Differential Networking with Path Weights in Gaussian Trees

Alberto Roverato ALBERTO.ROVERATO@UNIBO.IT
Università di Bologna, Italy

Robert Castelo ROBERT.CASTELO@UPF.EDU
Universitat Pompeu Fabra, Barcelona, Spain

Abstract

Marginal and partial correlations quantify the strength of the associations represented by the edges of a graphical Gaussian model. The identification of changes in these quantities across different multivariate distributions, defined on the same vector of random variables, is often used to analyze regulatory networks in molecular biology, doing what is popularly known as differential networking, or differential coexpression analysis. However, the strength of associations along the paths of a graphical model has remained largely unexplored in this type of analysis. Here we investigate how to quantify this strength over the paths of a Gaussian tree, leading to a factorization of what we shall call path weights. We show that tree structures allow for an intuitive interpretation of path weights and that the proposed factorization conveys information that is not captured by marginal or partial correlations alone. Path weights can help to improve our understanding of a multivariate system under study and provide a new tool for differential coexpression analysis.

Keywords: graphical Gaussian model; tree; path weight; inflation factor; inverse covariance.

1. Introduction

In many applications, data come from different distributions that share the same variables but differ in their dependence structure. In this case it is often of interest to identify differences in the association patterns of variables. A relevant example in the field of molecular biology is provided by the comparison of gene expression values measured on different experimental conditions. The motivation to follow such approaches is that the identification of changes in gene coexpression patterns provides information that is complementary to the identification of significant changes in mean levels of expression, commonly known as differential expression analysis (de la Fuente, 2010). The existing techniques for differential coexpression analysis involve the comparison, across different experimental conditions, of some measure of coexpression between two or more genes. Coexpression is commonly identified by using marginal measures of association such as Pearson or Spearman correlation coefficients, ignoring the multivariate structure of gene expression levels, which can potentially inform the quantification of coexpression.

Covariance and inverse covariance matrices, jointly with the mean vector, condense the information of multivariate Gaussian distributions. Scaling their rows and columns provides matrices of marginal and partial correlations, which have become the canonical units of interpretation of associations between continuous random variables in a multivariate system. In graphical Gaussian models (Whittaker, 1990; Lauritzen, 1996) an undirected graph is used to represent the association structure of variables as a network, and if a pair of variables is not joined by an edge in the graph, then the corresponding partial correlation is equal to zero. An analysis involving the comparison of networks is commonly known as differential networking and in the context of graphical models this has focused on learning and comparing graph structures (Guo et al., 2011; Danaher et al., 2014).

Although in graphical Gaussian models the structure of the network can be inferred from the zero pattern of the inverse covariance matrix, if the probability distribution of the variables is faithful to the network, then paths along the network connect random variables with non-zero entries in the covariance matrix. Paths of dependence were studied by Wright (1921) in directed graphs, to understand how marginal associations between two random variables decompose through their connecting paths. On the other hand, Jones and West (2005) developed the counterpart of Wright's decomposition, for undirected graphs, in terms of additive weights along undirected paths. In undirected graphical models, little is known about the interpretation of path weights and their relationship with covariance, inverse covariance, marginal and partial correlations. Recently, Roverato and Castelo (2017) provided an interpretation of path weights for the trivial case of single-edge paths in undirected graphical Gaussian models.

In this paper we focus on graphical Gaussian models with tree structure. Tree models can be used to approximate arbitrary undirected graphical models (Edwards et al., 2010; Lauritzen et al., 2018) and represent a starting point in the analysis of more complex graphical structures. The Pearson (marginal) correlation can be regarded as a natural weight to be associated with the path of a Gaussian tree because it can be computed as the product of the correlation coefficients associated with the edges of the path. Here we show that the information provided by such a path weight in a Gaussian tree, can be factorized as the product of a partial correlation and an inflation factor. The separate analysis of these two quantities can help to improve our understanding of a multivariate system under study and provide a new tool for performing differential networking.

The rest of this paper is organized as follows. Section 2 gives the required background on Gaussian tree models and inflation factors. In Section 3 we derive the path weight factorization and show some of its properties, which are illustrated through simulations in Section 4. Finally, Section 5 shows a differential networking analysis with real gene expression data from yeast.

2. Notation and Background

2.1 Graphs and Paths

An *undirected graph* is a pair $\mathcal{G} = (V, \mathcal{E})$, where V is a set of vertices and $\mathcal{E}$ is a set of *edges*, which are unordered pairs of vertices; formally $\mathcal{E} \subseteq V \times V$. The graphs we consider are simple, which implies that they have no multiple edges and have no loops, i.e. $\{v, v\} \notin \mathcal{E}$ for any $v \in V$. A *path* between x and y in $\mathcal{G}$ is a sequence $\pi = \langle x = v_1, \ldots, v_k = y \rangle$ of $k \geq 2$ distinct vertices such that $\{v_i, v_{i+1}\} \in \mathcal{E}$ for every $i = 1, \ldots, k-1$. We denote, respectively, by $V(\pi) \subseteq V$ and $\mathcal{E}(\pi) \subseteq \mathcal{E}$, the set of vertices and edges of the path π. To improve the readability of $V(\pi)$ in sub- and superscripts, we will set $P \equiv V(\pi)$. We write π_{xy} when we want to make more explicit which are the endpoints of the path. We denote by Π_{xy} the collection of all paths from x to y in $\mathcal{G}$. The length of a path, $|\pi_{xy}|$, is defined as the number of vertices forming the path, i.e. $|\pi_{xy}| = k$ for $\pi_{xy} = \langle x = v_1, \ldots, v_k = y \rangle$. A path π_{xy} becomes a *cycle* if the endpoints are allowed to be the same, i.e. $x = y$, and $|\pi_{xy}| \geq 3$. The degree of a vertex $v \in V$, denoted by $\deg(v)$, is the number of edges incident to v, and we define the degree of a path π as $\deg(\pi) = \Sigma_{v \in V(\pi)} \deg(v)$. A *tree* is an undirected graph where there is path between every pair of vertices but where there are no cycles, i.e. $|\Pi_{xy}| = 1$ for every $x, y \in V$, $x \neq y$. In the following, when we want to highlight that a graph is a tree we will write it as $\mathcal{T} = (V, \mathcal{E})$. Although in this paper we focus on trees, results given also hold true when the considered graph is the disjoint union of trees, i.e. it is a *forest*.

2.2 Inflation Factors

Let $X = X_V$ be a vector of continuous random variables indexed by a finite set $V = \{1, \ldots, p\}$ so that for $A \subseteq V$, X_A is the subvector of X indexed by A. The random vector X_V has probability distribution $\mathbb{P}_V$ and covariance matrix $\Sigma = \{\sigma_{uv}\}_{u,v \in V}$. For $x, y \in V$ with $x, y \notin A$ we denote by $\rho_{xy \cdot A}$ the *partial correlation coefficient* of X_x and X_y given X_A, which simplifies to the (marginal) correlation coefficient ρ_{xy} when $A = \emptyset$. In the literature, different quantities have been introduced in order to provide a generalization of the concept of (partial) correlation from pairs of variables to pairs of vectors; see Mardia et al. (1979, Section 6.5.4), Timm (2002, p. 485) and Kim and Timm (2006, Section 5.6) for a review of measures of correlation between vectors. Here we consider a coefficient, introduced by Rozeboom (1965). This turns out to be a building block for the computation of inflation factors, which arise naturally in the theory developed in this paper. Formally, if $B \subseteq V$, with $A \cap B = \emptyset$, the *vector correlation* of X_A and X_B is defined as $\rho_{(A)(B)} = \sqrt{1 - \lambda_{(A)(B)}}$, where $\lambda_{(A)(B)} = |\Sigma_{A \cup BA \cup B}| / \left(|\Sigma_{AA}| \times |\Sigma_{BB}|\right)$ is the *vector alienation coefficient* of Hotelling (1936). It is straightforward to check that when $A = \{x\}$, $\rho_{(A)(B)} = \rho_{(x)(B)}$ coincides with the *multiple correlation* of X_x on X_B, so that if also $B = \{y\}$ then $\rho^2_{(A)(B)} = \rho^2_{xy}$ (see also Timm, 2002, p. 485). We remark that the covariance matrices we consider are assumed to be positive definite so that $0 \leq \rho_{(A)(B)} < 1$; furthermore, we use the convention that $\lambda_{(A)(B)} = 1$, and therefore $\rho_{(A)(B)} = 0$, whenever either $A = \emptyset$ or $B = \emptyset$. Throughout this paper many quantities involve the computation of the determinant of a matrix whose rows and columns are indexed by a subset of V. When such a subset is empty we use the convention that the determinant is equal to one.

Linear regression diagnostics use a quantity called the *variance inflation factor* to help detect multicollinearity. More specifically, the variance inflation factor of X_v on $X_{V \setminus \{v\}}$ is defined as $\mathrm{IF}_v = 1/(1 - \rho^2_{(v)(V \setminus \{v\})})$ (see Belsley et al., 2005; Chatterjee and Hadi, 2012). The variance inflation factor equals 1 when X_v and $X_{V \setminus \{v\}}$ are uncorrelated so that $\rho_{(v)(V \setminus \{v\})} = 0$; otherwise $\mathrm{IF}_v > 1$ and its value increases as $\rho_{(v)(V \setminus \{v\})}$ increases. Here, we extend the definition of the variance inflation factor to a pair of subsets $A, B \subseteq V$, with $A \cap B = \emptyset$, as follows

$$\mathrm{IF}^B_A = \frac{1}{1 - \rho^2_{(A)(B)}} \tag{1}$$

and call this quantity the *inflation factor* of A on B, written as IF_A when $B = V \setminus A$. Note that the inflation factor is a monotonically increasing function of the vector correlation. However, while the vector correlation takes values between zero and one, it holds that $\mathrm{IF}^B_A \geq 1$. The name inflation factor comes from the fact that this quantity is used as a multiplicative term to "increase" the value of other quantities. It is desirable that the vector correlation, and therefore the inflation factor, should not decrease if more variables are added to either A or B. The following lemma shows that this basic requirement is satisfied.

Lemma 1 *Let Σ be the covariance matrix of a random vector X_V. If $A' \subseteq A \subseteq V$ and $B \subseteq V$ is a subset of V such that $A \cap B = \emptyset$ then it holds that $\mathrm{IF}^B_{A'} \leq \mathrm{IF}^B_A$.*

Proof We first notice that if we let $\Sigma_{AA \cdot B} = \Sigma_{AA} - \Sigma_{AB}\Sigma^{-1}_{BB}\Sigma_{BA}$ then $\mathrm{IF}^B_A = |\Sigma_{AA}||\Sigma_{BB}|/|\Sigma| = |\Sigma_{AA}|/|\Sigma_{AA \cdot B}|$, and that if we set $A'' = A \setminus A'$, it follows from the Schur's determinant identities that $|\Sigma_{AA}| = |\Sigma_{A''A'' \cdot A'}||\Sigma_{A'A'}|$ and that $|\Sigma_{AA \cdot B}| = |\Sigma_{A''A'' \cdot A' \cup B}||\Sigma_{A'A' \cdot B}|$. Hence, we have, $\mathrm{IF}^B_A = (|\Sigma_{A''A'' \cdot A'}||\Sigma_{A'A'}|)/(|\Sigma_{A''A'' \cdot A' \cup B}||\Sigma_{A'A' \cdot B}|) = (|\Sigma_{A''A'' \cdot A'}|)/(|\Sigma_{A''A'' \cdot A' \cup B}|) \times \mathrm{IF}^B_{A'}$, and the result follows because $|\Sigma_{A''A'' \cdot A'}| \geq |\Sigma_{A''A'' \cdot A' \cup B}|$. ■

2.3 Concentration Graph Models

The *concentration* (or precision) matrix $K = \{\kappa_{uv}\}_{u,v\in V}$ of the random vector X_V is the inverse of its covariance matrix, that is $K = \Sigma^{-1}$. As in Σ, rows and columns of K are indexed by V and we say that K is *adapted* to a graph $\mathcal{G} = (V, \mathcal{E})$ if for every $\kappa_{uv} \neq 0$, with $u \neq v$, it holds that $\{u, v\} \in \mathcal{E}$. The *concentration graph model* (Cox and Wermuth, 1996) with graph $\mathcal{G} = (V, \mathcal{E})$ is the family of multivariate normal distributions whose concentration matrix is adapted to $\mathcal{G}$. This model has also been called a *covariance selection model* (Dempster, 1972) and a *graphical Gaussian model* (Whittaker, 1990); we refer the reader to Lauritzen (1996) for details and discussion.

For $A, B \subseteq V$ with $A \cap B = \emptyset$ the *partial covariance* matrix $\Sigma_{AA\cdot B} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$ is the covariance matrix of $X_A|X_B$, that is the residual vector deriving from the linear least square predictor of X_A on X_B (see Whittaker, 1990, p. 134). We denote by $\sigma_{uv\cdot B}$, for $u, v \in A$, the entries of $\Sigma_{AA\cdot B}$ and recall that, in the Gaussian case, $\Sigma_{AA\cdot B}$ coincides with the covariance matrix of the conditional distribution of X_A given X_B. Note that we use the convention that $\Sigma_{AA}^{-1} = (\Sigma_{AA})^{-1}$ and, similarly, $\Sigma_{AA\cdot B}^{-1} = (\Sigma_{AA\cdot B})^{-1}$. Furthermore, if $\bar{A} = V \setminus A$ is the complement of a subset A with respect to V, then it follows from the rule for the inversion of a partitioned matrix that $\Sigma_{AA\cdot\bar{A}}^{-1} = K_{AA}$ and, accordingly, $\Sigma_{AA}^{-1} = K_{AA\cdot\bar{A}}$. We will focus on the case where the concentration matrix K is adapted to a tree $\mathcal{T} = (V, \mathcal{E})$, which we shall also call the *concentration tree model*.

3. Path Weights in Concentration Tree Models

In the framework of concentration tree models, a key role is played by the relationship existing between the correlation ρ_{xy} of two variables X_x and X_y, and the (unique) path π between x and y in $\mathcal{T}$. More specifically, ρ_{xy} can be factorized into the product of the marginal correlations along the edges of π in $\mathcal{T}$, as follows (see, e.g., Choi et al., 2011; Zwiernik, 2015),

$$\rho_{xy} = \prod_{\{u,v\}\in\mathcal{E}(\pi)} \rho_{uv}. \tag{2}$$

For this reason, the (marginal) correlation ρ_{xy} can be regarded as a natural weight to be associated with the path of a Gaussian tree.

Example 1 *Consider the trees $\mathcal{T}_a$ and $\mathcal{T}_b$ in Figure 1, and assume that in the respective concentration graph models the correlations for every pair of variables joined by an edge is constant and equal to $\rho = 0.5$. It follows that the correlation associated with a path π_{xy} is $\rho_{xy} = 0.5^{|\pi_{xy}|}$ both in $\mathcal{T}_a$ and $\mathcal{T}_b$ so that, for instance, because $\pi_{23}^a = \pi_{23}^b = \langle 2, 1, 3\rangle$ then , $\rho_{23}^a = \rho_{23}^b = 0.25$. Furthermore, the fold changes between these quantities for the paths π_{xy}^a and π_{xy}^b are always of the form 2^l, where $l = |\pi_{xy}^a| - |\pi_{xy}^b|$.*

Here we show that ρ_{xy} can be also written as the product of a partial correlation and an inflation factor, and describe the role played by these two quantities in the analysis of the information encoded by a path of a tree. If K is adapted to a tree $\mathcal{T} = (V, \mathcal{E})$ and π_{xy} is the unique path between x and y in $\mathcal{T}$, then we define the inflation factor associated with the path π_{xy} as follows,

$$\mathrm{IF}_{xy}^{\mathcal{T}} = \frac{\mathrm{IF}_P}{\sqrt{\mathrm{IF}_x^{\bar{P}} \times \mathrm{IF}_y^{\bar{P}}}}, \tag{3}$$

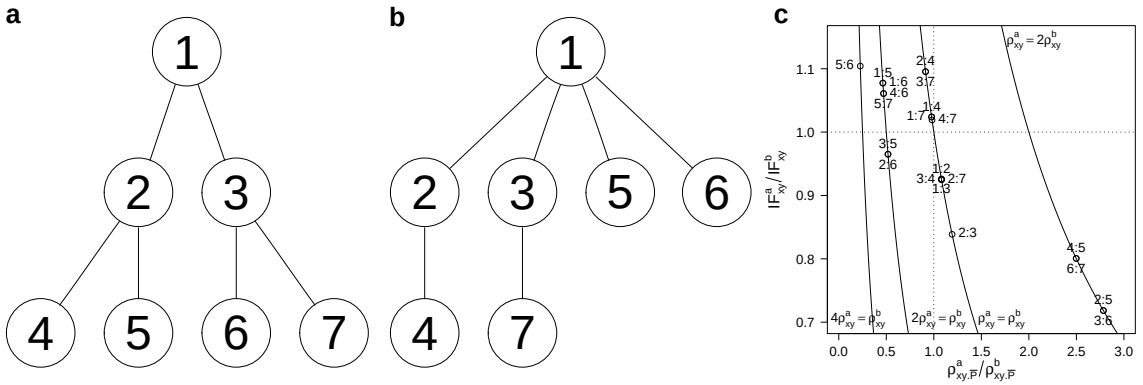

Figure 1: Comparison of two small trees. Ratios in the x and y-axis of panel (c) are calculated from trees $\mathcal{T}_a$ in (a) and $\mathcal{T}_b$ in (b). Numbers $x:y$ in (c) refer to the endpoints of paths π_{xy} in $\mathcal{T}_a$ and $\mathcal{T}_b$. Solid lines in (c) correspond to values in x and y-axis for which the ratio of marginal correlations ρ^a_{xy}/ρ^b_{xy} is constant.

where $P = V(\pi)$ and $\bar{P} = V \setminus P$. The application of Lemma 1 to (3) shows that $\mathrm{IF}^{\mathcal{T}}_{xy}$ is a proper inflation factor in the sense that it always takes values greater or equal to one.

Proposition 2 *Let π be the unique path between x and y in the tree $\mathcal{T} = (V, \mathcal{E})$ and let K be the concentration matrix of X_V such that K is adapted to $\mathcal{T}$. Then, it holds that* $\mathrm{IF}^{\mathcal{T}}_{xy} \geq 1$.

Proof This follows immediately from Lemma 1 because $x, y \in P$ so that it holds both that $\mathrm{IF}_P \geq \mathrm{IF}^{\bar{P}}_x$ and $\mathrm{IF}_P \geq \mathrm{IF}^{\bar{P}}_y$. This implies $(\mathrm{IF}_P)^2 \geq \mathrm{IF}^{\bar{P}}_x \times \mathrm{IF}^{\bar{P}}_y$ and therefore $\mathrm{IF}_P \geq (\mathrm{IF}^{\bar{P}}_x \times \mathrm{IF}^{\bar{P}}_y)^{\frac{1}{2}}$. ■

As well as IF_P, also $\mathrm{IF}^{\mathcal{T}}_{xy}$ is a measure of the association between X_P and $X_{\bar{P}}$. However, $\mathrm{IF}^{\mathcal{T}}_{xy}$ in (3) is computed by dividing IF_P by $(\mathrm{IF}^{\bar{P}}_x \times \mathrm{IF}^{\bar{P}}_y)^{\frac{1}{2}}$ with the consequence that $\mathrm{IF}^{\mathcal{T}}_{xy} \leq \mathrm{IF}_P$.

Example 2 *In the setting of Example 1, the path $\pi_{23} = \langle 2, 1, 3 \rangle$ is common to the two trees $\mathcal{T}_a$ and $\mathcal{T}_b$ and furthermore, $\rho^a_{23} = \rho^b_{23} = 0.25$. On the other hand, the two inflation factors are different because* $\mathrm{IF}^{\mathcal{T}_a}_{23} = 1.625$ *whereas* $\mathrm{IF}^{\mathcal{T}_b}_{23} = 1.938$ *thereby reflecting the different ways the two paths are joined with the remaining vertices in the respective trees.*

In a path, a different role is played by the endpoint vertices with respect to the inner vertices. In IF_P all the vertices of the path are considered to be on an equal footing whereas in the computation of $\mathrm{IF}^{\mathcal{T}}_{xy}$ the endpoints of the paths are considered explicitly and the value of IF_P is adjusted so as to reduce the relevance played by the endpoint vertices. The definition of the inflation factor $\mathrm{IF}^{\mathcal{T}}_{xy}$ associated with a path of a Gaussian tree $\mathcal{T}$ leads to the main result of this paper.

Theorem 3 *Let π be the unique path between x and y in the tree $\mathcal{T} = (V, \mathcal{E})$ and let K be the concentration matrix of X_V such that K adapts to $\mathcal{T}$. The correlation ρ_{xy} between X_x and X_y, corresponding to the weight of the path π, can be decomposed as,*

$$\rho_{xy} = \rho_{xy \cdot \bar{P}} \times \mathrm{IF}^{\mathcal{T}}_{xy}\,, \tag{4}$$

where $P = V(\pi)$.

Proof Firstly, we notice that $\Sigma_{PP\cdot\bar{P}} = K_{PP}^{-1}$ so that $\sigma_{xy\cdot\bar{P}} = c_{xy}/|K_{PP}|$ where c_{xy} is the (x,y)-cofactor of K_{PP}. Since K is adapted to $\mathcal{T}$, then if we order the rows and columns of K_{PP} from x to y, following the ordering of vertices on π, then K_{PP} is a tridiagonal matrix and it is straightforward to see that the (x,y)-cofactor of K_{PP} is equal to $c_{xy} = (-1)^{|P|+1}\prod_{\{u,v\}\in\mathcal{E}(\pi)}\kappa_{uv}$. Hence,

$$\sigma_{xy\cdot\bar{P}} = \frac{(-1)^{|P|+1}\prod_{\{u,v\}\in\mathcal{E}(\pi)}\kappa_{uv}}{|K_{PP}|}. \tag{5}$$

Secondly, we have

$$\mathrm{IF}_P = \frac{|\Sigma_{PP}|}{|\Sigma_{PP|\bar{P}}|} = \frac{|K_{PP|\bar{P}}|^{-1}}{|K_{PP}|^{-1}} = \frac{|K_{PP}||K_{\bar{P}\bar{P}}|}{|K_{PP|\bar{P}}||K_{\bar{P}\bar{P}}|} = \frac{|K_{PP}||K_{\bar{P}\bar{P}}|}{|K|}. \tag{6}$$

To obtain the desired result we exploit the formula of Jones and West (2005) for the covariance decomposition over the path of an undirected graph introduced. More precisely, Theorem 1 of Jones and West (2005) shows that one can write

$$\sigma_{xy} = \sum_{\pi\in\Pi_{xy}}(-1)^{|P|+1}\,\frac{|K_{\bar{P}\bar{P}}|}{|K|}\prod_{\{u,v\}\in\mathcal{E}(\pi)}\kappa_{uv}$$

that in the case where there is only one path π between x and y simplifies to

$$\begin{aligned}\sigma_{xy} &= (-1)^{|P|+1}\,\frac{|K_{\bar{P}\bar{P}}|}{|K|}\prod_{\{u,v\}\in\mathcal{E}(\pi)}\kappa_{uv}\\ &= \frac{(-1)^{|P|+1}\prod_{\{u,v\}\in\mathcal{E}(\pi)}\kappa_{uv}}{|K_{PP}|}\times\frac{|K_{PP}||K_{\bar{P}\bar{P}}|}{|K|} = \sigma_{xy\cdot\bar{P}}\times\mathrm{IF}_P,\end{aligned} \tag{7}$$

where (7) follows from (5) and (6). We can then divide both sides of (7) by $(\sigma_{xx}\,\sigma_{yy})^{\frac{1}{2}}$ and then multiply and divide the right hand side by $(\sigma_{xx\cdot\bar{P}}\,\sigma_{yy\cdot\bar{P}})^{\frac{1}{2}}$ to obtain $\rho_{xy} = \rho_{xy.\bar{P}}\times IF_P\times\sqrt{(\sigma_{xx\cdot\bar{P}}\,\sigma_{yy\cdot\bar{P}})/(\sigma_{xx}\,\sigma_{yy})} = \rho_{xy\cdot\bar{P}}\times\mathrm{IF}^{\mathcal{T}}_{xy}$, as required. ■

The identity in Equation (4) shows that ρ_{xy} can be decomposed into the product of $\rho_{xy\cdot\bar{P}}$ and $\mathrm{IF}^{\mathcal{T}}_{xy}$. The first term is the correlation between X_x and X_y computed after these two variables have been linearly adjusted for the variables outside the path in the network. On the other hand, $\mathrm{IF}^{\mathcal{T}}_{xy}$ is a measure of the strength of the association between the path and the rest of the network. The stronger the association of the path with the remaining variables the larger the inflation factor.

Example 3 *Let π_{xy} and π_{uv} be two paths such that the vertices of π_{xy} are disconnected from the rest of the network (so that $\mathcal{T}$ is in fact a forest) whereas the vertices of π_{uv} are highly connected with the rest of the network. Clearly, the path π_{xy} plays a different role in the multivariate system with respect to π_{uv}, but the values taken by the two correlation coefficients ρ_{xy} and ρ_{uv} fail to highlight this feature. Equation (4) clarifies this aspect by computing the correlation coefficient as an inflated partial correlation because $\mathrm{IF}_{xy} = 1$ whereas $\mathrm{IF}_{uv} > 1$.*

Example 4 *Consider again the setting of Example 2. In both trees it holds that $P = V(\pi_{23}) = \{2,1,3\}$, and therefore that $\bar{P} = \{4,5,6,7\}$. Furthermore, the partial correlations in (4) are given*

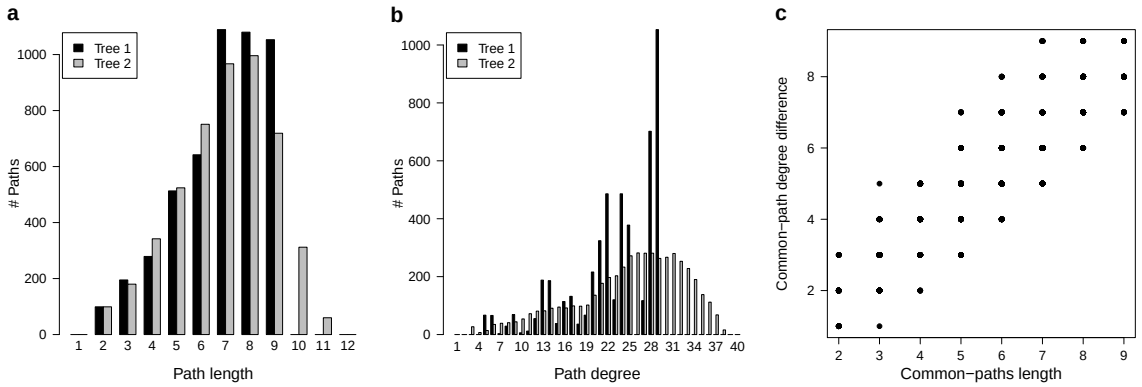

Figure 2: Comparison of structural properties of two large trees. Path length distribution (a); path degree distribution (b); and difference in path degree among the common paths between the two trees, as function of their length (c).

by $\rho^a_{23\cdot\bar{P}} = 0.154$ *and* $\rho^b_{23\cdot\bar{P}} = 0.129$*, respectively, so that* $\rho^a_{23} = \rho^a_{23\cdot\bar{P}} \times \mathrm{IF}^{\mathcal{T}_a}_{23} = 0.154 \times 1.625 = 0.25$ *and* $\rho^b_{23} = \rho^b_{23\cdot\bar{P}} \times \mathrm{IF}^{\mathcal{T}_b}_{23} = 0.129 \times 1.938 = 0.25$*. More generally, Figure 1c shows the ratios of the quantities in the factorization given in (4) and how these quantities change while the ratio of marginal correlations remains constant, even in the case of paths that are common to both trees.*

Interestingly, this decomposition preserves the factorization of the correlation along the path, in the sense that also the partial correlation is the product of partial correlations associated with the edges of the path. This shows that as well as ρ_{xy}, also $\rho_{xy\cdot\bar{P}}$ is a natural measure to be associated with a path in a tree.

Corollary 4 *Under the conditions of Theorem 3 the following relationships hold true*

$$\textit{(i)}\ \ |\rho_{xy}| \geq |\rho_{xy\cdot\bar{P}}|, \quad \textit{(ii)}\ \ \mathrm{sgn}(\rho_{xy}) = \mathrm{sgn}(\rho_{xy\cdot\bar{P}}) \quad \textit{and} \quad \textit{(iii)}\ \ \rho_{xy\cdot\bar{P}} = \prod_{\{u,v\}\in\mathcal{E}(\pi)} \rho_{uv\cdot\bar{P}}.$$

Proof Relationships (i) and (ii) follow from (4) because $\mathrm{IF}^{\mathcal{T}}_{xy} \geq 1$. The equality (iii) follows from (2) by noticing that $K_{PP} = \Sigma^{-1}_{PP\cdot\bar{P}}$ is adapted to the tree $\mathcal{T}_P$ and that π is a path in $\mathcal{T}_P$. ■

We close this section by showing that the relationships between marginal and partial correlations, given in Corollary 4, extend to every pair of variables.

Corollary 5 *Let* K *be the concentration matrix of* X_V*. If* K *is adapted to the tree* $\mathcal{T} = (V, \mathcal{E})$ *then for every* $u, v \in V$ *it holds that* $|\rho_{uv}| \geq |\rho_{uv\cdot V\setminus\{u,v\}}|$*; furthermore,* $\mathrm{sgn}(\rho_{uv}) = \mathrm{sgn}(\rho_{uv\cdot V\setminus\{u,v\}})$ *whenever* $\rho_{uv\cdot V\setminus\{u,v\}} \neq 0$*.*

Proof If the pair $\{u, v\}$ forms an edge in $\mathcal{T}$, i.e. $\{u, v\} \in \mathcal{E}$, then the result follows from Corollary 4 because every edge is a path. If $\{u, v\} \notin E$ then the result is trivially true because $\rho_{uv\cdot V\setminus\{u,v\}} = 0$. ■

4. Simulation Studies

We have conducted simulation studies to illustrate and understand the properties of the quantities described in the previous section. More concretely, we address the following three questions:

- How partial correlations and inflation factors change due to tree structural differences only.
- How partial correlations and inflation factors change due to differences between parameter values with a given tree structure.
- How partial correlations and inflation factors change due to differences in both, structure and parameter values of the two compared trees.

We simulate covariance matrices whose inverse adapts to a given tree $\mathcal{T} = (V, \mathcal{E})$ using the procedure described in (Tur et al., 2014, pg. 1380). Depending on the particular simulation, we will enforce a constant marginal correlation ρ associated with present edges in $\mathcal{T}$ or simulate covariance matrices with a mean correlation ρ among present edges.

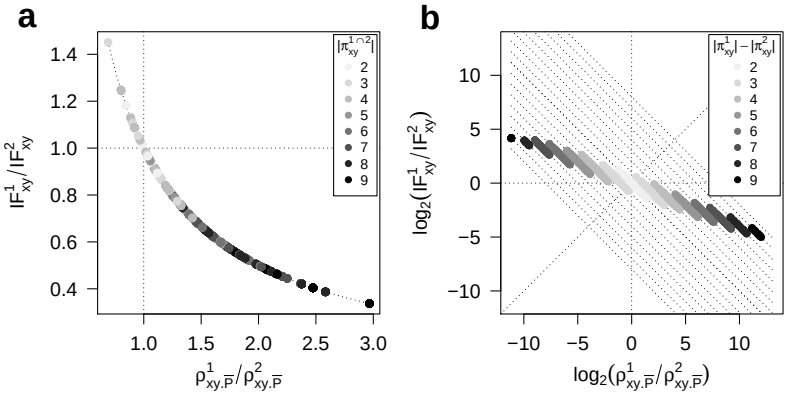

Figure 3: Comparison with constant marginal correlation. The ratio of path inflation factors is shown on the y-axis as function of the ratio of partial correlations given the vertices outside the path, on the x-axis. Values are shown separately for paths that are common (a) and different (b) between the two trees. Dotted lines across the diagonal indicate values of the x and y-axes at which the ratio of marginal correlations ρ^1_{xy}/ρ^2_{xy} is constant.

In the first simulation study we built a tree $\mathcal{T}_1$ of $p = 100$ vertices with a nearly constant degree $d = 3$ on each vertex. Then we randomly selected 50 vertices among those with degree $d = 1$, i.e. leafs in $\mathcal{T}_1$, and built a second tree $\mathcal{T}_2$ starting from $\mathcal{T}_1$ by removing the edges that connect leaf vertices and adding edges to them from 50 other randomly selected vertices. The tree $\mathcal{T}_2$ does not retain anymore the high degree of regularity in the connections of $\mathcal{T}_1$ as reflected in the comparison of path length and degree distributions shown in Figure 2. Using these two trees $\mathcal{T}_1$ and $\mathcal{T}_2$ we built two covariance matrices, whose inverse adapt to them with a constant marginal correlation $\rho = 0.5$, and calculate the ratio between partial correlations and inflation factors of the two trees, for every path. Figure 3 shows these ratios separately for common and disjoint paths between $\mathcal{T}_1$ and $\mathcal{T}_2$. In the case of common paths, we can see that despite $\rho^1_{xy} = \rho^2_{xy}$, there are paths with over 2-fold

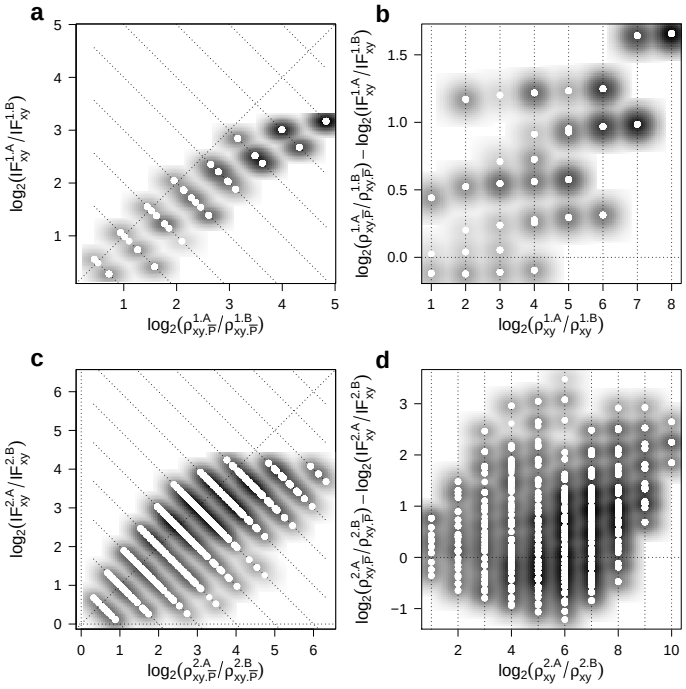

Figure 4: Comparison of marginal correlations within the same tree. Panels (a, c) show the ratio of inflation factors as function of the ratio of partial correlations, while panels (b, d) show differences between the ratio of the partial correlations and the ratio of inflation factors, as function of the marginal correlation. Ratios are compared between covariance matrices with constant $\rho = 0.5$ and $\rho = 0.25$ for tree $\mathcal{T}_1$ (a, b) and $\mathcal{T}_2$ (c, d). Shading indicates density of values and white dots are actual values.

differences in inflation and partial correlation ratios, associated with longer paths. In the case of paths that differ between $\mathcal{T}_1$ and $\mathcal{T}_2$, fold-changes in these quantities are much larger, reason why they are displayed in logarithmic scale, and also proportional to path length.

In the second simulation study we use the previously built trees $\mathcal{T}_1$ and $\mathcal{T}_2$, and simulate for each of them two covariance matrices with constant marginal correlations $\rho = 0.25$ and $\rho = 0.5$. Then, we compare partial correlations and inflation factors between covariance matrices with different ρ parameter values within each tree. We can see in Figure 4 that there are important differences in the compared quantities across identical ratios of the marginal correlation, specially in the case of $\mathcal{T}_2$, shown in panels (c) and (d). The smaller degree of regularity in the connections of $\mathcal{T}_2$ increases the diversity of these quantities.

In the third and last simulation study, we have again compared trees $\mathcal{T}_1$ and $\mathcal{T}_2$ simulating two covariance matrices with nonconstant marginal correlations on the present edges, but average $\rho = 0.5$. Thus, in this case, both structure and parameter values change. Figure 5 shows ratios of the investigated quantities and we can see that a fraction of their differences accumulate on $x = 0$, where $\rho^1_{xy} = \rho^2_{xy}$, but away from $y = 0$. These are paths that while they have similar marginal correlations, structural differences are reflected through the quantities introduced in this paper.

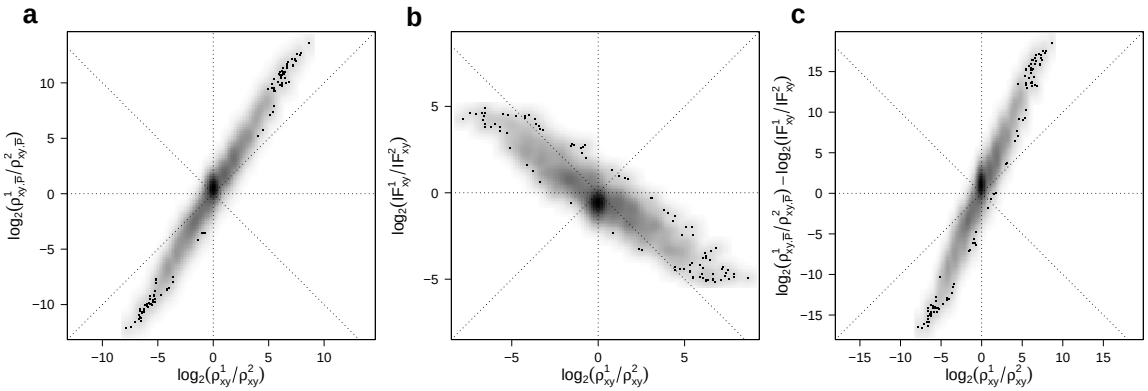

Figure 5: Comparison of two trees that differ in both, structure and marginal correlations. Differences between the ratio of the partial correlations and the ratio of inflation factors, as function of the marginal correlation, in logarithmic scale, where both covariance matrices have been simulated with marginal correlations that on average have $\rho = 0.5$ for present edges. Shading indicates density of values.

5. Differential Networking in Yeast Expression Data

Here we show a differential networking analysis from gene expression data in yeast, using the quantities introduced in this paper. We downloaded yeast RNA-seq count data from Schurch et al. (2016) and, using standard procedures (Robinson and Oshlack, 2010), we transformed integer read counts into normalized expression values in units of $\log_2$ counts per million (CPM) reads, adding a prior count value of 3 units to stabilize the variance. Schurch et al. (2016) produced the data using two strains of yeast, a wild type (BY4741 strain, WT) and a Δ*snf2* knock-out (KO) from the same genetic background. They grew these two strains in rich media and derived 48 independent biological replicates from each strain, resulting in two datasets of $n = 48$ observations each, on two genotype conditions, WT and KO. The *SNF2* gene is involved in the transcriptional activation of genes and its deletion leads to significant changes in expression of many other genes.

We reduced the initial set of $p = 5,983$ genes to $p = 5,600$ by removing genes that do not code for proteins or have co-linearities in either dataset. We further reduced the gene set to those with significant Pearson correlation (Holm's adjusted p-value $< 1\%$) in WT with the *SNF2* gene, leading to a final gene set of $p = 970$ genes, including *SNF2*. We applied the Chow-Liu algorithm (Chow and Liu, 1968) on the WT dataset using the absolute Pearson correlation as edge weight, obtaining a tree $\mathcal{T}_{WT}$. We did the same on the KO dataset, but excluding the *SNF2* gene, obtaining another tree $\mathcal{T}_{KO}$. We selected the paths traversing *SNF2* in $\mathcal{T}_{WT}$, which were 1,936, and the corresponding ones with the same endpoints in $\mathcal{T}_{KO}$. For each set of paths, we estimated inflation factors in the corresponding data set. To that end, because in our setting $p \gg n$, we relied on the same tecniques used by Roverato and Castelo (2017) to estimate vector correlations from such data.

Figure 6 shows the results where most paths have less than 50% difference in their marginal correlation. However, these paths show differences in the quantities introduced in this paper, the largest ones associated with differences in path length.

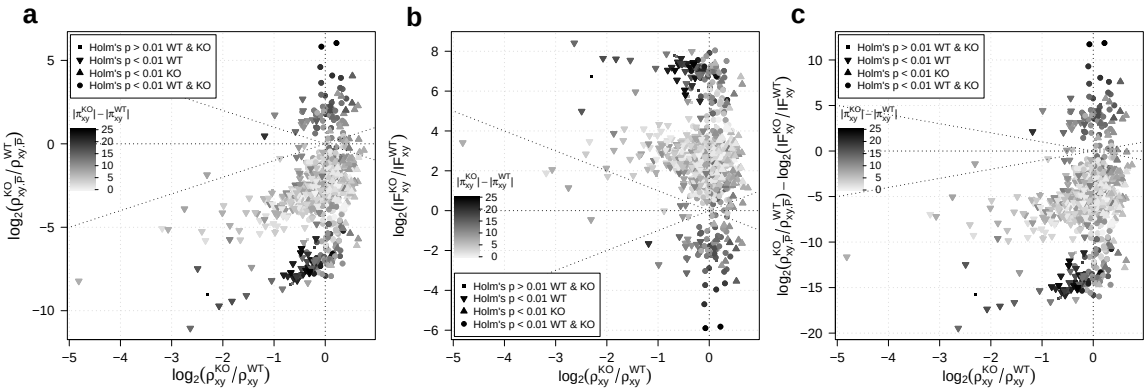

Figure 6: Differential networking in yeast. Quantities as in Figure 5, calculated from yeast gene expression data in two conditions, WT and KO. Marginal and partial correlations are shown in absolute value, where only 6 of them have opposite signs between WT and KO.

6. Conclusions

In this paper we have investigated the decomposition of the marginal correlation along paths of Gaussian trees. Our main result in Theorem 3 shows that marginal correlations between two vertices x and y in a tree, can be written as the product of the their partial correlation given the vertices outside the path, $\rho_{xy.\bar{P}}$, and an inflation factor associated with the path, $\mathrm{IF}_{xy}^{\mathcal{T}}$. Using simulations and real gene expression data, we have shown that these two quantities capture structural differences between two trees even when the path in question, or the ratio of marginal correlations, are identical between the trees. In our view, this result opens up new ways to identify differential networking events that we may miss with current methods to address this question.

Acknowledgments

We acknowledge the support of the Spanish MINECO/FEDER [TIN2015-71079-P] and the European Cooperation in Science and Technology (COST) [CA15109]. Alberto Roverato was also supported by the Air Force Office of Scientific Research under award number FA9550-17-1-0039.

References

D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*, volume 571. John Wiley & Sons, 2005.

S. Chatterjee and A. S. Hadi. *Regression analysis by example*. John Wiley & Sons, 2012.

M. J. Choi, V. Y. Tan, A. Anandkumar, and A. S. Willsky. Learning latent tree graphical models. *J. Mach. Learn. Res.*, 12:1771–1812, 2011.

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

D. R. Cox and N. Wermuth. *Multivariate Dependencies: Models, analysis and interpretation*. Chapman and Hall, London, 1996.

P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *J. Roy. Stat. Soc. Ser. B (Stat. Meth.)*, 76(2):373–397, 2014.

A. de la Fuente. From 'differential expression' to 'differential networking'–identification of dysfunctional regulatory networks in diseases. *Trends in Genetics*, 26(7):326–333, 2010.

A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

D. Edwards, G. C. De Abreu, and R. Labouriau. Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. *BMC Bioinformatics*, 11(1):18, 2010.

J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

B. Jones and M. West. Covariance decomposition in undirected Gaussian graphical models. *Biometrika*, 92(4):779–786, 2005.

K. Kim and N. Timm. *Univariate and Multivariate General Linear Models: Theory and applications with SAS*. CRC Press, 2006.

S. Lauritzen, C. Uhler, and P. Zwiernik. Maximum likelihood estimation in gaussian models under total positivity. *The Annals of Statistics, in press*, 2018.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. London: Academic Press, 1979.

M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, 11(3):R25, 2010.

A. Roverato and R. Castelo. The networked partial correlation and its application to the analysis of genetic interactions. *J. Roy. Stat. Soc. Ser. C (Appl. Stat.)*, 66(3):647–665, 2017.

W. W. Rozeboom. Linear correlations between sets of variables. *Psychometrika*, 30(1):57–71, 1965.

N. J. Schurch et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, 2016.

N. H. Timm. *Applied Multivariate Analysis*. Springer-Verlag, New York, 2002.

I. Tur, A. Roverato, and R. Castelo. Mapping eQTL networks with mixed graphical Markov models. *Genetics*, 198(4):1377–1393, 2014.

J. Whittaker. *Graphical Models in Applied Multivariate Analysis*. John Wiley & Sons, 1990.

S. Wright. Correlation and causation. *J. Agric. Res.*, 20(7):557–585, 1921.

P. Zwiernik. *Semialgebraic Statistics and Latent Tree Models*. CRC Press, 2015.