UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXV

Two-sample inference for graphical models

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Davide Risso

Co-supervisori: Prof. M. Chiogna, Dr. V. Djordjilović e Prof. M. Drton

Dottoranda: Erika Banzato

February 22, 2023

Abstract

One of the main goals of transcriptomics is the identification of genes that show a significant difference between two conditions. Biological processes underlying the basic functions of a cell involve complex interactions between genes, that can be represented through a graph where genes and their connections are, respectively, nodes and edges. Differential network analysis is a statistical tool to investigate how the network changes between two conditions.

The main research objective of this thesis is to improve some aspects of differential network analysis, accounting for the nature of the data and the network structure. To this aim, we propose a correction for the likelihood ratio test, with application to two-sample inference in decomposable Gaussian graphical models. We prove that the adjusted statistic leads to valid inference at different dimensionality regimes. Moreover, we study the necessary and sufficient conditions for the existence of the estimate in the Kullback-Leibler importance estimation procedure, with the aim of guiding the practitioner on the use of this tool in real data analyses and posing the basis for future works in the context of count data.

Sommario

Uno degli obiettivi principali della trascrittomica è l'identificazione di geni differenzialmente espressi in due condizioni. I processi biologici che regolano le funzioni di base delle cellule sono caratterizzati da complesse interazioni tra i geni. Tali processi possono essere rappresentati tramite dei grafi, dove i geni e le loro connessioni sono, rispettivamente, i nodi e gli archi. L'analisi delle differenze tra reti è un metodo statistico per studiare come cambia la rete tra diverse condizioni.

Il principale obiettivo di questa tesi è quello di migliorare alcuni aspetti dell'analisi delle differenze tra reti, tenendo in considerazione la natura dei dati e la struttura della rete coinvolta. A tal fine proponiamo una correzione per il test basato sul rapporto di verosimiglianza, che può essere applicato in problemi a due campioni in modelli grafici Gaussiani scomponibili. Dimostriamo che la statistica aggiustata porta a un'inferenza valida a diversi regimi di dimensionalità. Inoltre, studiamo le condizioni necessarie e sufficienti per l'esistenza della stima nella *Kullback-Leibler importance estimation procedure* (KLIEP), con l'obiettivo di dare direttive nell'uso di questo strumento nelle analisi di dati reali e ponendo le basi per futuri lavori nel contesto dei dati di conteggio.

Acknowledgements

The Ph.D. has been a rich and challenging period of academic and personal growth. Such an intense journey would not have been the same without the constant support of the people around me.

First of all, I would like to thank my supervisor, Prof. Davide Risso, for his guidance during the last two years. I am thankful for his advice, encouragement and opportunities he gave me. I am also extremely grateful for his patience and kindness, especially during the last period. A special thanks goes also to Prof. Monica Chiogna and Dr. Vera Djordjilović for being always present and for bringing new ideas. It was a pleasure working with you all.

I would like to thank my supervisor at TUM, Prof. Mathias Drton, for hosting me in his research group. Working with him was an honor and definitely challenged me to see things from another perspective. A special mention goes also to Dr. Irem Portakal for her valuable help.

To my colleagues of the XXXV cycle: Beppe, Pietro, Caizhu, Nicolas, Cristian, Marco, Touqeer and Fariborz. It was fun, and a pleasure, to share this journey with you. To Kim, for her help and friendship. To my lovely colleagues in Munich for making me feel welcome and part of the group from the first day.

I am also extremely grateful to my extra-university friends. To Cezara for being such an amazing friend even from miles away. To Marta, Vale, Marco, Giacomo, Lucy, Marika and Ile, for still being here after such a long time. To my Italian friends in Munich, I feel very lucky to have known you.

A special thanks to my parents, Nicoletta and Giovanni, for their support and infinite patience. To my brother, Marco, now it is your turn, good luck! To my grandparents, Italo and Antonietta, for having always believed in me.

My deepest gratitude goes to Alberto. Thank you for the constant presence, love, trust and support you give me every day. Thank you for just being you, this milestone is also yours.

And finally to Oliviero, the cutest.

Contents

List of Figures

List of Tables

Introduction

Overview

The interest in the two-sample problems in graphical models is increasing over the last years, especially in the biological field. In transcriptomic analysis, the main goal is to identify genes that show a significant difference between two conditions. Technically, we are in the context of a two-sample problem, where the aim of the analysis is to highlight differences in distribution between two conditions. For example, one might be interested in checking which genes are differently expressed comparing cells from a control group with those from patients with a certain disease, or one might be interested in comparing different stages of the same disease, such as cancer. In this type of analysis, common practice is to assume that genes are independent in order to perform, for each gene, a simple test on the means. However, this approach does not fully address the complexity of the data, since the difference in one gene can cause other genes to change, leading to a dysregulation in the entire gene network.

While it is of interest to identify groups of (marginally) altered genes, it is even more interesting and challenging to identify genes that are the source of this difference, conditionally on the others. In this setting, graphical models can be a useful tool to tackle the problem. The network representing the connections between genes can be described as a graph, where each node represents a gene and the connections are the edges. This makes it possible to take into account the conditional independence relations that occur in the network, leading to a more informative analysis.

When dealing with differential network analysis, we can distinguish two cases: the network is considered known and the network is unknown. In the genomic framework, known networks are typically in the form of biological pathways, which represent a series of interactions among molecules in a cell that leads to a change in a cell state or process, or to the creation of a new molecular product. Biological pathways do not always work properly and when the network is dysregulated, the result can be a disease. Hence, it is of interest to highlight *where* in that particular network the dysregulation occurred in the first place. Examples of pathways repositories are the KEGG database (Kanehisa and Goto, 2000), Reactome (Croft *et al.*, 2010) and WikiPathways (Pico *et al.*, 2008). However, pathways stored in these databases are a collection of manually drawn networks, based on biological knowledge, and may not be completely representative of the network underlying the biological phenomenon of the dataset under study. Moreover, one might be interested only in the differences in connections or distribution of the genes, without referring to a particular network, but considering it unknown. In this latter case, differences can be inferred directly.

A further reason for the complexity of the problem is the nature of the data. The previous technology for measuring gene expression was based on microarrays (Irizarry *et al.*, 2003a). These data were collected on a continuous scale and were usually assumed to be normally distributed on a logarithmic scale. Moreover, many microarray studies were performed on small cohorts of samples, resulting in the sample size n (typically less than 100) being much smaller than the dimension p (typically in the order of 10^4). Hence, most of the available statistical tools have been developed relying on the normality assumption of the data and assuming p much larger than n. Nowadays, however, RNA sequencing technology allows the analysis of gene expression at single-cell resolution (Wang *et al.*, 2009), yielding high dimensional count data, in which both the dimension p and the sample size n may be large. Furthermore, these data are characterized by skewed distributions with high variance and over-abundance of zeros. Methods specifically developed for Gaussian data are no longer appropriate and need to be adapted to the nature of the variables under study.

In this thesis, we contribute to on the field, by accounting for these two reasons of complexity: the underlying structure of the problem and the count nature of the data. The outline of the work is as follows. In Chapter 1 we present the motivating problem and a review of the literature on the predominant approach. Chapter 2 briefly reviews graphical models, with particular emphasis on Gaussian and Poisson graphical models. In Chapter 3 we propose our first contribution, a correction for the likelihood ratio test, with application to Gaussian graphical models. Chapter 4 deals with the Kullback-Leibler importance estimation procedure (KLIEP). In particular, we present a study on the existence of the estimate.Finally, in the last chapter, we draw the main conclusions from this work and possible directions for future research.

Main contributions of the thesis

The main contributions of the thesis can be summarized as follows.

1. Definition of a new correction factor that improves the asymptotic approximation of the likelihood ratio test in a two-sample problem to the chi-square (Banzato *et al.*, 2022). The proposed multiplicative correction factor is defined to be the ratio between the degrees of freedom of the asymptotic chi-square approximation and an approximation of the expected value of the likelihood ratio test statistic, under the null hypothesis. The expected value takes the form of a function of the dimension p and the sample size n, as defined in Jiang and Qi (2015). We study the *phase transition boundary* (He *et al.*, 2021), which characterizes the approximation accuracy by establishing the necessary and sufficient condition for the chi-square approximation to hold when p increases with n. We prove that the phase transition boundary of the corrected statistic, T_n, is equal to 1 so that the chi-square approximation holds in all situations in which $p/n \to 0$. We study the properties of T_n through a simulation study and we compared its performances to other competitors.

2. Extension of the new correction to the two-sample problem in decomposable Gaussian graphical models. Here, using the properties studied in

Djordjilović and Chiogna (2022), the problem of testing equality of two distributions, Markov with respect to a decomposable graph, can be broken up into testing equality of lower dimensional Gaussian distributions, at clique level. According to the structure of the graph, these lower dimensional problems can have different dimensions, and so it is crucial to rely on a test statistic that guarantees a good finite sample accuracy even in extreme cases, where p is close to n. We study the performance of the corrected statistic through a simulation study and show that the application of this correction gives good results in terms of control of the type I error rate at the clique level. Moreover, the computational time is improved over the original method of Salviato *et al.* (2019), *sourceSet*. This is because the use of the aforementioned correction allows overcoming the need for a permutation-based approach to control for both the multiplicity of the tests and the failure in the approximation of the limiting distribution, drastically reducing the time for computations.

3. Study of the properties of the existence of the KLIEP estimator (Liu *et al.*, 2017). This estimator is based on the direct estimation of the differences in parameters, using the ratio of the distributions in the two conditions. In particular, this is achieved by minimizing a specific loss function. We show that, for the minimum to be achieved, the sufficient statistic from one sample, say X, needs to be inside the convex hull generated by the rows of the sufficient statistics of the other sample, say Y. If the latter is not satisfied in the sample, the loss function is not strictly convex and the minimum cannot be achieved. This result gives an important indication of the possibility of applying the KLIEP algorithm to the sample at hand and opens possibilities for future developments.

Chapter 1

Motivating problem

1.1 Gene expression

The process that turns the information encoded in a gene into a function is called *gene expression*. Through the transcription of RNA molecules, gene expression controls when and where RNA molecules and proteins are made and determines how much of those products are made. This process changes considerably under different conditions and cell types, indeed the set of proteins synthesized by the cell is important in determining its phenotype. The RNA and protein products of many genes serve to regulate the expression of other genes, leading to a complex network of interactions. The protein abundance is not easily measurable and the measure of RNA content in a cell can be viewed as a proxy for this quantity.

The differential expression analysis has the primary goal of determining which genes are expressed at different levels between conditions, e.g. cancer vs. normal tissues or different stages of the same disease. These genes can offer biological insight into the processes affected by the conditions of interest. The idea is to better understand what characterizes a certain disease or which genes are involved in the development of the disease, in order to use that knowledge to develop new drugs or treatments that specifically act on that gene.

1.1.1 Microarray and RNA sequencing

A microarray is a laboratory tool used to detect the expression of thousands of genes at the same time and it has been the technology of choice for high-throughput gene expression studies in the early 2000's. DNA microarrays are microscope slides, printed with thousands of spots in defined positions, with each spot containing a known DNA sequence or gene. The microarray is then scanned and the expression of each gene is measured as spot intensity after hybridization. Typically, mRNA molecules for the analysis are collected from both an experimental sample and a reference sample, supplying a *relative* measure of expression. Microarray experiments are usually performed on small cohorts of samples, resulting in sample sizes (typically less than 100) much smaller than the dimension (the number of genes is usually in the order of 10^4).

In recent years, RNA sequencing (RNAseq) technology replaced microarrays as the assay of choice for measuring genome-wide transcription levels (Nagalakshmi *et al.*, 2008; Wang *et al.*, 2009). This technology allows the measurement of gene expression not only at the bulk level but also at the single-cell level for millions of cells in a single study, making it possible to characterize and distinguish each cell at the transcriptome level. Gene expression is measured as read counts and single-cell transcriptome measurements present low signal-to-noise ratios, a high abundance of zeroes, and very skewed distributions. Being possible to study the genome at single-cell resolution provides a huge amount of data, where both the sample size and the dimension might be large.

1.1.2 Differential expression analysis

In many experiments, a statistical test is performed to identify genes significantly associated with the experimental conditions, clinical response, or other sample attributes. The easiest statistical approach to select genes differentially expressed between two groups is to apply a series of t-tests, assuming the independence of all genes. With the development of techniques, more sophisticated methods for differential expression analysis have emerged, such as *limma* (Smyth, 2004), *DESeq2* (Anders and Huber, 2010), and *edgeR* (Robinson *et al.*,

2010). The first one is based on linear models and was first developed for microarray data and further extended to RNAseq data (Ritchie *et al.*, 2015); the latter two methods were meant for RNA sequencing data and are based on generalized linear models, assuming a negative binomial distribution. However, all these methods are based on the assumption that genes are independent. The resulting list of significant genes may be large and gene set analysis is then used as a biological summary of results. This approach detects over-representation of gene sets among the list of significant genes and it is often performed using a χ^2 or Fisher's Exact test (Zeeberg *et al.*, 2003; Boyle *et al.*, 2004; Beissbarth and Speed, 2004), which rely critically upon the assumption that individual genes, and their associated test statistics, are independent. However, this assumption has been shown to be unrepresentative of the real problem and to bring misleading results (Goeman and Bühlmann, 2007). Since genes are highly connected to each other, the difference in one gene can cause others to change and the network complexity should be taken into account in the analysis.

1.2 Biological pathways

Technological advances in high throughput analysis give access to a vast amount of data that can help enlighten the mechanisms underlying the complex interplay of different genes. These connections are collected in the form of diagrams, called biological pathways. Biological pathways consist of a set of linked biological components interacting with each other over time to generate a specific biological effect or a change in a cell. There are many types of biological pathways, the most well-known being the ones involved in metabolism, gene regulation, and signal transduction. Metabolic pathways make possible the chemical reactions that occur in our bodies.

Learning biological pathways is a complex task, because most pathways do not have real boundaries, and might also work together to accomplish tasks, forming a biological network. Identifying which genes, proteins, and other molecules are involved in a biological pathway can provide clues about mechanisms that generate certain diseases. Pathway analysis, and in particular identifying which pathways are involved in a disease (even in each patient), can

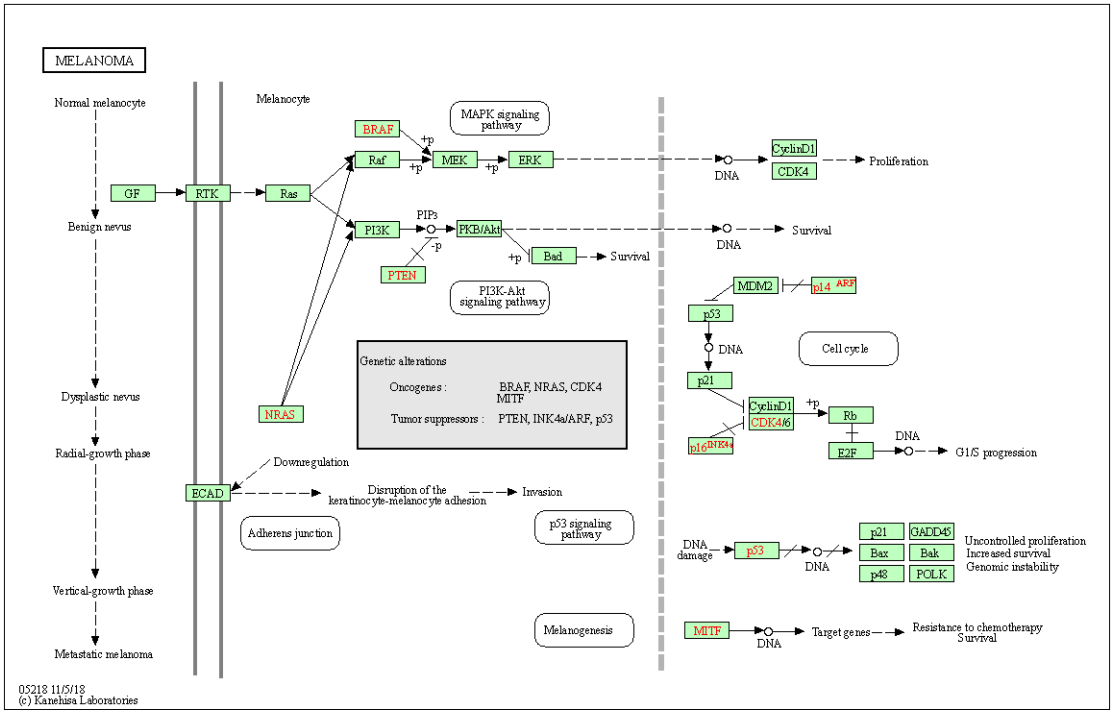

FIGURE 1.1: Melanoma pathway from the KEGG database.

have an enormous impact on the definition of more personalized strategies for diagnosing, treating, and preventing disease. This is the main reason why much emphasis is invested in the identification of *where* in a particular network the dysregulation occurred in the first place.

The most used pathway repositories are the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), Reactome (Croft *et al.*, 2010) and WikiPathways (Pico *et al.*, 2008). Figure 1.1 shows an example of a pathway from the KEGG repository.

1.2.1 Graphical representation

Friedman *et al.* (2000) introduced for the first time the idea of modeling gene networks using directed acyclic graphs (DAGs). Pathways, and gene networks in general, can be represented as graphs, where nodes and edges represent, respectively, the genes and the connections between them. A graphical model framework has the advantage of capturing the variability among variables in the

biological system, but it also allows learning and formulating new hypotheses on the relations between genes.

There is a rich literature that exploits graphs for pathway analysis, typically considering Gaussian graphical models and considering the network is known. See e.g. Rahnenführer *et al.* (2004); Draghici *et al.* (2007); Massa *et al.* (2010); Jacob *et al.* (2012); Grechkin *et al.* (2016) and Mukherjee *et al.* (2018) who adapted the latter to single-cell RNA-Seq data. We also mention the works of Salviato *et al.* (2019) and Djordjilović and Chiogna (2022), which will be further considered in Chapter 3.

Although pathways represent a useful tool for statistical analysis, they cannot be always assumed to be the best structure of a graphical model (Djordjilović, 2015). In fact, pathways represent the joint work of the scientific community and have been discovered through laboratory studies of cultured cells, bacteria, and other organisms. Thus, every graphical representation of a signaling pathway should be seen as a compromise between accuracy and complexity.

Chapter 2

Statistical background

2.1 Fundamentals of graphical models

This section is a review of key concepts and main terminology in graphical modeling. For a detailed treatment, we refer the interested reader to Lauritzen (1996) and Whittaker (1990).

Let $\mathcal{G} = (V, E)$ be a graph, where V is a finite set of vertices, also called nodes, and E is the set of edges. The set of edges, $E = \{(v, t) : v \neq t, (v, t) \in V \times V\}$, is a set of pairs of nodes, subset of $V \times V$. Edges can be *directed*, if exactly one of the edges $\{(v, t), (t, v)\}$ is in E and *undirected*, if both (v, t) and (t, v) are in E. If a graph $\mathcal{G}$ has only undirected edges is called *undirected* graph, whereas if it has only directed edges, the graph is said to be *directed*. If there is an edge between v and t, v and t are said to be *adjacent* or *neighbors*. The set of neighbors of a vertex v is denoted as $ne(v)$, for instance in Figure 2.1, $ne(1) = \{2, 3\}$. If there is an arrow from t pointing to v, t is said to be a *parent* of v and v a *child* of t. The set of parents of v is denoted by $pa(v)$ and the set of children of t as $ch(t)$, e.g in Figure 2.3 $pa(3) = \{1, 2\}$ and $ch(1) = ch(2) = \{3\}$.

If $A \subseteq V$ is a subset of the vertex set, it induces a subgraph $\mathcal{G}_A = (A, E_A)$, where the set of edges $E_A = E \cap A \times A$ is obtained from $\mathcal{G}$ by keeping only edges with both endpoints in A. An undirected graph is said to be *complete* if all vertices are joined by an edge. A subset is complete if it induces a complete subgraph. A complete subset that is maximal (with respect to $\subseteq$) is called a

clique. For instance, in Figure 2.1, we can recognize two cliques: $\{1, 2, 3\}$ and $\{3, 4, 5\}$.

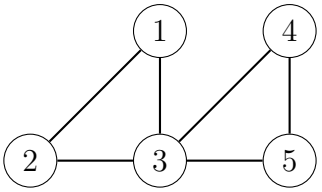

FIGURE 2.1: Example of an undirected graph with 5 nodes and 6 edges.

We say that a set of vertices S *separates* sets A and B in an undirected graph $\mathcal{G}$ if every path from a node in A to a node in B contains at least one node in S. In Figure 2.1, if $A = \{1, 2\}$, $B = \{4, 5\}$ and $C = \{3\}$, C separates A from B.

A *path* of length n from v to t is a sequence of distinct nodes $v = v_0, \ldots, v_n = t$, such that $(v_{i-1}, v_i) \in E$, for all $i = 1, \ldots, n$. A path can never cross itself and it can never go against the direction of the arrows. Paths can be defined as undirected, partially directed, and directed. An undirected path has all edges undirected, whereas if all edges are directed, we call it a directed path. A partially directed path is one that contains both directed and undirected edges. A *chain* of length n from v to t is a sequence of distinct nodes $v = v_0, \ldots, v_n = t$, such that $v_{i-1} \rightarrow v_i$ or $v_i \leftarrow v_{i-1}$ for all $i = 1, \ldots, n$. An *n-cycle* is a path of length n that begins and ends at the same point, such that $v = t$. The cycle is said to be directed if it contains an arrow.

Chain graphs contain both directed and undirected edges and can be seen as a generalization of both directed and undirected graphs. The vertex set V of the chain graph is partitioned into numbered subsets, forming the *dependence chain* $V = V_1 \cup \cdots \cup V_T$ such that all edges between nodes belonging to the same subset are undirected while all edges between different subsets are directed, pointing from the set with a lower number toward the set with a higher number. Such graphs are characterized by having no partially directed cycles. An undirected graph is a special case of a chain graph when there is a single chain component, while a directed acyclic graph (DAG) is a special case of a chain graph when all chain components consist of a single vertex. For a chain graph $\mathcal{G}$, we define its *moral* graph $\mathcal{G}_M$ as the undirected graph with the same vertex set but with

v and t adjacent in $\mathcal{G}_M$ if and only if either $v \to t$ or $t \to v$ or if there are z_1, z_2 in the same chain component such that $v \to z_1$ and $t \to z_2$.

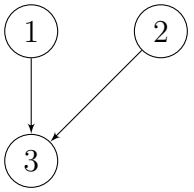

FIGURE 2.2: Example of a graph with V-structure.

In the special case of a DAG, a moral graph is used to find its equivalent undirected form. The moralization consists of first adding undirected edges between unmarried parents and then dropping directions by replacing arrows with undirected edges. As an example, the moralized version of the graph in Figure 2.3 consists of first marrying the nodes 1 and 2 and then replacing the arrows with undirected edges.

Definition 2.1. An undirected graph $\mathcal{G}$ is said to be *decomposable* if it is complete, or if there exists a proper decomposition $(A,\ B,\ C)$ into decomposable subgraphs $\mathcal{G}_{A\cup C}$ and $\mathcal{G}_{B\cup C}$.

Definition 2.2. A *triangulated* graph is an undirected graph with the property that every cycle of length $n \geq 4$ has a *chord*, which means two non-consecutive vertices that are neighbors.

Definition 2.3. Let $\mathcal{C} = (C_1, \ldots, C_k)$ be the set of cliques of the undirected graph $\mathcal{G}$. Let $J_j = C_1 \cup \cdots \cup C_j$, $R_j = B_j \setminus J_{j-1}$ and $S_j = J_{j-1} \cap B_j$. If for all $i > 1$ there is a $j < i$ such that $S_i \subseteq B_j$, the sequence is said to follow the *running intersection property*.

In this thesis, we focus on undirected graphs, without considering directionality, such that the presence of edges can be interpreted as a connection between the two vertices.

2.1.1 Conditional independence

In many statistical applications, graphs represent useful tools to describe interactions between variables. In fact, the set of conditional independence relations among a collection of random variables can be intuitively represented by connections among the set of vertices of a graph induced by a certain separation criterion. From the late '70s, the conditional independence of variables have started to be studied with the help of graphs, assigning a node to each variable, and using edges to encode the conditional dependencies (Speed, 1978; Knuiman, 1978; Pearl and Paz, 1987). This application of graphical methods gives rise to the so-called graphical models.

Definition 2.4. Let X, Y, Z be random variables with a joint distribution P. We say that the random variables X and Y are *conditionally independent* given the random variable Z and write $X \perp\!\!\!\perp Y|Z$ if and only if

$$\mathrm{P}(X \in A, Y \in B|Z) = \mathrm{P}(X \in A|Z)\,\mathrm{P}(Y \in B|Z), \tag{2.1}$$

for any A and B measurable in the sample space of X and Y, respectively.

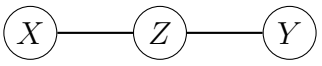

FIGURE 2.3: Example of conditional independence.

Equivalently, we can say that X and Y are conditionally independent given Z if and only if $\mathrm{P}(X \in A|Y, Z) = \mathrm{P}(X \in A|Z)$. This alternative definition has an intuitive interpretation: knowing the value of Z makes the distribution of X not further depending on Y. If Z is trivial we say that X is *independent* of Y and write $X \perp\!\!\!\perp Y$.

When X, Y, Z are discrete random variables, condition (2.1) simplifies as

$$\mathrm{P}(X = x, Y = y|Z = z) = \mathrm{P}(X = x|Z = z)\,\mathrm{P}(Y = y|Z = z),$$

where the equation holds for all z with $P(Z = z) > 0$. When X, Y, Z are continuous random variables, the condition (2.1) can be written as

$$f_{XY|Z}(x, y|z) = f_{X|Z}(x|z)\, f_{Y|Z}(y|z),$$

where equality holds almost surely. The definition of conditional independence can be extended to random vectors. Let A, B and C be three subsets of V. For discrete random vectors, we say that $\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B | \boldsymbol{X}_C$ if

$$\mathrm{P}(\boldsymbol{X}_A = \boldsymbol{x}_A, \boldsymbol{X}_B = \boldsymbol{x}_B | \boldsymbol{X}_C = \boldsymbol{x}_C) = \mathrm{P}(\boldsymbol{X}_A = \boldsymbol{x}_A | \boldsymbol{X}_C = \boldsymbol{x}_C)\, \mathrm{P}(\boldsymbol{X}_B = \boldsymbol{x}_B | \boldsymbol{X}_C = \boldsymbol{x}_C)$$

for any value of the realizations $\boldsymbol{x}_A$, $\boldsymbol{x}_B$, $\boldsymbol{x}_C$ of $\boldsymbol{X}_A, \boldsymbol{X}_B, \boldsymbol{X}_C$ respectively. For continuous random vectors, we have conditional independence if

$$f_{\boldsymbol{X}_A \boldsymbol{X}_B | \boldsymbol{X}_C}(\boldsymbol{x}_A, \boldsymbol{x}_B | \boldsymbol{x}_C) = f_{\boldsymbol{X}_A | \boldsymbol{X}_C}(\boldsymbol{x}_A | \boldsymbol{x}_C)\, f_{\boldsymbol{X}_B | \boldsymbol{X}_C}(\boldsymbol{x}_B | \boldsymbol{x}_C)$$

for all $\boldsymbol{x}_A, \boldsymbol{x}_B, \boldsymbol{x}_C$. The equation must hold almost surely with respect to P.

2.1.2 Markov properties on undirected graphs

In this section, we briefly review three Markov properties in the context of undirected graphs, see Lauritzen (1996) for a detailed presentation.

In what follows, we consider a $p-$dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ such that each random variable X_s corresponds to a node of the graph $\mathcal{G} = (V, E)$ with index set $V = \{1, 2, \ldots, p\}$. In an undirected graph $G = (V, E)$, an edge between two nodes, say v and t, is denoted by (v, t). Let $\boldsymbol{X} = \{X_i : i \in V\}$ be a random vector associated with the graph $\mathcal{G} = (V, E)$ and let $\boldsymbol{X}_A = \{X_j : j \in A \subset V\}$ be the random vector of the variables in $A \subset V$. A probability distribution of $\boldsymbol{X}$ is said to satisfy the

1. *pairwise Markov property* with respect to $\mathcal{G}$ if for any pair (v, t) of non-adjacent vertices
$$X_v \perp\!\!\!\perp X_t | \boldsymbol{X}_{V \setminus \{v,t\}},$$

2. *local Markov property* with respect to $\mathcal{G}$ if for any vertex $v \in V$
$$X_v \perp\!\!\!\perp \boldsymbol{X}_{V \setminus \{ne(v) \cup \{v\}\}} | \boldsymbol{X}_{ne(v)}$$

3. *global Markov property* with respect to $\mathcal{G}$ if for any triple (A, B, C) of disjoint subsets of V such that C separates A from B in $\mathcal{G}$

$$\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B | \boldsymbol{X}_C.$$

It can be shown that the global Markov property implies the local Markov property, which in turn implies the pairwise Markov property. Figure 2.4 shows an example of the Markov properties applied to a graph. From left to right: pairwise, local, and global Markov property. Let X_i, $i = 1, \ldots, 5$ be a set of random variables, the left panel shows the pairwise Markov property, where $X_1 \perp\!\!\!\perp X_5 | \{X_2, X_3, X_4\}$. In the middle panel, for the local Markov property, we have that $X_3 \perp\!\!\!\perp X_5 | \{X_2, X_4\}$ and $X_1 \perp\!\!\!\perp \{X_3, X_5\} | X_2$. The panel on the right shows the global Markov property, where $X_1 \perp\!\!\!\perp \{X_3, X_4\} | \{X_2, X_5\}$. The

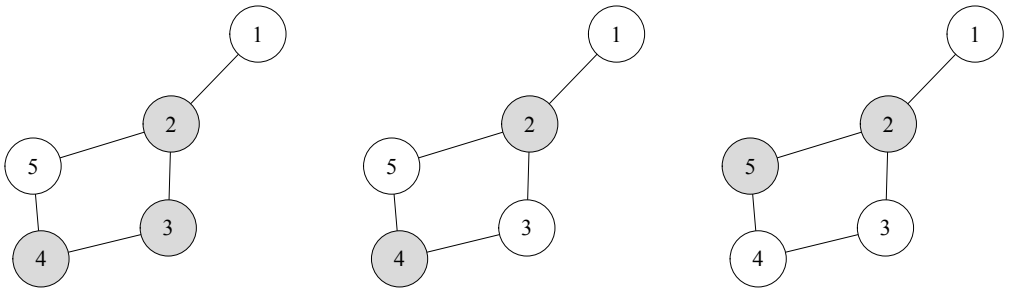

FIGURE 2.4: Example of the Markov properties applied to a graph. From left to right: pairwise, local, and global Markov property.

global Markov property gives a general criterion for deciding when two sets of variables, say A and B are conditionally independent given a third set C. Markov properties and conditional independence, are closely related to the factorization of the joint density. The latter in fact can be expressed as a product of clique-wise functions.

Definition 2.5. Let $\boldsymbol{X} = (X_1, \ldots, X_p)$ be a vector of random variables. A probability measure $\mathcal{P}$ on the sample space of $\boldsymbol{X}$, is said to *factorize* according

to $\mathcal{G} = (V, E)$ if for all complete subset $a \subseteq V$ there exists non-negative functions ψ_a that depend on $\boldsymbol{x}$ only through $\boldsymbol{x}_a$ and there exists a product measure $\mu = \otimes_{a \in V} \mu_a$ on the sample space of $\boldsymbol{X}$, such that $\boldsymbol{X}$ has density f with respect to μ where f has the form

$$f(\boldsymbol{x}) = \prod_{a \text{ complete}} \psi_a(\boldsymbol{x}). \tag{2.2}$$

The functions ψ_a are not uniquely determined and groups of functions ψ_a can be multiplied together or split up in different ways. Without loss of generality, one can assume that only cliques appear as sets a. Let $\mathcal{C}$ be the set of cliques of $\mathcal{G}$, (2.2) can be rewritten as

$$f(\boldsymbol{x}) = \prod_{c \in \mathcal{C}} \psi_c(\boldsymbol{x}).$$

It can be shown that for any undirected graph $\mathcal{G}$ and any probability distribution on the sample space of $\boldsymbol{X}$, it holds that if f factorizes according to $\mathcal{G}$ then it satisfies the global Markov property (and thus the local and pairwise Markov properties). If $\boldsymbol{X}$ has a positive and continuous density, all the Markov properties are equivalent.

Proposition 2.6. *The random vector $\boldsymbol{X}$ (or its density function) is decomposable if and only if its independence graph is triangulated. See Whittaker (1990).*

A benefit of these conditional models is that a graphical model can be understood as an exponential family distribution. The exponential family is specified by a vector of sufficient statistics, say $\boldsymbol{T}(\boldsymbol{x}) = \{T_1(\boldsymbol{x}), \ldots, T_m(\boldsymbol{x})\}$, the log-base measure $B(x)$ and the domain of the sample space, $\mathcal{D}$. The generic exponential family is defined as

$$\mathrm{P}_{EF}(\boldsymbol{x}|\boldsymbol{\eta}) = \exp\left\{ \boldsymbol{\eta}^T \boldsymbol{T}(\boldsymbol{x}) + B(\boldsymbol{x}) - A(\boldsymbol{\eta}) \right\},$$

where $\boldsymbol{\eta}$ is the vector of canonical parameters of the distribution and $A(\boldsymbol{\eta})$ is called the log normalization constant that normalizes the distribution over the

domain $\mathcal{D}$. By the Hammersley-Clifford theorem (Clifford, 1990), any distribution satisfying the global Markov property can be written as

$$\mathrm{P}(\boldsymbol{x}|\eta) = \exp\left\{\sum_{c\in\mathcal{C}} T_C(\boldsymbol{x}_C) - A(\eta)\right\}, \tag{2.3}$$

where $\mathcal{C}$ is the set of cliques of the undirected graph $\mathcal{G}$ and $T_C(\boldsymbol{x}_C)$ are the clique-wise sufficient statistics. A special case is a pairwise graphical model, where for a graph $\mathcal{G} = (V, E)$, $\mathcal{C}$ consists of merely V and E, with cliques $|C| = \{1, 2\}$, $\forall C \in \mathcal{C}$, so that we have

$$\mathrm{P}(\boldsymbol{x}|\eta) = \exp\left\{\sum_{i\in V} \eta_i T_i(x_i) + \sum_{(i,j)\in E} \eta_{ij} T_{ij}(x_i, x_j) - A(\eta)\right\}. \tag{2.4}$$

2.2 Gaussian graphical models

Gaussian graphical models are the undirected graphical models for the multivariate normal distribution. Gaussian distributions are probably the most known and used when analyzing network data, especially in the biological field, where microarray was the most used technique to collect gene expression data (Irizarry *et al.*, 2003a).

Given an undirected graph, $\mathcal{G}$, the Gaussian graphical model for the random vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ assumes that $\boldsymbol{X}$ follows a p-variate normal distribution under the conditional independence properties implied by the graph. The density is continuous and strictly positive, hence, the three Markov properties, as well as the factorization property, are all equivalent. Conditional independence relations implied by the graph $\mathcal{G}$ are easily represented by parameters of the multivariate normal distribution, through zero restrictions on the inverse of the covariance matrix. Let $S_{\mathcal{G}}^{+}$ be the set of all the $p \times p$ symmetric and positive definite matrices with null entries corresponding to missing edges in $\mathcal{G}$. The family of Gaussian graphical models can be defined as follows

$$\mathcal{M}_{\mathcal{G}} = \{\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{K}^{-1}) : \boldsymbol{\mu} \in \mathbb{R}^p,\, \mathbf{K}^{-1} \in S_{\mathcal{G}}^{+}\}.$$

The density of the multivariate normal distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\boldsymbol{\Sigma} = \mathbf{K}^{-1}$ is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{K}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\kappa_{ij}(x_i-\mu_i)(x_j-\mu_j)\right\},$$

where $\mathbf{K}$ is called precision matrix with elements κ_{ij}. The entries in the concentration matrix $\mathbf{K}$ have a simple interpretation. The diagonal elements κ_{ii} are the reciprocal of the conditional variances given the remaining elements,

$$\kappa_{ii} = \mathrm{Var}(X_i|\boldsymbol{X}_{V\setminus\{i\}})^{-1},$$

whereas the off-diagonal values κ_{ij} represent the interactions between variables,

$$\mathrm{Cov}(X_i, X_j|\boldsymbol{X}_{V\setminus\{i,j\}}) = \frac{-\kappa_{ij}}{\kappa_{ii}\kappa_{jj}-\kappa_{ij}^2}.$$

From the latter, it follows that

$$\kappa_{ij} = 0 \iff X_i \perp\!\!\!\perp X_j|\boldsymbol{X}_{V\setminus\{i,j\}}.$$

Whenever one of κ_{ij} is zero, this represents a missing edge in the corresponding graph and the joint density factorizes into two components: one containing x_i and the other containing x_j. According to the factorization criterion, the two variables X_i and X_j are conditionally independent given the others. Thus, we can use a graph $\mathcal{G}$ based on the concentration matrix, also called *concentration graph* (Cox and Wermuth, 1996), to represent a multivariate Gaussian distribution. This fundamental relation is the basis of the Gaussian graphical models and follows from the interpretation of the concentration matrix. The pairwise Markov property for the random vector $\boldsymbol{X}$ with respect to $\mathcal{G}$ is satisfied if and only if $\kappa_{ij} = 0$ for all pairs (X_i, X_j) non-adjacent in $\mathcal{G}$. At the same time, also the global Markov property is satisfied since the distribution of a normal random variable is positive.

The class of multivariate normal density functions is closed with respect to the operations of marginalization and conditioning, in fact, both the marginal and the conditional density functions of the multivariate normal are themselves

multivariate normal. The conditional distribution of $X_i|\boldsymbol{X}_{V\setminus\{i\}}$ is univariate normal. When no conditional independence restrictions are assumed to hold, the model is called the *saturated model*. Let A and B be two partitions of the random vector $\boldsymbol{X}$, such that $\boldsymbol{X} = (\boldsymbol{X}_A, \boldsymbol{X}_B)$. If the partitioned vector has a normal distribution parameterized by mean vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_A, \boldsymbol{\mu}_B)$ and variance

$$\boldsymbol{\Sigma}_{A\cup B} = \begin{pmatrix} \boldsymbol{\Sigma}_A & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_B \end{pmatrix}$$

then the marginal distribution of $\boldsymbol{X}_A$ is normal with mean $\boldsymbol{\mu}_A$ and variance $\boldsymbol{\Sigma}_A$, and the conditional distribution of $\boldsymbol{X}_B$ given $\boldsymbol{X}_A = \boldsymbol{x}_A$ is normal with mean vector $\mathrm{E}_{B|A}(\boldsymbol{X}_B) = \boldsymbol{\mu}_B + \boldsymbol{\Sigma}_{BA}\boldsymbol{\Sigma}_A^{-1}(\boldsymbol{x}_A - \boldsymbol{\mu}_A)$ and variance $\boldsymbol{\Sigma}_{B|A}(\boldsymbol{X}_B) = \boldsymbol{\Sigma}_B - \boldsymbol{\Sigma}_{BA}\boldsymbol{\Sigma}_A^{-1}\boldsymbol{\Sigma}_{AB}$.

2.2.1 Decomposable models

When interaction graphs are decomposable, the model shows special features. These models are built up from saturated models by successive direct joins and this makes it possible to break down the statistical analysis into smaller analyses of saturated submodels. The density itself can be decomposed in a clique-wise manner. Let $\mathcal{C}$ be the set of cliques of a decomposable graph $\mathcal{G}$. Cliques can be ordered to form a perfect sequence, i.e. $C_1, \ldots, C_k$, where each combination of subgraphs induced by $J_{j-1} = C_1 \cup, C_2, \cup, \ldots, C_{j-1}$ and C_j is a decomposition. Let $S_j = J_{j-1} \cap C_j$ be the sequence of separators. The density decomposes accordingly as follows

$$f(\boldsymbol{x}) = \frac{\prod_{j=1}^{k} f(\boldsymbol{x}_{C_j})}{\prod_{j=1}^{k} f(\boldsymbol{x}_{S_j})}. \tag{2.5}$$

The distribution of the maximum likelihood estimate obeys fundamental conditional independence. Whenever there are three sets A, B and C, such that C separates A from B in $\mathcal{G}$, we have

$$\hat{\boldsymbol{\Sigma}}_{A\cup C} \perp\!\!\!\perp \hat{\boldsymbol{\Sigma}}_{B\cup C}|\hat{\boldsymbol{\Sigma}}_C.$$

This property is called hyper Markov property and was studied by Dawid and Lauritzen (1993).

2.3 Poisson-type graphical models

The Poisson distribution is widely used to model univariate count-valued data and its multivariate generalizations to account for dependencies are starting to increase in popularity. In the last years, in fact, new technologies introduced new complexities in data measurement. Real-world high-dimensional data are more and more often expressed as counts. For example, word counts, crime statistics, and genomics. For the latter, the introduction of RNAseq technologies (Wang *et al.*, 2009) permits to measure gene expression as counts and in this case, variables are usually modeled according to a Poisson or a negative binomial distribution (Anders and Huber, 2010). These types of data are characterized by rich dependencies, highlighting the need for multivariate distributions to appropriately model these data.

The univariate Poisson distribution is the classical model for a count-valued random variable. Assume X is a random variable with sample space $\{0, 1, 2, \dots\}$, its probability distribution is

$$\mathrm{P}(x|\lambda) = \lambda \exp\{-\lambda\}/x!, \tag{2.6}$$

where λ is the mean parameter of the Poisson distribution.

Inouye *et al.* (2017) reviewed multivariate distributions derived from the univariate Poisson distribution. Based on their primary modeling assumptions, these models can be divided into three classes. The first one assumes that the univariate marginal distributions are derived from the Poisson. This assumption is based on the fact that in the multivariate Gaussian distribution, the marginals are univariate Gaussian distributed, see e.g. Teicher (1954). This marginal Poisson property can also be achieved in a more general way using copulas (Nikoloulopoulos and Karlis, 2009; Nikoloulopoulos, 2013; Xue-Kun Song, 2000). However, copula models paired with discrete marginal distributions are

theoretically and computationally more challenging than the corresponding developed for continuous distributions. The second class is derived as a mixture of independent multivariate Poisson distributions. Mixture models are often used to provide more flexibility by allowing the parameter to vary according to a mixing distribution. Moreover, mixture models can model overdispersion, which occurs when the variance is larger than the mean. In these cases, assuming distributions such as the log-normal or log-gamma give flexible dependency structures. Another key benefit of Poisson mixtures is that they permit both positive and negative dependencies. An extensive review of Poisson mixture distributions can be found in Karlis and Xekalaki (2005). The third class assumes that the univariate conditional distributions are derived from the Poisson distribution and they can be studied in the context of graphical models. In the multivariate Gaussian setting, the node-conditional distributions are univariate Gaussian and these models can be seen as an extension of this property to the Poisson case. These conditional models can be seen as undirected graphical models or Markov Random Fields, and they can be parameterized in a simple way. Estimation of these models generally reduces to estimating simple node-wise regressions (Allen and Liu, 2013; Yang *et al.*, 2015).

In this section, we mainly focus on these models, due to their relation to the graphical framework. Besag (1974) was the first one to consider the multivariate extension of the Poisson distribution assuming that conditional distributions are univariate exponential family distributions. The univariate Poisson distribution in (2.6) can be rewritten in the exponential family form (eq. 2.4) as

$$\mathrm{P}(x|\lambda) = \exp\Big\{\log(\lambda)x - \log(x!) - \lambda\Big\}, \tag{2.7}$$

where $\eta = \log(\lambda)$, $T(x) = x$, $B(x) = \log(x!)$ and $A(\eta) = exp(\eta)$. Suppose all node-conditional distributions are univariate Poisson. Then, there is a unique joint distribution consistent with these node-conditional distributions under some conditions. This joint distribution is a graphical model distribution that factors according to a graph specified by the node-conditional distributions. In particular, assume that the node-wise conditional distribution of every random variable X_i, $i = 1, \ldots, p$, follows a univariate Poisson distribution in the exponential family form as stated in (2.7). The pairwise Poisson graphical model

(PGM) is defined as follows

$$\mathrm{P}_{\mathrm{PGM}}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\Phi}) = \exp\left\{\boldsymbol{\theta}^T\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{\Phi}\boldsymbol{x} - \sum_{i=1}^{p}\log(x_i!) - A_{\mathrm{PGM}}(\boldsymbol{\theta},\boldsymbol{\Phi})\right\}, \quad (2.8)$$

where the edge parameters are collected into the symmetric matrix $\boldsymbol{\Phi} \in \mathbb{R}^{p\times p}$, such that $\phi_{ij} = \phi_{ji}$, $\forall(i,j) \in E$ and $\phi_{ij} = 0$, $\forall(i,j) \notin E$. For the PGM, $\boldsymbol{\Phi}$ has zero along the diagonal. The major drawback is that this model only permits negative conditional dependencies between variables, which entails a highly restrictive parameter space, with limited applicability.

In order to overcome this limitation, several extensions of the PGM have been proposed. The first one was introduced by Yang *et al.* (2013) and it is called the truncated Poisson graphical model (TPGM). Based on the idea in Kaiser and Cressie (1997), the authors suggested keeping the same parametric form as in (2.8), but truncating the domain to non-negative integers less or equal to a pre-specified value R, such that the domain becomes $\mathcal{D} = \{0, 1, 2, \ldots, R\}$. Hence, the only difference is that the node-conditional distributions belong to an exponential family that is Poisson-like but with the domain bounded by R, such that the log partition function only involves a finite number of summations. This allows having both negative and positive dependences. However, the major drawback is that as R increases, the pairwise parameters become increasingly negative or close to zero.

A second extension has been defined by Inouye *et al.* (2016), by substituting the sufficient statistics with the square root. This variant is called the square root Poisson graphical model (sqrPGM) and its density function takes the form

$$\mathrm{P}_{\mathrm{SQR}}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\phi}) = \exp\left\{\boldsymbol{\theta}^T\sqrt{\boldsymbol{x}} + \sqrt{\boldsymbol{x}}^T\boldsymbol{\Phi}\sqrt{\boldsymbol{x}} - \sum_{i=1}^{p}\log(x_i!) - A_{\mathrm{SQR}}(\boldsymbol{\theta},\boldsymbol{\Phi})\right\},$$

where ϕ_{ii} can be non-zero. When there are no edges and $\boldsymbol{\theta} = 0$, it reduces to the independent Poisson model. The node conditionals of this distribution are

$$\mathrm{P}(x_i|\boldsymbol{x}_{-i}) \propto \exp\left\{\phi_{ii}x_i + (\theta_i + 2\boldsymbol{\phi}_{i,-i}^T\sqrt{\boldsymbol{x}_{-i}})\sqrt{x_i} - \log(x_i!)\right\},$$

where $\phi_{i,-i}^T$ is the $i-$th column of $\boldsymbol{\Phi}$ with the $i-$th entry removed. The interaction term $\sqrt{\boldsymbol{x}}^T \boldsymbol{\phi} \sqrt{\boldsymbol{x}}$ is linear rather than quadratic and this allows both positive and negative dependencies.

2.4 Differential network analysis

Differential network analysis has become particularly popular in the last years, especially in the biological field. In many cases, the interest is focused on whether and how a particular network changes between two conditions. In this section we briefly review the literature of differential network analysis (Shojaie, 2021), highlighting the fact that most of the proposed methods are developed for Gaussian graphical models, exploiting their properties.

Let $\mathcal{G} = (V, E)$ be a graph with nodes $V = \{1, 2, \ldots, p\}$ and edge set $E \subset V \times V$. Changes in $\mathcal{G}$ can be due to changes in the set of nodes, edges or both. In this section, we focus on settings where the set of nodes V is common to both graphs $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ and the aim is the comparison of the set of edges $E^{(1)}$ and $E^{(2)}$, or equivalently, the adjacency matrices $A^{(1)}$ and $A^{(2)}$. Differences between $A^{(1)}$ and $A^{(2)}$ can be of various nature. One might be interested in the *global* difference between the two matrices, i.e. to check the null hypothesis that $A^{(1)} = A^{(2)}$. For this task, different norms or distances can be considered. The norm-based approach takes the *quantitative* values of estimated parameters, whereas one might be interested in the *qualitative* differences and considering the total number of different edges. In many applications, the focus of the analysis is on the *local* differences. Identifying local differences between graphs can also be of interest after a global test of the difference between the two networks. As in the case of global differences, local differences between two networks can be assessed qualitatively or quantitatively. For instance, in the Gaussian graphical models case, a quantitative analysis would ask to identify node pairs (i, j) such that $\kappa_{ij}^{(1)} \neq \kappa_{ij}^{(2)}$. Alternatively, one might be interested in identifying node pairs (i, j) such that $(i, j) \in E^{(1)}$ but $(i, j) \notin E^{(2)}$. In the Gaussian case, this would imply comparing the zero and non-zero patterns of $\hat{\boldsymbol{K}}^{(1)}$ and $\hat{\boldsymbol{K}}^{(2)}$. The choice of the most appropriate method depends on the application. Basically, differences in the *structure* of the underlying networks

are better captured by qualitative methods, whereas differences in *parameters* of graphical models used to estimate the graph benefit from the use of quantitative approaches. In the latter case, we should also point out that differences in a graph may only concern the structure, i.e. connections between nodes, but the node-level parameters could also change.

2.4.1 Global test

The global null hypothesis of no difference between two Gaussian graphical models is

$$H_0 : E^{(1)} = E^{(2)} \tag{2.9}$$

and can be tested by comparing the covariance matrices, or equivalently the concentration matrices, in the two populations. In fact, in the Gaussian case, (2.9) means testing

$$H_0 : \mathbf{\Sigma}^{(1)} = \mathbf{\Sigma}^{(2)} \quad \text{or, equivalently} \quad H_0 : \mathbf{K}^{(1)} = \mathbf{K}^{(2)}. \tag{2.10}$$

This matrix-based hypothesis can be tested by relying on different methods. The most traditional one is using the likelihood ratio tests (Anderson, 2003; Muirhead, 1982). In the high-dimensional case, more recent approaches compare correlation matrix using the Frobenius norm (Schott, 2007; Li and Chen, 2012) or use eigenstructure (Srivastava and Yanagihara, 2010) and leading eigenvalues (Zhu *et al.*, 2017). However, more recent approaches exploit graphical model properties accounting for the topology of the underlying network (Khatri *et al.*, 2012). Examples are topologyGSA (Massa *et al.*, 2010) and the NetGSA framework (Ma *et al.*, 2019).

2.4.2 The likelihood ratio test

The likelihood ratio test is a statistical procedure for testing the equality of distributions. In particular, in the perspective of testing the equality of

Gaussian graphical models, it can be used for testing the equality of several covariance matrices or for testing that multiple normal distributions are identical, see e.g. Anderson (2003); Muirhead (1982).

In a two-sample problem, testing the equality of covariance matrices reduces to test (2.10). For $j = 1, 2$, let $\boldsymbol{X}_j$ be i.i.d. $\mathcal{N}_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ distributed random vectors with sample sizes n_j. Let $n = n_1 + n_2$ and $A = \sum_{j=1}^{2} A_j$, where

$$\boldsymbol{A}_j = \sum_{i=1}^{n_j} (\boldsymbol{x}_{ji} - \bar{\boldsymbol{x}_j})(\boldsymbol{x}_{ji} - \bar{\boldsymbol{x}_j})^T \quad \text{and} \quad \bar{\boldsymbol{x}_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \boldsymbol{x}_{ji}, \quad i = 1, 2.$$

The modified likelihood ratio test with the unbiasedness property is

$$\Lambda_n = \frac{\prod_{j=2}^{2} \det(\boldsymbol{A}_j)^{(n_j-1)/2}}{\det(\boldsymbol{A})^{(n-2)/2}} \cdot \frac{(n-2)^{(n-2)p/2}}{\prod_{j=2}^{2}(n_j - 1)^{(n_j-1)p/2}}.$$

In order to ensure $\boldsymbol{A}_j$ is full rank we assume $p \leq n_i$ for all $i = 1, 2$. When p is fixed and $\min\{n_1, n_2\} \to \infty$ the approximation to the chi-square is $-2 \log \Lambda_n \xrightarrow{d} \chi^2_f$, where $f = p(p+1)/2$ are the degrees of freedom.

Testing the hypothesis that two normal distributions are identical is equivalent to joint testing the equality of the mean vectors and the covariance matrices. The hypothesis to test is

$$H_0 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}^{(1)} = \boldsymbol{\Sigma}^{(2)} \quad \text{vs.} \quad H_a : H_0 \text{ not true.}$$

For $j = 1, 2$, let $\boldsymbol{X}_j$ be i.i.d. $\mathcal{N}_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ distributed random vectors with sample sizes n_j. Define

$$\boldsymbol{B} = \sum_{j=1}^{2} n_j (\bar{\boldsymbol{x}_j} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}_j} - \bar{\boldsymbol{x}})^T.$$

The likelihood ratio test statistic is

$$\Lambda_n = \frac{\prod_{j=2}^{2} \det(\boldsymbol{A}_j)^{n_j/2}}{\det(\boldsymbol{A} + \boldsymbol{B})^{n/2}} \cdot \frac{n^{np/2}}{\prod_{j=2}^{2} n_j^{n_j p/2}}.$$

In order to ensure the covariance matrices are full rank, we assume $p \leq n_i$ for all $i = 1, 2$. When p is fixed and $\min\{n_1, n_2\} \to \infty$ the approximation to the

chi-square is $-2\log\Lambda_n \xrightarrow{d} \chi^2_f$, where $f = p(p+3)/2$ are the degrees of freedom.

2.4.3 Direct estimation of the difference

Previous methods require to first estimate the individual networks, or covariance matrices, and then compare them in order to estimate the differences. However, when the primary focus of the analysis is on learning the differences between the two graphs, a first step of network estimation might be unnecessary and computationally inefficient. In this case, it is more convenient to directly estimate the differences between the two graphs. In the Gaussian case, Zhao *et al.* (2014) proposed a method based on the work of Cai *et al.* (2011) that directly estimates the sparse difference of two precision matrices, $\boldsymbol{\Delta} = \mathbf{K}^{(1)} - \mathbf{K}^{(2)}$. This is motivated by the fact that the true covariance and precision matrices must satisfy the following relation

$$\boldsymbol{\Sigma}^{(1)}\boldsymbol{\Delta}\boldsymbol{\Sigma}^{(2)} - (\boldsymbol{\Sigma}^{(1)} - \boldsymbol{\Sigma}^{(2)}) = \mathbf{0}.$$

The advantage of this method is that sparsity can be assumed directly for the differences, i.e. on $\boldsymbol{\Delta}$, and not on the individual networks. To overcome existence issues when $n < p$, the author proposed an ℓ_1-norm constrained minimization. An estimate of $\boldsymbol{\Delta}$ can be found by solving the following minimization problem

$$\hat{\boldsymbol{\Delta}} = |\boldsymbol{\Delta}|_1 \quad \text{s.t.} \quad |(\boldsymbol{\Sigma}^{(1)} \otimes \boldsymbol{\Sigma}^{(2)})\text{vec}(\boldsymbol{\Delta}) - \text{vec}(\boldsymbol{\Sigma}^{(1)} - \boldsymbol{\Sigma}^{(2)})|_\infty \leq \rho_n,$$

where $\otimes$ is the Kronecker product and $|\cdot|_\infty$ indicates the sup-norm. Unlike global tests of network differences, methods for direct estimation of differences do not always provide measures of uncertainty, such as confidence intervals and p-values. This issue limits the utility in scientific applications; however, these models provide powerful tools for exploratory analysis and hypothesis generation.

In a similar way, Liu *et al.* (2014) proposed a method that directly estimates the differences in the networks without estimating the individual densities and Kim *et al.* (2021) developed a bootstrap-based procedure to make inference in

a high-dimensional context. The idea is to model the differences between two graphs as the ratio of their density functions, such that the ratio is estimated directly without estimating the densities themselves. The advantage of this approach is that it is suited for all distributions belonging to the exponential family. More details about this algorithm will be discussed in Chapter 4.

Chapter 3

A Bartlett-type correction for likelihood ratio tests

3.1 Introduction

Testing the equality of distributions in a two sample problem can conveniently be done resorting to the likelihood ratio test statistic, $W_n = -2 \log \Lambda_n$, where Λ_n is the likelihood ratio. In Wilks (1938), it is shown that for samples coming from p-variate normal distributions, W_n is asymptotically distributed as a chi-square with $f = p(p+3)/2$ degrees of freedom. It is well known (Muirhead, 1982) that the quality of the asymptotic approximation might be poor in finite sample problems, even at moderate sample sizes. However, convergence to the asymptotic distribution can be improved by multiplying the likelihood ratio test statistic by a constant (Van der Vaart, 1998). Under the low-dimensional setting, where the number of variables p is considered fixed and n is large, the correction factor ρ proposed in Muirhead (1982) improves the convergence rate, but when the value of p is close to n or increases with it, this correction is unable to provide an improvement. In the high-dimensional setting, where p is assumed to increase with n, Jiang and Qi (2015) proposed a standardization of the likelihood ratio test statistic that allows to resort to the central limit theorem and, therefore, to switch to a normal approximation. This solution, however, proves to be inaccurate for small p, given the asymmetry of the likelihood ratio test statistic.

In a recent work, He *et al.* (2021) studied the *phase transition boundary*, d in what follows, which characterizes the approximation accuracy by establishing the necessary and sufficient condition for the chi-square approximation to hold when p increases with n. The authors showed that the chi-square approximation holds if and only if $p/n^d \to 0$, with $d = 1/2$ for the raw likelihood ratio test statistic and $d = 2/3$ for its ρ-corrected version.

In this Chapter, we propose a new multiplicative correction factor, δ_n hereafter, defined to be the ratio between the degrees of freedom of the asymptotic chi-square approximation and an approximation of the expected value of the likelihood ratio test statistic, under the null hypothesis, as a function of p and n. We prove that its phase transition boundary d is equal to 1, so that the chi-square approximation holds in all situations in which $p/n \to 0$. We show the usefulness of our proposal in the context of Gaussian graphical models. Here, the problem of testing equality of two distributions, Markov with respect to a decomposable graph, can be broken up into testing equality of lower dimensional Gaussian distributions. According to the structure of the graph, these lower dimensional problems can lead to very different values of the p/n ratio. Hence, it becomes crucial to rely on an approximation that guarantees a good finite sample accuracy even in extreme cases, where p is close to n. Relying on the decomposability property of the graph allows both to test problems where $n < p$, as long as the dimension of the biggest clique is bigger than the sample size of the smallest sample, but also, it makes possible the identification of the source of a difference in the network. Motivating by the work of Djordjilović and Chiogna (2022), we show how the use of the correction improves the computation time of their algorithm, allowing for the use of asymptotic approximation. Note that testing equality of distribution in a high-dimensional regime is not a simple task and there are other methods that can be used, such as the ones in Gretton *et al.* (2012) and Städler and Mukherjee (2016). However, these latter methods are based on different assumptions on the network and cannot be directly used to localize the difference.

The outline of this Chapter is as follows. In Section 3.2, we introduce the likelihood ratio test for testing equality of multivariate normal distributions in a two-sample problem, and we introduce the most common corrections in

this setting. In Section 3.3 we propose and characterize a new Bartlett-type correction. The extension to testing equality of distribution of decomposable graphical models is described in Section 3.4. Section 3.5 deals with the extension to the algorithm of Djordjilović and Chiogna (2022), for the identification of the source of difference in the network. Proof of the theorem in Section 3.3 and some additional simulations are postponed to the Appendices (Sections 3.8, 3.9).

3.2 State of the art

Consider two p-dimensional multivariate normal distributions, $\mathcal{N}_p(\boldsymbol{\mu}^{(j)}, \boldsymbol{\Sigma}^{(j)})$, $j = 1, 2$, and the problem of testing their equality based on two independent random samples of size n_j. In detail, consider the hypothesis of equality of distributions

$$H_0 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)},\ \boldsymbol{\Sigma}^{(1)} = \boldsymbol{\Sigma}^{(2)} \quad \text{vs.} \quad H_a : H_0 \quad \text{is not true.} \tag{3.1}$$

The likelihood ratio test for testing (3.1), derived in Wilks (1938), can be written as

$$\Lambda_n = \frac{\prod_{j=1}^{2} \det\left(\hat{\boldsymbol{\Sigma}}^{(j)}\right)^{n_j/2}}{\det\left(\hat{\boldsymbol{\Sigma}}\right)^{n/2}},$$

where $n = n_1 + n_2$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Sigma}}^{(j)}$, $j = 1, 2$ are the maximum likelihood estimates of the covariance matrices under the null and alternative hypotheses, respectively, and $\det(\hat{\boldsymbol{\Sigma}})$ denotes the determinant of $\hat{\boldsymbol{\Sigma}}$. Under the null hypothesis in (3.1), the likelihood ratio test statistic $W_n = -2 \log \Lambda_n$, has an asymptotic chi-square distribution, with $f = p(p+3)/2$ degrees of freedom.

In settings where p is fixed and n is allowed to grow, a first correction of the statistic W_n was proposed by Bartlett (1937), based on a re-scaling aimed at making its mean exactly equal to the mean of the asymptotic chi-square distribution, i.e., equal to f. The corrected statistic, W_n^B say, takes the following

form

$$W_n^B = \frac{f}{\mathrm{E}_{H_0}(W_n)} W_n, \tag{3.2}$$

where $\mathrm{E}_{H_0}(W_n)$ is the expected value of W_n under the null hypothesis; see for example Van der Vaart (1998); Pace and Salvan (1997). Later, Muirhead (1982) proposed a version of Bartlett correction that leverages on an expansion of the correction factor, leading to the following correction

$$\rho = 1 - \frac{2p^2 + 9p + 11}{6(p+3)n} \left(\sum_{j=1}^{2} \frac{n}{n_j} - 1 \right). \tag{3.3}$$

The author showed that the resulting corrected statistic, W_n^ρ say, where $W_n^\rho = -2\rho \log \Lambda_n$, has a chi-square limit, with an improved approximation rate with respect to W_n. Both corrections, however, fail when p and n grow at comparable rates.

Recent studies have considered the problem when the dimension p changes with the sample size n. In these settings, Jiang and Yang (2013) and Jiang and Qi (2015) established the following result based on the central limit theorem (CLT):

$$\frac{\log \Lambda_n - \mu_n}{n\sigma_n} \xrightarrow{d} \mathcal{N}(0,1), \tag{3.4}$$

where μ_n and $\sigma_n > 0$ are functions of both n and p and are the asymptotic mean and standard deviation of $\log \Lambda_n$, respectively. The use of the central limit theorem has the advantage of being appropriate in a high-dimensional setting; however, it is less accurate when p is small, due to the asymmetric shape of the likelihood ratio test statistic distribution.

3.3 Our proposal

In this Section, we propose a Bartlett-type correction of the likelihood ratio test statistic, under the assumption that p changes with the sample size n. This correction replaces the denominator of (3.2) with a function of the approximated

mean given in equation (3.4). In a two sample problem, the term μ_n defined by Jiang and Qi (2015) is

$$\mu_n = \frac{1}{4}\left[-4p - \sum_{j=1}^{2}\frac{p}{n_j} + nr_n^2(2p-2n+3) - \sum_{j=1}^{2} n_j r_{n_j'}^2(2p-2n_j+3)\right], \quad (3.5)$$

where $n_j' = n_j - 1$ and $r_x = (-\log(1-p/x))^{1/2}$, for $x > p$, and $n = n_1 + n_2$. Let $\mu_{w_n} = -2\mu_n$, we define the adjusted statistic T_n as

$$T_n = \delta_n W_n, \quad \delta_n = \frac{f}{\mu_{w_n}}, \quad (3.6)$$

where $f = p(p+3)/2$ are the degrees of freedom of the chi-square asymptotic null distribution of W_n. We now prove that T_n is asymptotically chi-square distributed.

Theorem 3.1. *Let* $\mathbf{p} = (p_n)_{n\in N}$ *be a sequence of integers* $1 \leq p_n < n_j - 1$. *Under* H_0, *for* T_n *defined as in (3.6),* $\min_{j=1,2} n_j \to \infty$ *and* $p/n \to 0$, *we have that*

$$\sup_{-\infty<x<\infty} |P(T_n < x) - P(\chi^2_{f_n} < x)| \to 0$$

and the phase transition boundary of T_n *is* $d = 1$.

Proof. See Appendix 3.8. □

In Theorem (3.1), the condition $n_j > p+1$ is assumed to ensure the existence of the likelihood ratio test. Moreover, the condition $p/n \to 0$ defines the phase transition of the adjusted statistic, as introduced in He *et al.* (2021), which represents the boundary in which the chi-square approximation starts to fail as p increases and characterizes the approximation accuracy. This boundary is an improvement over W_n and W_n^ρ, whose approximations hold for $p/n^d \to 0$, with $d = 1/2$ and $d = 2/3$, respectively.

3.3.1 Simulation study

In this Section we present a simulation study to compare the performances of the likelihood ratio test statistics based on four different approximations: the

classic chi-square approximation, the ρ-adjusted approach of Muirhead (1982), the CLT approach of Jiang and Qi (2015) and our proposed δ-adjusted approach.

We study how the correction acts considering a fixed sample size and letting the dimension p change. Data are drawn from a multivariate normal distribution, with fixed covariance matrix and mean vector and we consider $n_1 = n_2 = 50$ and $p = 2, 30, 40$. For each scenario, five thousand simulations are run. Results are shown in Figure 3.1. For each value of p we plot the histograms of the empirical distribution of the four statistics, namely W_n, W_n^ρ, T_n and W_n^{clt}, and compare them with the chi-square distribution with $p(p+3)/2$ degrees of freedom in the first three cases and a standard normal in the last case. The top row of Figure 3.1 shows how the statistic W_n departs from the theoretical χ^2 distribution as p grows. This is expected and motivates the need of an adjustment when dealing with testing problems in which the dimension grows with n. In fact, if 50 observations might be enough for testing a problem of dimension 2, this is not the case for other values of p, especially when p and n have comparable values. The second row shows the results for the statistic corrected with ρ. Note that, also in this case, the approximation to the χ^2 fails as p approaches the group sample size, n_j. With respect to the previous case, however, the departure from the chi-square distribution occurs for higher values of p. The third row highlights the problem of applying the CLT when p is small. For example, when $p = 2$ the approximation to the normal distribution fails, while improves as p increases. This approach works well also for values of p very close to n_j. Finally, the bottom row shows the accuracy of the approximation of the proposed adjusted statistic T_n. Note that this correction leads to a good approximation regardless of the dimension of the testing problem, as long as $p/n \to 0$, and could be used as a unique tool for correcting W_n at different values of p and n.

Finally, we run some simulations to examine the phase transition boundary in Theorem 3.1, under the null hypothesis. We consider $p = \lfloor n_1^\varepsilon \rfloor$, $n_1 = n_2$, $n = \sum_{j=1}^2 n_j$ and $n_j \in \{100, 500, 1000\}$ and finally $\varepsilon \in \{6/24, \ldots, 23/24, 23.5/24\}$. $\lfloor \cdot \rfloor$ denotes the rounding to the nearest integer function. We plot the empirical

type-I error rate (over 1000 simulations) versus ε, for each chi-square approximation: W_n, W_n^ρ and T_n. Results are plotted in Figure 3.2. The first two panels confirm the results in He *et al.* (2021), while the one on the right hand side shows how the phase transition boundary of the adjusted statistic T_n is close to 1. The particular case with ε exactly equal to one is excluded, to ensure the identifiability of the covariance matrix.

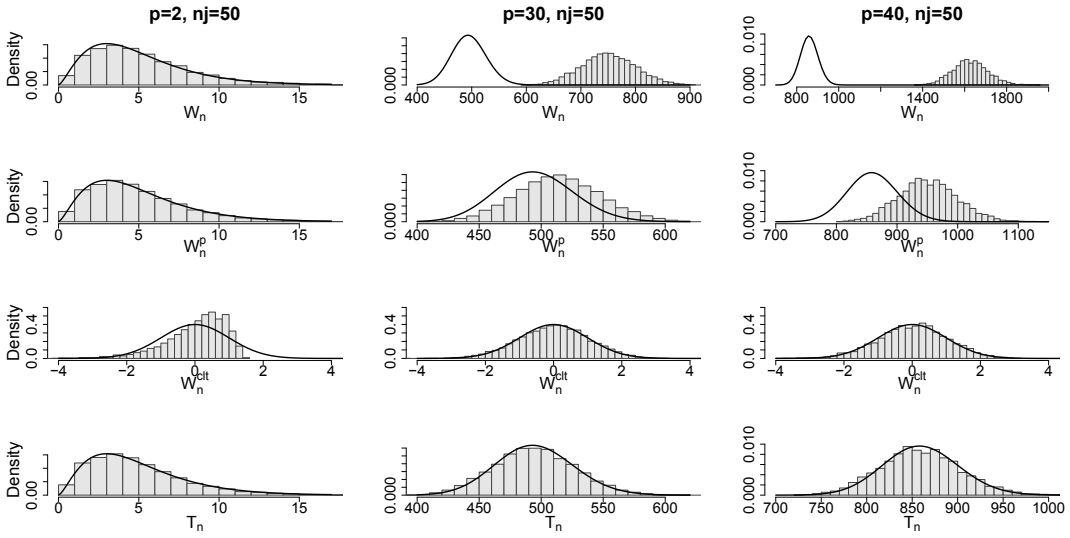

FIGURE 3.1: Simulation results with $n_1 = n_2 = 50$ and $p = 2, 30, 40$. From the top to the bottom row: empirical distribution of W_n, W_n^ρ, W_n^{clt}, and T_n. The solid line in the first, second, and fourth rows shows the nominal χ^2 distribution, with 5, 495 and 860 degrees of freedom (from left to right) respectively. The solid line in the third row, corresponding to the W_n^{clt} statistic, shows the standard normal distribution.

3.4 Testing equality of distributions in Gaussian graphical models

Our proposal finds a natural application in the context of decomposable graphical models. For an overview of the basic theory of (decomposable) undirected graphical models, we refer the reader to Chapter 2. One prominent advantage of decomposable graphs is that their cliques can be arranged so as to satisfy the running intersection property, and the joint probability distribution

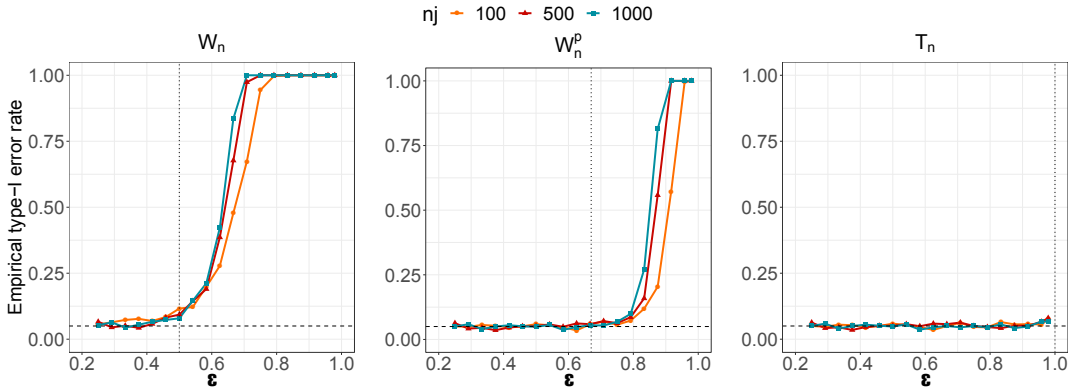

FIGURE 3.2: Chi-square approximation of W_n, W_n^ρ and T_n. Empirical type-I error rate for $n_j \in \{100, 500, 1000\}$, $j = 1, 2$ over 1000 simulations. The vertical dotted lines represents the phase transition boundaries for the three statistics: 1/2, 2/3 and 1, respectively. The horizontal dashed line represents the nominal significance level, 0.05.

of the associated random vectors factorizes accordingly. In detail, if a graph $\mathcal{G} = (V, E)$ decomposes into k, say, cliques, let C_i, $i = 1, \ldots, k$, be a sequence of cliques satisfying the running intersection property and $S_i = C_i \cap C_{i-1}$ and $R_i = C_i \setminus C_{i-1}$, $i = 2, \ldots, k$ the set of corresponding separators and residuals, respectively. Then, the probability distribution of the random vector $\boldsymbol{X}_V$ factorizes as

$$f(\boldsymbol{X}_V) = f(\boldsymbol{X}_{C_1}) f(\boldsymbol{X}_{R_2} | \boldsymbol{X}_{S_2}) \ldots f(\boldsymbol{X}_{R_k} | \boldsymbol{X}_{S_k}). \tag{3.7}$$

See Lauritzen (1996) for an exhaustive explanation. Such factorization renders tractable inference in the setting of large-scale graphical models, where the dimension p of the problem is higher that the available sample size n. Even when $p < n$, using the information on the graphical structure allows us both to improve the power of detecting a difference between the two distributions under study (the size of the model is reduced by constraints on the covariance matrix), and to localize that difference, thanks to the modular nature of graphical models (Djordjilović and Chiogna, 2022). This potential has fed the increasing prominence of graph–theoretic representations of probability distributions in fields such as statistical and quantum physics, bioinformatics, signal processing, econometrics and information theory. In our problem setting, this

factorization assumes a crucial role as it allows to decompose the global problem of testing equality of distribution in two samples into a sequence of local tests of equality of distributions defined on a smaller set of variables, as follows

$$H = \bigcap_{i=1}^{k} H_i, \quad H_i : \boldsymbol{X}_{R_i}^{(1)} | \boldsymbol{X}_{S_i}^{(1)} \stackrel{d}{=} \boldsymbol{X}_{R_i}^{(2)} | \boldsymbol{X}_{S_i}^{(2)}, \quad i = 1, \ldots, k, \tag{3.8}$$

with $S_1 = \varnothing$ and $R_1 = C_1$. Hence, to test the global hypothesis H, one can test the k local hypotheses $\{H_i, i = 1, \ldots, k\}$ of equality of the conditional distributions of $\boldsymbol{X}_{R_i} | \boldsymbol{X}_{S_i}$. In the case of strong meta Markov models (Lauritzen, 1996; Edwards, 2000), as is the Gaussian case, Djordjilović and Chiogna (2022) showed that the local hypotheses $H_i, i = 1, \ldots, k$, are independent and that the likelihood ratio test statistic for testing H also decomposes into k likelihood ratio test statistics, one for testing each local hypothesis. Specifically, the likelihood ratio test, W_n, factorizes as

$$W_n = \sum_{i=1}^{k} \left[W_n^{C_i} - W_n^{S_i} \right] = W_n^{C_1} + \sum_{i=2}^{k} W_n^{C_i | S_i}, \tag{3.9}$$

where W_n^A, $A \subseteq V$, represents the likelihood ratio test for the hypothesis of equality of distributions for $\boldsymbol{X}_A$, namely $H_{(A)} : \boldsymbol{\mu}_A^{(1)} = \boldsymbol{\mu}_A^{(2)}$, $\boldsymbol{\Sigma}_A^{(1)} = \boldsymbol{\Sigma}_A^{(2)}$, while $W_n^{A|B}$ is the likelihood ratio test for the hypothesis of equality of distributions for $\boldsymbol{X}_A | \boldsymbol{X}_B$, $B \subseteq V \setminus A$, namely $H_{(A|B)} : \boldsymbol{\mu}_{(A|B)}^{(1)} = \boldsymbol{\mu}_{(A|B)}^{(2)}$, $\boldsymbol{\Sigma}_{(A|B)}^{(1)} = \boldsymbol{\Sigma}_{(A|B)}^{(2)}$, where $\boldsymbol{\mu}_{(A|B)} = \boldsymbol{\mu}_A - \boldsymbol{\Sigma}_{AB} \boldsymbol{\Sigma}_B^{-1} \mu_B$ and $\boldsymbol{\Sigma}_{(A|B)} = \boldsymbol{\Sigma}_A - \boldsymbol{\Sigma}_{AB} \boldsymbol{\Sigma}_B^{-1} \boldsymbol{\Sigma}_{AB}$.

As proved in Theorem 1 of Djordjilović and Chiogna (2022), the k statistics $W_n^{C_1}$ and $W_n^{C_i | S_i}$, $i = 2, \ldots, k$, in the right-hand side of (3.9) are all asymptotically independent and chi-square distributed, with f_{C_1} and $f_{C_i} - f_{S_i}$, $i = 2, \ldots, k$, degrees of freedom, respectively, being f_{C_i} and f_{S_i} the degrees of freedom associated to the marginal test on the cliques and the separators, respectively. It is worth noting that, since $W_n^{A|B} = W_n^A - W_n^B$, the only quantities needed to compute W_n are the observed values of the likelihood ratio test on the marginal distributions defined over cliques and separators. It is easy to see

that

$$W_n^A = \sum_{j=1}^{2} n_j \log \frac{\det(\hat{\mathbf{\Sigma}}_A)}{\det(\hat{\mathbf{\Sigma}}_A^{(j)})} \tag{3.10}$$

for $A \in \{C_1, \ldots, C_k, S_1, \ldots, S_k\}$. Here, $\hat{\mathbf{\Sigma}}_A$ is the maximum likelihood estimate of $\mathbf{\Sigma}_A$, the block submatrix corresponding to the nodes in A in the null covariance matrix $\mathbf{\Sigma} = \mathbf{\Sigma}^{(1)} = \mathbf{\Sigma}^{(2)}$; and $\hat{\mathbf{\Sigma}}_A^{(j)}$ are the maximum likelihood estimates of $\mathbf{\Sigma}_A^{(j)}$, the block submatrices corresponding to the nodes in A of $\mathbf{\Sigma}^{(j)}$, $j = 1, 2$. Moreover, each W_n^A has a chi-square limit with $f_A = p_A(p_A + 3)/2$ degrees of freedom, where p_A is the cardinality of the set A. One remarkable side effect of the decomposition is that the dimension of each local problem is determined by the cardinality of the set of variables on which it is defined, so that, for a fixed sample size n, dimensionality regimes of local problems vary as a function of their cardinality. Local problems for which $p << n$ might coexist with problems for which $p \approx n$.

Our proposal naturally steps in this context, providing a convenient solution able to accommodate such a variety of situations. The extension of our correction to the test statistics of the kind $W_n^{C|S}$ does not represent an obstacle, resulting indeed being straightforward. In fact, being $\mathrm{E}(W_n^{C|S}) = \mathrm{E}(W_n^C) - \mathrm{E}(W_n^S)$, it results $\mu_n^{C|S} = \mu_n^C - \mu_n^S$. The corrected statistics for the tests relative to the decomposition (3.8) simply become

$$T_n^{C_1} = \delta_n^{C_1} W_n^{C_1}, \quad \delta_n^{C_1} = \frac{f_{C_1}}{\mu_n^{C_1}} \tag{3.11}$$

$$T_n^{C_i|S_i} = \delta_n^{C_i|S_i} W_n^{C_i|S_i}, \quad \delta_n^{C_i|S_i} = \frac{f_{C_i|S_i}}{\mu_n^{C_i|S_i}}, \quad i = 2, \ldots, k. \tag{3.12}$$

3.4.1 Simulation study in the graphical setting

In this section, we present a simulation study aimed at showing the performances of our corrected likelihood ratio tests versus ordinary likelihood ratio test statistics when working with Gaussian graphical models. Data are drawn from a p-variate Gaussian graphical model, Markov with respect to a graph with $p = 14$ nodes and $k = 4$ cliques (Figure 3.4). We consider a RIP-respecting

sequence C_1, C_2, C_3, C_4 of cliques, with cardinalities $|C_1| = 8$, $|C_2| = 5$, $|C_3| = 3$, $|C_4| = 2$, giving rise to the following cardinalities for the corresponding sequence of separators: $|S_2| = 2$, $|S_3| = 1$, $|S_4| = 1$. We generate data assuming

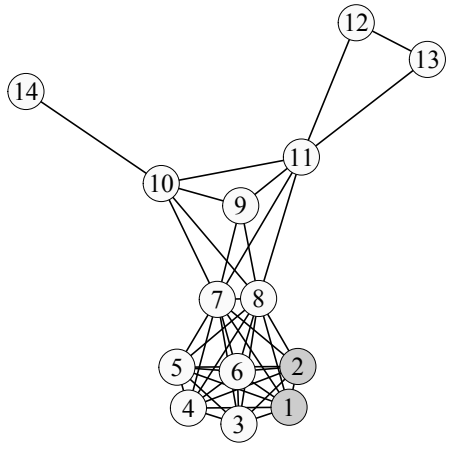

FIGURE 3.3: Graph for the simulation study. Nodes 1 and 2 (gray) are affected by a change in the second scenario.

that differences between the two conditions are attributable to nodes 1 and 2, located in C_1. In particular, in one condition the means of the two elected nodes is set to be 1.5 times greater than the means of the same nodes in the other condition, while the variances are decreased by 50%. It follows that the null hypothesis of equality of distribution for $\boldsymbol{X}_{C_1}$ is false, since C_1 includes the two altered nodes. All remaining null hypotheses of equality of distribution for $\boldsymbol{X}_{R_i}|\boldsymbol{X}_{S_i}$, $i = 2, 3, 4$, are true, thanks to the Markov properties of the graph. We run $10,000$ simulations assuming $n_1 = n_2 \in \{10, 50, 100, 250\}$. For each sample, we compute the following statistics: $W_n^{C_1}$, $W_n^{C_i|S_i}$, $T_n^{C_1}$, $T_n^{C_i|S_i}$, $i = 2, 3, 4$. The nominal Type I error rate is set to be $\alpha = 0.05$.

Results are reported in Table 3.1 (see also Appendix 3.9 for a simulation under the global null). Row 1 of Table 3.1 shows the empirical power of the test, while rows 2-4 show the empirical Type I error rates. For what concerns W_n, note that for small sample sizes, the empirical Type I error rate is significantly higher than the nominal one, due to a large number of false rejections. This

happens for all the local problems, but, for a fixed sample size, the number of false rejections largely depends on the dimension of the problem. As expected, this behavior decreases as the sample size increases, and asymptotically, the distribution of W_n can be approximated with a chi-square. On the other hand, the adjusted statistic T_n reaches the nominal size of the test for each considered sample size, regardless of the dimension of the local problems. The power of the test based on the adjusted statistic T_n on the clique C_1 increases with the sample size. The high power observed for W_n should not be misleading, as it highly depends on the false rejections due to the approximation issues already highlighted in Section 3.3.1. The adjusted statistic seems to meet expectations, being able to identify the altered clique, while controlling the Type I error of the remaining local tests.

	W_n				T_n			
n_j	10	50	100	250	10	50	100	250
C_1	0.985	0.730	0.970	1.000	0.066	0.535	0.946	1.000
$C_2 \vert S_2$	0.445	0.082	0.065	0.056	0.048	0.051	0.050	0.049
$C_3 \vert S_3$	0.167	0.061	0.056	0.051	0.049	0.044	0.048	0.049
$C_4 \vert S_4$	0.109	0.060	0.051	0.057	0.047	0.052	0.048	0.055

TABLE 3.1: Power and Type I error computed for each term of the decomposition. Number of rejected tests out of 10 thousand simulations, for different sample sizes, with significance level $\alpha = 0.05$.

3.5 Identifying the location of the difference

A further natural extension of this correction is in the setting of searching for the source of difference in a graphical model. A further result in the work of Djordjilović and Chiogna (2022) shows how the decomposability property of the graph can be used to identify the origin of the differences in the network. Their method has been also already implemented in an R package, SourceSet, by Salviato *et al.* (2019).

In this section, we first introduce the main algorithm behind their method and the theoretical reasoning, as a continuation of the results in Section 3.4. Then, we describe how our correction steps into this setting and shows the

main advantages of using it through a simulation study. We show how using the corrected statistic T_n as test statistic highly improves the computational time of the algorithm while maintaining comparable performances in terms of control of the type I error rate and power of the original version.

3.5.1 SourceSet: theory and algorithm

The main advantage of this method is the possibility to identify the source of difference in the network by exploiting the structural modularity of decomposable graphical models (Lauritzen, 1996; Frydenberg and Lauritzen, 1989). The set of conditional relevant variables, *seed set*, can be defined as follows (Djordjilović and Chiogna, 2022).

Definition 3.2. Consider the hypothesis of equality of distributions in (3.1) and let $\boldsymbol{\theta}^{(1)} = (\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)})$ and $\boldsymbol{\theta}^{(2)} = (\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}^{(2)})$. We call the set $D \in V$ seed set, if the collection of conditional laws $\boldsymbol{\theta}^{(1)}_{V\setminus D|D}$ and $\boldsymbol{\theta}^{(2)}_{V\setminus D|D}$ coincide. D is also a minimal seed set if no proper subset of it is a seed set itself.

In other words, the source of difference in the network, the *seed set* D, is the set of variables for which the conditional distribution of all the other variables in the network, given D, is equal in the two groups. It is called minimal if D is the minimal subset of variables explaining the difference between the two distributions.

This is made possible thanks to the decomposability property of the graph. The latter translates into the factorization of the density function, as described in equations (3.7), which leads to the factorization of the test statistic, as shown in (3.9). In Section 3.4, the decomposition of the test statistic was used to test the global hypothesis (H) of equality of distributions. Given the set of independent local hypotheses (H_i) in equation (3.8), the null hypothesis of equality of distribution is rejected if the null hypothesis is rejected for at least one local hypothesis. It is easy to see that this decomposition can have an important role in the determination of the location of the difference. The idea is to define an estimator, based on the decomposition of the graph, able to estimate the seed set. By using a clique-grained decomposition, it is not always

possible to identify the minimal seed set, however, it is possible to identify the *graphical seed set*, D_G, namely the superset of D.

Definition 3.3. Let D be a minimal seed set and $\mathcal{S} = \{S : S \text{ is a separator in } \mathcal{G}\}$ the collection of separators in G. The graphical seed set is the set

$$D_G = \{v \in V | \forall S \in \mathcal{S}, \text{ either } v \in \mathcal{S} \text{ or } S \text{ does not separate } v \text{ from } D \text{ in } \mathcal{G}\}.$$

Hence, the graphical seed set D_G is the smallest set containing the seed set D that can be identified by means of set operations on cliques and separators of $\mathcal{G}$. As seen in Section 3.4, the global null hypothesis can be decomposed into a set of independent hypotheses, $H = \bigcap_{i=1}^{k} H_i$. This decomposition is based on a perfect ordering of the cliques, but this is not unique. In fact, we can identify k perfect orderings, one for each clique C_i set as root clique. Each decomposition of the global hypothesis leads to a different factorization of the probability distribution. To identify the j-th decomposition, obtained with C_j as root clique, let $C_{1,j}, \ldots, C_{k,j}$ be the sequence of cliques satisfying the running intersection property and $S_{1,j}, \ldots, S_{k,j}$, the associated sequence of separators. The i-th null hypothesis in decomposition j is denoted by H_{ij}. Hence, an estimator of the graphical seed set is

$$\hat{D}_G = \bigcap_{i=1}^{k} \bigcup_{\{j:H_{ij}\, rejected\}} C_{ij}. \tag{3.13}$$

Due to the multiplicity of tests needed to compute the estimate, it is necessary to apply a correction for multiple testing. The implementation of the algorithm in the SourceSet package applies the *maxT* or *minP* procedure to this aim. See for example Goeman and Solari (2014) for a review of methods for multiple testing. Both these methods are permutation-based and have the advantage of controlling for the family-wise error rate avoiding the asymptotic approximation. The use of permutations however becomes computationally heavy as the number of tests and the sample size increase. Due to the huge amount of data available, the possibility of speeding up computations can be attractive. This is where the correction we defined in Section 3.2 steps in. Using the corrected test statistics as defined in (3.11) and (3.12) gives asymptotically valid p-values

for all the tests and permutations can be avoided when controlling for multiple testing.

3.5.2 Simulation study

In this section, we compare the *sourceSet* algorithm in its original implementation with our new proposal that exploits the corrected statistic T_n. For the simulation, we consider the same setting described in the work of Salviato *et al.* (2019). The graph showing the conditional independence structure of the variables is in Figure 3.4. This can be decomposed into 5 cliques, with 4 separators. Data are drawn from a multivariate normal distribution at different sample sizes. For each scenario, we run 1000 simulations. The family-wise error rate is controlled at a level $\alpha = 0.05$. To achieve this, for the corrected statistic T_n, we use the Hommel correction, while for the row likelihood ratio test statistic W_n we use *minP*, as implemented in the original algorithm, and Hommel as a comparison. See Goeman and Solari (2014) for a description of the multiple correction methods.

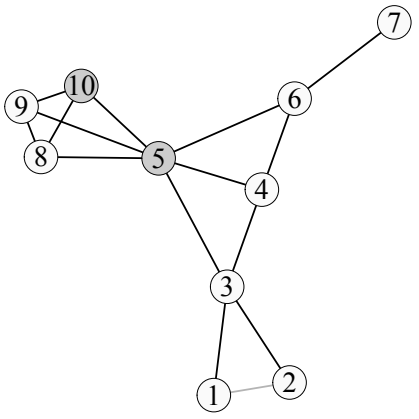

FIGURE 3.4: Graph for the simulation study. Nodes 5 and 10 (gray) and edges 1-2 (light gray) are affected by a change in the second condition of scenarios (ii), (iii), and (iv), respectively.

We define three simulation scenarios:

i. no differences in distribution between the two conditions (H_0);

ii. the source set is a separator, $\{5\}$. The mean and the variance of node 5 differ in the two conditions;

iii. the source set is a clique, $\{5, 8, 9, 10\}$. The mean and the variance of node 10 differ in the two conditions;

iv. the source set is a clique, $\{1, 2, 3\}$. The edge between nodes 1 and 2 is removed in the second condition.

Results for scenarios (i) and (ii) are shown in Figures 3.5 and 3.6, respectively, while results for scenarios (iii) and (iv) are postponed in the Appendix 3.9.3. Figure 3.5 reports the fraction of times the *sourceSet* algorithm identifies $\hat{D}_G$ as an empty set, under the null hypothesis of scenario (i). It is worth noticing that the implementation of the *sourceSet* algorithm with the corrected statistic T_n shows comparable results with the permutation-based approach on the W_n statistic. The gap between the green line (W_n) and the blue one (T_n) highlights the improvement in terms of control of the type I error rate (at the local level), arising from the use of the correction. As expected, the number of false rejections for the raw statistic (blue line) is very high for small values of the sample size n, as can be seen on the left side of the graph.

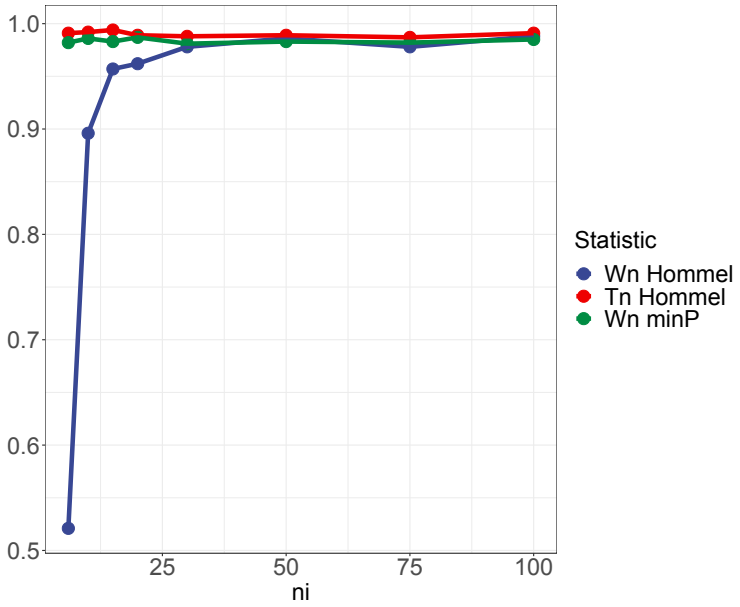

FIGURE 3.5: Fraction of times the *sourceSet* algorithm identifies an empty set, D_G under the null hypothesis of scenario (i). Comparison of the statistics T_n and W_n. The family-wise error rate is controlled at level 0.05 with the $minP$ and Hommel methods.

Figure 3.6 shows the results for scenario (ii). The plot on the left side shows the fraction of times the correct set of altered nodes ($\{5\}$) is identified as the source of difference. The one on the right side shows the rate of false positive discoveries. The corrected statistic T_n shows similar results to the permutation-based approach, in the case of correct discoveries, while it shows a lower number of false positives. As expected, using W_n (and thus relying on asymptotic results when n and p are comparable) leads to a higher number of false discoveries, as we can see in the plot on the right side, where for small n the statistic W_n wrongly identifies nodes as different a higher number of times with respect to the other two cases. The fraction of true set is then higher when n is small, but this result should not be misleading, since it is strongly related to the number of false discoveries.

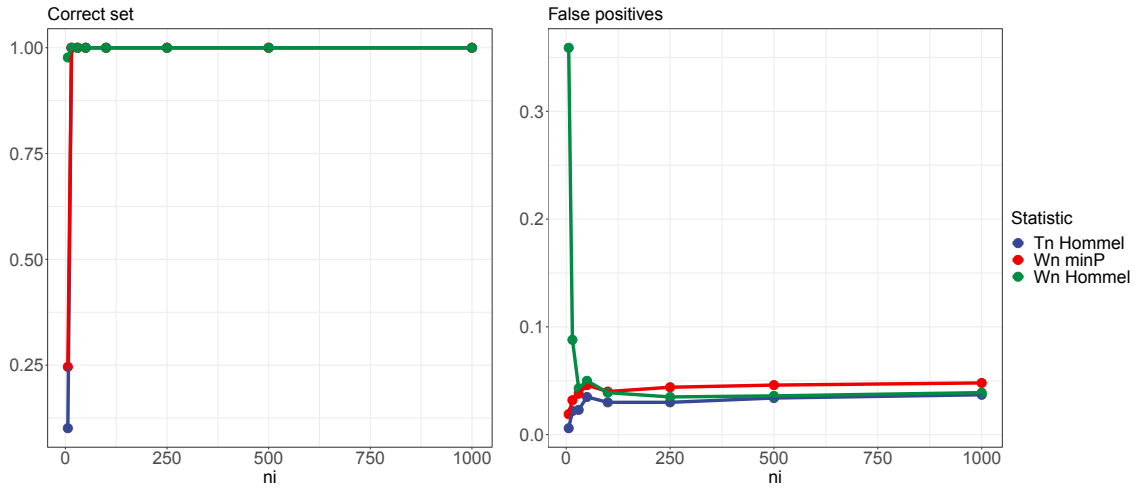

FIGURE 3.6: Comparison of the performance of the statistics T_n and W_n in scenario (ii) considering different multiple testing corrections. On the left-hand side: the fraction of times the correct set of altered nodes ($\{5\}$) is identified as the source of difference. In the right panel: rate of false positive discoveries.

3.5.3 Running time

In the previous Section, we showed how the performance of the permutation-based algorithm and the new version with the corrected statistic T_n presents similar results in terms of power and control of the number of false positives. The main advantage deriving from the use of the new methodology rather than the permutation-based algorithm is in terms of computation time. The number of computations of the *minP* approach has an order of magnitude that depends on n and p. Figure 3.7 shows the mean computation time needed for one run of the algorithms in the previous examples, as n increases. The computation time remains stable (and under one second) when the corrected statistic T_n is used, while it increases with n when using permutations.

It is worth noticing that the example at hand involves a simple small graph, while in real data, pathways can be much bigger with a high number of cliques. This leads, in turn, to a high number of tests to compute. It is clear that with thousands of observations, the computational cost of permutations becomes excessively time-demanding and in real data applications, it is not unusual that

data from single-cell RNA sequencing show a sample size in the order of 10000 observations.

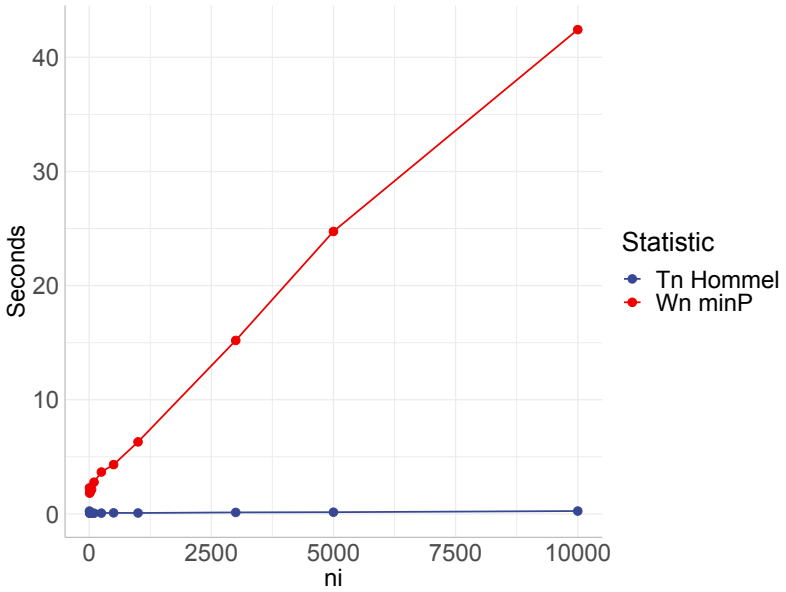

FIGURE 3.7: Running time for the two procedures, considering a graph with 10 nodes and 5 cliques.

3.6 Real data application

As a final step, the performances of the new method are assessed in a real data example. In Section 3.6.1 we apply the procedure described in Section 3.4, while in Section 3.6.2 we apply the *sourceSet* algorithm and compare the results with the ones obtained with the original implementation in Salviato *et al.* (2019). We considered the well-known dataset dealing with the ABL/BCR chimera in acute lymphocytic leukemia (ALL) patients (Chiaretti *et al.*, 2005), available from the R package ALL (Li, 2021). Expression values were normalized according to *rma* and *quantile* normalization (Irizarry *et al.*, 2003b). Genes were annotated using Affymetrix Human Genome U95 Set data and duplicated Entrez IDs were averaged for each sample. Two groups of ALL patients with and without ABL/BCR genomic rearrangement (37 and 42 patients, respectively), were compared.

3.6.1 Testing equality of distributions

The aim of the analysis presented in this Section is to verify the hypothesis of equality of distributions of the genes belonging to the chronic myeloid leukemia pathway, shown in Figure 3.8, whose function is highly impacted by the BCR and ABL genes. The corresponding graph was obtained using the R package graphite (Sales *et al.*, 2012). We finally moralized and triangulated the graph in order to obtain a decomposable graph. The obtained graph consisted of three unconnected sub-graphs, and for illustration reasons, we restricted the analysis to the largest connected component, which also included the two genes of interest, shown in Figure 3.9. The final graph consists of 60 nodes and 30 cliques. We can exploit the decomposability property of the graph and run a test for each sub-hypothesis, after factorizing the density, as described in Section 3.4. For the analysis, we considered one of the 30 possible decompositions of the global null hypothesis.

Results are shown in Table 3.2. Values of the corrected test statistic are reported along with the corresponding degrees of freedom of the test, the p-value, and the adjusted p-value. Adjusted p-values are obtained using the Hommel procedure (see e.g. Goeman and Solari (2014)) in order to control the family-wise error rate. Note that the hypothesis of equality of distribution is rejected for the clique C_1. Hence, we can conclude that the two graphs are different. The clique C_1 consists of three genes, two of which are the ABL and BCR genes, meaning that this method is able to highlight biologically meaningful differences between two sets of patients.

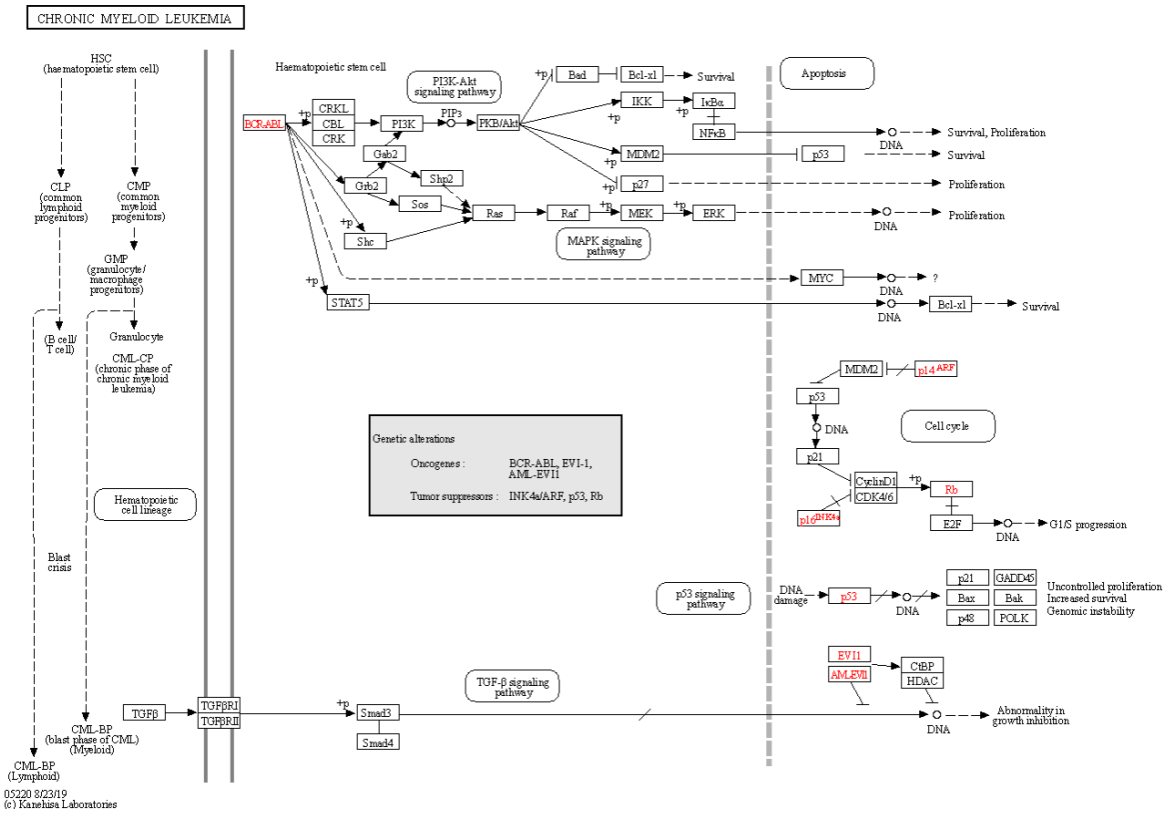

FIGURE 3.8: Chronic myeloid leukemia pathway from KEGG.

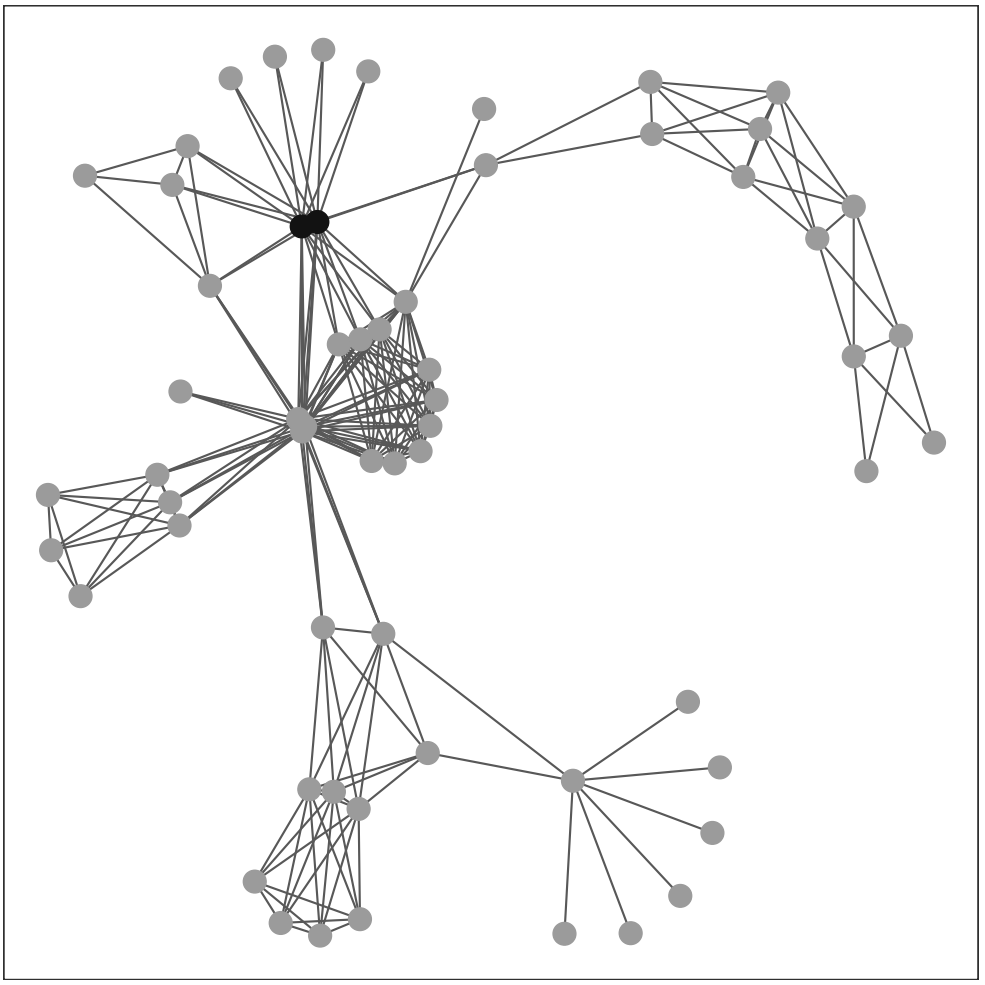

FIGURE 3.9: Undirected graph representing the chronic myeloid leukemia pathway, used for the analysis. Nodes in black represent the ABL and BCR genes.

	T_n	df	pvalue	adj.pvalue
C1	78.52	9	<0.001	<0.001
C2|S2	49.99	49	0.729	0.985
C3|S3	92.61	69	0.398	0.985
C4|S4	8.33	7	0.408	0.985
C5|S5	22.13	18	0.362	0.985
C6|S6	26.15	11	0.016	0.429
C7|S7	1.85	5	0.892	0.985
C8|S8	42.71	18	0.004	0.125
C9|S9	21.02	11	0.066	0.985
C10|S10	1.47	5	0.932	0.985
C11|S11	3.87	5	0.624	0.985
C12|S12	18.23	22	0.806	0.985
C13|S13	39.44	26	0.128	0.985
C14|S14	4.65	4	0.368	0.985
C15|S15	5.2	4	0.309	0.985
C16|S16	12.46	4	0.022	0.585
C17|S17	0.99	4	0.923	0.985
C18|S18	11.12	7	0.170	0.985
C19|S19	16.68	15	0.451	0.985
C20|S20	19.21	11	0.104	0.985
C21|S21	12.49	9	0.251	0.985
C22|S22	4.09	4	0.438	0.985
C23|S23	3.73	4	0.488	0.985
C24|S24	0.16	3	0.985	0.985
C25|S25	2.33	3	0.534	0.985
C26|S26	5.42	3	0.165	0.985
C27|S27	0.42	3	0.942	0.985
C28|S28	6.57	3	0.104	0.985
C29|S29	1.31	3	0.746	0.985
C30|S30	3.96	3	0.294	0.985

TABLE 3.2: Results of the local tests on cliques. Values of the statistic T_n are reported along with the corresponding degrees of freedom (df), the raw p-values, and the adjusted p-values. Adjusted p-values were obtained using the *hommel* procedure in order to control the family-wise error rate.

3.6.2 Studying the source of difference

In this Section we propose an analysis of the ALL dataset, looking at where the network is different. For the analysis, we select from KEGG (Kanehisa and Goto, 2000) all the pathways containing at least one of the chimera genes, and we used the graphite package (Sales *et al.*, 2012) for retrieving the graphical structure. In particular, we select the Chronic myeloid leukemia pathway (i.e., the target pathway) that describes the impact of the ABL/BCR fusion genes in the cell. We apply the *sourceSet* algorithm and compare the permutation-based version to the one that uses the corrected statistic T_n. Results are shown in Figure 3.10. The plot on the top shows the dysregulated genes identified using the corrected statistic T_n, while the bottom one shows the results obtained using the permutation-based algorithm. Plots are composed of a matrix whose rows represent pathways and columns represent genes. Given the structure of the tests needed for the estimate of D_G, results can be summarized as primary set and secondary set. The former is the estimate of D_G, the latter represents the set of nodes (not in the primary set) for which the null hypothesis was rejected in at least one decomposition. Each cell of the output matrix represents the result for each gene in the different pathways. In particular, the cell is blue (2) if the gene is in the primary set of the correspondent pathway; light blue (1) if the gene is in the secondary set of the correspondent pathway; and gray (0) if the gene belongs to the pathway, but was not identified as different considering that specific pathway. If white, the gene does not belong to the pathway. Pathways are vertically ordered (from top to bottom) according to the number of nodes in the source set. The genes are horizontally ordered (from left to right) based on the number of times they appear in a source set.

Results are similar for the two approaches and give a comparable estimate of the D_G. The method based on the corrected statistic identifies a slightly higher number of genes than the other one. The difference in terms of computational time is remarkable. Given the small entity of the problem, in terms of number of pathways and observations, using the corrected statistic gives a result in around 12 seconds, while using permutations needs almost 8 minutes.

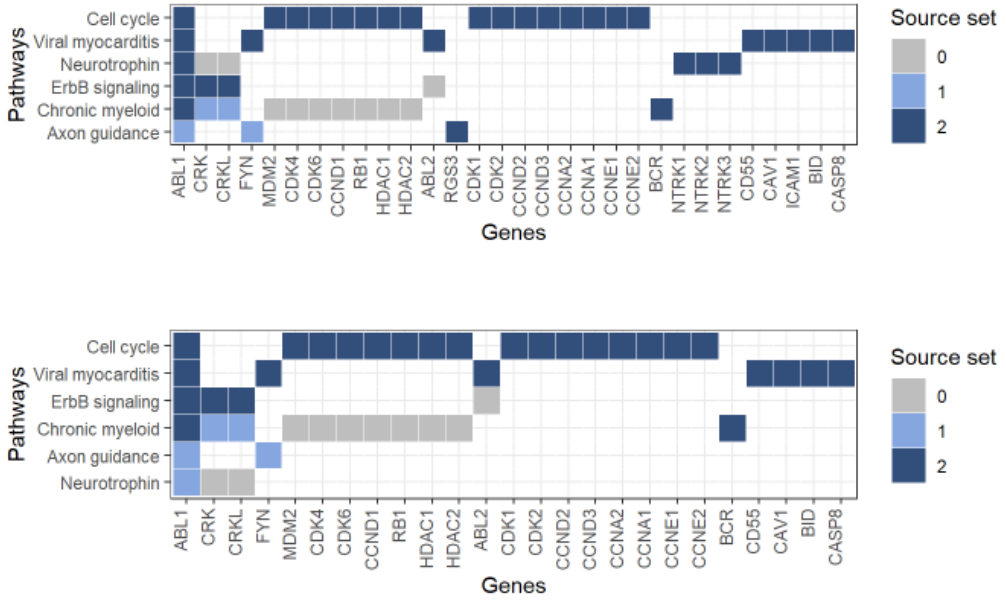

FIGURE 3.10: Sourceset results for the statistics W_δ (first panel) and W (second panel).

3.7 Discussion

In this Chapter, we proposed an adjusted likelihood ratio test, which leads to valid inference at different dimensionality regimes. Our proposal overcomes some weaknesses of alternative corrections reported in the literature, that occur at small sample sizes and, in particular, when the dimension p is close to n. We showed that the phase transition boundary of the likelihood ratio test statistic corrected following our proposal is $d = 1$, indicating that the only condition needed to work is $p/n \to 0$. Simulations confirmed that the adjusted test statistic is well approximated by a chi-square distribution both for small and large values of p.

In the context of decomposable Gaussian graphical models, where the problem of testing the equality of two networks breaks down into a sequence of problems defined on smaller sets of variables, our correction can help tackle the

possibly high heterogeneity resulting from the decomposition in terms of dimensionality regimes. Our simulation study showed that the size of the test was reached for different configurations of p and n and, in the presence of a difference in two conditions, the adjusted statistic is able to detect it, still controlling the Type I error in the other cliques.

This can be extended to the localization of the source of difference in a fixed network. Exploiting the decomposability property of undirected graphs allows to define a series of tests on cliques and separators, such that the estimate of the graphical seed set is identified by means of set operations. Simulations showed that the results obtained using the corrected statistic T_n are similar and comparable to the ones obtained using a permutation-based approach. This drastically improves the computation time for the analysis and makes possible the analysis of big datasets in a reasonable amount of time.

3.8 Appendix 1: proof of Theorem 1

First of all, let $T_n = \delta_n W_n$ as define in (3.6), with $\delta_n = f/\mu_{w_n}$ and $f = p(p+3)/2$. Define now the two main quantities μ_{w_n} and σ_{w_n}, respectively mean and variance of W_n, from the quantities defined in Jiang and Qi (2015), for the specific case of comparison of two populations. Let

$$\mu_{w_n} = \frac{1}{2}\left[4p + \sum_{j=1}^{2}\frac{p}{n_j} + n(2p-2n+3)\log\left(1-\frac{p}{n}\right) - \sum_{j=1}^{2} n_j(2p-2n_j+3)\log\left(-\frac{p}{n_j-1}\right)\right]$$

$$\sigma^2_{w_n} = 2n^2\left[-\sum_{j=1}^{2}\frac{n_j^2}{n^2}\log\left(1-\frac{p}{n_j-1}\right) + \log\left(1-\frac{p}{n}\right)\right]$$

where $n = n_1 + n_2$. Hence, $\mathrm{E}(T_n) = f$ and $\mathrm{Var}(T_n) = \sigma^2_{T_n} = \frac{f^2}{\mu_{w_n}}\sigma^2_{w_n}$.

We prove Theorem (3.1) under two assumptions:

1. $p_n = p$ fixed.
2. $\lim_{n\to\infty} p_n = \infty$

Assumption 1 $f_n = f$ is a fixed integer, such that $f = p(p+3)/2$. It suffices to show that T_n converges in distribution to a χ^2_f. First of all, we show that $f/\mu_{w_n} \to 1$. We use $\log(1-x) = -x - x^2/2 - x^3/3 - x^4/4 + O(x^5)$ and write

$$\begin{aligned} n(2p-2n+3)\log\left(1-\frac{p}{n}\right) &= n(2p-2n+3)\left(-\frac{p}{n} - \frac{p^2}{2n^2} - \frac{p^3}{3n^3} - \frac{p^4}{4n^4} + O\left(\frac{p^5}{n^5}\right)\right) \\ &= 2pn - 3p - p^2 - \frac{p^3}{3n} - \frac{3p^2}{2n} - \frac{p^4}{6n^2} - \frac{3p^4}{4n^3} + O\left(\frac{p^5}{n^3}\right) \end{aligned} \tag{3.14}$$

Similarly, by using Taylor's expansion and $n_j = \Theta(n)$ and using $1/(n_j-1) = 1/n_j + 1/n_j^2 + o(n_i^{-3})$ and $1/(n_j-1)^a = 1/n_j^a + o(n_i^{-3})$ for $a \geq 2$ we have

$$\begin{aligned} n_j(2p-2n_j+3)\log\left(1-\frac{p}{n_j-1}\right) = \\ &= n_j(2p-2n_j+3)\left(-\frac{p}{n_j-1}-\frac{p^2}{2(n_j-1)^2}-\frac{p^3}{3(n_j-1)^3}+O\left(\frac{p^4}{n^4}\right)\right) \\ &= n_j(2p-2n_j+3)\left(-\frac{p}{n_j}-\frac{p}{n_j^2}-\frac{p^2}{2n_j^2}-\frac{p^3}{3n_j^3}+O\left(\frac{p^4}{n^3}\right)\right) \\ &= p^2+p-2pn_j+\frac{3p}{n_j}+\frac{7p^2}{2n_j}+\frac{p^3}{n_j}+\frac{p^3}{3n_j^2}+O\left(\frac{p^4}{n}\right) \end{aligned} \tag{3.15}$$

Hence, as $n \to \infty$, using (3.14) and (3.15) we have

$$\begin{aligned} \mu_{w_n} &= \frac{1}{2}\left[4p-3p-p^2+2p^2+2p+o(n^{-1})\right] \\ &= p(p+3)/2+o(1) = f+o(1). \end{aligned} \tag{3.16}$$

Then, $T_n = W_n(1+o(1))$ and using Slutsky theorem we have that since $W_n \xrightarrow{d} \chi^2_f$, also $T_n \xrightarrow{d} \chi^2_f$.

Assumption 2 If $p_n \to \infty$, as a consequence, also $f_n = p_n(p_n+3)/2 \to \infty$ and we can write

$$\begin{aligned} \lim_{n\to\infty}\sup_x |P(T_n<x)-P(\chi^2_{f_n}<x)| &= \lim_{n\to\infty}\sup_x \left|P\left(\frac{T_n-f_n}{\sigma_{T_n}}<x\right)-P\left(\frac{\chi^2_{f_n}-f_n}{\sigma_{T_n}}<x\right)\right| \\ &= \lim_{n\to\infty}\sup_x \left|P\left(\frac{T_n-f_n}{\sigma_{T_n}}<x\right)-\phi(x)+\phi(x)-P\left(\frac{\chi^2_{f_n}-f_n}{\sigma_{T_n}}<x\right)\right| \end{aligned} \tag{3.17}$$

$(T_n-f_n)/\sigma_{T_n}$ converges in distribution to a $\mathcal{N}(0,1)$ because $(T_n-f_n)/\sigma_{T_n} = (-2\log\Lambda-\mu_n)/\sigma_n \to N(0,1)$ as shown in Jiang and Qi (2015). Moreover, applying Berry-Esseen theorem to $\chi^2_{f_n}$ variable we obtain

$$\lim_{n\to\infty}\sup_x \left|P\left(\frac{\chi^2_{f_n}-f_n}{\sqrt{2f_n}}<x\right)-\phi(x)\right| \to 0 \tag{3.18}$$

Hence, to show (3.17) it is enough to prove that $\sigma_{T_n}^2/(2f_n) \to 1$ as $n \to \infty$. Using (3.14) and (3.15) μ_{w_n} can be written as

$$\begin{aligned}
\mu_{w_n} =& \frac{1}{2}\left\{4p + \sum_{j=1}^{2}\frac{p}{n_j} + 2pn - 3p - p^2 - \frac{p^3}{3n} - \frac{3p^2}{2n} - \frac{p^4}{6n^2} - \frac{3p^4}{4n^3} + O\left(\frac{p^5}{n^3}\right)\right. \\
& \left. 2p^2 + 2p - 2pn + \sum_{j=1}^{2}\left[3\frac{p}{n_j} + 7\frac{p^2}{2n_j} + \frac{p^3}{n_j} + \frac{p^3}{3n_j^2} + O\left(\frac{p^4}{n_j}\right)\right]\right. \\
=& \frac{1}{2}\left[3p + p^2 + O\left(\frac{p}{n}\right) + O\left(\frac{p^2}{n}\right) + O\left(\frac{p^3}{n}\right) + O\left(\frac{p^3}{n^2}\right)\right] \\
=& \frac{1}{2}p(p+3) + O\left(\frac{p^3}{n}\right)
\end{aligned} \tag{3.19}$$

Moreover,

$$\begin{aligned}
\sigma_{w_n}^2 &= 2\sum_{j=1}^{2} n_j^2\left(\frac{p}{n_j - 1} + \frac{p^2}{2(n_j-1)^2} + O\left(\frac{p^3}{n_i^3}\right)\right) - 2n^2\left(\frac{p}{n} + \frac{p^2}{2n^2} + O\left(\frac{p^3}{n^3}\right)\right) \\
&= 2\sum_{j=1}^{2} n_j^2\left(\frac{p}{n_j} + \frac{p}{n_j^2} + \frac{p^2}{2n_j^2} + O\left(\frac{p^3}{n_j^3}\right)\right) - 2\left(pn + \frac{p^2}{2} + O\left(\frac{p^3}{n}\right)\right) \\
&= 2np + 4p + 2p^2 - 2pn - p^2 + O\left(\frac{p^3}{n}\right) \\
&= p^2 + 4p + O\left(\frac{p^3}{n}\right)
\end{aligned} \tag{3.20}$$

Hence, for $\lim_{n\to\infty} p_n/n = 0$, we have

$$\begin{aligned}
\frac{f\sigma_{w_n}^2}{2\mu_{w_n}^2} &= \frac{\frac{1}{2}p(p+3)(4p + p^2 + O(p^3/n))}{2\left(\frac{1}{2}p(p+3) + O(p^3/n)\right)^2} = \frac{4p^3 + p^4 + 12p^2 + 7p^3 + O(p^5/n)}{p^4 + 6p^3 + 9p^2 + O(p^5/n)} \\
&= \frac{p^4(1 + O(p/n))}{p^4(1 + O(p/n))} \to 1
\end{aligned} \tag{3.21}$$

3.9 Appendix 2: additional simulations

3.9.1 Phase transition boundary

In this section, we extend the study of the phase transition boundary under the same assumptions of Figure 3.2, but considering different proportions of the group sample sizes. We set $n_1 = 500$ and $n_2 \in \{1000, 2500, 4000, 10000\}$ such that $n_2/n_1 \in \{2, 5, 8, 20\}$. We take $p = \lfloor n_1^{\varepsilon} \rfloor$ and $\varepsilon \in \{6/24, \ldots, 23/24, 23.5/24\}$, where $\lfloor \cdot \rfloor$ denotes the rounding to the nearest integer function. Figure 3.11 shows the results of the empirical type-I error rate (over 1000 simulations) versus ε, for each chi-square approximation: W_n, W_n^{ρ} and T_n. Simulations show that the relative size of the groups sample sizes does not influence the accuracy of the approximation.

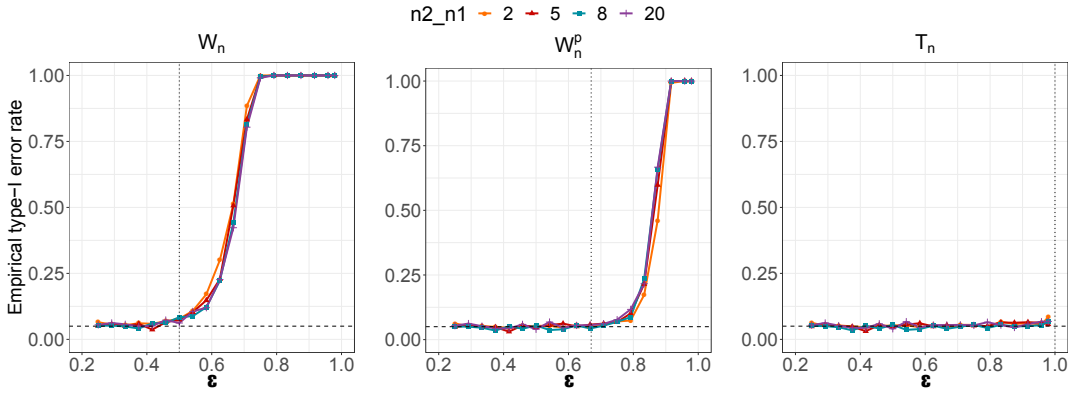

FIGURE 3.11: Chi-square approximation of W_n, W_n^{ρ} and T_n. Empirical type-I error rate over 1000 simulations for $n_1 = 500$ and n_2 such that $n_2/n_1 \in \{2, 5, 8, 20\}$. Phase transition boundaries (vertical dashed lines) for the three statistics respectively: 1/2, 2/3 and 1.

3.9.2 Graphical setting

In this section, we extend the simulation study of Section 3.4.1, showing the results under the global null hypothesis. Data were generated following the same scheme used for Table 3.1, but without considering any changes in the node distribution for the second condition. Results of the empirical type I error rate are shown in Table 3.3. The nominal Type I error rate was set to

be $\alpha = 0.05$. For the clique C_1, the empirical Type I error rate of W_n is higher than the nominal one, especially for low sample sizes. This confirms the lack of Type I error control of the W_n statistic, while T_n controls the Type I error at all sample sizes.

	W_n				T_n			
n_j	10	50	100	250	10	50	100	250
C_1	0.974	0.125	0.085	0.064	0.050	0.049	0.051	0.053
$C_2 \vert S_2$	0.446	0.085	0.064	0.055	0.047	0.051	0.050	0.049
$C_3 \vert S_3$	0.169	0.059	0.058	0.050	0.048	0.044	0.050	0.048
$C_4 \vert S_4$	0.109	0.059	0.050	0.056	0.049	0.050	0.046	0.054

TABLE 3.3: Type I error computed for each term of the decomposition. Number of rejected tests out of 10 thousand simulations, for different sample sizes, with significance level $\alpha = 0.05$.

3.9.3 SourceSet

This section presents the simulation results of scenarios (iii) and (iv), described in Section 3.5.2, respectively reported in Figure 3.12 and 3.13. Plots on the left side show the fraction of times the correct set of altered nodes ($\{5, 8, 9, 10\}$ in Figure 3.12 and $\{1, 2, 3\}$ in Figure 3.13) is identified as the source of difference. The ones on the right side show the rate of false positive discoveries. Also in these cases, the corrected statistic T_n shows similar results to the permutation-based approach, in the case of correct discoveries, while it shows a slightly lower number of false positives.

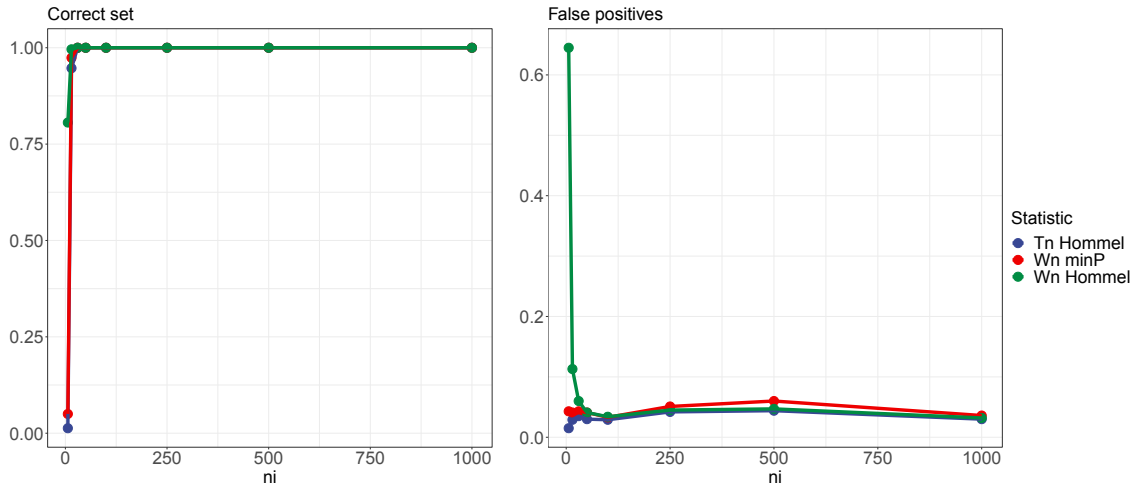

FIGURE 3.12: Comparison of the performance of the statistics T_n and W_n in scenario (iii) considering different multiple testing corrections. On the left-hand side: the fraction of times the correct set of altered nodes ($\{5, 8, 9, 10\}$) is identified as the source of difference. On the right-hand panel: rate of false positive discoveries.

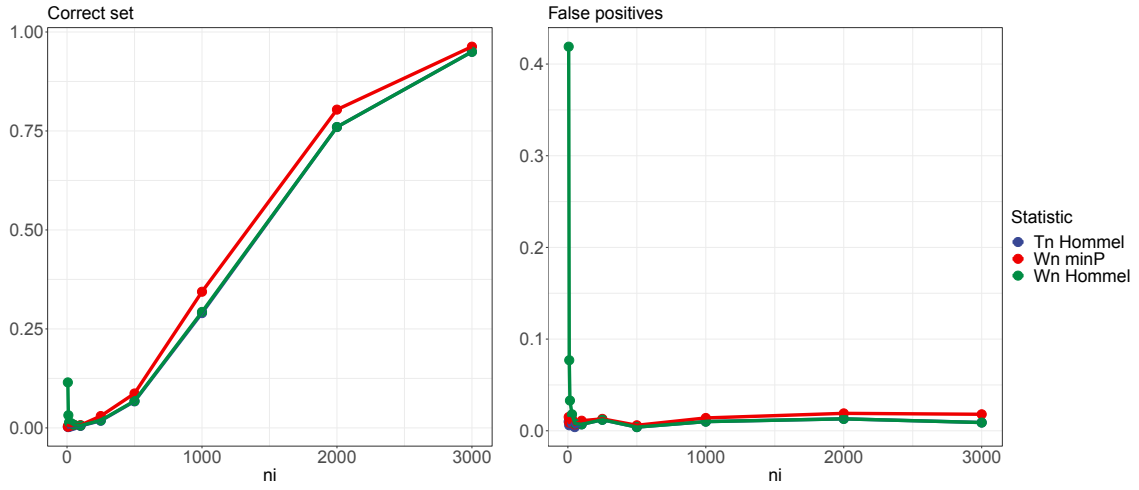

FIGURE 3.13: Comparison of the performance of the statistics T_n and W_n in scenario (iv) considering different multiple testing corrections. On the left-hand side: the fraction of times the correct set of altered nodes ($\{1, 2, 3\}$) is identified as the source of difference. On the right-hand panel: rate of false positive discoveries.

Chapter 4

On the existence of the KLIEP estimator

4.1 Introduction

In this Chapter, we study the problem of learning the differences between two undirected graphical models. The focus is on learning the differences in the networks *directly*, without estimating the individual graphs. This is achieved by using the density ratio approach. For simplicity, in this Chapter we will use the term *density* to indicate both the probability density function and the probability mass function, following the terminology in Liu *et al.* (2014).

Differential network analysis using the density ratio approach gives the opportunity to study the differences in a network without restricting the analysis to a specific structure in advance. Using this approach allows considering in the analysis the connections between variables, without explicitly modeling them. Moreover, it is useful for studying all those types of data for which we can assume a data-generating mechanism belonging to the exponential family.

As described in the previous Chapter, the goal of the analysis is to study and describe if and where the network structure changes between two conditions. When nothing can be assumed about the structure of the interactions between the random variables at hand, a two-step approach is a possible way to deal with the problem. Many statistical approaches have been developed to learn the graph structure over the last decades, see Drton and Maathuis (2017) for a

recent review. A two-step approach for learning differences in the networks is to apply one of these methods separately to the two samples, learn the individual structure, and compare the final models. However, this method can be restrictive in those situations where the individual networks are dense, even if the differences are sparse. In a high-dimensional setting, in fact, an assumption of sparsity is required in order to have a consistent estimation of the network (Cai *et al.*, 2011; Ravikumar *et al.*, 2011; Friedman *et al.*, 2008). Hence, a two-step approach can only work if both individual networks are sparse. Furthermore, how to deal with the two separate tuning steps in the graph estimation process is not fully clear.

Moreover, the literature on difference estimation is mainly developed assuming a particular observation model (Xia *et al.*, 2015; Cai *et al.*, 2019; Zhao *et al.*, 2014). This makes the extension to other parametric models not straightforward and in some cases computationally intractable due to the normalization term.

Liu *et al.* (2014) proposed a method that overcomes these problems by directly estimating the differences in the networks without estimating the individual densities. The idea is to tackle the estimation problem by focusing on the ratio of the two density functions of the two samples, such that the ratio is estimated directly without estimating the densities themselves. In this approach, the parameters of the models represent the difference between the two densities and this allows to directly impose sparsity constraints on the changes. This approach has been developed for any distribution belonging to the exponential family, hence it is suited for *general* Markov random field, with the advantage of avoiding developing different methods for different distributions. The properties of the algorithm have been studied in Liu *et al.* (2017), where the authors provided sufficient conditions for successful change detection with respect to the number of samples in the two groups, the data dimension, and the number of changed edges. In order to keep the density ratio model well-behaved, the magnitude of the change should not be too drastic. Recently, Kim *et al.* (2021) developed a bootstrap-based method to make inference in a high-dimensional context, when the number of observed variables increases with the sample size.

Application of this algorithm can be useful to model differences in gene regulation when conditions change, or to highlight variations in the brain areas connectivity due to particular activity the subject is performing.

This methodology, and differential network analysis in general, mostly focuses on detecting changes in the network structure, but not in the single nodes. For example, in the Gaussian context, it is common to study the differences in the concentration matrix, by first centering the variables. However, this is possible in the Gaussian case due to the orthogonality of the density parameters, but this is not the case for other distributions, such as Poisson graphical models. Using the density ratio approach permits extending the analysis to recover differences in node-wise parameters, simply by treating them as the edge-wise ones.

Even though in principle it is possible to apply this methodology to any statistical model, as long as the density belongs to the exponential family, examples presented in the literature mainly focus on Gaussian or Ising models. Since nowadays more complex data are available, differential network analysis needs to be extended to these new kinds of data. In models for counts data (see Chapter 2) it is not possible to study only the differences in the network structure, and assuming node-wise differences are all zero might be too restrictive. Thus, it is important to study the behavior of this method in this setting.

This Chapter aims to study the performance of the density ratio method when applied to models for count data and in general to any distribution in the exponential family. In particular, we study the necessary and sufficient conditions for the estimate to exist, in finite sample problems. We consider the particular case where changes in both the network structure and node-wise parameters are of interest, leading to a more flexible model.

4.2 Statement of the problem

Consider two independent samples, X and Y, from probability distributions P and Q on $\mathbb{R}^m$. Assume they belong to the family of pairwise Markov networks

and that their densities p and q belong to the exponential family, expressed by

$$p(\boldsymbol{x};\boldsymbol{\theta}^{(p)}) = \frac{1}{Z(\boldsymbol{\theta}^{(p)})}\exp\left(\sum_{v=1}^{m}\boldsymbol{\theta}_v^{(p)}\mathbf{t}_v(x_v) + \sum_{u,v=1,v\geq u}^{m}\boldsymbol{\theta}_{uv}^{(p)}\mathbf{t}_{uv}(x_u,x_v)\right) \tag{4.1}$$

where m is the dimension of the random variable $\boldsymbol{X}$, $Z(\boldsymbol{\theta}^{(p)})$ is the normalization constant and $\boldsymbol{\theta}^{(p)}=(\theta_v)_{v=1}^m \cup (\theta_{uv})_{1\leq u\leq v\leq m}$ is the set of parameters. The normalization factor is defined as

$$Z(\boldsymbol{\theta}^{(p)}) = \int \exp\left(\sum_{v=1}^{m}\boldsymbol{\theta}_v^{(p)}\mathbf{t}_v(x_v) + \sum_{u,v=1,v\geq u}^{m}\boldsymbol{\theta}_{uv}^{(p)}\mathbf{t}_{uv}(x_u,x_v)\right)d\boldsymbol{x}.$$

The density $q(\boldsymbol{Y};\boldsymbol{\theta}^{(q)})$ is defined analogously. The idea in Liu *et al.* (2014) is to look at the ratio between the two densities P and Q,

$$\frac{p(\boldsymbol{x};\boldsymbol{\theta^{(p)}})}{q(\boldsymbol{x};\boldsymbol{\theta^{(q)}})} \propto \exp\left(\sum_{v=1}^{m}(\theta_v^{(p)}-\theta_v^{(q)})t_v(x_v) + \sum_{u,v=1,v\geq u}^{m}(\theta_{uv}^{(p)}-\theta_{uv}^{(q)})t_{uv}(x_u,x_v)\right).$$

Note that the difference between the two densities is represented directly through $\boldsymbol{\Delta} = \boldsymbol{\theta}^{(p)} - \boldsymbol{\theta}^{(q)}$, such that $\theta_j^{(p)} - \theta_j^{(q)}$ is zero if there is no change in the corresponding factor $t_j(\cdot)$. Since the distributions come from the same parametric exponential family, the density ratio still has the exponential form and can be modeled as

$$r(\boldsymbol{x};\boldsymbol{\Delta}) = \frac{1}{N(\boldsymbol{\Delta})}\exp\left(\boldsymbol{\Delta}^T\mathbf{t}(\boldsymbol{x})\right), \tag{4.2}$$

where $\mathbf{t}(\boldsymbol{x}) = (t_v(x_v), t_{uv}(x_v,x_u))$, $v,u = 1,\ldots,m$, $v\geq u$. The term $N(\boldsymbol{\Delta})$ is the normalization constant, which fulfills

$$\begin{aligned} N(\boldsymbol{\Delta}) &= \frac{Z(\boldsymbol{\theta^{(p)}})}{Z(\boldsymbol{\theta^{(q)}})} \\ &= \int \frac{\exp\left(\boldsymbol{\theta}^{(q)T}\mathbf{t}(\boldsymbol{x})\right)}{Z(\boldsymbol{\theta^{(q)}})}\cdot\frac{\exp\left(\boldsymbol{\theta}^{(p)T}\mathbf{t}(\boldsymbol{x})\right)}{\exp\left(\boldsymbol{\theta}^{(q)T}\mathbf{t}(\boldsymbol{x})\right)}d\boldsymbol{x} \\ &= \int q(\boldsymbol{x})r(\boldsymbol{x};\boldsymbol{\Delta})d\boldsymbol{x}. \end{aligned}$$

An estimate of $N(\boldsymbol{\Delta})$ can be obtained by the sample average over $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n_y} \overset{iid}{\sim} q(\boldsymbol{\theta}^{(q)})$ of

$$\hat{N}(\boldsymbol{\Delta}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n_y}) = \frac{1}{n_y} \sum_{i=1}^{n_y} \exp\left(\boldsymbol{\Delta}^T \mathbf{t}(\boldsymbol{y_i})\right). \tag{4.3}$$

Remark 4.1 The setting in Section 4.2 refers to distributions that only allow pairwise interactions between variables. However, it is worth noticing that any other interaction of three or more nodes can be considered. The density ratio in equation (4.2) is rather general and can tackle any difference of parameters, as long as they linearly enter in equation (4.1).

4.2.1 Direct density ratio estimation

The *Kullback-Leibler importance estimation procedure* (KLIEP) minimizes the distance between $p(\boldsymbol{x})$ and $\hat{p}(\boldsymbol{x}; \boldsymbol{\Delta}) = q(\boldsymbol{x}) r(\boldsymbol{x}; \boldsymbol{\Delta})$. Let $D_{KL}(p \parallel q)$ be the Kullback-Leibler divergence for probability densities p and q. It holds that $D_{KL}(p \parallel q) \geq 0$, with equality if and only if $p = q$ almost everywhere and we can write

$$D_{KL}(p \parallel r_\Delta\, q) = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x}) r(\boldsymbol{x}; \boldsymbol{\Delta})} d\boldsymbol{x} \tag{4.4}$$

$$= Const. - \int p(\boldsymbol{x}) \log r(\boldsymbol{x}; \boldsymbol{\Delta}) d\boldsymbol{x}. \tag{4.5}$$

Hereafter, $\log(\cdot)$ is the natural logarithm. An estimate of Δ can be obtained by minimizing the negative empirical approximation of the rightmost term in equation (4.5), such that

$$\begin{aligned} \Delta &= \arg\min_{\Delta} D_{KL}(p \parallel r_\Delta\, q) \\ &= \arg\min_{\Delta} \left(- \mathrm{E}_p[\boldsymbol{\Delta}^T \boldsymbol{t}(\boldsymbol{x})] + \log \mathrm{E}_q[\exp\{\boldsymbol{\Delta}^T \boldsymbol{t}(\boldsymbol{y})\}] \right), \end{aligned}$$

where E_p and E_q represent the expectation with respect to P and Q. The empirical KLIEP loss function (ℓ_{KL}) is obtained by replacing each expectation

with the corresponding sample average, such that

$$\hat{\Delta} = \arg\min_{\Delta} \ell_{KL}(\Delta; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_x}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n_y}) \tag{4.6}$$

$$= \arg\min_{\Delta} \left(-\frac{1}{n_x} \sum_{i=1}^{n_x} \boldsymbol{\Delta}^T \boldsymbol{t}(\boldsymbol{x_i}) + \log \left[\frac{1}{n_y} \sum_{j=1}^{n_y} \exp\{\boldsymbol{\Delta}^T \boldsymbol{t}(\boldsymbol{y_j})\} \right] \right). \tag{4.7}$$

Since the log-sum-exp function is convex (see e.g. Boyd and Vandenberghe (2004)), the loss function ℓ_{KL} itself is a convex function in $\boldsymbol{\Delta}$ and its global minimizer can be found using standard optimization techniques. However, in a finite sample problem, the existence of a minimum is subject to some conditions on the samples. Liu *et al.* (2017) discussed some practical advices on choosing P and Q when datasets are given. In order to guarantee the boundedness of the density ratio, Q should be *wide* and more *spread out* compared to P. The density ratio approach is in fact asymmetric and the performances can be easily affected by the choice of samples. Besides that, even when P and Q are chosen accordingly to the latter indications, the existence of the estimate is subject to more strict conditions on the characteristic of the samples, which are described in the next section.

4.3 On the existence of the estimate of $\boldsymbol{\Delta}$

In this section, we study the properties of the samples that guarantee the possibility of reaching the minimum when it exists. We already know from Liu *et al.* (2017) that the sample chosen to be from Q should be wider than the one from P. However, this result is vague; even when it is satisfied, reaching the minimum is not guaranteed. Hence, assuming the minimum for the parameter Δ exists, we study the necessary and sufficient conditions of the sufficient statistics for the samples X and Y that ensure the minimum, and thus that the estimate is reached.

To ease the notation, let $\boldsymbol{S} = (\boldsymbol{R}, \boldsymbol{T})$ be the set of sufficient statistics for (4.7), from the samples X and Y, respectively. In particular, $\boldsymbol{R} = (R_1, \ldots, R_m)$ is a vector $1 \times m$, where m is the dimension of $\boldsymbol{\Delta}$, of sample means such that

$R_k = \frac{1}{n_x}\sum_{i=1}^{n_x} t_k(\boldsymbol{x_i})$, $k = 1, \ldots, m$. $\mathbf{T}$ is a matrix of dimension $n_y \times m$, where each entrance is $T_{ik} = t_k(\boldsymbol{y_i})$, $k = 1, \ldots, m$ and $i = 1, \ldots, n_y$.

Example 4.1. *Two-dimensional square root Poisson graphical model. Let* $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\Phi})$, *with* $\boldsymbol{\eta} = (\eta_1, \eta_2)$ *and* $\boldsymbol{\Phi} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix}$ *be the parameters. The distribution function for the single observation* $\boldsymbol{x}_i$ *is*

$$f(\boldsymbol{x}_i, \boldsymbol{\theta}) \propto \exp\{\boldsymbol{\eta}^T \sqrt{\boldsymbol{x}_i} + \sqrt{\boldsymbol{x}_i}^T \boldsymbol{\Phi} \sqrt{\boldsymbol{x}_i} - \sum_{k=1}^{2} \log(x_{ki})\}$$

and for the entire sample of dimension n is

$$f(\boldsymbol{x}, \boldsymbol{\theta}) \propto \exp\Bigg\{\eta_1 \sum_{i=1}^{n} \sqrt{x_{1i}} + \eta_2 \sum_{i=1}^{n} \sqrt{x_{2i}} + \phi_{11} \sum_{i=1}^{n} x_{1i} + \phi_{22} \sum_{i=1}^{n} x_{2i} + \\ + 2\phi_{12} \sum_{i=1}^{n} \sqrt{x_{1i}}\sqrt{x_{2i}} - \sum_{i=1}^{n}\sum_{k=1}^{2} \log(x_{ki})\Bigg\}.$$

Let $\boldsymbol{\theta} = (\eta_1, \eta_2, \phi_{11}^2, \phi_{12}, \phi_{22}^2, \phi_{12})$ *be the vector of parameters of the model. The parameters of interest of the density ratio model are* $\boldsymbol{\Delta} = \boldsymbol{\theta}^{(p)} - \boldsymbol{\theta}^{(q)} = (\Delta_1, \ldots, \Delta_m)$, $m = 5$. *Assume the sample X comes from the distribution P and the sample Y is from the distribution Q. The sufficient statistics for the sample X are* $\boldsymbol{R} = (R_1, \ldots, R_5)$, *where* R_1 *and* R_2 *are* $\frac{1}{n_x}\sum_{i=1}^{n_x} \sqrt{x_{ki}}$, *with* $k = 1, 2$ *respectively;* R_3 *and* R_4 *are* $\frac{1}{n_x}\sum_{i=1}^{n_x} x_{ki}$ *with* $k = 1, 2$ *respectively; and* $R_5 = \frac{2}{n_x}\sum_{i=1}^{n_x} \sqrt{x_{1i}}\sqrt{x_{2i}}$. *The sufficient statistics for the sample Y are represented by the* $n_y \times 5$ *matrix* $\boldsymbol{T}$, *with columns* $(\sqrt{\boldsymbol{x}_1}, \sqrt{\boldsymbol{x}_2}, \boldsymbol{x}_1, \boldsymbol{x}_2, 2\sqrt{\boldsymbol{x}_1}\sqrt{\boldsymbol{x}_2})$.

If the minimum of the loss function ℓ_{KL} exists, the following result needs to be satisfied.

Theorem 4.1. *Let X and Y be two samples, and let* $\boldsymbol{S} = (\boldsymbol{R}, \boldsymbol{T})$ *be the sufficient statistics for* (4.7) *from the samples X and Y respectively. Where* $\boldsymbol{R} \in \mathbb{R}^m$ *and* $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_n \in \mathbb{R}^m$. *If the minimum of the loss function* ℓ_{KL} *exists, then* $\boldsymbol{R}$ *needs to lie inside the relative interior of the rows of* $\boldsymbol{T}$.

Proof. Let us writing (4.7) as function of the sufficient statistics $\boldsymbol{S} = (\boldsymbol{R}, \boldsymbol{T})$.

$$\begin{aligned}\ell_{KL}(\Delta) &= -\boldsymbol{\Delta}^T\boldsymbol{R} + \log\left(\frac{1}{n_y}\sum_{i=1}^{n_y}\exp\{\boldsymbol{\Delta}^T\boldsymbol{T}_i\}\right)\\ &= -\sum_{k=1}^{d}\Delta_k R_k + \log\left(\frac{1}{n_y}\sum_{i=1}^{n_y}\exp\left\{\sum_{k=1}^{d}\Delta_k T_{ik}\right\}\right).\end{aligned}$$

The function is convex (see Boyd and Vandenberghe (2004); Sugiyama *et al.* (2012)). To ensure the existence of a minimum, it is required that the vector of the first derivatives is a zero vector. Let the gradient be $\nabla\ell_{KL}(\Delta)$, with j-th component, $k = 1, \ldots, m$,

$$\begin{aligned}\frac{\partial\ell_{KL}(\Delta)}{\partial\Delta_k} &= -R_k + \frac{\sum_{j=1}^{n_y}\exp\left\{\boldsymbol{\Delta}^T\boldsymbol{T}_j\right\}T_{jk}}{\sum_{i=1}^{n_y}\exp\left\{\boldsymbol{\Delta}^T\boldsymbol{T}_i\right\}}\\ &= -R_k + \sum_{j=1}^{n_y}\frac{\exp\left\{\boldsymbol{\Delta}^T\boldsymbol{T}_j\right\}}{\sum_{i=1}^{n_y}\exp\left\{\boldsymbol{\Delta}^T\boldsymbol{T}_i\right\}}T_{jk}\\ &= -R_k + \sum_{j=1}^{n_y}\alpha_j T_{jk},\end{aligned}$$

where

$$\alpha_j = \frac{\exp\left\{\boldsymbol{\Delta}^T\boldsymbol{T}_j\right\}}{\sum_{i=1}^{n_y}\exp\left\{\boldsymbol{\Delta}^T\boldsymbol{T}_i\right\}}, \quad \text{for } j \in [n_y].$$

Note that $\alpha_j > 0$, $\forall j \in [n_y]$ and $\sum_{j=1}^{n_y}\alpha_j = 1$.

Now let $\boldsymbol{\Delta}$ be a minimizer of the loss function ℓ_{KL}. Then the gradient needs to be zero such that the following is satisfied

$$\frac{\partial\ell_{KL}(\boldsymbol{\Delta})}{\partial\boldsymbol{\Delta}} = \begin{pmatrix} -R_1 + \sum_{j=1}^{n_y}\alpha_j T_{j1} \\ -R_2 + \sum_{j=1}^{n_y}\alpha_j T_{j2} \\ \vdots \\ -R_m + \sum_{j=1}^{n_y}\alpha_j T_{j\,m}\end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0\end{pmatrix}.$$

Hence, if it exists a minimum of $\boldsymbol{\Delta}$, then $\boldsymbol{R}$ must lie in the relative interior of the convex hull of the rows of $\boldsymbol{T}$. In other words, $\boldsymbol{R}$ is a point in the relative interior of the polytope that is the convex hull of the rows of $\boldsymbol{T}$. □

To better understand the reasoning, we now analyze the one-dimensional case. Figure 4.1 shows the function in the one-dimensional case when R lies inside, on the boundary, and outside the convex hull of $\boldsymbol{T}$. The x-axis reports the possible values for the parameter Δ, while the y-axis the values of the loss function ℓ_{KL}. As can be seen from the picture, when R is a point in the interior of the polytope generated by $\boldsymbol{T}$, the function is strictly convex, meaning that the minimum, when achieved is also a global minimum. In the other two cases, when either R lies on the boundary or outside the convex hull of the points in $\boldsymbol{T}$, the function is convex but not strictly convex and a global minimizer cannot be found.

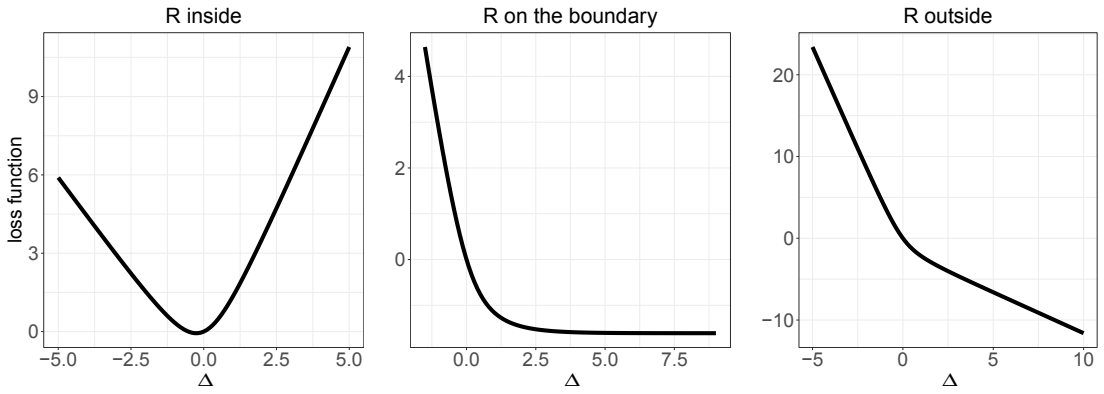

FIGURE 4.1: The loss function ℓ_{KL} when R lies inside, on the boundary, and outside the convex hull of $\boldsymbol{T}$, in the one-dimensional case.

In fact, in the one-dimensional case, Δ is of dimension 1. The sufficient statistics take the form $R = \sum_{i=1}^{n_x} t(x_i)$, while $\boldsymbol{T}$ is of dimension $1 \times n_y$.

Following the result in Theorem 4.1, if the minimum is achieved, R needs to be a convex combination of $\boldsymbol{T}$, see equation (4.3). That is $R = \sum_{j=1}^{n} \alpha_j T_j$. To study the convexity of the function ℓ_{KL} we refer to the second derivative, that is

$$\frac{\partial^2}{\partial \Delta^2} \ell_{KL} = \sum_{j=1}^{n_y} \left[\alpha_j T_j^2 - \alpha_j T_j \cdot \sum_{k=1}^{n_y} \alpha_k T_k \right].$$

A function is strictly convex if the second derivative is strictly positive, that is

$$\sum_{j=1}^{n_y} \alpha_j T_j^2 > \sum_{j=1}^{n_y} \alpha_j T_j \cdot \sum_{k=1}^{n_y} \alpha_k T_k.$$

Using the result in Theorem 4.1, if the minimum is achieved, then $R = \sum_{i=1}^{n_y} \alpha_i T_i$ and we can write

$$\sum_{j=1}^{n_y} \alpha_j T_j^2 > \sum_{j=1}^{n_y} (\alpha_j T_j) R = R^2$$

This is always verified, with the exception of trivial cases. In fact, using the Cauchy-Schwartz inequality we know that

$$\left(\sum_{j=1}^{n_y} \alpha_j\right)\left(\sum_{j=1}^{n_y} \alpha_j T_j^2\right) \geq \left(\sum_{j=1}^{n_y} \sqrt{\alpha_j}\sqrt{\alpha_j}\, T_j\right)^2 = \left(\sum_{j=1}^{n_y} \alpha_j T_j\right)^2 = R^2$$

$$\sum_{j=1}^{n_y} \alpha_j T_j^2 \geq R^2$$

The equality is verified if and only if $\alpha_j = k\alpha_j T_j$ for a non-zero constant $k \in \mathcal{R}$. This would imply $k = 1/T_j$. Hence the equality holds if and only if all T_j are equal, that is a trivial case and we can conclude

$$\sum_{j=1}^{n_y} \alpha_j T_j^2 > R^2.$$

Hence, strict convexity is ensured if and only if R is in the interior of $\boldsymbol{T}$.

4.4 Discussion

In this chapter, we studied the properties of the estimate existence when using the KLIEP algorithm for direct estimation of the differences in a network.

The methodology was first introduced by Liu *et al.* (2014) and recently Kim *et al.* (2021) proposed a bootstrap routine to make inference in this context. The advantage of using the density ratio estimation is derived from the fact that

the differences in the network can be directly estimated without estimating the single networks. This allows relaxing the hypothesis of independence between variables that are usually assumed for example when analyzing transcriptomic data. In this kind of setting, it is known that variables interact with each other, but the underlying network is not always available and known. If in many situations it is possible to restrict the analysis to a subset of variables, for which the connections are studied and known, in many others this is not possible or of interest. Moreover, this methodology is also very flexible due to the fact that it can be applied to any distribution belonging to the exponential family.

Although it is a promising technique, KLIEP shows some limitations in finite samples. It is known (Liu *et al.*, 2017) that this method works well when differences are sparse and relatively small and that the assignment of the samples X and Y to the distributions P and Q plays an important role. In order to ensure the density ratio behaves well, the sample assigned to the distribution Q should be wider and more spread out than the other one. This is a vague indication and the aim of the work presented in this Chapter was to characterize more in detail this statement.

We showed that when the minimum is achieved, the sufficient statistics from the sample X needs to lie inside the polytope generated by the rows of the sufficient statistics of the sample Y. This result is related to the indication in Liu *et al.* (2017), to assign Y to be the widest sample between the two. In fact, in many cases, this would ensure the latter property is satisfied.

This result can be very useful in real data applications, to check in advance if the samples at hand are suitable for having an estimate of the differences, and for choosing which of the two samples to assign to the sample X and which to the sample Y.

Conclusions

Discussion

Differential network analysis plays an important role in studying biological data, especially in transcriptomics, where the main goal is to identify genes that show a significant difference between two conditions. Two-sample problem inference in the context of graphical models has become widely popular in the last decades.

When the first data on gene expression became available, statistical analysis of this type of data was based on the assumption of independence of genes. However, biological processes in a cell involve complex interactions between genes, and these dependencies can be usefully represented by a graph, where nodes and edges represent the genes and their connections, respectively. The state-of-the-art inference procedures usually assume that data arise from a multivariate Gaussian distribution. However, high-throughput omics data are usually discrete, high-dimensional, show a large number of zeros, and come from skewed distributions.

In this thesis, we addressed and studied some problems arising in two-sample problems for graphical models. First of all, we proposed an adjusted likelihood ratio test, useful in differential network analysis of decomposable Gaussian graphical models. We proved that the corrected statistic leads to valid inference at different dimensionality regimes and overcomes some weaknesses of alternative corrections reported in the literature, in particular, when the dimension p is close to n. In the context of decomposable Gaussian graphical models, where the problem of testing the equality of two networks breaks down into a sequence of problems defined on smaller sets of variables, we showed that

the corrected statistic can help tackle the possibly high heterogeneity resulting from the decomposition.

Secondly, we studied the properties of the existence of the estimate in the Kullback-Leibler importance estimation procedure (KLIEP). This approach permits direct estimation of the differences in a network when data are assumed to come from any distribution belonging to the exponential family. In finite sample problems, we studied the characteristics of the samples that need to be satisfied in order to ensure the existence of the estimate. We showed that when the minimum is achieved, the sufficient statistics from sample X needs to lie inside the relative interior of the polytope generated by the rows of the sufficient statistics of sample Y. This result can be very useful in real data applications, to check in advance if the samples at hand are suitable for having an estimate of the differences, and for choosing which of the two samples to assign to sample X and which to sample Y.

Future directions of research

The implementation of the corrected statistic in the context of the *sourceSet* algorithm showed promising results in terms of accuracy and computation time. Future work might be the implementation of the new version of the algorithm in the package, as a first step in the analysis. When dealing with real data, it might be more reliable to use the permutation-based approach anyway, but since it is a time-demanding routine, the asymptotic adjusted procedure might be used as a first step to make an initial selection of pathways and genes to analyze. Moreover, the extension of the algorithm to non-Gaussian data is still an open question. A popular choice for adapting the results available under the Gaussian assumption to non-Gaussian data is data transformation. It would be interesting to study and compare different transformations for count data to better understand which one is the most suited for real data application. Following a recent work of Ahlmann-Eltze and Huber (2022), more popular and simple transformations such as the logarithm and the square root might be compared to more complex ones, such as the copula (Liu *et al.*, 2009) or the randomized quantile residuals (Dunn and Smyth, 1996).

The results on the Kullback-Leibler importance estimation procedure can be seen as a basis for the future development of the algorithm. In fact, given the result stated in Theorem 4.1, improvements of the algorithm can try to solve this finite sample problem. Future directions will be the implementation of a penalty that has to be tuned to account for the convexity problem. In a recent work, Kim *et al.* (2021) suggested the use of a penalty to account for sparsity, but in some cases, this might not be enough if not tuned in the right way. Another direction might be the implementation of a data augmentation step to meet the necessary conditions to guarantee the existence of the estimate. Connected to the latter proposal, it would also be interesting to study the accuracy of the estimate as sufficient statistic progressively approaches the boundary of the polytope.

Bibliography

Ahlmann-Eltze, C. and Huber, W. (2022) Comparison of Transformations for Single-Cell RNA-Seq Data. *bioRxiv* p. 2021.06.24.449781.

Allen, G. I. and Liu, Z. (2013) A local Poisson graphical model for inferring networks from sequencing data. *IEEE Transactions on Nanobioscience* **12**(3), 189–198.

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Nature Precedings* pp. 1–10.

Anderson, T. W. (2003) *An Introduction to Multivariate Statistical Analysis.* Wiley.

Banzato, E., Chiogna, M., Djordjilović, V. and Risso, D. (2022) A Bartlett-type correction for likelihood ratio tests with application to testing equality of Gaussian graphical models. *Statistics & Probability Letters* **193**, e109732.

Bartlett, M. S. (1937) Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences* **160**(901), 268–282.

Beissbarth, T. and Speed, T. P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**(9), 1464–1465.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(2), 192–225.

Boyd, S. and Vandenberghe, L. (2004) *Convex optimization*. Cambridge university press.

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M. and Sherlock, G. (2004) GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**(18), 3710–3715.

Cai, T., Li, H., Ma, J. and Xia, Y. (2019) Differential Markov random field analysis with an application to detecting differential microbial community networks. *Biometrika* **106**(2), 401–416.

Cai, T., Liu, W. and Luo, X. (2011) A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106**(494), 594–607.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Wang, K. S., Mandelli, F., Foà, R. and Ritz, J. (2005) Gene Expression Profiles of B-lineage Adult Acute Lymphocytic Leukemia Reveal Genetic Patterns that Identify Lineage Derivation and Distinct Mechanisms of Transformation. *Clinical Cancer Research* **11**(20), 7209–7219.

Clifford, P. (1990) Markov random fields in statistics. *Disorder in Physical Systems: A volume in honour of John M. Hammersley* pp. 19–32.

Cox, D. and Wermuth, N. (1996) *Multivariate Dependencies: Models, Analysis and Interpretation*. Volume 67. CRC Press.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P. and Stein, L. (2010) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research* **39**(suppl_1), D691–D697.

Dawid, A. P. and Lauritzen, S. L. (1993) Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics* **21**(3), 1272–1317.

Djordjilović, V. (2015) *Graphical modeling of biological pathways.* Ph.D. Thesis, University of Padova.

Djordjilović, V. and Chiogna, M. (2022) Searching for a source of difference in graphical models. *Journal of Multivariate Analysis* **190**, 1–13.

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., Georgescu, C. and Romero, R. (2007) A systems biology approach for pathway level analysis. *Genome Research* **17**(10), 1537–1545.

Drton, M. and Maathuis, M. H. (2017) Structure Learning in Graphical Modeling. *Annual Review of Statistics and Its Application* **4**(1), 365–393.

Dunn, P. K. and Smyth, G. K. (1996) Randomized quantile residuals. *Journal of Computational and Graphical Statistics* **5**(3), 236–244.

Edwards, D. I. (2000) *Introduction to Graphical Modelling.* Springer.

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

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000) Using Bayesian networks to analyze expression data. *Journal of Computational Biology* **7**(3-4), 601–620.

Frydenberg, M. and Lauritzen, S. L. (1989) Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika* **76**(3).

Goeman, J. J. and Bühlmann, P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**(8), 980–987.

Goeman, J. J. and Solari, A. (2014) Multiple hypothesis testing in genomics. *Statistics in Medicine* **33**(11), 1946–1978.

Grechkin, M., Logsdon, B. A., Gentles, A. J. and Lee, S.-I. (2016) Identifying network perturbation in cancer. *PLOS Computational Biology* **12**(5), e1004888.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2012) A kernel two-sample test. *The Journal of Machine Learning Research* **13**(1), 723–773.

He, Y., Meng, B., Zeng, Z. and Xu, G. (2021) On the phase transition of Wilks' phenomenon. *Biometrika* **108**(3), 741–748.

Inouye, D., Ravikumar, P. and Dhillon, I. (2016) Square root graphical models: Multivariate generalizations of univariate exponential families that permit positive dependencies. In *International Conference on Machine Learning*, volume 48, pp. 2445–2453.

Inouye, D. I., Yang, E., Allen, G. I. and Ravikumar, P. (2017) A review of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics* **9**(3), e1398.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B. and Speed, T. P. (2003a) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* **31**(4), e15.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**(2), 249–264.

Jacob, L., Neuvial, P. and Dudoit, S. (2012) More power via graph-structured tests for differential expression of gene networks. *The Annals of Applied Statistics* **6**(2), 561–600.

Jiang, T. and Qi, Y. (2015) Likelihood ratio tests for high-dimensional normal distributions. *Scandinavian Journal of Statistics* **42**(4), 988–1009.

Jiang, T. and Yang, F. (2013) Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. *The Annals of Statistics* **41**(4), 2029–2074.

Kaiser, M. S. and Cressie, N. (1997) Modeling Poisson variables with positive spatial dependence. *Statistics & Probability Letters* **35**(4), 423–432.

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**(1), 27–30.

Karlis, D. and Xekalaki, E. (2005) Mixed poisson distributions. *International Statistical Review/Revue Internationale de Statistique* **73**(1), 35–58.

Khatri, P., Sirota, M. and Butte, A. J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLOS Computational Biology* **8**(2), e1002375.

Kim, B., Liu, S. and Kolar, M. (2021) Two-sample inference for high-dimensional Markov networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **83**(5), 939–962.

Knuiman, M. (1978) Covariance selection. *Advances in Applied Probability* **10**, 123–130.

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

Li, J. and Chen, S. X. (2012) Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics* **40**(2), 908–940.

Li, X. (2021) *ALL: A data package*.

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

Liu, S., Quinn, J. A., Gutmann, M. U., Suzuki, T. and Sugiyama, M. (2014) Direct learning of sparse changes in Markov networks by density ratio estimation. *Neural Computation* **26**(6), 1169–1197.

Liu, S., Suzuki, T., Relator, R., Sese, J., Sugiyama, M. and Fukumizu, K. (2017) Support consistency of direct sparse-change learning in Markov networks. *The Annals of Statistics* **45**(3), 959–990.

Ma, J., Shojaie, A. and Michailidis, G. (2019) A comparative study of topology-based pathway enrichment analysis methods. *BMC Bioinformatics* **20**(1), 1–14.

Massa, M. S., Chiogna, M. and Romualdi, C. (2010) Gene set analysis exploiting the topology of a pathway. *BMC Systems Biology* **4**(1), 1–15.

Muirhead, R. J. (1982) *Aspects of Multivariate Statistical Analysis.* Wiley.

Mukherjee, S., Carignano, A., Seelig, G. and Lee, S.-I. (2018) Identifying progressive gene network perturbation from single-cell RNA-seq data. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5034–5040. IEEE.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349.

Nikoloulopoulos, A. K. (2013) Copula-based models for multivariate discrete response data. In *Copulae in Mathematical and Quantitative Finance*, pp. 231–249.

Nikoloulopoulos, A. K. and Karlis, D. (2009) Modeling multivariate count data using copulas. *Communications in Statistics-Simulation and Computation* **39**(1), 172–187.

Pace, L. and Salvan, A. (1997) *Principles of statistical inference: from a Neo-Fisherian perspective*. Volume 4. World scientific.

Pearl, J. and Paz, A. (1987) GRAPHOIDS: A Graph-based Logic for Reasoning About Relevance Relations. *Advances in Artificial Intelligence* **2**.

Pico, A. R., Kelder, T., Van Iersel, M. P., Hanspers, K., Conklin, B. R. and Evelo, C. (2008) WikiPathways: pathway editing for the people. *PLOS Biology* **6**(7), e184.

Rahnenführer, J., Domingues, F. S., Maydt, J. and Lengauer, T. (2004) Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical Applications in Genetics and Molecular Biology* **3**(1), 1–29.

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011) High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5**, 935–980.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. and Smyth, G. K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**(7), e47.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1), 139–140.

Sales, G., Calura, E., Cavalieri, D. and Romualdi, C. (2012) graphite-a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics* **13**(1), 1–12.

Salviato, E., Djordjilović, V., Chiogna, M. and Romualdi, C. (2019) SourceSet: A graphical model approach to identify primary genes in perturbed biological pathways. *PLOS Computational Biology* **15**(10), e1007357.

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**(12), 6535–6542.

Shojaie, A. (2021) Differential network analysis: A statistical perspective. *Wiley Interdisciplinary Reviews: Computational Statistics* **13**(2), e1508.

Smyth, G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**(1).

Speed, T. (1978) Relations between models for spatial data, contingency tables and Markov fields on graphs. *Advances in Applied Probability* **10**, 111–122.

Srivastava, M. S. and Yanagihara, H. (2010) Testing the equality of several covariance matrices with fewer observations than the dimension. *Journal of Multivariate Analysis* **101**(6), 1319–1329.

Städler, N. and Mukherjee, S. (2016) Two-sample Testing in High Dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **79**(1), 225–246.

Sugiyama, M., Suzuki, T. and Kanamori, T. (2012) *Density ratio estimation in machine learning*. Cambridge University Press.

Teicher, H. (1954) On the multivariate Poisson distribution. *Scandinavian Actuarial Journal* **1954**(1), 1–9.

Van der Vaart, A. W. (1998) *Asymptotic Statistics*. Cambridge University Press.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**(1), 57–63.

Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics*. Wiley.

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

Xia, Y., Cai, T. and Cai, T. T. (2015) Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika* **102**(2), 247–266.

Xue-Kun Song, P. (2000) Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics* **27**(2), 305–320.

Yang, E., Ravikumar, P., Allen, G. I. and Liu, Z. (2015) Graphical models via univariate exponential family distributions. *The Journal of Machine Learning Research* **16**(1), 3813–3847.

Yang, E., Ravikumar, P. K., Allen, G. I. and Liu, Z. (2013) On Poisson graphical models. *Advances in Neural Information Processing Systems* **26**.

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S. and others (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology* **4**(4), 1–8.

Zhao, S. D., Cai, T. T. and Li, H. (2014) Direct estimation of differential networks. *Biometrika* **101**(2), 253–268.

Zhu, L., Lei, J., Devlin, B. and Roeder, K. (2017) Testing high-dimensional covariance matrices, with application to detecting schizophrenia risk genes. *The Annals of Applied Statistics* **11**(3), 1810.

Erika Banzato

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4174
e-mail: erika.banzato@phd.unipd.it

Current Position

Since January 2023
Research Fellow
University of Padova
Department of Statistical Sciences

Research interests

- Graphical models
- Two-sample inference
- High-dimensional statistical inference
- Biological networks

Education

October 2019 – December 2022
PhD (*dottorato*) in Statistical Sciences
University of Padova, Department of Statistical Sciences
Thesis title: "Two-sample inference for graphical models"
Supervisor: Prof. Davide Risso
Co-supervisors: Prof. Monica Chiogna, Dr. Vera Djordjilović and Prof. Mathias Drton

October 2015 – September 2018
Master (*laurea magistrale*) degree in Statistical Sciences.
University of Padova, Department of Statistical Sciences
Title of dissertation: "Accounting for uncertainty in the predictive calibration of prognostic models constructed using multiple imputations with cross-validatory assessment"
Supervisor: Prof. Livio Finos Co-supervisors: Prof. Bart J. A. Mertens and Prof. Liesbeth de Wreede Final mark: 107/110

October 2012 – September 2015
Bachelor degree (*laurea triennale*) in Statistics and Management.
University of Padova, Department of Statistical Sciences
Title of dissertation: "Performance scolastiche e contest familiar. Risultati da un'indagine Istat sulle famiglie" ("School performance and family context. Results from an Istat survey on families")
Supervisor: Prof. Fausta Ongaro
Final mark: 108/110.

Visiting periods

February 2022 – July 2022
Technical University of Munich,
Munich, Germany.
Supervisor: Prof. Mathias Drton

February 2018 – July 2018
Leiden University Medical Center,
Leiden, the Netherlands.
Erasmus+ program, Master's thesis writing.
Supervisors: Prof. Bart J. A. Mertens and Prof. Liesbeth de Wreede

Further education

April 2022 – July 2022
High-dimensional Statistics
Technical University of Munich
Instructor: Prof. Mathias Drton (Technical University of Munich)

September 2021
Graphical models for categorical data with R
13th Virtual Conference of the Italian Region of the International Biometric Society
Instructor: Prof. Monia Lupparelli (University of Florence)

October 2020 – November 2020
Probabilistic Graphical Models
University of Florence
Instructor: Prof. Manfred Jaeger (Aalborg University)

Work experience

March 2019 – September 2019
Reserch Fellow
Center of Biostatistics for Clinical Epidemiology, University of Milano-Bicocca

November 2018 – February 2019
Collaborator
Center of Biostatistics for Clinical Epidemiology, University of Milano-Bicocca

Computer skills

- R, Python, Stata, SAS
- LaTeX, Markdown

Language skills

Italian: native; English: fluent; German: basic.

Awards and Scholarship

2019-2022
Ph.D. Scholarship in Statistical Sciences, XXXV cycle, University of Padova, Padova, Italy

2019
Special mention for best Master's thesis, Premio Oliviero Lessi (SIS)

February 2018 - July 2018
Erasmus+ Scholarship, University of Padova, Padova, Italy

Publications

Articles in journals

Banzato, E., Chiogna, M., Djordjilović, V., Risso, D. (2023). A Bartlett-type correction for likelihood ratio tests with application to testing equality of Gaussian graphical models. *Statistics & Probability Letters*, **193**, e109732.

Mertens, B. J., Banzato, E., de Wreede, L. C. (2020). Construction and assessment of prediction rules for binary outcome in the presence of missing predictor data using multiple imputation and cross-validation: Methodological approach and data-based evaluation. *Biometrical Journal*, **62**(3), 724-741.

Robba, C., Banzato, E., Rebora, P., Iaquaniello, C., Huang, C. Y., Wiegers, E. J., Meyfroidt, G., Citerio, G. (2020). Acute kidney injury in traumatic brain injury patients: results from the collaborative European neurotrauma effectiveness research in traumatic brain injury study. *Critical care medicine*, **49**(1), 112-126.

Robba, C., Rebora, P., Banzato, E., Wiegers, E. J., Stocchetti, N., Menon, D. K., Citerio, G. (2020). Incidence, risk factors, and effects on outcome of ventilator-associated pneumonia in patients with traumatic brain injury: analysis of a large, multicenter, prospective, observational longitudinal study. *Chest*, **158**(6), 2292-2303.

Cardim, D., Robba, C., Czosnyka, M., Savo, D., Mazeraud, A., Iaquaniello, C., Banzato, E., Rebora, P., Citerio, G. (2020). Noninvasive intracranial pressure estimation with transcranial Doppler: a prospective observational study. *Journal of Neurosurgical Anesthesiology*, **32**(4), 349-353.

Working papers

Nguyen, T. K. H., Chiogna, M., Risso, D., Banzato, E. (2022). Guided structure learning of DAGs for count data. arXiv preprint arXiv:2206.09754.

Conference presentations

Banzato, E., Chiogna, M., Djordjilović, V., Risso, D. (2023). A Bartlett-type correction for likelihood ratio tests with application to testing equality of Gaussian graphical models. (poster) *Statistical methods and models for complex data*, Padova, Italy, September 21, 2022

Banzato, E., Chiogna, M., Djordjilović, V., Risso, D. (2021). A modified Bartlett correction for likelihood ratio tests in graphical models. (contributed) *13th Virtual Conference of the Italian Region of the International Biometric Society*, online, September 29, 2021.

Teaching experience

October 2022 – January 2023
Applied Statistics
Master's degree in Molecular Biology
Laboratory, 21 hours
University of Padova
Instructor: Prof. Davide Risso

September 2022 – November 2022
Statistical Methods for Genomics
Bachelor in Genomics, Department of Pharmacy and Biotechnology
Exercises, 12 hours
University of Bologna, Bologna, Italy
Instructor: Prof. Monica Chiogna

September 2022
Introduction to LaTeX
Master's degree in Statistical Sciences
Laboratory, 2.5 hours
University of Padova
Instructor: Prof. Francesco Lisi

October 2021 – January 2022
Applied Statistics
Master's degree in Molecular Biology
Laboratory assistance, 21 hours
University of Padova
Instructor: Prof. Davide Risso

October 2021 – November 2021
Statistical Methods for Genomics
Bachelor in Genomics, Department of Pharmacy and Biotechnology
Exercises, 12 hours
University of Bologna, Bologna, Italy
Instructor: Prof. Monica Chiogna

References

Prof. Davide Risso
Department of Statistical Sciences
University of Padova
via Cesare Battisti, 241-243, Padova, Italy
e-mail: davide.risso@unipd.it

Prof. Monica Chiogna
Department of Statistical Sciences
University of Bologna
Via Belle Arti, 41, Bologna, Italy
e-mail: monica.chiogna2@unibo.it

Dr. Vera Djordjilović
Department of Economics
Ca' Foscari University of Venice
Cannaregio 873, Venice, Italy
e-mail: vera.djordjilovic@unive.it

Prof. Mathias Drton
Department of Mathematics
Technical University of Munich
Boltzmannstr. 3, Garching b. München
Munich, Germany
e-mail: mathias.drton@tum.de