Journal of Machine Learning Research 7 (2006) 2621-2650 Submitted 4/06; Revised 10/06; Published 12/06

A Robust Procedure For Gaussian Graphical Model Search From Microarray Data With p Larger Than n

Robert Castelo ROBERT.CASTELO@UPF.EDU
Departament de Ciències Experimentals i de la Salut
Universitat Pompeu Fabra
Dr. Aiguader 88, E-08003 Barcelona, Spain

Alberto Roverato ALBERTO.ROVERATO@UNIBO.IT
Dipartimento di Scienze Statistiche
Università di Bologna
Via Belle Arti 41, I-40126 Bologna, Italy

Editor: Max Chickering

Abstract

Learning of large-scale networks of interactions from microarray data is an important and challenging problem in bioinformatics. A widely used approach is to assume that the available data constitute a random sample from a multivariate distribution belonging to a Gaussian graphical model. As a consequence, the prime objects of inference are *full-order partial correlations* which are partial correlations between two variables given the remaining ones. In the context of microarray data the number of variables exceed the sample size and this precludes the application of traditional structure learning procedures because a sampling version of full-order partial correlations does not exist. In this paper we consider *limited-order partial correlations*, these are partial correlations computed on marginal distributions of manageable size, and provide a set of rules that allow one to assess the usefulness of these quantities to derive the independence structure of the underlying Gaussian graphical model. Furthermore, we introduce a novel structure learning procedure based on a quantity, obtained from limited-order partial correlations, that we call the *non-rejection rate*. The applicability and usefulness of the procedure are demonstrated by both simulated and real data.

Keywords: Gaussian distribution, gene network, graphical model, microarray data, non-rejection rate, partial correlation, small-sample inference

1. Introduction

High-throughput experimental technologies developed within the field of molecular biology allow one to observe in real time the activity of thousands of biomolecules in the cell under tens of different experimental conditions. These technologies, known as *microarray* technologies, are able to put together in a solid substrate (a chip) of a few squared centimeters a bidimensional matrix (an array) formed by tens of thousands of probes. Each probe is specific to a nucleic acid sequence that recognizes (hybridises) marked samples (biomolecules) of complementary RNA (coming from the experimental conditions under study), quantifying the abundance of each recognized biomolecule. An open question within molecular biology research is to be able to describe the set of interactions, or biomolecular network, between the different functional elements in the genome that mediate the production of the biomolecules we observe through these high-throughput platforms. These data,

the so-called *microarray data*, can be seen as a random sample of a multivariate distribution defined by a set of random variables associated to the genome functional elements under study (e.g., genes). Each record corresponds to a vector of values describing the abundance of a particular kind of biomolecule (e.g., messenger RNA) produced by each genome functional element under a specific experimental condition (e.g., a specific tissue or cell line). Thus, a way to describe the interactions among the genome functional elements is by using conditional independencies and, more concretely, graphical models (see Pearl, 1988; Whittaker, 1990; Lauritzen, 1996) which have emerged as a powerful tool for the learning, description and manipulation of conditional independencies.

However, in a typical microarray data set the number of observations n (on the order of tens) is substantially smaller than the number of variables p (on the order of hundreds or even thousands) and this prevents us from applying directly most of the existing multivariate methods for structure learning of graphical models due to the difficulties in obtaining estimates of the joint probability distribution.

In this paper, we focus in Gaussian graphical models and investigate the role of marginal distributions in their structure learning. Firstly, we formally introduce the concept of *q-partial graph* that is a graph associated with the set of all marginal distributions of dimension $q+2$ and, furthermore, we provide a comprehensive description of the connection between a q-partial graph and the graph associated with the Gaussian graphical model of interest. Secondly, we propose a novel q-partial-correlations based procedure, *qp-procedure* hereafter, for structure learning of q-partial graphs based on a quantity that we call the *non-rejection rate*. The results of this paper can be applied also outside the biological context because they can be more generally useful whenever structure learning of a Gaussian graphical model is carried out in the special context in which (i) p is large compared to n, (ii) the underlying structure of the graphical model is sparse. Furthermore, the qp-procedure can also be regarded as a method to obtain shrinkage estimators of the covariance matrix. We remark that the theory of q-partial graphs is developed under the assumption of *faithfulness* of the probability distribution to its independence graph, however the qp-procedure is robust with respect to this assumption as we shall discuss at the end of the paper.

The paper is organized as follows. Sections 2 and 3 give the theory of Gaussian graphical models and their application to learning of biomolecular networks from microarray data, respectively. The theory of q-partial graphs is given in Section 4 whereas the required graph theory is provided in the Appendix. The qp-procedure is introduced in Section 5 where instances of its application to both simulated and real data are given and, finally, Section 6 contains a brief discussion.

2. Gaussian Graphical Models

In this section we review the Gaussian graphical model theory required for this paper. For a full account of graphical model theory we refer to Cox and Wermuth (1996), Lauritzen (1996) and Whittaker (1990) whereas, for the theory relating to structure learning of graphical models we refer to Cowell et al. (1999), Edwards (2000), Jones et al. (2005) and Whittaker (1990).

Let $X_V \equiv X$ be a random vector indexed by $V = \{1,\ldots,p\}$ with probability distribution P_V and let $G = (V,E)$ be an undirected graph; see Appendix A for the graph theory used here. For a subset $A \subseteq V$, we denote by X_A the subvector of X indexed by A, and by P_A the associated marginal distribution. For a triplet $I,J,U \subseteq V$ we write $X_I \perp\!\!\!\perp X_J | X_U$ to denote that X_I is conditionally independent of X_J given X_U; we allow U to be the empty set to denote the marginal independence of X_I and X_J.

We say that P_V is (undirected) *Markov* with respect to G if it holds that $X_I \perp\!\!\!\perp X_J | X_U$ whenever U separates I and J in G; in particular this implies that if $(i,j) \in \bar{E}$ then $X_i \perp\!\!\!\perp X_j | X_{V\setminus\{i,j\}}$. Here $\bar{E}$ denotes the set of missing edges of $G = (V,E)$ as formally defined in Appendix A. We say that P_V is *faithful* to G if all the conditional independence relationships in P_V can be read off the graph G through the Markov property. Consider a graph $G' = (V,E')$ larger than G, $G \subseteq G'$. It is straightforward to check that if P_V is Markov with respect to G then it is also Markov with respect to G'. However, if P_V is faithful to G then it is faithful to G' if and only if $G = G'$.

Throughout this paper X_V is assumed to have a multivariate normal distribution with mean vector μ_V and positive definite covariance matrix $\Sigma_{VV} \equiv \Sigma$. Furthermore, we assume that P_V is both Markov and faithful with respect to an undirected graph $G = (V,E)$. Hence, for a subset $Q \subset V$ with $i, j \notin Q$ it holds that $X_i \perp\!\!\!\perp X_j | X_Q$ if and only if the partial correlation coefficient

$$\rho_{ij.Q} = \frac{-\kappa^A_{ij}}{\sqrt{\kappa^A_{ii}\kappa^A_{jj}}}$$

is equal to zero, where $A = Q \cup \{i,j\}$ and $K^A = \{\kappa^A_{ij}\}$ is the *concentration matrix* of X_A, $K^A = (\Sigma_{AA})^{-1}$ (Lauritzen, 1996, p. 130). Of special interest is the case $A = V$ because the concentration matrix $K^V \equiv K = \{\kappa_{ij}\}$ is the inverse of Σ and the structure of $G = (V,E)$ can be derived from the zero pattern of K. More specifically, it holds that (Lauritzen, 1996, Proposition 5.2)

$$k_{ij} = 0 \quad \Leftrightarrow \quad \rho_{ij.V\setminus\{i,j\}} = 0 \quad \Leftrightarrow \quad (i,j) \in \bar{E}, \tag{1}$$

and for this reason G is called the *concentration graph* of X_V. For $|Q| = q$, the parameter $\rho_{ij.Q}$ is called a *q-order partial correlation* of X_i and X_j, and if $q = p-2$, that is, $Q = V\setminus\{i,j\}$, we say that $\rho_{ij.Q}$ is the *full-order partial correlation* of X_i and X_j.

A *Gaussian graphical model* (Dempster, 1972) is the family of p-variate normal distributions that are Markov with respect to a given undirected graph $G = (V,E)$. Let $X^{(n)} = (X^1, \ldots, X^n)$ be a random sample from P_V. For a Gaussian graphical model with graph G the sufficient statistics are given by the sample mean vector and by the sample covariance matrices S_{CC} for $C \in \mathcal{C}$ where $\mathcal{C}$ is the set of cliques of G (Lauritzen, 1996, p. 132). It follows that, when G is complete the sufficient statistics are the sample mean and the sample covariance matrix S. Here, we consider problems in which the sample size is small, and it is thus important to recall that, for $A \subseteq V$, the sample covariance matrix S_{AA} from $X_A^{(n)}$ has full rank, with probability one, if and only if $n > |A|$ (Dykstra, 1970) and that a necessary condition for the computation of several statistical quantities such as the maximum likelihood estimates of K and of the partial correlations in (1) is that S_{CC} has full rank for all $C \in \mathcal{C}$.

Structure learning aims at identifying the structure $G = (V,E)$ with the fewest number of edges on the basis of the available data such that the underlying distribution P_V is undirected Markov over G. In a frequentist approach to inference, a basic operation to be performed in structure learning procedures is a statistical test for the hypothesis that a given partial correlation is zero, $\rho_{ij.Q} = 0$, since for $Q = V\setminus\{i,j\}$ this is equivalent to the hypothesis that $(i,j) \in \bar{E}$. If, for $A = Q \cup \{i,j\}$, X_A has an (unrestricted) normal distribution then the generalized likelihood ratio test for the hypothesis that $\rho_{ij.Q} = 0$ has form $L = -n\log(1-\hat{\rho}^2_{ij.Q})$ where $\hat{\rho}_{ij.Q} = -\hat{\kappa}^A_{ij}/\sqrt{\hat{\kappa}^A_{ii}\hat{\kappa}^A_{ii}}$ and $\widehat{K^A} = (S_{AA})^{-1}$ is the maximum likelihood estimate of K^A (Whittaker, 1990, p. 175). Under the null hypothesis, the asymptotic distribution of L is χ^2_1, even though for a small sample size the exact distribution of the

statistical test may be preferred; see Schäfer and Strimmer (2005a). An alternative way to verify the above hypothesis is provided by the connection between partial correlations and regression coefficients. More specifically, in the regression of X_i on $X_{A\setminus\{i\}}$ the regression coefficient associated with X_j is zero if and only if $\rho_{ij.Q} = 0$ (see Cox and Wermuth, 1996, p. 69). In the structure learning procedure proposed in this paper, to verify the absence of an edge from the unrestricted model we will apply the usual t test for zero regression coefficients because it is optimal, in the sense that it is Uniformly Most Powerful Unbiased (UMPU) (see Lehmann, 1986, p. 397).

3. Gaussian Graphical Models For biomolecular Networks

Microarray data quantify the abundance of biomolecules, commonly known as expression level, by probing functional elements along the genome which, without loss of generality, we shall hereafter refer to as genes. A set of p genes being probed define a vector of random variables X_i, $i = 1, \ldots, p$, that take normalized values of the expression levels of the corresponding genes. For every variable X_i there is vector of n values coming from n different experimental conditions forming the so-called expression profile. The microarray data consist of the expression profiles of a set of genes and form a snapshot of the interactions between the genes in terms of statistical (in)dependencies which, in principle, could be inferred through structure learning of Gaussian graphical models and thus leading to a description of the underlying biomolecular network in these terms. Hence, the prime object of interest is the inverse of the covariance matrix, also known as concentration matrix, whose zero pattern defines the structure of the graphical model, known then as concentration graph.

However, in contrast with the usual data sets found in the literature, on which structure learning of Gaussian graphical models is applied, microarray data constitute a challenging problem because microarray experiments typically measure the expression level of a large number of genes across a small number of experimental conditions. As a consequence of the scarcity of the data, the maximum likelihood of the inverse covariance matrix does not exist because the sample covariance matrix has full rank, with probability one, if and only if $n > p$ (Dykstra, 1970). This paper tackles this specific circumstance under which we perform structure learning of Gaussian graphical models with *small n and large p*.

An important observation in this context is that a growing body of biological evidence suggests that biomolecular networks have a sparse structure. This feature, usually regarded as an advantage, has been exploited in a number of ways to enable learning of Gaussian graphical models from microarray data (see, among others, Wong et al., 2003; Dobra et al., 2004; Wille et al., 2004; Wille and Bühlmann, 2006; Shäfer and Strimmer, 2005a, 2005b, 2005c) among which some methods work by obtaining shrinkage estimators of the covariance matrix (Wong et al., 2003; Shäfer and Strimmer, 2005c) while some other have made an attempt to learn an approximate version of the biomolecular network by using marginal distributions of dimension smaller than n. We shall discuss this latter approach in more detail below.

Instead of trying to learn the concentration graph of a Gaussian graphical model from microarray data, a tool employed by the bioinformatics community to describe interactions between genes is the *relevance network*; see Butte et al. (2000) and Steuer et al. (2003a, 2003b). In relevance networks missing edges denote zero correlations between pairs of genes, that in the Gaussian case imply marginal independence. In these graphs, edges are typically represented by undirected lines; nevertheless in the graphical model literature these models are known as *covariance graphs* (Cox and Wermuth, 1993, 1996) and edges are represented by either bidirected arrows or dashed undi-

rected lines. A correlation coefficient is zero if and only if the corresponding covariance is zero and therefore the structure of a covariance graph is derived from the zero pattern of the covariance matrix Σ. Although structure learning of covariance graphs is not straightforward (Drton and Perlman, 2004; Drton and Richardson, 2004), a statistical test for the hypothesis that a single correlation coefficient is zero can be easily carried out for $n > 2$. This allows the implementation of naive learning procedures that consider separately every edge of the graph overcoming the *large p and small n* problem. In a similar vein to the relevance network approach see also the ARACNE algorithm by Margolin et al. (2006).

More recently, other families of graphical models have been used to describe biomolecular networks (see Friedman, 2004) and among these, an important role is played by Gaussian graphical models where missing edges correspond to zero partial correlations and, therefore, to conditional independence relationships. In these models an edge between two genes represents a direct association and, more generally, a path connecting two genes represents an undirect association mediated by other genes in the path (see Jones and West, 2005). The reason why concentration graphs are more adequate than covariance graphs to describe gene networks is that, even though two genes may present a non-zero correlation because they belong to a common biological pathway, they should not be joined by an edge when they influence each other only indirectly through other observed genes that act as confounders.

The Pearson correlation is a marginal measure of association between two genes, regardless of other genes in the network. On the other hand, partial correlation is a measure of association between two genes that keeps into account all the remaining observed genes. Consequently, partial correlations cannot be computed by only looking at bivariate marginal distributions but require the full joint distribution of genes, and this is problematic when n is small. More formally, the network structure is derived from the zero pattern of the concentration matrix $K = \Sigma^{-1}$ whose maximum likelihood estimate is $\widehat{K} = S^{-1}$ which requires that S has full rank and this holds, with probability one, if and only if $n > p$ (Dykstra, 1970). Furthermore, the statistical properties of procedures for fitting and testing partial correlations depend on $n - p$ and, as pointed out for instance by Yang and Berger (1994) and Dempster (1969), the estimators based on scalar multiples of S tend to distort the Eigenstructure of the true covariance matrix, unless $n \gg p$.

Several solutions have been proposed in the literature to carry out structure learning of biomolecular networks by means of concentration graphs; see Jones et al. (2005) and Shäfer and Strimmer (2005c) for a review. A popular approach is based on *limited-order partial correlations*, that is q-order partial correlations with $q < (n-2)$. Procedures based on limited-order partial correlations have been applied, among others, by de la Fuente et al. (2004), Magwene and Kim (2004), Wille et al. (2004), Wille and Bühlmann (2006) and are also implemented in the statistical software MIM (Edwards, 2000). The key point here is that if a set of $q+2$ genes such that $(q+2) < n$ is considered, then a test for the hypothesis of a zero q-order partial correlation can be carried out with standard techniques such as those described in Section 2. Consequently, it seems somehow sensible to replace full-order partial correlations with lower-order partial correlations so as to obtain a graph that can be regarded as an approximation of the entire concentration graph G. The procedures proposed in the literature for learning such an approximating graph are based on the application of the following rule to every distinct pair of vertices $i, j \in V$:

> Test the hypotheses $\rho_{ij.Q} = 0$ for every $Q \subseteq V \backslash \{i, j\}$ such that $|Q| = q$. Then, i and j are joined by an edge if and only if all of such hypotheses of zero q-order partial correlations are rejected.

In principle, q-order partial correlations can be computed for any $q < (n-2)$; however, in practice, testing $\binom{p-2}{q}$ partial correlations for each of the $p \times (p-1)/2$ pairs of genes is computationally intensive unless q is small and, to our knowledge, the above procedure has only be applied for $q \leq 3$. For instance, Wille and Bühlmann (2006) proposed a modified version of the above procedure that considers all q-order partial correlations for $q \leq 1$. We remark that this learning procedure presents two main drawbacks. Firstly, as shown in the next section, the usefulness of q-order partial correlations increases with q, so that a procedure that can be applied for larger values of q is called for. More seriously, however, an edge is added to the graph if all of $\binom{p-2}{q}$ null hypotheses are rejected. The statistical tests are performed separately so that the well-known problems deriving from the sequential application of several tests may occur. In particular, the probability that at least one hypothesis of zero q-order partial correlation is wrongly non-rejected increases with the number of performed tests and, consequently, if the value of $\binom{p-2}{q}$ is large then one should expect that most, or even all, of the edges are removed.

In the next section we provide a formal definition of the graph associated with q-order partial correlations that we call the *q-order partial correlation graph* of X_V, *q-partial graph* hereafter, denoted by $G^{(q)} = (V, E^{(q)})$, and derive some of its properties. In this way we generalize the results of Wille and Bühlmann (2006) given for $q = 1$ to an arbitrary value of q. In particular, it is easy to check that, under the assumption of faithfulness, it holds that $G \subseteq G^{(q)}$, and consequently that P_V is undirected Markov with respect to $G^{(q)}$. This means that every pair of vertices separated in $G^{(q)}$ corresponds to a conditional independence relation between the two corresponding variables and, more specifically, every missing edge corresponds to a pairwise conditional independence. In practice, however, the usefulness of $G^{(q)}$ depends on its closeness to G, that is, on the number of edges that are present in $G^{(q)}$ but are missing in G, and we will formally address this point.

Even though the q-partial graph $G^{(q)}$ of X_V may provide a good approximation to the concentration graph G, our standpoint is that the real object of interest is the concentration graph and that the q-partial graph is useful as an intermediate step of the analysis. In fact, if the dimension of the largest clique of $G^{(q)}$ is smaller than the sample size, then the corresponding graphical model, as well as all its submodels, can be fitted and, consequently, it is possible to apply traditional search procedures to learn the concentration graph by using the fitted q-partial graph as a starting point. In this perspective, in Section 5 we propose a novel procedure to learn q-partial graphs from data. This is based on limited-order partial correlations but can be used with larger values of q and, furthermore, it does not suffer of the problems deriving from multiple testing. Since the selected graph is the starting point for further investigation, our procedure is designed to be conservative, that is, it aims at keeping the number of wrongly removed edges small and, consequently, the probability of breaking the Markov condition of P_V low. It follows that the selected graph may still contain edges that should be removed. However, if the underlying concentration graph is sparse the procedure will remove a large number of edges leading to a great simplification of the learning problem. Furthermore, as shown by examples carried out on both simulated and real data, the resulting graph is manageable with standard techniques. We remark that our procedure neither imposes any constraints to induce a dimensionality reduction nor makes any assumption of sparseness of the graph. However, the usefulness of the proposed procedure does depend on the sparseness of G. It provides an indication whether the underlying concentration graph is sparse and, in this case, it will lead to a great simplification of the structure learning problem.

4. q-Partial Graphs

The use of limited-order partial correlations in structure learning is appealing when either $p > n$ or the available data are too scarce to produce reliable estimates of the concentration matrix. However, the object of interest is the concentration graph G of X_V and it is not clear which graph can be learnt by using q-order partial correlations, and what is the connection between such a graph and G. In this section we formally approach this question: firstly, we introduce the *q-partial graph* of X_V, that is a graph in which missing edges correspond to zero q-order partial correlations. Secondly, we characterize the class of graphs for which concentration graphs and q-partial graphs coincide and, in particular, we show how information on the concentration graph of X_V can be extracted from the q-partial graph of X_V. The theory here developed relies on the graph theory described in Appendix A and more specifically on the concepts of the outer connectivity of two vertices i and j, $d(i,j|G)$, the outer connectivity of the edges of G, $d(E|G)$, the outer connectivity of the missing edges of G, $d(\bar{E}|G)$, and finally, the outer connectivity of G, $d(G)$.

The concentration graph of X_V is associated with the probability distribution of X_V and we define the q-partial graph of X_V as a graph associated with the set of all marginal distributions of X_V of dimension $(q+2)$.

Definition 1 *For a random vector X_V and an integer $0 \leq q \leq (p-2)$ we define the* q-partial graph *of X_V, denoted by $G^{(q)} = (V, E^{(q)})$, as the undirected graph where $(i,j) \in \bar{E}^{(q)}$ if and only if there exists a set $U \subseteq V$ with $|U| \leq q$ and $i, j \notin U$ such that $X_i \perp\!\!\!\perp X_j | X_U$ holds in P_V.*

We first observe that $G^{(p-2)}$ and $G^{(0)}$ are the concentration graph and the covariance graph of X_V respectively, whereas $G^{(1)}$ is the 0-1 conditional independence graph introduced by Wille and Bühlmann (2006, Definition 3). It is also easy to show that that $G^{(q)}$ is larger than G, $G \subseteq G^{(q)}$, that is every edge in G is also an edge in $G^{(q)}$. This follows from the fact that if $(i,j) \in E$ then the faithfulness of X_V to G implies that there is no set $U \subseteq V$ with $i, j \notin U$ such that $X_i \perp\!\!\!\perp X_j | X_U$, and therefore it holds that $(i,j) \in E^{(q)}$; see also Wille and Bühlmann (2006).

The relation $G \subseteq G^{(q)}$ implies that X_V is Markov with respect to $G^{(q)}$. However, the usefulness of $G^{(q)}$ as a surrogate of G depends on the closeness of the two graphs. Every edge of G is present in $G^{(q)}$ and in the following proposition we characterize the missing edges of G that are also missing in $G^{(q)}$.

Proposition 1 *Let $G = (V,E)$ and $G^{(q)} = (V,E^{(q)})$ be the concentration and the q-partial graph of X_V respectively. If $(i,j) \in \bar{E}$ then $(i,j) \in \bar{E}^{(q)}$ if and only if $d(i,j|G) \leq q$.*

Proof Sufficiency. If $d(i,j|G) \leq q$ then there exists a nontrivial minimal $\{i,j\}$-separator $S \in \mathcal{S}_{(i,j|G)}$ such that $|S| \leq q$. By the Markov property, it holds that $X_i \perp\!\!\!\perp X_j | X_S$ so that $(i,j) \in \bar{E}^{(q)}$ by definition of q-partial graph. Necessity. If $(i,j) \in \bar{E}^{(q)}$ then there exists a set $U \subseteq V$ with $|U| \leq q$ and $i, j \notin U$ such that $X_i \perp\!\!\!\perp X_j | X_U$. By the faithfulness assumption, such a conditional independence relation can be also read off the graph G through the Markov property. In other worlds, U is a nontrivial $\{i,j\}$-separator in G so that there exists a subset $S \subseteq U$ such that $S \in \mathcal{S}_{(i,j|G)}$ and, consequently, $d(i,j|G) \leq |S| \leq q$. ■

The result stated in the above proposition is very intuitive. A missing edge in G is missing also in $G^{(q)}$ if and only if the outer connectivity of the corresponding vertices is smaller or equal to q or, that is, if and only if there exists a marginal distribution of X_V of dimension $(q+2)$ in which the

corresponding variables are conditionally independent. If this relation is satisfied for all the missing edges of G then the q-partial graph and the concentration graph are identical.

Proposition 2 *Let $G = (V,E)$ and $G^{(q)} = (V,E^{(q)})$ be the concentration and the q-partial graph of X_V respectively. Then $G = G^{(q)}$ if and only if $d(\bar{E}|G) \leq q$.*

Proof We have already shown that the inclusion relation $G \subseteq G^{(q)}$ is always satisfied. Consequently, we have only to show that $G \supseteq G^{(q)}$ if and only if $d(\bar{E}|G) \leq q$. The condition $G \supseteq G^{(q)}$ is satisfied if and only if $(i,j) \in \bar{E}$ implies $(i,j) \in \bar{E}^{(q)}$, and in the following we consider the latter formulation of the condition. Sufficiency. By Equation (9) in the Appendix, $d(\bar{E}|G) \leq q$ implies $d(i,j|G) \leq q$ for all $(i,j) \in \bar{E}$ and, by Proposition 1, this implies that $(i,j) \in \bar{E}^{(q)}$ for every $(i,j) \in \bar{E}$. Necessity. By Proposition 1, if $(i,j) \in \bar{E}^{(q)}$ for all $(i,j) \in \bar{E}$, then $d(i,j|G) \leq q$ for all $(i,j) \in \bar{E}$, and it follows from (9) that $d(\bar{E}|G) \leq q$. ■

The result of Proposition 2 clarifies that the concentration graph G and the q-partial graph $G^{(q)}$ of X_V coincide when $d(\bar{E}|G)$ is not greater than q so that a natural question concerns the connection between the sparseness of G and the value of $d(\bar{E}|G)$. This is discussed at the end of Appendix A where it is shown that there is no direct connection between the degree of sparseness of G and outer degree of missing edges. In particular it is possible to find examples in which the condition of Proposition 2 is satisfied for a graph G' but is not satisfied for a sparser graph $G \subset G'$. Note also that the condition of Proposition 2 is always satisfied when G is the complete graph. The point here is that sparseness is useful as long as it implies small separators for non-adjacent vertices, however it is not difficult to draw a very sparse graph in which two non-adjacent vertices have high value of outer connectivity.

It is somehow intuitive that larger values of q should be preferred and, in fact, an immediate consequence of Proposition 1 is the following relation of inclusion between partial graphs of different order.

Corollary 3 *Let $G^{(q)} = (V,E^{(q)})$ and $G^{(r)} = (V,E^{(r)})$ be the q-partial and the r-partial graph of X_V respectively. If $r \leq q$ then $G^{(q)} \subseteq G^{(r)}$.*

Proof We show that if $r \leq q$ and $(i,j) \in \bar{E}^{(r)}$ then $(i,j) \in \bar{E}^{(q)}$. From the definition of outer connectivity (see Appendix) $(i,j) \in \bar{E}^{(r)}$ implies $d(i,j|G^{(r)}) \leq r$. Since $r \leq q$, $d(i,j|G^{(r)}) \leq q$ and therefore by Proposition 1 $(i,j) \in \bar{E}^{(q)}$. ■

The results provided so far allow to understand in which cases q-partial graphs may be useful. They give a set of necessary and sufficient conditions, however such conditions are stated with respect to G, which is unknown, and therefore their usefulness is limited in practice to situations in which background knowledge on the problem under analysis may provide information on the structure of G. Also $G^{(q)}$ is typically unknown but it can be learnt from data and in the rest of this section we show how information on the structure of G can be extracted from $G^{(q)}$.

The fact that $G^{(q)}$ is larger than G implies that if an edge is missing in $G^{(q)}$ then it is also missing in G and the next theorem provides a sufficient condition to check whether an edge that is present in $G^{(q)}$ is also present in G.

Theorem 4 *Let $G = (V,E)$ and $G^{(q)} = (V,E^{(q)})$ be the concentration and the q-partial graph of X_V respectively. If $(i,j) \in E^{(q)}$ then a sufficient condition for the relation $(i,j) \in E$ to hold is $d(i,j|G^{(q)}) \leq q$.*

Proof Assume $(i,j) \in E^{(q)}$ and $d(i,j|G^{(q)}) \leq q$. As mentioned earlier in the paper, from the faithfulness of P_V it follows $G \subseteq G^{(q)}$ and thus by Equation (13) in Theorem 6 $d(i,j|G) \leq d(i,j|G^{(q)}) \leq q$. By Proposition 1, $d(i,j|G) \leq q$ implies that if $(i,j) \in \bar{E}$ then $(i,j) \in \bar{E}^{(q)}$ which would contradict the initial assumption and therefore $(i,j) \in E$. ■

Note that the condition of Theorem 4 can be checked on $G^{(q)}$, and an immediate consequence of Theorem 4 is the following corollary that provides a sufficient condition for checking the identity $G = G^{(q)}$ directly from $G^{(q)}$.

Corollary 5 *Let $G = (V,E)$ and $G^{(q)} = (V,E^{(q)})$ be the concentration and the q-partial graph of X_V respectively. A sufficient condition for the relation $G = G^{(q)}$ to hold is that $d(E^{(q)}|G^{(q)}) \leq q$.*

Assuming that $G^{(q)}$ is known, then Corollary 5 gives a condition to check the identity $G = G^{(q)}$. In the case one cannot conclude that G is equal to $G^{(q)}$ then Theorem 4 can be applied to decide which edges of $G^{(q)}$ belong also to G and which edges of $G^{(q)}$ may be spurious. Theorem 4 and Corollary 5 should be compared with Propositions 1 and 2. The former give weaker results but are of more practical use because if an estimate $\widehat{G}^{(q)} = (V,\widehat{E}^{(q)})$ of $G^{(q)}$ is available, then one can estimate $d(E^{(q)}|G^{(q)})$ with $d(\widehat{E}^{(q)}|\widehat{G}^{(q)})$ and $d(i,j|G^{(q)})$ with $d(i,j|\widehat{G}^{(q)})$.

The computation of the outer connectivity of two vertices is known to be a NP-hard problem. Nevertheless several algorithms are available to derive both upper and lower bounds to this number (Rosenberg and Heath, 2001) and, since all the results stated in this section involve inequalities, then such upper and lower bounds may be sufficient to check the required conditions. Note also that equations (7), (10), (11) and (12) in Appendix A are instances of easily computable upper bounds.

We close this section by noticing that the outer connectivity of edges and the outer connectivity of missing edges play a different role with respect to $G^{(q)}$. The quantities that determine the "closeness" of $G^{(q)}$ to G are $d(i,j|G)$ for $(i,j) \in \bar{E}$. Indeed, both the value of $d(E|G)$ and of $d(E^{(q)}|G^{(q)})$ are irrelevant here, and a concentration graph can coincide with a q-partial graph even if its edges have a very high maximal degree of outer connectivity; recall that $d(E|G) \leq d(E^{(q)}|G^{(q)})$ by (14). On the other hand, the values of $d(i,j|G^{(q)})$ for $(i,j) \in E^{(q)}$ are important for the practical usefulness of q-partial graphs: the larger the number of edges of $(i,j) \in E^{(q)}$ with $d(i,j|G^{(q)}) \leq q$ the larger is the amount of information that $G^{(q)}$ provides with respect to G. Note also that, unlike $d(\bar{E}|G)$, the value of $d(E|G)$ is related with the sparseness of G (see Theorem 6 in Appendix A).

5. The qp-Procedure

We now introduce a novel procedure to learn the q-partial graph $G^{(q)}$ of X_V, that we name the *qp-procedure*. This is based on limited-order partial correlations and, more specifically, on a quantity that we call the *non-rejection rate*. The latter is a probability associated with every pair of variables X_i and X_j, and turns out to be useful in discriminating between present and missing edges in $G^{(q)}$. The qp-procedure firstly estimates the value of all the $p \times (p-1)/2$ non-rejection rates and then a graph $\widehat{G}^{(q)}$ is constructed by removing from the complete graph all the edges corresponding to the pairs of variables whose fitted value of the non-rejection rate is above a given threshold. In Section 5.1 we formally introduce the non-rejection rate. In Section 5.2 we describe the procedure in more detail by means of two examples and, finally, in Section 5.3 we provide istances of the application of the procedure on both simulated and real data.

5.1 The Non-Rejection Rate

For a pair of vertices $i, j \in V$, with $i \neq j$, and an integer $q \leq (p-2)$ let $\mathcal{Q}_{ij}$ be the set made up of all the subsets Q of $V\backslash\{i,j\}$ such that $|Q| = q$; thus the cardinality of $\mathcal{Q}_{ij}$ is $m = \binom{p-2}{q}$. Furthermore, let T_{ij}^q be the random variable resulting of the two stage experiment in which firstly an element Q is sampled from $\mathcal{Q}_{ij}$ according to a (discrete) uniform distribution and then the data $X^{(n)}$ are used to test the null hypothesis $H_0 : \rho_{ij.Q} = 0$ against the alternative hypothesis $H_A : \rho_{ij.Q} \neq 0$. The random variable T_{ij}^q takes value 0 if the above null hypothesis is rejected and 1 otherwise. It follows that T_{ij}^q has a Bernoulli distribution and the non-rejection rate is defined as follows.

Definition 2 *For a random sample $X^{(n)}$ from X_V the* non-rejection rate *for the variables X_i and X_j with $i, j \in V$, $i \neq j$, is given by*

$$E\left[T_{ij}^q\right] = Pr(T_{ij}^q = 1).$$

In order for the non-rejection rate to be unambiguously defined, we have to specify the statistical test we use. In the following, we always take $q < (n-2)$ and apply the t test for zero regression coefficient as described at the end of Section 2.

If $Pr(T_{ij}^q = 1|Q)$ denotes the probability that H_0 is not rejected for a given set $Q \in \mathcal{Q}_{ij}$, then

$$Pr(T_{ij}^q = 1|Q) = \begin{cases} (1-\alpha) & \text{if } Q \text{ separates } i \text{ and } j \text{ in } G; \\ \beta_{ij.Q} & \text{otherwise;} \end{cases} \tag{2}$$

where α and $\beta_{ij.Q}$ are the probability of the first and the second type error of the test respectively. The value of α can be arbitrarily specified and we take it constant over all pairs of vertices and all elements of $\mathcal{Q}_{ij}$. The value of $\beta_{ij.Q}$ is usually unknown because it depends on the true value of the parameters. Nevertheless, the effectiveness of the *qp*-procedure depends on the statistical properties of the power function of the test, and for this reason we use a UMPU test; in particular, recall that $\beta_{ij.Q} \leq (1-\alpha)$.

The non-rejection rate for X_i and X_j can thus be computed by using the law of total probability as follows

$$\begin{aligned} Pr(T_{ij}^q = 1) &= \sum_{Q \in \mathcal{Q}_{ij}} Pr(T_{ij}^q = 1|Q) Pr(Q) \\ &= \frac{1}{m} \sum_{Q \in \mathcal{Q}_{ij}} Pr(T_{ij}^q = 1|Q). \end{aligned} \tag{3}$$

An element Q of $\mathcal{Q}_{ij}$ can either separate i and j in G or not separate them. We denote by $1_{ij}(Q)$ the indicator function that is 1 if $Q \in \mathcal{Q}_{ij}$ separates i and j in G and 0 otherwise. Furthermore, we denote by π_{ij} the proportion of elements of $\mathcal{Q}_{ij}$ which separate i and j in G so that

$$\pi_{ij} = \frac{1}{m} \sum_{Q \in \mathcal{Q}_{ij}} 1_{ij}(Q) \qquad \text{and} \qquad (1-\pi_{ij}) = \frac{1}{m} \sum_{Q \in \mathcal{Q}_{ij}} \{1 - 1_{ij}(Q)\}.$$

The second type error is defined only for the sets $Q \in \mathcal{Q}_{ij}$ such that $1_{ij}(Q) = 0$ and we define the average value of the second type error for the pair i and j over $\mathcal{Q}_{ij}$ as

$$\beta_{ij} := \frac{1}{m(1-\pi_{ij})} \sum_{Q \in \mathcal{Q}_{ij}} \beta_{ij.Q} \{1 - 1_{ij}(Q)\} \tag{4}$$

with $\beta_{ij} = 0$ if $\pi_{ij} = 1$.

We can now turn to the computation of the non-rejection rate in (3). By (2) it holds that

$$Pr(T_{ij}^q = 1) \quad = \quad \frac{1}{m} \sum_{Q \in \mathcal{Q}_{ij}} [\beta_{ij.Q}\{1 - 1_{ij}(Q)\} + (1-\alpha)1_{ij}(Q)]$$

and, by (4),

$$Pr(T_{ij}^q = 1) \quad = \quad \frac{1}{m} \left\{\beta_{ij}\, m\,(1-\pi_{ij}) + (1-\alpha)\, m\, \pi_{ij}\right\}$$

so that we obtain the final form

$$Pr(T_{ij}^q = 1) \quad = \quad \beta_{ij}\,(1-\pi_{ij}) + (1-\alpha)\,\pi_{ij}. \tag{5}$$

Equation (5) can be used to clarify the usefulness of the non-rejection rate in the statistical learning of $G^{(q)}$.

Consider first the situation in which the vertices i and j are joined by an edge in $G^{(q)} = (V, E^{(q)})$, that is, $(i, j) \in E^{(q)}$. In this case no element of $\mathcal{Q}_{ij}$ separates i and j in $G = (V, E)$ so that $\pi_{ij} = 0$ and $Pr(T_{ij}^q = 1) = \beta_{ij}$ where β_{ij} is the mean value of $\beta_{ij.Q}$ for $Q \in \mathcal{Q}_{ij}$. Since for every $Q \in \mathcal{Q}_{ij}$, $\beta_{ij.Q}$ belongs to the interval $(0, 1-\alpha)$ then also $0 \leq \beta_{ij} \leq (1-\alpha)$ but, more interestingly, β_{ij} is close to the boundary $(1-\alpha)$ only if the distribution of the $\beta_{ij.Q}$ for $Q \in \mathcal{Q}_{ij}$ is highly asymmetric on the interval $(0, 1-\alpha)$ with most of the values very close to the boundary $(1-\alpha)$; in other words, if the second type error $\beta_{ij.Q}$ is uniformly very high over $\mathcal{Q}_{ij}$. It follows that a value of $Pr(T_{ij}^q = 1)$ "close" to $1-\alpha$ means either that $(i, j) \in \bar{E}^{(q)}$ or that $(i, j) \in E^{(q)}$ but that such an edge is very difficult to identify on the basis of q-order partial correlations and of the available data. The qp-procedure aims at identifying some of, but not necessarily all the, missing edges of $G^{(q)}$ by keeping the number of wrongly removed edges low and thus trying to avoid breaking the Markov condition of the underlying probability distribution. In this perspective, it makes sense to remove the edges with $Pr(T_{ij}^q = 1)$ above a given threshold β^*. By keeping the value β^* very close to the boundary $(1-\alpha)$ the procedure will wrongly remove a present edge only when data strongly support its removal.

We now turn to the situation in which $(i, j) \in \bar{E}^{(q)}$. In this case $Pr(T_{ij}^q = 1)$ belongs to the interval $(\beta_{ij}, 1-\alpha)$ and, although it can take any value in such interval, it is important to notice that it will be closer to the boundary $(1-\alpha)$ for larger values of π_{ij}.

A missing edge is identified by the qp-procedure if its non-rejection rate is above β^*; however, the procedure does not aim at removing all missing edges and it is only important that the value of the non-rejection rate is above β^* for a large number of missing edges. A sufficient condition for this to happen is that (i) $G^{(q)}$ has a large number of missing edges and (ii) for a large number of such missing edges, the value of π_{ij} is high. Condition (i) can obviously be satisfied only if G is sparse but also the value of q plays a fundamental role because as shown in Corollary 3 a larger value of q increases the sparseness of the q-partial graph and, consequently, the values of the π_{ij}'s. On the other hand, a present edge is correctly identified by the procedure if the value of β_{ij} is below β^* and, in turn, this depends on the second type errors $\beta_{ij.Q}$ for $Q \in \mathcal{Q}_{ij}$. The statistical properties of inferential procedures involving q-order partial correlations depend on $n-q$. In the context we are considering, the sample size n cannot be easily increased but a way to make $n-q$ larger is to

decrease the value of q. We can conclude that a larger value of q allows us to identify a larger number of missing edges but also decreases the power of the statistical tests, making present edges more difficult to identify; see Section 5.3.

An interesting observation is that, in general, the effectiveness of inferential procedures in multivariate problems depends on the quantity $n-p$ being sufficiently large. The effectiveness of procedures based on the non-rejection rate also depends on $n-p$ but split such quantity into two parts:

$$(n-p) = (n-q) - (p-q) \tag{6}$$

the term $n-q$ has to be sufficiently large to guarantee the required power of statistical tests and $(p-q)$ has to be sufficiently small to guarantee the required sparseness of $G^{(q)}$, and there is a trade-off between these two requirements. However, for problems in which G is very sparse, the q-partial graph $G^{(q)}$ can be sufficiently sparse also for small values of q and, in turn, this leads to satisfactory values of $(n-q)$ even in the case $n-p$ is very small or even negative.

5.2 Description Of The Procedure

The qp-procedure is made up of five steps:

1. Specify a value $q < (n-2)$;
2. estimate the non-rejection rate $E[T_{ij}^q]$ for every pair of variables;
3. on the basis of the estimated non-rejection rates, decide whether to go
   - 3.1 on to step 4
   - 3.2 back to step 1 and modify the value of q (if possible);
4. specify a threshold β^*;
5. return a graph $\widehat{G}^{(q)}$ obtained by removing from the complete graph all the edges whose estimated non-rejection rate is greater than β^*.

We now describe every step in detail by means of an example. Figure 1 gives the image of a partial correlation matrix for 164 variables. It is made up of 20 diagonal blocks of size 12×12 and there is a 4×4 submatrix overlap between every two adjacent blocks. The associated concentration graph, that we denote by G, has 1206 edges corresponding to 9% of all possible edges. We used this matrix as a concentration matrix to generate $n = 40$ independent observations from a multivariate normal distribution with zero mean.

It is straightforward to check, by using the results of Section 4, that $G^{(20)} = G$ whereas $G^{(3)}$ is the complete graph and in this example we compare the qp-procedure for both $q = 3$ and $q = 20$.

We have thus set the value of q, and the second step of the procedure requires the estimation of the non-rejection rates. In principle, an unbiased estimate of the non-rejection rate for a pair of variables X_i and X_j can be easily obtained by first testing the hypothesis $\rho_{ij.Q} = 0$ for all $Q \in \mathcal{Q}_{ij}$, on the basis of the available data $X^{(n)}$, and then by computing the proportion of such tests in which the null hypothesis is not rejected. In practice, however, this requires the computation of $\binom{p-2}{q}$ statistical tests for every one of the $p \times (p-1)/2$ pairs of variables and may be computationally unfeasible. In order to overcome this difficulty we use a Monte Carlo method in which, for every pair X_i and

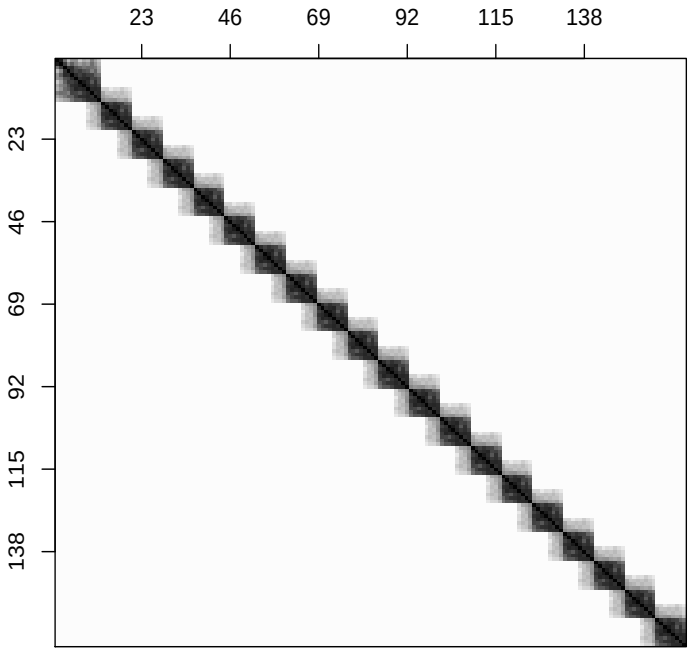

Figure 1: Image of a partial correlation matrix for 164 variables. Every entry of the matrix is represented as a gray-scaled point between zero (white points) and ± 1 (black points).

X_j, the required statistical tests are computed for a large number of sets randomly sampled from Q_{ij} according to a uniform distribution. In the example we are considering, the non-rejection rate is estimated by sampling 500 elements from Q_{ij}, for all of the 13366 pairs of variables. For the case $q = 20$, Figure 2 gives the boxplots of the estimates of the non-rejection rate for the present and missing edges of $G^{(20)}$. This picture provides a clear example of the different behavior of the non-rejection rate for present and missing edges and it is also worth recalling that that there is a large difference in the number of present and missing edges: 1206 versus 12160.

The third step involves a decision on the adequateness of the chosen value of q and possibly on the effectiveness of the non-rejection rate for the considered problem. The main tools used here are two plots that we call the *qp-hist plot* and the *qp-clique plot* respectively. The first is the histogram of estimated values of the $p \times (p-1)/2$ non-rejection rates, see Figure 3. The latter is more complex, see Figure 4, and provides information on the graphs potentially selected by specifying different values of the threshold β^*. More specifically, every circle in the plot corresponds to a graph and has three values associated with it: the threshold value used to construct the graph (horizontal axis); the number of vertices of the largest clique of the graph (vertical axis); the percentage of present edges in the graph (number inside the plot, beside the circle). Furthermore, adjacent circles are joined by a line and the dotted horizontal line corresponds to the sample size n. To understand the usefulness of this plot one has to recall that in Gaussian graphical models the real dimension of the problem is given by the size of the largest clique of the concentration graph. The *qp-clique* plot gives the dimension of the largest cliques of the graphs associated with different values of the threshold thus providing a way to assess the effectiveness of the non-rejection rate as a tool for

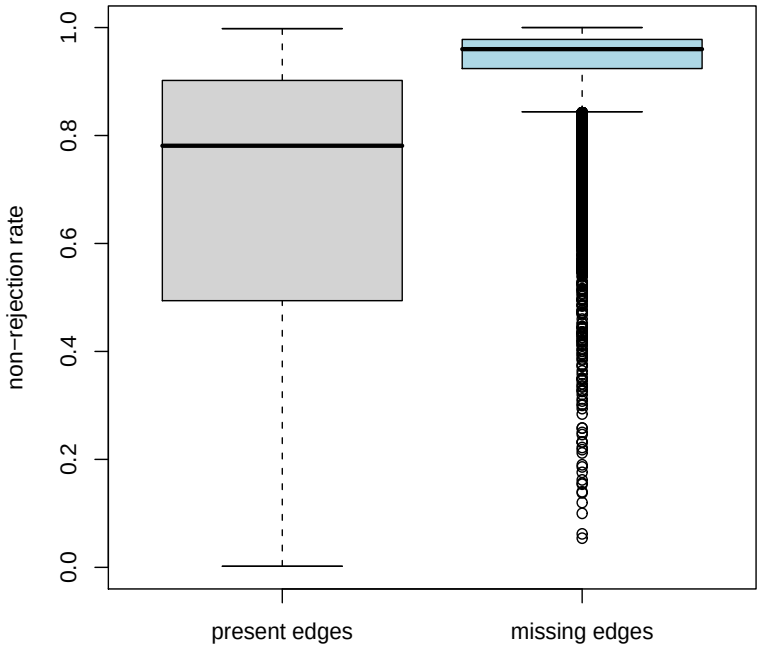

Figure 2: Boxplots of the estimated values of the non-rejection rate for the 1206 present edges and for the 12 160 missing edges of $G = G^{(20)}$.

dimensionality reduction. In particular, every circle below the dotted horizontal line corresponds to a model whose dimension is smaller than the sample size, and therefore that can be dealt with standard techniques.

We now analyze these two types of plots for the example considered. Both histograms in Figure 3 are asymmetric but the first histogram, for $q = 3$, is less asymmetric with a heavier left tail, and this is a first indication that for the case $q = 3$ the non-rejection rate may be of limited usefulness because we will not be able to remove many edges that are really missing without removing many others that should not be removed.

However, a more clear difference between the two cases can be derived from Figure 4. The dimension of models grows almost linearly for $q = 3$ whereas, for the case $q = 20$, it grows exponentially, increasing drastically only for threshold values larger than 0.975. For instance, for $q = 20$, a threshold equal to 0.9 would lead to the removal of 77% of edges, returning a graph with 23% of edges left. The same threshold for $q = 3$ would only lead to the removal of 43% of edges, returning a graph with 57% of edges left. Furthermore, the largest threshold that produces a graph for which the dimension of the largest clique is smaller than the sample size is 0.5 for $q = 3$ and 0.975 for $q = 20$. The qp-clique plot provides an indication of the sparseness of the q-marginal graph as well as of the usefulness of the non-rejection rate in statistical learning. As explained in Section 5.1, in the qp-procedure the threshold β^* has to be a value very close to one, and in the example for $q = 3$ any value close to one would lead to an insufficient dimensionality reduction. In this case, one should go back to the first step and, if possible, to increase the value of q. If the value of q cannot be increased, then one can conclude that the use of q-partial graphs is not appropriate for the problem

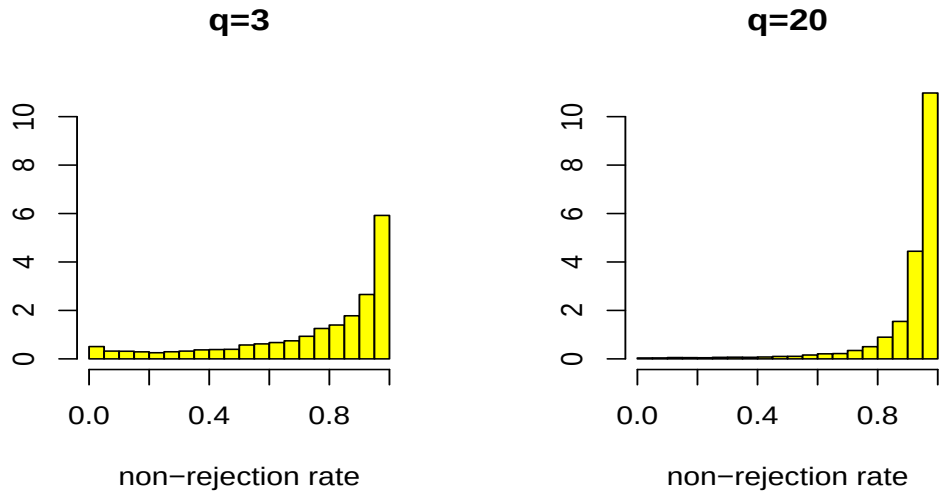

Figure 3: Histograms of the estimated values of the non-rejection rates.

under analysis. For the case $q = 20$ we can set $\beta^* = 0.975$ selecting in this way a graph $\widehat{G}^{(20)}$ with 9751 out of 13366 possible edges and whose largest clique has size 32. Figure 5 gives the adjacency matrix of $\widehat{G}^{(20)}$ and shows that, although this is clearly an overparameterized model, a substantial dimensionality reduction has been achieved while preserving the block diagonal structure of $G^{(20)}$. Indeed, only 34 of the 1206 present edges are wrongly removed corresponding to an error of 2.8%.

5.3 Experimental Results

In this section we use simulated data to describe the behavior of the non-rejection rate for different values of q, n and different degrees of sparsity of the concentration graph. Furthermore, we present the application of the procedure to a real data set.

For the simulations, we set $p = 150$ and constructed two graphs, $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ which have been randomly generated by imposing that every vertex has at most 5 and 20 adjacencies respectively. In this way, it follows from the results of Section 4 that for all $q \geq 5$ it holds that $G_1^{(q)} = G_1$ whereas for all $q \geq 20$ it holds that $G_2^{(q)} = G_2$. The graph G_1 has 375 edges whereas G_2 has 1499 edges that correspond to 3.36% and 13.4% of the 11175 possible edges respectively. Successively, an inverse covariance matrix with the zero pattern induced by G_1 has been randomly constructed (see Roverato, 2002) and then two samples, of size 20 and 150 respectively, have been randomly generated from a normal distribution with zero mean and the given covariance matrix. The same procedure was used to generate two random samples of size 20 and 50 for G_2.

We first consider G_1 and $n = 20$ and independently apply the qp-procedure with six different values of q, ranging from 1 to 17; recall that the latter is the maximum possible value of q when $n = 20$. Figure 6 shows the six qp-hist plots, which are displayed for increasing values of $(n-q)$, that is, for decreasing values of q, because the power of the statistical test we use increases with $(n-q)$. For $q = 17$ the tests have very low power and this results in a qp-hist plot where the

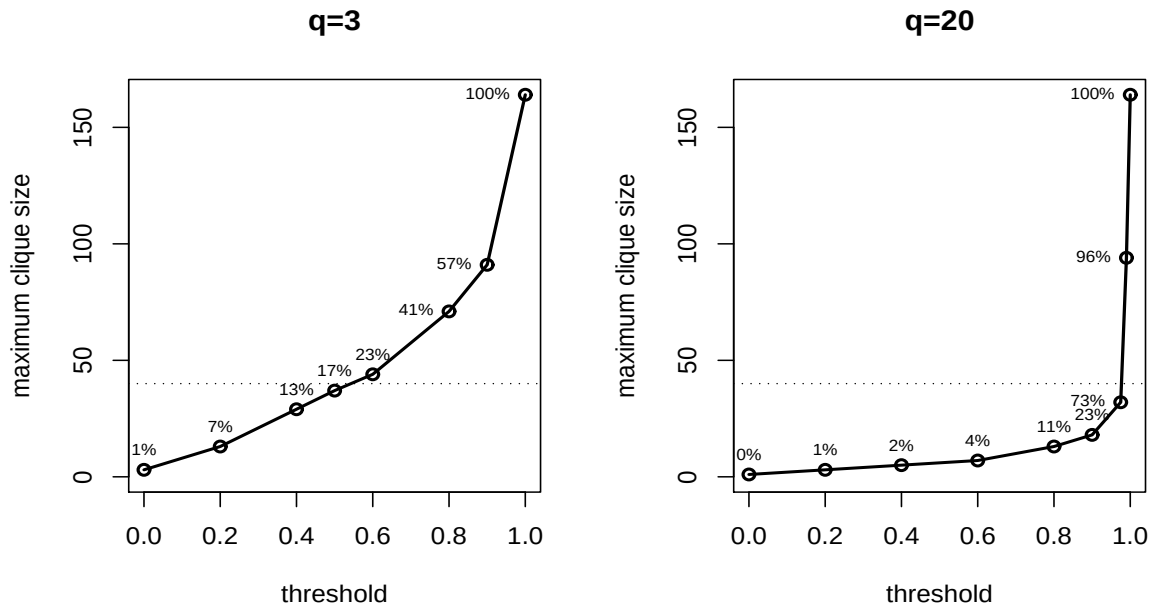

Figure 4: Plots giving the largest clique sizes of the graphs selected with different threshold values. For every graph the percentage of present edges is given and the dotted horizontal line is the sample size n.

non-rejection rate is very high for all pairs of variables. As the value of $(n-q)$ increases the qp-hist plots show heavier left tails while maintaining a strong negative asymmetric form. As Figure 7 clarifies, this happens because the distributions of the non-rejection rate for present and missing edges become more and more separated as $(n-q)$ increases. We remark that the present and missing edges in Figure 7 are relative to G_1 and not to $G_1^{(q)}$.

A numerical description of the results of these simulations is given in Tables 1 and 2. The first part of these tables gives the quantities used in the construction of the qp-clique plots: some threshold values (thr.) and, for every threshold, the size of the largest clique (l.c.) and the percentage of present edges (% pre.) of the corresponding graph. The remaining columns provide measures of goodness of the graph associated with each threshold. More specifically, "err." gives the number of wrongly removed edges, "% err." is the percentage of wrongly removed edges with respect to all the removed edges and, finally, "% imp." is the rate of improvement with respect to the random removal of edges: a learning procedure based on the random removal of edges would lead to a relative error whose expected value is the proportion of edges in the graph, that is 3.36% for G_1, and the improvement rate of a graph is the relative difference between "% err." and the proportion of present edges in the concentration graph. We remark that the last three columns of these tables are not available in real applications where the concentration graph is unknown.

Figures 6 and 7 seem to indicate that the value of q should be chosen as low as possible; nevertheless, as described in Section 5.1 the value of q should not be chosen too small in order to

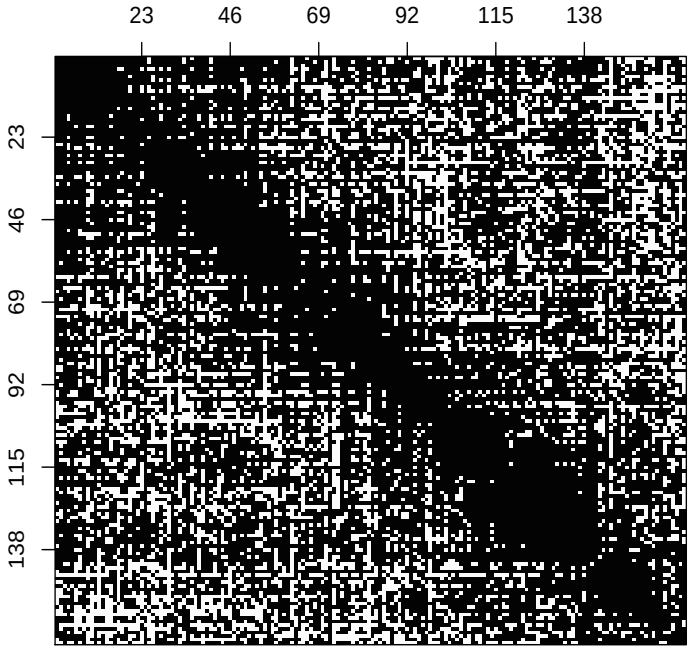

Figure 5: Adjacency matrix of the graph selected by the qp-procedure with $q = 20$ and $\beta^* = 0.975$. Black points are present edges (value 1 in the adjacency matrix) and white points missing edges (value 0 in the adjacency matrix).

guarantee an adequate sparseness of $G_1^{(q)}$. If in Tables 1 and 2 one takes, for the different values of q and $n = 20$, the largest threshold corresponding to a graph whose largest clique size is smaller than n, then the best solution is provided by $q = 10$ with a graph in which 6601 edges are missing, the largest clique has size 13 and the absolute error is 97 with a 56.21% improvement rate. However, also the case $q = 5$ provides a good solution with a graph in which 7194 edges are missing, the largest clique has size 19 and the absolute error is 103 with a 57.33% improvement rate. A value of q equal either to 5 or to 10 represents the most natural choice in the trade-off between $(n-q)$ and $(p-q)$ in (6), however we notice that, apart from $q = 17$ where the relative improvement is only 38.32%, all the other considered values of q provide satisfying solutions. This seems to suggest that the qp-procedure is not very sensitive to the choice of q. We can conclude that the qp-procedure is very effective despite the fact that we are considering an extremely challenging problem where the sample size is very small, $n = 20$, compared to the number of variables, $p = 150$. In order to show the behavior of the non-rejection rate as the sample size increases, in Figure 8 and Table 2 we provide an example in which the sample size is larger, $n = 150$, but still too low to permit the computation of sample full-order partial correlations. The boxplots in Figure 8 highlights the great effectiveness of the non-rejection rate in this case. Table 2 shows that one can either select the largest graph manageable with standard techniques, choosing in this way a graph with only 12 wrongly removed edges, or select a sparser graph; for instance, the threshold 0.60 gives a graph with 9365 out of 11 175 missing edges, absolute error 85 and a 72.94% improvement rate. It is also interesting to compare Figure 8 with the case $q = 17$ in Figures 6 and 7.

n	q	thr.	l.c.	% pre.	err.	% err.	% imp.
20	1						
		0.30	10	10.4	187	1.87	44.37
		0.60	13	14.2	177	1.85	45.00
		0.80	14	17.1	169	1.82	45.63
		0.85	14	18.5	166	1.82	45.68
		0.90	15	21.3	155	1.76	47.50
		0.95	17	27.2	136	1.67	50.18
		0.97	19	32.4	123	1.63	51.51
		0.98	19	36.9	111	1.58	53.05
		0.99	22	46.9	88	1.48	55.81
20	3						
		0.30	7	4.7	228	2.14	36.18
		0.60	9	10.1	191	1.90	43.35
		0.80	12	16.7	170	1.83	45.59
		0.85	14	19.8	156	1.74	48.15
		0.90	14	24.5	143	1.69	49.50
		0.95	17	34.2	120	1.63	51.36
		0.97	20	42.7	96	1.50	55.36
		0.98	22	50.4	79	1.43	57.49
		0.99	27	63.8	53	1.31	60.99
20	5						
		0.30	6	2.9	235	2.16	35.49
		0.60	8	6.9	195	1.87	44.13
		0.80	11	13.8	163	1.69	49.57
		0.85	12	17.3	152	1.65	50.98
		0.90	13	22.9	138	1.60	52.27
		0.95	19	35.6	103	1.43	57.33
		0.97	23	47.1	83	1.40	58.15
		0.98	28	57.0	65	1.35	59.70
		0.99	36	74.2	38	1.32	60.80

Table 1: Graph $G_1 = (V, E_1)$. Numerical description of the output of the qp-procedure applied for $n = 20$ and $q = 1, 3, 5$. The first part of the table gives the quantities used in the construction of the qp-clique plots: some threshold values (thr.) and, for every threshold, the size of the largest clique (l.c.) and the percentage of present edges (% pre.) of the corresponding graph. The last three columns give the number of wrongly removed edges (err.), the percentage of wrongly removed edges with respect to all the removed edges (% err.) and the rate of improvement with respect to the random removal of edges (% imp.).

n	q	thr.	l.c.	% pre.	err.	% err.	% imp.
20	10						
		0.30	4	0.7	313	2.82	15.94
		0.60	5	2.5	244	2.24	33.26
		0.80	7	7.6	199	1.93	42.59
		0.85	8	11.4	174	1.76	47.66
		0.90	9	19.0	149	1.65	50.93
		0.95	13	40.9	97	1.47	56.21
		0.97	25	67.2	58	1.58	52.83
		0.98	45	85.6	26	1.62	51.82
		0.99	99	98.1	6	2.82	16.06
20	15						
		0.30	2	0.1	371	3.32	1.03
		0.60	3	0.3	347	3.11	7.20
		0.80	5	1.0	303	2.74	18.36
		0.85	6	1.9	278	2.54	24.45
		0.90	6	5.5	233	2.21	34.28
		0.95	11	45.5	104	1.71	49.08
		0.97	50	94.2	10	1.53	54.29
		0.98	124	99.6	0	0.00	100.00
		0.99	150	100.0	0	0.00	100.00
20	17						
		0.30	1	0.0	375	3.36	0.00
		0.60	1	0.0	375	3.36	0.00
		0.80	1	0.0	375	3.36	0.00
		0.85	2	0.1	366	3.28	2.31
		0.90	3	0.4	339	3.05	9.23
		0.95	11	53.3	108	2.07	38.32
		0.97	89	98.7	2	1.38	58.90
		0.98	149	99.9	0	0.00	100.00
		0.99	150	100.0	0	0.00	100.00
150	17						
		0.30	6	7.0	118	1.14	66.17
		0.60	9	16.2	85	0.91	72.94
		0.80	13	29.4	60	0.76	77.32
		0.85	15	35.6	53	0.74	78.07
		0.90	17	44.3	44	0.71	78.93
		0.95	23	60.4	34	0.77	77.10
		0.97	34	70.7	30	0.92	72.72
		0.98	44	77.5	21	0.84	75.09
		0.99	62	86.3	12	0.78	76.61

Table 2: Graph $G_1 = (V, E_1)$. Numerical description of the output of the qp-procedure applied with different values of n and q. See Table 1 for a description of columns.

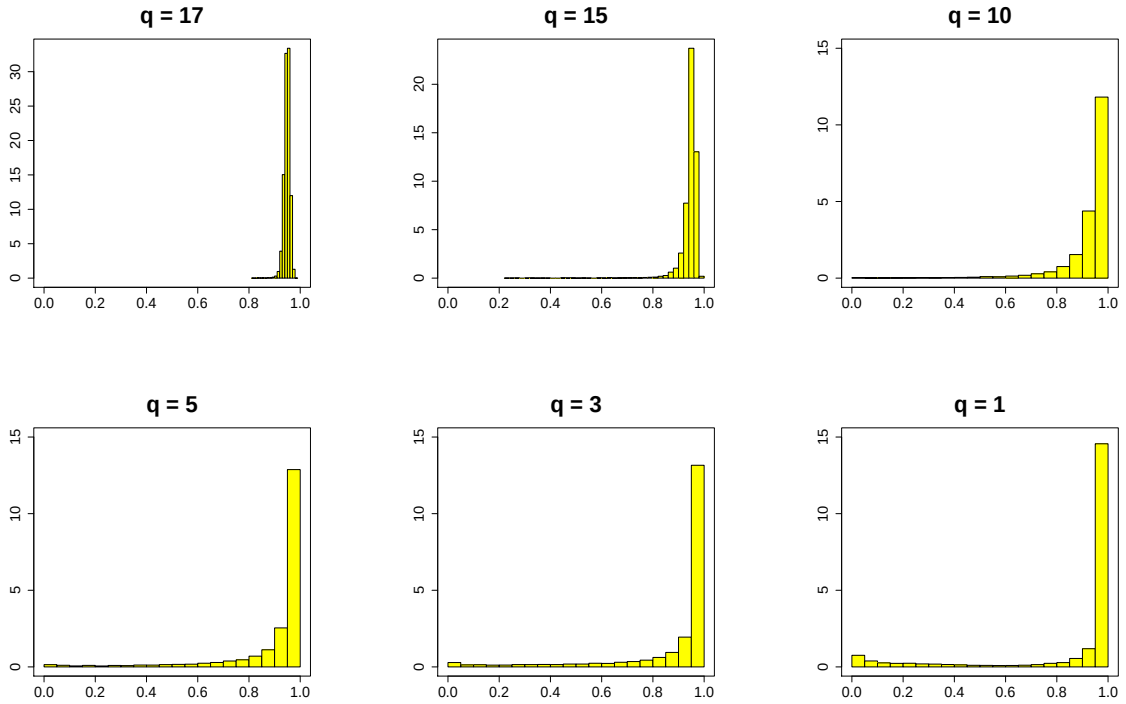

Figure 6: qp-hist plots for $G_1 = (V, E_1)$ with $n = 20$.

We now apply the qp-procedure for the case with concentration graph G_2, $n = 20, 50$ and $q = 5, 10$; see Figure 9 and Table 3. The graph G_2 is not sparse and both $G_2^{(5)}$ and $G_2^{(10)}$ are even more dense, and this affects the shape of the qp-hist plots in Figure 9. Indeed, all the three histograms are clearly less asymmetric than the corresponding histograms in Figure 6; note also that this is less evident in the case $n = 20$ and $q = 10$ because the quantity $(n - q)$ is smaller than in the other two cases.

We deem that this kind of behavior of the qp-hist plot should be read as an indication that the considered q-partial graphs do not provide satisfying approximations of the required concentration graphs. Hence, if the value of q cannot be increased then we suggest that the application of any learning procedure based on limited-order partial correlations should be avoided for the problem under analysis.

We close this section applying the qp-procedure to a subset of the gene expression data from the study by West et al. (2001). This subset was extracted and analysed originally by Jones et al. (2005) and contains the expression profiles for $p = 150$ genes associated with the estrogen receptor pathway coming from $n = 49$ breast tumor samples.

We have applied the qp-procedure with $q = 20$ and the qp-hist and qp-clique plots, given in Figure 10, provide a strong indication that $G^{(20)}$ is sparse. Hence, we set $\beta^* = 0.975$ and, in this way, we identify a graph with 7240 out of 11 175 possible edges and whose largest clique has size 24 which can be taken as an estimate of the maximum size of the highly interconnected sets of interacting genes. Such sets are a class of the so-called network motifs (Milo et al., 2002) which are characteristic network patterns whose identification can be used to draw hypotheses on basic cellular

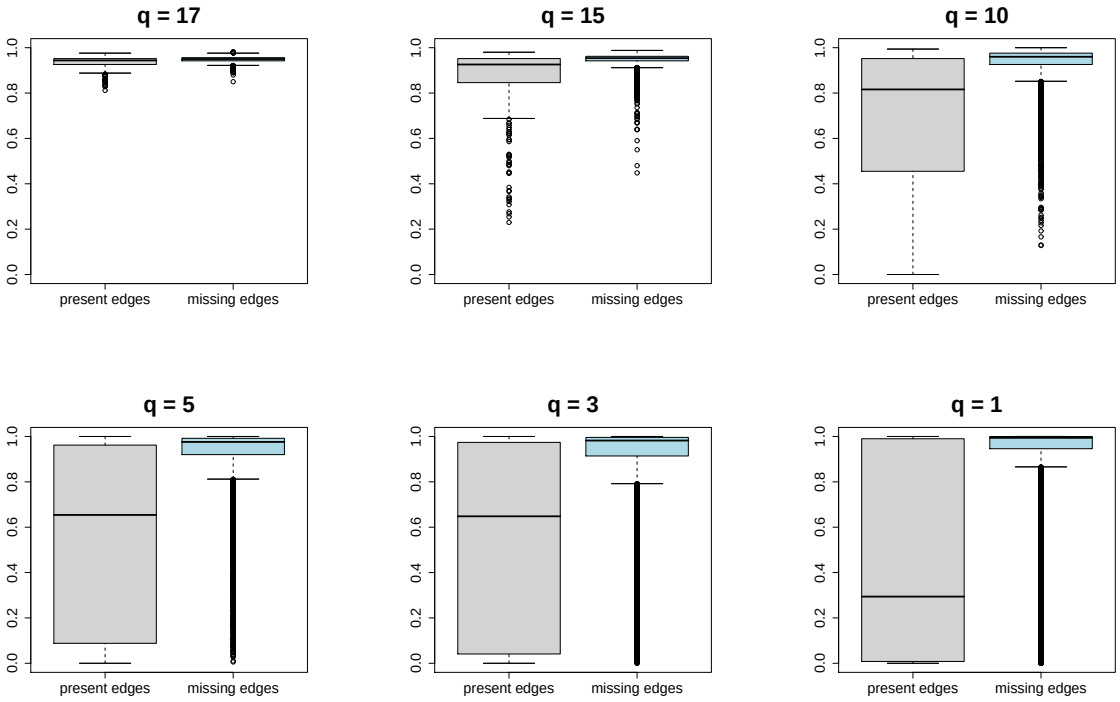

Figure 7: Distribution of the non-rejection rate for present and missing edges of $G_1 = (V, E_1)$, to be associated with the corresponding histograms in Figure 6.

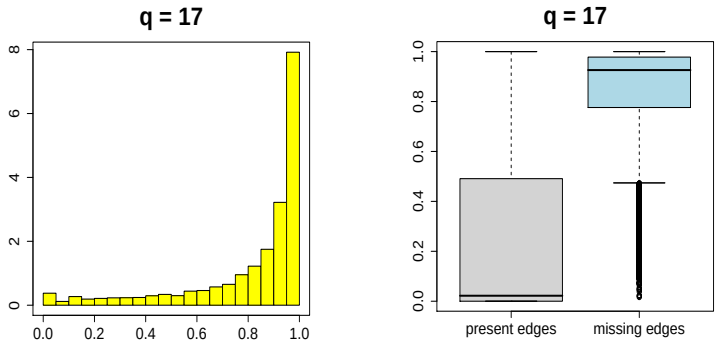

Figure 8: qp-hist plot and associated distributions of the non-rejection rate for present and missing edges of $G_1 = (V, E_1)$, resulting from the application of the qp-procedure where $n = 150$ and $q = 17$.

mechanisms (Yeger-Lotem et al., 2005). Note that the theory of q-partial graphs developed in this paper, and implemented through the qp-procedure, allows us to obtain this estimate, and eventually explore other ones, in relationship to the amount of true interactions we are willing to remove and

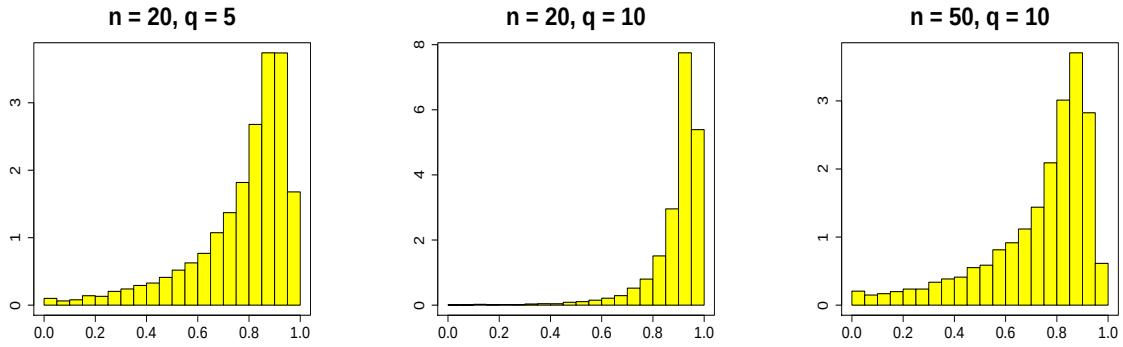

Figure 9: *qp*-hist plots and associated distributions of the non-rejection rate for present and missing edges of $G_2 = (V, E_2)$, resulting from the application of the *qp*-procedure for different values of n and q.

the dimension of the data. Such a feature may be a critical piece of information when dealing with real data for which we lack background knowledge on its underlying structure of interactions.

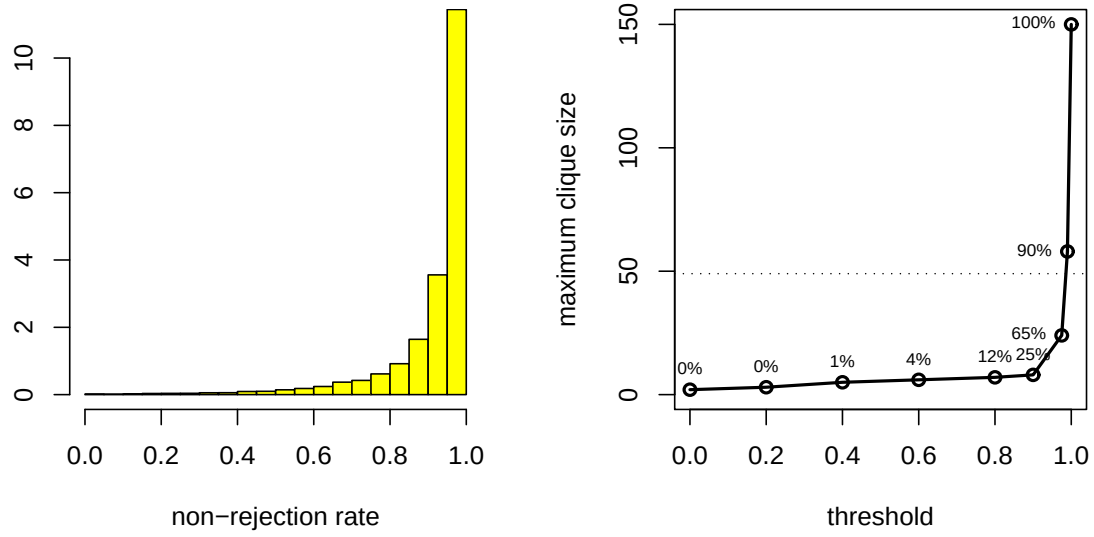

Figure 10: Estrogen receptor data of West et al. (2001): *qp*-hist and *qp*-clique plots for $q = 20$.

n	q	thr.	l.c.	% pre.	err.	% err.	% imp.
20	5						
		0.30	5	3.6	1342	12.45	6.78
		0.60	10	15.7	1099	11.66	12.72
		0.80	21	40.8	735	11.11	16.82
		0.85	29	54.2	580	11.33	15.16
		0.90	55	72.9	328	10.84	18.89
		0.95	103	91.6	90	9.59	28.18
		0.97	123	96.5	31	7.81	41.55
		0.98	134	98.3	23	12.30	7.94
		0.99	144	99.5	6	10.00	25.15
20	10						
		0.30	3	0.5	1451	13.05	2.36
		0.60	5	2.8	1333	12.27	8.13
		0.80	7	11.9	1094	11.12	16.77
		0.85	9	19.5	971	10.80	19.19
		0.90	12	34.3	758	10.32	22.72
		0.95	43	73.1	292	9.69	27.44
		0.97	88	92.4	76	8.91	33.31
		0.98	116	97.8	20	8.16	38.90
		0.99	141	99.7	2	6.90	48.38
50	10						
		0.30	6	6.0	1171	11.14	16.59
		0.60	9	21.4	869	9.89	25.96
		0.80	17	49.2	518	9.13	31.69
		0.85	27	64.3	351	8.79	34.20
		0.90	62	82.8	152	7.91	40.81
		0.95	120	96.9	27	7.87	41.08
		0.97	134	99.4	7	9.59	28.23
		0.98	143	99.8	3	12.50	6.44
		0.99	148	100.0	0	0.00	100.00

Table 3: Graph $G_2 = (V, E_2)$. Numerical description of the output of the qp-procedure applied for different values of n and q. See Table 1 for a description of columns.

6. Discussion

This paper provides two main contributions: the theory related to q-partial graphs and the qp-procedure.

The theory of q-partial graphs clarifies the connection between the sparseness of the concentration graph and the usefulness of marginal distributions in structure learning, under the assumption of faithfulness.

The qp-procedure is designed to learn q-partial graphs and overcomes the main drawbacks of the existing procedures based on limited-order partial correlations. Furthermore, our procedure has several advantages. Most importantly, it is robust with respect to the assumption of faithfulness because the estimation of the non-rejection rate is based on a large number of statistical tests involving different marginal distributions and, therefore, a zero q-order partial correlation deriving from the

lack of faithfulness has a very weak impact on the resulting estimate. Apart from faithfulness, the *qp*-procedure does not require any additional assumptions with respect to traditional structure learning procedures and, in particular, the sparseness of the concentration graph, despite being crucial for the effectiveness of the procedure, is not assumed but exploited when present. In the case the *qp*-hist and *qp*-clique plots provide and indication that the concentration graph is not sparse, then this should be read as a warning on the real usefulness of limited-order partial correlations in the problem under analysis. The fact that the *qp*-procedure is designed to select an overparameterized model might be regarded as a limitation, but in fact we deem that this is a useful feature that adds additional flexibility in its use. Indeed, the *qp*-procedure can be used as an explorative tool to assess the sparseness of the concentration graph and, therefore, the usefulness of *q*-partial correlations in structure learning. Furthermore, the result of the procedure may be applied to obtain a shrinkage estimate of the covariance matrix useful both in the case n is larger, but close, to p and in the case n is smaller than p. Finally, the set of all the submodels of the selected model may identify a restricted search space where a traditional structure learning procedure, either in a Bayesian or in a frequentist approach to inference, can be applied. In Gaussian graphical models it is assumed that X_V follows a multivariate normal distribution, and the normality of microarray data is a disputed question. We refer to Wit and McClure (2004; Section 6.2.2) for a discussion of this point, but we remark that the non-rejection rate is a quantity that can be obtained from any test for conditional independence computed on marginal distributions, and therefore it constitutes a general tool that can be used also outside the multivariate normal case.

The *qp*-procedure, jointly with other functions showing the *qp*-hist and *qp*-clique plots, has been implemented in a package, named *qp*, for the statistical software R (`http://www.r-project.org`). This package can be downloaded from The Comprehensive R Archive Network (CRAN) at `http://cran.r-project.org/src/contrib/PACKAGES.html`.

The *qp*-procedure is implemented in this package through the R and C programming languages requiring 10 minutes in a laptop 1.33GHz PowerPC G4 with 1.25 Gbyte RAM running Mac OS X, as well as in a desktop Intel 1.60GHz P4 with 1 Gbyte RAM running Linux, to perform the calculations of one of the simulations involving $p = 150$ variables, $n = 50$ observations, and $q = 15$ sampling 500 conditioning subsets to estimate the non-rejection rate for each of the 11 175 adjacencies. Note also that the $p \times (p-1)/2$ non-rejection rates could be estimated in parallel and thus such an implementation would greatly improve the performance.

Acknowledgments

We would like to thank David Madigan and David Edwards for useful discussions and the anonymous reviewers whose remarks and suggestions have improved this paper. Part of this paper was written when the second author was visiting the first author at the Universitat Pompeu Fabra supported by a mobility grant (ref. SAB2003–0197) from the Spanish Ministerio de Educación y Ciencia (MEC). Financial support to the second author has also been provided by MIUR, grant number 134079, 2005 and by the MIUR-FISR grant number 2982/Ric (Mitica). The first author is a researcher from the Ramon y Cajal program of the Spanish MEC (ref. RYC–2006–000932).

Appendix A. Graph Theory

In this appendix we present the graph theory required for this paper and, in particular, we introduce the novel concept of *outer connectivity* that is used in Section 4 to describe the properties of q-partial graphs. We refer to Cowell et al. (1999) for a full account of graph theory usually applied in graphical models, to Diestel (2005) for the theory relating separators and independent paths and, finally, to Rosenberg and Heath (2005) for a comprehensive description of the techniques for obtaining upper and lower bounds on the sizes of graph separators.

An undirected *graph* is a pair $G = (V,E)$, where $V = \{1,\ldots,p\}$ is a finite set of *vertices* and in this paper E, called the *edge set*, is a subset of the set of unordered distinct pair of vertices. If two vertices $i, j \in V$ form an edge then we say that i and j are *adjacent* and write $(i, j) \in E$; recall that edges are unordered pairs, so that $(i, j) = (j, i)$. Graphs are usually represented by drawing a dot for each vertex and joining two of these dots by a line if the corresponding two vertices form an edge; see Figure 11 for a few examples. For a subset $A \subseteq V$ the *subgraph* of G induced by A is $G_A = (A, E_A)$ with $E_A = E \cap (A \times A)$. For two graphs with common vertex set, $G = (V,E)$ and $G' = (V,E')$, we say that G' *is larger than* G, and write $G \subseteq G'$, if $E \subseteq E'$; when the inclusion is strict, that is, $E \subset E'$, we write $G \subset G'$. The *boundary* of a vertex $v \in V$, denoted by $\mathrm{bd}_G(v)$, is the set of vertices adjacent to v. A subset $C \subseteq V$ with all vertices being mutually adjacent is called *complete*, and when V is complete then we say that G is complete. A subset $C \subseteq V$ is called a *clique* if it is maximally complete, that is, C is complete, and if $C \subset D$, then D is not complete. An undirected graph can be identified by the set $\mathcal{C}$ of its cliques. The set $\bar{E}$ is the *set of missing edges* of G; that is, for a pair $i, j \in V$, $(i, j) \in \bar{E}$ if and only if $i \neq j$ and $(i, j) \notin E$. A *path* of length $l > 0$ from v_0 to v_l is a sequence $v_0, v_1, \ldots, v_l$ of distinct vertices such that $(v_{k-1}, v_k) \in E$ for all $k = 1,\ldots,l$. Two or more paths from v_0 to v_l are *independent* if they have no common vertices other then v_0 and v_l. We can define an equivalence relation on V as

$$i \sim_p j \Leftrightarrow \text{ there is a path } v_0, v_1, \ldots, v_l \text{ with } v_0 = i, v_l = j.$$

The subgraphs induced by the equivalence classes are the *connected components* of G. If there is only one equivalence class, we say that G is connected. The subset $U \subseteq V$ is said to separate $I \subseteq V$ from $J \subseteq V$ if for every $i \in I$ and $j \in J$ all paths from i to j have at least one vertex in U. For a pair of vertices $i \neq j$ with $(i, j) \in \bar{E}$, a set $U \subseteq V$ is called a $\{i, j\}$-*separator* if it separates $\{i\}$ and $\{j\}$ in G. If either $i \in U$ or $j \in U$ then we say that U is *trivial*. If no proper subset of U is a $\{i, j\}$-separator we say that U is *minimal*; see also Cowell et al. (1999). Note that the unique possible minimal $\{i, j\}$-separators that are trivial are $\{i\}$ and $\{j\}$. Hereafter, to stress that a separator is nontrivial and minimal we denote it by S; furthermore, we denote by $\mathcal{S}_{(i,j|G)}$ the set of all nontrivial minimal $\{i, j\}$-separators in G, so that $\mathcal{S}_{(i,j|G)} = \{\emptyset\}$ if and only if i and j are in different connected components. There is a close connection between the concepts of connectivity and separation: the dimension of the smallest $\{i, j\}$-separator, that is the cardinality of the smallest (possibly non unique) set in $\mathcal{S}_{(i,j|G)}$, is called the *connectivity of i and j* because it represents both the maximum number of independent paths between i and j in G and the minimum number of vertices that need to be removed from G to make i and j disconnected (see Theorem 3.3.1 of Diestel, 2005). In order to deal with q-partial graphs we need to introduce a slightly different definition of connectivity of two vertices.

Definition 3 *Let $i \neq j$ be a pair vertices of an undirected graph $G = (V,E)$. The* outer connectivity of i and j *is defined as*

$$d(i,j|G) = \min_{S \in \mathcal{S}_{(i,j|G_{ij})}} |S|$$

where G_{ij} is the graph with vertex set V and edge set $E_{ij} = E \setminus \{(i,j)\}$.

Hence, $d(i,j|G)$ is the connectivity of i and j in G_{ij}. The latter graph is constructed by removing the edge (i,j) from G, so that if $(i,j) \in \overline{E}$ then $G = G_{ij}$. The idea here is that the edge (i,j) represents an *inner*, or direct, connection between i and j and it should not be considered when *outer*, or indirect, connectivity is of concern.

Example 1 *For the vertex set $V = \{1,\ldots,6\}$ let $G_i = (V,E_i)$, $i = 1,\ldots,3$ be the graphs in Figure 11 and let G_4 be the complete graph. Then*

- $d(2,3|G_i) = 0$ *for* $i = 1,2,3$ *whereas* $d(2,3|G_4) = 4$*;*
- $d(1,6|G_1) = 0$, $d(1,6|G_i) = 1$ *for* $i = 2,3$ *whereas* $d(1,6|G_4) = 4$*;*
- $d(3,4|G_i) = 0$ *for* $i = 1,2$ *whereas* $d(3,4|G_3) = 1$*;*
- $d(3,6|G_1) = 0$, $d(3,6|G_2) = 1$, $d(3,6|G_3) = 2$.

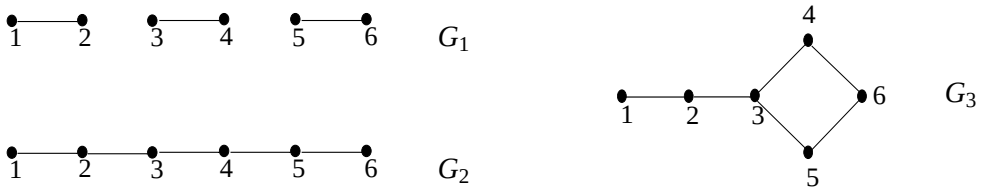

Figure 11: Examples of undirected graph.

Computing the connectivity of two vertices is known to be a NP-hard problem, however several algorithms are available to derive both upper and lower bounds to this number; see Rosenberg and Heath (2001). Here we remark that the cardinality of any $\{i,j\}$-separator in G_{ij} is an upper bound to the connectivity of i and j; consequently, since $\mathrm{bd}_{G_{ij}}(i)$ and $\mathrm{bd}_{G_{ij}}(j)$ are both $\{i,j\}$-separators in G_{ij}, then the number

$$\widetilde{d}(i,j|G) := \min\{|\mathrm{bd}_{G_{ij}}(i)|, |\mathrm{bd}_{G_{ij}}(j)|\} \tag{7}$$

provides an easy-to-compute upper bound to the outer connectivity of i and j; formally

$$d(i,j|G) \leq \widetilde{d}(i,j|G) \quad \text{for all} \quad i,j \in V;\ i \neq j. \tag{8}$$

It is useful to consider separately the pairs of vertices that define an edge in G from the pairs of vertices that are not adjacent in G. Hence, we define the *outer connectivity of the edges of $G = (V,E)$* as

$$d(E|G) := \max_{(i,j) \in E} d(i,j|G),$$

with the understanding that $d(E|G) = 0$ if $E = \emptyset$; that is if G as no edges. Similarly, the *outer connectivity of the missing edges of* $G = (V,E)$ is defined as

$$d(\bar{E}|G) := \max_{(i,j)\in E} d(i,j|G), \tag{9}$$

with the understanding that $d(\bar{E}|G) = 0$ if $\bar{E} = \emptyset$; that is if G is complete. Finally, the *outer connectivity of* $G = (V,E)$ is given by

$$\begin{aligned} d(G) &:= \max_{i,j\in V; i\neq j} d(i,j|G) \\ &= \max\{d(E|G), d(\bar{E}|G)\}. \end{aligned}$$

It is a straightforward consequence of (8) that the quantities

$$\widetilde{d}(\bar{E}|G) := \max_{(i,j)\in E} \widetilde{d}(i,j|G), \tag{10}$$

$$\widetilde{d}(E|G) := \max_{(i,j)\in E} \widetilde{d}(i,j|G), \tag{11}$$

and

$$\widetilde{d}(G) := \max\left\{\widetilde{d}(E|G), \widetilde{d}(\bar{E}|G)\right\} \tag{12}$$

are upper bounds to $d(\bar{E}|G)$, $d(E|G)$ and $d(G)$ respectively.

Example 2 *For the graphs in Figure 11 it holds that*

G_1: $d(\bar{E}|G_1) = 0,\ d(E|G_1) = 0,\ d(G_1) = 0$;

G_2: $d(\bar{E}|G_2) = 1,\ d(E|G_2) = 0,\ d(G_2) = 1$;

G_3: $d(\bar{E}|G_3) = 2,\ d(E|G_3) = 1,\ d(G_3) = 2$;

There is no strict distinction between *sparse* and *dense* graphs, however a sparse graph can be informally defined as a graph in which the number of edges is much less than the possible number of edges. Thus the complete graph is dense and the graph in which the edge set is empty is sparse; furthermore, if $G \subset G'$ than we can say that G is sparser than G'. Since G is obtained by removing edges from the larger graph G' the intuition suggests that G has a smaller number of independent paths between vertices and consequently smaller values of outer connectivity. This is formally stated in the following theorem.

Theorem 6 *Let $G = (V,E)$ and $G' = (V,E')$ be two undirected graphs such that $G \subseteq G'$. For any pair of vertices $i, j \in V$ with $i \neq j$ it holds that*

$$d(i,j|G) \leq d(i,j|G') \tag{13}$$

furthermore,

$$d(E|G) \leq d(E'|G') \qquad \textit{and} \qquad d(G) \leq d(G'). \tag{14}$$

Proof Let S be a smallest nontrivial $\{i,j\}$-separator in G'_{ij} so that $d(i,j|G') = |S|$ and every path from i to j in G'_{ij} has a vertex in S. By construction, every edge in G_{ij} is an edge in G'_{ij} and this implies that every path form i to j in G_{ij} is also a path from i to j in G'_{ij} and, consequently, that every path form i to j in G_{ij} has a vertex in S. Thus, S is a nontrivial $\{i,j\}$-separator in G_{ij} so that $d(i,j|G) \leq |S| = d(i,j|G')$, that proves (13). We consider now the first inequality in (14). Let $i,j \in V$ be two vertices such that $(i,j) \in E$ and $d(E|G) = d(i,j|G)$; recall that $(i,j) \in E$ implies $(i,j) \in E'$. Then, $d(E|G) = d(i,j|G) \leq d(i,j|G') \leq d(E'|G')$ where the first inequality holds by (13) and the second holds for every $(i,j) \in E'$. A similar reasoning can be used to prove the second inequality in (14): if i and j are such that $d(G) = d(i,j|G)$, then $d(G) = d(i,j|G) \leq d(i,j|G') \leq d(G')$ where the first inequality holds by (13) and the second is always true. ■

Note that neither the inequality $d(\bar{E}|G) \geq d(\bar{E}'|G')$ nor the inequality $d(\bar{E}|G) \leq d(\bar{E}'|G')$ are satisfied in general. For a counterexample, let $G_1 = (V,E_1)$ and $G_3 = (V,E_3)$ be the empty and the complete graph respectively, and let $G_2 = (V,E_2)$ be the graph with exactly one edge missing. Clearly, $G_1 \subseteq G_2 \subseteq G_3$, however

$$\{d(\bar{E}_1|G_1) = 0\} \leq \{d(\bar{E}_2|G_2) = p-2\} \quad \text{and} \quad \{d(\bar{E}_2|G_2) = p-2\} \geq \{d(\bar{E}_3|G_3) = 0\}.$$

References

A.J. Butte, P. Tamayo, D. Slonim, T.R. Golub and I.S. Kohane. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97(22): 12182-12186, 2000.

R.G. Cowell, A.P. Dawid, S.L. Lauritzen and D.J. Spiegelhalter. *Probabilistic networks and expert systems*. Springer-Verlag, New York, 1999.

D.R. Cox and N. Wermuth. Linear dependencies represented by chain graphs (with discussion). *Statist. Sci.*, 8: 204–283, 1993.

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

A. de la Fuente, N. Bing, I. Hoeschele and P. Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20: 3565-3574, 2004.

A.P. Dempster. *Elements of continuous multivariate analysis*. Addison-Wesley, Reading, Massachusetts, 1969.

A.P. Dempster. Covariance selection. *Biometrics*, 28: 157–75, 1972.

R. Diestel. (2005). *Graph theory*. Springer-Verlag, Heidelberg, 2005.

A. Dobra, C. Hans, B. Jones, J.R. Nevins and M. West. Sparse graphical models for exploring gene expression data. *J. Mult. Anal.* 90: 196-212, 2004.

M. Drton and M.D. Perlman. Model selection for Gaussian concentration graphs. *Biometrika*, 91(3): 591–602, 2004.

M. Drton and T. Richardson. Iterative conditional fitting for estimation of a covariance matrix with zeros. Technical Report no. 469, Department of Statistics, University of Washington, 2004.

R.L. Dykstra. Establishing the positive definiteness of the sample covariance matrix. *Ann. Math. Statist.*, 41(6): 2153–2154, 1970.

D.E. Edwards. *Introduction to graphical modelling*. Springer-Verlag, New York, 2000.

N. Friedman. Inferring cellular network using probabilistic graphical models. *Science*, 33: 799–805, 2004.

B. Jones, A. Dobra, C. Carvalho, C. Hans, C. Carter and M. West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20(4): 388–400, 2005.

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

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

E.L. Lehmann. *Testing statistical hypotheses, 2nd edition*. Wiley, New York, 1986.

P.M. Magwene and J. Kim. Estimating genomic coexpression networks using first-order conditional independence. *Genome Biology*, 5: R100, 2004.

A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R.D. Favera and A. Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks ina mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006.

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298: 824–827, 2002.

J. Pearl. *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann, San Mateo, 1988.

A.L. Rosenberg and L.S. Heath. *Graph separators, with applications*. Kluwer Academic Publishers, New York, 2001.

A. Roverato. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Statist.*, 29: 391–411, 2002.

J. Schäfer and K. Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6): 754-764, 2005a.

J. Schäfer and K. Strimmer. Learning large-scale graphical Gaussian models from genomic data. In: J.F. Mendes. (Ed.). Proceeding of *Science of Complex Networks: from Biology to the Internet and WWW* (CNET 2004), Aveiro, PT, (Publisher: The American Institute of Physics), 2005b.

J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1): article 32, 2005c.

R. Steuer, J. Kurths, O. Fiehn and W. Weckwerth. Interpreting correlations in metabolomic networks. *Bioch. Soc. Trans.*, 31: 1476-1478, 2003a.

R. Steuer, J. Kurths, O. Fiehn and W. Weckwerth. Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19: 1019-1026, 2003b.

M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson, J.R. Marks and J.R. Nevings. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, 98(20): 11462–11467, 2001.

A. Wille and P. Bühlmann. Low-order conditional independence graphs for inferring genetic networks. *Statistical Applications in Genetics and Molecular Biology*, 5(1): article 1, 2006.

A. Wille, P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelić, P. von Rohr, L. Thiele, E. Zitzler, W. Gruissem and P. Bühlmann. Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biology*, 5:R92, 2004.

E. Wit and J. McClure. *Statistics for microarrays. Design, analysis and inference.* Wiley, Chichester, 2004.

J. Whittaker. *Graphical models in applied multivariate statistics*. Wiley, Chichester, 1990.

F. Wong, C.K. Carter and R. Kohn. Efficient estimation of covariance selection models. *Biometrika*, 90: 809–830, 2003.

R. Yang, and J.O. Berger. Estimation of a covariance matrix using the reference priors. *Ann. Statist.*, 3: 1195–1211, 1994.

E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R.Y. Pinter, U. Alon and H. Margalit. Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction. *Proc. Natl. Acad. Sci.*, 101(16): 5934–5939, 2004.