Working Paper Series, N. 4, December 2017

Department of Statistical Sciences
University of Padua
Italy

# Searching for a Source of Difference: a Graphical Model Approach

**Vera Djordjilović**
Department of Biostatistics
University of Oslo
Norway

**Monica Chiogna**
Department of Statistical Sciences
University of Padua
Italy

**Abstract:** In this work, we look at a two-sample problem within the framework of Gaussian graphical models. When the global hypothesis of equality of two distributions is rejected, the interest is usually in localizing the source of difference. Motivated by the idea that diseases can be seen as system perturbations, and by the need to distinguish between the origin of perturbation and components affected by the perturbation, we introduce the concept of a *minimal seed set*, and its graphical counterpart a *graphical seed set*. They intuitively consist of variables driving the difference between the two conditions. We propose a simple and fast testing procedure to estimate the graphical seed set from data, and study its finite sample behavior with a stimulation study. We illustrate our approach in the context of gene set analysis by means of a publicly available gene expression dataset.

**Keywords:** Gaussian graphical models, Two sample problem, Decomposition, Hyper Markov laws, Gene set analysis.

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

## Contents



**Department of Statistical Sciences**
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

**Corresponding author:**
Monica Chiogna
tel: +39 049 827 4183
monica@stat.unipd.it
http://www.stat.unipd.it/monicachiogna

# Searching for a Source of Difference: a Graphical Model Approach

**Vera Djordjilović**
Department of Biostatistics
University of Oslo
Norway

**Monica Chiogna**
Department of Statistical Sciences
University of Padua
Italy

**Abstract:** In this work, we look at a two-sample problem within the framework of Gaussian graphical models. When the global hypothesis of equality of two distributions is rejected, the interest is usually in localizing the source of difference. Motivated by the idea that diseases can be seen as system perturbations, and by the need to distinguish between the origin of perturbation and components affected by the perturbation, we introduce the concept of a *minimal seed set*, and its graphical counterpart a *graphical seed set*. They intuitively consist of variables driving the difference between the two conditions. We propose a simple and fast testing procedure to estimate the graphical seed set from data, and study its finite sample behavior with a stimulation study. We illustrate our approach in the context of gene set analysis by means of a publicly available gene expression dataset.

**Keywords:** Gaussian graphical models, Two sample problem, Decomposition, Hyper Markov laws, Gene set analysis.

## 1 Introduction

An immense amount of data from studies of biological systems reshaped our view of human diseases. The prevailing view is that a living cell is a complex, dynamical system whose functioning is to a large extent determined by interactions of its components. Diseases are then seen as perturbed states of these systems, and our knowledge about relevant genetic and environmental factors is often represented in the form of disease pathways (see for instance KEGG disease database Kanehisa et al., 2009). Network models are often used to describe these systems, and within this framework, diseases are represented as network perturbations (Del Sol et al., 2010).

Living cells are generally quite robust: they are continuously experiencing a wide range of perturbations, but more often than not, elaborate defense mechanisms allow cells to react and adapt to changing conditions. However, there are certain pertur-

bations towards which cells are extremely fragile; consider Huntington's disease, sickle cell anemia, and cystic fibrosis, that all develop due to single gene mutations. In these cases, small, local perturbations trigger a cascade of failures, disrupting normal cell functioning, and resulting in the development of the disease.

Diseases caused by mutations of single genes – monogenic diseases – are the exception rather than the rule. The pathogenesis of most chronic diseases: different cancers, diabetes, and cardiovascular diseases, involve multiple genes, of which many are still unknown. In order to advance our knowledge, scientists usually rely on exploratory research and experimental studies. One possibility is to measure and compare the abundance of different molecular markers – mRNA, protein or epigenomic – between healthy and affected subjects. The aim is to identify differentially expressed markers between the two groups and recommend them for further study in subsequent experiments. Provided sufficient statistical power, the set of identified markers will include the origin of perturbation, or the so-called driver of the differential behavior, as well as the markers affected by the cascade initiated by that perturbation. It goes without saying that, in terms of identifying new drug targets, and developing new therapeutic strategies, it would be highly beneficial to be able to distinguish between these two types of markers. However, standard statistical approaches to two sample problems are not designed to answer this question. In this work, we try to fill in this gap by targeting the problem of identifying the origin of perturbation within the framework of Gaussian graphical models.

In what follows, we will focus on gene set analysis, although we feel that much of the methodology developed here is readily applicable in a wide range of other contexts. The idea of using Gaussian graphical models in the context of gene set analysis is not new, and most recently, Städler and Mukherjee (2015); Zhao et al. (2014) and Xia et al. (2015) focused on identifying changes in the interplay between genes by comparing network structures and precision matrices in two conditions. Our starting point is different: we assume that the structure of a Gaussian graphical model is given by expert knowledge and shared across the two conditions. When the equality of two joint distributions is rejected, the key question is to localize the source of the difference. Massa et al. (2010) compare marginal distributions associated to cliques of the underlying graph and identify those cliques for which the hypothesis of equality is rejected. We go a step further and try to identify, among the genes whose marginal distribution is different, those that are driving the differential behavior. We start by formalizing the notion of driver genes and introducing a novel quantity of interest: the *minimal seed set*, i.e., the smallest set of variables/genes such that after conditioning on it, the distribution of the remaining variables remains the same in two conditions. We then consider its estimation. For a given pair of multivariate distributions, the number of potential seed sets is exponential in the number of variables, and hence estimation from observed data would require performing a large number of tests. We deal with this issue by restricting our search space and considering only *graphical seed sets*, i.e. sets obtainable by basic set operations from cliques of the underlying graph. We then propose a fast and simple testing procedure that exploits the modularity of graphical models and estimates the graphical seed set in linear time.

The layout of the paper is as follows. In Section 2, we motivate our work with

a problem of comparing subjects with and without a specific chromosome rearrangement. In Section 3, we introduce the minimal and the graphical seed set. In Section 4, we describe the main result that underlies our approach – we show how the global hypothesis of equality of two Gaussian graphical distributions, Markov with respect to the same graph, can be decomposed into a set of local, independent hypotheses. In Section 5, we propose an estimator of the graphical seed set, and study its theoretical properties. The performance of the proposed estimator is evaluated through a simulation study in Section 6. The problem introduced in Section 2 is addressed with a proposed approach in Section 7, and a closing discussion is offered in Section 8. Readers less familiar with Gaussian graphical models can find some essential notions and relevant references in the Appendix.

## 2 A motivating example

Chromosome rearrangements can have substantial effects on the regulation of gene expression through a variety of different mechanisms. Therefore, when comparing populations with and without a given gene rearrangement, sound gene set analysis tools are expected to flag most pathways including genes with the rearrangement as statistically significant. As an example, consider the BCR/ABL fusion gene, formed by rearrangement of the breakpoint cluster region (BCR) on chromosome 22 with the c-ABL proto-oncogene on chromosome 9. This rearrangement causes production of an abnormal tyrosine kinase molecule with increased activity, postulated to be responsible for the development of leukemia and is present in virtually all chronic myelogenous leukemia patients. It is also identified in some cases of acute lymphocytic leukemia (ALL), in which it is associated with poor prognosis.

Martini et al. (2013) consider a well-known dataset (Chiaretti et al., 2005) available from an `R` package `ALL` (Li, 2009), already analyzed in Dudoit and van der Laan (2008); Chen et al. (2010); Li et al. (2012); Martini et al. (2013). Data refer to gene expression signatures of two groups of ALL patients: a first group of 37 subjects with BCR/ABL gene rearrangement, and a second group of 41 subjects without the BCR/ABL gene rearrangement. By applying the approach of Massa et al. (2010), almost all pathways containing BCR and/or ABL genes are found to be statistically different.

Nevertheless, identifying BCR/ABL as a driver of the observed dysregulation might be difficult. In the same paper, Martini et al. (2013) propose an empirical algorithm to extract from a dysregulated pathway the portion mostly affected by the dysregulation. With specific reference to the Chronic myeloid leukemia pathway, a pathway whose functioning is highly impacted by BCR and ABL genes, the algorithm arrives at identifying 23 genes as involved in the dysregulation. This certainly allows to zoom into the functioning of the system; still, the special role of ABL and BCR genes in driving the dysregulation is far from being recognized. Tackling this limitation is the aim of this paper.

## 3 The seed set and the graphical seed set

We start by formalizing the notion of the set of variables driving the difference between two conditions under study. We limit our attention to normal random vectors.

**Definition 1** (Seed set)**.** Let $X^{(1)}$ and $X^{(2)}$ be two normal random vectors indexed by a set $V$. We call the set $D \subseteq V$ the *seed* set, if

1. the distribution of $X_D^{(1)}$ differs from that of $X_D^{(2)}$,
2. the conditional distributions $X_{\bar{D}}^{(1)} \mid X_D^{(1)}$ and $X_{\bar{D}}^{(2)} \mid X_D^{(2)}$ coincide, where $\bar{D} = V \setminus D$.

Furthermore, we say that $D$ is a *minimal* seed set, if no proper subset of it is itself a seed set.

*Remarks.*

- The minimal seed set is thus the smallest subset of variables that explains the difference between the two conditions: after conditioning on it, the distributions of the remaining variables are identical.
- If $D$ is a seed set, then any $D' \supset D$ is also a seed set.
- In case of regular normal distributions, a minimal seed set always exists and is unique.
- The number of potential seed sets for a pair of $p$-dimensional distributions is $2^p$.
- Motivation behind this definition is closely related to the casual concepts of intervention and invariance; nevertheless, the seed set is defined in purely mathematical terms.

Since there are $2^p$ potential seed sets for any given pair of $p$-dimensional distributions, identifying the minimal seed set on the basis of observed data is computationally challenging for all but small $p$. However, when comparing two normal distributions Markov with respect to the same graph, significant computational relief is possible. We therefore turn our attention to Gaussian graphical models and the identification of the seed set that takes advantage of the graphical structure.

**Definition 2** (Graphical seed set)**.** Let $D$ be a minimal seed set for $X^{(1)}$ and $X^{(2)}$, two Gaussian graphical distributions Markov with respect to a decomposable, undirected graph $G = (V, E)$, where $V$ is a set of nodes and $E$ is a set of edges. Let $\mathcal{S} = \{S : S \text{ is a separator in } G\}$ be a collection of separators in $G$. Then we call the set

$$D_G = \{v \in V \mid \nexists S \in \mathcal{S}, \text{s.t } v \notin S \text{ and } S \text{ separates } v \text{ from } D \text{ in } G\} \tag{1}$$

a *graphical seed set.*

*Remarks.*

- Note that $S$ separates $v$ from $D$ when all paths between $v$ and any element of $D$ pass through some element of $S$. We allow for non-empty intersection between $S$ and $D$, as well as $S = D$.
- For $v \in D$, the condition (1) is trivially satisfied ($v$ cannot be separated from $D$ by any set), and therefore $D_G \supseteq D$.
- When the minimal seed set is a separator, we can set $S = D$ in (1), to obtain $D = D_G$. In general, $D$ and $D_G$ will coincide whenever $D$ can be expressed as an intersection of two or more cliques. In other instances, $D_G$ will be a seed set, but not a minimal one.

The graphical seed set $D_G$ is thus the smallest set containing the seed set $D$ that can be identified by means of set operations on cliques and separators of $G$. In general, $D_G$ will be larger than the set of interest, i.e. the minimal seed set $D$; however, in what follows we will show that if we focus on $D_G$, we can exploit the graphical structure and obtain an estimating procedure linear in the number of variables. Before we have a look at the proposed estimator in Section 5, we dedicate the next Section to the theoretical result underpinning our approach. In particular, we show how the global hypothesis of equality of two distributions belonging to the same Gaussian graphical model decomposes into a set of independent local hypotheses.

# 4 Decomposition of the global hypothesis of equality of two Gaussian graphical distributions

Let $G = (V, E)$ be a decomposable undirected graph on $p$ vertices. Let $C_1, \ldots, C_k$ be a sequence of its cliques satisfying a running intersection property, and let $S_2, \ldots, S_k$ be an associated sequence of separators.

**Theorem 1.** *Let $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$ and $X_1^{(2)}, \ldots, X_{n_2}^{(2)}$ be two random samples from $\mathsf{N}(\mu^{(1)}, \Sigma^{(1)})$ and $\mathsf{N}(\mu^{(2)}, \Sigma^{(2)})$, $\mu^{(l)} \in \mathbb{R}^p$, $\left(\Sigma^{(l)}\right)^{-1} \in S^+(G)$, $l = 1, 2$, and consider the hypothesis of equality of distributions*

$$H : \mu^{(1)} = \mu^{(2)} \quad \textit{and} \quad \Sigma^{(1)} = \Sigma^{(2)}. \tag{2}$$

*Let $\lambda(V)$ denote the log likelihood ratio criterion for testing (2) and let $\lambda(A)$ denote the log likelihood ratio criterion for testing $H_A : \mu_A^{(1)} = \mu_A^{(2)}$ and $\Sigma_A^{(1)} = \Sigma_A^{(2)}$ for $A \subseteq V$. The following equality holds*

$$\lambda(V) = \lambda(C_1) + \sum_{j=2}^{k} \left[\lambda(C_j) - \lambda(S_j)\right], \tag{3}$$

*Moreover, the $k$ terms on the right hand side of (3) are asymptotically independent under the null hypothesis.*

It is worth noting that the terms in the summation on the right handside represent the likelihood ratio test statistic for the test of equality of conditional distributions $X_{C_j \setminus S_j} \mid X_{S_j}$, $j = 2, \ldots, k$. This feature plays a crucial role when estimating $D_G$ in Section 5.

*Proof.* Let $\lambda := \lambda(V)$. We recall the log likelihood ratio statistic for testing $H$

$$\lambda = \sum_{l=1}^{2} n_l \log \frac{|\hat{\Sigma}|}{|\hat{\Sigma}^{(l)}|},$$

where $|\hat{\Sigma}|$ is determinant of the maximum likelihood estimate of $\Sigma$ under $H$, $\hat{\Sigma}$, with

$$\hat{\Sigma} = \frac{1}{n_1 + n_2} \left[ \sum_{i=1}^{n_1} (X_i^{(1)} - \bar{X})(X_i^{(1)} - \bar{X})^\top + \sum_{j=1}^{n_2} (X_j^{(2)} - \bar{X})(X_j^{(2)} - \bar{X})^\top \right],$$

where

$$\bar{X} = \frac{1}{n_1 + n_2} \left( n_1 \bar{X}^{(1)} + n_2 \bar{X}^{(2)} \right), \quad \bar{X}^{(l)} = \frac{1}{n_l} \sum_{i=1}^{n_l} X_i^{(l)}, \quad l = 1, 2,$$

and $\hat{\Sigma}^{(l)}$, $l = 1, 2$, are maximum likelihood estimates of $\Sigma^{(l)}$ under the general alternative, given by

$$\hat{\Sigma}^{(l)} = \frac{1}{n_l} \sum_{i=1}^{n_l} (X_i^{(l)} - \bar{X}^{(l)})(X_i^{(l)} - \bar{X}^{(l)})^\top, \quad l = 1, 2.$$

We recall that the asymptotic distribution of $\lambda$ is chi square with $\mathrm{card}(E) + 2p$ degrees of freedom, where $\mathrm{card}(E)$ denotes the cardinality of $E$.

Since the determinant of every $\Omega$ for which $\Omega^{-1} \in S^+(G)$ can be decomposed with respect to the graph as $|\Omega| = \prod_{i=1}^{k} |\Omega_{C_i}| / \prod_{i=2}^{k} |\Omega_{S_i}|$, and therefore also the determinants of $\hat{\Sigma}$ and $\hat{\Sigma}^{(l)}$, $l = 1, 2$, the above equality can be equivalently written as

$$\lambda = \sum_{i=1}^{k} \lambda(C_i) - \sum_{i=2}^{k} \lambda(S_i),$$

from which the equality (3) follows.

The asymptotic independence of terms in the right-hand side of (3) can be seen as an immediate consequence of the hyper Markov property and the well known results regarding the maximum likelihood estimation in Gaussian graphical models (see Appendix for details). Let us first consider the case $k = 2$. Let $C_1$, $C_2$ be the two cliques satisfying the running intersection property, $S_2$ be the associated separator, $A = C_1 \setminus S_2$, $S = S_2$ and $B = C_2 \setminus S_2$, so that $(A, S, B)$ is a decomposition of $G$. It is easy to see that $H = H_1 \cap H_2$, where $H_1 : \theta^{(1)}_{A \cup S} = \theta^{(2)}_{A \cup S}$ and $H_2 : \theta^{(1)}_{B|S} = \theta^{(2)}_{B|S}$ concern variation independent parameters.

Exploiting the block structure of $\hat{\Sigma}_{B\cup S}$, we obtain $|\hat{\Sigma}_{B\cup S}| = |\hat{\Sigma}_S|/|\hat{K}_B|$, and the equality (3) becomes

$$\lambda = \sum_{l=1}^{2} n_l \log \frac{|\hat{\Sigma}_{A\cup S}|}{|\hat{\Sigma}^{(l)}_{A\cup S}|} + \sum_{l=1}^{2} n_l \log \frac{|\hat{K}^{(l)}_B|}{|\hat{K}_B|}. \tag{4}$$

The first term on the right hand side, $\lambda(A \cup S)$, corresponds to the likelihood ratio test for the hypothesis of equality of marginal distributions induced by $A \cup S$, i.e., $H_{01} : \theta^{(1)}_{A\cup S} = \theta^{(2)}_{A\cup S}$. The second term, that we might informally denote as $\lambda(B \mid S)$, corresponds to the likelihood ratio test for the hypothesis of equality of conditional distributions induced by variables in $B$ given the variables in $S$, i.e. $H_2 : \theta^{(1)}_{B|S} = \theta^{(2)}_{B|S}$. It is $\lambda(A \cup S) = \lambda(C_1)$ and $\lambda(B \mid S) = \lambda(C_2) - \lambda(S_2)$. Thanks to variation independence of the parameters in $H_1$, $H_2$ and to their L-independence, this implies that $\lambda(A \cup S)$ and $\lambda(B \mid S)$ are asymptotically independent not only under $H$, but whenever one of the two hypotheses is true, i.e., under $H_1 \cup H_2$.

For $k > 2$, asymptotic independence for all pairs of subsequent components of (3) is proven analogously, which together with the characterizing property of the chi-square distribution (Tan, 1977) suffices to prove the joint asymptotic independence. □

## 5 Estimating the graphical seed set

We have seen above that, within the framework of Gaussian graphical models, the global hypothesis of equality can be decomposed according to a specified perfect ordering into a set of local independent hypotheses. By independent hypotheses, we mean that there are no logical relations between them, and that all combinations of true and false hypotheses are possible. However, the perfect ordering is not unique. In fact, there are multiple decompositions of the global hypothesis, each corresponding to a different factorization of the same distribution. It is this multiplicity that we exploit when estimating the graphical seed set.

For a given graph, the enumeration of all decompositions might resemble the problem of enumerating its junction trees (Thomas and Green, 2009), but a closer look reveals that it is a far simpler task. Given the uniqueness of the sequence of separators, it is not difficult to show that there is exactly one decomposition for each choice of the root clique – the clique labeled $C_1$ – leading to a total of $k$ decompositions.

Before we show how these different decompositions relate to the graphical seed set in Proposition 1, we introduce some notation and restate the testing problem (2) in decision theory terms. Let $\Theta$ be the unrestricted parameter space of $(\mu^{(l)}, \Sigma^{(l)})$, $l = 1, 2$ where $\left(\Sigma^{(l)}\right)^{-1} \in S^+(G)$, $l = 1, 2$; let $\Theta_0$ denote the space restricted by $H$ in (1), and let $\Theta_1 = \Theta \setminus \Theta_0$. We want to test $H : \theta \in \Theta_0$ against a general alternative $\theta \in \Theta_1$. Let the decision taken on $H$ be denoted by $d$, where $d = 0$ means that the null hypothesis is not rejected and $d = 1$ means that the null hypothesis is rejected. A test $\phi$ is a mapping from the sample space to the set $\{0, 1\}$ (we rule out the trivial case that the test makes no decisions). Let $d^*$ denote the correct

decision (the truth) for the null hypothesis in (2). As seen in the previous Section, the null hypothesis can be decomposed into a set of independent local hypotheses, i.e., $H = \bigcap_{j=1}^{k} H_j$, and we denote by $d_j^*$ the correct decision for $H_j$, $j = 1, \dots k$, so that $d^* = (d_1^*, \dots, d_k^*)$. To identify the $i-$th decomposition, obtained when $C_i$ is set as the root clique, we let $C_{i,1}, \dots, C_{i,k}$ denote a sequence of cliques satisfying the running intersection property. Let $S_{i,2}, \dots, S_{i,k}$ be an associated sequence of separators, and set $S_{i,1} = \varnothing$, $i = 1, \dots, k$. In this notation, $H_{i,j}$ will denote the $j-$th null hypothesis in decomposition $i$, $\phi_{i,j}$ the corresponding test, and $d_{i,j}^*$ the associated correct decision.

We now show the connection between the graphical seed set and the decompositions obtained from the graph $G$.

**Proposition 1.** Let $d_i^* = \left(d_{i,1}^*, \dots, d_{i,k}^*\right)$ be the vector of correct decisions for the hypotheses $H_{i,j}$ of equality of distribution of $X_{C_{i,j} \setminus S_{i,j}} \mid X_{S_{i,j}}$ in the $i-$th decomposition. We then have

$$D_G = \bigcap_{i=1}^{k} \bigcup_{\{j:\, d_{i,j}^*=1\}} C_{i,j}.$$

*Proof.* Let $P = \bigcap_{i=1}^{k} \bigcup_{\{j:\, d_{i,j}^*=1\}} C_{i,j}$. Let $D_G$ be the graphical seed set defined in (1). We want to show $P = D_G$. Let $v \notin D_G$. Then there is a $S \in \mathcal{S}$ separating $v$ from $D$, and we choose $S$ such that $v$ and $S$ are connected in $G$. Note that this is always possible for any $v \notin D_G$. Let $C$ be a clique containing $v$ and $S$. Then $S$ must also be separating $C \setminus S$ and $D$. Using the properties of conditional independence and its connection to the graph separations, we have

$$\mathcal{L}(X_{C \setminus S} \mid X_S) = \mathcal{L}(X_{C \setminus S} \mid X_{D \cup S})$$

for any $X_V$ Markov with respect to $G$. Since $D \cup S$ is a seed set, the distribution $\mathcal{L}(X_{C \setminus S} \mid X_S)$ is the same in two conditions and the associated null hypothesis is true leading to $d_{i,j}^* = 0$ for some $i, j = 1, \dots, k$. We therefore have $v \notin P$. All the steps relied on equivalence relations and thus $P = D_G$. □

The above proposition gives an oracle procedure for recovering the graphical seed set from the knowledge of the two joint distributions. In practice, we need to rely on statistical tests. Let $\phi_i = (\phi_{i,1}, \dots, \phi_{i,k}) \in \{0,1\}^k$ be a vector indicating the results of the statistical tests performed in the $i$-th decomposition, $i = 1, \dots k$, with $\phi_{i,j} = 1$ when the hypothesis $H_{i,j}$ is rejected, and $\phi_{i,j} = 0$ otherwise. The following definition naturally follows.

**Definition 3** (Graphical seed set estimator)**.** The random set $\hat{D}_G$, defined as

$$\hat{D}_G = \bigcap_{i=1}^{k} \bigcup_{\{j:\, \phi_{i,j}=1\}} C_{i,j} \tag{5}$$

is an estimator of $D_G$.

*Remark.* As already stated, the number of potential seed sets, growing exponentially with the dimension $p$, renders the estimation of the minimal seed set computationally expensive. In contrast, thanks to Theorem 1, estimating the graphical seed set requires computing at most $2k-1$ test statistics in the marginal models induced by cliques and separators. Furthermore, no conditional distribution is actually estimated: all test statistics pertaining to conditional distributions are computed from those of appropriate marginal models.

## 5.1 Asymptotic behavior

Estimator $\hat{D}_G$ is different from classical estimators in that its values depend on data through the results of sequences of tests. Properties of the estimator will ultimately depend on the properties of the tests which are used. A treatment of these properties in the limit of infinite data benefits from the introduction of a more general notion of consistency of tests, that we give in general terms as follows.

**Definition 4.** A sequence of tests $\phi(n)$ for the hypothesis $H : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ is consistent if for each $\theta \in \Theta$ there exists a sequence of significance levels $\alpha_n$ s.t.

(1) for each $\theta \in \Theta_0$, $\lim_{n\to\infty} \mathbb{P}_\theta(\phi(n) = 1) = 0$;

(2) for each $\theta \in \Theta_1$, $\lim_{n\to\infty} \mathbb{P}_\theta(\phi(n) = 0) = 0$.

In other words, a sequence of tests is consistent if, at least asymptotically, it reports a correct decision.

Let us now consider testing $H_{i,j}$ in the above given framework. Let $n = n_1 + n_2$ and assume that as $n \to \infty$, $n_l/n \to \gamma_l$ such that $0 < \gamma_l < 1,\ l = 1, 2$, and $\gamma_1 + \gamma_2 = 1$. Moreover, let the test statistic $\phi_{i,j}(n)$ be defined as

$$\phi_{i,j}(n) = \begin{cases} 0 & \lambda_{i,j;n} < q_n \\ 1 & \lambda_{i,j;n} > q_n \end{cases}$$

where $\lambda_{i,j;n}$ is the log likelihood ratio for $H_{i,j}$ and $q_n$ a suitable sequence of quantiles. Standard results assure that, under the null hypothesis, the sequence $\lambda_{i,j;n}$ converges to a chi-square distribution with $f$ degrees of freedom, where $f$ is the difference between the dimensions of the unrestricted parameter space and the restricted parameter space implied by the hypothesis of equality of the distributions of $X_{C_{i,j}\setminus S_{i,j}} \mid X_{S_{i,j}}$ in the two groups. Then, the test that rejects the null hypothesis if $\lambda_{i,j;n}$ exceeds the upper $\alpha$-quantile of the chi-square distribution is asymptotically of level $\alpha$. We can state the following proposition.

**Proposition 2.** In the framework of the problem stated above, for each $H_{i,j}$, there exists a sequence of significance levels $\alpha_n$, s.t. the sequence of tests $\phi_{i,j}(n)$ is consistent.

*Proof.* Choose $\alpha_n = (1 - F_U(n^d))$, with $0 < d < 1/2$, $U \sim \chi^2_f$, and let $q_n = F_U^{-1}(\alpha_n)$. Under the null hypothesis, $\lambda_{i,j;n} \xrightarrow{d} \lambda$, with $\lambda \sim \chi^2_f$. Thanks to the Slutsky theorem, we can write

$$\mathbb{P}_{\theta\in\Theta_0}(\phi_{i,j}(n) = 1) = \mathbb{P}_{\theta\in\Theta_0}\left(\frac{\lambda_{i,j;n}}{n^d} > 1\right) \longrightarrow 0.$$

Furthermore, for each $\theta_1 \in \Theta_1$, it is known that the log likelihood ratio test is degenerate with the order $O(\sqrt{n})$. With the choice of $\alpha_n$ above,

$$\mathbb{P}_{\theta_1}(\phi_{i,j}(n) = 0) = \mathbb{P}_{\theta_1}\left(\frac{\lambda_{i,j;n}}{n^d} < 1\right) \longrightarrow 0, \quad \forall \theta_1 \in \Theta_1.$$

□

**Theorem 2.** *The estimator $\hat{D}_G$ is a pointwise consistent estimator of $D_G$, i.e., $\mathbb{P}_{\theta\in\Theta}(\hat{D}_G = D_G) \to 1$.*

*Proof.* For a fixed $i$, we have that $\phi_i(n) = (\phi_{i,1}(n), \ldots, \phi_{i,k}(n)) \to d_i^* = (d_{i,1}^*, \ldots, d_{i,k}^*)$, since the inequality

$$\mathbb{P}_{\theta\in\Theta}(\phi_i(n) = d_i^*) \geq 1 - \sum_{j=1}^{k} \mathbb{P}_{\theta\in\Theta}(\phi_{i,j}(n) \neq d_{i,j}^*)$$

in conjunction with Proposition 2 implies $\mathbb{P}_{\theta\in\Theta}(\phi_i(n) = d_i^*) \longrightarrow 1$. Convergence of $\hat{D}_G$ to $D_G$ follows straightforwardly.

□

### 5.2 Finite sample behavior

With finite samples, it is customary to assign a bound to the probability of incorrectly rejecting the null hypothesis by imposing conditions such as $\mathbb{P}_{\theta\in\Theta_0}(\phi_{i,j}(n) = 1) \leq \alpha$. Estimation of $D_G$ requires performing a collection of $k + \sum_{i=1}^{k} \nu(C_i)$ tests, where $\nu(C_i)$ denotes the number of separators contained within the clique $C_i$. Finite sample behavior of $\hat{D}_G$ thus hinges on the proper control of the multiplicity issue. If we wish to control the inclusion of false positives in $\hat{D}_G$ by controlling the familywise error rate (FWER), the simplest approach would be to apply the Bonferroni correction with a factor of $k + \sum_{i=1}^{k} \nu(C_i)$. However, the Bonferroni correction can be overly conservative in this situation since intricate logical relations among subsets of hypotheses result in a high positive dependence between the associated $p$-values. To address this issue, we employ the min$P$ method of Westfall and Young (1993), which uses permutations to obtain the joint distribution of the $p$-values and, by accounting for the dependence among $p$-values, attenuates the conservativeness of the Bonferroni procedure. In our setting, the condition of subset pivotality is satisfied, and the Westfall and Young procedure controls the FWER in the strong sense.

## 6 Simulation study

We studied the finite sample behavior of $\hat{D}_G$ with a simulation study. The graph we used, shown in Figure 1, consists of 10 nodes grouped in 5 cliques. We set the parameters of the first, i.e. control, condition in the following way. The means of 10 variables were drawn randomly from a normal distribution centered at 0.5 (standard deviation 1). The matrix with all off-diagonal elements equal to 0.4 and all diagonal elements equal to 1 was modified so that its inverse has zeros corresponding to the missing edges of $G$.

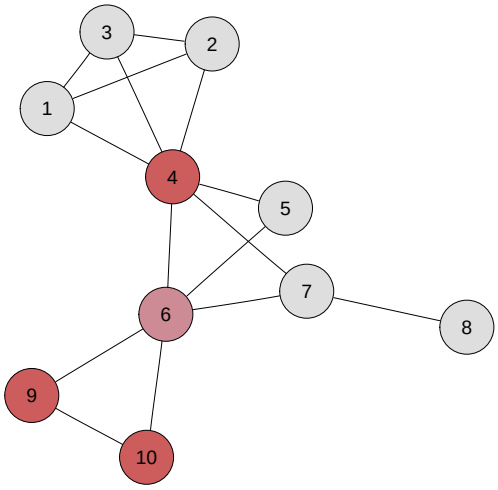


**Figure 1:** An undirected graph used in the simulation study. The minimal seed set is set to $D = \{4, 9, 10\}$, corresponding to the graphical seed set $D_G = \{4, 6, 9, 10\}$.

For the second or post-intervention condition, we considered four different scenarios: one with no intervention, corresponding to the global null hypothesis, and three scenarios with interventions of different strength. Namely, the minimal seed set was set to $D = \{4, 9, 10\}$, and in the post-intervention distribution the mean of the targeted variables was multiplied by a constant $\lambda \in \{1.1, 1.5, 2\}$ corresponding to a mild, moderate, and strong intervention, respectively. The variance of the three seed set variables was also manipulated: it was decreased by 50% in the post-intervention distribution. In this setting, the graphical seed set does not coincide with the minimal seed set since there is no separator of $G$ that separates node 6 from $D$. We thus have $D_G = \{4, 6, 9, 10\}$.

For each combination of the sample size $n = 50, 100, 200$, and the intervention scenario, we simulated 1000 pairs of samples. For each simulated pair, we considered all 5 decompositions of the global null hypothesis, and computed the estimate $\hat{D}_G$. The FWER was controlled at 5% by the min$P$ method (Westfall and Young, 1993) with $B = 500$ permutations. We have thus relied on permutation, rather than asymptotic $p$-values. To evaluate the performance of our procedure, we looked at the number of times the estimated seed set $\hat{D}_G$ coincided with the true seed set $D_G$. The results are shown in Figure 2.

We see that under the global null hypothesis the true seed set, $D_G = \varnothing$, is correctly identified approximately 99% of times, irrespective of the sample size. This is a consequence of the fact that by controlling the FWER, we are controlling the probability of including false positives in $\hat{D}_G$. On the other hand, not surprisingly, the performance under the alternative hypothesis depends on the strength of the intervention. When the intervention is strong, the power of the employed tests approaches 1 even for the smallest sample size ($n = 50$), and the seed set is identified correctly more than 76% of the times. When the intervention is weak, the power to detect it is low, and larger sample sizes are needed. This is evident from the mild

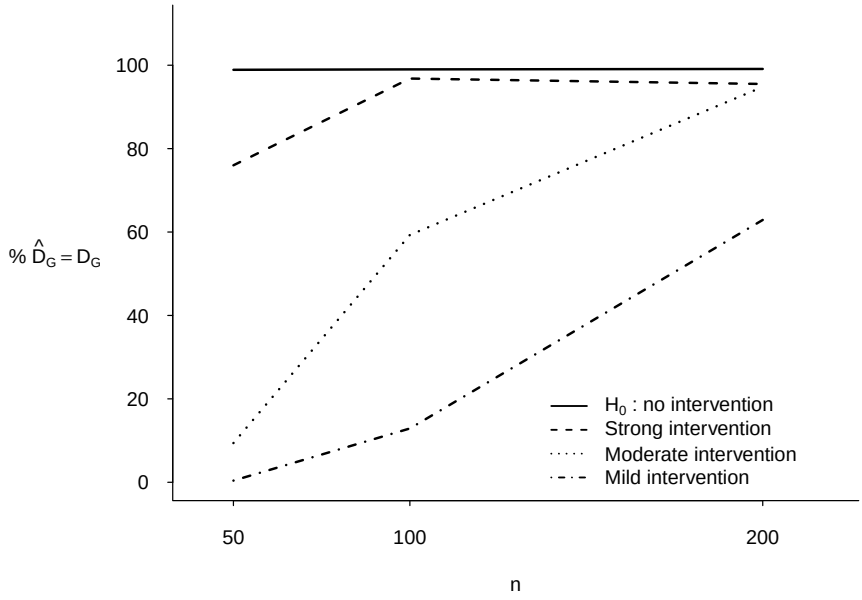


**Figure 2:** Simulation study: The relative frequency of the correct identification of the graphical seed set for different sample sizes under four different scenarios: no intervention ($H_0$), mild, moderate and strong intervention.

intervention setting, where even for $n = 200$, the seed set was identified correctly only 62.9% of times: in approximately one third of instances the estimated seed set was $\{6, 9, 10\}$ which is a subset of the true graphical seed set. This is a consequence of our choice to control the inclusion of false positives: in case of low power, we are bound to obtain an estimate which is a subset of the true seed set.

## 7 Chronic myeloid leukemia

We pick up on the example introduced in Section 2. Following Martini et al. (2013), we focus our attention on genes participating in the Chronic myeloid leukemia pathway, reproducing the preprocessing used in the paper. In particular, to derive the underlying undirected graph, we used the `R` package `graphite` (Sales et al., 2016), which transforms KEGG pathways into graph objects. We moralized and triangulated this graph to obtain a decomposable graph. For graph operations, we relied on the package `gRbase` (Dethlefsen and Højsgaard, 2005). The obtained graph consists of three connected components, and for illustration purposes, we restricted our attention to the largest connected component, consisting of 28 nodes and 16 cliques, shown in Figure 3 (colors can be ignored for now). The number associated to each node is a unique gene identifier from the Entrez Gene database at the National Center for Biotechnology Information (Maglott et al., 2005). Note that nodes 25 and

**Table 1:** Chronic myeloid leukemia dataset: decomposition of a two sample testing problem. Tests for which the null hypothesis was rejected are highlighted.

| No. | Test | $p-$value | No. | Test | $p-$value |
|---|---|---|---|---|---|
| 1 | 1147, 207, 3551 | 0.54 | 21 | 207, 5294, 5295, 8503 | 0.20 |
| 2 | 4790, 4792 \| 1147, 3551 | 0.01 | 22 | 1398, 1399, 5294, 5295, 8503, 867, 9846 | 0.88 |
| 3 | 4193 \| 207 | 0.03 | 23 | 207 \| 5294, 5295, 8503 | 0.09 |
| 4 | 5294, 5295, 8503 \| 207 | 0.20 | 24 | 1398, 1399, 25, 613, 867, 9846 | $< 0.01$ |
| 5 | 1398, 1399, 867, 9846 \| 5294, 5295, 8503 | 0.92 | 25 | 5294, 5295, 8503 \| 1398, 1399, 867, 9846 | 0.64 |
| 6 | 25, 613 \| 1398, 1399, 867, 9846 | $< 0.01$ | 26 | 25, 2885, 613, 9846 | $< 0.01$ |
| 7 | 2885 \| 25, 613, 9846 | 0.95 | 27 | 1398, 1399, 867 \| 25, 613, 9846 | 0.44 |
| 8 | 6776 \| 25, 613 | $< 0.01$ | 28 | 25, 613, 6776 | $< 0.01$ |
| 9 | 6777 \| 25, 613 | 0.93 | 29 | 1398, 1399, 867, 9846 \| 25, 613 | 0.34 |
| 10 | 25759 \| 25, 613 | 0.82 | 30 | 25, 613, 6777 | $< 0.01$ |
| 11 | 4609 \| 25, 613 | 0.21 | 31 | 25, 25759, 613 | $< 0.01$ |
| 12 | 6654, 6655 \| 2885 | 0.45 | 32 | 25, 4609, 613 | $< 0.01$ |
| 13 | 3265, 3845, 4893 \| 6654, 6655 | 0.96 | 33 | 2885, 6654, 6655 | 0.60 |
| 14 | 369 \| 3265, 3845, 4893 | 0.54 | 34 | 25, 613, 9846 \| 2885 | $< 0.01$ |
| 15 | 5894 \| 3265, 3845, 4893 | 0.49 | 35 | 3265, 3845, 4893, 6654, 6655 | 0.87 |
| 16 | 7157 \| 4193 | 0.16 | 36 | 2885 \| 6654, 6655 | 0.96 |
| 17 | 1147, 3551, 4790, 4792 | 0.05 | 37 | 3265, 369, 3845, 4893 | 0.66 |
| 18 | 207 \| 1147, 3551 | 0.40 | 38 | 6654, 6655 \| 3265, 3845, 4893 | 0.91 |
| 19 | 207, 4193 | 0.05 | 39 | 3265, 3845, 4893, 5894 | 0.63 |
| 20 | 1147, 3551 \| 207 | 0.59 | 40 | 4193, 7157 | 0.01 |
| | | | 41 | 207 \| 4193 | 0.47 |

613 represent, ABL and BCR genes, respectively.

The hypothesis, shown in (2), of equality of distributions in the two groups is rejected by the likelihood ratio test ($p$-value $= 3.65 \times 10^{-8}$). This is, of course, expected, since the two groups are defined on the basis of differences in genes 25 and 613. To see whether these differences are propagated over to the other genes, we can perform a test of equality of conditional distributions of the remaining genes given the central two. The obtained $p$-value, $6.65 \times 10^{-3}$, suggests rejecting the hypothesis of equality. We therefore decomposed the graph into a succession of cliques, in order to estimate the underlying graphical seed set.

In this case, there are 16 cliques, and thus 16 decompositions of the global null hypothesis. Across different decompositions, there are 41 unique local hypotheses. We controlled the FWER at 5% level by the min$P$ method with $B = 3000$ permutations. Obtained $p$-values are shown in Table 1, in which tests whose null hypothesis is rejected are highlighted (the threshold found by min$P$ method was $2.3 \times 10^{-3}$). The results of these tests are then combined according to (5), and the result is represented in Figure 3. Highlighted nodes (either gray or red) belong to cliques that result significantly different in two conditions, while the red nodes form the estimated graphical seed set $\hat{D}_G = \{25, 613, 6776\}$. These three genes, thus, explain the marked difference between the two groups, but their effect does not seem to propagate towards other genes in the network (the majority of white nodes in Figure 3).

## 8 Discussion

Motivated by the differential analysis of gene expression data, we have proposed a method for identifying the set of variables driving the difference between two

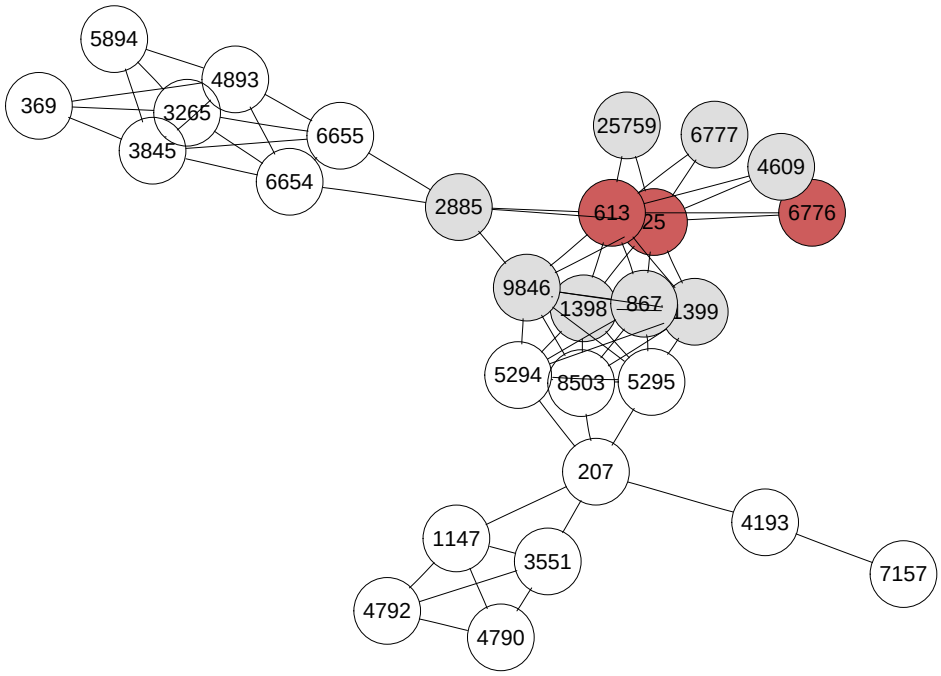


**Figure 3:** An undirected graph representing the Chronic myeloid leukemia pathway. Genes belonging to cliques for which the hypothesis of equality of distributions is rejected are highlighted. Genes belonging to the estimated graphical seed set are colored red.

multivariate normal distributions Markov with respect to the same graph. Our approach compares both marginal and conditional distributions between the two conditions, and uses the resulting evidence to infer the seed set – a set of variables consisting of potential sources of differential behavior. Such an approach would in general require an exponential number of tests, but the modularity of graphical models and the focus on the graphical seed set allow for a linear solution.

The proposed approach assumes that the graphical structure is known a priori which might be uncommon in practice. To relax this assumption, one could resort to a sample splitting strategy, in which one half of the data is used for the estimation of the graphical structure, while the second half is used for hypothesis testing (see for instance Städler and Mukherjee, 2015). However, estimation of the graphical structure is a highly complex inferential task, and one should aim to make use of subject matter knowledge whenever possible.

The basic building block of our method is the likelihood ratio statistic. Maximum likelihood estimates exist if and only if $\min\{n_1, n_2\} > \max_{i=1,\dots,k} |C_i|$, which implies that the method is applicable when the largest clique of the underlying graph is small enough with respect to the sizes of the two samples. Note that this also includes cases for which $\max\{n_1, n_2\} < p$, as long as cliques are sufficiently small.

We focused our attention on the hypothesis of equality of distributions. Anal-

ogous considerations apply to the hypotheses of equality of means, given that the covariance matrices are the same, i.e. $H : \mu^{(1)} = \mu^{(2)}$ given $\Sigma^{(1)} = \Sigma^{(2)}$, and the equality of covariance matrices $H : \Sigma^{(1)} = \Sigma^{(2)}$. In case of the latter hypothesis, special care is needed when addressing the multiple testing issue: observations are no longer exchangeable under the null hypothesis rendering the permutation approach not applicable.

The idea of decomposing the global null hypothesis in a sequence of independent tests, similar to the one in Section 4, can be found in Anderson (2003, p. 423). Although there the goal was different – testing the global hypothesis of equality – the decomposition proposed here can be seen as an extension to Gaussian graphical models, where cliques play the role of individual variables, and the structure of conditional independence is used to reduce the dimensionality of the testing problem.

# A Basics in graphical models

Here, we briefly review key notions regarding Gaussian graphical models, relevant for our work. For a detailed exposition, see Lauritzen (1996).

Consider an undirected graph $G = (V, E)$ where $V$ is a set of nodes and $E$ is a set of edges. A subset of vertices $A$ defines an induced subgraph $G_A = (A, E \cap A \times A)$. A subgraph is said to be complete if all pairs of its vertices are connected in $G$. A clique is a maximal complete subgraph, that is, it is not a subset of any other complete subgraph.

Two disjoint subsets $A, B \subset V$ are said to be *separated* by a subset $S$ (disjoint from $A$ and $B$) if all paths from $A$ to $B$ contain vertices from $S$.

A graph $G$ is decomposable if and only if the set of cliques of $G$ can be ordered so as to satisfy the *running intersection property*, that is, for every $i = 2, \ldots, k$

$$\text{if} \quad S_i = C_i \cap \bigcup_{j=1}^{i-1} C_j \quad \text{then} \quad S_i \in C_l \quad \text{for some} \quad l < i - 1.$$

Although this ordering is generally not unique, the structure of the graph $G$ uniquely determines the set of cliques $\{C_1, \ldots, C_k\}$ and the set of *separators* $\{S_2, \ldots, S_k\}$. For ease of notation, it is often set $S_1 = \varnothing$, so that the set of separators becomes $\{S_1, \ldots, S_k\}$.

For simplicity, we consider only graphs consisting of a single connected component, although most of the presented notions remain valid for more general graphs. We also restrict our attention to decomposable graphs, and this assumption is central to our approach. We assume throughout that cliques have been ordered in an order satisfying the running intersection property. Since, in the following, we deal with different partitions of the set of vertices, we note that such an ordering naturally leads to several *partitions* of $V$. Recall that $(A, S, B)$ is said to be a partition of $V$ if $A, S$ and $B$ are disjoint and $V = A \cup S \cup B$. Partitions of $V$ that correspond to *decompositions* of the graph $G$ are of particular interest. For a graph $G = (V, E)$, a partition $(A, S, B)$ of $V$ is a decomposition of $G$ if $A$ and $B$ are separated by $S$ in $G$, and $S$ is complete.

Denote $p = |V|$ and let $X \sim \mathsf{N}(\mu, \Sigma)$ be a $p$-variate normal random vector indexed by vertices of $G$. If $\Sigma$ is invertible and such that its inverse, $K = \Sigma^{-1}$, has zeroes corresponding to missing edges of $G$, we say that $X$ is a Gaussian graphical model. Let $S^+(G)$ denote the set of all symmetric $p \times p$ positive definite matrices with zeros corresponding to the missing edges of $G$. Moreover, for $A \subset V$, let $\Sigma_A$ denote the corresponding block submatrix of $\Sigma$. In Gaussian graphical models, decompositions of the graph $G$ correspond to special properties of the induced statistical models and associated inference procedures, as we will review in what follows.

Consider first the parameter $\theta = (\mu, \Sigma)$ of the model. If $(A, S, B)$ is a decomposition of $G$, then $X$ can partitioned as $(X_A, X_S, X_B)$, where $X_A \perp\!\!\!\perp X_B \mid X_S$. Here, the conditional laws $\mathcal{L}(X_B \mid X_A, X_S)$ and $\mathcal{L}(X_B \mid X_S)$ coincide and are equal to

$$X_B \mid X_S \sim \mathsf{N}\left[\mu_B + \Sigma_{BS}\Sigma_S^{-1}(X_S - \mu_S), (K_B)^{-1}\right],$$

where $K_B = \left(\Sigma_B - \Sigma_{BS}\Sigma_S^{-1}\Sigma_{SB}\right)^{-1}$. Split $\mu$ into two components, $\mu = (\mu_{A\cup S}, \mu_B)$ and partition $\Sigma$ correspondingly. Then, parameters $\theta_{A\cup S} = (\mu_{A\cup S}, \Sigma_{A\cup S})$ and

$$\theta_{B|S} = \left(\mu_B - \Sigma_{BS}\Sigma_S^{-1}\mu_S, \Sigma_{BS}\Sigma_S^{-1}, K_B^{-1}\right),$$

(i.e., parameters of the marginal law of $(X_A, X_S)$ and of the conditional law of $X_B \mid X_S$) are variation independent (Barndorff-Nielsen, 2014, p.28). It is worth noting that, on exploiting the symmetry of $A$ and $B$ with respect to $S$, we can analogously say that conditional laws $\mathcal{L}(X_A \mid X_B, X_S)$ and $\mathcal{L}(X_A \mid X_S)$ coincide and are equal to

$$X_A \mid X_S \sim \mathsf{N}\left[\mu_A + \Sigma_{AS}\Sigma_S^{-1}(X_S - \mu_S), (K_A)^{-1}\right],$$

where $K_A = \left(\Sigma_A - \Sigma_{AS}\Sigma_S^{-1}\Sigma_{SA}\right)^{-1}$. Accordingly, parameters $\theta_{B\cup S}$ and $\theta_{A|S}$ are variation independent.

Consider now a random sample $X_1, \ldots, X_n$ from the same model and maximum likelihood estimation of $\theta$. To estimate $\theta$, we go through the estimation of $\theta_{A\cup S}$ and $\theta_{B|S}$. It is known that, beside being variation independent, these parameters are also L-independent since the likelihood function is of the product form: $L(\theta_{A\cup S}, \theta_{B|S}) = L(\theta_{A\cup S})L(\theta_{B|S})$, causing the covariance between their maximum likelihood estimators to vanish asymptotically. Therefore, $\hat{\theta}_{A\cup S}$ and $\hat{\theta}_{B|S}$ are asymptotically independent. Note that although this independence is reminiscent of the strong hyper Markov property, it holds only asymptotically, since the distribution of the maximum likelihood estimator is not strong hyper Markov unless $\Sigma$ is diagonal (Dawid and Lauritzen, 1993). However, the sampling distribution of $\hat{\theta}$ defines a *weak hyper Markov law* on the parameter space (Dawid and Lauritzen, 1993). A weak hyper Markov property ensures that separations in the graph, reflected in the distribution of the original variables, are also reflected in the distribution of the maximum likelihood estimator. More precisely, it trivially holds

$$\hat{\mu}_{A\cup S} \perp\!\!\!\perp \hat{\mu}_{B\cup S} \mid \hat{\mu}_S,$$

and, more importantly,

$$\hat{\Sigma}_{A\cup S} \perp\!\!\!\perp \hat{\Sigma}_{B\cup S} \mid \hat{\Sigma}_S.$$

## References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. Wiley, New Jersey.

Barndorff-Nielsen, O. (2014). *Information and exponential families in statistical theory*. John Wiley & Sons, New York.

Chen, S. X., Y.-L. Qin, et al. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics 38*(2), 808–835.

Chiaretti, S., X. Li, R. Gentleman, A. Vitale, K. S. Wang, F. Mandelli, R. Foa, and J. Ritz (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.

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

Del Sol, A., R. Balling, L. Hood, and D. Galas (2010). Diseases as network perturbations. *Current Opinion in Biotechnology 21*(4), 566–571.

Dethlefsen, C. and S. Højsgaard (2005). A common platform for graphical models in R: The gRbase package. *Journal of Statistical Software 14*(17), 1–12.

Dudoit, S. and M. J. van der Laan (2008). Multiple tests of association with biological annotation metadata. In *Multiple Testing Procedures with Applications to Genomics*, pp. 413–476. Springer.

Kanehisa, M., S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa (2009). Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research 38*(suppl_1), D355–D360.

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

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

Li, X. (2009). *ALL: A data package*. R package version 1.16.0.

Maglott, D., J. Ostell, K. D. Pruitt, and T. Tatusova (2005). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research 33*(suppl 1), D54–D58.

Martini, P., G. Sales, M. S. Massa, M. Chiogna, and C. Romualdi (2013). Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Research 41*(1), e19.

Massa, M. S., M. Chiogna, and C. Romualdi (2010). Gene set analysis exploiting the topology of a pathway. *BMC Systems Biology 4*(1), 121.

Sales, G., E. Calura, and C. Romualdi (2016). *graphite: GRAPH Interaction from pathway Topological Environment*. R package version 1.20.1.

Städler, N. and S. Mukherjee (2015). Multivariate gene-set testing based on graphical models. *Biostatistics 16*(1), 47–59.

Tan, W. (1977). On the distribution of quadratic forms in normal random variables. *Canadian Journal of Statistics 5*(2), 241–250.

Thomas, A. and P. J. Green (2009). Enumerating the junction trees of a decomposable graph. *Journal of Computational and Graphical Statistics 18*(4), 930–940.

Westfall, P. H. and S. S. Young (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, Volume 279. John Wiley & Sons, New York.

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

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

## Acknowledgements

The authors wish to thank Gianfranco Adimari for fruitful discussions on a preliminary draft of the paper.

**Working Paper Series**
**Department of Statistical Sciences, University of Padua**

You may order paper copies of the working papers by emailing wp@stat.unipd.it
Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Department of Statistical Sciences
*University of Padua*
*Italy*