B_1}, \tag{6.10}$$

where $\lambda_{1_\sigma}(\alpha_1^*,B_0,\psi_{ij})$ and $\lambda_{1_\sigma}(\alpha_1^{**},B_1,\psi_{ij})$ are the solution of

$$\alpha_1^* \quad = \quad 2[1-\Phi(\sqrt{2}\lambda_{1_\sigma}(1+\psi_{ij})^{-1/2}/\sigma)], \tag{6.11}$$

$$\alpha_1^{**} \quad = \quad \int_{|x+\lambda_2 v_{ij}/2|>\lambda_{1_\sigma}} \sigma^{-1}\varphi(x/\sigma)\Phi\left(\frac{-x(1-\psi_{ij})^{1/2}-\lambda_2}{\sigma\,(1+\psi_{ij})^{1/2}}\right)dx, \tag{6.12}$$

respectively. The proportion of false rejections $\alpha_1$ is then given by

$$\alpha_1 \quad = \quad \alpha_1^*(1-\alpha_2')+\alpha_1^{**}\alpha_2'. \tag{6.13}$$

Here we assume that $\alpha_1^*=\alpha_1^{**}=\alpha_1$, therefore, given $\alpha_1$ and $\alpha_2'$, we set $\lambda_2=\sqrt{2}\sigma\Phi^{-1}(1-\alpha_2'/2)$, set $\lambda_{1_\sigma}(\alpha_1,B_0,\psi_{ij})=\sigma\Phi^{-1}(1-\alpha_1/2)(1-\psi_{ij})^{1/2}/\sqrt{2}$, solve numerically eq. (6.13) to obtain $\lambda_{1_\sigma}(\alpha_1,B_1,\psi_{ij})$, and evaluate $\alpha_2$ using (6.9).

In practice, $S_0$, $[\psi_{ij}]$ and $\sigma$ are unknown, then $\alpha_2$ is approximated using all pairs $\{(i,j),\, i<j\}$, $[\psi_{ij}]$ is estimated as proposed in Section 6.2.3, and $\sigma$ is estimated by a robust estimator (Rousseeuw and Croux,1993), i.e., any of the following three estimators could be used: (1) Absolute deviation around the median (50% breakdown point with $|S_0|>p\,(p-1)/4$ needed for consistency), $\hat{\sigma}_x=1.483\text{mad}(x)$, where $\text{mad}(x)=\text{med}(|x_i-\text{med}(x)|)$; (2) Interquartile range (25% breakdown point with $|S_0|>p\,(p-1)/8$), $\hat{\sigma}_x=\text{IQR}(x)/1.349$, where $\text{IQR}(x)=q_{0.75}(x)-q_{0.25}(x)$ with $\alpha$-quanitle $q_\alpha(x)$; (3) Rousseeuw and Croux (RC) mad alternative (50% breakdown point with $|S_0|>p\,(p-1)/4$), $\hat{\sigma}_x=1.1926\text{RCmad}(x)$, where $\text{RCmad}(x)=\text{med}_i\{\text{med}_j|x_i-x_j|\}$.

Note that the error rate $\alpha_2'$ is interpreted as the proportion of falsely estimated differential edges before sparsity thresholding operations are applied. It considers dense estimates of the individual matrices, which links with the proposed method in Zhao et al. (2014) of directly estimating the difference matrix $\Omega_d$ since it does not assume sparsity of $\{\Omega_m\}_{m=1,2}$ either. This, as well as numerical simplicity, motivates us to control $\alpha_2'$, and estimate $\alpha_2$.

Default values for $\alpha_1$ and $\alpha_2'$ as 0.01 or 0.05 could be used. An immediate upper bound for $\alpha_2$ is $\alpha_2\leq 2\alpha_1\alpha_2'$ however the numerical integration gives a more precise value.

### 6.2.3 Weights in the similarity penalization term

Recall from Corollary 6.1 that we consider $\lambda_{ij}^{(2)} = \lambda_2 v_{ij} = \lambda_2(1-\psi_{ij})^{1/2}$. When $Y^{(1)}$ and $Y^{(2)}$ are independent, then the correlation coefficients $\psi_{ij} = 0$ for all pairs $(i, j)$, and the penalty coincides for all elements in the matrix, $\lambda_{ij}^{(2)} = \lambda_2$ for any $i \neq j$. Otherwise, $\psi_{ij}$ are approximated by

$$\psi_{ij} = \text{cor}(\hat{\Omega}_{ij}^{(1)}, \hat{\Omega}_{ij}^{(2)}) = \text{cor}((\hat{\Omega}_{ii}^{(1)})^{1/2}(\hat{\Omega}_{jj}^{(1)})^{1/2}\hat{w}_{ij}^{(1)}, (\hat{\Omega}_{ii}^{(2)})^{1/2}(\hat{\Omega}_{jj}^{(2)})^{1/2}\hat{w}_{ij}^{(2)}) \doteq \text{cor}(\hat{w}_{ij}^{(1)}, \hat{w}_{ij}^{(2)}), \tag{6.14}$$

where $\hat{W} = [\hat{w}_{ij}]$ is the partial correlation matrix determined by the scaled estimated precision matrix $\hat{\Omega}$. A mathematical expression for $\text{cor}(\hat{w}_{ij}^{(1)}, \hat{w}_{ij}^{(2)})$ is derived in Olkin and Finn (1990), among others, and uses the true partial correlation coefficients by

$$\begin{aligned}\psi_{ij} \doteq \frac{1}{(1-(w_{ij}^{(1)})^2)(1-(w_{ij}^{(2)})^2)}[&w_{ii}^{(12)}w_{jj}^{(12)} + w_{ij}^{(12)}w_{ji}^{(12)} + w_{ij}^{(1)}w_{ij}^{(2)}((w_{ii}^{(12)})^2 + (w_{jj}^{(12)})^2 + (w_{ij}^{(12)})^2 \\ &+ (w_{ji}^{(12)})^2)/2 - \{w_{ij}^{(1)}(w_{ij}^{(1)}w_{ij}^{(12)} + w_{ji}^{(12)}w_{ij}^{(2)}) + w_{ij}^{(2)}(w_{ji}^{(12)}w_{ii}^{(12)} + w_{jj}^{(12)}w_{ij}^{(12)})\}\end{aligned} \tag{6.15}$$

Expression (6.15), which excludes the perfect dependence case where $w_{ij}^{(1)} = 1$ and $w_{ij}^{(2)} = 1$, is found to provide a good approximation of $\text{cor}(\hat{\Omega}_{ij}^{(1)}, \hat{\Omega}_{ij}^{(2)})$. In particular, if $\text{cor}(\hat{\Omega}_{ij}^{(1)}, \hat{\Omega}_{ij}^{(2)}) = 0$ then $\text{cor}(\hat{w}_{ij}^{(1)}, \hat{w}_{ij}^{(2)}) = 0$.

The expression of weights $[\psi_{ij}]$ given in (6.15) depends on $\Omega_{12}^J$, defined in eq. (6.1), and its estimation requires higher sample sizes, which is not always possible in practice. Hence, for practical purposes, its nonzero structure is assumed to be known and the number of unknown elements is assumed to be relatively small. In Xie et al. (2016), the authors fix the structure of $R_{12}$ by considering an additive model. As we do not have such prior information about the data, we assume that $\Omega_{12}^J$ is a diagonal matrix as proposed by Wit and Abbruzzo (2015) in a similar context, i.e., we assume that any variable of the first dataset $Y_{ki}^{(1)}$ is conditionally independent from any variable of the other dataset $Y_{kj}^{(2)}$, if $i \neq j$, $k \in \{1, \ldots, n\}$, given the rest of the variables $(Y_{kh}^{(1)})_{h\neq i}$ and $Y_{ki}^{(2)}$. In such case, the expression for weights defined in eq. (6.15) can be simplified to

$$\psi_{ij} \doteq \frac{w_{ii}^{(12)}w_{jj}^{(12)} + w_{ij}^{(1)}w_{ij}^{(2)}((w_{ii}^{(12)})^2 + (w_{jj}^{(12)})^2)/2}{(1-(w_{ij}^{(1)})^2)(1-(w_{ij}^{(2)})^2)}. \tag{6.16}$$

Under subject-dependence, we propose the following two estimators of $\psi_{ij}$,

1. Regression-based estimator (Reg-based):

$$\hat{\psi}_{ij} = \frac{\hat{w}_{ii}^{(12)}\hat{w}_{jj}^{(12)} + \hat{w}_{ij}^{(1)}\hat{w}_{ij}^{(2)}((\hat{w}_{ii}^{(12)})^2 + (\hat{w}_{jj}^{(12)})^2)/2}{(1-(\hat{w}_{ij}^{(1)})^2)(1-(\hat{w}_{ij}^{(2)})^2)}. \tag{6.17}$$

   where $\hat{w}_{ij}^{(1)}$ and $\hat{w}_{ij}^{(2)}$ are estimators of $w_{ij}^{(1)}$ and $w_{ij}^{(2)}$, respectively, which are found using eq. (6.4) on the initial iteration of the ADMM Algorithm 8. Coefficients $\hat{w}_{ii}^{(12)}$ and $\hat{w}_{jj}^{(12)}$ are computed by considering a regression-type partial correlation coefficient estimation, i.e., $\hat{w}_{ii}^{(12)} = \widehat{\text{cor}}(Y_{.i}^{(1)} -$

$Y^{(1)}_{.,-i}\hat{\beta}^{(1)}_{i,-i}, Y^{(2)}_{.i} - Y^{(2)}_{.,-i}\hat{\beta}^{(2)}_{i,-i}$), with regression coefficients $\hat{\beta}^{(m)}_{i,-i} = -/hat\Omega^{(m)}_{i,-i}/\hat{\Omega}^{(m)}_{i,i}$ for $m = 1,2$.

2. Regression-based simplified estimator (Reg-based-sim):

$$\hat{\psi}_{ij} = \hat{w}^{(12)}_{ii}\hat{w}^{(12)}_{jj}(1-(\hat{w}^{(1)}_{ij})^2)^{-1}(1-(\hat{w}^{(2)}_{ij})^2)^{-1}. \tag{6.18}$$

for same regression-based estimators of $\hat{w}^{(12)}_{ii}$ and $\hat{w}^{(12)}_{jj}$ as well as partial correlation estimators $\hat{w}^{(1)}_{ij}$ and $\hat{w}^{(2)}_{ij}$, which are defined in the Reg-based estimator, using the leading term in the numerator in eq. (6.17).

The performance of the two estimators is compared on simulated data in Appendix B.6.

## 6.3 Weighted fused regression lasso

### 6.3.1 Model setting, assumptions and link with joint precision matrices

Consider that $n$ i.i.d. pairs of $q$-dimensional samples $(Y^{(1)}, Y^{(2)}) : (Y^{(1)}_1, Y^{(2)}_1), \ldots, (Y^{(1)}_n, Y^{(2)}_n)$ are observed. For the same individuals, we assume there are data measurements of pairs of $p$-dimensional vectors of covariates $(X^{(1)}, X^{(2)}) : (X^{(1)}_1, X^{(2)}_1), \ldots, (X^{(1)}_n, X^{(2)}_n)$. For the motivating data of colon cancer (Hinoue et al., 2012) introduced in Section 6.1, $Y^{(1)}$ would correspond to the $n \times q$ matrix with the gene expression for healthy samples and $X^{(1)}$ would be the $n \times p$ matrix with the methylation presence information for healthy samples. Similarly, $Y^{(2)}$ and $X^{(2)}$ would refer to the gene expression and methylation presence data for tumor samples.

We take gene expression samples $(Y^{(1)}, Y^{(2)})$ as response variables and methylation presence samples $(X^{(1)}, X^{(2)})$ as explanatory variables, and we assume these are associated by a joint Gaussian linear model

$$(Y^{(1)}_k, Y^{(2)}_k) \sim N_{2q}\left(\begin{bmatrix} X^{(1)}_k\beta^{(1)} \\ X^{(2)}_k\beta^{(2)} \end{bmatrix}', R_\epsilon\right),\ R_\epsilon = \begin{bmatrix} R^{(1)}_\epsilon & R^{(12)}_\epsilon \\ R^{(21)}_\epsilon & R^{(2)}_\epsilon \end{bmatrix},\ q \gg n,\quad p \gg n, \tag{6.19}$$

where $\beta^{(1)}$ (first condition, i.e., healthy) and $\beta^{(2)}$ (second condition, i.e., tumor) describe the $p \times q$ regression coefficient matrices. Define the residual matrices $(Y^{(1)} - X^{(1)}\beta^{(1)}, Y^{(2)} - X^{(2)}\beta^{(2)})$, here, $R_\epsilon$ is the joint covariance matrix of the residuals with $R^{(1)}_\epsilon$ being the covariance sub-matrix for the residuals in samples on the first condition, $R^{(2)}_\epsilon$ being the covariance sub-matrix for residuals in samples on the second condition and $R^{(12)}_\epsilon$ being the cross-covariance matrix relating residuals in the two conditions. The regression method we propose in Section 6.3.2 has a rather strict assumption on the non-zero structure of these matrices: it assumes that $R^{(1)}_\epsilon$ and $R^{(2)}_\epsilon$ are diagonal matrices, and if data are paired, it assumes that $R^{(12)}_\epsilon$ is also a diagonal matrix. Hence, it considers linear independence between genes once conditioning for methylation. If this assumption does not hold, especially if residuals are highly correlated, then the predictive error can increase. Rothman et al. (2010) propose to account

for the residual's linear dependence structure to estimate a regression coefficient matrix in a single class of observations. We provide some initial insights on a similar approach to jointly estimating two regression coefficient matrices in Appendix B.2, but the inversion of a $pq \times pq$ matrix is needed, which can be computationally unfeasible for large dimensions.

Assuming multivariate normal distributions in both $X$ and $Y$, conditional dependence structures can be found by jointly estimating two precision matrices

$$\Omega^{(1)} = \begin{bmatrix} \Omega_Y^{(1)} & \Omega_{YX}^{(1)} \\ \Omega_{XY}^{(1)} & \Omega_X^{(1)} \end{bmatrix} \quad \text{and} \quad \Omega^{(2)} = \begin{bmatrix} \Omega_Y^{(2)} & \Omega_{YX}^{(2)} \\ \Omega_{XY}^{(2)} & \Omega_X^{(2)} \end{bmatrix}.$$

The elements in the cross-precision matrices $(\Omega_{XY}^{(l)})_{l=1,2}$ describe the linear dependence between a gene and a methylation site once conditioning on the linear dependence between the rest of the genes and sites. These have a slightly different interpretation to the regression coefficient matrices $(\beta^{(l)})_{l=1,2}$ defined in eq. (6.19) since a regression coefficient finds the linear relationship between a gene and a methylation site accounting for the rest of methylation sites but ignoring the dependence in the rest of the genes. In this sense, we consider a directed graphical representation of the non-zero structure of the regression coefficient matrix which has to be interpreted as methylation presence driving gene expression. A concern that is raised in the causality literature in which the proposed marginal regression model may be incurring is known by faithfulness (Robins et al., 2003). Unfaithfulness, or cancellation of correlations, can occur when ignoring covariates in the model. This can be shown in a simple example in which the model $y_1 \sim N(\beta_1 x_1 + \sum_{j\neq 1} \beta_j y_j, \sigma_1^2)$ is under consideration. The correlation between $y_1$ and $x_1$ might be zero even when $\beta_1$ is large if $\sum_{j\neq 1} \text{cor}(x_1, y_j)\beta_j \approx -\beta_1$.

In the following sections we present an initial method to jointly estimate $\beta^{(1)}$ and $\beta^{(2)}$. Similarly, matrices $\Omega_{XY}^{(1)}$ and $\Omega_{XY}^{(2)}$ could be estimated employing the joint graphical lasso approach proposed in Section 6.2. The link between precision matrices and regression coefficient matrices is studied in Section 2.1.

### 6.3.2 Estimation of joint regression coefficient matrices

We propose a weighted fused regression lasso estimator (WFRL) to find $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$. This solves the following penalized marginal least squares optimization problem, which encourages sparsity in the individual estimated regression coefficient matrices and commonality between the two such matrices,

$$(\hat{\beta}^{(1)}, \hat{\beta}^{(2)})_{FRL}^{\Lambda_1,\Lambda_2} = \arg\min_{\beta^{(1)},\beta^{(2)}} \left[ \frac{1}{2n} \sum_{l=1,2} ||Y^{(l)} - X^{(l)}\beta^{(l)}||_2^2 + P_{\Lambda_1,\Lambda_2}(\beta) \right], \tag{6.20}$$

with

$$P_{\Lambda_1,\Lambda_2}(\beta) = ||\Lambda_1 \circ \beta^{(1)}||_1 + ||\Lambda_1 \circ \beta^{(2)}||_1 + ||\Lambda_2 \circ (\beta^{(2)} - \beta^{(1)})||_1. \tag{6.21}$$

The tuning parameters in the $p \times q$ matrix $\Lambda_1 = [\lambda_{ij}^{(1)}]$ provide a trade off between sparsity and fit to the data, and $\Lambda_2 = [\lambda_{ij}^{(2)}]$ controls the similarity between $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$.

The maximization problem defined in eq. (6.20) and eq. (6.21) is solved by optimizing its Lagrangian formulation

$$L_\rho = P_{\Lambda_1,\Lambda_2}(Z^{(1)}, Z^{(2)}) + \frac{1}{2n} \sum_{l=1,2} ||Y^{(l)} - X^{(l)}\beta^{(l)}||_2^2 + \frac{\rho}{2}||\beta^{(l)} - Z^{(l)} + U^{(l)}||_F^2$$

using the ADMM-type algorithm (Boyd, 2010) described in Algorithm 9. Here, $U^{(l)}$ are the dual variables, $Z^{(l)}$ corresponds to $\beta^{(l)}$, for $l = \{1,2\}$, and $\rho$ is a positive constant that is used as a regularization parameter with default value equal to 1.

**Algorithm 9** Weighted Fused Regression Lasso

1: Input: $\Lambda_1, \Lambda_2, \rho$.
2: Initialization: $t = 0$, $U_t^{(l)} = 0$ and $Z_t^{(l)} = 0$, for $l = 1,2$, repeat 3-5 until convergence.
3: Find $\hat{\beta}_t^{(l)}, \hat{\beta}_t^{(2)}$ by solving the minimization problem:

$$[\hat{\beta}_t^{(1)}, \hat{\beta}_t^{(2)}] = \arg\min_{\beta^{(1)},\beta^{(2)}} \left\{ \sum_{l=1,2} \frac{1}{2n} ||Y^{(l)} - X^{(l)}\beta^{(l)}||_2^2 + \frac{\rho}{2}||\beta^{(l)} - Z_{t-1}^{(l)} + U_{t-1}^{(l)}||_F^2 \right\}.$$

4: Find $Z_t^{(1)}, Z_t^{(2)}$ such that

$$\sum_{l=1,2} \frac{\rho}{2}||\hat{\beta}_t^{(l)} - Z_l + U_t^{(l)}||_F^2 + P_{\Lambda_1,\Lambda_2}(Z_t^{(l)}, Z_t^{(l)})$$

is minimized.
5: Set $t = t+1$. Update dual variables $U_t^{(l)} = U_{t-1}^{(l)} + \hat{\beta}_t^{(l)} - Z_t^{(l)}$, for $l = 1,2$. Stop if convergence.
6: Output: $\hat{\beta}^{(1)} = \hat{Z}_{t-1}^{(1)}$, $\hat{\beta}^{(2)} = \hat{Z}_{t-1}^{(2)}$ and $\hat{\beta}^{(d)} = \hat{Z}_{t-1}^{(2)} - \hat{Z}_{t-1}^{(1)}$.

The optimization problem in step 3 of the algorithm is solved by a ridge type matrix inversion. For $l = 1,2$,

$$\hat{\beta}_t^{(l)} = \left(\frac{1}{n} X^{(l)'} X^{(l)} + \rho I\right)^{-1} \left(\frac{1}{n} X^{(l)'} Y^{(l)} - Z_{t-1}^{(l)} + U_{t-1}^{(l)}\right), \tag{6.22}$$

with $\rho > 0$ such that $\frac{1}{n} X^{(l)'} X^{(l)} + \rho I$ is a positive definite matrix. Moreover, step 4 is determined by the following thresholding operations:

(i) Given $\hat{A}^{(1)} = \hat{\beta}_t^{(1)} + U_{t-1}^{(1)}$ and $\hat{A}^{(2)} = \hat{\beta}_t^{(2)} + U_{t-1}^{(2)}$, set regression coefficients between two classes to their average value if $|\hat{A}_{ij}^{(1)} - \hat{A}_{ij}^{(2)}| \leq \lambda_{ij}^{(2)}/\rho$, and furthermore:

$$[Z_{t_{ij}}^{'(1)}, Z_{t_{ij}}^{'(2)}] = \begin{cases} .5[\hat{A}_{ij}^{(1)} + \hat{A}_{ij}^{(2)}, \hat{A}_{ij}^{(1)} + \hat{A}_{ij}^{(2)}] & \text{if } |\hat{A}_{ij}^{(1)} - \hat{A}_{ij}^{(2)}| \leq \lambda_{ij}^{(2)}/\rho; \\ [\hat{A}_{ij}^{(1)} + \lambda_{ij}^{(2)}/(2\rho), \hat{A}_{ij}^{(2)} - \lambda_{ij}^{(2)}/(2\rho)] & \text{if } \hat{A}_{ij}^{(1)} - \hat{A}_{ij}^{(2)} > \lambda_{ij}^{(2)}/\rho; \\ [\hat{A}_{ij}^{(1)} - \lambda_{ij}^{(2)}/(2\rho), \hat{A}_{ij}^{(2)} + \lambda_{ij}^{(2)}(2\rho)] & \text{if } \hat{A}_{ij}^{(2)} - \hat{A}_{ij}^{(1)} > \lambda_{ij}^{(2)}; \end{cases} \tag{6.23}$$

(ii) Set regression coefficients to zero by soft-thresholding (Rothman et al., 2009) with exceedances

threshold given by $\lambda_1$:

$$Z^{(1)}_{t_{ij}} = \text{sign}(Z'^{(1)}_{t_{ij}})\left(Z'^{(1)}_{t_{ij}} - \lambda^{(1)}_{ij}\right)_+, \quad Z^{(2)}_{t_{ij}} = \text{sign}(Z^{(2)}_{t_{ij}})\left(Z'^{(2)}_{t_{ij}} - \lambda^{(1)}_{ij}\right)_+. \tag{6.24}$$

Under subject-independence between $X^{(1)}$ and $X^{(2)}$ (and also between $Y^{(1)}$ and $Y^{(2)}$), we consider $\Lambda_1 = \lambda_1 J$ and $\Lambda_2 = \lambda_2 J$, and we refer to the underlying estimator of $(\beta^{(l)})_{l=1,2}$ by fused regression lasso (FRL). Under the paired data setting, see Section 6.3.1, FRL estimators $\hat{\beta}^{(1)}_{ij}$ and $\hat{\beta}^{(2)}_{ij}$ might be correlated for some pairs $(i, j)$ with $i \in [1, p]$ and $j \in [1, q]$. Motivated by the results obtained in Section 6.2.2 for the estimation of joint precision matrices, here we also consider a weighted procedure (WFRL) where $\lambda^{(2)}_{ij} = \lambda_2 v_{ij}$ depends on weights $v_{ij} = (1-\theta_{ij})^{1/2}$ with $\theta_{ij} = \text{cor}(\hat{\beta}^{(1)}_{ij}, \hat{\beta}^{(2)}_{ij})$. Furthermore, Similarly to the WFGL, error rates $\alpha_1$ $\alpha_2$ and $\alpha'_2$ could also be adapted to the WFRL method considering $\hat{\beta}^{(1)}$, $\hat{\beta}^{(2)}$ and $\hat{\beta}^{(d)}$ instead of $\hat{\Omega}_1$, $\hat{\Omega}_2$ and $\hat{\Omega}_d$.

### 6.3.3 Weights in the similarity penalization term

Weights defined in the $p \times q$ matrix $V = [v_{ij}]$ adjust the similarity penalization term $\lambda_2$ for every pair of variables, and we propose to use $v_{ij} = (1-\theta_{ij})^{1/2}$ where $\theta_{ij} = \text{cor}(\hat{\beta}^{(1)}_{ij}, \hat{\beta}^{(2)}_{ij})$ describes the correlation between estimated regression coefficients in the two conditions. If subject-independence is known, then $\theta_{ij}$ can be set to zero for all pairs $(i, j)$, and the weights are constant for all pairs of variables. Otherwise, the expression derived in Olkin and Finn (1990), which is used to determine the correlation of sample correlation coefficients, is applied here to regression coefficients by

$$\begin{aligned}\theta_{ij} \doteq \frac{1}{(1-\rho^{(1)^2}_{(Y|X)_{ij}})(1-\rho^{(2)^2}_{(Y|X)_{ij}})} \Big[ & \rho^{(1,2)}_{(X)_{ii}} \rho^{(1,2)}_{(Y|X)_{jj}} + \rho^{(1,2)}_{(X,Y|X)_{ij}} \rho^{(2,1)}_{(X,Y|X)_{ij}} + \rho^{(1)}_{(Y|X)_{ij}} \rho^{(2)}_{(Y|X)_{ij}} \\ & \times (\rho^{(1,2)^2}_{(X)_{ii}} + \rho^{(1,2)^2}_{(Y|X)_{jj}} + \rho^{(1,2)^2}_{(X,Y|X)_{ij}} + \rho^{(2,1)^2}_{(X,Y|X)_{ij}})/2 - \{\rho^{(1)}_{(Y|X)_{ij}} (\rho^{(1)}_{(Y|X)_{ij}} \rho^{(1,2)}_{(X,Y|X)_{ij}} \\ & + \rho^{(2,1)}_{(X,Y|X)_{ij}} \rho^{(2)}_{(Y|X)_{ij}}) + \rho^{(2)}_{(Y|X)_{ij}} (\rho^{(2,1)}_{(X,Y|X)_{ij}} \rho^{(1,2)}_{(X)_{ii}} + \rho^{(1,2)}_{(Y|X)_{jj}} \rho^{(1,2)}_{(X,Y|X)_{ij}})\}\Big], \end{aligned} \tag{6.25}$$

where

$$\begin{cases} \rho^{(1,2)}_{(X)_{ii}} = \text{cor}(X^{(1)}_i - X^{(1)}_{-i}\beta^{(1)}_{X_{-i,i}}, X^{(2)}_i - X^{(2)}_{-i}\beta^{(2)}_{X_{-i,i}}), \\ \rho^{(1,2)}_{(Y|X)_{jj}} = \text{cor}(Y^{(1)}_j - X^{(1)}\beta^{(1)}_{.,j}, Y^{(2)}_j - X^{(2)}\beta^{(2)}_{.,j}), \\ \rho^{(1,2)}_{(X,Y|X)_{ij}} = \text{cor}(X^{(1)}_i - X^{(1)}_{-i}\beta^{(1)}_{X_{-i,i}}, Y^{(2)}_j - X^{(2)}\beta^{(2)}_{.,j}), \\ \rho^{(2,1)}_{(X,Y|X)_{ij}} = \text{cor}(X^{(2)}_i - X^{(2)}_{-i}\beta^{(2)}_{X_{-i,i}}, Y^{(1)}_j - X^{(1)}\beta^{(1)}_{.,j}), \\ \rho^{(1)}_{(Y|X)_{ij}} = \beta^{(1)}_{ij}\text{var}(Y^{(1)}_j - X^{(1)}\beta^{(1)}_{.,j}), \quad \rho^{(2)}_{(Y|X)_{ij}} = \beta^{(2)}_{ij}\text{var}(Y^{(2)}_j - X^{(2)}\beta^{(2)}_{.,j}). \end{cases}$$

We estimate the correlation coefficients $[\theta_{ij}]$, $i \in [1, p]$ and $j \in [1, q]$, by plugging in the sample estimators of $\rho$'s instead of the true values in eq. (6.25). For instance, we take $\hat{\beta}^{(1)} = \hat{\beta}^{(1)}_1$ and $\hat{\beta}^{(2)} = \hat{\beta}^{(2)}_1$ being the solution of step 3 in Algorithm 9 at the initial iteration, which provides dense estimators of the regression coefficient matrices. Besides, we approximate $\rho^{(1,2)}_{(X)_{ii}}$ considering marginal estimates of $\Omega^{(1)}_X$ and $\Omega^{(2)}_X$, for instance using the proposed estimator in Section 6.2, follow description of $\hat{w}_{12}$ in eq. (6.17).

For simplicity, we make the further assumption that $\rho^{(1,2)}_{(X,Y|X)_{ij}} = \rho^{(2,1)}_{(X,Y|X)_{ij}} = 0$ for any $i \neq j$. This assumes that there is no direct link between any pair given by a gene and methylation site in which one is in a normal tissue and the other corresponds to a tumor tissue. The leading term of expression (6.25) is given by the product $\rho^{(1,2)}_{(X)_{ii}} \rho^{(1,2)}_{(Y|X)_{jj}}$. Thus, both dependence in explanatory variables and dependence in residuals are needed for $\theta_{ij}$ to be influential.

The performance of the proposed estimator for $[\theta_{ij}]$ is assessed on simulated data in Appendix B.7. We compare the estimator to an approximated value of $\text{cor}(\hat{\beta}^{(1)}_{ij}, \hat{\beta}^{(2)}_{ij})$ using 5000 i.i.d Monte Carlo instances, so in a way the sensibility of expression (6.25), which is adapted here for regression coefficients, is also evaluated.

## 6.4 Overestimation of triangular motifs

### 6.4.1 Problem and toy example

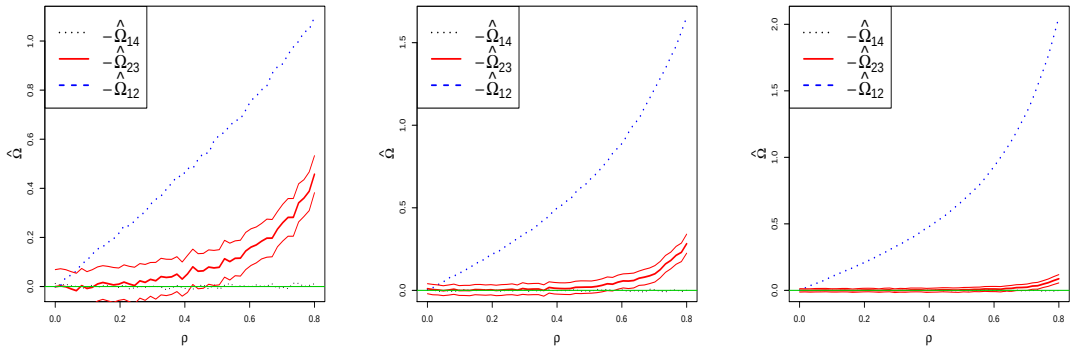
We have discovered that for the WFGL and FGL estimators, the overestimation of triangles is a major issue. If there are 3 nodes $i, j, h$ and it is known that pairs $i, h$ and $j, h$ are connected, then a connection between $i$ and $j$ is more often falsely predicted than expected. The reason for this is that the Algorithm 8 used to find the estimates, see eq. (6.4), considers a regularization with rate $\rho$ for the eigenvalues $[D_{jj}]_{j=1}^{p}$ of the covariance/correlation matrix to approximate its inverse denoted by $[\tilde{D}_{jj}]_{j=1}^{p}$. It can be proved that when $D_{jj} \gg (\rho/n)^{1/2}$ then $\tilde{D}_{jj} \approx 1/D_{jj}$ and when $D_{jj} \leq c(\rho/n)^{1/2}$ then $\tilde{D}_{jj} \approx \tilde{c}(n/\rho)^{1/2}$. In the second such scenario, which happens when eigenvalues $[D_{jj}]$ are small, the estimated coefficients are biased.

This is illustrated using a toy graph structure example described by: $G_x : (1 \longleftrightarrow 2)$, $(1 \longleftrightarrow 3)$, $(4 \longleftrightarrow \emptyset)$; hence, here the edge $2 \longleftrightarrow 3$ is the one missing to complete a triangle. Assuming that the correlations between $1 \longleftrightarrow 2$ and $1 \longleftrightarrow 3$ have the same strength $r$, the correlation matrix and its inverse are expressed by

$$R_1 = \begin{pmatrix} 1 & & & \\ r & 1 & & \\ r & r^2 & 1 & \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad R_1^{-1} = \Omega_1 = \frac{1}{1-r^2} \begin{pmatrix} 1+r^2 & & & \\ -r & 1 & & \\ -r & 0 & 1 & \\ 0 & 0 & 0 & 1-r^2 \end{pmatrix},$$

To show the behavior of the regularized precision matrix estimator defined by (6.4) we simulate data from a multivariate normal distribution with mean vector equal to zero and covariance matrix equal to $R_1$. Figure 6.1 shows the trend of $-\hat{\Omega}_{12}$ (true edge), $-\hat{\Omega}_{14}$ (false edge) and $-\hat{\Omega}_{23}$ (false triangle edge) for different sample sizes and over 1000 simulations. Note that $\hat{\Omega}_{12}$ is shrunk towards zero for small $n$ as expected, also $\hat{\Omega}_{14}$ is centered at zero as expected but $\hat{\Omega}_{23}$ is biased. The true $\Omega_{23} = 0$, but for $r$ large enough, $\hat{\Omega}_{23}$ is different from zero.

The algorithm 8, when weights $[v_{ij} = 1]$, leading to FGL is implemented in R (Danaher et al., 2013).




**(a)** $n = 25$ **(b)** $n = 100$ **(c)** $n = 500$

**Figure 6.1.** Average values (times $-1$ for positive representation) of the estimated precision matrix elements for true edge, $\hat{\Omega}_{12}$, false triangle edge, $\hat{\Omega}_{23}$ (with confidence intervals), and false edge, $\hat{\Omega}_{14}$. A bias $\hat{\Omega}_{23}$ is observed as $\rho$ increases.

We have realized that the authors make an additional consideration to the formula (6.4) which finds dense precision matrices by quadratic inversions, i.e., if the sample sizes of the two datasets are equal, then $\tilde{D}^{(t)}_{m_{jj}} = (2\rho)^{-1}[-D^{(t)}_{m_{jj}} + \{(D^{(t)}_{m_{jj}})^2 + 4\rho\}^{1/2}]$ is considered for both $m = 1,2$. The reason is that even though using $n$ in expression (6.4) reduces the bias, it also gives much larger variances for edges equal to zero producing more false positive edges. However, replacing $n$ by 1 in eq. (6.4) causes the principle problem of detecting too many false positive triangular motifs.

### 6.4.2 Reducing overestimation of triangular motifs

Consider the three nodes $i, j, h$ with partial correlation coefficients $w^{(1)}_{ij} = \text{cor}(Y^{(1)}_i, Y^{(1)}_j \mid Y^{(1)}_h)$, $w^{(1)}_{ih} = \text{cor}(Y^{(1)}_i, Y^{(1)}_h \mid Y^{(1)}_j)$ and $w^{(1)}_{ih} = \text{cor}(Y^{(1)}_j, Y^{(1)}_h \mid Y^{(1)}_i)$. Assume that $w^{(1)}_{ih} \neq 0$ and that $w^{(1)}_{jh} \neq 0$. Here we focus on the hypothesis testing problem defined by $\text{H}_0$: $w^{(1)}_{ij} = 0$ (not a triangle) and $\text{H}_1$: $w^{(1)}_{ij} \neq 0$ (triangle) using the sample partial correlation matrix $\hat{W}_1$ which contains the three variables $i, j, h$. The p-value of the test is given by

$$\text{p-val} = P(|Z| \geq |g(\hat{w}^{(1)}_{ij})|) \doteq 2 - 2\Phi\left(\sqrt{n-5}(|g(\hat{w}^{(1)}_{ij})|)\right), \tag{6.26}$$

where $g : (-1,1) \to \mathbb{R}$, $g(z) = \log\{(1+z)/(1-z)\}/2$ is the Fisher transformation function, that is applied to the partial correlation coefficient $\hat{w}_{ij}$ (Fisher, 1924), and $Z$ is the standard normal r.v. with cumulative distribution $\Phi$.

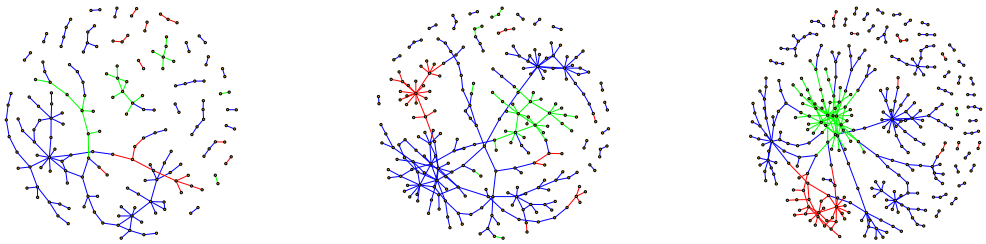
In practice, the pair $(i, j)$ with $w^{(1)}_{ij} = 0$ might be unknown. Hence, a p-value for the test is approximated by applying (6.26) on the smallest estimated coefficient in absolute value $\min(|\hat{w}^{(1)}_{ij}|, |\hat{w}^{(1)}_{ih}|, |\hat{w}^{(1)}_{jh}|)$. This results to a conservative p-value, i.e., $\Pr(|Z| \geq |g(\hat{w}^{(1)}_{ij})| \cup |Z| \geq |g(\hat{w}^{(1)}_{ih})| \cup |Z| \geq |g(\hat{w}^{(1)}_{jh})|) \geq \Pr(|Z| \geq |g(\hat{w}^{(1)}_{ij})|)$. For sufficiently large sample sizes, and large true non-zero partial correlation coefficients, the equality holds.

Here we assess the weakest edges of all observed triangular motifs independently for $Y^{(1)}$ and $Y^{(2)}$, and we eliminate those with large p-values (default threshold equal to $\alpha_1$, see eq. (6.6)). In case an edge is tested more than once, we only count its smallest p-value. However, multiple testing correction and another interpretation for triangles overlap could be used instead.

## 6.5 Simulated data analysis

### 6.5.1 Generation of joint precision matrices

Simulated data are obtained from multivariate normal distributions with zero mean vector and almost-block diagonal precision matrices, where each block has a power-law underlying graph structure and some extra random connections between blocks, i.e., we follow the same strategy proposed in Section 4.5.2. We generate datasets with several dimension sizes $p = 200, 300, 400$ and sample sizes $n = 25, 100, 250, 500$ to assess the performance of the WFGL approach and compare it to standard methods in Sections 6.5.3-6.5.4. Figure 6.2 shows the network representation of some of the simulated non-zero precision matrix structures. It distinguishes between common edges (blue) and differential edges (green and red).



**(a)** Network example p=200 **(b)** Network example p=300 **(c)** Network example p=400

**Figure 6.2.** Graph structure examples: green edges are zero elements in the precision matrix for second class and non zero for first class; Red edges are zero in first class and non-zero in second class; Finally, blue edges are non-zero and equal in both conditions.

### 6.5.2 Generation of joint regression coefficient matrices

Given p-dimensional random vectors $X^{(1)}$ and $X^{(2)}$, which can be obtained as described in Section 6.5.1, we assume a Gaussian linear model to relate explanatory variables $X = [X^{(1)}, X^{(2)}]$ and response variables $Y = [Y^{(1)}, Y^{(2)}]$ by

$$(Y_k^{(1)}, Y_k^{(2)}) \sim N_{2q}\left(\begin{bmatrix} X_k^{(1)}\beta^{(1)} \\ X_k^{(2)}\beta^{(2)} \end{bmatrix}', R_\epsilon\right), \qquad R_\epsilon = \begin{bmatrix} R_\epsilon^{(1)} & R_\epsilon^{(1,2)} \\ R_\epsilon^{(2,1)} & R_\epsilon^{(2)} \end{bmatrix}, \tag{6.27}$$

where, for the sake of simplicity, we assume same dimension for response and explanatory variables ($q = p$). We determine $R_\epsilon$ so that $R_\epsilon^{(1)} = R_\epsilon^{(2)} = I\sigma^2$ and $R_\epsilon^{(1,2)}$ is a diagonal matrix with $R_{\epsilon_{ii}}^{(1,2)} = 0.6\sigma^2$ for $\lfloor q/2 \rfloor$ diagonal elements and $R_{\epsilon_{ii}}^{(1,2)} = 0$ for the other $\lceil q/2 \rceil$. Moreover, we distinguish between the following two patterns for $\beta^{(1)}$ and $\beta^{(2)}$:

1. Scenario 1: we assume that $\beta^{(1)} = I\kappa^{(1)}$ and $\beta^{(2)} = I\kappa^{(2)}$ are diagonal matrices which have $m = \theta p$ elements equal in the diagonal coefficients $\kappa^{(1)}$ and $\kappa^{(2)}$. For the other $p(1-\theta)$ elements, we take $\kappa^{(1)} \neq 0$ and $\kappa^{(2)} = 0$. We use $\theta = 0.1, 0.4, 0.7$. This is a simple dependence structure which might be unrealistic for our application. For instance, note that several methylation sites might be related to the same gene promoter.

2. Scenario 2: we assume sparse regression coefficient matrices $\beta^{(1)} = \rho\Omega^{(1)}$ and $\beta^{(2)} = \rho\Omega^{(2)}$ for some $0 < \rho < 1$, in which the linear relationships between response and explanatory variables is proportional to the conditional linear relationship within explanatory variables. We use several proportions of differential edges: $\theta = 0.1, 0.4, 0.7$.

In both scenarios, the proportion of differential edges $\theta = 0.7$ is only added to compare the joint estimation approach against estimating two separate regression lasso in Section 6.5.5, but we do no expect such large proportions in the application to genomic data.

### 6.5.3 Differential network recovery for the precision matrices

In this section we focus on the recovery of differential edges by using two joint graphical lasso algorithms in the simulated datasets: FGL (Danaher et al., 2014) and WFGL -without triangle correction- (proposed in this chapter). Initially we had though about using ROC curves to compare the two methods, by keeping fix $\lambda_2$ and moving $\lambda_1$ from low to high values. The comparison resulted to be difficult though, as for instance, the same $\lambda_2$ might induce different graph structure complexities in the two approaches. Then differences in the ROC curves might be due to the $\lambda_2$ specification rather than real differences among methods.

For this reason, in order to make the structures of the estimated matrices comparable, we select estimated graphs (or $\lambda_1$ and $\lambda_2$) that have the same number of common edges and differential edges in the two approaches, i.e., we select the pair $[\lambda_1, \lambda_2]$ for the WFGL approach by setting the expected false positive rate by the parameters $[\alpha_1 = 0.05, \alpha_2' = 0.05]$ following the strategy proposed in Section 6.2.2, and we find $\lambda$'s such that the FGL graphs have the same sizes as WFGL. We compare the performance of the methods using a simple measure as the Youden's index, which is defined by $\text{YI}_\lambda^M = \text{TP}_\lambda^M - \text{FP}_\lambda^M$, with $M$ = FGL or WFGL, where $\text{TP}_\lambda^M = \sum_{i<j} I[\hat{\Omega}_{ij}^{(1)}(M) - \hat{\Omega}_{ij}^{(2)}(M) \neq 0, \Omega_{ij}^{(1)} - \Omega_{ij}^{(2)} \neq 0)]$ and $\text{FP}_\lambda^M = \sum_{i<j} I[\hat{\Omega}_{ij}^{(1)}(M) - \hat{\Omega}_{ij}^{(2)}(M) \neq 0, \Omega_{ij}^{(1)} - \Omega_{ij}^{(2)} = 0)]$ are the numbers of true positives and false positive of the estimated differential graphs with $\lambda = [\lambda_1, \lambda_2]$ and method $M$. Then we compute $\delta = \text{YI}_\lambda^{\text{WFGL}} - \text{YI}_\lambda^{\text{FGL}}$, which defines the Youden's index differences between the two methods to estimate the joint networks.

In Table 6.1 we present the average value of $\delta$ (with a Student's $t$-test p-value), and also the average sign of $\delta$ (with a Wilcoxon test p-value). In total we use 200 instances for each model, 4 different sample sizes $n = 25, 100, 250, 500$ and three dimension sizes $p = 200$, 300, 400. The proposed method, that assumes a dependence structure, achieves better TP-FP ratios for the differential network than the original FGL in most of the models when $n$ is large ($\geq 100$). For small $n$ ($n = 25$), the FP-TP are similar between the two algorithms even if there exists a dependence structure in the data. This may be due to the lack of data to estimate additional parameters $\psi_{ij}$.

**Table 6.1.** Youden Index differences between WFGL and FGL algorithm: average (p-value for t-test) and average sign (p-value for sign test). WFGL finds better estimates than FGL, especially for $n \geq 100$.

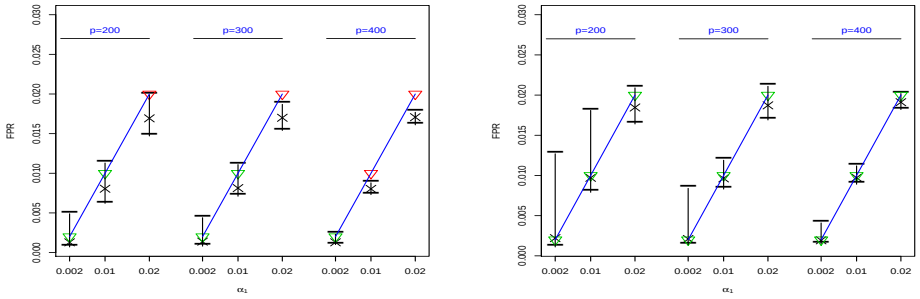
| | p= 200 | | p=300 | | p=400 | |
|---|---|---|---|---|---|---|
| $n$ | $\bar{\delta}$ (p-val) | $\bar{sgn}(\delta)$ (p-val) | $\bar{\delta}$ (p-val) | $\bar{sgn}(\delta)$ (p-val) | $\bar{\delta}$ (p-val) | $\bar{sgn}(\delta)$ (p-val) |
| 25 | 0.3 (.02) | .15 (.03) | 0.2 (0.22) | .07 (.22) | 0.2 (.18) | .09 (.19) |
| 100 | 2.0 (<.01) | .55 (<.01) | 2.6 (<.01) | .55 (<.01) | 2.4 (<.01) | .53 (<.01) |
| 250 | 3.1 (<.01) | .70 (<.01) | 5.6 (<.01) | .86 (<.01) | 5.7 (<.01) | .84 (<.01) |
| 500 | 1.9 (<.01) | .54 (<.01) | 4.0 (<.01) | .68 (<.01) | 5.6 (<.01) | .77 (<.01) |

### 6.5.4 Tuning parameter selection and testing and removing triangular motifs

In Figure 6.3 we compare the expected proportion of false positive edges determined by the value of $\alpha_1$ against the observed false positive rate (with median and 95% confidence) using the RCmad estimator described in Section 6.2.2 to approximate $\sigma_1$. To construct the confidence interval we replicate the procedure in 100 simulated datasets using different sample sizes and dimensions. The approximated false positive rate is close to the true one, given by $\alpha_1$, and it is only for small $n$ ($n = 25$) that the true value is not always included in the confidence interval.

Similarly, in Figure 6.4 we compare the expected proportion of false positive edges in the differential network (as defined in Section 3.2 of the article) determined by the value of $\alpha_2$ against the observed false positive rate (with median and 95% confidence) using the RCmad estimator to approximate $\sigma_2$. As for $\alpha_1$, the approximated false positive rate is close to the desired $\alpha_2$ and again it is only for the smallest tested $n$ that the true value is not included in the confidence interval.

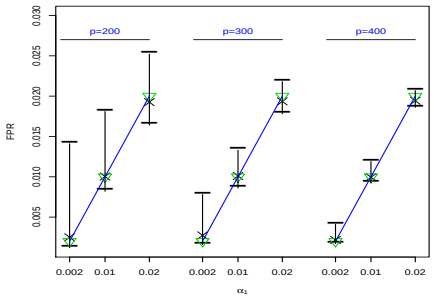
As we discussed in Section 6.4, using the eigenvalue decomposition regularization forces an overestimation of some non existing edges in the true network that complete triangular motifs. In Table 6.2 we present the average TP-FP behaviour for the weakest edge of estimated triangles for simulated models with different sample sizes, dimensions and error rates $\alpha$, distinguishing the triangles that take part in a common network and triangles in a differential network. The initial estimated triangles contain more false positives than true positives increasingly with $p \,\&\, n$ increasing. This is corrected by our triangle detection procedure (particularly for common edges), which notably reduces the number of false positives without losing many true positive edges.



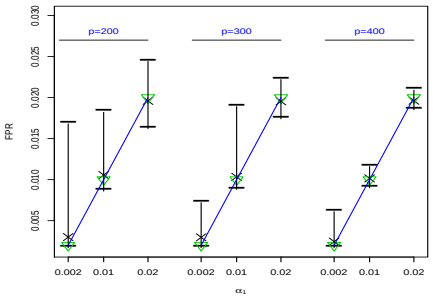

**(a)** n=25




**(b)** n=100




**(c)** n=250




**(d)** n=500

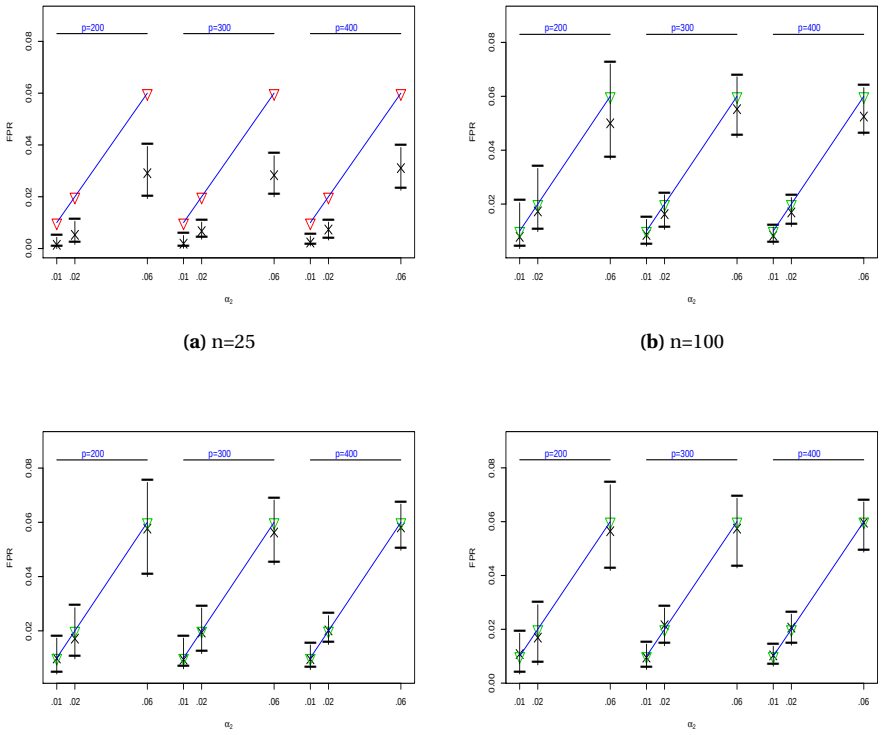
**Figure 6.3.** FPR vs $\alpha_1$: average (cross) + CI is plotted together with the expected values (triangle). For visualization reasons, x-axis and y-axis are not in the same scale (i.e. $2x : y$).

### 6.5.5 Graph recovery for the regression coefficient matrices

In this section we evaluate the performance of the proposed joint regression lasso approach against the standard lasso regression that finds estimates in the two classes independently. To do so we consider four values for the similarity tuning parameter $\alpha_2'$ (see Section 6.2.2 for definition): $\alpha_2' = 0.05$, $\alpha_2' = 0.10$, $\alpha_2' = 0.20$ and $\alpha_2' = 1$. Note that using $\alpha_2' = 1$ is equivalent to not penalizing the similarity of the two regression coefficient matrices. We further consider a sequence of values for $\alpha_1$ that goes from 0.001 (highly sparse) to 0.5 (dense). For each combination of $\alpha_1$ and $\alpha_2'$ we fit the weighted fused regression lasso (WFRL) model and we measure the graph recovery by calculating the false positive rate and the true positive rate:

$$TPR = \frac{\sum_{l=1,2}\sum_{i,j} I(\hat{\beta}_{ij}^{(l)} \neq 0 \ \& \ \beta_{ij}^{(l)} \neq 0)}{\sum_{l=1,2}\sum_{i,j} I(\beta_{ij}^{(l)} \neq 0)}, \quad FPR = \frac{\sum_{l=1,2}\sum_{i,j} I(\hat{\beta}_{ij}^{(l)} \neq 0 \ \& \ \beta_{ij}^{(l)} = 0)}{\sum_{l=1,2}\sum_{i,j} I(\beta_{ij}^{(l)} = 0)}.$$

For every $\alpha_2'$, we approximate the AUC coefficient, which estimates the area under the curve given by the FPR and TPR relationship as function of $\alpha_1$ (with 1 being perfect recovery and 0.5 being recovery by chance). We consider 20 instances for each combination of sample size, dimension and scenario described in Section 6.5.2, and in Table 6.3 we present AUC estimates for the three models with their respective average ranks: rank = 1 is assigned to the best AUC, and rank = 4 is given to the worst AUC.




**(a)** n=25 **(b)** n=100

**(c)** n=250 **(d)** n=500

**Figure 6.4.** FPR vs $\alpha_2$: average (cross) + CI is plotted together with the expected values (triangle). For visualization reasons, x-axis and y-axis are not in the same scale (i.e. $2x : y$).

The ranks are added to directly compare the methods since the AUC levels are close to 1 and similar for some of the cases. This is due to the strong sparsity levels assumed at matrices $(\beta^{(l)})_{l=1,2}$, which lead to very small FPR values.

The joint methodology, especially when the number of differential coefficients is small ($\theta = 0.1$ and $\theta = 0.4$), produces better graph recovery levels than the standard lasso regression. In scenario 1, the joint model with large $\alpha_2'$ ($= 0.20$) turns out to achieve better rates than the other joint models with smaller $\alpha_2'$ whereas in scenario 2, $\alpha_2' = 0.05$ and $\alpha_2' = 0.10$ find the best results. In both scenarios, the best $\alpha_2'$ tends to increase with $\theta$, and we find that for the setting $n = 100$, $p = 170$ and scenario 1, $\alpha_2' = 1$ achieves the highest ranks. AUC levels are found to be quite similar among joint estimators, and present visible difference against the individual estimates.

### 6.5.6 Differential network recovery for the regression coefficient matrices

We compare the performance of FRL (fused regression lasso with constant weights) and WFRL (proposed weights for dependent datasets) for data generated as presented in Section 6.5.2 using a proportion of differential edges equal to $\theta = 0.4$. In order to make the structures of the estimated matrices comparable, we select estimated graphs (or $\lambda_1$ and $\lambda_2$) that have the same number of com-

**Table 6.2.** True positives vs false positives for weakest estimated triangle edges using WFGL + triangular motifs elimination at levels $\alpha = 0.01$, $\alpha = 0.03$ and $\alpha = 0.05$. The results are compared to the initial estimate, without the triangle correction (labeled as NO row).

| | common edges | | | | | | | | differential edges | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | 25 | | 100 | | 250 | | 500 | | 25 | | 100 | | 250 | | 500 | |
| | TP | FP | TP | FP | TP | FP | TP | FP | TP | FP | TP | FP | TP | FP | TP | FP |
| | dimension p=200 | | | | | | | | | | | | | | | |
| NO | 3.53 | 26.40 | 6.03 | 40.04 | 6.39 | 61.02 | 5.75 | 57.06 | 0.38 | 3.68 | 0.76 | 4.30 | 0.51 | 4.75 | 0.48 | 5.10 |
| $\alpha = .01$ | 0.00 | 0.00 | 0.09 | 0.53 | 1.72 | 0.97 | 3.38 | 1.26 | 0.00 | 0.00 | 0.00 | 0.12 | 0.07 | 0.48 | 0.12 | 0.76 |
| $\alpha = .03$ | 0.00 | 0.02 | 0.61 | 1.25 | 2.93 | 2.40 | 4.21 | 3.21 | 0.00 | 0.00 | 0.03 | 0.49 | 0.17 | 0.94 | 0.29 | 1.14 |
| $\alpha = .05$ | 0.00 | 0.08 | 1.21 | 1.87 | 3.67 | 4.30 | 4.55 | 5.37 | 0.00 | 0.00 | 0.09 | 0.82 | 0.19 | 1.27 | 0.37 | 1.47 |
| | dimension p=300 | | | | | | | | | | | | | | | |
| NO | 5.92 | 60.20 | 9.43 | 74.25 | 8.12 | 91.25 | 7.23 | 114.64 | .51 | 9.13 | 0.84 | 8.35 | 0.67 | 7.76 | 0.38 | 8.20 |
| $\alpha = .01$ | 0.00 | 0.00 | 0.30 | 0.71 | 2.33 | 1.09 | 4.37 | 1.87 | 0.00 | 0.00 | 0.00 | 0.16 | 0.09 | 0.67 | 0.08 | 0.92 |
| $\alpha = .03$ | 0.00 | 0.02 | 1.03 | 1.60 | 3.65 | 3.89 | 5.19 | 5.86 | 0.00 | 0.00 | 0.02 | 0.65 | 0.25 | 1.22 | 0.21 | 1.43 |
| $\alpha = .05$ | 0.04 | 0.10 | 1.93 | 3.28 | 4.52 | 7.48 | 5.63 | 11.11 | 0.00 | 0.02 | 0.08 | 1.05 | 0.36 | 1.83 | 0.28 | 2.09 |
| | dimension p=400 | | | | | | | | | | | | | | | |
| NO | 11.90 | 232.4 | 18.36 | 241.2 | 16.43 | 259.4 | 13.29 | 274.3 | 0.56 | 17.20 | 1.14 | 17.86 | 0.92 | 16.69 | 0.64 | 17.7 |
| $\alpha = .01$ | 0.00 | 0.08 | 0.7 | 1.7 | 4.31 | 3.36 | 7.49 | 5.46 | 0.00 | 0.00 | 0.00 | 0.44 | 0.05 | 0.98 | 0.25 | 1.21 |
| $\alpha = .03$ | 0.00 | 0.12 | 2.09 | 5.09 | 6.74 | 13.22 | 9.23 | 19.8 | 0.00 | 0.00 | 0.02 | 1.26 | 0.30 | 1.79 | 0.40 | 2.52 |
| $\alpha = .05$ | 0.01 | 0.37 | 3.75 | 12.19 | 8.30 | 27.03 | 9.95 | 38.5 | 0.00 | 0.00 | 0.14 | 1.95 | 0.43 | 2.85 | 0.47 | 4.27 |

mon edges and differential edges in the two approaches, i.e., we select the pair $[\lambda_1, \lambda_2]$ for the WFRL approach by setting the expected false positive rate by the parameters $[\alpha_1 = 0.05, \alpha_2' = 0.05]$ following the strategy proposed in Section 6.2.2, and we find $\lambda$'s such that the FRL graphs have the same sizes as WFRL. In total we use 200 instances for each model, 3 different sample sizes $n = \{25, 50, 100\}$, and two dimension sizes $p = \{120, 170\}$ with $q = p$.

The Youden's index for the estimated regression coefficient matrices is found by $\mathrm{YI}_\lambda^M = \mathrm{TP}_\lambda^M - \mathrm{FP}_\lambda^M$, $M = \mathrm{FRL}, \mathrm{WFRL}$, where $\mathrm{TP}_\lambda^M = \sum_{i,j} I[\hat{\beta}_{ij}^{(1)}(M) - \hat{\beta}_{ij}^{(2)}(M) \neq 0, \beta_{ij}^{(1)} - \beta_{ij}^{(2)} \neq 0)]$ and $\mathrm{FP}_\lambda^M = \sum_{i,j} I[\hat{\beta}_{ij}^{(1)}(M) - \hat{\beta}_{ij}^{(2)}(M) \neq 0, \beta_{ij}^{(1)} - \beta_{ij}^{(2)} = 0)]$ are the number of true positives and false positive of the estimated differential graphs with $\lambda = [\lambda_1, \lambda_2]$ and method $M$. Then we compute $\delta = \mathrm{YI}_\lambda^{\mathrm{WFRL}} - \mathrm{YI}_\lambda^{\mathrm{FRL}}$, which defines the Youden's index differences between the two methods to estimate the joint networks. In Table 6.4 we present the average difference (with a $t$-test p-value) and also the average sign of the differences $\delta$ (with a Wilcoxon test p-value) for network pattern described in Scenario 1 and Scenario 2. The proposed method, that assumes a dependence structure, achieves better TP-FP ratios for the differential network than the original FRL for any combination of sample size and dimension, being highly significant for $n \geq 50$. However, these represent very small differences in magnitude as the total number of possible non-zero coefficients is of $O(10000)$.

## 6.6 Estimation of sparse networks using gene expression data

We apply the proposed WFGL method with $\lambda_1$ and $\lambda_2$ selected by the FDR procedure (Section 6.2.2), and with triangular motif correction (Section 6.4) to two different real case studies of gene expression data. We present detailed analysis for the first dataset, which contains the gene expression profiling of 82 patients with the psoriasis vulgaris disease in a paired lesional and non-lesional samples (Suárez-Fariñas et al., 2012). We also show the main results of the analysis of a gene expression dataset that represents a paired tumor and healthy samples from 60 female non-smoker patients with lung

**Table 6.3.** Average ranks for AUC estimates (and their AUC average value) for models generated as defined in scenario 1 (diagonal matrices of regression coefficients) and scenario 2 (proportional coefficients to precision matrix). Rank = 1 corresponds to the best AUC and rank = 4 is for the worst AUC.

| | Scenario 1 | | | Scenario 2 | | |
|---|---|---|---|---|---|---|
| $\theta$ | 0.1 | 0.4 | 0.7 | 0.1 | 0.4 | 0.7 |
| | | | n=25, p=120 | | | |
| joint ($\alpha_2' = 0.05$) | **1.8** (.96) | 2.1 (.94) | 3.8 (.88) | **1.5** (.68) | 2.1 (.70) | 2.5 (.69) |
| joint ($\alpha_2' = 0.10$) | 2.1 (.96) | 2.1 (.94) | 2.8 (.89) | 1.8 (.68) | **1.7** (.70) | 2.0 (.69) |
| joint ($\alpha_2' = 0.20$) | 2.1 (.96) | **1.8** (.94) | **1.3** (.91) | 2.7 (.68) | 2.2 (.69) | **1.5** (.70) |
| ind. ($\alpha_2' = 1$) | 4.0 (.91) | 4.0 (.89) | 2.1 (.90) | 4.0 (.65) | 4.0 (.67) | 4.0 (.67) |
| | | | n=25, p=170 | | | |
| joint ($\alpha_2' = 0.05$) | 2.1 (.97) | 2.9 (.93) | 3.9 (.87) | **1.8** (.67) | **1.5** (.70) | 2.1 (.70) |
| joint ($\alpha_2' = 0.10$) | **1.8** (.97) | 2.0 (.94) | 2.8 (.89) | 1.9 (.67) | 2.2 (.70) | **1.9** (.70) |
| joint ($\alpha_2' = 0.20$) | 2.1 (.97) | **1.1** (.95) | **1.5** (.90) | 2.3 (.67) | 2.3 (.69) | 2.0 (.69) |
| ind. ($\alpha_2' = 1$) | 4.0 (.91) | 4.0 (.91) | 1.8 (.90) | 4.0 (.65) | 4.0 (.68) | 4.0 (.68) |
| | | | n=50, p=120 | | | |
| joint ($\alpha_2' = 0.05$) | 2.4 (.99) | 2.9 (.99) | 3.7 (.97) | **1.8** (.67) | **1.5** (.68) | 2.1 (.70) |
| joint ($\alpha_2' = 0.10$) | **1.7** (.99) | 2.1 (.99) | 2.4 (.98) | 1.9 (.67) | 2.2 (.68) | **1.9** (.70) |
| joint ($\alpha_2' = 0.20$) | 1.9 (.99) | **1.6** (.99) | **1.9** (.98) | 2.3 (.67) | 2.3 (.68) | 2.0 (.69) |
| ind. ($\alpha_2' = 1$) | 4.0 (.98) | 3.4 (.98) | 2.0 (.98) | 4.0 (.65) | 4.0 (.66) | 4.0 (.68) |
| | | | n=50, p=170 | | | |
| joint ($\alpha_2' = 0.05$) | 2.0 (.99) | 2.8 (.99) | 3.7 (.97) | **1.3** (.72) | **1.8** (.72) | 2.0 (.72) |
| joint ($\alpha_2' = 0.10$) | **1.8** (.99) | 2.3 (.99) | 2.6 (.97) | 1.8 (.71) | 1.8 (.72) | **1.8** (.72) |
| joint ($\alpha_2' = 0.20$) | 2.2 (.99) | **1.1** (.99) | **1.5** (.98) | 2.9 (.71) | 2.4 (.72) | 2.2 (.72) |
| ind. ($\alpha_2' = 1$) | 4.0 (.98) | 3.8 (.97) | 2.2 (.98) | 4.0 (.68) | 4.0 (.69) | 4.0 (.70) |
| | | | n=100, p=120 | | | |
| joint ($\alpha_2' = 0.05$) | 2.5 (.99) | 2.8 (.99) | 3.4 (.99) | **1.3** (.77) | **1.3** (.78) | **1.6** (.79) |
| joint ($\alpha_2' = 0.10$) | **2.0** (.99) | 2.4 (.99) | 2.5 (.99) | 2.0 (.77) | 2.0 (.78) | 1.8 (.79) |
| joint ($\alpha_2' = 0.20$) | 2.2 (.99) | **2.2** (.99) | 2.4 (.99) | 2.7 (.77) | 2.7 (.78) | 2.6 (.79) |
| ind. ($\alpha_2' = 1$) | 3.3 (.99) | 2.6 (.99) | **1.7** (.99) | 4.0 (.71) | 4.0 (.73) | 4.0 (.75) |
| | | | n=100, p=170 | | | |
| joint ($\alpha_2' = 0.05$) | **1.8** (.99) | 2.7 (.99) | 3.4 (.99) | **1.2** (.77) | **1.3** (.79) | **1.9** (.77) |
| joint ($\alpha_2' = 0.10$) | 2.1 (.99) | 2.4 (.99) | 3.0 (.99) | 2.0 (.77) | 1.7 (.79) | **1.9** (.77) |
| joint ($\alpha_2' = 0.20$) | 2.3 (.99) | **1.7** (.99) | 2.1 (.99) | 2.8 (.77) | 3.0 (.78) | 2.2 (.77) |
| ind. ($\alpha_2' = 1$) | 3.8 (.99) | 3.2 (.99) | **1.6** (.99) | 4.0 (.71) | 4.0 (.73) | 4.0 (.73) |

cancer (Lu et al., 2010). In both cases, there are 19,507 different genes which have been identified by the biomaRt R package (Durinck et al., 2005). In the original data, some genes are represented by more than one probe. These are aggregated at the gene level by taking the average. The main objective is to make inference about the gene interconnections in the two medical conditions and relate common and differential estimated networks to functions in biological processes. Moreover, the WFRL approach is applied to colon cancer data in Chapter 8.

### 6.6.1 Network analysis of psoriasis vulgaris disease gene expression data

*Reduction of the number of genes for network analysis*

For computational needs in the joint estimation procedures, we reduce the dimension of the data set by considering two filters with the objective to keep only the most relevant genes in the gene dependence networks, i.e., we select highly correlated genes and differentially correlated genes. As a first filter we use the hypothesis testing problem described in Section 4.4.3 that assesses if a correlation matrix row is the identity vector. As a second filter we consider the hypothesis testing problem described in Section 4.4.1 for equality of two correlation rows. In both cases, we employ the average of squares test statistic. The null distribution is approximated using 300 permuted samples.

We correct the p-values by multiple testing using the false discovery rate (FDR) approach of

**Table 6.4.** Youden Index differences $\delta = \mathrm{YI}_{\lambda}^{\mathrm{WFRL}} - \mathrm{YI}_{\lambda}^{\mathrm{FRL}}$ between WFRL and FRL algorithm for data generated by scenario 1 (diagonal matrices of regression coefficients) and by scenario 2 (proportional coefficients to precision matrix). The WFRL method obtains better rates than FRL for $n \geq 50$.

Scenario 1 pattern

| | p = 120 | | p = 170 | |
|---|---|---|---|---|
| $n$ | $\delta$ (p-val) | $sgn(\delta)$ (p-val) | $\delta$ (p-val) | $sgn(\delta)$ (p-val) |
| 25 | .24 (.03) | .11 (.03) | .06 (0.36) | .04 (0.33) |
| 50 | 1.9 (< .01) | .62 (< .01) | 2.52 (< 0.01) | .69 (< 0.01) |
| 100 | 4.5 (< .01) | .93 (< .01) | 6.86 (< 0.01) | .95 (< 0.01) |

Scenario 2 pattern

| | p = 120 | | p = 170 | |
|---|---|---|---|---|
| $n$ | $\delta$ (p-val) | $sgn(\delta)$ (p-val) | $\delta$ (p-val) | $sgn(\delta)$ (p-val) |
| 25 | .01 (0.42) | .01 (0.43) | .04 (0.31) | .05 (0.24) |
| 50 | .68 (< .01) | .32 (< .01) | .94 (< 0.01) | .34 (< 0.01) |
| 100 | 1.17 (< .01) | .42 (< .01) | 1.68 (< 0.01) | .46 (< 0.01) |

Benjamini and Hochberg (1995). The following genes are selected with the threshold of 0.01

$$g^* = \{g : \text{p-val}(g)^{NL} < 0.001\} \cup \{g : \text{p-val}(g)^{L} < 0.001\} \cup \{g : \text{p-val}(g)^{D} < 0.001\},$$

where p-val$(g)^{NL}$ and p-val$(g)^{L}$ are the adjusted p-values using the first filter for healthy and lesional datasets respectively, and p-val$(g)^{D}$ are the adjusted p-values for the difference matrix using the second filter.

The total number of selected genes is 17,967, which is a reduction of the 8% of the original variables (for extended results see Section 4.6 ). We further use a clustering procedure on the reduced dataset to estimate joint networks separately for different groups of genes. We consider the hierarchical clustering algorithm presented in Müllner (2013) since it provides a fast procedure even for large dimensions. We use 1 minus the matrix of absolute correlations for healthy genes as dissimilarity matrix to find 6 large clusters of size [5335, 1697, 781, 879, 1017, 4694] genes. Other clusters are found but their sizes are very small (less than 100 genes) and are not considered for estimation.

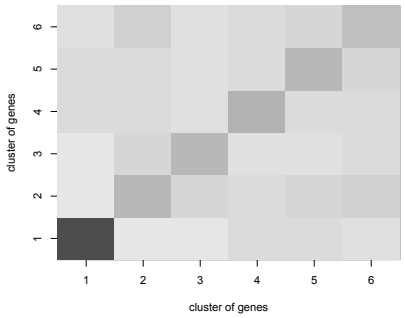



**Figure 6.5.** Heatmap of gene clusters linear dependence for psoriasis data: square darkness is related to the average of absolute correlation within and between clusters.

*Network estimation of lesional and healthy gene expression data*

We fit the weighted fused graphical lasso model to each of the 6 clusters of genes defined above, so we assume conditional independence for genes between clusters, as the estimation of the whole network requires extremely demanding computational efforts. We use error rates $\alpha_1$ and $\alpha_2'$ (defined in Section 6.2.2) to tune the penalization parameters $\lambda_1$ and $\lambda_2$. For $\alpha_1$ we set the underlying expected number of false positive edges (EFP) with EFP = 200, 150, 100, 100, 100, 200 respectively for each cluster as we found these represent well the graphical complexity of the observed cluster sizes. Then, $\alpha_{1k} = EFP/p_k'$ with $p_k' = p_k(p_k-1)/2$ ($p_k$ cluster size for $k = 1:6$). By setting $\alpha_1$ in this way, we permit more false positives for small dimensions to control the graph complexity. Note that if we were going to consider equal $\alpha_1$ for all clusters, for the EFP = 100 of cluster 3 we would expect about EFP = 5000 for cluster 1, which would make the graphical interpretation fairly difficult. Besides, we use three different values for $\alpha_2'$ which are specified in Table 6.5.

Table 6.5 provides the number of estimated edges common to the two medical conditions and the number of differential edges: "healthy only" for edges only present in the network for healthy samples; and "les only" for edges only present in the network for lesional samples. The total number of edges is much larger than the expected number of false positives which suggests reasonable confidence in the results. Moreover, the number of differential edges is remarkably larger for healthy samples than for lesional samples in cluster 1, 2, 4, 5 and 6, and the other way around for cluster 3.

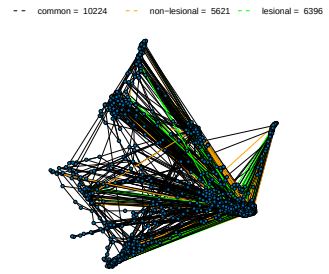
**Table 6.5.** Number of edges for common networks and differential edges using similarity tuning parameters $\alpha_2' = 0.001$, $\alpha_2' = 0.01$ and $\alpha_2' = 0.05$ in psoriasis dataset.

| | Cluster 1 | | | Cluster 2 | | | Cluster 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.001 | 0.01 | 0.05 | 0.001 | 0.01 | 0.05 | 0.001 | 0.01 | 0.05 |
| common | 11,771 | 10,224 | 8,891 | 2,413 | 2,407 | 2,388 | 1,028 | 1,021 | 1,000 |
| healthy only | 3,646 | 5,621 | 7,737 | 0 | 0 | 11 | 7 | 23 | 44 |
| les only | 4,259 | 6,339 | 8,493 | 2 | 8 | 17 | 4 | 10 | 16 |

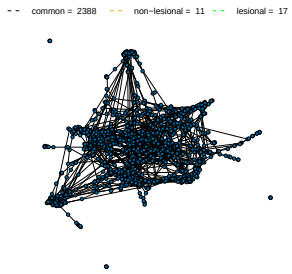
| | Cluster 4 | | | Cluster 5 | | | Cluster 6 | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.001 | 0.01 | 0.05 | 0.001 | 0.01 | 0.05 | 0.001 | 0.01 | 0.05 |
| common | 946 | 920 | 897 | 1,320 | 1,320 | 1,311 | 7,674 | 7,633 | 7,485 |
| healthy only | 0 | 0 | 4 | 0 | 1 | 5 | 3 | 18 | 69 |
| les only | 14 | 29 | 55 | 1 | 5 | 10 | 29 | 70 | 136 |

Figure 6.6 shows the graphical representation of some of the estimated networks. The black edges are common edges, whereas in orange there are "healthy only" edges and in green there are "les only" edges. In general, in almost all clusters we detect presence of hub genes (genes with much higher degree than the rest). Furthermore, we can see a clustered graph structure in each estimated networks, which could be expected in biological data (Eisen and Spellman, 1998) with some specific groups of genes that are uniquely present in one medical condition. For instance, the genes with a largest number of differential edges are ABCC6P2, CALB1, CATSPER3, CYP1A2, IDI2-AS1, JARID2-AS1, KRT3, NBAS, NPY4R, PHACTR2-AS1, SYT13, TNNC2, TRAV20, UNC13C, XKR6, DICER1-AS1, LYPD5, OSR2, RIMBP2, SIAE, USH1G , C1orf61, DNMBP, PCDHB11, SNORD38A, BEND7, FOXD3 (for "healthy only") and BAALC-AS2, C2CD4A, CD244, CDH17, CFLAR-AS1, CPB1, DNAH2, FCRL3, FITM1, FRRS1, IGHD, KCNK17, LINC00491, LINC00847, PAEP, PRR15, PWRN1, RNF144A-AS1, SLC26A4-AS1, SLC6A18,

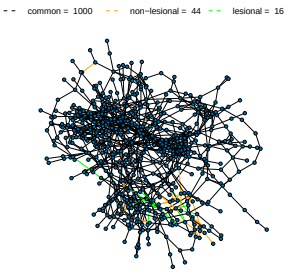
STRA6, SYNPO2L, TAS2R38, TECRL, TRG-AS1, TTTY15, ZFP42, ZNF671, ZNHIT2, ZP4, ARV1, CHERP, ERICH1, PRKCE, REEP3, SGPP2, ZCCHC10, GSG1L, PSMA7, PKN3, ZNF438, LHFPL2, RNU6-125P, CCDC168, FNTA, GIPC1 (for "les only"). Most of the genes in this list were not identified as important genes (differentially expressed analysis) in the study by Suárez-Fariñas et al. (2012) but are found to be relevant for gene interaction network.



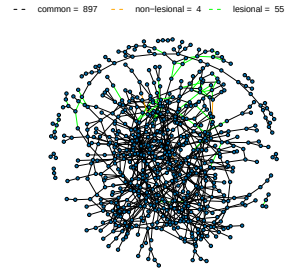

**(a)** Network estimation in cluster 1



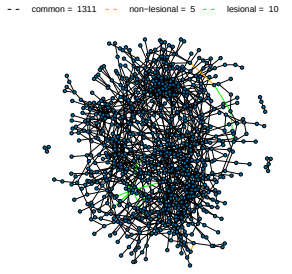

**(b)** Network estimation in cluster 2



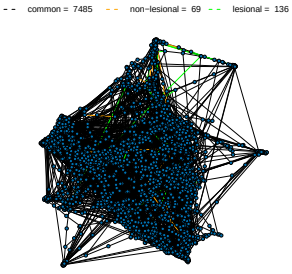

**(c)** Network estimation in cluster 3




**(d)** Network estimation in cluster 4




**(e)** Network estimation in cluster 5




**(f)** Network estimation in cluster 6

**Figure 6.6.** Estimated joint networks for four groups of genes in psoriasis dataset: in black there are the common edges and in orange ("healthy only") and green ("les only") the differential connections.

*Integration with biological pathway lists*

We are particularly interested in knowing how standard gene pathways change in different medical conditions. To assess which biological processes might be linked to changes in the gene connections we download 1,320 gene sets from the MSig database (Subramanian et al., 2005), which represent canonical pathways compiled from two sources: KeGG (Kanehisa et al., 2016) and Reactome (Milacic et al., 2012). To integrate and analyze the estimated networks within the pathway lists, we count which pairs of connected genes in the estimated networks are both present in a specific pathway list (see Table 6.6). Using the 17,967 genes as background, we find that approximately 1% of estimated connections are expected to be included by chance. Thus, we also evaluate how likely it is to obtain at least the same number of biological relevant connections in a random process. Common network associations are significantly present in all pathways except cluster 4. Moreover, differential edges overlapping with the pathway list could be expected by chance in all clusters except for cluster 1

(significance level of 0.01).

**Table 6.6.** Total number of estimated edges whose pair of genes are both in the same pathway list (p-value) using psoriasis data.

| | Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.001 | 0.01 | 0.05 | 0.001 | 0.01 | 0.05 |
| common | 121 (.35) | 111(.17) | 97 (.17) | 115 (< .01) | 115 (< .01) | 116 (< .01) |
| healthy only | 26 (.96) | 47 (.90) | 54 (.86) | 0 | 0 | 1(.10) |
| les only | 57 (.62) | 61 (.64) | 67 (.67) | 0 | 0 | 1 (.16) |

| | Cluster 3 | | | Cluster 4 | | |
|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.001 | 0.01 | 0.05 | 0.001 | 0.01 | 0.05 |
| common | 82 (< .01) | 83 (< .01) | 82 (< .01) | 15 (.06) | 15 (.05) | 14 (.06) |
| healthy only | 0 | 0 | 0 | 0 | 0 | 0 |
| les only | 0 | 1 (.10) | 2 (.01) | 0 | 0 | 0 |

| | Cluster 5 | | | Cluster 6 | | |
|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.001 | 0.01 | 0.05 | 0.001 | 0.01 | 0.05 |
| common | 82 (< .01) | 82 (< .01) | 82 (< .01) | 332 (< .01) | 332 (< .01) | 328 (< .01) |
| healthy only | 0 | 0 | 0 | 0 | 0 | 0 |
| les only | 0 | 0 | 0 | 0 | 0 | 1 (.74) |

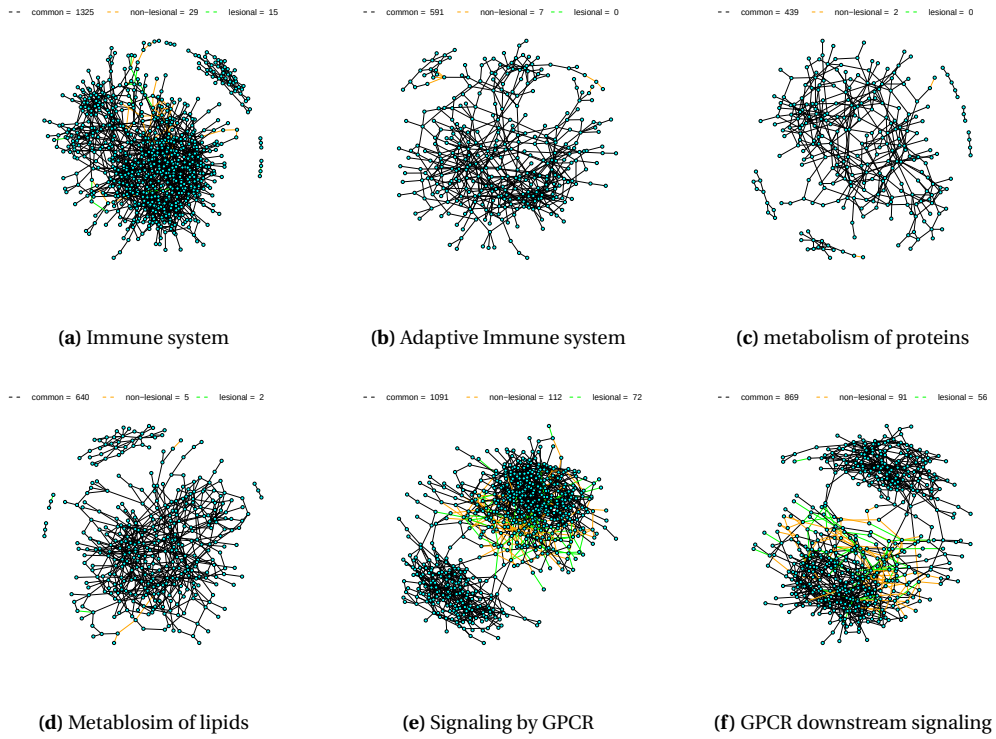
We perform further investigation for genes in six of the most important canonical pathways: *immune system, adaptive immune system, metabolism of proteins, metabolism of lipids and lipoproteins, signaling by GPCR* and *GPCR downstream signaling*. We estimate joint CD structures only considering the genes in each of the six pathways. In Figure 6.7 we show the graphical representation of *immune system, metabolism of proteins* and *signaling by GPCR* using $\alpha_1$ so the expected number of false positive edges is about 100 and we set $\alpha_2' = 0.05$. In all cases we observe more "healthy only" estimated edges than "les only" edges, which is a behavior seen in the previous section exclusively in cluster 3.

A permuted samples based procedure presented in Appendix B.3 is used to assess the uncertainty in the number of estimated differential edges under the hypothesis of equal conditional dependence structures using 100 instances in every pathway list. For the immune system, the number of "healthy only" edges is not expected by chance (with non of the permuted sample estimations exceeding the 29 edges). Similarly for the adaptive immune system, the maximum number of "healthy only" edges in permuted samples is 5 for the 7 obtained using the original data. In both metabolism pathways, 20% of the permuted samples statistics exceed the total number of "healthy only" edges. Finally, for both GPCR pathways "healthy only" edges are much more present than expected by chance. In contrast, in all pathway lists, the number of "les only" edges is largely exceeded by the replicates.

### 6.6.2 Network analysis of lung cancer gene expression data

*Reduction of the number of genes for network analysis*

We applied the same procedure presented in Section 6.6.1 to the lung cancer gene expression data. The datasets are reduced to a total of 15,459 genes (80% of the original dimension). Clustering is applied to the reduced data leading to 6 large clusters of size [942, 2302, 1722, 784, 768, 6276] genes respectively and other small clusters that are not considered for estimation. In Figure 6.8 there is the heatmap of the average correlation between and within clusters.




**(a)** Immune system **(b)** Adaptive Immune system **(c)** metabolism of proteins

**(d)** Metablosim of lipids **(e)** Signaling by GPCR **(f)** GPCR downstream signaling

**Figure 6.7.** Estimated joint networks using psoriasis data in pathways (a) Immune system, (b) Adaptive Immune system (c) metabolism of proteins (d) metabolism of lipids (e) signaling by GPCR and (f) GPCR downstram signaling. In black there are the common edges and in orange ("healthy only") and green ("les only") the differential connections.

*Network estimation of lesional and healthy gene expression data*

We fit a weighted fused graphical lasso to each of the 6 clusters of genes with different values of error rates $\alpha_1$ and $\alpha_2'$. We use $\alpha_1 = EFP/p_k'$, $p_k' = p_k(p_k-1)/2$ ($p_k$ cluster size), with EFP = $100, 150, 150, 100, 100, 200$ respectively for each cluster and several $\alpha_2'$ specified in Table 6.7.

**Table 6.7.** Number of edges for common networks and differential edges using similarity tuning parameters $\alpha_2' = 0.001$, $\alpha_2' = 0.01$ and $\alpha_2' = 0.05$ in lung cancer dataset.

| | Cluster 1 | | | Cluster 2 | | | Cluster 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 |
| common | 748 | 738 | 724 | 1,788 | 1,765 | 1,726 | 1,646 | 1,619 | 1,597 |
| heal only | 0 | 1 | 3 | 2 | 9 | 16 | 0 | 5 | 8 |
| tum only | 0 | 1 | 8 | 12 | 36 | 53 | 4 | 17 | 36 |

| | Cluster 4 | | | Cluster 5 | | | Cluster 6 | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 |
| common | 610 | 610 | 503 | 678 | 670 | 679 | 10,050 | 8,912 | 8,016 |
| heal only | 0 | 1 | 2 | 0 | 0 | 1 | 1,143 | 2,124 | 2,840 |
| tum only | 0 | 1 | 2 | 2 | 4 | 9 | 1,426 | 2,622 | 3,444 |

We observe a common behavior of more "tumor only" differential edges than "healthy only" differential edges for the six clusters except cluster 4, where not many differential edges are estimated. Figure 6.9 presents the network representation of the estimated precision matrices. Genes with more than 15 non-common edges are ARHGAP11A, C14orf105, FAM47A, IMPG2, LINC01537,

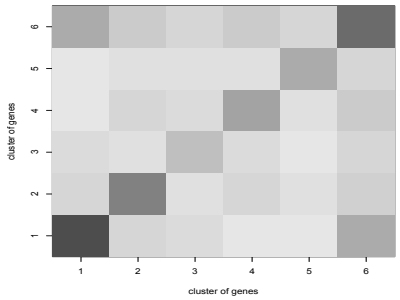
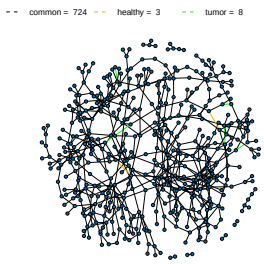
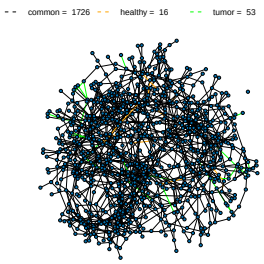



**Figure 6.8.** Heatmap of gene clusters linear dependence for for lung cancer data: square darkness is related to the average of absolute correlation within and between clusters.

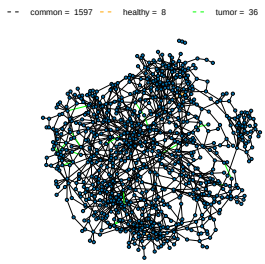
LINC01592, MIP, PAPOLB, PIWIL2 and TRIM42 (only healthy network), and C12orf42, LINC00648, PEX5L, PRR23D2, RELL1, RPTN and SLC30A8 (only tumor network).



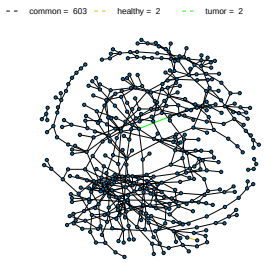

**(a)** Network estimation in cluster 1



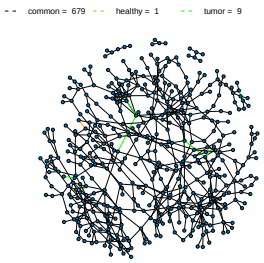

**(b)** Network estimation in cluster 2



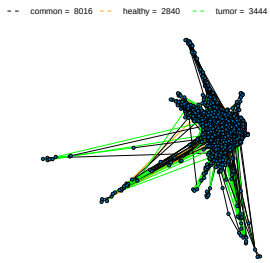

**(c)** Network estimation in cluster 3




**(d)** Network estimation in cluster 4




**(e)** Network estimation in cluster 5




**(f)** Network estimation in cluster 6

**Figure 6.9.** Estimated joint networks for four groups of genes in lung cancer dataset: in black there are the common edges and in orange ("healthy only") and green ("les only") the differential connections.

*Integration with biological pathway lists*

We integrate the estimated networks with 1,320 pathway lists by counting the number of estimated gene associations whose pair of genes is present in a specific pathway list. Common networks have significant overlap with the pathway lists for all 6 clusters. Differential edges overlap could be observed by chance in "healthy only" edges (found using testing approach in Appendix B.3). For "tumor only"

edges, cluster 5 and 6 present significant overlap.

**Table 6.8.** Total number of estimated edges whose pair of genes are both in the same pathway list (p-value) using lung cancer data.

| | Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 |
| common | 29 (< .01) | 29 (< .01) | 30 (< .01) | 74 (< .01) | 72 (< .01) | 73 (< .01) |
| heal only | 0 | 0 | 0 | 0 | 0 | 0 |
| tum only | 0 | 0 | 1 (.07) | 1 (.13) | 2 (.05) | 2 (.10) |

| | Cluster 3 | | | Cluster 4 | | |
|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 |
| common | 74 (< .01) | 72 (< .01) | 74 (< .01) | 45 (< .01) | 45 (< .01) | 44 (< .01) |
| heal only | 0 | 0 | 0 | 0 | 0 | 0 |
| tum only | 0 | 1 (.16) | 3 (.01) | 0 | 0 | 1 (.01) |

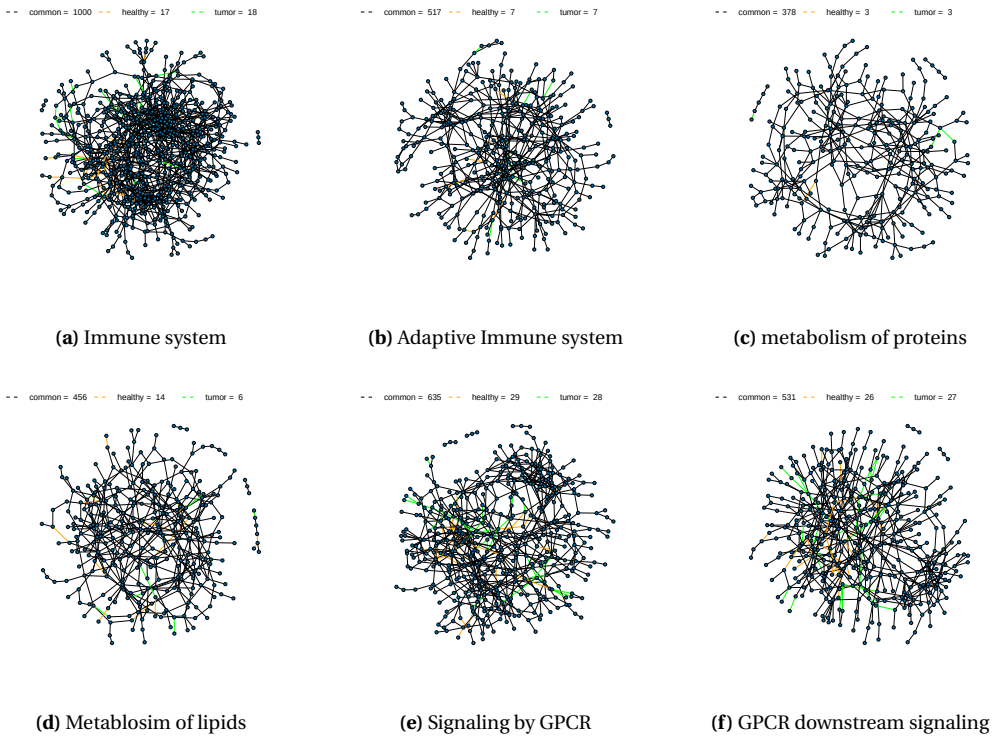
| | Cluster 5 | | | Cluster 6 | | |
|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 |
| common | 32 (< .01) | 32 (< .01) | 31 (< .01) | 201 (< .01) | 179 (< .01) | 167 (< .01) |
| heal only | 0 | 0 | 0 | 7 (.94) | 12 (.99) | 18 (.99) |
| tum only | 2 (< .01) | 2 (< .01) | 2 (< .01) | 21 (.05) | 49 (< .01) | 60 (< .01) |

As for the psoriasis data, we estimate the gene networks using subgroups of genes determined by six canonical pathways: *immune system*, *adaptive immune system*, *metabolism of proteins*, *metablosim of lipids and lipoproteins*, *signaling by GPCR* and *GPCR downstream signaling*. In all estimated pathway networks except *immune system* we observe more "healthy only" estimated edges than "tumor only" edges. This contrasts with the results we obtained for the six estimated networks (by clusters) where "tumor only" edges are more frequently estimated than "healthy only" edges.

## 6.7 Discussion

Motivated by genomic data where gene expression is obtained for the same individual in two different medical conditions, in this chapter we develop a weighted fused graphical lasso method (WFGL) that jointly estimates two precision matrices. As in the fused graphical lasso (FGL) approach proposed by Danaher et al. (2014), we consider a penalized maximum marginal likelihood estimator that assumes both sparsity and similarity between precision matrices. To account for dependence between observations, we extend FGL by weighting the similarity tuning parameters for each pair of variables. Our method, WFGL, improves the recovery rates of the original FGL for sufficiently large sample sizes ($n \geq 100$) in simulated data. For small sample size ($n = 25$) we find similar rates for WFGL and FGL as the variances of the estimators of the correlation coefficients $\psi_{ij}$, which are needed to weight the tuning parameters, can be quite high. WFGL also provides a less biased procedure than FGL in the sense that all differential connections with same magnitude in the differential precision matrix have approximately the same chance to be recovered (see Appendix B.5.

Furthermore, we propose a method to simultaneously estimate two regression coefficient matrices, and their underlying graphical structure, corresponding to samples in two different classes, whose observations can be paired, and where both response and explanatory variables are high-dimensional. The method, which is called WFRL, finds a penalized marginal least squares estimator with a lasso




**(a)** Immune system **(b)** Adaptive Immune system **(c)** metabolism of proteins

**(d)** Metablosim of lipids **(e)** Signaling by GPCR **(f)** GPCR downstream signaling

**Figure 6.10.** Estimated joint networks using psoriasis data in pathways (a) Immune system, (b) Adaptive Immune system (c) metabolism of proteins (d) metabolism of lipids (e) signaling by GPCR and (f) GPCR downstram signaling. In black there are the common edges and in orange ("healthy only") and green ("tumor only") the differential connections.

penalization term to encourage sparsity in the estimated networks as well as a fused penalization term to favor similarity between regression coefficients, and it is also solved employing an ADMM based algorithm. The proposed joint estimator is proven to give better network recovery rates than estimating the two networks separately when the true regression coefficient matrices are fairly similar. This is not a rare assumption in our application to genomic data where even for such different states as healthy and tumor tissues, we expect a large part of the gene connections to be equal. Moreover, we have applied a correction on the fused penalization to account for data settings where observations in the two classes are paired. This adjustment is found to improve the recovery of differential networks for paired data using simulations.

We present a method to select the tuning parameters in the two joint estimation algorithms, WFGL and WFRL, which is motivated by practical needs for controlling the expected false positive rates. We transform the selection problem to the more intuitive selection of expected proportion of false positive edges (EPFR) which works well for reasonably sparse graphs. This requires the assumption of normality in the estimated precision matrix elements and should be tested for other datasets. If the assumptions hold, we see in the simulated data analysis that the proposed method produces results near the desired EFPR for a sufficiently large sample size. The numerical integration of expression (6.13) is computed to control the error rate for sparsity $\alpha_1$. To avoid this computation, we have investigated

using an upper bound for $\alpha_1$ instead of $\alpha_1$, i.e., see Fayed and Atiya (2014), but the simplification found gave very crude results. Finding accurate bounds is left as future work.

Finally we address the problems of FGL and WFGL in estimating triangular motif graph structures using an hypothesis testing approach on the weakest edge in a triangle of variables just after the estimation process. Using simulated data we corroborate that our proposed strategy reduces the number of false positive edges without missing many estimated true positives.

The analysis of the motivating gene expression data with healthy and lesional (and also tumor) classes underlines some interesting results. We estimate 6 joint networks corresponding to 6 clusters of genes in the two datasets. As a general pattern, we observe that in each cluster, genes interact between each other in groups, suggesting a clustering sub-structure. Connections between genes in lesional tissue appear to occur more often than in healthy tissue. Furthermore, pathway integration analysis suggests that common edges, which are estimated using a larger effective sample size than the original number of patients, have a strong significant overlap with some of the considered pathway lists. Main pathways listed such as *immune system* or *GPCR* contain more "healthy only" edges than "lesional/tumor".

We have realized that, recently, Cai et al. (2016) proposed a method to estimate multiple precision matrices which proved to outperform FGL in graph structure recovery using simulations. As future work, we will compare our methods to such novel proposal. Besides, we could use these techniques to other type of similarity penalizations, i.e., the group lasso approach (Guo et al. 2011), and we could extend the methods to jointly estimating $K$ precision (or regression coefficient) matrices, with $K > 2$, for datasets with paired observations.

# Chapter 7

# ldstatsHD: an R package for estimation and testing linear dependence in high-dimensional data

## 7.1 Motivation for creating ldstatsHD

Omics datasets obtained as a result of genomic, metabolomic or proteomic experiments produce generally cases of high-dimensional data, where the dimension (e.g., number of genes) is much larger than the sample size (e.g., number of patients). The analysis of this particular type of data has been the focus of attention of many authors in the statistics literature. In Chapter 2, we review some of the statistical approaches in the context of testing and estimating linear dependence measures related to the correlation matrix and its inverse matrix when data are high-dimensional. Some of these methods are implemented in the statistical software R as part of the CRAN (`https://cran.r-project.org/`) and Bioconductor (`https://bioconductor.org`) repositories. Some of the most relevant R packages are: **WGCNA** (Langfelder and Horvath, 2008) employs the sample correlation matrix for network reconstruction, module detection (clustering) and statistical significance; **DiffCorr** (Fukushima, 2013) contains an hypothesis testing approach for equality of correlation coefficients with false discovery rate (FDR) multiple testing significance correction; **MixOmics** (Lê Cao et al., 2009; González et al., 2012) consists of different multivariate analysis procedures as principal components analysis (PCA), partial least square (PLS), independent principal component analysis (IPCA) and other visualization techniques for high-dimensional datasets; A remarkable R package for the estimation of partial correlation matrices and their underlying conditional dependence networks is the package **huge** (Zhao et al., 2012), which estimates the inverse covariance matrix by lasso penalized maximum likelihood (Meinshausen and Bühlmann, 2006; Friedman et al., 2007); **camel** (Li et al., 2013) implements a sparse precision matrix estimator based on the tiger algorithm presented in Liu and Wang (2012); **JGL**

(Danaher et al., 2014) extends the lasso methodology to jointly estimating multiple partial correlation matrices; Finally, **MRCE** (Rothman et al., 2010) is an R package that finds a sparse estimator of a multivariate regression coefficient matrix when both response and predictors are high-dimensional.

In this chapter we present the R package **ldstatsHD**, which consists of functions with statistical methods for the estimation and testing of multiple correlation matrices, precision matrices and regression coefficient matrices from high-dimensional data when these matrices can come from paired observations. The methodological and algorithmic contributions are mainly discussed in Chapters 4, 5, and 6. With the creation of this package we intend to document all the generated code and make it accessible to the R community for its use.

The chapter is organized as follows. In Section 7.2 we separate the package in three modules that correspond to data simulators, testing methods and estimation methods. In Section 7.3 we describe the main functions in each of these modules. We complete the description of the package in Section 7.4, where we present the user interface of **ldstatsHD** by exploiting several simulated data case studies.

## 7.2 Modules of ldstatsHD

The package **ldstatsHD** can be installed and loaded from the comprehensive R archive Network (CRAN) by entering in the R command

```
R> install.packages("ldstatsHD")
R> library(ldstatsHD)
```

By doing so, some other packages/functions used in **ldstatsHD** are automatically downloaded. For instance, it depends on the packages **huge** (Zhao et al., 2012) and **igraph** (Csárdi and Nepusz, 2006). Moreover, it imports functions from packages **evd** (Stephenson, 2002), **fExtremes** (Wuertz, 2013), **corpcor** (Schäfer et al., 2015), **Matrix** (Bates and Maechler, 2016), **MASS** (Venables and Ripley, 2002), **robustbase** (Rousseeuw et al., 2016), **VGAM** (Yee, 2010), **cluster** (Maechler et al., 2016), **RBGL** (Carey et al., 2016), **camel** (Li et al., 2013) and **qvalue** (Storey et al., 2015). Alternatively, the root files are available at `http://cran.r-project.org/packages=ldstatsHD`. The package is under the public license GPL-3 and the code is implemented using the S3 class (which is the most employed class in the R community).

Below, we introduce the main functions available in **ldstatsHD** which can be classified in three modules: data simulators, testing methods and estimation methods.

**Module 1. Data simulators:** it provides two functions for generating positive definite partial correlation matrices. The first is `pcorSimulator`, which simulates a single partial correlation matrix in which the underlying graph structure can be defined by power-law, hub-based or random graphs (see Section 5.5.1). The second function is `pcorSimulatorJoint`, which extends `pcorSimulator` for generating a joint partial correlation matrix that relates two classes

of observations. Several paired data structures, which are discussed in Section 2.2, are proposed to account for dependence between the two datasets.

**Module 2. Testing methods:** this includes statistical methods that test global dependence characteristics. It implements a test for equality of two correlation matrices as well as a test for identity correlation matrix. These methods are described in Section 4.2 and Section 4.4.2 respectively and are coded in the function `eqCorrMatTest`. Moreover, it provides a test for equality of two correlation matrix rows as well as a test to determine if a variable is not correlated to any other variable in a dataset. These approaches are presented in Section 4.4.1 and Section 4.4.3 respectively and can be found in the function `eqCorTestByRows`.

**Module 3. Estimation methods:** joint estimation of two precision matrices is implemented in the function `wfgl` and joint estimation of two regression coefficient matrices is found in function `wfrl`. These use a weighted-fused lasso penalized maximum likelihood estimator that enforces both sparsity and similarity between estimated matrices (see Chapter 6). **ldstatsHD** also contains approaches to select the sparsity tuning parameter of graphical lasso estimators (which can be found by packages **huge** or **camel.tiger**). Several risk functions based on characteristics of the estimated networks are available (see Chapter 5). Among others, statistics that measure clustering structure or network connectivity are used to choose an estimated network for its analysis in function `lambdaSelection`.

All considered approaches permit cases where datasets come from paired observations. For visualization purposes, S3 methods like plot and print are also implemented for objects created using these functions.

## 7.3 The ldstatsHD R package

In this section we present the main functions available in **ldstatsHD**. We only describe some of the most relevant arguments and values of the functions. A more detailed explanation and use of all the other arguments/values is given in the documentation of the package. The functions are grouped in three blocks corresponding to the three modules specified in Section 7.2.

### 7.3.1 Module 1 functions: data simulators

*Description of pcorSimulator*

The function `pcorSimulator` creates an (almost) block diagonal positive definite precision matrix with three possible graph structures: hub-based, power-law (default) or random. It allows for a percentage of connections between blocks to increase the complexity of the networks and make it closer to real applications in biological data. It also generates samples from a multivariate normal

distribution with covariance matrix given by the inverse of such precision matrix. The function is called in R using the following arguments

```
pcorSimulator(nobs, nclusters, nnodesxcluster, pattern = "powerLaw",
              low.strength = 0.5, sup.strength = 0.9, nhubs = 5,
              degree.hubs = 20, nOtherEdges = 30, alpha = 2.3, plus = 0,
              prob = 0.05, perturb.clust = 0, mu = 0,
              probSign = 0.5, seed = 2313)
```

The parameter `nclusters` defines the number of block diagonal matrices with `nnodesxcluster` nodes/variables for each block. The `seed` argument permits simulations to be reproducible by setting the random number generator.

Hub-based networks (`pattern = "hubs"`) are graphs where only a small number (defined in `nhubs`) of nodes have a much higher degree (or connectivity) than the rest (`degree.hubs`). For a power-law network (`pattern = "powerLaw"`), the degree of the nodes follows a power-law distribution determined by the exponent `alpha`. Both hub-based and power-law networks are described in Section 5.5.1. Random networks are included (`pattern = "random"`) for their mathematical interest, though real networks are usually non-random (Newman, 2003). These networks consider that the degree of the nodes follows a binomial distribution where the success probability determines the probability of existing an edge connecting two nodes and is specified in argument `prob`.

The function returns an object of class `pcorSim` containing the generated positive definite precision matrix and a dataset with `nobs` observations. The plot function for an object of class `pcorSim` produces the graphical representation of the network using the **igraph** package style (Csárdi and Nepusz, 2006).

*Description of pcorSimulatorJoint*

The function `pcorSimulatorJoint` is an extension of `pcorSimulator` for the more general case of creating two similar positive definite precision matrices. It allows for three types of differential graph structures: random differences, clustered differences (default) or a mixture of the two. Then, it generates datasets from a multivariate normal distribution defined by the inverse of such precision matrices with the possibility of considering linear dependence between datasets. The function is called in R by

```
pcorSimulatorJoint(nobs, nclusters, nnodesxcluster, pattern = "hubs",
             diffType = "cluster", dataDepend = "ind", low.strength = 0.5,
             sup.strength = 0.9, pdiff = 0, nhubs = 5, degree.hubs = 20,
             nOtherEdges = 30, alpha = 2.3, plus = 0, prob = 0.05,
             perturb.clust = 0, mu = 0, diagCCtype = "dicot",
             diagNZ.strength = .5, mixProb = 0.5, probSign = 0.5,
             exactZeroTh = 0.05, seed = 2313)
```

The argument `dataDepend` determines the model used to characterize paired/independent sample design. If `dataDepend = "ind"`, it assumes independence. It offers three models with a paired data structure: `"diagOmega"` , `"mult"` or `"add"` which correspond to a diagonal cross-partial cor-

relation matrix, a multiplicative model and an additive model, respectively (see Section 2.2 for description). The argument `diagCCtype` defines the relationship between the same variable in the two datasets. Following the notation in Section 2.2, this corresponds to the diagonal elements in matrix $\Delta$ if `dataDepend = "mult"` or `dataDepend = "add"`, or diagonal elements in matrix $\Omega_{12}^J$ if `dataDepend = "diagOmega"`. Two options are available: `diagCCtype="dicot"`, where half of the variables are assumed to be independent between the two datasets, and the other half are assumed to be linearly dependent by a magnitude defined in `diagNZ.strength`; `diagCCtype="beta"`, where the dependence between variables in the two datasets is randomly generated by a beta(1,3).

When `diffType = "cluster"`, differential edges are included using two additional block diagonal structures. For instance, let $\Omega^{(0)}$ be a common structure generated by `pcorSimulator` and let $D_1$ and $D_2$ be two unique partial correlation matrices also simulated by `pcorSimulator`. Following the notation in Section 2.2, $\Omega_1^J = \text{diag}(\Omega^{(0)}, D_1, I)$ and $\Omega_2^J = \text{diag}(\Omega^{(0)}, I, D_2)$ determine the two precision matrices. When `diffType = "random"`, connections between pairs of variables (in the initial $\Omega^{(0)}$ generated by `pcorSimulator`) are removed randomly with probability `pdiff` in only one condition.

The value of the function is an object of class `pcorSimJoint` with two simulated datasets that follow a multivariate normal distribution determined by the generated joint precision matrix. The S3 plot function provides the network visualization of the common network (corresponding to non-zero partial correlation coefficients in the two matrices) as well as the differential edges (zero partial correlation coefficients in one matrix and non-zero in the other matrix).

### 7.3.2 Module 2 functions: testing methods

*Description of eqCorrMatTest*

The function `eqCorrMatTest` performs hypothesis testing (HT) of equality of two correlation matrices coming from two Gaussian datasets, that can possibly be high dimensional and linearly dependent. It also contemplates the simpler hypothesis testing problem of a correlation matrix being the identity matrix, thus testing linear independence between any pair of variables in a dataset. Three test statistics are available: `AS` (average squares), `max` (extreme value test), `exc` (sum of exceedances). The function is called in R by

```
eqCorrMatTest(D1, D2 = NULL, testStatistic = c("AS", "max", "exc"),
      testNullDist = c("asyIndep","asyDep", "np"), nite = 500,
      paired = FALSE, threshold = 2.3, excAdj = FALSE, exact = FALSE,
      conf.level = 0.95, saddlePoint = FALSE, MINint = 2, MAXint = 100, ...)
```

By default, equality of two correlation matrices HT is performed. The arguments `D1` and `D2` have to have the same number of columns (defining variables), and in case `paired = TRUE`, they must also contain the same number of rows (defining samples). The identity correlation matrix HT is employed when `D2` is NULL. The parameter `testNullDist` is used to select the method to determine the null distribution. For `"asyIndep"`, it considers an asymptotic null distribution for the test statistics assuming independence between elements in the sample differential correlation matrix (see Section

4.2). Dependence is accounted by "asyDep" (which also takes parametric distributions) or by "np" (that uses permuted-based samples to approximate an empirical null distribution). The sum of exceedances test statistic depends on a weight $w$ (Section 4.3.3) where $w = 0$ if `excAdj = FALSE` and $w = 1$ if `excAdj = TRUE`, and also requires the threshold of exceedances $u$ which is specified by the argument `threshold`.

The function returns an object of class `eqCorrMatTest` containing the value of the test statistic with the underlying hypothesis testing p-values and confidence intervals at an specified `conf.level`. For this function, only the print S3 function is provided.

*Description of eqCorTestByRows*

The function `eqCorTestByRows` performs hypothesis testing to assess whether the $g$th row (for all $g \in [1, p]$) of a correlation matrix is equal or not to the same row of another correlation matrix. It also considers the simpler hypothesis testing that checks if the $g$th row of a correlation matrix (except the $g$th element) contains only zero coefficients, thus testing linear independence of a variable against all the rest of the variables. In this case, it provides AS (average squares) and max (maximum) test statistics. Both tests are conducted as permutation tests to assess significance. The complete call of the function in R is defined by

```
eqCorTestByRows(D1, D2 = NULL, testStatistic = c("AS", "max"), nite = 200,
         paired = FALSE, exact = TRUE, whichRows = NULL, conf.level = 0.95)
```

By default all rows are tested which can be computationally intensive for large dimensions. Through the argument `whichRows`, the function allows to perform HT in only the variables defined in such argument. Even though it is not implemented in the function, parallel computations could be done, e.g., using function `mclapply` from package **parallel**. The aim of this function is the screening of global dependence levels for each variable, thus adjustments for multiple testing are not included but can be applied to the resulting p-values *a posteriori*, for instance using R function `p.adjust`.

The function returns an object of class `eqCorTestByRows` containing test statistics, p-values and confidence intervals. The plot of an object of this class shows the confidence intervals for all computed test statistics corresponding to all tested rows.

### 7.3.3 Module 3 functions: estimation methods

*Description of wfgl*

The function `wfgl` provides a joint estimator of two precision matrices corresponding to the conditional dependence structure of two sets of multivariate normal distributed observations which can be linearly dependent. It uses the ADMM algorithm presented in Section 6.2. The function is called in R by

```
wfgl(D1, D2, lambda1, lambda2, paired = TRUE, automLambdas = TRUE,
    sigmaEstimate = "CRmad", pairedEst = "Reg-based-sim", maxiter = 30,
    tol = 1e-05, nsubset = 10000, weights = c(1,1), rho=1, rho.increment = 1,
```

```
    triangleCorrection = TRUE, alphaTri = 0.01, temporalFolders = FALSE,
    notOnlyLambda2 = TRUE, roundDec = 4, burst = 0, lambda1B = NULL,
    lambda2B = NULL)
```

It accounts for linear dependence between observations in the two datasets when `paired = TRUE`. Tuning parameters can be selected by setting error rates for individual and difference matrices when `automLambdas = TRUE` (see Section 6.2.2). Otherwise, the parameters `lambda1` and `lambda2` are equivalent to the interpretation in Algorithm 8. In case `lambda2` is a single value and `lambda1` is a vector with several values, then lambda selection approaches implemented in the function `lambdaSelection` (defined below) can also be used. As studied in Section 6.4, the algorithm to estimate joint precision matrices recovers more triangular motifs than expected by chance. Hypothesis testing for the weakest edges of these estimated triangular motifs is performed if `triangleCorrection = TRUE` with rejection level determined by `alphaTri`.

The function returns an object of class `wfgl` containing the two estimated precision matrices. The plot function is the same as the one defined for objects of class `pcorSimJoint` and represents the non-zero structures of both common and differential estimated precision matrices.

*Description of wfrl*

The function `wfrl` permits the joint estimation of two regression coefficient matrices from multivariate normal distributed samples using an ADMM based algorithm (see Section 6.3). As for `wfgl`, it accounts for cases where observations from the two datasets are paired. The function is called in R by

```
wfrl(D1, D2, lambda1, lambda2, automLambdas = TRUE, paired = TRUE,
     sigmaEstimate = "CRmad", maxiter=30, tol=1e-05, nsubset = 10000,
     rho = 1, rho.increment = 1, notOnlyLambda2 = TRUE)
```

Here `D1` and `D2` are lists containing two matrices: response variables and explanatory variables for the first condition in `D1` and response variables and explanatory variables for the second condition in `D2`. The tuning parameter selection options are equivalent to the ones explained above for the `wfgl` function.

The function returns an object of class `wfrl` containing the two regression coefficient matrices. The plot function is similar to the one for objects of class `pcorSimJoint` or `wfrl`. The only difference is that here the networks are directed (edges going from explanatory variables to response variables).

*Description of lambdaSelection*

The function `lambdaSelection` is designed to select the sparsity regularization parameter $\lambda$ in graphical models. Eight different criteria are available to select $\lambda$ with risk functions based on network characteristics: path connectivity (PC), AGlommerative NESted (AGNES), Augmented-MSE (A- MSE), Vulnerability (VUL), AIC/BIC/eBIC and StARS (from the **huge** package). The algorithms for all these options are described in Chapter 5. The function is called by

```
lambdaSelection(obj, criterion = c("PC","AGNES","A-MSE","VUL","STARS",
               "AIC", "BIC", "eBIC"), ...)
```

Depending on each criterion, several parameters are specified:

```
pcLambdaSelection(obj)
agnesLambdaSelection(obj, way = "direct", nite = 10, subsvec = NULL,
                     eps = 0.05, until = NULL, minNodes = 30,
                     distF = c("correlation","shortPath"))
amseLambdaSelection(obj, pathIni, y, generator = c("subsampling",
                "montecarlo"), pB = 0.7, nite = 10, method = "mb", from = 1,
                until = NULL,  distF = c("correlation","shortPath"),
                oneByone = FALSE, many = 3)
vulLambdaSelection(obj, loo = FALSE, subOut =  10, nite = 50)
icLambdaSelection(obj,  y, criterion = c("AIC", "BIC", "eBIC"))
```

The argument `obj` must be an object generated by functions `huge`, `camel.tiger`, `wfgl` or `wfrl`, and has to contain at least five different estimated precision/adjacency matrices for five different tuning parameters. For `AIC`, `BIC` and `eBIC` criterion, neighborhood selection (`"mb"` option in function `huge`) is not a suitable object since precision matrix elements are not explicitly estimated and therefore likelihoods cannot be calculated.

The function returns an object of class `lambdaSelection` describing the selected tuning parameter. The plot function for an object of class `lambdaSelection` reproduce the observed values of the selected risk function for all the tuning parameters that are used.

## 7.4 User interface in simulated data

In this section we present a brief tutorial on the functionality and capability of the **ldstatsHD** package. As in Section 7.3, we organize the functions in three different modules: data simulators, testing methods and estimation methods. Simulated data examples described in the first module are used to illustrate the usage of testing and estimation functions in second and third modules.

### 7.4.1 Module 1 functions: data simulators

*Example of pcorSimulator use*

We simulate three precision matrices using the function `pcorSimulator` corresponding to power-law, hubs and random graph structures. We set a seed in each one of them to make all results reproducible. We give a vector of values to be consistent with an early version of the package but declaring a single value is also possible. For power-law networks we take a 3 block diagonal matrix with 200, 140 and 60 variables each block. We use the power-law parameter `alpha` to be 2.3 (Peng et al., 2009). The R command and print is given by

```
R> EX1 <- pcorSimulator(nobs = 70, nclusters = 3, nnodesxcluster = c(200,
          140,60),  pattern = "powerLaw", alpha = 2.3, seed = c(5,22,50))
R> EX1
```

```
pattern: "powerLaw", Number of nodes = 400,  Number of edges = 356,
   Sparsity = 0.99555
```

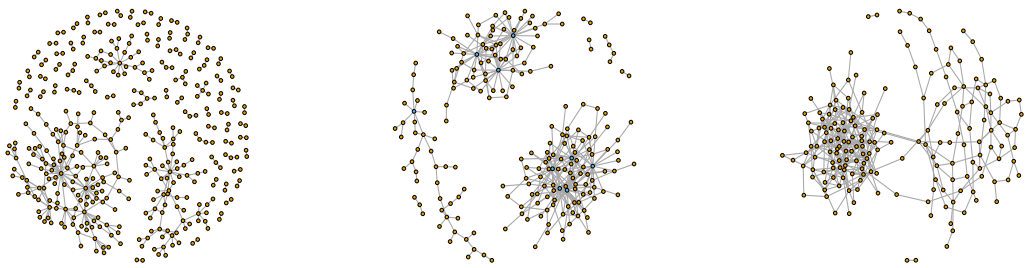
Sparsity levels in the print function are defined by the proportion of zero elements in the lower triangular precision matrices. For instance in `EX1`, 356 non-zero partial correlation elements are considered from a total of $400 \times 399 / 2 = 79,800$ possible edges. Similarly, for hub-based networks we define 5, 3 and 1 hub nodes for the three clusters respectively with degree 20, 20 and 5. The other generated edges in the three clusters (100, 50, 40) are selected randomly. The R command and print is given by

```
R> EX2 <- pcorSimulator(nobs = 70, nclusters = 3, nnodesxcluster = c(100,80,
           60), pattern = "hubs", nhubs = c(5,3,1), degree.hubs = c(20,20,5),
           nOtherEdges = c(100,50, 40), seed = c(10,20,20))
R> EX2
pattern: "hubs", Number of nodes = 240,  Number of edges = 355,
   Sparsity = 0.98767
```

The generated graph structure for `EX2` is denser than the one for the first example in `EX1`. Finally, for random networks, we use two clusters with the same size (100 nodes each) and edge probabilities 0.05 and 0.02. The R command and print is given by

```
R> EX3 <- pcorSimulator(nobs = 70, nclusters = 2, nnodesxcluster = c(100,
          100), prob=c(0.05,0.02), perturb.clust = 0.05, pattern = "random",
          seed = c(3,4))
R> EX3
pattern: "random", Number of nodes = 200, Number of edges = 356,
   Sparsity = 0.9822
```

This generated network is the densest of the three. Plots for each of the three examples are shown in Figure 7.1 and can be obtained typing `plot(EX1)`, `plot(EX2)` and `plot(EX3)` in the R prompt.



**(a)** Power law graph example **(b)** Hub-based graph example **(c)** Random graph example

**Figure 7.1.** Graphical representation of generated precision matrices using function `pcorSimulator`.

*Example of pcorSimulatorJoint use*

We simulate joint precision matrix structures using different definitions for the parameters of the model. We first generate a joint power-law graph structure where the difference matrix is clustered.

For the paired data, here we use a diagonal cross-partial correlation matrix where the diagonal components are generated by a beta distribution (with parameters 1 and 3)

```
R> EXJ1 <- pcorSimulatorJoint(nobs = 80, nclusters = 3, nnodesxcluster = c(30,
           30,30), pattern = "pow", diffType = "cluster", dataDepend = "diag",
           pdiff = 0.5, perturb.clust = 0.2, mixProb = 0.5,
           diagCCtype = "beta", seed = c(20,3,50,52,23))
R> EXJ1
Pattern: "powerLaw",  DataDepend = "diagOmega",  DiagCCtype = "beta13"
Number of nodes = 134,  Common edges = 92,  Sparsity common network = 0.98975
Differential edges = 38,  Sparsity differential network = 0.99577
```

In the print, the number of edges and sparsity levels are specified for both common and differential networks. In this case, 92 edges are common in the two conditions, whereas 38 edges are only present in either one of the two conditions. A second example is considered when the differential edges are randomly generated (`diffType = "random"`) and the paired structure is determined by a multiplicative model. In this case, a two-block precision matrix with 160 and 60 nodes each is used

```
R> EXJ2 <- pcorSimulatorJoint(nobs = 50, nclusters = 2,
        nnodesxcluster = c(160, 60), pattern = "pow", diffType = "random",
        dataDepend = "mult", pdiff = 0.2, perturb.clust = 0.2, mixProb = 0.5,
        seed = 56)
R> EXJ2
Pattern: "powerLaw",  DataDepend = "mult",  DiagCCtype = "dicot"
Number of nodes = 220,  Common edges = 78,  Sparsity common network = 0.9968
Differential edges = 119,  Sparsity differential network = 0.9951
```

In `EXJ2`, a larger number of differential edges than `EXJ1` is observed for the same number of common edges. We also generate data by assuming a mixture of random (80%) and clustered (20%) differential edges and paired structure determined by an additive model

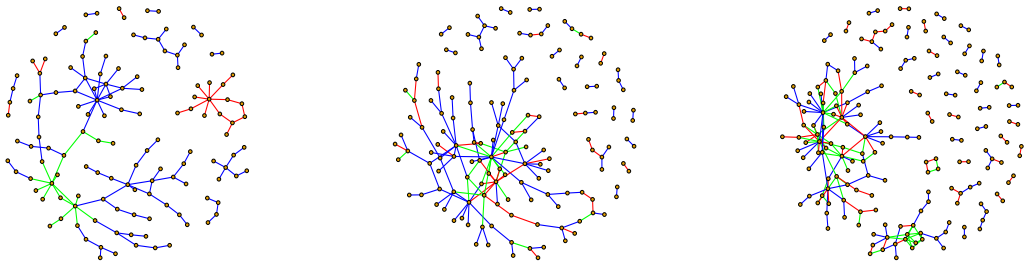
```
R> EXJ3 <- pcorSimulatorJoint(nobs = 50, nclusters = 2,
        nnodesxcluster = c(160, 130), pattern = "pow", diffType = "mixed",
        dataDepend = "add", pdiff = 0.4, perturb.clust = 0, mixProb = 0.8,
        seed = 43)
R> EXJ3
Pattern: "powerLaw", DataDepend = "add",  DiagCCtype = "dicot"
Number of nodes = 382,  Common edges = 73,  Sparsity common network = 0.999
Differential edges = 186,  Sparsity differential network = 0.99745
```

In this last object, there are more differential edges than common edges. Plots for each of the three examples are shown in Figure 7.2.

### 7.4.2 Module 2 functions: testing methods

*Example of eqCorrMatTest use*

We consider simulated data defined in object `EXJ1`, which is the first example declared in Section 7.4.1 for function `pcorSimulatorJoint`. We test whether the correlation matrix that generates the data for the first class `D1` is equal to the correlation matrix for the second class `D2`. We initially consider all test statistics (with $w = 0$ in the sum of exceedances test) and also the three ways to describe the



**(a)** Clustered differences **(b)** Random differences **(c)** Mixed differences

**Figure 7.2.** Graphical representation of generated precision matrices using the plot function for objects obtained by `pcorSimulatorJoint`. Blue edges are common edges in the two conditions. Red edges are only present in the first condition and green edges are only present in the second condition.

null distribution. We use 500 permuted samples to estimate dependence parameters in asymptotic dependence null distributions and to approximate the non-parametric null distributions. The R call and print are given by

```
R> (test1 <- eqCorrMatTest(EXJ1$D1, EXJ1$D2, testStatistic = c("AS",
       "max", "exc"), testNullDist = c("asyIndep","asyDep", "np"), nite= 500,
       paired = TRUE, threshold = 2.3, excAdj = FALSE, exact = FALSE,
       conf.level = 0.95))
    Test for equality of two correlation matrices using independent data
asyIndep  Tas = 0.013,  pval =  0.198,  95 percent CI: -0.012 0.067
asyDep    Tas = 0.016,  pval =  0.198,  95 percent CI: -0.015 0.082
np        Tas = 0.016,  pval =  0.19,  95 percent CI: -0.013 0.072

asyIndep  Tm = 0.276,  pval =  0.25,  95 percent CI: -0.395 0.778
asyDep    Tm = 0.458,  pval =  0.214,   95 percent CI:  -0.349 0.85
np        Tm = 0.197,  pval =  0.212,   95 percent CI: -0.363 0.847

asyIndep  thr = 2.3,  Texc = 5.037,  pval =  0.48,
          95 percent CI: -159.399 376.828
asyDep    thr = 2.3,  Texc = 25.563,  pval =  0.399,
          95 percent CI: -157.886 440.342
np      thr = 2.3,  Texc = 25.563,  pval =  0.404,
          95 percent CI: -161.363 340.44
```

The print of `test1` shows the value for the test statistics, p-values and confidence intervals. None of the tests shows any evidence against the null hypothesis. We also perform the same hypothesis testing on the object `EXJ2`. In this case, only the exceedances-based test (with $w = 1$) is used with an asymptotic dependence null distribution and three different thresholds: `threshold = c(0,1,2)`,

```
R> (test2 <- eqCorrMatTest(EXJ2$D1, EXJ2$D2, testStatistic = "exc",
               testNullDist = "asyDep", nite= 300, paired = TRUE,
               threshold = c(0,1,2), excAdj = TRUE, exact = FALSE,
               conf.level = 0.95))
     Test for equality of two correlation matrices using paired data
asyDep  thr = 0,  Texc = 151.777, pval =  0.25, 95 percent CI: -219.8 991.9
asyDep  thr = 1,  Texc = 86.879,  pval =  0.133, 95 percent CI: -41.1 376.4
```

```
asyDep  thr = 2,  Texc = 27.014,  pval =  0.061, 95 percent CI: -1.1 90.7
```

For the three tested thresholds, the p-values are small, especially when `thr = 2`. To show the usage of the HT for identity correlation matrix, we consider the dataset generated in object `EX3`. Here we leave the argument `D2=NULL` and we use a non-parametric null distribution. We change the confidence level at 99%. The R call and print are given by

```
R> (test3 <- eqCorrMatTest(EX3$y, NULL, testStatistic = c("AS", "max",
             "exc"),  testNullDist = "np", nite= 300, threshold = 2,
             excAdj = FALSE, conf.level = 0.99))
  Test for non-Identity correlation matrix
np   Tas =  0.522,  pval =  0,  99 percent CI: 0.495 0.553
np   Tm = 5.931,    pval =  0,  99 percent CI: 5.111 6.474
np   thr = 2,  Texc = 9095.521,  pval =  0,  99 percent CI: 8729.044 9483.248
```

The confidence intervals do not include zero in any of the three test statistics and null hypothesis could be rejected at 0.01 significance level.

*Example of eqCorTestByRows use*

In the first example of usage of the function `eqCorTestByRows` we intend to test the equality of correlation rows between the two datasets defined in object `EXJ1`. We use both test statistics, `AS` and `max`, and 200 permuted samples:

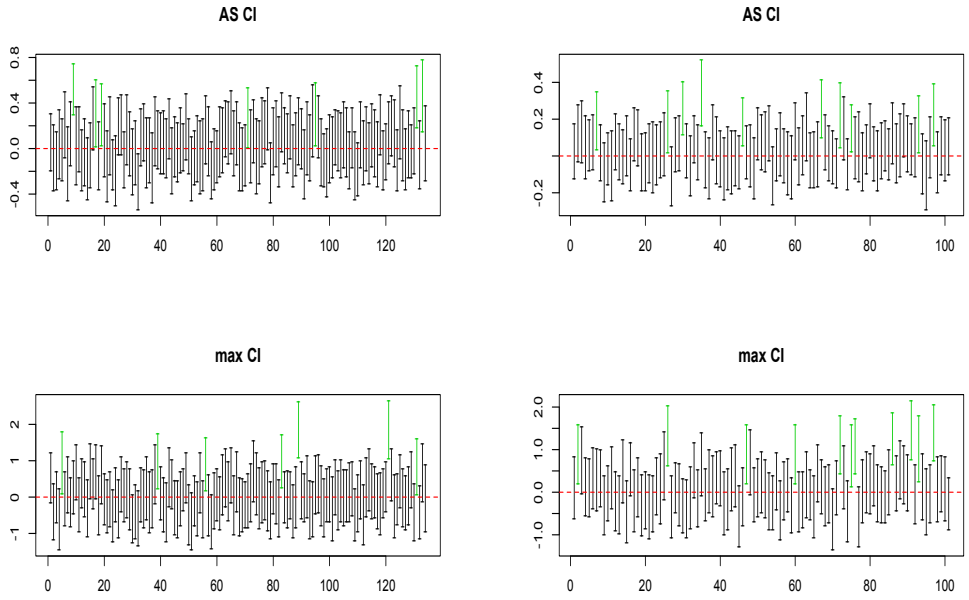
```
R> (testr1 <- eqCorTestByRows(EXJ1$D1, EXJ1$D2, testStatistic = c("AS",
      "max"), nite = 200,  paired = TRUE, exact = FALSE, whichRows = NULL,
      conf.level = 0.95))
   Test for equality of correlation matrix rows using paired data
number of significant rows for Tas: 13 at 0.95 conf.level, expected 6.7
number of significant rows for Tm: 6 at 0.95 conf.level, expected 6.7
```

The print gives the number of tested rows with a p-value smaller than 1−`conf.level` against the expected number of significant rows under $H_0$. In this case, expected is much lower than observed in only `Tas`. We also perform similar tests for the datasets in object `EXJ3`. We only test the rows 100 to 200 by setting `whichRows = c(100:200)`,

```
R> (testr2 <- eqCorTestByRows(EXJ3$D1, EXJ3$D2, testStatistic = c("AS",
            "max"), nite = 1000, paired = TRUE, exact = FALSE,
            whichRows = c(100:200), conf.level = 0.95))
          Test for equality of correlation matrix rows using paired data
number of significant rows for Tas: 2 at 0.95 conf.level, expected 5.05
number of significant rows for Tm: 8 at 0.95 conf.level, expected 5.05
```

In contrast to the first tested dataset, for `EXJ3`, `Tm` gives more significant rows than `Tas`. The plots for `testr1` and `testr2` are presented in Figure 7.3. The confidence intervals are shown for all tested rows and, in green, the significant tests at 0.05 significance level are highlighted.

We finally provide an example for the HT problem of linear independence between a variable and all the rest using the dataset in object `EX2`. We only use the average of squares test statistic here with a 99% confidence level,




**(a)** Object `testr1` **(b)** Object `testr2`

**Figure 7.3.** Confidence intervals for tested rows in objects `testr1` and `testr2`. Green lines correspond to variables whose confidence intervals do not include zero.

```
R> (testr3 <- eqCorTestByRows(EX2$y, NULL, testStatistic = "AS", nite = 200,
                  paired = TRUE, exact = FALSE, whichRows = NULL,
                  conf.level = 0.99))
          Test for non-zero correlation matrix rows
number of significant rows for Tas: 78 at 0.99 conf.level, expected 2.4
```

In the print we observe that 78 of the rows have a p-value smaller than 0.01, when only 2.4 where expected by chance.

### 7.4.3 Module 3 functions: estimation methods

*Example of wfgl use*

First of all, we show using data defined in object `EXJ1`, that `wfgl`, when arguments `paired = FALSE`, `automLambdas = FALSE`, and `triangleCorrection = FALSE`, coincides with function `JGL` (Danaher et al., 2014),

```
R> fgl1 <- wfgl(EXJ1$D1, EXJ1$D2, lambda1=0.2, lambda2=0.1, paired = FALSE,
           automLambdas = FALSE, maxiter = 30, tol = 1e-05,
           triangleCorrection = FALSE)
R> fgl2 <-JGL(list(scale(EXJ1$D1), scale(EXJ1$D2)), penalty="fused",
      lambda1=0.2, lambda2=0.1, return.whole.theta=TRUE, maxiter=31,
      penalize.diagonal = FALSE)
R> c(sum(abs(fgl2$theta[[1]]-fgl1$omega[[1]])), sum(abs(fgl2$theta[[2]]-
   fgl1$omega[[2]])))
```

```
[1]0.0004 0.0003
```

Otherwise, controlling the error rates to select tuning parameters, adjusting for paired data and correcting for triangular motifs, the R call is

```
R> (wfgl1 <- wfgl(EXJ1$D1, EXJ1$D2, lambda1 = 0.05, lambda2 = 0.05,
         paired = TRUE, automLambdas = TRUE, maxiter = 30, tol = 1e-05,
         triangleCorrection = TRUE, alphaTri = 0.05))
           joint partial correlation estimator using paired data
Number of nodes = 134, Total number of possible edges = 8911
Est. common edges = 331, Sparsity est. common network = 0.96285
Est. differential edges = 25, Sparsity est. differential network = 0.99719
Est. edges for only pop.1 = 12, Est. edges for only pop.2 = 13
alpha2 = 0.0043

R>  plot(wfgl1, col = c("blue","red","green"), vertex.size = 3,
     edgesThickness = TRUE, zoomThick = 10)
```

The print shows some basic information about the estimated network sizes. It also provides an approximation of the error rate $\alpha_2$ defined in Section 6.2.2. A useful visualization tool for the estimated network is provided with the plot function of a `wfgl` object (see Figure 7.4). By setting the attribute `edgesThickness` to `TRUE`, we account for different widths in the estimated edges that are proportional to the magnitude of their underling estimated precision matrix elements.

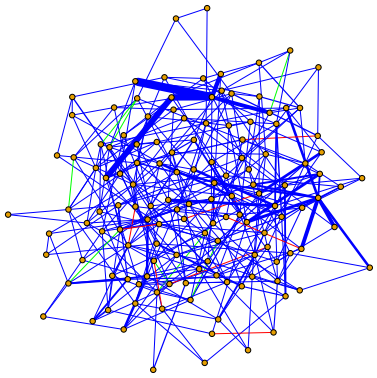


**Figure 7.4.** Estimated network using function wfgl for data example `EXJ1`. Blue edges are common edges in the two conditions. Red edges are only present in the first condition and green edges are only present in the second condition. The thickness of the edges is proportional to the underlying estimated precision matrix elements.

Keeping `lambda2` fix (our $\alpha_2'$ defined in Section 6.2.2), we give several values for `lambda1`,

```
 R> (wfgl1 <- wfgl(EXJ1$D1, EXJ1$D2, lambda1 = c(0.01, 0.05, 0.1),
      lambda2 = 0.05, paired = TRUE, automLambdas = TRUE,
      maxiter = 30, tol = 1e-05, triangleCorrection = TRUE, alphaTri = 0.05))
   joint partial correlation estimator using paired data
lambda1 sequence of length 3
Est. com. edges : 122 -> 360,  Sparsity est. com. network : 0.9596 -> 0.9863
Est. diff. edges : 10 -> 45,  Sparsity est. diff. network : 0.995 -> 0.999
Est. edges for only pop.1 : 5 -> 22, Est. edges for only pop.2 : 10 -> 46
```

In this case, the print provides the range of number of estimated edges in common and differential networks, from the smallest to the largest $\lambda_1$. Differential networks only change in two estimated edges whereas common network goes from 122 edges (`lambda1` = 0.01) to 360 edges (`lambda1`=0.1).

*Example of wfrl use*

We design a simple example to show the usage of `wfrl`. We consider the data created in object `EXJ1` that contains 135 variables as our explanatory variables. We consider the same number of response variables linking covariates by a linear model with regression coefficient matrices being diagonal matrices. We employ tuning parameter selection by setting the underlying error rates in `lambda1` and `lambda2` (denoted by $\alpha_1$ and $\alpha_2'$ in Chapter 6). Besides, we account for paired data and stop the ADMM algorithm in a maximum of 10 iterations to avoid high computational burden.

```
R> P           <- EXJ1$P
R> q           <- P
R> N          <- dim(EXJ1$D1)[1]
R> BETA1       <- array(0, dim = c(P, q))
R> diag(BETA1) <- rep(0.35,q)
R> BETA2       <- BETA1
R> diag(BETA2)[c(1:floor(q/2))] <- 0
R> sigma2      <- 1.3
R> Q           <- scale(EXJ1$D1)
R> W           <- scale(EXJ1$D2)
R> set.seed(231)
R> X                  <- Q%*%BETA1 + mvrnorm(N,rep(0,q),diag(rep(sigma2,q)))
R> set.seed(2234)
R> Y                  <- W%*%BETA2 + mvrnorm(N,rep(0,q),diag(rep(sigma2,q)))
R> D1                 <- list(scale(X), scale(Q))
R> D2                 <- list(scale(Y), scale(W))

R> (wfrl1 <- wfrl(D1, D2, lambda1=0.01, lambda2=0.05, automLambdas = TRUE,
    paired = FALSE, sigmaEstimate = "CRmad", maxiter=10, tol=1e-05,
    nsubset = 10000,  rho = 1, rho.increment = 1, notOnlyLambda2 = TRUE))

          joint regression coefficients estimator using independent data
Number of response variables = 134, Number of explanatory variables = 134,
   Number of possible edges = 17956
Estimated common edges = 221, Sparsity estimated common network = 0.9877
Estimated differential edges = 15, Sparsity estimated diff. network = 0.99916
Estimated edges for only pop.1 = 13, Estimated edges for only pop.2 = 2
```

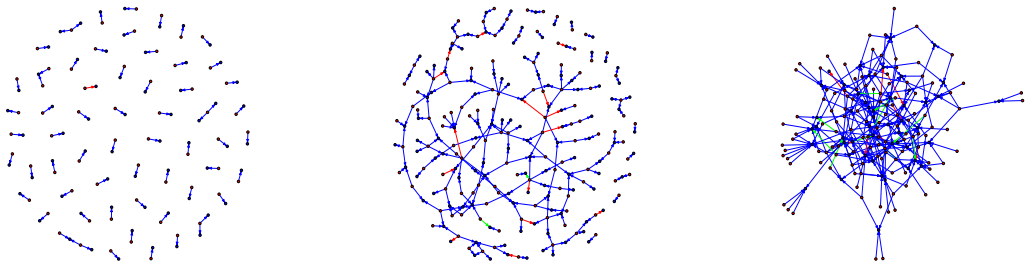
The print reflects the graphical representation of the estimated networks, both common (221 edges) and differential (15 edges) networks. Note that the expected number of false positive edges in each of the two networks under the specified `lambda1` (or $\alpha_1$) is 179. We also consider setting a vector of values for `lambda1` keeping `lambda2` fixed at 0.05. The R call is defined by

```
R> (wfrl2 <- wfrl(D1, D2, lambda1 = c(.001,.01,.04), lambda2=0.10,
   automLambdas = TRUE, paired = FALSE, sigmaEstimate = "mad",
   maxiter=30, tol=1e-05, nsubset = 10000,  rho = 1, rho.increment = 1,
   notOnlyLambda2 = TRUE))

          joint regression coefficients estimator using independent data
lambda1 sequence of length 3
Number of response variables = 134, Number of explanatory variables = 134,
```

```
   Number of possible edges = 17956
Est. com. edges: 63 -> 647, Sparsity est. com. network: 0.96397 -> 0.99649
Est. diff. edges: 3 ->  95, Sparsity est. diff. network: 0.99471 -> 0.99983
Est. edges for only pop.1 : 3 -> 55, Est. edges for only pop.2 : 0 -> 40
```

The number of estimated common edges ranges from 63 (`lambda1=0.001`) to 670 (`lambda1=0.04`). The plots of object `wfrl2` are shown in Figure 7.5.



**(a)** $\alpha_1 = 0.001$ **(b)** $\alpha_1 = 0.01$ **(c)** $\alpha_1 = 0.04$

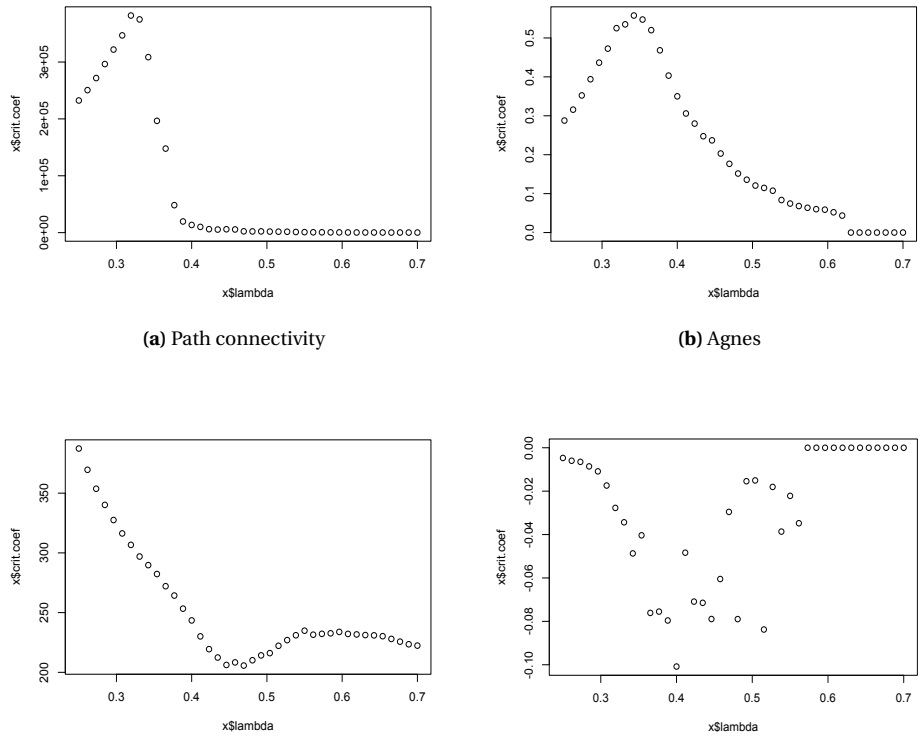
**Figure 7.5.** Graphical representation of the nonzero structure of regression coefficient matrices defined in object `wfrl2`. Blue edges are common edges in the two conditions. Red edges are only present in the first condition and green edges are only present in the second condition.

*Example of lambdaSelection use*

We first select the optimal hyper-parameter of graphical lasso models (employing neighborhood selection) with data defined in object `EX1`. We use the following risk functions: PC, AGNES, A-MSE (with AGNES estimate), VUL and STARS. Note that the outcome of the huge function is the non-zero structure of the estimated precision matrix, thus likelihood-based methods as AIC and BIC are not well defined.

```
R> y <- EX1$y
R> Lambda.SEQ  <- seq(.25,0.70,length.out = 40)
R> out3 <- huge(y, method = "mb", lambda = Lambda.SEQ)
R> (lamPC <- lambdaSelection(out3, criterion = c("PC")))
          lambda selection by optimizing PC risk function
optimal lambda = 0.3769,          Sparsity graph structure = 0.9969
R> (lamAG <- lambdaSelection(out3, criterion = c("AGNES")))
          lambda selection by optimizing AGNES risk function
optimal lambda = 0.3423,          Sparsity graph structure = 0.9945
R> (lamAAG <- lambdaSelection(out3, criterion = c("A-MSE"), y=y,
            pathIni =out3$path[[which(lamAG$opt.lambda == Lambda.SEQ)]] ))
          lambda selection by optimizing A-MSE risk function
          with subsampling generator
optimal lambda = 0.4692,          Sparsity graph structure = 0.9987
R> (lamVUL <- lambdaSelection(out3, criterion = c("VUL"))) # do not run
     #(computationally intensive)
          lambda selection by optimizing VUL risk function
optimal lambda = 0.4,          Sparsity graph structure = 0.9977
R> (lamST <- lambdaSelection(out3, criterion = c("STARS")))
optimal paramter: 0.25, sparsity level: 0.02382206.
```

The print shows the selected lambda by the given criterion as well as the sparsity level of the selected graph structure. In this case, STARS produce the densest estimated graph structure, followed by AGNES, PC, VUL and finally A-MSE. The plots for the latter four risk functions are shown in Figure 7.6.




**(a)** Path connectivity

**(b)** Agnes

**(c)** A-MSE

**(d)** Vulnerability

**Figure 7.6.** Obtained coefficients for the tuning parameter selection risk functions path connectivity, agnes, A-MSE and vulnerability using data in object `EX1`.

The lambda selection function can also be used for objects of class `wfgl` and `wfrl`. For instance here we use `PC` for selecting `lambda1` in jointly estimating two precision matrices,

```
R> wfgl1 <- wfgl(EXJ1$D1, EXJ1$D2, lambda1 = seq(0.001,0.05,length.out=30),
        lambda2 = 0.05, paired = TRUE, automLambdas = TRUE, maxiter = 5)
R> (lam1PC <- pcLambdaSelection(wfgl1))
          lambda selection by optimizing PC risk function
optimal lambda = 0.0061,          Sparsity graph structure = 0.988
```

The optimal tuning parameter is 0.0061, which is one of the sparsest estimated graph structures.

## 7.5 Discussion

In this chapter we have presented the R package **ldstatsHD** which consists of data simulators, testing methods for two correlation matrices, and joint estimation methods for two conditional dependence

structures as precision matrices and regression coefficient matrices. It also contains functions to select the sparsity tuning parameter in graphical models. These implemented approaches are especially useful when the two datasets are high-dimensional and come from paired observations.

The algorithms are efficiently implemented in R by taking advantage, when possible, of sparsity properties. Nevertheless, the computational time and memory used is still a major issue when analyzing datasets with very large dimensions (order of thousands). Particularly, joint estimation methods implemented in functions `wfgl` and `wfrl`, due to estimating dense matrices in every iteration of the ADMM recursive algorithms (see Chapter 6), turn out to be computationally intensive when the dimension is larger than $5,000$. Regularization parameter selection methods as A-MSE and VUL (see Chapter 5) are also slow for similar dimension sizes. As future work, the algorithms and code could be refined to speed up the procedures.

The user interface of the proposed functions tries to mirror other leading R functions in the topic. For instance, all the attributes in `wfgl` that have the same meaning to the analogous attributes in the function `JGL` (Danaher et al., 2014) can be identified by the same name. The S3 method print is available for all the methods to summarize the output of the functions. Moreover, when required, the plot function is also implemented for visualization purposes.

# Chapter 8

# Testing and estimation of linear dependence structures for colon cancer data

## 8.1 Introduction

The main idea of this chapter is to present the data analysis of a real case study employing the developed methods in this thesis. Our motivating data are presented in Hinoue et al. (2012) and are freely available at the Gene Expression Omnibus (GEO) database (Edgar et al., 2002) with accession numbers GSE25070 and GSE25062. In total there are 50 samples from 25 patients, a tumor and a normal colon tissue samples from each subject, which contain the gene expression information in $24,526$ genes as well as methylation presence in $27,578$ sites. The aims of this analysis are (a) finding known biological processes which can be linked to changes in the gene linear dependence structures between the two sample populations (healthy and tumor), (b) finding common and unique gene-to-gene networks among the two classes of observations, and (c) integration of the two types of omics data to find connections between genes and specific methylation sites.

### 8.1.1 Methylation and gene expression

DNA Methylation is an epigenetic process that occurs when a methyl (CH3) group is bounded to DNA. In humans, this is mostly found when the cytosine nucleotide is followed by the guanidine nucleotide (creating CpG-sites) and can be associated with the start of the gene (the promoter). In the data, the $27,578$ CpG sites are located at the promoter regions of about $15,000$ protein-coding genes. Regions with large concentration of CpG-sites are called CpG-islands and are expected to be strongly negatively correlated with the expression of the gene promoter due to silencing. We aim to investigate

this biological behaviour in our data.

Methylation presence is measured in a continuous scale that ranges from 0 -not present at all- to 1 -100% present-, where something in between indicates the strength of methylation. We apply a logit transformation of methylation presence so the values are defined in the whole real line and are closer to Gaussianity (Wahl et al., 2014). Besides, the gene expression data are log2-transformed and normalized using robust spline normalization (Schmid et al., 2010).

### 8.1.2 Summary of the chapter

The advances in technology in the field of omics (i.e., genomics, metabolomics or proteinomics) have allowed the collection and storage of different data profiles on the same individual. This has encouraged the development of integration techniques to incorporate all data for a joint analysis (Kislinger et al., 2006; Fagan et al., 2007; Lê Cao et al., 2008; Depuydt et al., 2009). Particularly, integration and analysis of methylation with gene expression data have been recently studied in Gadaleta and Bessonov (2015), who integrate gene expression and methylation presence for a dataset with 215 individuals affected with glioblastoma cancer. The authors apply lasso-penalized maximum likelihood approaches to estimate two networks: the non-zero structure of the regression coefficients using gene expression as response variables and methylation presence as explanatory variables; and the non-zero structure of the precision matrix (inverse of covariance matrix) using only gene expression data. Other related contributions include Wang et al. (2014), who employ biological knowledge of gene interactions to estimate associations between methylation presence and gene expression on individuals with primary ovarian tumours; Renner et al. (2013), who analyze the behaviour of DNA methylation in different sarcoma subtypes; Wagner et al. (2014), who study the relationship between the two types of data in healthy human cells, or List et al. (2014), who combine methylation and gene expression data to classify several breast cancer subtypes.

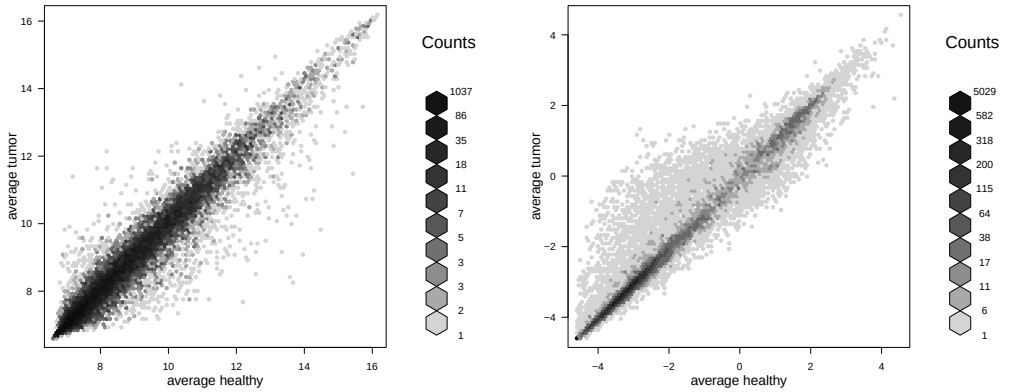
In this chapter we employ the methodology presented in Chapters 4, 5 and 6 to fully analyze and integrate both gene expression and methylation presence datasets. In Section 8.2 we perform an exploratory analysis of the datasets in which we visualize the differences between samples in the two medical conditions and we relate gene expression and methylation presence using some basic summary statistics. In Section 8.3 we consider hypothesis testing for the equality of correlation matrices on subgroups of genes determined by 1,320 biological pathway lists. We also test if each of the 24,526 measured genes interact similarly in the two conditions considering both sum of squares and extreme value test statistics. This is used to reduce the number of genes prior to estimation, which is done separately for healthy and tumor gene expression datasets in Section 8.4. In Section 8.5 and Section 8.6 we refine the estimations by applying joint graphical lasso techniques (for both precision matrices and regression coefficient matrices) to find common and unique conditional dependence structures among the two classes of observations. In Section 8.7 we integrate all estimated networks and we compare the estimated edges with some of the most relevant pathway lists in Section 8.8.

## 8.2 Exploratory analysis of the data

An initial summary table with measures of central tendency, range and dispersion for the gene expression and methylation presence datasets are presented in Table 8.1. The differences between tumor and healthy samples in both average and variance are in the third decimal for gene expression data whereas tumor samples contain substantially larger average/variance than normal samples for methylation presence data. Figure 8.1 shows the relationship between gene expression (and methylation presence) mean vectors on the two classes of observations, healthy and tumor. For visualization purposes, note that the number of genes/sites is huge, we approximate a bivariate density distribution by diving the plot space in equidistant hexagon bins whose colors are related to the number of points that occur in each bin, i.e., see R package **hexbin** (Carr et al., 2015). In the figure, in spite of observing a clear positive correlation between mean vectors in the two medical conditions, some genes/sites are located away from the common tendency.

**Table 8.1.** Summary for gene expression and methylation presence (logit transformed) datasets. Basics statistics as the minimum, maximum, quantiles, median, mean and variance are presented for both healthy and tumor samples.

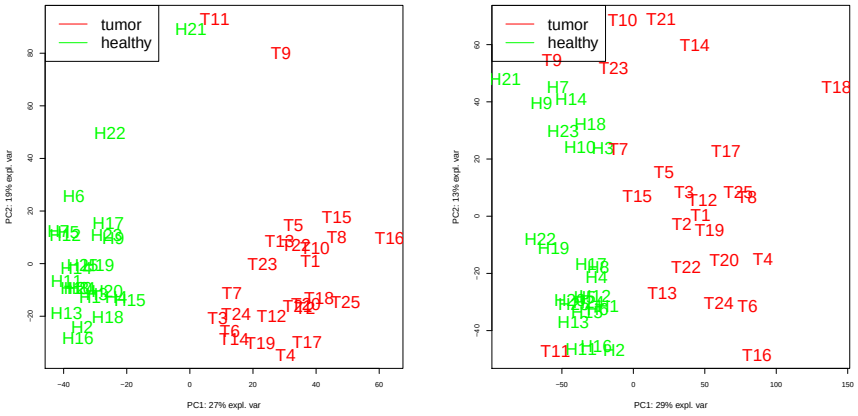
| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Var. |
|---|---|---|---|---|---|---|---|
| | | | Gene expression | | | | |
| Healthy | 6.378 | 7.081 | 7.714 | 8.426 | 9.360 | 17.240 | 3.090 |
| Tumor | 6.363 | 7.081 | 7.716 | 8.427 | 9.362 | 17.040 | 3.083 |
| | | | Methylation presence | | | | |
| Healthy | -4.595 | -3.925 | -2.885 | -2.300 | -0.924 | 4.595 | 4.276 |
| Tumor | -4.595 | -3.977 | -2.913 | -2.246 | -0.580 | 4.595 | 4.556 |




**(a)** gene expression **(b)** methylation

**Figure 8.1.** Global relationship between normal and tumor tissues for (a) gene expression and (b) methylation presence: mean vectors for gene expression (or methylation presence) on healthy samples are in the x-axis and the ones on tumor samples are on the y-axis; Hexagon bin colors indicate the frequency of points in that region going from white (low frequency) to black (high frequency). A positive linear relationship is observed for the majority of genes and methylation sites.

A sparse principal component analysis (Zou et al., 2006) is applied to the two datasets, and Figure 8.2 illustrates the individual projections of the first two components which explain the 46% (for gene expression) and 42% (for methylation) of variability in the data. The first component in either methylation or gene expression distinguishes between tumor (red) or normal (green) samples. It also shows a potential outlier in the methylation subfigure that corresponds to observation 11 for tumor samples. To obtain a good representation of the differences between the two classes of observations we do not consider this sample for estimating regression coefficient matrices in Section 8.6 and 8.8.

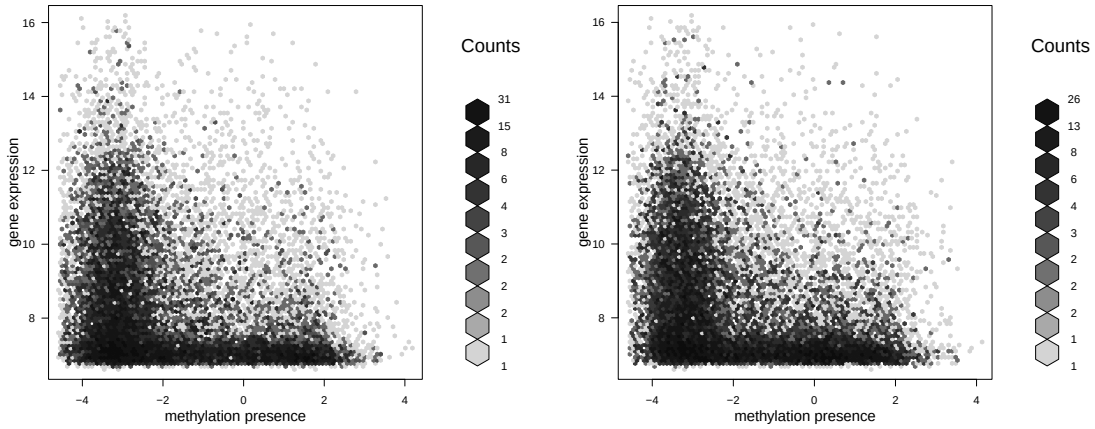



**(a)** gene expression **(b)** methylation presence (logit)

**Figure 8.2.** Projections on the first two sparse principal components for (a) gene expression dataset and (b) methylation presence dataset. The samples are colored by disease, tumor in red and healthy in green, in the first component.

In order to measure the relationship between methylation presence and gene expression in the 50 samples, methylation sites are matched to their gene promoters. The average correlation between gene expression and methylation presence of those matched genes and sites is $-0.04$, for healthy, and $-0.08$, for tumor (both values being significantly smaller than zero -using a t-test-). This negative correlation is stronger when looking at the linear relationship between the gene expression and methylation presence mean vectors ($-0.27$ for healthy and $-0.33$ for tumor), as shown in Figure 8.3. While for low methylation presence (from -4 to -2), the gene expression often reaches high values ($\geq 10$), these are rarely exceeded when the methylation is high (from 0 to 2).

Finally, we compare the four sample correlation matrices that correspond to the four datasets filtered by genes and sites that are matched: these are the gene expression with healthy or tumor samples, and methylation presence with healthy or tumor samples. Considering only pairs of genes whose sample correlation coefficient in the gene expression dataset is larger than 0.5 in absolute value, it turns out that the proportion of correlation coefficients whose signs are the same in both gene expression and methylation presence is about 0.52 for healthy and 0.54 for tumor. Even though this rate is significant, it does not seem to be highly informative. For instance, the same coefficient computed matching normal and tumor gene expression correlations is approximately 0.75.




**(a)** healthy samples

**(b)** tumor samples

**Figure 8.3.** Global relationship between gene expression and logit transformed methylation presence for (a) healthy and (b) tumor tissues in which the average gene expression over 25 samples is matched with the average methylation presence of a site near the gene promoter. Silencing is observed with higher gene expression values for low methylation.

## 8.3 Hypothesis testing problems in gene expression data

### 8.3.1 Testing differentially expressed genes

We test whether, in average, the expression of a gene $g$ in healthy samples (denoted by $Y_g^{(1)}$) is equal or not to the expression of the same gene $g$ in tumor samples (denoted by $Y_g^{(2)}$). We assume a Gaussian likelihood on the gene expression differences

$$(Y_g^{(1)} - Y_g^{(2)}) \sim N(\mu_g, \sigma_g^2),$$

where $\mu_g$ is the parameter of interest that describes the differential expression mean for a specific gene $g$. We test the hypothesis

$$H_0 : \mu_g = 0, \quad \text{vs} \quad H_1 : \mu_g \neq 0,$$

independently for all genes $g \in [1, p]$. We consider the hiereachical Bayesian model described in Bochkina and Richardson (2007), who place a $\mathrm{N}(0, 10^4)$ distributed prior for $\mu_g$ and a lognormal distributed prior $LN(a, b)$ on $\sigma_g$. The hypeparameters $a$ and $b$ follow, independently, $a, b \sim \Gamma(\epsilon, \epsilon)$ with $\epsilon = 10^{-4}$. We compile the model in the R package **jags** (Plummer, 2016) and we generate $10,000$ MCMC samples from the posterior distribution of $\mu_g$. We approximate the probability $p_g = P(\mu_g > 0 | X_g, Y_g)$, and in Figure 8.4 we show, for all $g \in [1, p]$, the tail probabilities $t_g = 2\,(1 - \max(p_g, 1 - p_g))$. The 26% of the genes have a tail probability smaller than 0.01. Among them, the ten genes with the smallest tail probabilities are E2F5, CSF3R, CEP72, CKS2, IDH3A, PLXNA1, ODF2, WDR53, KIAA0513 and PHYH.

Besides, the distribution for non-significant tail probabilities resembles to the uniform distribution, which is expected when $H_0$ is true.

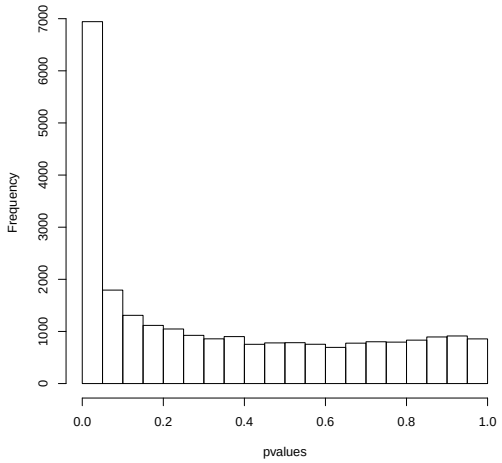



**Figure 8.4.** Evidence on differential expression tests: tail probabilities $(t_g)_{g=1}^{p}$ for the posterior distribution of the mean vector $(\mu_g)_{g=1}^{p}$.

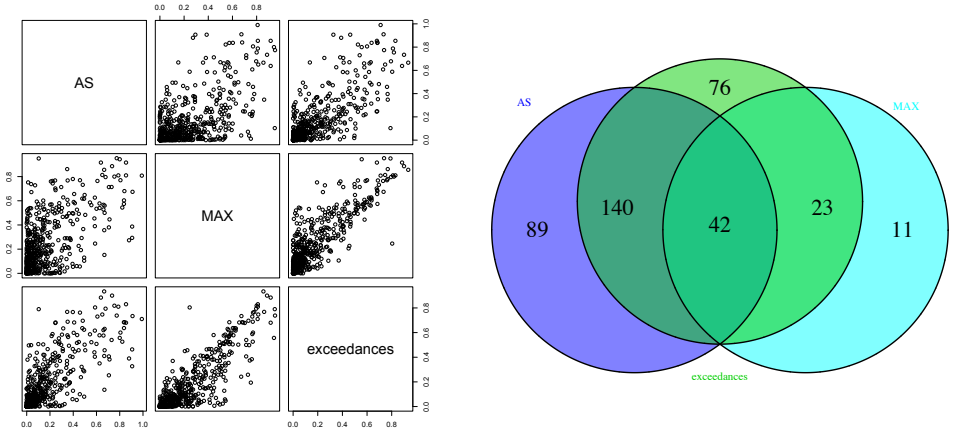
### 8.3.2 Testing the equality of gene expression correlation matrices

In this section we employ the hypothesis testing of equality of two correlation matrices (see Chapter 4) to assess differences in tumor/normal linear dependence structures for multiples subgroup of genes (of the total $25 \times 10^3$ that consists our data). These correspond to 1,320 standard gene pathways obtained from the MSig database (`http://www.broadinstitute.org/gsea/msigdb/index.jsp`).

In Figure 8.5(a) we present the permutation-approximated p-values using average of squares, extreme value and sum of exceedances test statistics. In the sum of exceedances test, we give the results for $w = 0$, though they are very similar to the p-values found for $w = 1$. 18% of the average of squares test p-values, 9% of the extreme value test p-values and 19% of the sum of exceedances test p-values are smaller than 0.01 and under $H_0$ we were expecting only 1%. About 4% of the lists have the three tests with p-values smaller than 0.01. Moreover, about 35% of the lists have the three p-values larger than 0.10, indicating some similarity in the correlation matrices even with conditions as different as cancer and healthy.

We further adjust the p-values for multiple testing by using a Benjamini-Hochberg (BH) correction (Benjamini and Hochberg, 1995), and in Figure 8.5(b) we present a Venn's diagram of the adjusted p-values smaller than 0.05. Among others, some of the pathway lists that had highly significant adjusted p-values (0.0003 significance level) in the three tests are: [1] "KEGG SPLICEOSOME", [2] "KEGG JAK STAT SIGNALING PATHWAY", [3] "BIOCARTA INFLAM PATHWAY" , [4] "BIOCARTA ERYTH PATHWAY", [5] "BIOCARTA STEM PATHWAY", [6] "REACTOME SIGNALING BY GPCR", [7] "REACTOME GPCR

DOWNSTREAM SIGNALING", [8] "REACTOME SIGNALING BY ILS", [9] "REACTOME CYTOKINE SIGNALING IN IMMUNE SYSTEM" and [10] "REACTOME TELOMERE MAINTENANCE". Of these 10 significant pathways lists, more than 50% of the genes within list [3] and [5] are also present in list [2].




**(a)** p-values **(b)** Venn's diagram

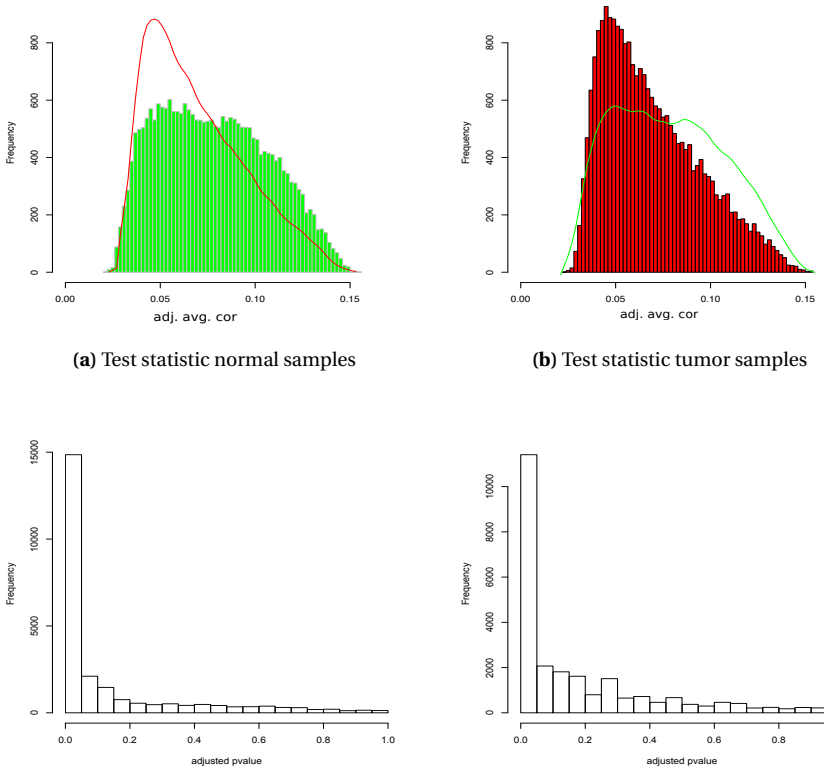
**Figure 8.5.** Evidence representation of equality of correlation matrices testing for 1,320 pathway list. In (a) there are the test p-values for each pathway list. In (b) Venn's diagram shows the number of rejected lists with an adjusted p-value smaller than 0.05.

### 8.3.3 Testing correlation matrix rows and reducing the number of genes

The two omic datasets analysed in this chapter are typical cases of very high dimensional data where the number of variables $p$ is of order of thousands. The statistical analysis of the whole data (e.g., estimation of precision matrices) involves dealing with matrices of size $p \times p$ which supposes a challenge for both number of operations and memory space. In this section we use the hypothesis testing procedures for correlation matrix rows studied in Section 4.4 on the gene expression dataset to reduce the dimension of the data by only keeping both highly correlated genes as well as highly differentially (tumor - normal) correlated genes.

We apply both adjusted average of squares and maximum test statistics to assess the evidence of highly correlated genes independently for all 24,526 genes and then we adjust the p-values to account for multiple testing using a BH correction. Figure 8.6 shows the adjusted average of squares test statistic in each gene, distinguishing between healthy and tumor samples as well as an histogram with the p-values of the underlying hypothesis testing procedure. In general, it seems that normal samples have larger correlations than tumor samples. For instance, 12,992 genes (53% of the total) have an adjusted p-value smaller than 0.01 for healthy samples whereas only 8,637 of the p-values for tumor samples genes (35%) are smaller than 0.01. Similarly, for the maximum test, 11,142 genes (45%) and 6,361 (26%) have adjusted p-values smaller than 0.01 for healthy and tumor samples, respectively. Hence, as for testing equality of correlation matrices, average of squares test finds a larger number of

genes with small p-values than maximum test. The ten genes with largest test statistics for healthy and tumor are: FAM96A, M-RIP, RRAGA, PITPNB, B2M, TGFB3, SULF1, CHST3, SCARA3 and DTNA for healthy; ATP8B2, PLEKHO1, KIAA0495, HOM-TES-103, MBNL1, PRMT2, GIMAP8, NNMT, CAST and RHOJ.




**(a)** Test statistic normal samples

**(b)** Test statistic tumor samples

**(c)** P-value normal samples

**(d)** P-value tumor samples

**Figure 8.6.** Adjusted average square correlation test statistic and p-values for the 24,526 genes in the the two datasets that distinguish between healthy and tumor samples.

Hypothesis testing to assess the evidence of differentially correlated genes is done using the permutation method for the average of squares test statistic. The p-values are also adjusted by the BH multiple testing correction. In total, 1,573 genes (6%) have adjusted p-values smaller than 0.01 of whose, only 87 genes were not highly significant in the non-zero correlation test described above. Among the differentially correlated genes, ten genes with largest test statistics are PCBD1, TMEM185B, RPL8, PPIL1, BYSL, SNRPC, EIF3S1, RALGDS, DDX21 and GCNT2. The correlation between the p-values found by testing differential expressed genes (in Section 8.3.1) and differentially correlated genes is 0.14. This is a significant but low level of dependence between the two hypothesis testing procedures.

In the following three sections we consider the problem of estimating conditional dependence structures for both tumor and healthy samples. The algorithms used are computationally demanding so to speed up the process we reduce the dimension size of the datasets such that we only select highly

correlated genes and differentially correlated genes. Let p-val$(g)^H$ and p-val$(g)^T$ be the adjusted p-values for healthy and tumor datasets respectively, and let p-val$(g)^D$ be the adjusted p-values for the difference matrix, we keep genes $g^*$ such that

$$g^* = \{g : \text{p-val}(g)^H < 0.01\} \cup \{g : \text{p-val}(g)^T < 0.01\} \cup \{g : \text{p-val}(g)^D < 0.01\},$$

where the three sets of p-values are found using average of squares test statistics.

The total number of remaining genes is 14,978 which is a reduction of the 39% of the data. We further use a hierarchical clustering procedure described in Müllner (2013) on the reduced dataset to separate the genes in different clusters. We use 1 minus the matrix of correlations for healthy genes as dissimilarity matrix to find 4 large clusters of size 1900, 5728, 5984 and 437 genes respectively. Other clusters are found but their sizes are very small (less than 100 genes) and are not considered for estimation. Figure 8.7 shows the heat map of the average squared correlation between and within clusters. Note that the darkest squares are given in the diagonal indicating large within cluster correlation magnitudes in comparison to between correlation magnitudes. Estimation of conditional dependence structures are done in the following sections within clusters, thus assuming conditional independence for genes between clusters. The only reason is the huge computational needs of the proposed joint estimation methods which make implausible the estimation of the whole network.

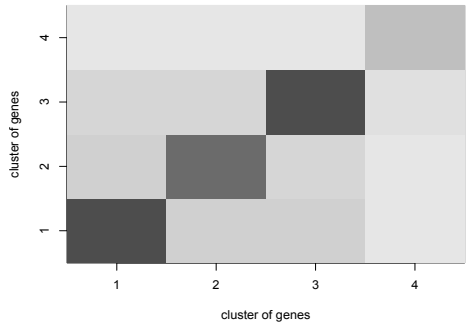



**Figure 8.7.** Heat-map that represents a measure of linear dependence between and within gene clusters, i.e., the darkness of the bins is proportional to the the average squared pairwise correlation between genes.

## 8.4 Graphical lasso to estimate network of genes

We estimate four gene expression conditional dependence structures separately for samples in the two medical conditions corresponding to genes within the four clusters found in Section 8.3.3. To do so, we use the neighbourhood selection lasso-penalized maximum likelihood approach (Meinshausen and Bühlmann, 2006) which is presented in Section 5.2. For each cluster and class of observations, we estimate 70 different graphs corresponding to different values for the tuning parameter $\lambda$ following

an equidistant sequence that ranges between 0.5 and 0.95. Values of $\lambda$ lower than 0.5 produce fully connected graphs and values above 0.95 produce no edges in the graph. We use AGNES, path connectivity (PC), A-MSE (subsampling) and vulnerability (VUL) regularization parameter selection approaches, see Chapter 5, to choose only few graphs, from the initial 70, to be analyzed. Table 8.2 shows the number of estimated edges in each of these graph structures. AGNES provides the densest graphs, PC and VUL find similar network sizes and A-MSE achieves the sparsest estimators. Besides, estimated networks for healthy samples tend to be denser than estimated networks for tumor samples in the VUL and PC approaches. AGNES and A-MSE are approaches that optimize risk functions based on clustering characteristics, but here hierarchical clustering is previously applied to separate the data in four groups of genes for estimation. Thus, the two selection methods turn out to produce uninformative networks (either too dense or too sparse).

**Table 8.2.** Number of estimated edges for either healthy or tumor selected graph structures by PC, AGNES, A-MSE and VUL. The number of healthy edges is larger than the number of tumor edges, especially for the PC and VUL estimated networks.

| | Cluster 1 | | | | Cluster 2 | | | |
|---|---|---|---|---|---|---|---|---|
| method | PC | AGNES | A-MSE | VUL | PC | AGNES | A-MSE | VUL |
| healthy | 700 | 2,351 | 0 | 700 | 1,706 | 7,580 | 85 | 1,018 |
| tumor | 377 | 2,209 | 1 | 591 | 1,023 | 7,809 | 130 | 752 |

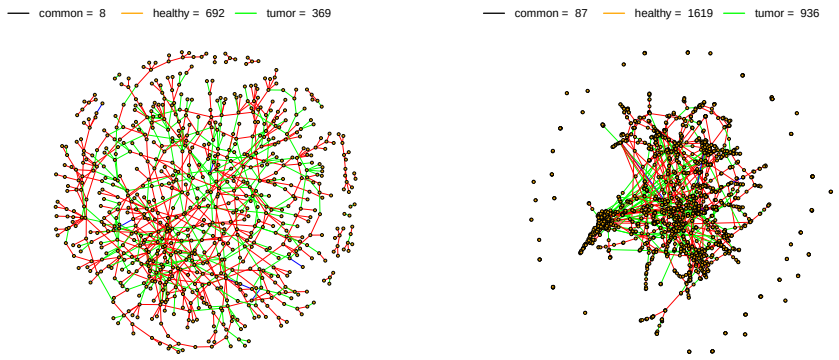
| | Cluster 3 | | | | Cluster 4 | | | |
|---|---|---|---|---|---|---|---|---|
| method | PC | AGNES | A-MSE | VUL | PC | AGNES | A-MSE | VUL |
| healthy | 2,900 | 7,510 | 117 | 1,183 | 395 | 458 | 2 | 409 |
| tumor | 1,916 | 8,391 | 75 | 509 | 86 | 400 | 4 | 233 |

Figure 8.8 illustrates the joint graphical representation of some of the estimated networks (only employing VUL and PC approaches) by finding the common edges and unique edges for each medical condition. Even though the number of common edges is small in comparison to the number of differential edges, it is still much larger than expected by chance (this is assessed by a resampling approach). As our analysis looks into healthy and tumor samples separately it is not well suited to establish how the network actually changes between the two conditions. This require a more refined approach that models both networks and their differences simultaneously and it is the focus of attention in the following sections.

Important genes, i.e., genes that interact with at least 7 other genes, include ADH6, ATP2B4, CHST9, CSEN, CYP2C9, FLJ20125, HAPLN4, HTRA3, MAP1LC3C, MAWBP, NR1H4, SCN3B, SMUG1, STX5A, TYROBP and VWCE.

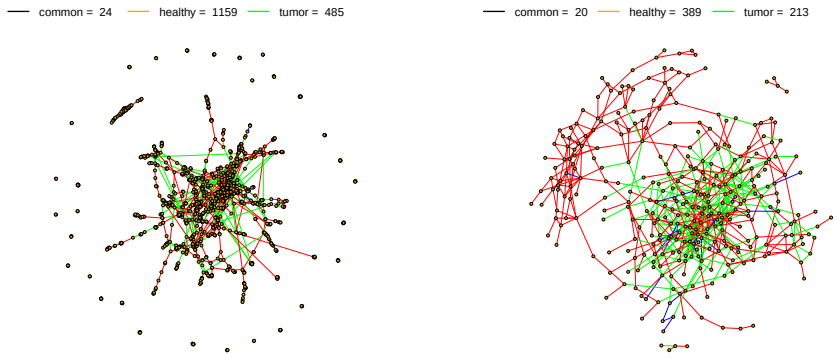
## 8.5 Estimation of joint gene expression networks

We estimate four fused-lasso precision matrices (following methodology described in Section 6.2) and their underlying gene-to-gene networks corresponding to the 4 clusters of genes described in




**(a)** PC best graph for cluster 1

**(b)** PC best graph for cluster 2




**(c)** VUL best graph for cluster 3

**(d)** VUL best graph for cluster 4

**Figure 8.8.** Estimated gene expression networks distinguishing between healthy edges (red), tumor edges (green) and common edges (blue). Only some of the PC and VUL estimated graphs are shown for visualization purposes.

Section 8.3.3 for only gene expression samples ($Y^{(1)}$: healthy and $Y^{(2)}$: tumor). We use significant levels $\alpha_1$ (which determines sparsity of the estimates) and $\alpha_2'$ (which determines similarity of the non-zero estimates) to tune the penalization parameters. For $\alpha_1$ we set the underlying expected number of false positive edges (EFP) with EFP = $150, 200, 200, 50$ respectively for each cluster, with $\alpha_1 = 2EFP/p(p-1)$. In terms of $\alpha_2'$ (see interpretation in Section 6.2.2) we use three different levels: $\alpha_2' = \{0.01, 0.05, 0.1\}$.

Table 8.3 provides the number of estimated edges common to the two medical conditions and the number of estimated differential edges: "healthy only" for edges only present in the network for healthy samples; and "tumor only" for edges only present in the network for tumor samples. The total number of estimated edges is much larger than the expected number of false positives which suggests certain strength in the results. Moreover, we observe that the number of differential edges is remarkably greater for healthy samples than for tumor samples in cluster 2 and cluster 3 whereas it is

slightly larger for tumor samples in cluster 1. Figure 8.9 shows the graphical representation of some of the estimated gene-to-gene networks, where black, orange and green edges differentiate between common, “healthy only” and “tumor only” edges respectively.

**Table 8.3.** Joint estimation of gene-to-gene networks in four clusters of genes: number of estimated edges, both common and differential edges, using several similarity tuning parameters $\alpha_2'$.

| | Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 |
| common | 459 | 441 | 414 | 2,791 | 2,487 | 2,263 |
| healthy only | 0 | 4 | 7 | 357 | 765 | 1,036 |
| tumor only | 2 | 16 | 41 | 92 | 272 | 421 |

| | Cluster 3 | | | Cluster 4 | | |
|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 |
| common | 4,001 | 3,340 | 3,027 | 107 | 107 | 109 |
| healthy only | 670 | 1,410 | 1,719 | 0 | 0 | 0 |
| tumor only | 294 | 921 | 1,193 | 0 | 0 | 0 |

## 8.6 Estimation of joint regression coefficient matrices

We consider the four sets of genes/sites described by the four clusters in Section 8.3.3. For each one of them, we match genes and methylation sites that are closeby so the analysis can be done at gene level. Besides, there are some methylation sites with variance equal to zero that are eliminated from the analysis. We consider gene expression samples as response variables ($Y^{(1)}$: healthy; and $Y^{(2)}$: tumor) and methylation presence samples as explanatory variables ($X^{(1)}$: healthy; and $X^{(2)}$: tumor). We estimate four fused-lasso regression coefficient matrices and their underlying site-to-gene directed networks following the methodology described at Section 6.3. We use different combinations of the tuning parameters $\alpha_1$ (for sparsity) and $\alpha_2'$ (for similarity of non-zero estimates). For $\alpha_1$ we set the expected number of false positive edges (EFP) with EFP = $150, 200, 200, 50$ for the four clusters, respectively. Then, $\alpha_1 = EFP/(pq)$. For $\alpha_2'$ (see interpretation in Section 6.3.2) we use the following three levels: $\alpha_2' = \{0.01, 0.05, 0.10\}$. Table 8.4 provides the estimated number of site-to-gene edges distinguishing among common, "'healthy only" and "tumor only" as defined previously in Section 8.5. The results resemble the estimated graph structures found in Table 8.3, i.e. "healthy only" edges are more frequent than "tumor only" in the large clusters 2 and 3 and less present in cluster 1.

Figure 8.10 shows the graphical representation of four of the estimated site-to-gene directed networks corresponding to the four clusters of genes with $\alpha_2' = 0.05$. Nodes in blue represent genes and nodes in white are methylation sites. Moreover, black, orange and green edges going from methylation sites to genes differentiate between common, "healthy only" and "tumor only" edges respectively. We identify several hub-methylation sites which are connected to many different genes. Moreover, "healthy only" and "tumor only" edges are found in clusters where almost all connections from one methyl site to genes are either black, green or orange.


**(a)** Cluster 1

**(b)** Cluster 2

**(c)** Cluster 3

**(d)** Cluster 4

**Figure 8.9.** Graph structure for the estimated gene-to-gene networks: in black there are the common edges, and in orange ("healthy only" ) and green ("tumor only" ) the differential connections.

## 8.7 Integration of estimated gene-to-gene and site-to-gene networks

In this section we consider a joint analysis using both gene-to-gene networks found in Section 8.5 and site-to-gene directed networks found in Section 8.6. Recall that methylation sites can be matched to genes that are nearby. These matching elements are expected to be negatively related. We corroborate this using data by counting the number of estimated edges in the site-to-gene network that link methylation sites with their matching genes. In Table 8.5 we separate the number of such estimated non-zero elements by the sign of their underlying regression coefficients. For instance, using the most conservative $\alpha_2' = 0.01$, summing up all clusters, a total of 17, for healthy, and 20, for tumor, matching genes and methylation sites are non-zero with 15 and 18 of them, respectively, being with a negative coefficient. Although the percentage of these estimated edges is very small, it is much larger (about 4, 12, 10 and 185 times for healthy and 4, 15, 17 and 185 for tumor) than expected by chance (whose levels can be found considering the estimated sparsity levels in the whole network).

We integrate the gene-to-gene networks with the site-to-gene directed networks by using the ANDnet approach (Gadaleta and Bessonov, 2015). This corresponds to the network where edges in

**Table 8.4.** Joint estimation of site-to-gene directed networks in four clusters of genes: number of estimated edges, both common and differential edges, using several similarity tuning parameters $\alpha_2'$.

| | Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 |
| common | 746 | 714 | 663 | 2,339 | 2,097 | 1,892 |
| healthy only | 6 | 40 | 79 | 398 | 943 | 1,395 |
| tumor only | 11 | 76 | 137 | 66 | 193 | 305 |

| | Cluster 3 | | | Cluster 4 | | |
|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 |
| common | 2,880 | 2,487 | 2,218 | 57 | 55 | 56 |
| healthy only | 556 | 1,196 | 1,633 | 0 | 5 | 7 |
| tumor only | 282 | 646 | 943 | 0 | 2 | 3 |

**Table 8.5.** Number of estimated edges that match methyl site (for explanatory variables) and gene nearby (for response variables). In + positive estimated regression coefficients, in − negative estimated regression coefficients.

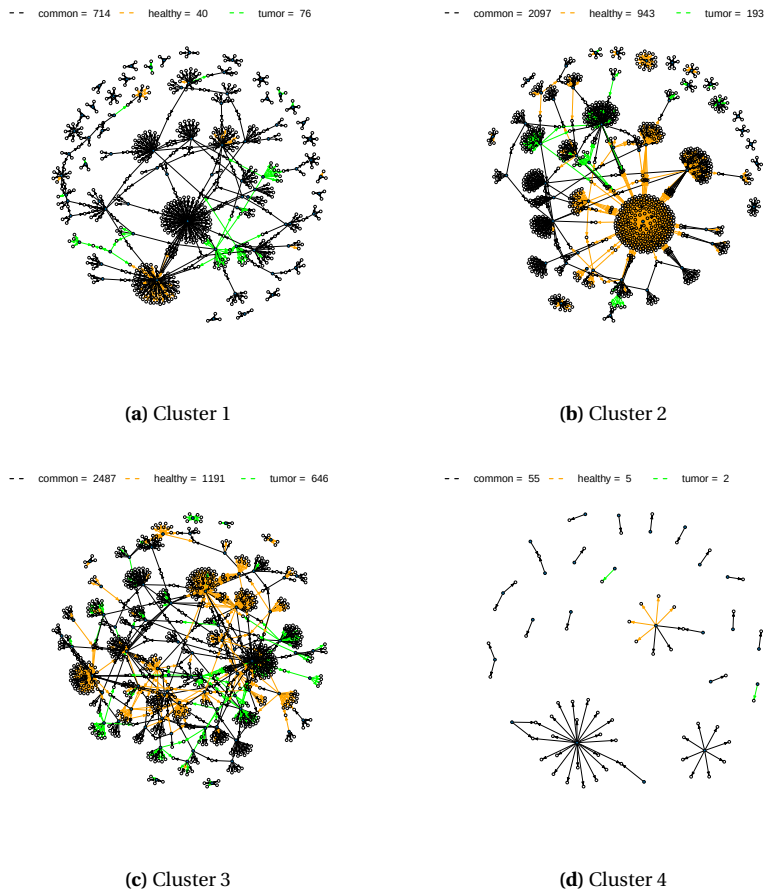
| | Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 |
| healthy | −1, +0 | −0, +0 | −0, +0 | −4, +1 | −4, +1 | −4, +1 |
| tumor | −1, +0 | −0, +0 | −0, +0 | −4, +1 | −4, +1 | −4, +1 |

| | Cluster 3 | | | Cluster 4 | | |
|---|---|---|---|---|---|---|
| $\alpha_2'$ | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 |
| healthy | −5, +1 | −4, +1 | −4, +1 | −5, +0 | −5, +0 | −5, +0 |
| tumor | −8, +1 | −8, +2 | −8, +2 | −5, +0 | −5, +0 | −5, +0 |

the gene-to-gene network and edges in the site-to-gene network coincide, i.e., the methylation sites are matched to the genes that are nearby so both networks are at gene-to-gene level. It turns out that the total number of coincidences between the two types of networks is low but larger than expected by chance in clusters 3 and 4. We use an exact Fisher test to assess the significance of the common links. Cluster 3 has at most 3/8 (healthy/tumor) shared associations (p-val = 0.09/ p-val $\ll 0.001$) and cluster 4 has at most 4/4 shared associations (p-val $\ll 0.001$ in both cases). For clusters 1 and 2, the number of shared edges is very low and could be obtained by chance.

## 8.8 Integration with biological pathway lists

We download 314 gene sets from the MSig database (Subramanian et al., 2005), which represent canonical pathways compiled from two sources: KeGG (Kanehisa et al., 2016) and Reactome (Milacic et al., 2012), and that contain at least 50 genes. For every gene set we estimate its gene-to-gene and site-to-gene joint networks. In order to determine which biological processes might be linked to changes in the gene/site associations between healthy and colon cancer samples, we use the hypothesis testing procedure described in Appendix B.3 which assesses whether the conditional dependence structures (i.e, gene-to-gene and site-to-gene) vary or do not vary in the presence of tumor cells. In terms of the gene-to-gene network, out of the 314 lists of genes, 119 and 19 contain more “healthy only" edges and




**(a)** Cluster 1

**(b)** Cluster 2

**(c)** Cluster 3

**(d)** Cluster 4

**Figure 8.10.** Graph structure for the estimated site-to-gene directed networks: in black there are the common edges, and in orange (”healthy only” ) and green (”tumor only” ) the differential connections. Methylation sites in blue circles map genes in whites circles.

“tumor only" edges, respectively, than expected by chance (at significance level 0.01). Not as many important pathway lists are present for the site-to-gene directed network where 11 out of the 314 pathway list are significant for “healthy only" edges and 5 lists are significant for “tumor only" edges (also at significance level 0.01). Especially for“healthy only” networks, there are more significant lists than expected under the null hypothesis of equality of the edges in the two medical conditions (where only 3 lists are expected to have a significance level lower than 0.01 under some mild independence conditions).

Table 8.6 presents some the most important lists that show enough evidence against the null hypothesis of equality of gene-to-gene networks between healthy and tumor samples. Among the significant lists, metabolism of proteins, cell cycle, immune system or signaling by GPCR are expected to change in carcinogen processes. Genes that are connected to many other genes in these statistically relevant networks are JAK1, KPNA4, DEFB103A, CD46, PRKCSH, PRSS2, SOS1, PFDN4, NUDC, EIF4G2 and TIAM2 (for “healthy only" edges, gene-to-gene network), MAPK12, MAPK11, GNB3, SLC3A1, SLC6A12, SLC24A6, HIST1H2BI, OR2B11 and OR2L8 (for “tumor only" edges, gene-to-gene network).

**Table 8.6.** Pathway lists with “tumor only" edges (T.) or “healthy only" edges (H.) being significantly different from zero (significance level 0.01) in gene-to-gene jointly estimated network.

```
-----------------------------------------------------------------------
[T1]  "KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY"
[T2]  "REACTOME_TRANSMEMBRANE_TRANSPORT_OF_SMALL_MOLECULES"
[T3]  "REACTOME_TRANSCRIPTION"
[T4]  "PID_IL12_2PATHWAY"
[T5]  "REACTOME_SIGNALING_BY_GPCR"
[H1] "REACTOME_IMMUNE_SYSTEM"
[H2] "REACTOME_METABOLISM_OF_PROTEINS"
[H3] "REACTOME_ADAPTIVE_IMMUNE_SYSTEM"
[H4] "REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM"
[H5] "REACTOME_CELL_CYCLE"
[H6] "REACTOME_INTERFERON_SIGNALING"
[H7]  "REACTOME_METABOLISM_OF_LIPIDS_AND_LIPOPROTEINS"
[H8] "REACTOME_METABOLISM_OF_RNA"
-----------------------------------------------------------------------
```

In Table 8.7 there are the top significant lists for the site-to-gene directed networks. These include gene sets as Tgf-beta signaling alterations which have been widely associated to colorectal cancer (Drabsch and Ten Dijke, 2012). Others as Gaba receptor acitivation, mRNA splicing or EGFR pathways are also link to have roles in tumor cells. Among the highly connected genes there are EGFR, TGFB1, TGFBR1, PIK3R1, PRKACA, SNRPB2, SNRPE and GNG8 (for “healthy only" edges, site-to-gene network), ITGA4, MYD88, IFNGR2 and FZD6 (for “tumor only" edges, site-to-gene network).

**Table 8.7.** Pathway lists with “tumor only" edges (T.) or “healthy only" edges (H.) being significantly different from zero (significance level 0.01) in site-to-gene jointly estimated network.

```
-----------------------------------------------------------------------
[T1] "KEGG_HEDGEHOG_SIGNALING_PATHWAY"
[T2] "KEGG_COMPLEMENT_AND_COAGULATION_CASCADES"
[T3] "WNT_SIGNALING"
[H1] "ST_FAS_SIGNALING_PATHWAY"
[H2] "PID_TGFBRPATHWAY"
[H3] "REACTOME_SIGNALING_BY_ERBB2"
[H4] "REACTOME_SIGNALING_BY_EGFR_IN_CANCER"
[H5] "REACTOME_RECRUITMENT_OF_MITOTIC_CENTROSOME_PROTEINS_AND_COMPLEXES"
[H6] "REACTOME_MRNA_SPLICING"
[H7] "REACTOME_GABA_RECEPTOR_ACTIVATION"
[H8] "REACTOME_SIGNALING_BY_TGF_BETA_RECEPTOR_COMPLEX"
-----------------------------------------------------------------------
```

## 8.9 Discussion

In this chapter we have considered both hypothesis testing and estimation methods to analyze two types of genomic data: gene expression and methylation presence. We have used hypothesis testing approaches on the gene expression data with the aim to reduce the number of genes for

estimation. Given the reduced datasets, we have estimated two types of networks, gene-to-gene network (employing WFGL on gene expression data) and site-to-gene network (employing WFRL to map methylation presence to gene expression). In general, these estimated networks contain more "healthy only" edges than "tumor only" edges, which may indicate that some of the gene-to-gene (site-to-gene) associations vanish on the appearance of the disease.

Focusing on the site-to-gene networks, we have confirmed in data the hypothesis that methylation presence can silence the expression of its gene promoter. Besides, we have observed that the estimated networks tend to present hub-based structures in which methylation sites are connected to many different genes. This can be due to genes (in gene expression level) being highly correlated between each other, and might suggest to find more accurate estimators that also account for the residuals linear dependence structure (see discussion in Chapter 6). Finally, we have estimated the same two types of networks using more than 300 gene sets that are known to have functions in biological processes. Particularly interesting is the comparison of differential network sizes for the studied pathway lists that corroborates previous findings in the literature that relate Tgf-beta signaling, Gaba receptor acitivation or mRNA splicing to colon cancer mutations.

# Chapter 9

# Conclusions

The objectives of this work were to develop statistical methodology for the testing and estimation of linear dependence structures such as correlation matrices, precision matrices and regression coefficient matrices when data are both paired and high-dimensional. This is motivated by the application to genomic data, where high-throughput technology is able to measure the whole genome profile of an organism for a specific location/tissue leading to datasets with large dimensions. Besides, the paired data setting is due to experimenting with samples on different tissues, that can be under different medical conditions (e.g., cancer and normal states), for the same individual.

Testing and estimation methods for gene interactions using high-dimensional data have been extensively studied in the literature in the past 20 years. Firstly, testing methods for global dependence structures are proposed in Li and Chen (2012) and Cai et al. (2016), among others, to assess whether two correlation matrices, which can represent the linear dependence structure of a group of genes on healthy and unhealthy tissues, are equal or not. Secondly, penalized maximum likelihood estimation approaches like lasso (Tibshirani, 1996; Lauritzen, 1996) are applied to infer (conditional dependence) gene associations in high-dimensional data by encouraging sparse graphical structures. The extension of these techniques to jointly estimating multiple matrices are considered in Danaher et al. (2014) or Lam et al. (2016), and can be relevant to finding gene interactions that distinguish between samples under several medical conditions. These methods in the literature are suitable when data are high-dimensional but ignore the dependence structure between datasets, which can be present when analyzing paired data. In this thesis, the main goal was to design convenient global testing approaches and joint estimation techniques that accounted for cases where there are two high-dimensional datasets whose observations can be paired.

In Chapter 4 we studied the hypothesis testing problem of equality of two correlation matrices for high-dimensional data with paired observations. We proposed test statistics that are based on the average of squares, maximum and sum of exceedances using the element-wise difference of Fisher transform sample correlation coefficients. The sum of exceedances test is a novel approach in this hypothesis testing problem that was introduced to link maximum test (which only uses the largest

magnitude of these transformed coefficients) and average of squares (which uses all the elements no matter their magnitude). For a threshold close to zero, the sum of exceedances test achieves similar powers to the average of squares test, whereas as the threshold increases, it finds powers that are closer to the maximum test. The null distributions of the three suggested test statistics were approximated by their limiting parametric distributions as well as by non-parametric distributions based on permutations. When determining the parametric distributions we considered both the assumption of asymptotic independence among correlation coefficients and a correction to account for dependence among elements in the differential sample correlation matrix. Although asymptotic independence distributions are remarkably fast to obtain, we developed dependence-corrections since we showed that estimates of the empirical size assuming asymptotic independence can be strongly biased.

In Chapter 5 and 6 we studied the related problem of estimating conditional dependence between variables when data are high-dimensional. In Chapter 5 we considered the estimation of sparse precision matrices in a single dataset whereas in Chapter 6 we extended the methodology for the more challenging problem of simultaneously estimating two precision matrices whose samples come from paired observations. Moreover, we also developed joint estimation methods for regression coefficient matrices which can be used for both independent and paired observations. The design of appropriate algorithms to estimate sparse precision matrices for single datasets was already well studied in the literature and we did not provide any other competitive method. However, in Chapter 5 we focused on the crucial issue of selecting the tuning parameter $\lambda$ in the lasso estimator which totally controls the complexity of the non-zero structure of the estimated precision matrix. We suggested to use several risk functions that optimize network characteristics as graph connectivity or clustering for selecting $\lambda$. These approaches only consider the graphical structures of the precision matrices, thus ignoring the value of such estimated matrices, and contrast to widely used likelihood based procedures like AIC, BIC (or its high-dimensional extension eBIC) and RIC which we found that tend to overestimate the size of the non-zero structures.

In Chapter 6 we employed joint estimation procedures to obtain conditional dependence relationships among variables for samples on two different classes. These procedures considered a larger sample size than $n$ for the estimation of a common network of variables in which edges coincide in the two classes. Besides, they were found to improve graph recovery rates when the two dependence structures that generate the data are similar, i.e., when many non-zero elements in the conditional dependence matrix for the first class of observations are equal to the same elements in the conditional dependence matrix for the second class of observations. The main contribution of the work in this chapter is that we adapted a current joint estimation algorithm to account for paired observations which led to better estimates of connections that vary between the two types of samples. Tuning parameters in these joint estimation problems were selected by controlling error rates related to the expected number of false positive edges in individual and differential networks. We argued that this is more informative than the initial selection problem as, for example, the number of estimated edges

can be compared to the expected number of false positive edges.

We find that the proposed methods complement well previous work done in the statistical literature for testing and estimation problems in high-dimensional data, and that these can be particularly useful to assess the dependence structures of multiple types of genomic data when observations are paired. This is not a rare situation in real data, and throughout the thesis we have provided the analysis of three different cases studies (examining colon cancer, lung cancer and psoriasis vulgaris disease) in which our techniques can help to answer questions that arise from biological processes. For instance, for the colon cancer data, in Chapter 8 we fully analyzed and integrated two types of genomic data representing gene expression and methylation presence for patients with colon cancer in which samples were provided for both tumor and healthy tissues.

The methodology and application work has been completed with the implementation of an R package, called ldstatsHD, which consists of functions that permit to conveniently employ the testing and estimation methods developed throughout the thesis. We though this could be an important contribution for the R scientific community so we have made it available in the CRAN repository for its use.

The presented methods have some limitations: (a) Testing methods are fast when assuming asymptotic null distributions without accounting for dependence between elements in the sample correlation matrices. However, these are only useful when the correlation matrices that generate the data are very sparse, which is not always verifiable in practice. For this reason, we presented methods that account for dependence employing permuted samples. While this assures correct representations of the p-values' distribution under the null hypothesis, it greatly increases the computational time. (b) In terms of the regularization parameter selection methods, we found that measuring network characteristics was useful to select a graph structure, as the interpretation of the network could be directly linked to the features used. However, we shall remark that the corresponding risk functions do not optimize the differences between true and estimated network characteristics, e.g., see A-MSE in Section 5.3.2, which would be the oracle solution. (c) For the joint estimation procedures, the main problem of the presented ADMM recursive procedures (Boyd, 2010) can be the lack of memory space in the machine. For each iteration of the algorithm, a dense estimator of two precision matrices (or also two regression coefficient matrices) is needed temporarily, which for large dimensions (more than $5,000$) requires the storage of numerical vectors of order of the square of the dimension and can slow down the computations.

Continuing the line of research of the thesis, there are some statistical problems that could be considered for a future work. In the testing methodology, we want to contemplate the usage of the sum of exceedances test statistic for higher criticism testing (Donoho and Jin, 2004), which would avoid the threshold selection problem, that is extensively discussed in Section 4.3.3, whilst obtaining optimal (or near optimal) power for the test. For sparsity tuning parameter selection methods we suggest to employ some network characteristics defining clustering, graph connectivity or graph vulnerability. However, other features of interest like the Estrada index (Estrada, 2011) or the degree distribution

(Costa and Rodrigues, 2007) could also be implemented. For the joint estimation methods proposed in Chapter 6, a logical extension would be considering a more general case when $K$ datasets, $K \geq 2$, are available and may also be dependent among each other. Moreover, to avoid memory issues when doing intensive computations we intend to employ efficient tools for big matrix storage as proposed in the package **bigmemory** (Kane et al., 2013).

# Acronyms

**AD** asymptotic-dependence

**ADMM** alternating direction method of multipliers

**AI** asymptotic-independence

**AIC** Akaike information criterion

**A-MSE** augmented mean square error for regularization parameter selection

**BIC** Bayesian information criterion

**CCA** canonical correlation analysis

**CD** conditional dependence

**CI** confidence interval

**CIA** co-inertia analysis

**CLIME** constrained L1-minimization for inverse (covariance) matrix estimation

**CV** cross-validation

**eBIC** extended Bayesian information criterion

**EFP** expected number of false positives

**EFPR** expected false positive rate

**FGL** fused graphical lasso

**FRL** fused regression lasso

**GEO** gene expression omnibus

**GGL** group graphical lasso

**HD** high dimensional

**HT** hypothesis testing

**ICA** independent components analysis

**JGL** joint graphical lasso

**LASSO** least absolute shrinkage and selector operator

**LARS** least angle regression

**LSE** least squares estimator

**MB** Meinshausen and Buhlmann neighborhood selection approach

**MCMC** Markov chain Monte Carlo

**MLE** maximum likelihood estimator

**MRCE** multivariate regression with covariance estimation

**NP** non-parametric

**PC** path connectivity for regularization parameter selection

**PCA** principal component analysis

**PLS** partial least squares

**RCON** row-column overlap norm

**SCAD** smoothly clipped absolute deviation

**StARS** stability approach to regularization selection

**TIGER** tuning-insensitive graph estimation and regression

**VUL** graph vulnerability for regularization parameter selection

**WFGL** weighted fused graphical lasso

**WFRL** weighted fused regression lasso

# Bibliography


Abegaz, F. and E. Wit (2015). Copula Gaussian graphical models with penalized ascent Monte Carlo EM algorithm. *Statistica Neerlandica 69*(4), 419–441.

Albieri, V. and V. Didelez (2014). Comparison of statistical methods for finding network motifs. *Statistical Applications in Genetics and Molecular Biology 13*(4), 403–422.

Aldous, D. (1989). *Probability Approximations via the Poisson Clumping Heuristic*. New York: Springer-Verlag.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, and G. O. Consortium (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics 25*(1), 25–29.

Augugliaro, L., A. M. Mineo, and E. C. Wit (2013). Differential geometric least angle regression: A differential geometric approach to sparse generalized linear models. *Journal of the Royal Statistical Society: Series B 75*(3), 471–498.

Augugliaro, L., A. M. Mineo, and E. C. Wit (2014). dglars: An R package to estimate sparse generalized linear models. *Journal of Statistical Software 59*(8), 1–40.

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

Banerjee, S. and S. Ghosal (2014). Posterior convergence rates for estimating large precision matrices using graphical models. *Electronic Journal of Statistics 8*(2), 2111–2137.

Bates, D. and M. Maechler (2016). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-7.1.

Bebek, G., M. Koyutürk, N. D. Price, and M. R. Chance (2012). Network biology methods integrating biological data for translational science. *Briefings in Bioinformatics 13*(4), 446–459.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B 57*(1), 289–300.

Bickel, P. and E. Levina (2008). Covariance regularization by thresholding. *The Annals of Statistics 36*(6), 2577–2604.

Bochkina, N. and S. Richardson (2007). Tail posterior probability for inference in pairwise and multiclass gene expression data. *Biometrics 63*, 1117–1125.

Boyd, S. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning 3*(1), 1–122.

Buhlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. New York: Springer-Verlag.

Caballe, A. (2017). *ldstatsHD: linear dependence statistics for high-dimensional data*. R package version 1.0.1.

Cai, T., H. Li, W. Liu, and J. Xie (2016). Joint estimation of multiple high-dimensional precision matrices. *Stat Sin 26*(2), 445–465.

Cai, T. and W. Liu (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association 106*(494), 672–684.

Cai, T., W. Liu, and X. Luo (2011). A constrained L1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*, 594–607.

Cai, T., W. Liu, and Y. Xia (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association 108*, 265–277.

Cai, T., W. Liu, and Y. Xia (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B 76*, 349–372.

Candes, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics 35*(6), 2313–2351.

Carey, V., L. Long, and R. Gentleman (2016). *RBGL: An interface to the BOOST graph library*. R package version 1.48.1.

Carr, D., N. Lewin-Koh, and M. Maechler (2015). *hexbin: Hexagonal Binning Routines*. R package version 1.27.1.

Chen, J. and Z. Chen (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Chun, H., M. Chen, B. Li, and H. Zhao (2013). Joint conditional Gaussian graphical models with multiple sources of genomic data. *Frontiers in Genetics 4*, 294.

Comon, P. (1994). Independent component analysis, A new concept? *Signal Processing 36*(3), 287–314.

Costa, L. and F. Rodrigues (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics 56*(1), 167–242.

Csárdi, G. and T. Nepusz (2006). The igraph software package for complex network research. *InterJournal Complex Systems 1695*, 1695.

Danaher, P. (2013). *JGL: Performs the Joint Graphical Lasso for sparse inverse covariance estimation on multiple classes*. R package version 2.3.

Danaher, P., P. Wang, and D. Witten (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B* (2006), 1–20.

Daniels, M. and R. Kass (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association 94*(448), 1254–1263.

Daniels, M. J. and R. E. Kass (2001). Shrinkage estimators for covariance matrices. *Biometrics 57*(4), 1173–84.

Depuydt, S., S. Trenkamp, A. R. Fernie, S. Elftieh, J.-P. Renou, M. Vuylsteke, M. Holsters, and D. Vereecke (2009). An integrated genomics approach to define niche establishment by Rhodococcus fascians. *Plant Physiology 149*(3), 1366–86.

Diedenhofen, B. and J. Musch (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS One 10*(4), 1–12.

Dobra, A., C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis 90*(1), 196–212.

Dokuzoglu, D. and V. Purtucuoglu (2017). Comprehensive analyses of gaussian graphical model under different biological networks. *Acta Physica Polonica 132*(3), 1106–1111.

Doledec, S. and D. Chessel (1994). Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biology 31*, 277–294.

Donoho, D. and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics 32*(3), 962–994.

Drabsch, Y. and P. Ten Dijke (2012). TGF-B signalling and its role in cancer progression and metastasis. *Cancer and Metastasis Reviews 31*(3-4), 553–568.

Dunn, O. and V. Clark (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association 64*, 366–377.

Durinck, S., Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber (2005). BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics 21*(16), 3439–3440.

Eastoe, E. F. and J. A. Tawn (2012). Modelling the distribution of the cluster maxima of exceedances of subasymptotic thresholds. *Biometrika 99*, 43–55.

Edgar, R., M. Domrachev, and A. E. Lash (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research 30*(1), 207–10.

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, H. Ishwaran, K. Knight, J. M. Loubes, P. Massart, D. Madigan, G. Ridgeway, S. Rosset, J. I. Zhu, R. A. Stine, B. A. Turlach, S. Weisberg, T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*(2), 407–499.

Eisen, M. and P. Spellman (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America 95*, 14863–14868.

Elston, R. (1975). On the correlation between correlations. *Biometrika 62*, 133–140.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics 29*, 751–760.

Estrada, E. (2011). *The Structure of Complex Networks*. New York: OXFORD University press.

Ezkurdia, I., D. Juan, J. M. Rodriguez, A. Frankish, M. Diekhans, J. Harrow, J. Vazquez, A. Valencia, and M. L. Tress (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics 23*(22), 5866–5878.

Fagan, A., A. C. Culhane, and D. G. Higgins (2007). A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics 7*(13), 2162–71.

Fan, J., Y. Feng, and Y. Wu (2009). Network exploration via the adaptive LASSO and SCAD penalties. *Annals of Applied Statistics 3*(2), 521–541.

Fan Jianqing, L. R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fayed, H. A. and A. F. Atiya (2014). An evaluation of the integral of the product of the error function and the normal probability density with application to the bivariate normal integral. *Mathematics of Computation 83*(285), 235–250.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation an found from a small sample. *Metron 1*, 3–32.

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

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics 1*(2), 302–332.

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

Fukushima, A. (2013). DiffCorr: An R package to analyze and visualize differential correlations in biological networks. *Gene 518*(1), 209–214.

Gadaleta, F. and K. Bessonov (2015). Integration of gene expression data and methylation reveals genetic networks for glioblastoma. *arXiv preprint 1506.00080*, 1–7.

Goldenberg, A. (2007). *Scalable Graphical Models for Social Networks*. Carnegie mellon university: Doctoral dissertation.

González, I., K.-A. L. Cao, M. J. Davis, and S. Déjean (2012). Visualising associations between paired 'omics' data sets. *BioData mining 5*(1), 19.

Gosline, S. J. C., S. J. Spencer, O. Ursu, and E. Fraenkel (2012). SAMNet: a network-based approach to integrate multi-dimensional high throughput datasets. *Integrative Biology: Quantitative Biosciences from Nano to Macro 4*(11), 1415–27.

Greene, C. S., J. Tan, M. Ung, J. H. Moore, and C. Cheng (2014). Big data bioinformatics. *Journal of Cellular Physiology 229*(12), 1896–1900.

Gu, X., G. Yin, and J. Lee (2013). Bayesian two-step Lasso strategy for biomarker selection in personalized medicine development for time-to-event endpoints. *Contemp Clin Trials 36*(2), 642–650.

Guo, J., E. Levina, G. Michailidis, and J. Zhu (2011). Joint estimation of multiple graphical models. *Biometrika 98*(1), 1–15.

Hans, C. (2009). Bayesian lasso regression. *Biometrika 96*(4), 835–845.

Hastie, T. and B. Efron (2013). *lars: Least Angle Regression, Lasso and Forward Stagewise.* R package version 1.2.

Hinoue, T., D. J. Weisenberger, C. P. E. Lange, H. Shen, H.-M. Byun, D. Van Den Berg, S. Malik, F. Pan, H. Noushmehr, C. M. van Dijk, R. a. E. M. Tollenaar, and P. W. Laird (2012). Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Research 22*, 271–82.

Hoerl, A. and R. Kennard (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*, 417–441.

Hughey, J. J. and A. J. Butte (2015). Robust meta-analysis of gene expression using the elastic net. *Nucleic Acids Research 43*(12), 1–11.

James, G. M. and P. Radchenko (2009). A generalized Dantzig selector with shrinkage tuning. *Biometrika 96*(2), 323–337.

Jameson, A. (1968). Solution of equation AX + XB = C by inversion of an M x M or N x N matrix. *SIAM J. Appl. Math 16*(5), 1020–1023.

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

Jolliffe, I. T., N. T. Trendafilov, and M. Uddin (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics 12*(3), 531–547.

Joyce, A. R. and B. Ø. Palsson (2006). The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology 7*(3), 198–210.

Kamburov, A., R. Cavill, T. M. D. Ebbels, R. Herwig, and H. C. Keun (2011). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics 27*(20), 2917–2918.

Kane, M. J., J. W. Emerson, and S. Weston (2013). Scalable strategies for computing with massive data. *Journal of Statistical Software 55*(14), 1–19.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research 44*(D1), D457–D462.

Kaufman, L. and P. Rousseeuw (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. New Jersey: John Wiley & sons.

Kim, Y., H. Choi, and H.-S. Oh (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association 103*(484), 1665–1673.

Kislinger, T., B. Cox, A. Kannan, C. Chung, P. Hu, A. Ignatchenko, M. S. Scott, A. O. Gramolini, Q. Morris, M. T. Hallett, J. Rossant, T. R. Hughes, B. Frey, and A. Emili (2006). Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell 125*(1), 173–86.

Kolar, M., L. Song, A. Ahmed, and E. P. Xing (2012). Estimating time-varying networks. *Annals of Applied Statistics 6*(1), 94–123.

Kullback, S. (1967). On testing correlation matrices. *Applied Statistics*, 80–85.

Kuznetsov, I, B., S. Hwang, and M. J. Zaki (2009). Integration of multiple types of genome-wide datasets and analysis of functional relationships among genes in the human genome. *University at Albany,*.

Kyung, M., J. Gill, M. Ghosh, and G. Casella (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis 5*(2), 369–411.

Lafferty, J., H. Liu, and L. Wasserman (2012). Sparse nonparametric graphical models. *Statistical Science 27*(4), 519–537.

Lam, K. Y., Z. M. Westrick, L. M. Christian, L. Christiaen, and R. Bonneau (2016). Fused regression for multi-source network inference. *PLoS Comput Biol 12*(12), e1005157.

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

Larntz, K. and M. D. Perlman (1985). A simple test for the equality of correlation matrices. *University of Minnesota*.

Lauritzen, S. (1996). *Graphical Models*. New York: OXFORD University press.

Lê Cao, K.-A., P. G. P. Martin, C. Robert-Granié, and P. Besse (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC bioinformatics 10*, 34.

Le Cao, K.-A., F. Rohart, I. Gonzalez, S. D. with key contributors Benoit Gautier, F. Bartolo, contributions from Pierre Monget, J. Coquery, F. Yao, and B. Liquet. (2016). *mixOmics: Omics Data Integration Project*. R package version 6.1.1.

Lê Cao, K.-A., D. Rossouw, C. Robert-Granié, and P. Besse (2008). A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology 7*(1), Article 35.

Leadbetter, M., G. Lindgren, and H. Rootzen (1983). *Extremes and Related Properties of Random Sequences and Processes*. New York: Springer-Verlang.

Ledoit, O. and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis 88*(2), 365–411.

Lee, W. (2015). Joint estimation of multiple precision matrices with common structures. *Journal of Machine Learning Research 16*, 1035–1062.

Li, J. and S. X. Chen (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics 40*, 908–940.

Li, X., T. Zhao, and H. Liu (2013). *camel: Calibrated Machine Learning*. R package version 0.2.0.

List, M., A.-C. Hauschild, Q. Tan, T. A. Kruse, J. Mollenhauer, J. Baumbach, and R. Batra (2014). Classification of Breast Cancer Subtypes by combining Gene Expression and DNA Methylation Data. *Journal of Integrative Bioinformatics 11*(2), 236.

Liu, H., F. Han, M. Yuan, and M. L. Jul (2012). High dimensional semiparametric gaussian copula graphical models. *Annals of statistics 40*(4), 2293–2326.

Liu, H., K. Roeder, and L. Wasserman (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. *Adv. Neural Inf. Process. Syst. 23*, 1432–1440.

Liu, H. and L. Wang (2012). TIGER : A tuning-insensitive approach for optimally estimating Gaussian graphical models. *arXiv preprint 1209.2437*.

Lu, T.-P., M.-H. Tsai, J.-M. Lee, C.-P. Hsu, P.-C. Chen, C.-W. Lin, J.-Y. Shih, P.-C. Yang, C. K. Hsiao, L.-C. Lai, and E. Y. Chuang (2010). Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer epidemiology, biomarkers & prevention 19*(10), 2590–7.

Lu, X., V. V. Jain, P. W. Finn, and D. L. Perkins (2007). Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. *Molecular systems biology 3*, 98.

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik (2016). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.5.

Mardia, K., J. Kent, and J. Bibby (1979). *Multivariate Analysis*. New York: Academic Press.

Marx, V. (2013). Biology: The big challenges of big data. *Nature 498*(7453), 255–260.

Mayer, C.-D., J. Lorent, and G. W. Horgan (2011). Exploratory analysis of multiple omics datasets using the adjusted RV coefficient. *Statistical Applications in Genetics and Molecular Biology 10*(1), 1–27.

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics & Data Analysis 52*, 374–393.

Meinshausen, N. and P. Bühlman (2010). Stability Selection. *Journal of the Royal Statistical Society: Series B 72*, 417–473.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics 34*, 1436–1462.

Meng, C., B. Kuster, A. C. Culhane, and A. Moghaddas Gholami (2014). A multivariate approach to the integration of multi-omics datasets. *BMC bioinformatics 15*, 162.

Milacic, M., R. Haw, K. Rothfels, G. Wu, D. Croft, H. Hermjakob, P. D'Eustachio, and L. Stein (2012). Annotating cancer variants and anti-cancer therapeutics in Reactome. *Cancers 4*(4), 1180–1211.

Mitra, R., P. Muller, and Y. Ji (2016). Bayesian graphical models for differential pathways. *Bayesian Analysis 11*(1), 99–124.

Mohammadi, A. and E. C. Wit (2015a). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis 10*(1), 109–138.

Mohammadi, A. and E. C. Wit (2015b). BDgraph: An R Package for Bayesian Structure Learning in Graphical Models. *arXiv preprint 1501.05108*.

Mohan, K., P. London, M. Fazel, D. Witten, and S.-I. Lee (2014). Node-based learning of multiple Gaussian graphical models. *Journal of Machine Learning Research 15*, 31.

Müllner, D. (2013). fastcluster: fast Hierarchical, agglomerative. *Journal of Statistical Software 53*(9), 1–18.

Newman, M. (2003). The structure and function of complex networks. *SIAM REVIEW 45*(2), 167–256.

O'Brien, G. (1987). Extreme values for stationary and Markov sequences. *The Annals of Probability*, 281–291.

Olkin, I. and J. Finn (1990). Testing correlated correlations. *Psychological Bulletin 108*(2), 330.

Park, T. and G. Casella (2008). The Bayesian lasso. *Journal of the American Statistical Association 103*(482), 681–686.

Peng, J., P. Wang, N. Zhou, and J. Zhu (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association 104*, 735–746.

Peterson, C., F. Stingo, and M. Vannucci (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association 110*(509), 159–174.

Plummer, M. (2016). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-6.

Pourahmadi, M. (2007). Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance correlation parameters. *Biometrika 94*(4), 1006–1013.

Pourahmadi, M. (2011). Covariance estimation: the GLM and regularization perspectives. *Statistical Science 26*(3), 369–387.

Raghunathan, T. (2003). An approximate test for homogeneity of correlated correlation coefficients. *Quality and Quantity*, 99–110.

Reinert, G. (2009). *Statistical Inference for Networks*. Oxford university,.

Renner, M., T. Wolf, H. Meyer, W. Hartmann, R. Penzel, A. Ulrich, B. Lehner, V. Hovestadt, E. Czwan, G. Egerer, T. Schmitt, I. Alldinger, E. K. Renker, V. Ehemann, R. Eils, E. Wardelmann, R. Büttner, P. Lichter, B. Brors, P. Schirmacher, and G. Mechtersheimer (2013). Integrative DNA methylation and gene expression analysis in high-grade soft tissue sarcomas. *Genome Biology 14*(12), r137.

Reynolds, A. P., G. Richards, B. De La Iglesia, and V. J. Rayward-Smith (2006). Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms 5*(4), 475–504.

Robins, B. J. M., R. Scheines, P. Spirtes, and L. Wasserman (2003). Uniform consistency in causal inference. *Biometrika 90*(3), 491–515.

Rothman, A. J. (2013). *MRCE: Multivariate regression with covariance estimation.* R package version 2.0.

Rothman, A. J., E. Levina, and J. Zhu (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association 104*(485), 177–186.

Rothman, A. J., E. Levina, and J. Zhu (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics 19*(4), 947–962.

Rousseeuw, P., C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, M. Koller, and M. Maechler (2016). *robustbase: Basic Robust Statistics.* R package version 0.92-6.

Rousseeuw, P., A. Struyf, and M. Hubert (2013). cluster: Cluster Analysis Basics and Extensions.

Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics 29*, 391–411.

Sánchez, A. and M. C. R. D. Villa (2008). A tutorial review of microarray data analysis. *Universitat de Barcelona.*

Schäfer, J., R. Opgen-Rhein, and K. Strimmer (2006). Reverse engineering genetic networks using the GeneNet package. *R News*, 50–53.

Schäfer, J., R. Opgen-Rhein, V. Zuber, M. Ahdesmäki, A. P. D. Silva, and K. Strimmer. (2015). *corpcor: Efficient Estimation of Covariance and (Partial) Correlation.* R package version 1.6.8.

Schäfer, J. and K. Strimmer (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics 21*(6), 754–64.

Schmid, R., P. Baum, C. Ittrich, K. Fundel-clemens, W. Huber, B. Brors, R. Eils, A. Weith, D. Mennerich, and K. Quast (2010). Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3. *BMC Genomics 11*(1), 349.

Schott, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Computational Statistics & Data Analysis 51*, 6535–6542.

Shen, H. and J. Z. Huang (2008). Sparse principal component analysis via low rank matrix approximation. *Journal of Multivariate Analysis 99*(6), 1015–1034.

Sheng, J., H. Deng, V. Calhoun, and Y. Wang (2011). Integrated Analysis of Gene Expression and Copy Number Data on Gene Shaving using Independent Component Analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 8*, 1568–1579.

Sibuya, M. (1959). Bivariate extreme statistics, I. *Annals of the Institute of Statistical Mathematics 11*, 195–210.

Simeonov, K. P. and D. S. Himmelstein (2015). Lung cancer incidence decreases with elevation: evidence for oxygen as an inhaled carcinogen. *PeerJ 2*, 1 – 24.

Srivastava, M. S., H. Yanagihara, and T. Kubokawa (2014). Tests for covariance matrices in high dimension with less sample size. *Journal of Multivariate Analysis 130*, 289–309.

Steiger, J. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin 87*, 245–251.

Stephenson, A. G. (2002). evd: Extreme Value Distributions. *R News 2*.

Storey, J. D., D. A. Bass, and D. Robinson (2015). *qvalue: Q-value estimation for false discovery rate control*. R package version 2.4.2.

Suárez-Fariñas, M., K. Li, J. Fuentes-Duculan, K. Hayden, C. Brodmerkel, and J. G. Krueger (2012). Expanding the Psoriasis disease profile: interrogation of the skin and serum of patients with moderate-to-severe Psoriasis. *Journal of Investigative Dermatology 132*(11), 2552–2564.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. a. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America 102*(43), 15545–50.

Thioulouse, J., D. Chessel, S. Doledec, and J. M. Olivier (1997). ADE-4: A multivariate analysis and graphical display software. *Statistics and Computing 7*(1), 75–83.

Tiago de Oliveira, J. (1962). Structure theory of bivariate extremes, extensions. *Estudos Math. Estat. Econom. 7*, 165–195.

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

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, K. Knight, and I. B. M. T. J. Watson (2005). Via the fused lasso and smoothness sparsity. *Journal of the Royal Statistical Society: Series B 67*(1), 91–108.

Timpe, L., D. Li, T. Yen, J. Wong, R. Yen, B. Macher, and A. Piryatinska (2015). Mining the breast cancer proteome for predictors of drug sensitivity. *Journal of Proteomics & Bioinformatics 8*(9), 204.

Van De Geer, S. A. (2008). High-dimensional generalized linear models and the Lasso. *Annals of Statistics 36*(2), 614–645.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer.

Wagner, J. R., S. Busche, B. Ge, T. Kwan, T. Pastinen, and M. Blanchette (2014). The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biology 15*(2), R37.

Wahl, S., N. Fenske, S. Zeilinger, K. Suhre, C. Gieger, M. Waldenberger, H. Grallert, and M. Schmid (2014). On the potential of models for location and scale for genome-wide DNA methylation data. *BMC Bioinformatics 15*(1), 232.

Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis 7*(4), 867–886.

Wang, T., Z. Ren, Y. Ding, Z. Fang, Z. Sun, M. L. MacDonald, R. A. Sweet, J. Wang, and W. Chen (2016). FastGGM: an efficient algorithm for the inference of Gaussian graphical model in biological networks. *PLoS Computational Biology 12*(2), 1–16.

Wang, Z., E. Curry, and G. Montana (2014). Network-guided regression for detecting associations between DNA methylation and gene expression. *Bioinformatics 30*(19), 2693–2701.

Wanichthanarak, K., J. F. Fahrmann, and D. Grapov (2015). Genomic, proteomic, and metabolomic data integration strategies. *Biomarker Insights 10*(4), 1–6.

Wasserman, L. and K. Roeder (2009). High-dimensional variable selection. *The Annals of Statistics 37*(5A), 2178–2201.

Wit, E. and A. Abbruzzo (2015). Factorial graphical models for dynamic networks. *Network Science 3*, 37–57.

Witten, D. M., R. Tibshirani, and T. Hastie (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics 10*(3), 515–34.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*, 391–420.

Wong, F., C. Carter, and R. Kohn (2003). Efficient estimation of covariance selection models. *Biometrika 90*(4), 809–830.

Wuertz, D. (2013). *fExtremes: Rmetrics - Extreme Financial Market Data*. R package version 3010.81.

Xie, Y., Y. Liu, and W. Valdar (2016). Joint estimation of multiple dependent Gaussian graphical models with applications to mouse genomics. *Biometrika 103*(3), 493–511.

Yang, J. Y., K. Yoshihara, K. Tanaka, M. Hatae, H. Masuzaki, H. Itamochi, M. Takano, K. Ushijima, J. L. Tanyi, G. Coukos, Y. Lu, G. B. Mills, and R. G. W. Verhaak (2013). Predicting time to ovarian carcinoma recurrence using protein markers. *Journal of Clinical Investigation 123*(9), 3740–3750.

Yao, F., J. Coquery, and K.-A. Lê Cao (2012). Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. *BMC bioinformatics 13*(1), 24.

Yee, T. W. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software 32*, 1–34.

Yelland, P. (2010). An introduction to correspondence analysis. *The Mathematica Journal 12*, 1–23.

Yi, G., S. H. Sze, and M. R. Thon (2007). Identifying clusters of functionally related genes in genomes. *Bioinformatics 23*(9), 1053–1060.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research 11*, 2261–2286.

Yuan, M. and Y. Lin (2005). Efficient empirical bayes variable selection and estimation in linear models. *Journal of the American Statistical Association 100*(472), 1215–1225.

Yuan, M. and Y. Lin (2007). Model selection and estimation in the gaussian graphical model. *Biometrika 94*(1), 19–35.

Zeng, L. and J. Xie (2012). Group variable selection via SCAD-L2. *Statistics 48*(1), 49–66.

Zhang, B. and Y. Wang (2012). Learning structural changes of Gaussian graphical models in controlled experiments. *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*.

Zhang, Y. and X. Shen (2010). Model selection procedure for high-dimensional data. *Statistical Analysis and Data Mining 3*(5), 350–358.

Zhao, P. and B. Yu (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research 7*, 2541–2563.

Zhao, S., T. Cai, and H. Li (2014). Direct estimation of differential networks. *Biometrika 101*(2), 253–268.

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

Zhou, C., F. Han, X. Zhang, and H. Liu (2015). An extreme-value approach for testing the equality of large U-statistic based correlation matrices. *arXiv preprint 1502.03211*.

Zhou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

Zhou, S., J. Lafferty, and L. Wasserman (2010). Time varying undirected graphs. *Machine Learning 80*(2), 295–319.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

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

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics 15*(2), 265–286.

# Appendix A

# Proofs and derivations of hypothesis testing methods

## A.1 Variance of mean of squares for dependent samples

Here we proof the result in Lemma 1 that gives the expression of the variance of the average of squares for dependent random variables. Consider $n$ dependent random variables $Z = (z_1, \ldots, z_n)$ which marginally follow a standard normal distribution. Take $\mathbf{E}[z_i^2] = \mu_2 = 1$ and $\mathbf{E}[z_i^4] = \mu_4 = 3$ for any $z_i \in Z$ and $\bar{\gamma}_2 = 2(n(n-1))^{-1} \sum_{i<j} \text{cov}(z_i^2, z_j^2)$ which is function of the dependence structure between variables.

The mean square of elements in $Z$ is found by $S^2 = n^{-1} \sum_{i=1}^n z_i^2$ and has variance $\text{var}[S^2] = \mathbf{E}[S^4] - \mathbf{E}[S^2]^2$. The second term is determined by $\mu_2$ such that $\mathbf{E}[S^2]^2 = \mu_2^2$. Moreover, the first term is expressed as

$$\mathbf{E}[S^4] = \mathbf{E}[n^{-2}(\sum_{i=1}^n z_i^2)^2] = \mu_4 / n + (\bar{\gamma}_2 + \mu_2)(n-1)/n.$$

Hence, $\text{var}[S^2] = (\mu_4 - \mu_2^2)/n + \bar{\gamma}_2 (n-1)/n$.

## A.2 First and second order statistics for estimated exceedances

We show the expected value and variance of $(|\hat{d}_t| - w_u u)^2 | \hat{d}_t^2 > u^2$ for a general case of $d_t$ being any value. This is used in the paper to obtain the lower bound of the power of the sum of exceedances test, and also to select the threshold $u$.

**Scenario** $w_u = 0$

Take $x_t = \hat{d}_t \sim N(d_t, 1)$. Expected value is determined by

$$\begin{aligned} E[x_t^2 \mid x_t^2 > u^2] &= \frac{\int_u^\infty x_t^2 (2\pi)^{-1/2} e^{-\frac{(x_t-d_t)^2}{2}} dx_t + \int_{-\infty}^{-u} x_t^2 (2\pi)^{-1/2} e^{-\frac{(x_t-d_t)^2}{2}} dx_t}{\Phi(d_t-u)+\Phi(-d_t-u)} \\ &= 1 + d_t^2 + \frac{(u-d_t)\varphi(u-d_t)}{\Phi(d_t-u)+\Phi(-d_t-u)} + \frac{(u+d_t)\varphi(-u-d_t)}{\Phi(d_t-u)+\Phi(-d_t-u)} \\ &+ 2d_t \frac{\varphi(u-d_t)-\varphi(-u-d_t)}{\Phi(d_t-u)+\Phi(-d_t-u)} \\ &= 1 + d_t^2 + A + B, \end{aligned} \tag{A.1}$$

where $A = u\{\varphi(u-d_t)+\varphi(u+d_t)\}/\{\Phi(d_t-u)+\Phi(-d_t-u)\}$ and $B = d_t\{\varphi(u-d_t)-\varphi(u+d_t)\}/\{\Phi(d_t-u)+\Phi(-d_t-u)\}$. If $|d_t| > u$, then $E[x_t^2 \mid x_t^2 > u^2] \geq d_t^2 + 1$. Under $H_0$, where $d_t = 0$, $\mu_0 = 1 + u\frac{\varphi(u)}{1-\Phi(u)}$.

The expression for the variance is

$$\begin{aligned} \text{var}[x_t^2 \mid x_t^2 > u^2] &= \frac{(2\pi)^{-1/2}[\int_u^\infty x_t^4 e^{-\frac{(x_t-d_t)^2}{2}} dx_t + \int_{-\infty}^{-u} x_t^4 e^{-\frac{(x_t-d_t)^2}{2}} dx_t]}{\Phi(d_t-u)+\Phi(-d_t-u)} - E[x_t^2 \mid x_t^2 > u^2]^2 \\ &= d_t^4 + d_t^3 D + d_t^2(6+uC) + d_t(u^2+5)D + (u^3+3u)C + 3 \\ &- E[x_t^2 \mid x_t^2 > u^2]^2, \end{aligned} \tag{A.2}$$

where $C = \{(\varphi(u+d_t)+\varphi(u-d_t)\}/\{\Phi(d_t-u)+\Phi(-d_t-u)\}$ and $D = \{(\varphi(u+d_t)-\varphi(u-d_t)\}/\{\Phi(d_t-u)+\Phi(-d_t-u)\}$. Under $H_0$, $\sigma_0^2 = 3 + (u^3+3u)\frac{\varphi(u)}{1-\Phi(u)} - \mu_0^2$.

**Scenario** $w_u = 1$

Take $x_t = \hat{d}_t \sim N(d_t, 1)$. Expected value is determined by

$$\begin{aligned} E[(|x|-u)_t^2 \mid x_t^2 > u^2] &= \frac{1}{\sqrt{2\pi}} \left[ \frac{\int_u^\infty (x_t-u)^2 e^{-\frac{(x_t-d_t)^2}{2}} dx_t + \int_{-\infty}^{-u} (-x_t-u)^2 e^{-\frac{(x_t-d_t)^2}{2}} dx_t}{\Phi(d_t-u)+\Phi(-d_t-u)} \right] \\ &= E[x_t^2 \mid x_t^2 > u^2] + u^2 - 2u\frac{\varphi(d_t-u)+\varphi(-d_t-u)}{\Phi(d_t-u)+\Phi(-d_t-u)} \\ &- 2d_t u \frac{\Phi(d_t-u)-\Phi(-d_t-u)}{\Phi(d_t-u)+\Phi(-d_t-u)} \\ &= 1 + d_t^2 + u^2 + A + B - E, \end{aligned} \tag{A.3}$$

where A and B are defined above, and

$$E = 2u\frac{\varphi(d_t-u)+\varphi(-d_t-u)}{\Phi(d_t-u)+\Phi(-d_t-u)} - 2d_t u\frac{\Phi(d_t-u)-\Phi(-d_t-u)}{\Phi(d_t-u)+\Phi(-d_t-u)}.$$

Note that if $|d_t| > u$, then $E[(|x|-u)_t^2 \mid x_t^2 > u^2] \geq (|d_t|-u)^2 + 1$ can be used as a lower bound. Under $H_0$, $\mu_1 = (u^2+1) - u\frac{\varphi(u)}{1-\Phi(u)}$.

The expression for the variance is

$$\begin{aligned}\text{var}[(|x|-u)_t^2 \mid x_t^2 > u^2] = &\frac{1}{\sqrt{2\pi}}\left[\frac{\int_u^\infty (x_t-u)^4 e^{-\frac{(x_t-d_t)^2}{2}}\,dx_t + \int_{-\infty}^{-u}(-x_t-u)^4 e^{-\frac{(x_t-d_t)^2}{2}}\,dx_t}{\Phi(d_t-u)+\Phi(-d_t-u)}\right] \\ &- E[(|x|-u)_t^2 \mid x_t^2 > u^2]^2 \\ =&E[x_t^4 \mid x_t^2 > u^2] + 6uE[x_t^2 \mid x_t^2 > u^2] + u^4 + 4u^3(d_t C - D) - F,\end{aligned} \tag{A.4}$$

where

$$\begin{aligned}F = &8uC + 12ud_t^2 C + (4ud_t^3 + 12d_t u)(\Phi(d_t-u)-\Phi(-d_t-u))/\{\Phi(d_t-u)+\Phi(-d_t-u)\} \\ &+ \frac{4u\{(u-d_t)^2\varphi(u-d_t)+(u+d_t)\varphi(u+d_t)\} + 12d_t u\{(u-d_t)\varphi(u-d_t)-(u+d_t)\varphi(u+d_t)\}}{\Phi(d_t-u)+\Phi(-d_t-u)}.\end{aligned}$$

Under $H_0$, $\sigma_0^2 = 3+u^4+6u^2-(5u+u^3)\frac{\varphi(u)}{1-\Phi(u)} - \mu_1^2$.

## A.3 Gumbel approximation of extreme value test statistic

Let $V_{tj} = \text{cov}(\hat{d}_t, \hat{d}_j)$ be the covariance between two elements in the matrix $\hat{D}$. For op $\in \{=, \neq\}$, we define

$$v_t^{\text{op}} = \sum_{j\in A} I(V_{tj} \text{ op } 0), \;\; A = M \setminus \{t\},$$

so $v_t^{=} + v_t^{\neq} = m-1$. Following sparsity constrains in Meinshausen and Bühlmann (2006), the sparsity level $v_t^{\neq}$ is assumed to be

$$v_t^{\neq} = O(m_t^{\eta}) = L(m)\, m^{\eta_t},$$

where $0 \le \eta_t < 1$ and $L(m)$ is a slowly varying function, i.e., $\lim_{m\to\infty} L(mx)/L(m) \to 1$. Moreover,

$$v_t^{=} = m - 1 - O(m_t^{\eta}) = m\,(1 - m^{-1} - L(m)\, m^{\eta_t - 1}) = m\,(1+o(1)) = L(m)\, m.$$

Assume that $\max_{i<j}|V_{tj}| < 1$ and that there exists a permutation $\hat{D}^*$ of elements in $\hat{D}$ such that $V^* = [\text{cov}(\hat{d}_t^*, \hat{d}_j^*)]$ is block diagonal. Then for all rows in $V^*$ there exists $h$ such that for all $j > h : V_{tj}^* = 0$. Let $\epsilon_n \in o(1/\log n)$ and take $\epsilon$ any positive number such that $\max_{i<j}|V_{ij}^*| + \epsilon < 1$. Define

$$\rho_n = \begin{cases} \max_{t<j}|V_{tj}| + \epsilon, & n < |j-t| \\ \epsilon_n, & n \ge |k-t|. \end{cases}$$

It then follows that $|V_{tj}^*| < \rho_{|j-t|}$, and $\rho_n \log n \to 0$ as $n \to \infty$. This is a sufficient condition (Leadbetter et al., 1983) for the distribution of $T_{MAX} = \max_{t\in M}|\hat{d}_t|$ to converge weakly to a Gumbel distribution.

## A.4 Sub-asymptotic model for structured non-stationary processes

### A.4.1 Heuristic

The heuristic approach proposed in this section follows results and notation from Aldous (1989). Let $\mathscr{S}_x = \{t \in M : |\hat{d}_t| \geq x\}$ be a random set that, for large $x$, defines a sparse mosaic on the sub-integer lattice $\mathbb{Z}^2$ corresponding to the lower triangular matrix $M$ (defined in eq.(4.2)). We assume a structured dependence structure on the process $(\hat{d}_t : t \in M)$ such that $\mathscr{S}_x$ contains several (near) independent clusters defined by a compound Bernoulli process with cluster intensity $\lambda_x(t)$. Let $C_x(t)$ denote the cluster area (or cardinality) at point $t$, and assume that as the number of variables increase, $C_x(t)$, in any position $t \in M$, is finite and does not exceed a given constant $\kappa$. Besides, assume that $\lambda_x(t)$ and $C_x(t)$ do not vary much as $t$ moves around the same cluster. For $x(m) = \mu(m) + \sigma(m)\, x$, $x \in \mathbb{R}$, the distribution of $T_M = \max_{t \in M} |\hat{d}_t|$ can be approximated by

$$
\begin{aligned}
\Pr(T_M < x(m)) &= \Pr(\mathscr{S}_{x(m)} \cap M \text{ empty}) \\
&\doteq \exp\left( - \int_M \lambda_{x(m)}(t)\, \mathrm{d}t \right) \\
&\doteq \exp\left\{ - \int_M \frac{\Pr(|\hat{d}| > x(m))}{\mathbb{E}(C_t^{x(m)})}\, \mathrm{d}t \right\} \\
&= \exp\left\{ - \Pr(|\hat{d}| > x(m)) \int_M \frac{1}{\mathbb{E}(C_t^{x(m)})}\, \mathrm{d}t \right\} \\
&= \exp\left\{ - \Pr(|\hat{d}| > x(m)) \sum_{t \in M} \frac{1}{\mathbb{E}(C_t^{x(m)})} \right\},
\end{aligned}
$$

where $\hat{d} \sim N(0,1)$, $\mathbb{E}(C_t^x)$ is the expected cluster area at cell $t$ and threshold level $x$. The result obtained above is equivalent to the cumulative distribution function of the cluster maxima for sub-asymptotic models ( $u < \sup\{|\hat{d}_t| : \Phi(|\hat{d}_t|) < 1\}$) in a stationary process (Eastoe and Tawn, 2012),

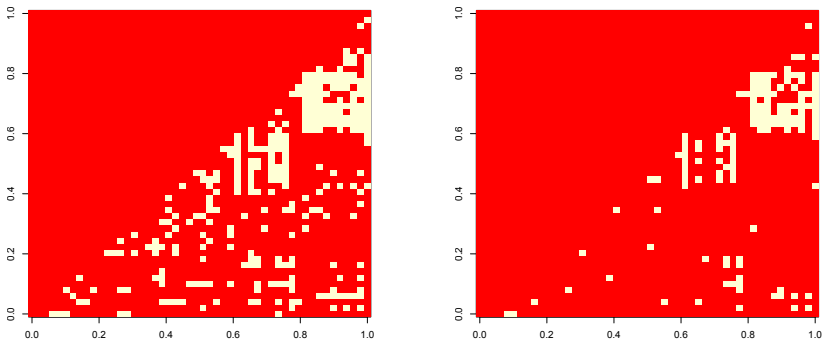
$$
\begin{aligned}
\Pr(T_M < x) &= \exp\left\{ - m\theta_x \Pr(|\hat{d}_t| > x) \right\} \\
&\doteq \exp\left[ - m p_u \theta_x \exp\{-(x-u)/\sigma_u\} \right], \quad (x \geq u)
\end{aligned}
$$

when $m\theta_x = \sum_{t=1}^{m} \frac{1}{\mathbb{E}(C_t^x)}$ and with $p_u = \Pr(|\hat{d}_t| > u)$.

### A.4.2 Exceedances for simulated data using block diagonal correlation matrices

We consider a simple toy example to show the behavior of sparse mosaics $\mathscr{S}_x$ over different values $x$. We use a block-diagonal correlation matrix with 5 blocks of 10 variables each. We take the same structure within every block so off-diagonal elements are equal to 0.7 (this can be varied to see the impact on $\mathscr{S}_x$). For first condition $Y^{(1)}$, we generate data by a multivariate normal distribution with zero mean and the correlation matrix specified above. We do the same and independently for $Y^{(2)}$. Figure A.1 shows some of the observed sparse mosaics $\mathscr{S}_x$ for a single realization of the process, where

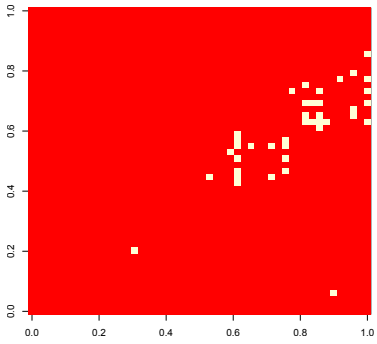
exceedances (highlighted by white squares) are clustered in the lower-triangular matrix $M$.



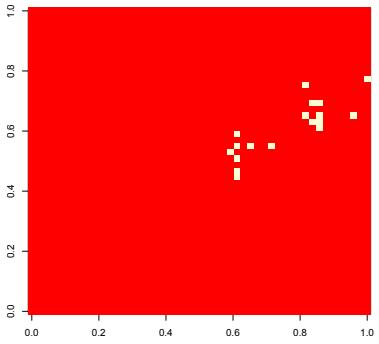
**(a)** $x = 1$



**(b)** $x = 1.5$



**(c)** $x = 2$



**(d)** $x = 2.3$

**Figure A.1.** Observed sparse mosaic $\mathscr{S}_x$ for several threshold values $x = 1, 1.5, 2, 2.3$. In red there are elements that have not exceed $x$, whereas in white there are the exceedances over $x$.

## A.5 Saddle point approximation for sum of exceedances test

We propose to use a saddle point approximation for the distribution of $(T_E^w(u) < x \mid H_0, N_u = k)$ when the $E[N_u]$ is low, in which case normal approximations might fail, with pdf

$$f_{T_E^{(w)}|H_0,N_u=k}(x) \approx \hat{f}_{T_E^{(w_u)}|H_0,N_u=k}(x) \equiv \frac{1}{(2\pi K_w''(\hat{t}))^{1/2}} e^{kK_w(\hat{t})-\hat{t}x}$$

where $K_w(\hat{t})$ is the cumulant moment generating function evaluated at point $t = \hat{t}$, $K_w''(\hat{t})$ is the second derivative of $K_w(t)$ at point $\hat{t}$ with first derivative $K_w'(\hat{t}) = x/k$. The saddle point approximation is suitable when there always exist $\hat{t}$ such that $K_w'(\hat{t}) = x/k$. This is proven to work well for $w = 0$ but might be undefined for high values $x/k$ when $w = 1$. The cdf of $T_E^{(w)} \mid N_u = k$ is found by numerical integration. Moments and cumulants generating functions are provided below.

**Scenario** $w_u = 0$**.** Take $y_t = \hat{d}_t^2$ so $y_t \sim \chi_1^2$. The moment generating function of $y$ is defined by

$$M_y = \frac{1}{(1-2t)^{1/2}} \frac{1-\Phi\{u(1-2t)\}}{1-\Phi(u)},$$

with cumulant generating functions

$$K_y = -\frac{1}{2}\log(1-2t) + \log[\Phi\{-u(1-2t)\}] - \log\{\Phi(-u)\},$$

$$K_y' = \frac{1}{1-2t} + \frac{\Phi'\{-u(1-2t)\}}{\Phi\{-u(1-2t)\}} = \frac{1}{1-2t} + u\frac{1}{(1-2t)^{1/2}} \frac{\varphi'\{u(1-2t)\}}{\Phi\{-u(1-2t)\}},$$

and

$$K_y'' = \frac{2}{(1-2t)^2} + \frac{u}{(1-2t)^{1/2}} \frac{\varphi'\{u(1-2t)\}}{\Phi\{-u(1-2t)\}} \left[ u^2 + \frac{1}{1-2t} - \frac{u}{(1-2t)^{1/2}} \frac{\varphi'\{u(1-2t)\}}{\Phi\{-u(1-2t)\}} \right].$$

**Scenario** $w_u = 1$**.** Take $(|\hat{d}_t| - u)^2/u^2 = x_t$, so $|\hat{d}_t| = ux_t^{1/2} + u$ and $f_x(x) = f_{|\hat{d}_t|}(s)\left|\frac{dx}{d\hat{d}_t}\right|$. The moment generating function of $x$ is defined by

$$M_x = 2ue^{\frac{1}{2}\frac{u^4}{u^2-2t}} \frac{1}{(u^2-2t)^{1/2}} \frac{1-\Phi\{u^2(u^2-2t)^{-1/2}\}}{1-\Phi(u)},$$

with cumulant generating functions

$$K_x = \log(2u) - \frac{1}{2}\frac{u^4}{u^2-2t} - \frac{1}{2}\log(u^2-2t) + \log(1-\Phi\{u^2(u^2-2t)^{-1/2}\}) - \log(\Phi(-u)),$$

$$K_y' = \frac{u^4}{(u^2-2t)^2} + \frac{1}{u^2-2t} - \frac{u^2}{(u^2-2t)^{3/2}} \frac{\varphi\{u^2(u^2-2t)^{-1/2}\}}{\Phi\{-u^2(u^2-2t)^{-1/2}\}},$$

and

$$K_x'' = \frac{u^2}{(u^2-2t)^{3/2}} \left[ -\frac{\varphi\{u^2(u^2-2t)^{-1/2}\}}{\Phi\{-u^2(u^2-2t)^{-1/2}\}} \frac{u^4}{(u^2-2t)^2} + \frac{\varphi^2\{u^2(u^2-2t)^{-1/2}\}}{\Phi^2\{-u^2(u^2-2t)^{-1/2}\}} \frac{u^2}{(u^2-2t)^{3/2}} \right].$$

## A.6 Threshold selection for sum of exceedances test

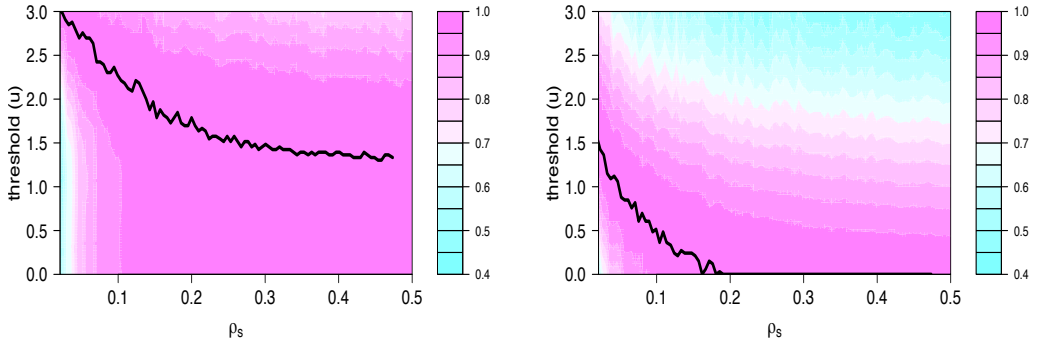
### A.6.1 Optimizing the asymptotic power

The threshold $u$ is key to find the test statistic that maximizes the power and its selection is the focus of attention of this section. Under notation in Theorem 3, take

$$f(\delta_t, s, u, n, m, w) = \frac{\sum_{t\in\mathscr{S}_d} \mu_{t_w}\eta_t - s\eta_0\mu_w - z_\alpha[m\eta_0\{(1-\eta_0)\mu_w^2 + \sigma_w^2\}]^{1/2}}{[\sum_{t\in\mathscr{S}_d} \eta_t\{(1-\eta_t)\mu_{t_w}^2 + \sigma_{t_w}^2\} + (m-s)\eta_0\{(1-\eta_0)\mu_w^2 + \sigma_w^2\}]^{1/2}}, \tag{A.5}$$

so the lower bound for the asymptotic power is $1-\exp(-f(\delta_t, s, u, n, m, w)^2/2)$, where $f(\delta_t, s, u, n, m, w)$ depends on parameters $n, m, w, u$ (known), and $s, \delta_t$ (unknown). Let $\rho_s = s/m$ be the proportion of

non-zero elements in $R_2 - R_1$. To show the influence that $\rho_s$ has in the asymptotic power, the function $f$, defined in eq. (A.5), is evaluated for several values of $\rho_s$, $u$, with fixed sizes $n = 100$, $m = 10000$ and generating values of $\delta_t$ from a Gamma$(a, b)$ distribution with parameters $a = 3$ and $b = 10$. In Figure A.2, the optimal threshold, defined by the value of $u$ that maximizes $f(\delta_t, s, u, n, m, w)$, is decreasing with $\rho_s$ for both $w = 0$ and $w = 1$.

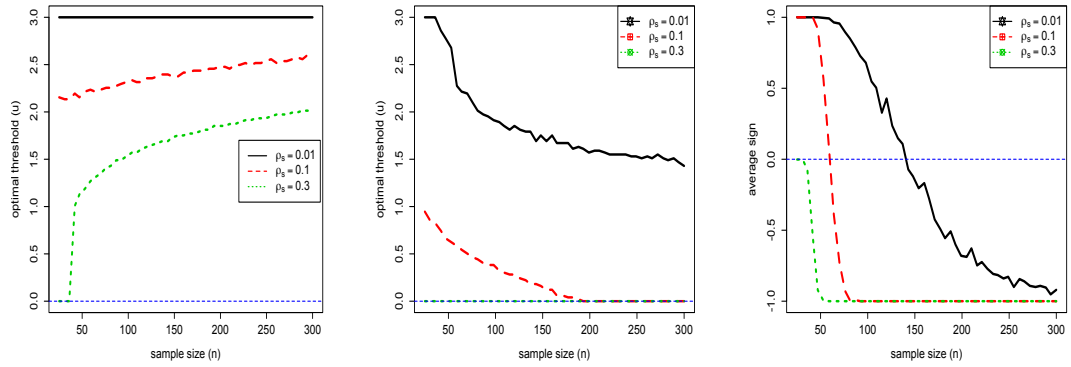



**(a)** $w = 0$ **(b)** $w = 1$

**Figure A.2.** Relative power of sum of exceedances test with respect to threshold ($u$) and proportion of non-zero correlation differences ($\rho_s$) for (a) $w = 0$ and (b) $w = 1$. The black line corresponds to the threshold with highest power.

Moreover, in panels (a) and (b) of Figure A.3, the optimal values for $u$ using a range of sample sizes and three different values for $\rho_s \in \{0.01, 0.1, 0.3\}$ are obtained. We also considered several dimension sizes, but their impact on the threshold selection was very low and for simplicity we only show the cases for $m = 1000$, which corresponds to $p \approx 43 - 44$. For $w = 0$, the optimal threshold increases with the sample size, whereas for $w = 1$, the optimal threshold decreases with the sample size. In panel (c) of Figure A.3, we show the lower bound of the power differences between $w = 0$ and $w = 1$. We consider the best power for both $w = 0$ and $w = 1$ and then we take the difference between the two. In the figure we present the average sign of such power differences over 1000 simulations for the set of parameters $(\delta_t : t \in \mathscr{S}_d)$. Only for small sample sizes ($n < 100$) and low $\rho_s$, $w = 1$ reaches better rates than $w = 0$. Otherwise, $w = 0$ dominates the asymptotic power.

As Figure A.2 and Figure A.3 show, the fraction of zero elements in $R_2 - R_1$ denoted by $\rho_s$ is essential to find the best threshold. We propose to find an estimator for $\rho_s$ using the q-values approach of Storey et al. (2015) where the input are approximated p-values $2(1 - \Phi(|\hat{d}_t|))$ for all $t \in M$. Even though testing if $\rho_s = 0$ is the same as our hypothesis testing of $R_1 = R_2$, here we only use this testing procedure to find a first crude estimation of $\rho_s$. This estimator is shown to be asymptotically unbiased with $n \to \infty$ but biased downwards when $\delta_t \sqrt{n-3}$ is small for all $t \in \mathscr{S}_d$ under mild dependence assumptions. However, in the application to biological data we generally have a relatively small $n$ and we have seen that the dependence process in $(\hat{d}_t : t \in M)$ can bias quite heavily the testing procedures in simulated




**(a)** $w = 0$ **(b)** $w = 1$ **(c)** Difference

**Figure A.3.** Optimal threshold in sum of exceedances test with respect to several values of the sample size for (a) $w = 0$ and (b) $w = 1$. In (c) is shown the average sign for the difference between best power using $w = 1$ and best power using $w = 0$ over 1000 simulated sets of differential correlation coefficients.

data (see Section 4.5).

The other unknown parameters are the Fisher transform correlation differences $\delta_t$, for all $t \in \mathscr{S}_d$. Below we propose a prior specification for $\delta_t$ to control the amount of elements that might be masked by the coefficients $\hat{d}_k$, $k \notin \mathscr{S}_d$, when $\delta_k = 0$. However, other distributions or other specifications for the hyper-parameters could be employed instead. We assume that $(\delta_t)$ are i.i.d. random variables with a known distribution, for instance we explore $\delta_t \sim \text{gamma}(a, b)$, with hyper-parameters satisfying $mode = (a-1)/b = Z_\alpha\,(n-3)^{-1/2}$, so the mode is assumed to be at the $1-\alpha$ quantile of the marginal distribution of $\hat{d}_t\,(n-3)^{-1/2}$ under $H_0$. Moreover, we set the variance of the prior, $var = a/b^2$, so $a$ and $b$ are fully defined.

We numerically integrate out $\delta_t$ from the function $f(\delta_t, s, u, n, m, w)$ defined in eq. (A.5) for threshold selection, i.e.,

$$\hat{u}^w = \arg\max_u \int_{\Omega_{\delta_t}} f(\delta_t, m\hat{\rho}_s, u, n, m, w)\, p(\delta_t)\, d\delta_t.$$

As final estimate we use the minimum between the optimal threshold and the $1-\alpha$ quantile of a standard normal distribution with default value $\alpha = 0.05$ in order to prevent cases with infinite thresholds.

### A.6.2 Threshold selection on simulated data

In Section A.6.1 we propose to select the threshold that maximizes the lower bound of the power by integrating out some of the unknown parameters. We use a q-values approach to estimate the important parameter $\rho_s$, which (as detailed in Section A.6.1) it defines the proportion of correlation coefficients that are different in the two matrices. In table A.1 we show the relative bias levels of the estimator for several sample sizes and true $\rho_s$. The bias is generally negative and it decreases with the

sample size for all $\rho_s$ levels.

**Table A.1.** Relative bias for estimator $\hat{\rho}_s$ given by $((\hat{\rho}_s - \rho_s)/\rho_s)$ using a q-values approximation. Several values for the sample size and true value of $\rho_s$ are employed.

| n | 50 | 100 | 200 | 500 |
|---|---|---|---|---|
| $\rho_s$ =0.01 | -0.227 | -0.168 | 0.040 | -0.022 |
| $\rho_s$ =0.08 | -0.350 | -0.209 | -0.104 | -0.047 |
| $\rho_s$ =0.16 | -0.358 | -0.213 | -0.113 | -0.037 |
| $\rho_s$ =0.23 | -0.364 | -0.218 | -0.110 | -0.038 |
| $\rho_s$ =0.30 | -0.364 | -0.215 | -0.114 | -0.044 |

In Table A.2 we compare the power (at 0.05 rejection level) of the sum of exceedances test using both estimated threshold and best threshold (found employing the true parameters) when $\delta_t$ deviates from the chosen prior distribution. For instance we generate $\delta_t$ values by a gamma(3,10) and then divide the resulting replicates by 2, 1, 2/3, and 1/2. In the table we show a measure of efficiency given by the ratio between the power of the test using the estimated threshold and the power for the optimal threshold. Only for small sample sizes (see $n = 25$), as we deviate from the $\delta_t$ prior distribution, the proportion of explained power decreases substantially.

**Table A.2.** Efficiency of the test defined as the explained power of the sum of exceedances test using estimated threshold against the sum of exceedances test using the optimal threshold.

| | n=25 | | | | n=50 | | | | n=100 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho_s$ | 0.01 | 0.08 | 0.15 | 0.23 | 0.01 | 0.08 | 0.15 | 0.23 | 0.01 | 0.08 | 0.15 | 0.23 |
| $\delta/2$ | 99 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| $\delta$ | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| $3\delta/2$ | 98 | 92 | 94 | 99 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| $2\delta$ | 94 | 86 | 79 | 85 | 98 | 97 | 99 | 100 | 100 | 100 | 100 | 100 |

## A.7 Asymptotic power

Let's first acknowledge the Mill's ratio which approximates $\Phi(-x) \doteq \frac{\varphi(x)}{x}$, where $\varphi(x) = e^{-\frac{1}{2}x^2}$, when $x$ is large. We recall that we use the set of variables $(\hat{d}_t : t \in M)$, with $m = \text{card}(M)$ such that $\mathscr{S}_d = \{t \in M : d_t \neq 0\}$ and $s = \text{Card}(\mathscr{S}_d)$ is the sparsity level. We assume that $|g(r_{2_t}) - g(r_{1_t})| = \delta_t$ for all $t \in \mathscr{S}_d$ with $d_t = \sqrt{n-3}\delta_t$. Moreover, we consider normality for the Fisher transform correlation differences such that for all $t \in \mathscr{S}_d$, $\hat{d}_t \sim N(\delta_t, (n-3)^{-1})$ and for all $t \notin \mathscr{S}_d$, $\hat{d}_t \sim N(0, (n-3)^{-1})$.

The power of the test is given by the probability of rejecting the null hypothesis when the $H_1$ is true. Hence, the objective is to find the test that provides the maximum power. For all tests $(q = s, m, e)$, we define a rejecting level $t_{q,\alpha}$ such that we reject the null hypothesis when the observed test statistic is larger than $t_{q,\alpha}$ at significance level $\alpha$.

### A.7.1 Asymptotic power of the average of squares test

Here we assume that the test statistic $T_S$ defined in eq. 4.6 of the main paper is well approximated by a normal distribution under both $H_0$ and $H_1$. We define $\mu_{H_0}$ and $\sigma^2_{H_0}$ as the expected value and

variance of $T_S$ when $H_0$ holds. Moreover, $\mu_{H_1}$ and $\sigma^2_{H_1}$ are the correspondent expected value and variance of $T_S$ when $H_1$ holds. The power of the average of squares test is

$$\Pr(T_S \geq t_{S,\alpha} \mid H_1) \doteq \Pr\left( Z \geq \frac{\mu_{H_1} - t_{S,\alpha}}{\sqrt{\sigma^2_{H_1}}} \right), \tag{A.6}$$

approximated using the Mill's ratio, with rejecting level given by $t_{S,\alpha} = \mu_{H_0} + z_\alpha \sqrt{\sigma^2_{H_0}}$.

Denote $\delta_0^2 = \sum_{t \in \mathscr{S}_d} \delta_t^2$ and recall that $\bar{\gamma}_2 = 2(m^2-m)^{-1} \sum_{t<h} \operatorname{cov}(\hat{d}_t^2, \hat{d}_h^2 \mid H_0)$. Under $H_0$, the parameters $\mu_{H_0} \doteq 1$ and $\sigma^2_{H_0} \doteq \frac{2}{m}\{1 + (m-1)\bar{\gamma}_2/2\}$. The expected value of $T_S$ under $H_1$ is found by a weighted average $\mu_{H_1} = (m-s)\mu_0/m + s\mu_1/m$ with $\mu_0 = \mathbf{E}[\hat{d}_t^2 \mid t \notin \mathscr{S}_d] \doteq 1$ and $\mu_1 = \operatorname{var}[\hat{d}_t \mid t \in \mathscr{S}_d] + \mathbf{E}[\hat{d}_t \mid t \in \mathscr{S}_d]^2 \doteq 1 + d_t^2$. Similarly, the parameter $\sigma^2_{H_1}$ can be found by the variance of a weighted average, so $\sigma^2_{H_1} = 2/m(1 + 2s(n-3)\delta_0^2/m + (m-1)\bar{\gamma}_2'/2)$ where $\bar{\gamma}_2' = 2(m^2-m)^{-1} \sum_{t<h} \operatorname{cov}(\hat{d}_t^2, \hat{d}_h^2 \mid H_1)$. Note that $\bar{\gamma}_2'$ is different to $\bar{\gamma}_2$ as it depends on the values $(d_t, t \in \mathscr{S}_d)$. Plugging in the expressions for $t_{S,\alpha}$, $\mu_{H_1}$ and $\sigma^2_{H_1}$ in (A.6), we obtain the stated expression for the power.

### A.7.2 Asymptotic power of the extreme value test

We assume $(\hat{d}_t) \sim MVN$, $t \in M$ under both $H_0$ and $H_1$. Hence, the maximum $T_M = \max_{t \in M} |\hat{d}_t|$, in the limit, is well represented by a Gumbel distribution. We further define the parameters $\mu_t = \mathbf{E}[\hat{d}_t \mid t \in \mathscr{S}_d]$, $\sigma_t^2 = \operatorname{var}[\hat{d}_t \mid t \in \mathscr{S}_d]$ with $|\mu_t|$ being sufficiently large. Assume independence on the sequence $(\hat{d}_t)$, the power of the extreme value test is defined by

$$\begin{aligned}
\Pr(T_M \geq t_{M,\alpha} \mid H_1) &= 1 - \Pr(|d_t| < t_{M,\alpha},\ \forall\, t) \geq 1 - \Pr(|d_t| < t_{M,\alpha} :\ t \in \mathscr{S}_d) \\
&= 1 - \Pr\left( \frac{-t_{M,\alpha} - \mu_t}{\sigma_t} < Z_t < \frac{t_{M,\alpha} - \mu_t}{\sigma_t},\ t \in \mathscr{S}_d \right) \\
&\geq 1 - \Pr\left( Z_t < \frac{t_{M,\alpha} - |\mu_t|}{\sigma_t},\ t \in \mathscr{S}_d \right),
\end{aligned}$$

where $Z_t = (|d_t| - \mu_t)/\sigma_t$. The rejecting level $t_{M,\alpha}$ is found using the quantile function of the Gumbel distribution that in the limit ascertains that

$$Q_G(\alpha) \doteq (2\log 2m)^{1/2} - \frac{\log\log 2m + \log(4\pi \log_2 2)}{2(2\log 2m)^{1/2}} - \frac{\log(-\log(\alpha))}{(2\log 2m)^{1/2}}.$$

We use the main term of the expression to find $Q_G(\alpha)$ such that

$$t_{M,\alpha} = (2\log 2m)^{1/2} - \frac{\log(-\log(\alpha))}{(2\log 2m)^{1/2}} > Q_G(\alpha).$$

For the expected value of the test statistic under $H_1$ we use $|\mu_t| \doteq \delta_t \sqrt{n-3}$, and for the variance we approximate $\sigma_t^2 \doteq \operatorname{var}(\hat{d}_t) \doteq 1$, for all $t \in \mathscr{S}_d$.

If $s = |\mathscr{S}_d| \to \infty$ and the conditions of the Gumbel approximation described in Section A.3 hold (namely that the maximum correlation between pairs of $d_t$, $t \in \mathscr{S}_d$, is bounded above by a constant

strictly less than 1), we have

$$\begin{aligned}\Pr(T_M \geq t_{M,\alpha} \mid H_1) &\geq 1 - \Pr\left(Z_t < \frac{t_{M,\alpha} - \min_{t\in\mathscr{S}_d}|\mu_t|}{\sigma_t},\ t \in \mathscr{S}_d\right) \\ &\geq 1 - \exp\{-\exp\{-(2\log 2s)^{1/2}[(n-3)^{1/2}\min_{t\in\mathscr{S}_d}\delta_t - (2\log 2m)^{1/2} + (2\log 2s)^{1/2}]\}\} \\ &\approx 1 - \exp\{-\exp\{-(2\log 2s)^{1/2}[(n-3)^{1/2}\min_{t\in\mathscr{S}_d}\delta_t - (2\log 2m)^{1/2}]\}\}.\end{aligned}$$

If $s = |\mathscr{S}_d|$ is a constant, then, using the Mill's ratio to approximate the normal probabilities,

$$\begin{aligned}\Pr(T_M \geq t_{M,\alpha} \mid H_1) &\geq 1 - \Pr\left(Z_t < \frac{t_{M,\alpha} - |\mu_t|}{\sigma_t},\ t \in \mathscr{S}_d\right) \geq 1 - \min_{t\in\mathscr{S}_d}\Pr\left(Z_t < \frac{t_{M,\alpha} - |\mu_t|}{\sigma_t}\right) \\ &\geq 1 - \min_{t\in\mathscr{S}_d}\exp\left[-\frac{1}{2}\left\{(n-3)^{1/2}\delta_t - \left((2\log 2m)^{1/2} - \frac{\log(-\log(\alpha))}{(2\log 2m)^{1/2}}\right)\right\}^2\right] \\ &\approx 1 - \min_{t\in\mathscr{S}_d}\exp\left[-\frac{1}{2}\left\{(n-3)^{1/2}\delta_t - (2\log 2m)^{1/2}\right\}^2\right].\end{aligned}$$

### A.7.3 Asymptotic power of the exceedances test

We set an arbitrary large threshold $u$, such that we define set $\mathscr{S}_u = \{t \in M : |\hat{d}_t| > u\}$. We define the probabilities $\eta_0 = \Pr(t \in \mathscr{S}_u \mid t \notin \mathscr{S}_d)$ and $\eta_t = \Pr(t \in \mathscr{S}_u \mid t \in \mathscr{S}_d)$ as well as the standard normal distribution density function at quantile $u$ which we denote by $\varphi(u)$. Under both $H_0$ and $H_1$, we approximate the test statistic $T_E^{(w)}$ described in eq. (4.8) by a normal distribution. We define $\mu_{H_0}(m,w)$ and $\sigma^2_{H_0}(m,w)$ as the expected value and variance of $T_E^{(w)}$ when $H_0$ holds. Moreover, $\mu_{H_1}(m,w)$ and $\sigma^2_{H_1}(m,w)$ are the correspondent expected value and variance of $T_E^{(w)}$ when $H_1$ holds. To find both $\mu_{H_1}(m,w)$ and $\sigma^2_{H_1}(m,w)$, we redefine the measures in eq.(4.20) by assuming that the expected value of $\hat{d}_t$ can be different from zero for some $t \in M$:

$$\begin{aligned}\gamma^{(H1,w)}_{u_{tj}} &= \mathrm{cov}((|\hat{d}_t| - uw)^2, (|\hat{d}_j| - uw)^2 \mid \hat{d}_t^2 > u, \hat{d}_j^2 > u), \\ \eta_t &= \Pr(|\hat{d}_t| > u), \\ \phi^{H1}_{tj} &= \Pr(\hat{d}_t^2 > u^2, \hat{d}_j^2 > u^2),\end{aligned}$$

The power is described by

$$\Pr(T_E^{(w)} \geq t^{(w)}_{E,\alpha} \mid H_1) \doteq \Pr\left(Z \geq \frac{\mu_{H_1}(m,w) - t^{(w)}_{E,\alpha}}{\sqrt{\sigma^2_{H_1}(m,w)}}\right),$$

where $\mu^{(w)}_{H_1} = (m-s)\eta_0\mu_w + \sum_{t\in\mathscr{S}_d}\eta_t\mu_{t_w}$, rejecting level $t^{(w)}_{E,\alpha} = \mu^{(w)}_{H_0} + z_\alpha\sqrt{\sigma^2_{H_0}(m,w)}$, and

$$\sigma^2_{H_1}(m,w) = \sum_{t\in\mathscr{S}_d}\eta_t\{(1-\eta_t)\,\mu^2_{t_w} + \sigma^2_{t_w}\} + (m-s)\eta_0\{(1-\eta_0)\,\mu^2_w + \sigma^2_w\} + C_w,$$

where $C_w = \sum_{t,h\in M, t\neq h}(\gamma_{u_{th}}^{(H1,w)} + \mu_{t_w}\mu_{h_w})\phi_{th}^{H1} - \eta_t\mu_{t_w}\eta_{h_w}\mu_{h_w}$ is different from zero if elements in $\hat{D}^2$ are dependent. Let $\mu_{H_0}(m,w) = \mu(m,w)$ and $\sigma^2_{H_0}(m,w) = \sigma^2(m,w)$ defined by eq. (4.21). The lower bound for the asymptotic power of sum of exceedances test, with $w = \{0,1\}$, is

$$\Pr(T_E^{(w)} \geq t_{E,\alpha}^{(w)} \mid H_1) \geq 1 - \exp\left\{-\frac{1}{2}\left(\frac{\sum_{t\in\mathscr{S}_d}\mu_{t_w}\eta_t - s\eta_0\mu_w - z_\alpha\sigma_{H_0}(m,w)}{\sigma_{H_1}(m,w)}\right)^2\right\}.$$

Let $\mathscr{S}_{du} = \{t \in M, |d_t| \gg u\}$ with $s_u = |\mathscr{S}_{du}|$. For $w = 0$, when $(n,m,u) \to \infty$, under weak independence, i.e., $C_w \ll \sigma^2_{H_1}(m,w)$, the asymptotic power leading terms ares

$$\frac{\sum_{t\in\mathscr{S}_{du}} d_t^2 - B_0(s\eta_0^{1/2} + z_\alpha(2m)^{1/2})}{\sqrt{\sum_{t\in\mathscr{S}_{du}} d_t^2 + mB_0^2}},$$

where $B_0 = u^2\eta_0^{1/2}$. Let $\delta_{00}^2 = s_u^{-1}\sum_{t\in\mathscr{S}_{du}} d_t^2$, asymptotic recovery condition is

$$\delta_{00}^2 \gg \frac{u^2}{n}\frac{\max(1, s\eta_0, (2m\eta_0)^{1/2})}{s_u},$$

If $s_u = k\max(1, s\eta_0, (2m\eta_0)^{1/2})$, for any positive integer $k$, and $d_t^2/u^2 \to \infty$, for any $t \in \mathscr{S}_{du}$, $\Pr(T_E^{(0)} \geq t_{E,\alpha}^{(0)} \mid H_1) \to 1$.

Similarly for $w = 1$, when $(n,m,u) \to \infty$, $\mu_1 \approx 2/(u^2+1)$ and $\sigma_1^2 \approx 4/(u^2+1)^2$ (these rates can be found using L'Hospital rule), and similar weak independence conditions, the asymptotic power leading terms are

$$\frac{\sum_{t\in\mathscr{S}_{du}} |d_t| - u^2 - B_1(s_u\eta_0^{1/2} + z_\alpha(2m)^{1/2})}{\sqrt{\sum_{t\in\mathscr{S}_{du}} |d_t| - u^2 + 2mB_1^2}},$$

where $B_1 = 2\eta_0/(u^2+1)$. Let $\delta_{01}^2 = s_u^{-1}\sum_{t\in\mathscr{S}_{du}}(|d_t| - u)^2$, asymptotic recovery condition is

$$\delta_{01}^2 \gg 2/(u^2+1)\frac{\max(1, s\eta_0, (2m\eta_0)^{1/2})}{s_{du}},$$

If $s_u = k\max(1, s\eta_0, (2m\eta_0)^{1/2})$, for any positive integer $k$, and $d_t^2/u^2 \to \infty$, for any $t \in \mathscr{S}_{du}$, $\Pr(T_E^{(1)} \geq t_{E,\alpha}^{(1)} \mid H_1) \to 1$.

# Appendix B

# Proofs and supplementary material of joint estimation methods

## B.1 Approximating error rates for tuning parameter selection

Expressions in Lemma 6.1 are found as follows. Under assumption 6.7, $\Pr(Q_{ij} - Z_{ij} > \lambda_2 v_{ij}) = 1 - \Phi(\lambda_2/(\sqrt{2}\sigma))$. In the non-differentially connected event, $B_0$, where $|Q_{ij} - Z_{ij}| \leq v_{ij}\lambda_2$, since we assume that $\sigma^2_{Q_{ij}} = \sigma^2_{Z_{ij}} = \sigma^2$, then $\text{cov}(Q_{ij} + Z_{ij}, Q_{ij} - Z_{ij}) = 0$. Assumption 6.7 in the main paper implies that

$$0.5(Q_{ij} + Z_{ij}) \mid (|Q_{ij} - Z_{ij}| \leq v_{ij}\lambda_2) \sim N(0, \sigma^2(1+\psi_{ij})/2),$$

and so $\Pr(|0.5(Q_{ij} + Z_{ij})| > \lambda_1 \mid |Q_{ij} - Z_{ij}| \leq v_{ij}\lambda_2) = 2[1 - \Phi(\sqrt{2}\lambda^{(1)}_{ij}(1+\psi_{ij})^{-1/2}/\sigma)]$.

The relationship between $Q_{ij}$ and $Z_{ij}$ can be expressed by a linear model

$$Z_{ij} = Q_{ij}\psi_{ij} + \epsilon_{ij}, \quad \text{where } Q_{ij} \sim N(0, \sigma^2) \text{ and } \epsilon_{ij} \sim N(0, \sigma^2(1-\psi^2_{ij})).$$

Hence, for any $a < b$, $c < d \leq b + v_{ij}\lambda_2$, $\Pr(Q_{ij} \in [c,d] \,\&\, Z_{ij} \in [a,b] \,\&\, Q_{ij} - Z_{ij} > v_{ij}\lambda_2)$ can be expressed in terms of $Q_{ij}$ and $\epsilon_{ij}$ by $\Pr(Q_{ij} \in [c,d] \,\&\, \epsilon_{ij} \in [a - Q_{ij}\psi_{ij}, Q_{ij}(1-\psi_{ij}) - \lambda_2 v_{ij}])$. Since $Q_{ij}$ and $\epsilon_{ij}$ are independent, then

$$\Pr(Q_{ij} \in [c,d] \,\&\, Z_{ij} \in [a,b] \,\&\, Q_{ij} - Z_{ij} > v_{ij}\lambda_2) =$$

$$\Pr(Q_{ij} \in [c,d] \,\&\, \epsilon_{ij} \in [a - Q_{ij}\psi_{ij}, Q_{ij}(1-\psi_{ij}) - \lambda_2 v_{ij}]) =$$

$$\int_c^d \sigma^{-1}\varphi(x/\sigma)\left[\Phi\left(\frac{x(1-\psi_{ij}) - \lambda_2(1-\psi_{ij})^{1/2}}{\sigma(1-\psi^2_{ij})^{1/2}}\right) - \Phi\left(\frac{a - x\psi_{ij}}{\sigma(1-\psi^2_{ij})^{1/2}}\right)\right] dx =$$

$$\int_c^d \sigma^{-1}\varphi(x/\sigma)\left[\Phi\left(\frac{x(1-\psi_{ij})^{1/2} - \lambda_2}{\sigma(1+\psi_{ij})^{1/2}}\right) - \Phi\left(\frac{a - x\psi_{ij}}{\sigma(1-\psi^2_{ij})^{1/2}}\right)\right] dx.$$

For any $A=[c,d]$ and $a=-\infty$ and $b=+\infty$, Lemma 1 implies that

$$\Pr(Q_{ij}\in A \& Q_{ij}-Z_{ij}>v_{ij}\lambda_2)=\int_{x\in A}\sigma^{-1}\varphi(x/\sigma)\Phi\left(\frac{x(1-\psi_{ij})^{1/2}-\lambda_2}{\sigma\,(1+\psi_{ij})^{1/2}}\right)dx,$$
$$\Pr(Q_{ij}\in A \& Q_{ij}-Z_{ij}<-v_{ij}\lambda_2)=\int_{x\in A}\sigma^{-1}\varphi(x/\sigma)\Phi\left(\frac{-x(1-\psi_{ij})^{1/2}-\lambda_2}{\sigma\,(1+\psi_{ij})^{1/2}}\right)dx.$$

Expression for $I_\sigma(\lambda^{(1)}_{ij},\psi_{ij},\lambda_2)$ in eq. (6.8) can be derived as follows:

$$\begin{aligned}
I_\sigma(\lambda^{(1)}_{ij},\psi_{ij},\lambda_2) &= \Pr(|\hat{A}^{''(t)}_{1_{ij}}|\le\lambda^{(1)}_{ij}\,\&\,|\hat{A}^{''(t)}_{2_{ij}}|\le\lambda^{(1)}_{ij}\mid \hat{A}^{''(t)}_{1_{ij}}-\hat{A}^{''(t)}_{2_{ij}}>v_{ij}\lambda_2)\\
&= \Pr(|\hat{A}^{''(t)}_{1_{ij}}|\le\lambda^{(1)}_{ij}\,\&\,|\hat{A}^{''(t)}_{2_{ij}}|\le\lambda^{(1)}_{ij}\mid \hat{A}^{''(t)}_{1_{ij}}-\hat{A}^{''(t)}_{2_{ij}}<-v_{ij}\lambda_2)\\
&= \Pr(|Q_{ij}+v_{ij}\lambda_2/2|\le\lambda^{(1)}_{ij}\,\&\,|Z_{ij}-v_{ij}\lambda_2/2|\le\lambda^{(1)}_{ij}\mid Q_{ij}-Z_{ij}>v_{ij}\lambda_2)\\
&= \int_{-\lambda^{(1)}_{ij}-v_{ij}\lambda_2/2}^{\lambda^{(1)}_{ij}-v_{ij}\lambda_2/2}\sigma^{-1}\varphi(x/\sigma)\left[\Phi\left(\frac{x(1-\psi_{ij})-\lambda_2 v_{ij}}{\sigma\,(1-\psi^2_{ij})^{1/2}}\right)-\Phi\left(\frac{\lambda_2 v_{ij}/2-\lambda^{(1)}_{ij}-x\psi_{ij}}{\sigma\,(1-\psi^2_{ij})^{1/2}}\right)\right]dx.
\end{aligned}$$

In Figure B.1 we present the values of $\lambda^{(1)}_{ij}$ obtained as function of $\psi_{ij}$ (see eq. 6.10) when $\sigma=1$, $\alpha_1=0.1$ and $\alpha_2=0.05$. This distinguishes between events in $B_0$ and events in $B_1$, which recall are defined by $B_0=\{(i,j)\in S_0,|Q_{ij}-Z_{ij}|>\lambda_2 v_{ij}\}$ and $B_1=\{(i,j)\in S_0,|Q_{ij}-Z_{ij}|\le\lambda_2 v_{ij}\}$. In our data, we use estimated values for $\psi_{ij}$, which are found to range between $-0.1$ and 0.4. Note that in that range of values, $\lambda^{(1)}_{ij}$ can be approximated well by a linear function of $\psi_{ij}$. Moreover, as expected, i.e., the variance of $0.5(Q_{ij}+Z_{ij})$ is smaller than the variance of $Q_{ij}+\lambda_2 v_{ij}/2$, then $\lambda_{1_\sigma}(\alpha_1,B_1,\psi_{ij})\ge\lambda_{1_\sigma}(\alpha_1,B_0,\psi_{ij})$ for any $\psi_{ij}\in(-1,1)$.

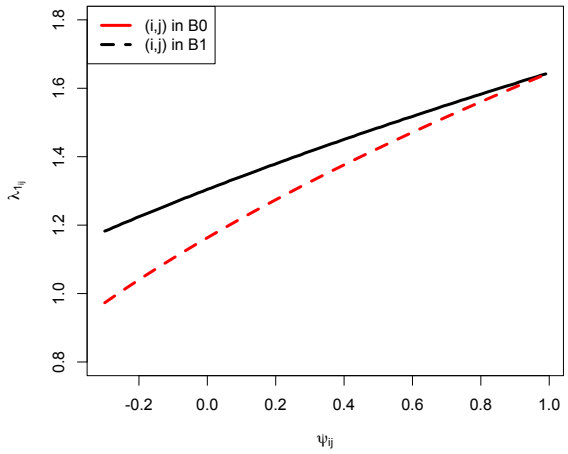



**Figure B.1.** Obtained values for $\lambda_{ij}$ as function of $\psi_{ij}$. In red it is considered $\lambda_{1_\sigma}(\alpha_1,B_0,\psi_{ij})$ whereas in black it is considered $\lambda_{1_\sigma}(\alpha_1,B_1,\psi_{ij})$.

## B.2 Joint estimation of regression coefficient matrices with linear dependent residuals

In order to find a joint estimation of regression coefficient matrix $\beta = [\beta^{(1)}, \beta^{(2)}]$, we consider an optimization problem that minimizes the standardized least square errors, i.e.,

$$(\hat{\beta}, \hat{\Omega}_e)^{\Lambda_1,\Lambda_2}_{WFRL} = \arg\min_{\beta,\Omega_e} \left[ \frac{1}{2n} \sum_{l=1,2} \text{tr}\left\{ (Y^{(l)} - X^{(l)}\beta^{(l)})^\intercal (Y^{(l)} - X^{(l)}\beta^{(l)})\, \Omega_e^{(l)} \right\} + P_{\Lambda_1,\Lambda_2}(\beta, \Omega_e) \right], \quad \text{(B.1)}$$

where $\Omega_e = [\Omega_e^{(1)}, \Omega_e^{(2)}]$ refers to the errors conditional dependence structure and

$$P_{\Lambda_1,\Lambda_2}(\beta) = ||\Lambda_1 \circ \beta^{(1)}||_1 + ||\Lambda_1 \circ \beta^{(2)}||_1 + ||\Lambda_2 \circ (\beta^{(2)} - \beta^{(1)})||_1 + ||\Lambda_1 \circ \Omega_e^{(1)}||_1 + ||\Lambda_1 \circ \Omega_e^{(2)}||_1 + ||\Lambda_2 \circ (\Omega_e^{(2)} - \Omega_e^{(1)})||_1. \quad \text{(B.2)}$$

We simplify the notation be assuming that the same tuning parameters $\Lambda_1$ and $\Lambda_2$ are applied to $\beta$ and $\Omega_e$. Nevertheless, different penalization parameters for the two type of conditional dependence structures could be employed instead with no major changes in the solution.

The optimization problem in eq. (B.1) is only convex if either $\beta$ or $\Omega_e$ is known. Let $\hat{\beta}_0$ be an initial estimate for $\beta$, which could be found by WFRL (Section 6.3.2). A common strategy is finding $\hat{\Omega}_{e_t}$ and $\hat{\beta}_t$ iteratively, for $t = 1, \ldots, T$, fixing the other to the solution on the current iteration until convergence. A solution for $\hat{\Omega}_e$ can be obtained by weighted fused graphical lasso (Section 6.2) applied to $q$-dimensional residual vectors $[Y^{(1)} - X^{(1)}\hat{\beta}^{(1)}]$ and $[Y^{(2)} - X^{(2)}\hat{\beta}^{(2)}]$. Besides, following approaches described in Chapter 6, $\hat{\beta}$ is found by optimizing the Lagrangian formulation of expression (B.1)

$$L_\rho = P_{\Lambda_1,\lambda_2 V}(Z^{(1)}, Z^{(2)}) + \left[ \frac{1}{2n} \sum_{l=1,2} \text{tr}\left\{ (Y^{(l)} - X^{(l)}\beta^{(l)})^\intercal (Y^{(l)} - X^{(l)}\beta^{(l)})\, \Omega_e^{(l)} \right\} + \frac{\rho}{2} ||\beta^{(l)} - Z^{(l)} + U^{(l)}||_F^2 \right].$$

using the ADMM-type algorithm (Boyd, 2010) described in Algorithm 10. Here, $U^{(l)}$ are the dual variables, $Z^{(l)}$ corresponds to $\beta^{(l)}$, for $l = \{1, 2\}$, and $\rho$ is a positive constant that is used as a regularization parameter with default value equal to 1.

The main difference with respect to Algorithm 9, where $\Omega_e$ was not contemplated, is in step 3 of the algorithm. Let $X$ be any $X^{(l)}$, $Y$ be any $Y^{(l)}$, $\Omega_e$ be any $\Omega_e^{(l)}$ and $\beta$ be any $\beta^{(l)}$, for $l = 1, 2$. Solving by $\beta$ eq. (B.3), the following solution can be obtained:

$$X^\intercal X\beta\Omega_e - X^\intercal Y\beta\Omega_e + \rho\beta - \rho Z + \rho U = 0$$

$$X^\intercal X\beta + \rho\beta\Omega_e^{-1} = X^\intercal Y - \rho(Z + U)\Omega_e^{-1}$$

$$\text{vec}(\beta) = [(1_q \otimes \Sigma_X) + \rho(\Omega_R^{-1} \otimes 1_p)]^{-1} \text{vec}(X^\intercal Y - \rho(Z + U)\Omega_e^{-1}).$$

[Going from line 2 to 3 can be done following eq.(2) of Jameson (1968)]. Hence, we consider as dense

**Algorithm 10** Weighted Fused Regression Lasso

1: Input: $\Lambda_1, \lambda_2, \rho, V, \Omega_e$.
2: Initialization: $t = 0$, $U_t^{(l)} = 0$ and $Z_t^{(l)} = 0$, for $l = 1,2$, repeat 3-5 until convergence.
3: Find $\hat{\beta}_t^{(l)}, \hat{\beta}_t^{(2)}$ by solving the minimization problem:

$$[\hat{\beta}_t^{(1)}, \hat{\beta}_t^{(2)}] = \arg\min_{\beta^{(1)},\beta^{(2)}} \left[ \frac{1}{2n} \sum_{l=1,2} \mathrm{tr}\left\{ (Y^{(l)} - X^{(l)}\beta^{(l)})^\intercal (Y^{(l)} - X^{(l)}\beta^{(l)}) \Omega_e^{(l)} \right\} + \frac{\rho}{2} ||\beta^{(l)} - Z^{(l)} + U^{(l)}||_F^2 \right]. \tag{B.3}$$

4: Find $Z_t^{(1)}, Z_t^{(2)}$ such that

$$\sum_{l=1,2} \frac{\rho}{2} ||\hat{\beta}_t^{(l)} - Z_l + U_t^{(l)}||_F^2 + P_{\Lambda_1,\lambda_2 V}(Z_t^{(l)}, Z_t^{(l)})$$

is minimized.
5: Set $t = t+1$. Update dual variables $U_l^{(t)} = U_{t-1}^{(l)} + \hat{\beta}_t^{(l)} - Z_t^{(l)}$, for $l = 1,2$. Stop if convergence.
6: Output: $\hat{\beta}^{(1)} = \hat{Z}_{t-1}^{(1)}$, $\hat{\beta}^{(2)} = \hat{Z}_{t-1}^{(2)}$ and $\hat{\beta}^{(d)} = \hat{Z}_{t-1}^{(2)} - \hat{Z}_{t-1}^{(1)}$.

estimator for $\beta$ in step 3,

$$vec(\hat{\beta}) = [(1_q \otimes \Sigma_X) + \rho(\Omega_R^{-1} \otimes 1_p)]^{-1} \mathrm{vec}(X^\intercal Y - \rho(Z+U)\Omega_e^{-1}). \tag{B.4}$$

Thresholding operations in step 4 of the Algorithm are the same as described in Section 6.3.2.

## B.3 Hypothesis testing for the number of differential edges

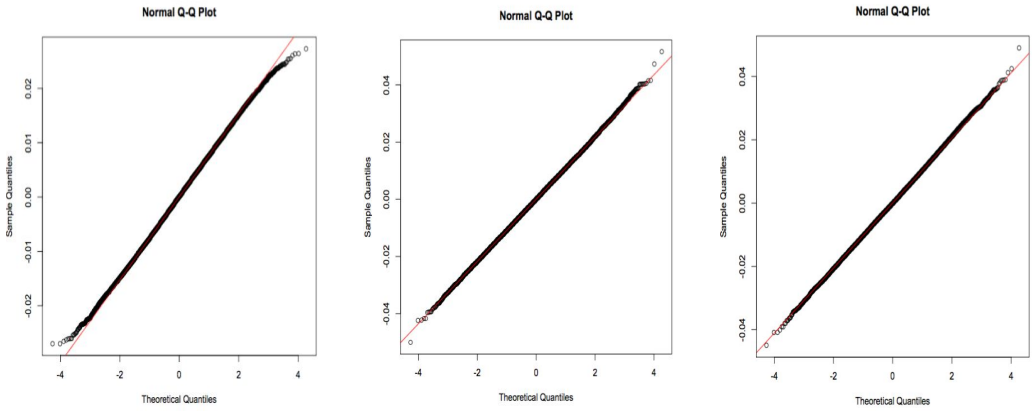
Differential network estimators incorporate the variability of the two individual estimated networks and tend to be much more uncertain than the underlying estimated common network. Define the set $S_d = \{(i,j), i < j : \Omega_{ij}^{(1)} - \Omega_{ij}^{(2)} \neq 0\}$ with $|S_d| = \mathrm{card}(S_d)$. To check if there is any differential edges, we propose to test hypothesis $\mathrm{H}_0$: $|S_d| = 0$ against $\mathrm{H}_1$: $|S_d| > 0$ by employing the test statistic $T_d = \sum_{i<j} I(\hat{\Omega}_{ij}^{(1)} \neq \hat{\Omega}_{ij}^{(2)})$.

A permuted samples based approach is used to assess the uncertainty in the number of estimated differential edges under the hypothesis of equality in the two precision matrices $\mathrm{H}_0$. Data are permuted as follows to ensure that the dependence structure between datasets is maintained: $[(Z_1^{\pi_1}, \ldots, Z_n^{\pi_n}), (Z_1^{\bar{\pi}_1}, \ldots, Z_n^{\bar{\pi}_n})]$ where $\bar{\pi}_i = 1 - \pi_i$ and $Z_i^{\pi_i} = Y_i^{(1)}$ if $\pi_i = 0$ and $Z_i^{\pi_i} = Y_i^{(2)}$ if $\pi_i = 1$, with $\Pr(\pi_i = 1) = 0.5$. Given the new permuted data, a weighted fused graphical lasso estimate is found by solving eq. (6.2) using the same combination for $\lambda$'s as for the original estimate, and the number of estimated differential edges is recorded. By repeating this permutation and estimation process $B$ times with $(T_d^{(b)})_{b=1}^B$ being the obtained test statistics, the p-value of the test is computed, i.e., p-val $= B^{-1} \sum_{b=1}^B I(T_d^{(b)} \geq T_d)$. Similar tests are applied to real data in Section 6.5.4 to assess the evidence of "healthy only" edges or "unhealthy only" edges. These would consider test statistics $T_{d_T} = \sum_{i<j} I(\hat{\Omega}_{ij}^{(1)} \neq 0 \,\&\, \hat{\Omega}_{ij}^{(2)} = 0)$ or $T_{d_H} = \sum_{i<j} I(\hat{\Omega}_{ij}^{(1)} = 0 \,\&\, \hat{\Omega}_{ij}^{(2)} \neq 0)$ instead of $T_d$.

The same procedure is done for the regression coefficient matrices, i.e., change $\Omega_{ij}^{(1)}$ and $\Omega_{ij}^{(2)}$ for $\beta_{ij}^{(1)}$ and $\beta_{ij}^{(1)}$ above and solve eq. (6.20) for new permuted data.

## B.4 Normality assumption for estimated precision matrix elements

In section 6.2.2 we discuss a way to select the regularization parameters $\lambda$'s based on setting their correspondent error rates $\alpha_1$ and $\alpha_2$. We make an assumption of normality for the estimated precision matrix elements in each iteration of the joint estimation algorithm. Figure B.2 shows some of the obtained normality qqplots employing all the estimated coefficients $[\hat{A}^{'(0)}_{1_{ij}}, \hat{A}^{'(0)}_{2_{ij}}]$ using simulated datasets with $p = 300$ and $n = 25, 100, 200$. This represents a general observed behavior in many tested datasets. We shall see that for sufficiently large $n$ the Gaussian assumption is well justified.




**(a)** n=25 **(b)** n=100 **(c)** n=200

**Figure B.2.** `qqnorm` plots for several examples of estimated precision matrices coefficients. We distinguish among three sample sizes $n$.

## B.5 Showing fairness of WFGL in simulated data

Assuming that differential edges can occur with same probability independently of the values $[\psi_{ij}]$, WFGL produces, even for small $n$, a less biased procedure than FGL in which edges with high correlation have similar chances to be recovered as edges with low correlation. We illustrate this using the model defined in Section 6.5.1 with dimension $p = 300$ and several sample sizes. We divide pairs of variables $(i, j)$ in two groups: $L = \{(i, j) : \psi_{ij} < 0.1\}$ and $U = \{(i, j) : \psi_{ij} > 0.1\}$. Consider partial estimates in the ADMM algorithm 8 $\hat{\Omega}^{(0)}_m$ for $m = 1, 2$. For all pairs $(i, j)$, we compute $h_{ij} = v^{-1}_{ij} |(\hat{\Omega}^{(0)}_2)_{ij} - (\hat{\Omega}^{(0)}_1)_{ij}|$ using $v_{ij} = 1$ (Indep.) as well as $v_{ij} = (1 - \hat{\psi}_{ij})^{1/2}$ (paired) with $[\hat{\psi}_{ij}]$ estimated by the Reg-based-sim method discussed in Section 6.2.2. Denote the ranks of $h_{ij}$ by $k_{ij}$ in the decreasing order ($k_{ij} = 1$ for the largest $h_{ij}$ and $k_{ij} = p(p-1)/2$ for the smallest $h_{ij}$). In Figure B.3 we show the differences between the average ranks in the two groups, i.e., $|L|^{-1} \sum_{(i,j)\in L} k_{ij} - |U|^{-1} \sum_{(i,j)\in U} k_{ij}$. We can see that the independent method encourages recovery of differential edges with small $\psi_{ij}$ (seen in the plot by large negative rank differences) and this bias is corrected by the dependent data adjustment, which for relatively large sample size gives very similar ranks in the two groups.

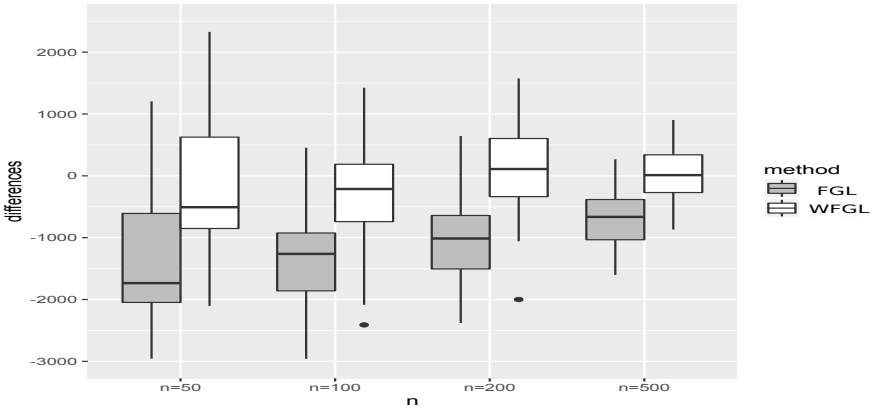



**Figure B.3.** Differences between average ranks of $\Omega_d$ among large $\psi_{ij}$ and small $\psi_{ij}$ over 50 simulations in the first iteration of the ADMM algorithm correcting for weights $v_{ij} = 1$ (Indep.) and weights $v_{ij} = (1 - \hat{\psi}_{ij})^{1/2}$ (paired).

## B.6 Estimation of weights in simulated data for WFGL

The performance of the two estimators (Reg-based and Reg-based-sim) described in Section 6.2.3 is analyzed using simulation. We calculate the mean square error of $[\hat{\psi}_{ij}]$ against $[\psi_{ij}]$ as well as the correlation $\text{cor}(\psi, \hat{\psi})$. We compare the Reg-based and Reg-based-sim estimator results with $[\hat{\psi}_{ij} = 0]$ (which assumes independence between samples). The values $\psi_{ij}$ are approximated by the sample correlation using 5,000 i.i.d. Monte Carlo replicates of the theoretical model. Table B.1 provides the average ranks (average MSE) for the mean square error and Table B.2 gives the average ranks (average correlation) for the correlation levels. Rank = 1 is assigned to the best estimator and Rank = 3 is given to the worst estimator.

For very small sample sizes ($n = 25$), the estimators' MSE are very large, and can even find worse results than assuming independence. However, for all other investigated sample sizes, the Reg-based and its simplified version find the lowest MSE. Correlation-wise, the two proposed estimators give large positive correlations consistently for large $p/n$ ratios.

**Table B.1.** Ranks and average for the sum of MSE.

| n | 25 | 50 | 150 | 300 | 500 |
|---|---|---|---|---|---|
| dimension p=50 | | | | | |
| Reg-based | 2.04 (0.86) | 1.83 (0.42) | 1.52 (0.16) | **1.40 (0.09)** | **1.28 (0.06)** |
| Reg-based-sim | **1.04 (0.86)** | **1.17 (0.41)** | **1.48 (0.16)** | 1.60 (0.09) | 1.72 (0.06) |
| Independence | 2.91 (1.24) | 3.00 (1.30) | 3.00 (1.38) | 3.0 (1.41) | 3.00 (1.42) |
| dimension p=170 | | | | | |
| Reg-based | 2.74 (0.74) | 2 (0.33) | **1.17 (0.13)** | **1.01 (0.08)** | **1.06 (0.06)** |
| Reg-based-sim | 1.74 (0.74) | **1.00 (0.33)** | 1.83 (0.13) | 1.99 (0.08) | 1.94 (0.06) |
| Independence | **1.52 (0.77)** | 3.00 (0.70) | 3.00 (0.65) | 3.00 (0.67) | 3.00 (0.69) |
| dimension p=290 | | | | | |
| Reg-based | 2.30 (0.74) | 2.00 (0.33) | **1.00 (0.13)** | **1.00 (0.08)** | **1.00 (0.06)** |
| Reg-based-sim | **1.30 (0.74)** | **1.00 (0.33)** | 2.00 (0.13) | 2.00 (0.08) | 2.00 (0.06) |
| Independence | 2.40 (0.77) | 3.00 (0.70) | 3.00 (0.65) | 3.00 (0.67) | 3.00 (0.69) |
| dimension p=500 | | | | | |
| Reg-based | 2.80 (0.72) | 2.00 (0.32) | **1.00 (0.13)** | **1.00 (0.09)** | **1.00 (0.06)** |
| Reg-based-sim | 1.79 (0.72) | **1.00 (0.32)** | 2.00 (0.13) | 2.00 (0.09) | 2.00 (0.06) |
| Independence | **1.42 (0.68)** | 3.00 (0.64) | 3.00 (0.58) | 3.00 (0.59) | 3.00 (0.61) |

**Table B.2.** Ranks and average for the average correlations between approximated and estimated $\psi$.

| n | 25 | 50 | 150 | 300 | 500 |
|---|---|---|---|---|---|
| | | | dimension p=50 | | |
| Reg-based | **1.16 (0.63)** | **1.23 (0.80)** | **1.5 (0.93)** | 1.67 (0.96) | 1.57 (0.97) |
| Reg-based-sim | 1.84 (0.63) | 1.77 (0.80) | **1.5 (0.93)** | **1.33 (0.96)** | **1.43 (0.97)** |
| Independence | 3.00 (0) | 3.00 (0) | 3.00 (0) | 3.00 (0) | 3.00 (0) |
| | | | dimension p=170 | | |
| Reg-based | **1.04 (0.57)** | **1.09 (0.69)** | **1.34 (0.86)** | 1.55 (0.92) | 1.94 (0.95) |
| Reg-based-sim | 1.96 (0.57) | 1.90 (0.69) | 1.66 (0.86) | **1.45 (0.92)** | **1.05 (0.95)** |
| Independence | 3.00 (0) | 3.00 (0) | 3.00 (0) | 3.00 (0) | 3.00 (0) |
| | | | dimension p=290 | | |
| Reg-based | **1.07 (0.62)** | **1.03 (0.72)** | 1.51 (0.86) | 1.90 (0.92) | 1.94 (0.95) |
| Reg-based-sim | 1.92 (0.62) | 1.97 (0.72) | **1.49 (0.86)** | **1.10 (0.92)** | **1.05 (0.95)** |
| Independence | 3.00 (0) | 3.00 (0) | 3.00 (0) | 3.00 (0) | 3.00 (0) |
| | | | dimension p=500 | | |
| Reg-based | **1.28 (0.61)** | **1.08 (0.73)** | **1.06 (0.85)** | **1.16 (0.91)** | **1.47 (0.94)** |
| Reg-based-sim | 1.72 (0.61) | 1.92 (0.73) | 1.94 (0.85) | 1.84 (0.91) | 1.53 (0.94) |
| Independence | 3.00 (0) | 3.00 (0) | 3.00 (0) | 3.00 (0) | 3.00 (0) |

## B.7 Estimation of weights in simulated data for WFRL

In Section 6.3.3 we propose a way to estimate the correlation between same coefficients in $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ for similarity penalization. Here we analyze the performance of the estimator using simulations. We calculate the mean square error of $[\hat{\theta}_{ij}]$ against $[\theta_{ij}]$ as well as the correlation $\text{cor}(\theta, \hat{\theta})$. We compare the performance of our proposed estimator against setting $\hat{\theta}_{ij} = 0$, for all pairs $i, j$, which assumes independence between samples. The values $\theta_{ij}$ are approximated by the sample correlation using 5,000 i.i.d. Monte Carlo replicates of the theoretical model. In Table B.3 we present the average mean square error and also the average correlation over 100 instances of simulations. For all investigated sample sizes (even for $n = 25$) the proposed estimator finds the lowest MSE. Besides, the proposed estimator gives large positive correlations consistently for large $p/n$ ratios.

**Table B.3.** Average mean square errors (average correlation) over 100 instances between approximated $\psi$ (using 5,000 i.i.d. Monte Carlo replicates of the true model) and estimated $\psi$ (proposed -found following Section 6.3.3 approach). These statistics are also obtained by considering $\hat{\psi} = 0$ (Independence).

| n | 25 | 50 | 75 | 100 |
|---|---|---|---|---|
| | | dimension p=120 | | |
| Proposed | **0.0042** (0.86) | **0.0023** (0.90) | **0.0018** (0.93) | **0.0015** (0.94) |
| Independence | 0.0181 (-) | 0.0145 (-) | 0.0142 (-) | 0.0135 (-) |
| | | dimension p=200 | | |
| Proposed | **0.0050** (0.86) | **0.0025** (0.91) | **0.0019** (0.93) | **0.0016** (0.94) |
| Independence | 0.021 (-) | 0.0173 (-) | 0.0150 (-) | 0.0151 (-) |