STATISTICA, anno LXVI, n. 4, 2006

STRUCTURAL LEARNING OF GAUSSIAN GRAPHICAL MODELS FROM MICROARRAY DATA WITH p LARGER THAN n

A. Roverato, R. Castelo

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 propose a novel q-partial-correlations based procedure, qp-procedure hereafter, for structure learn-

ing 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. The paper is organized as follows. Section 2 gives the background on graph theory, Gaussian graphical models and q-partial graphs. Section 3 describes the application of Gaussian graphical models to biomolecular networks. The qp-procedure is introduced in Section 4 where instances of its application to both simulated and real data are given and, finally, Section 5 contains a brief discussion.

2. BACKGROUND

2.1. *Graph theory*

We present here the graph theory required for this paper; we refer to Cowell *et al.* (1999) for a full account of graph theory usually applied in graphical models.

An undirected graph is a pair $G = (V,E)$, where $V = \{1, ..., p\}$ is a finite set of vertices and 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 $(i, j) = (j, i)$. 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, i.e. $E \subset E'$, we write $G \subset G'$. 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, i.e., C is complete, and if $C \subseteq D$, then D is not complete. An undirected graph can be identified by the set C of its cliques. The set $\overline{E}$ is the set of missing edges of G; that is, for a pair $i, j \in V$, $(i, j) \in \overline{E}$ if and only if $i \neq j$ and $(i, j) \notin E$. A path of length $l > 0$ from v_0 to v_1 is a sequence $v_0, v_1, ..., v_l$ of distinct vertices such that $(v_{k-1}, v_k) \in E$ for all $k = 1, ..., l$. 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.

2.2. *Gaussian graphical model*

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 $V = \{1, ..., p\}$ with probability distribution P_V and let $G = (V,E)$ be an undirected graph. 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 $X_I \perp X_J | X_U$ 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 $X_I \perp X_J | X_U$ whenever U separates I and J in G; in particular this implies that if $(i,j) \in \overline{E}$ then $Xi \perp Xj | X_{V \setminus \{i,j\}}$.

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 μv 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 $Xi \perp Xj | X_Q$ if and only if the partial correlation coefficient

$$\rho_{ij.Q} = \frac{-k_{ij}^A}{\sqrt{k_{ii}^A k_{jj}^A}}$$

is equal to zero, where $A = Q \cup \{i,j\}$ and $K^A = \{k_{ij}^A\}$ 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 $K_V \equiv K = \{k_{ij}\}$ 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)

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

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 $q = p-2$, i.e. $Q = V \setminus \{i,j\}$, 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 $X^{(n)} = (X^1, ..., X^n)$ 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 $C \in C$ where 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 $C \in C$.

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 $(i,j) \in \overline{E}$. 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 $\hat{\rho}_{ij.Q} = -\hat{k}^A_{ij} \Big/ \sqrt{\hat{k}^A_{ij}\hat{k}^A_{ii}}$ and $\hat{K}^A = (S_{AA})^{-1}$ 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).

2.3. *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. Structural learning procedures based on q-order partial correlations aim at identifying the q-partial graph of X_V, that is a graph in which missing edges correspond to zero q-order partial correlations. Here, we give the definition of q-partial graph and refer to Castelo and Roverato (2006) for the theory of q-partial graphs 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 $(i,j) \in \overline{E}^{(q)}$ if and only if there exists a set $U \subseteq V$ with $|U| \leq q$ and $i,j \notin U$ such that $X_i \perp X_j | X_U$ holds in P_V.

3. GAUSSIAN GRAPHICAL MODELS FOR BIOMELOCULAR NETWORKS

Microarray data quantify the abundance of biomolecules, commonly known as expression level, by probing functional elements along the genome which, with-

out 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, $i = 1, ..., p$, 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 tipically 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.

More recently, a number of different 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. 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 seem to be adequate 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.

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 con-

centration matrix $K = \Sigma^{-1}$ whose maximum likelihood estimate is $\hat{K} = S^{-1}$ 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 Buhlmann (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 2.2.

In the next section we propose a novel procedure to learn q-partial graphs from data. 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. 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. THE *qp*-PROCEDURE

In this section we introduce the *qp*-procedure which 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 $\hat{G}^{(q)}$ 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 4.1 we formally introduce the non-rejection rate. In section 4.2 we describe the procedure in more detail by means of two examples and, finally, in section 4.3 we provide istances of the application of the procedure on both simulated and real data.

4.1. *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 $V \setminus \{i,j\}$ 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 1 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

$$E[T_{ij}^{q}] = Pr(T_{ij}^{q} = 1)$$

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

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

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 func-

tion of the test, and for this reason we use a UMPU test; in particular, recall that $\beta_{ij.\mathcal{Q}} \leq (1-\alpha)$.

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

$$Pr(T_{ij}^q = 1) = \sum_{\mathcal{Q} \in \mathcal{Q}_{ij}} Pr(T_{ij}^q = 1 \mid \mathcal{Q}) Pr(\mathcal{Q}) = \frac{1}{m} Pr(T_{ij}^q = 1 \mid \mathcal{Q}) \tag{3}$$

An element $\mathcal{Q}$ of $\mathcal{Q}_{ij}$ can either separate i and j in G or not separate them. We denote by $1_{ij}(\mathcal{Q})$ the indicator function that is 1 if $\mathcal{Q} \in \mathcal{Q}_{ij}$ 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

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

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

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

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

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

and, by (4),

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

so that we obtain the final form

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

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 $G^{(q)} = (V, E^{(q)})$, i.e. $(i, j) \in E^{(q)}$. 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 $Q \in Qij$. Since for every $Q \in Qij$, $\beta_{ij.Q}$ 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 $Q \in Q_{ij}$ 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 Q_{ij}. It follows that a value of $Pr(T_{ij}^q = 1)$ "close" to $1-\alpha$ means either that $(i,j) \in \overline{E}^{(q)}$ 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 $(i,j) \in \overline{E}^{(q)}$. 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 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 $Q \in Q_{ij}$. 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 4.3.

4.2. *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 $\hat{G}^{(q)}$ 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.

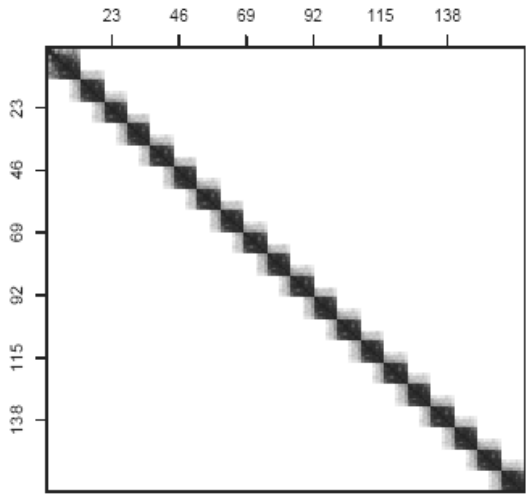

Figure 1 – Image of a partial correlation matrix for 164 variables. Every entry of the matrix is represented as a gray-scaled point between zero (white points) and ±1 (black points).

It is straightforward to check, by using the results of section 2.3, 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 X_j, the required statistical tests are computed for a large number of sets randomly sampled from $\mathcal{Q}_{ij}$ according to a uniform distribution. In the example we are considering, the non-rejection rate is estimated by sampling 500 elements from $\mathcal{Q}_{ij}$, for all of the 13 366 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 12 160.

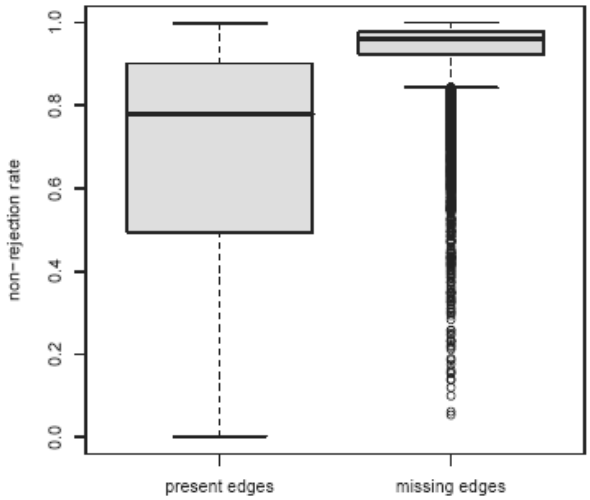

Figure 2 – Boxplots of the estimated values of the non-rejection rate for the 1206 present edges and for the 12160 edges of G= $G^{(20)}$.

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 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.

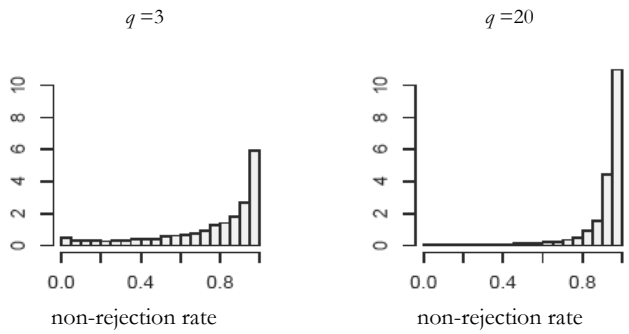

Figure 3 – Histograms of the estimated values of the non-rejection rates.

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 4.1, 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 under analysis. For the case q=20 we can set β^* =0.975 selecting in this way a graph $\hat{G}^{(20)}$ with 9751 out of 13 366 possible edges and whose largest clique has size 32. Figure 5 gives the adjacency matrix of $\hat{G}^{(20)}$ and shows that, although this is clearly an overparameterized model, a substantial dimensionality reduction has been achieved while preserving the block diagonal structure of $\hat{G}^{(20)}$. Indeed, only 34 of the 1206 present edges are wrongly removed corresponding to an error of 2.8%.

4.3. *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.

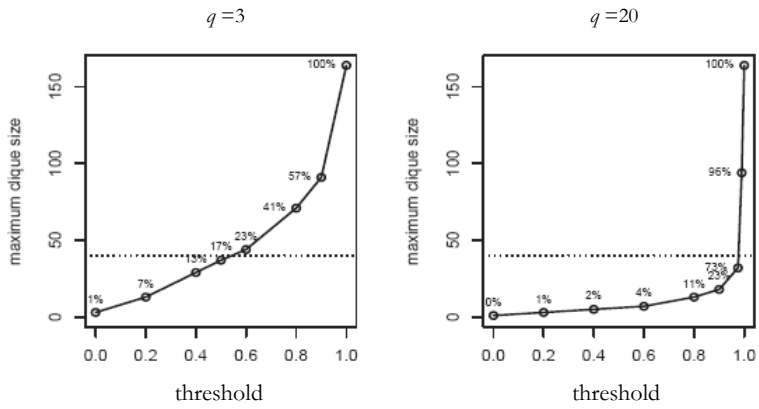

Figure 4 – Plots giving the largest clique sizes of the graphs selected with different threshold values. For every graph the percentage edges is given and the dotted horizontal line is the sample size n.

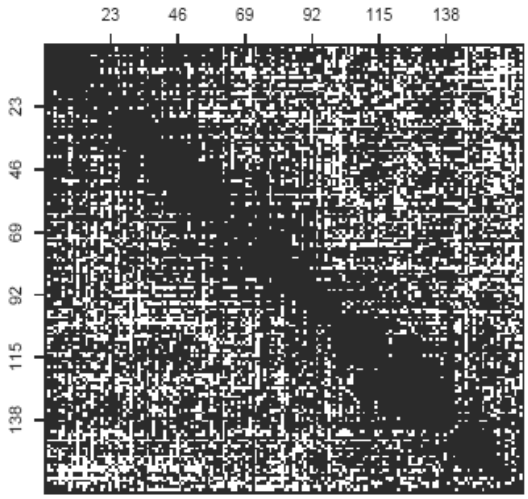

Figure 5 – Adjacency matrix of the graph selected by the *qp*-procedure with q=20 and $\beta^* = 0.975$. Black points are present edges (value 1 in the adjacency matrix) and white points missing edges (value 0 in the adjacency matrix).

In this way, it follows from the results of section 2.3 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 11 175 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 $(n-q)$, i.e. 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 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

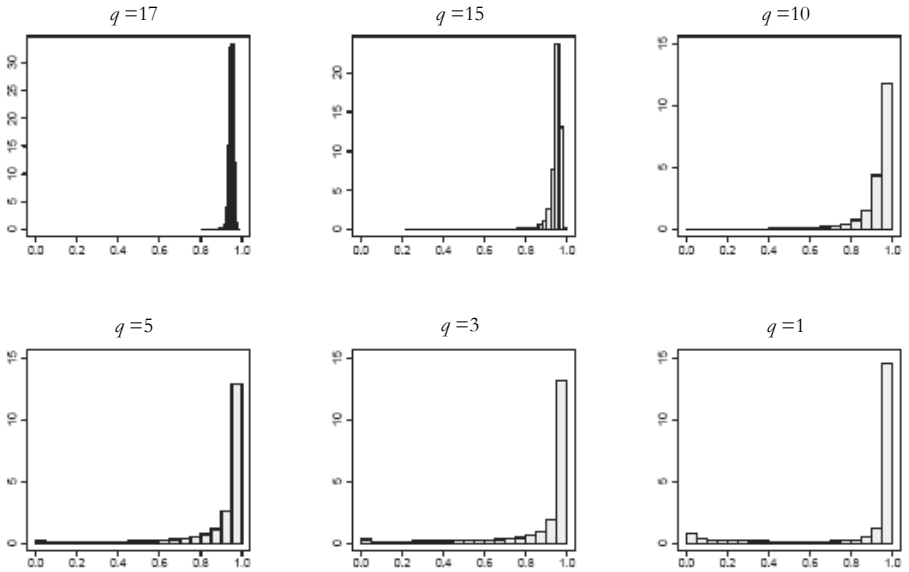

Figure 6 – qp-hist plots for $G_1 = (V,E_1)$ with n=20.

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 4.1 the value of q should not be chosen too small in order to guarantee an adequate sparseness of $G^{(q)}$. 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 $(p-q)$, however we notice that, apart from q=17 where the relative improvement

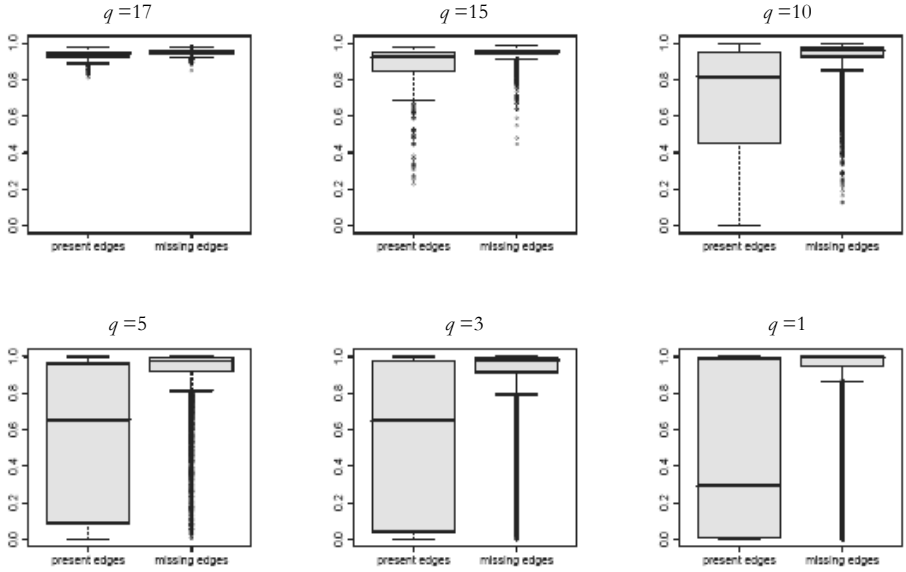

Figure 7 – Distribution of the non-rejection rate for present and missing edges of $G_1 = (V,E_1)$, to be associated with the corresponding histograms in figure 6.

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.

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 satis-

TABLE 1

Graph $G_1 = (V,E1)$. 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	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

fying 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 β^*=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 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 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.

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

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
20	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

4.4. *The qp-package*

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

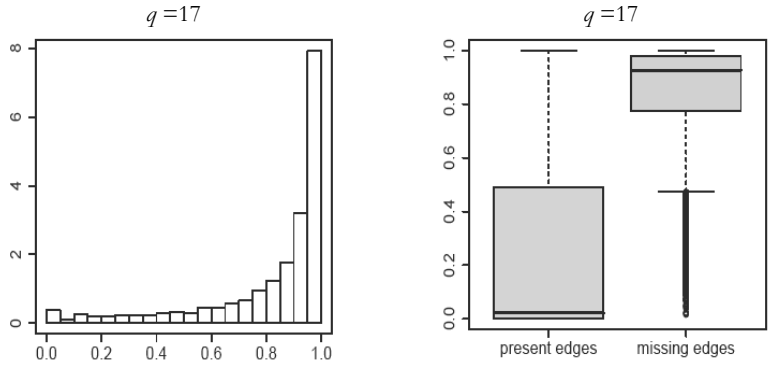

Figure 8 – *qp*-hist plot and associated distributions of the non-rejection rate for present and missing edges of $G_1 = (V,E_1)$, resulting from the application of the *qp*-procedure where n=150 and q =17.

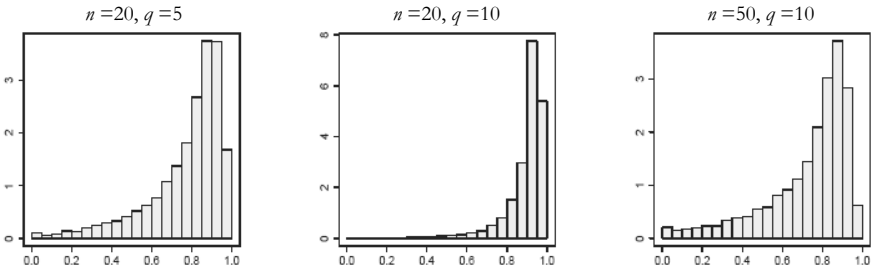

Figure 9 – *qp*-hist plots and associated distributions of the non-rejection rate for present and missing edges of $G_2 = (V,E_2)$, resulting from the application of the *qp*-procedure for different values of n and q.

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.

5. DISCUSSION

The *qp*-procedure is designed to learn q-partial graphs. Its main advantage is that 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 re-

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

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

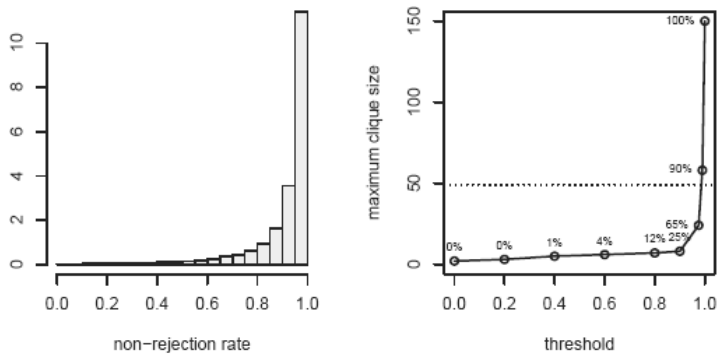

Figure 10 – Estrogen receptor data of West *et al.* (2001): *qp*-hist and *qp*-clique plots for $q = 20$.

sulting 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 XV follows a multivariate normal distribution, and the normality of microarray data is a disputed question.

We refer toWit 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.

Dipartimento di Scienze Statistiche ALBERTO ROVERATO
Università di Bologna

Department de Ciències Experimental i de la Salut ROBERT CASTELO
Universitat Pompeu Fabra

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).

REFERENCES

R. CASTELO, A. ROVERATO (2006), *A robust procedure for gaussian graphical model search from microarray data with p larger than n*, "Journal of Machine Learning Research", 7, pp. 2621-2650.

R.G. COWELL, A.P. DAWID, S.L. LAURITZEN, D.J. SPIEGELHALTER (1999), *Probabilistic networks and expert systems*, Springer-Verlag, New York.

D. R. COX, N. WERMUTH, (1996), *Multivariate dependencies: Models, analysis and interpretation*, Chapman and Hall, London.

A. DE LA FUENTE, N. BING, I. HOESCHELE, P. MENDES (2004), *Discovery of meaningful associations in genomic data using partial correlation coefficients*, "Bioinformatics", 20, pp. 3565-3574.

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

A.P. DEMPSTER (1972), *Covariance selection,* "Biometrics", 28, pp. 157-75.

A. DOBRA, C. HANS, B. JONES, J.R. NEVINS, M. WEST (2004), *Sparse graphical models for exploring gene expression data*, "J. Mult. Anal.", 90, pp. 196-212.

R.L. DYKSTRA (1970), *Establishing the positive definiteness of the sample covariance matrix*, "Ann. Math. Statist.", 41, pp. 2153-2154.

D.E. EDWARDS (2000), *Introduction to graphical modelling*, Springer-Verlag, New York.

N. FRIEDMAN (2004), *Inferring cellular network using probabilistic graphical models*, "Science", 33, pp. 799-805.

B. JONES, A. DOBRA, C. CARVALHO, C. HANS, C. CARTER, M. WEST (2005), *Experiments in stochastic computation for high-dimensional graphical models*, "Statistical Science", 20, pp. 388-400.

B. JONES, M. WEST (2005), *Covariance decomposition in undirected Gaussian graphical models*, "Biometrika", 92, pp. 779-786.

S.L. LAURITZEN (1996), *Graphical models*, Oxford University Press, Oxford.

E.L. LEHMANN (1986), *Testing statistical hypotheses, 2nd edition*, Wiley, New York.

P.M. MAGWENE, J. KIM (2004), *Estimating genomic coexpression networks using firstorder conditional independence Genome Biology*, 5, R100.

R. MILO, S. SHEN-ORR, S. ITZKOVITZ, N. KASHTAN, D. CHKLOVSKII, U. ALON (2002), *Network motifs: simple building blocks of complex networks*, "Science", 298, pp. 824-827.

J. PEARL (1988), *Probabilistic reasoning in intelligent systems*, Morgan Kaufmann, San Mateo.

A. ROVERATO (2002), *Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models*, "Scand. J. Statist.", 29, pp. 391-411.

J. SCHÄFER, K. STRIMMER (2005a), *An empirical Bayes approach to inferring large-scale gene association networks*, "Bioinformatics", 21, pp. 754-764.

J. SCHÄFER, K. STRIMMER (2005b), *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, August 2004. (Publisher: The American Istitute of Physics).

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

M. WEST, C. BLANCHETTE, H. DRESSMAN, E. HUANG, S. ISHIDA, R. SPANG, H. ZUZAN, J.A. OLSON, J.R. MARKS, J.R. NEVINGS (2001), *Predicting the clinical status of human breast cancer by using gene expression profiles*, "Proceedings of the National Academy of Sciences", 98, pp. 11462-11467.

J. WHITTAKER (1990), *Graphical models in applied multivariate statistics*, Wiley, Chichester.

A. WILLE, P. BÜHLMANN (2006), *Low-order conditional independence graphs for inferring genetic networks*, "Statistical Applications in Genbetics and Molecular Biology", 5, article 1.

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, P. BÜHLMANN (2004), *Sparse graphical Gaussian modeling of the isoprenoid gene network*, "Arabidopsis thaliana. Genome Biology", 5:R92.

E. WIT, J. MCCLURE (2004), *Statistics for microarrays. Design, analysis and inference*, Wiley, Chichester.

F. WONG, C.K. CARTER, R. KOHN (2003), *Efficient estimation of covariance selection models*, "Biometrika", 90, pp. 809-830.

R. YANG, J.O. BERGER (1994), *Estimation of a covariance matrix using the reference priors*, "Ann. statist.", 3, pp. 1195-1211.

E. YEGER-LOTEM, S. SATTATH, N. KASHTAN, S. ITZKOVITZ, R. MILO, R.Y. PINTER, U. ALON, H. MARGALIT (2004), *Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction*, "Proc. Natl. Acad. Sci.", 101, 16, pp. 5934-5939.

RIASSUNTO

Apprendimento strutturale di modelli grafici gaussiani sulla base di dati di microarray con p maggiore di n

L'apprendimento di reti di interazioni tra variabili sulla base di dati rilevati su microarray è un tema di grande interesse in bioinformatica. Un approccio che ha riscosso elevata attenzione si basa sull'assunzione che i dati a disposizione rappresentino un campione casuale da una distribuzione normale multivariata che appartiene ad un modello grafico gaussiano. In questo caso il principale oggetto di inferenza è costituito dai *coefficienti di correlazione parziale di ordine completo*, ossia dai coefficienti di correlazione parziale tra due variabili al netto di tutte le variabili rimanenti. Per i dati da microarray, solitamente il numero di variabili rilevate eccede la dimensione campionaria e questo preclude l'applicazione delle procedure tradizionali per l'apprendimento di modelli perchè non è possibile calcolare i coefficienti di correlazione campionaria. In questo articolo si propone una procedura di apprendimento strutturale, denominata *procedura qp*, che utilizza i *coefficienti di correlazione parziale di ordine limitato*. La procedura è implementata in un pacchetto di pubblico dominio per il software statistico R.

SUMMARY

Structural learning of Gaussian graphical models from microarray data with p larger than n

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 exceeds 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 introduce a structure learning procedure, that we call the qp-procedure, based on limited-order partial correlations. The procedure is implemented in a freely available package for the statistical software R.

DISCUSSION

Nanny Wermunt, Elena Stanghellini

First, we want to congratulate Professor Roverato on his new position at the most beautiful and arguably oldest European University of Bologna, for having obtained jointly with Robert Castelo some remarkable theoretical results and for an excellent presentation in today's lecture. Our comments are observations about some of the assumptions, and to properties of partial correlations and their multiples, the linear least-squares regression coefficients; we have some further specific questions.

The assumption of a joint Gaussian distribution is extremely strong and most likely to be unrealistic for most sets of observable variables. In microarray applications, even the marginal distributions of individual genes will rarely be symmetric. Thus, there is a strong need to investigate whether there is evidence of substantial nonlinear or interactive effects in any set of data (see e.g. Cox and Wermuth, 1994) or to investigate whether relations will be at least quasi-linear, see Wermuth and Cox (1998a), so that nonlinear relation have a strong linear component and the vanishing of a partial correlation coefficient, $\rho_{yx,z}$, say, coincides with the conditional expectation of Y on X given Z not depending on X.

In general, if the assumption of a joint Gaussian distribution does not apply, it may not only happen that there is strong nonlinear dependence of Y on given Z if $\rho_{yx,z} = 0$, but more surprisingly, Y may also be conditionally independent of X given Z if $\rho_{yx,z}$ takes on a high value (see Wermuth and Cox, 1998b for an example).

Suppose however that a joint Gaussian distribution is given, then the assumption of a faithful concentration graph needs attention. In this case, almost no constraints are imposed on corresponding data if a linear stepwise data generating process is assumed, for which missing edges in a directed acyclic graph mean zero partial correlations and edges present correspond to a nonvanishing dependence. However, it becomes difficult to judge from a concentration graph whether all independencies present in the joint density are reflected in it.

Let the linear system in a mean-centred vector variable Y be given by

$$AY = \varepsilon, \text{ with } \operatorname{cov}(\varepsilon) = \Delta \text{ diagonal,}$$

where A is a unit-upper-triangular matrix in which all nonzero off-diagonal *ij*-elements are proportional to partial correlation coefficients and it is nonzero if and only if there is an *ij*-arrow present in the directed acyclic graph in which node i corresponds to variable Y_i for i=1,....,d, say. Then, for a missing *ij*-arrow, the corresponding concentration graph has an additional edge *ij*-edge if and only if Y_i and Y_j have a common response Y_h, with (h<i<j). And, an additional asso-

ciation is induced with the partial correlation $\rho_{ij.C}$ for C the set of all other d-2 variables, i.e. a nonzero concentration. Thus, more dependencies will often show in the concentration matrix $\sum^{-1} = A^T \Delta^{-1} A$ than those needed to generate the joint density; more precisely, whenever the generating graph has a 3-node-2-arrow subgraph such that $i \to h \leftarrow j$.

It may also happen that some concentrations vanish due to special parametric constellations.

A simple example is with the following vector of residual covariances, containing the diagonal elements of Δ:

$$(\delta_{11,}\delta_{22,}\delta_{33,}\delta_{44,}) = (1/2, 2/3, 3/4, 1)$$

and the matrices A and $\sum^{-1}$ being

$$A = \begin{pmatrix} 1 & -1/2 & -1/2 & 0 \\ 0 & 1 & -1/3 & 2/3 \\ 0 & 0 & 1 & -1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \sum^{-1} = \begin{pmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & 0 & 1 \\ -1 & 0 & 2 & -1 \\ 0 & 1 & -1 & 2 \end{pmatrix}$$

Here, $\rho_{23.1} = -1/3$, $\rho_{23.4} = 1/3$ and $\rho_{23.14} = \rho_{23} = 0$ so that for pair Y_2, Y_3 negative dependence, positive dependence or independence holds, depending on the conditioning set. The concentration graph is here not faithful. But, the question is how could one check in models for large graphs this assumption that is essential for the theoretical results obtained by the authors?

Related to this issue is the fact that the partial correlation graph G^q defined by the authors can be regarded as a graph with latent variables, where the set of latent variables may change with q. This contrasts with classical latent variable problems in which the set of latent variables remains unchanged. Thus for the partial correlation graphs, there is additional knowledge that may allow to check whether two models are consistent with one another.

Therefore, a question is whether and how this additional knowledge can be or is incorporated in the structure learning procedure using the partial correlation graph.

Another question is, how the authors see the relation to the variable selection method for Gaussian concentration graph models which has been proposed and studied recently (Meinshausen and Bühlmann, 2006) for the same type of situation in which the number of variables is much larger than the number of observation.

For all such methods, one main issue is reproducibility of the conclusions. In general, the direction and strength of observed associations vary the more the smaller the number of observations. For instance in a recent study, Drton and Richardson (2003) study a sample of eight observation from a joint Gaussian distribution. Though they sample, for instance, for three variables Y_2, X_1, and X_2

which satisfy independence of Y_2 from X_1 given X_2, the observed partial correlation coefficient between Y_2 and X_1 given X_2 takes on the value 0.9007. Such strong deviations from population values are quite common if, as in their case, seven parameters are to be estimated from eight observations.

And, more generally, it may be possible to find an association structure in observed data, but this structure can be quite unrelated to the structure in the population whenever the sample size is small compared to the number of parameters that are to be estimated.

Mathematical Statistics, NANNY VERMUNT
Göteborgs Universitet, Sweden

Dipartimento di Economia, Finanza e Statistica ELENA STANGHELLINI
Università di Perugia

REFERENCES

D.R. COX, N. WERMUTH (1994), *Tests of linearity, multivariate normality and adequacy of linear scores*, "Applied Statistics, Journal of the Royal Statistical Society", C, 43, pp. 347-355.

M. DRTON, T. RICHARDSON (2003), *Multimodality of the likelihood in the bivariate seemingly unrelated regression model*, "Biometrika", 91, pp. 383-392.

N. MEINSHAUSEN, P. BÜHLMANN (2006), *High-dimensional graphs and variable selection with the lasso*, "Annals of Statistics", 34, pp. 1436-1462.

N. WERMUTH, D.R. COX (1998a), *Statistical dependence and independence*, in P. Armitage and T. Colton (eds), *Encyclopedia of Biostatistics*, Wiley, New York, pp. 4260-4267.

N. WERMUTH, D.R. COX (1998b), *On association models defined over independence graphs*, "Bernoulli", 4, pp. 477-495.

Henry P. Wynn

The use of projection operators for Gaussian graphical models is useful. One way to derive these is via conditional expectations. Thus if $EX = E(\cdot|X)$ is the conditional expectation operator on a random variable X, so that $E(Y|X)$ is considered as a random variable, there are several equivalent ways to express conditional independence. Thus with three random variables (X, Y, Z), X and Y conditionally independent given Z is equivalent to any of the following:

1. $E_{X,Z} + E_{Y,Z} - E_Z = I$, where I is the identity operator for span(X, Y, Z)
2. $(E_{X,Z} - E_Z)(E_{Y,Z} - E_Z) = 0$
3. $E_{X,Y}\, E_{Y,Z} = E_{Y,Z} E_{X,Z}$

Condition (2) is the statement that the "innovations" are independent, which is familiar, for example, from time series analysis. Condition (3) is an important commutativity condition which lies at the heart of the Gaussian theory. It derives

this importance from to operator theory. Suppose projections P_1 and P_2 are the projections onto the linear subspaces V_1 and V_2. Then P_1 and P_2 commute if and only if $P_1 P_2$ is the projection onto $V_1 \cap V_2$. If all the relevant operators in the subspace lattice commute, which is sometimes called a Boolean lattice, then we can link the lattice to certain kinds of transitive directed graphs in the DAG case (TDAG). This is the theory of Lattice Conditional Independence (LCI) of Anderrson and Perlman (1993) (Annals of Statistics, 21, 1318-1358) and later papers.

It would be useful to make use of this algebra in the case when we are handling sample covariances, and indeed Anderrson and Perlman and others address this issue. Should one, for example, impose the projection conditions (or equivalent) on the sample covariances in the null testing case? In fact by imposing a special zero structure on the influence matrix (inverse covariance) as in the paper the authors may be doing exactly that. In other words one may have a projection or subspaces lattice of the true model which is shadowed by that for the "sample model", in the null case. Also, the conditions imply certain groups structures so that multivariate invariant tests are a natural framework, in a classical testing environment. It would be interesting to link the "partial covariances" and the partial graphs of the present paper to these lattice and group structures. As the various tests give rejection or acceptance the subspace structure will change and one may be able to track this with an appropriate graphic alongside the DAG graphic. The undirected graph case will be similar.

The sparseness discussion in the present paper are very interesting, both theoretically and computationally.

The recent book by Rue and Held (2005) (Gaussian Markov Random Fields, Chapman and Hall/CRC), makes impressive use of sparse matrix methods and may have some useful ideas. To link the lattice structure, the graph structure and the sparseness theoretically, computationally and graphically seems like an exciting research programme and this important paper travels some distance along the road.

London School of Economics HENRY P. WYNN

Angela Grassi, Ernst Wit

We congratulate Alberto Roverato and Robert Castelo for this paper, which has done a valuable service in formalizing the theory of *q*-partial graphs and in providing the so called *qp*-procedure to learn the structure of *q*-partial graphs.

In our comment we concentrate on the *non rejection rate*, a quantity at the base of the proposed *qp*-procedure. Castelo and Roverato exhaustively discuss the influence of the order of partial correlation, *q*, on the estimates of the non rejection rate and discuss its role as a tuning parameter of the learning procedure.

We have been wondering about the dependence of the non rejection rate estimates from α, the probability of first type error.

Studying this dependence has been possible thanks to the easily manageable package, *qp*, for the statistical software R, provided by the authors.

We use the *qp*-package applied to the same simulated data of Roverato and Castelo (available in the *qp*-package itself), in particular to G_1, the graph with p = 150 and 357 edges. Fixing q = 10, and n = 20, we independently apply the *qp*-procedure with three different value of α, 0.01, 0.05, and 0.1.

In figure 1 we represent the histograms with the corresponding distribution of the non-rejection rate estimates. As the value of the significance level α increases, the *qp*-hist plot shows heavier left tails while maintaining a strong negative asymmetric form.

In figure 2 we represent the boxplots associated with the corresponding histograms. The distribution of present and missing edges becomes more and more separated as the significance level α increases.

As we can see in figure 3, in correspondence to a significance level 0.01 it is very difficult to discriminate between present and missing edges even if we decrease the order of partial correlation to q = 1.

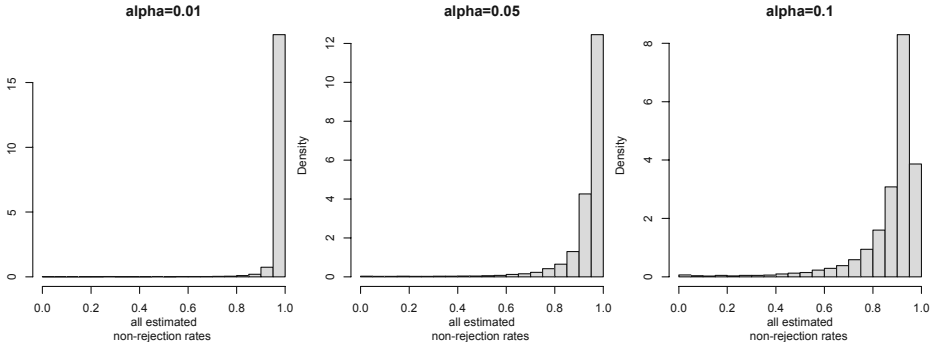

Figure 1 – *qp*-hist plots for G_1 with n = 20, q = 10.

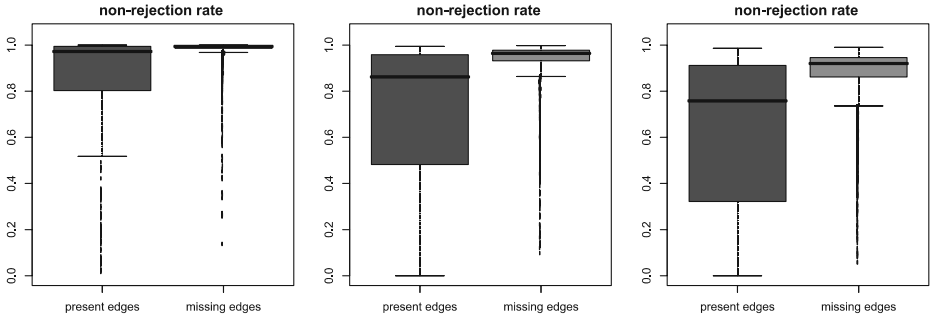

Figure 2 – Boxplots with the distribution of the non-rejection rate for present and missing edges of G_1 to be associated to the corresponding histograms in Figure 1.

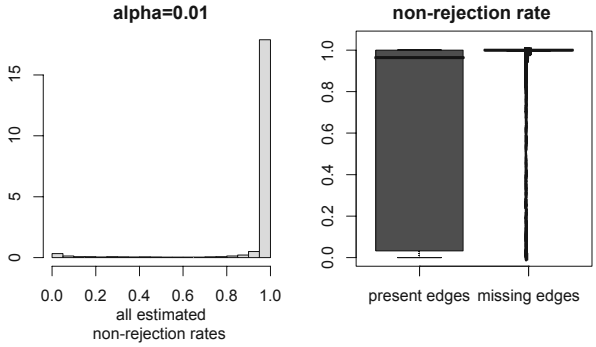

Figure 3 – *qp*-hist plot and associated boxplots for G_1 with $n = 20$, $q = 1$, $\alpha = 0.01$.

What one can clearly see from these plots is that α is an additional tuning parameter which potentially has a large impact on the functioning of the algorithm and so one should carefully think of how to select it.

We would be grateful to the authors if they could give some insight on the figures we presented above and explain us if they have already thought of the possibility of introducing a way to optimally choose α.

Istituto di Ingegneria Biomedica ANGELA GRASSI
CNR - Padova

Department of Mathematics and Statistics ERNST WIT
University of Lancaster

Reply by the Authors

First, we would like to thank all the discussants for their interesting comments and suggestions which have given us the opportunity to consider the material in our paper from new angles. Below we give brief replies to the issues that are raised although many would deserve longer and elaborate answers.

We are grateful to Professor Wynn for highlighting the connection of the theory of *q*-partial graphs with conditional expectation operators, projections, Lattice Conditional Independence models and sparse matrix methods. These are all relevant issues whose importance becomes crucial within the *small n and large p paradigm*. In particular, sparseness is a very general concept that can be specified into different graphical structures. For instance, a network may be sparse because the number of independent paths between every pair of vertices is small, because every vertex has a small number of adjacencies, because it has small cliques and so forth. Sparseness due to a small number of independent paths imply small separators and is therefore useful when conditional independencies relationships have to be identified. On the other hand, a graph with small cliques may have large separators but it

has the advantage that the corresponding model can be fitted also when the sample size is small. We agree with Professor Wynn that it would be interesting to identify model subspaces that induce an appropriate dimensionality reduction so that both hypothesis testing and model fitting can be carried out efficiently also when sampling size is small compared to the number of variables.

The analysis conducted by Professors Grassi and Wit shows the role of the significance level α. The value of α is constant throughout the paper and set equal to 0.05 but, in fact, there is a trade-of between the value of α and the average second type error β. A smaller value of α leads to a larger value of the non-rejection rate for missing edges, but also to a larger value of the average second type error β for present edges. The identification of an optimal value of α may lead to reduction in the selection error and this is always desirable. However, it is also possible to carry out the *qp*-procedure for different values of α and combine the resulting networks. This would lead to an improvement of the robustness properties of our procedure but at the cost of an increased computational complexity.

Professors Wermuth and Stanghellini raise a number of relevant points. Faithfulness is a strong assumption and they illustrate an example of multivariate normal distribution which is not faithful to its undirected independence graph. Although we assume that the underlining distribution is faithful, we try to construct a search algorithm that is robust with respect to such assumption. Specifically, the non-rejection rate for a pair of variables is estimated by considering a large number of different test procedures based on different sets of conditioning variables; as a consequence, such estimate is not affected by an occasional failure of the faithfulness assumption. A second point concerns the precision of partial correlation estimates when the sample size is small compared to the number of variables involved in the analysis. The *qp*-procedure aims at identifying missing edges and the power of statistical tests is improved by considering a small value of q, that is by testing for zero partial correlations in marginal distributions of small dimension. This makes sense only when independence graph is very sparse and, more precisely, when network sparseness is such that marginal distributions of small dimension allow to identify the graph structure. This turns the problem into the investigation of the connection structure of biomolecular networks, which is an open problem of prime interest detailed in our answer to Professor Wynn.

The suggested connection between q-partial graphs and latent variable models is intriguing. In fact, every step of the *qp*-procedure only makes use of a marginal distributions of size $q+2$ and the remaining variables may be regarded as "unobserved". However, in the following steps the observed values of the previously "unobserved" variables are used and, in this way, the additional knowledge is indeed incorporated in the structural learning procedure.

We close by underlining that a proper comparison of our procedure with existing procedure designed to deal with the *small n and large p paradigm* is difficult because the *qp*-procedure is not a traditional structural learning algorithm, but rather a device to restrict the searching procedure to a subset of models of manageable dimension.