*

Assessing the Validity Domains of Graphical Gaussian Models in order to Infer Relationships among Components of Complex Biological Systems

by

Fanny Villers, Brigitte Schaeffer, Caroline Bertin
and Sylvie Huet

Research Report No. 16
July 2008

Statistics for Systems Biology Group
Jouy-en-Josas/Paris/Evry, France
`http://genome.jouy.inra.fr/ssb/`

Assessing the Validity Domains of Graphical Gaussian Models in order to Infer Relationships among Components of Complex Biological Systems

Fanny Villers[1]
Brigitte Schaeffer[2]
Caroline Bertin[3]
Sylvie Huet[4]

Unité Mathématiques et Informatique Appliquées
INRA
Domaine de Vilvert
F-78352 Jouy-en-Josas Cedex

[4] sylvie.huet@jouy.inra.fr

Abstract. The study of the interactions of cellular components is an essential base step to understand the structure and dynamics of biological networks. So, various methods were recently developed in this purpose. While most of them combine different types of data and ¡em¿a priori¡/em¿ knowledge, methods based on Graphical Gaussian Models are capable of learning the network directly from raw data. They consider the full-order partial correlations which are partial correlations between two variables given the remaining ones, for modelling direct links between variables. Statistical methods were developed for estimating these links when the number of observations is larger than the number of variables. However, the rapid advance of new technologies that allow to simultaneous measure genome expression, led to large-scale datasets where the number of variables is far larger than the number of observations. To get round this dimensionality problem, different strategies and new statistical methods were proposed. In this study we focused on statistical methods recently published. All are based on the fact that the number of direct relationship between two variables is very small in regards to the number of possible relationships, ¡em¿p(p-1)/2¡/em¿. In the biological context, this assumption is not always satisfied over the whole graph. So it is essential to precisely know the behaviour of the methods in regards to the characteristics of the studied object before applying them. For this purpose, we evaluated the validity domain of each method from wide-ranging simulated datasets. We then illustrated our results using recently published biological data.

Subjclass. 62H12

Keywords. Graphical Gaussian Model, Estimation, Simulation

1. Introduction

Biological systems involve complex cellular processes built up from physical and functional interactions between molecular entities (genes, proteins, small molecules,...). Thus, to understand how these processes are regulated, it is necessary to study the behavior of the molecular machinery. Recently, biotechnological developments were focused on the characterization and the quantification of cellular system components leading to produce a huge amount of various data. So, one of major challenges is nowadays to understand from these data, how molecular entities interact i.e. what the functional links are, in the context of a whole system. To this end, several mathematical and computational approaches are developing. Some methods based on correlations or clustering can reveal proximities between variables but do not bring to light the direct or functional links. Other methods, such as kernel-based methods (Okamoto et al., 2007; Yellaboina et al., 2007) imply a learning phase and so need a training data set. Bayesian approaches are also used to infer relations between biological entities in order to understand the regulatory mechanisms of living cells (Husmeier, 2003; Werhli & Husmeier, 2007). However these methods have to deal with the prior probability that has a non-negligible influence on the posterior probability when data are sparse and noisy.

A valuable complement to all of these methods is graphical Gaussian modeling (Kishino & Waddell, 2000; Dobra et al., 2004; Wu & Ye, 2006) that can infer direct relations between variables from a set of repeating observations of these variables without any *a priori* knowledge. Graphical modeling is the use of a graph to represent a model. A graph is a set of nodes and edges which can be represented as a graphic for a visual study or as a matrix for computer processing. Graphical modeling is based on the conditional independence concept. In other words, a direct relation between two variables exists if those two variables are conditionally dependent given all remaining variables. In the Gaussian setting, a direct relation between two variables corresponds to a non-zero entry in the partial correlation matrix. As the partial correlation matrix is related to the inverse of the covariance matrix, a direct relation between two variables also corresponds to a non-zero entry in the inverse of the covariance matrix.

Graphical models are classically used when the number of observations, denoted n, is larger than the number of variables, denoted p. This is generally the case in financial or sociological studies where surveys concern few variables and a lot of observations. But it is not the case in the post-genomic context where each experiment is costly in time and money. So the number of repetitions is limited; moreover, each experiment generates numerous data. Then the data set structure, $p \gg n$, does not match with the assumptions of the classical graphical modeling approach and the empirical covariance matrix cannot be inverted. Over the last years, some mathematical and computational researches were developed for surrounding that dimensionality problem and various methods were proposed. Most of them are based on the fact that the number of direct relations between two variables is very small in regards to the number of possible relations, $p(p-1)/2$.

The purpose of our study is to determine the validity domain of some of these methods recently proposed. The reason for this work is to give biologists hints for using the most appropriate methods. Indeed, biologists are very interested in infering biological networks but they generally have a small number of repetitions, the order of ten.

The core of this document is divided in three parts. The first one describes the statistical methodology involved in Schäfer & Strimmer (2005a), Schäfer & Strimmer (2005b) Wille & Bühlmann (2006), Meinshausen & Bühlmann (2006), Friedman et al. (2007), Kalisch & Bühlmann (2007) and Giraud (2008) approaches. The second part presents simulations carried out with each of these methods, under different conditions of dataset structure. The third part illustrates the interest of the graphical Gaussian modeling with an application to flow cytometry data produced by Sachs et al. (2005). In the conclusions we discuss the performances of each method and we bring some recommendations according to their validity domain.

2. STATISTICAL METHODS

Let $\Gamma = \{1, \ldots, p\}$ be the set of nodes of the graph. The p nodes of the graph are identified with p Gaussian random variables. Let us denote by $\boldsymbol{X} = (X_1, \ldots, X_p)^T$, a p random vector distributed as a multivariate Gaussian $\mathcal{N}(0, \Sigma)$. For m a subset of $\{1, \ldots, p\}$ with cardinality $|m|$, we denote by $\boldsymbol{X}^m$ the $|m|$ random vector whose components are the variables X_c, where $c \in m$. Moreover we denote by Γ^{-m} the set of nodes that are not in m, $\Gamma^{-m} = \Gamma \setminus m$, and by $\boldsymbol{X}^{-m}$ the $p - |m|$ random vector whose components are the variables X_c, where $c \in \Gamma^{-m}$. There exists an edge between nodes a and b if and only if, the random variables X_a and X_b are not independent conditionally to $\boldsymbol{X}^{-\{a,b\}}$. In other words, assuming that the matrix Σ is nonsingular, there exists an edge between nodes a and b if and only if the component (a, b) of the concentration matrix $K = \Sigma^{-1}$ is non zero. These graphs are called concentration graphs or full conditional independence graphs. For each node a, the set of neighbors of a is defined as the set of nodes in $\Gamma^{-\{a\}}$ that are connected with a. Finally let us denote by E, the set of edges of the graph.

The statistical challenge is to detect the edges in the graph on the basis of a n-sample from a multivariate distribution $\mathcal{N}(0, \Sigma)$. For each $i = 1, \ldots, n$ we denote by $\mathcal{X}_i = (X_{i1}, \ldots, X_{ip})$ the i^{th} observation. When the number of observation n is large enough, at least $n \geq p+1$, in order to guarantee that the sample covariance matrix S is nonsingular, several methods have been proposed. A detailed review can be found in a recent paper by Drton & Perlman (2007). However, when the interest lies on genomic networks, we are generally dealing with data where the number of variables p is large and the number of experiments n is small. Several methods have been proposed recently in that context.

Some of these methods aim at estimating the concentration matrix K. For instance, Schäfer & Strimmer (2005b,a) proposed methods based on bagging or shrinkage in order to stabilize either the estimator of P, the correlation matrix associated to Σ, or the estimator of Π the partial correlation matrix. Then they estimate the probability of an edge between two nodes (a, b) by estimating the density of the estimated partial correlation coefficient. More recently some authors (Yuan & Lin, 2007; Banerjee et al., 2008; Huang et al., 2006; Friedman et al., 2007) proposed algorithms to estimate K by maximizing the penalized log-likelihood, the penalty term being proportionnal to the sum of the absolute values of the components of K. The coefficient of proportionality may be chosen such as to control the probability of error in estimating the graphical model.

Other methods are based on the estimation of a graph that is an approximation of the full conditional graph. Wille & Bühlmann (2006) suggested to estimate a lower-order conditional independence graph in place of the full conditional independence graph. They use a multiple testing procedure for detecting edges. Kalisch & Bühlmann (2007) considered the PC-algorithm (Spirtes et al. (2000)) to estimate a graph defined through conditional dependencies on any subset of the variables. The PC-algorithm starts from the complete graph and deletes recursively edges based on conditional independencies.

We finally consider a third kind of methods, based on neighborhood estimation. Meinshausen & Bühlmann (2006) proposed to estimate the neighbors of each node using a model selection procedure based on the LASSO method. The choice of the penalty parameter allows to control the probability of falsy joining distinct connectivity components in the graph. More recently, Giraud (2008) suggested to estimate graphs using a model selection procedure based on a penalized empirical risk. The procedure leads to control the mean square error of prediction and its performances are established in a non-asymptotic setting.

In the next section we briefly describe these methods, specifying their theoretical properties if any.

2.1. **Estimating the concentration matrix.**

2.1.1. *Bagging or shrinkage for improving the covariance estimator.*

Schäfer and Strimmer proposed to use bagging (Schäfer & Strimmer, 2005a) or shrinkage (Schäfer & Strimmer, 2005b) for obtaining accurate and reliable estimates of the covariance matrix Σ or its inverse K.

The bagging approach. Bootstrap aggregation (bagging) is used in order to reduce the variance of the estimator of the correlation matrix P. For each bootstrap sample $\mathcal{X}^*$, the empirical correlation matrix $\widehat{P}^*$ is calculated. The bagged estimator is the empirical mean of the $\widehat{P}^*$'s from the bootstrap samples. The partial correlation matrix Π is estimated from the pseudo inverse of the bagged correlation matrix estimator and is denoted by $\widehat{\Pi}^{\texttt{bagged}}$.

The shrinkage approach. The shrinkage estimator is a linear combination of the empirical covariance matrix S and of a *target* estimator denoted $\widehat{\Omega}$ chosen for its very low variability. Precisely $\widehat{\Sigma}(\lambda) = \lambda\widehat{\Omega} + (1-\lambda)S$ where the parameter λ is chosen such as to minimize the quadratic risk function defined as $R(\lambda) = \mathrm{E}\left\{\sum_a \sum_b (\widehat{\Sigma}_{a,b}(\lambda) - \Sigma_{a,b})^2\right\}$. The parameter λ can be explicetly calculated and is estimated using the data only. Let $\widehat{\lambda}$ be this estimator. The partial correlation matrix Π is estimated by $\widehat{\Pi}^{\texttt{shrinked}}$ from the inverse of the matrix $\widehat{\Sigma}(\widehat{\lambda})$.

Estimating the graph. It remains to define a decision rule for detecting the significant components of Π. Let us denote by $\widehat{\Pi}$ either $\widehat{\Pi}^{\texttt{bagged}}$ or $\widehat{\Pi}^{\texttt{shrinked}}$. Schäfer and Strimmer assume that the distribution of the $\widehat{\Pi}_{a,b}$'s is known up to some parameters that are estimated. They deduce from this estimator the posterior probability of an edge to be present in the graph and decide to keep edges such that the posterior probability is greater than a given threshold $1-\alpha$.

2.1.2. *Penalized maximum likelihood.* Banerjee et al. (2008) considered the problem of estimating the parameters of a Gaussian distribution solving a maximum likelihood problem with an added ℓ_1-norm penalty term. Precisely they proposed to estimate the inverse covariance matrix K by maximizing with respect to Ω in the set of positive definite matrices the following criteria:

$$C(\Omega, \lambda) = \log(\det\Omega) - \mathrm{trace}(S\Omega) - \lambda \sum_a \sum_b |\Omega_{ab}|.$$

Friedman et al. (2007) recently proposed a performant algorithm allowing to estimate K by showing that solving this optimisation problem comes to recursively solving and updating a regression LASSO problem. For a given parameter λ, let us denote by $\widehat{K}(\lambda)$ the estimator of K. The set of pairs (a, b) such that $\widehat{K}_{a,b}(\lambda)$ is non zero constitutes the set of edges in the graph. Banerjee and collaborators proposed a choice of λ for which the probability to connect two distinct connectivity components of the graph is bounded by some α. Precisely

$$\lambda(\alpha) = \frac{\mathcal{T}_{n-2}^{-1}(1-\alpha/2p^2)}{\sqrt{n-2+\mathcal{T}_{n-2}^{-1}(1-\alpha/2p^2)}} \tag{1}$$

where $\mathcal{T}_{n-2}$ is the distribution function of a Student variable with $n-2$ degrees of freedom.

2.2. **Approximation of the concentration graph.**

2.2.1. *The 0-1 conditional independence graph.* Wille and Bühlmann proposed to infer the first-order conditional independence graph instead of the full conditional independence graph. Their method has nice computational properties but the drawback is that 0-1 conditional independence graphs do not generally coincide with concentrations graphs, though the links between both graphs can be established in some cases. The 0-1 conditional independence graph is defined as follows: for each pair of nodes (a, b), let $R_{ab/\emptyset}$ be the correlation between the variables X_a and X_b, and for each $c \in \Gamma^{-\{a,b\}}$, let $R_{ab/c}$ be the correlation between X_a and X_b conditionally to X_c; there exists an edge between nodes (a, b) if $R_{ab/\emptyset} \neq 0$ and $R_{ab/c} \neq 0$ for all $c \in \Gamma^{-\{a,b\}}$, or equivalently if

$$\phi_{a,b} = \min\left\{|R_{ab/c}|, c \in \Gamma^{-\{a,b\}} \cup \emptyset\right\} \tag{2}$$

is non zero. Therefore, detecting edges in the graph remains to testing $p(p-1)/2$ statistical hypotheses: For each (a, b), $1 \leq a < b \leq 1$, there exists an edge between nodes (a, b) if the hypothesis "$\phi_{ab} = 0''$ is rejected. Wille and Bühlmann propose the following testing procedure: For each (a, b) and $c \in \Gamma^{-\{a,b\}} \cup \emptyset$ the likelihood ratio test statistic of the hypothesis "$R_{ab/c} = 0$" is calculated as well as the corresponding p-value denoted $P(a, b/c)$. Then the hypothesis "$\phi_{ab} = 0''$ is rejected at level α if

$$P_{\max}(a, b) = \max\left\{P(a, b/c), c \in \Gamma^{-\{a,b\}} \cup \emptyset\right\} \leq \alpha.$$

It remains to calculate the adjusted p-values to take into account the multiplicity of hypotheses to test, considering for example the Bonferroni procedure or the Benjamini-Hochberg one's.

Considering 0-1 conditional independence in place of full conditional independence has several advantages. The test statistics are very easy to calculate. For each hypothesis to test, "$R_{ab/c} = 0$", one considers the marginal distribution of the 3-random Gaussian vector $(X_a, X_b, X_c)^T$.

Therefore, provided that n is large enough, the distribution of the likelihood ratio test statistic of the hypothesis "$R_{ab/c} = 0$" is well approximated by the distribution of a χ^2 with 1 degree of freedom. Note that it is not necessary to assume that p is small. It follows, that, for each (a, b), the probability to detect an edge between a and b when it does not exist is smaller than α, if n is large (see Proposition 3 in Wille & Bühlmann (2006)). Moreover it can be shown that if p increases with n in such a way that $\log(p)/n$ tends to 0 when n tends to infinity, then the estimators of the $R_{ab/c}$'s are uniformly convergent for all $a, b \in \Gamma$ and $c \in \Gamma^{-\{a,b\}} \cup \emptyset$.

Castelo & Roverato (2006) and Malouche & Sevestre (2007) proposed a similar approach for estimating "up to q"-order conditional independence graphs where the presence/absence of edges is associated to all marginal distributions up to the order q. We will only present the method proposed by Wille and Bühlmann in our simulation study.

2.2.2. *The strong conditional independence graph.* Let us consider graphs defined as follows : there exists an edge between nodes a and b if and only if for all set of nodes $m \subset \Gamma^{-\{a,b\}}$, the random variables X_a and X_b are not independent conditionally to $\boldsymbol{X}^m$. This graph is subset of the full conditional independence graph and will be called strong conditional independence graph.

Such graphs can be estimated using an iterative procedure called the PC-algorithm proved to be computationnally very fast for sparse graphs. The procedure starts with the complete graph and removes edges with zero order conditional independence relations. Then edges with one order conditional independence relations are removed and so on. For each step s, let us denote by E^s the set edges and for each node a, by V_a^s the set of neighbors of a. At step $s+1$, we need only to consider the ordered pairs of nodes $(a, b) \in E^s$, such that the cardinality of V_a^s is strictly greater than s. For each of these pairs (a, b), the procedure consists in keeping an edge between nodes a and b if X_a and X_b are not independent conditionally to $\boldsymbol{X}^m$ for all subsets of nodes m contained in V_a^s with cardinality equal to $s + 1$.

Kalisch & Bühlmann (2007) considered a sample version of the PC-algorithm as follows : the testing procedure for deciding to keep an edge between nodes a and b at step s consists in testing, for each subset of nodes m to be considered, that the correlation between X_a and X_b conditionally to $\boldsymbol{X}^m$ is zero. The test statistic is based on the Fisher's z-transform of the sample partial correlations $\widehat{R}_{ab/m}$. Precisely

$$Z_{a,b/m} = \frac{1}{2}\log\left(\frac{1+\widehat{R}_{ab/m}}{1-\widehat{R}_{ab/m}}\right).$$

and for some $\alpha > 0$, the null hypothesis is rejected if $\sqrt{n-|m|-3}Z_{a,b/m} > \Phi^{-1}(1-\alpha/2)$, where $|m|$ denotes the cardinality of m and Φ the distribution function of a Gaussian centered variable with unit variance. The edge between nodes a and b is removed at step s of the algorithm, if there exists m with cardinality s such that the test is not rejected,

Under some conditions on the distribution of $\boldsymbol{X}$, the estimated graph is a consistent estimate of the strong conditional independence graph. The asymptotic framework considers sparse graphs of high dimension: when n tends to infinity, the maximum number of neighbors tends to infinity slower than n, while the number of nodes p may grow like any power of n and the parameter α has to tend to zero.

For practical issues the choice of the parameter α is an open problem. Kalisch & Bühlmann (2007) discussed this point on the basis of a simulation study for estimating the skeleton of a directed acyclic graph.

2.3. **Estimating the neighbors.**

2.3.1. *LASSO procedure.* Detecting the neighbors of all nodes leads to detecting the edges in the graph. Because of the Gaussian assumption on the distribution of $\boldsymbol{X}$, for each variable X_a, a conditional regression model can be defined as follows:

$$X_a = \sum_{b\in\Gamma^{-\{a\}}} \theta_{a,b}X_b + \varepsilon_a \tag{3}$$

where the parameters $\theta_{a,b}$ are equal to $-K_{a,b}/K_{a,a}$. The variable ε_a is distributed as a centered Gaussian variable and is independent from the X_b's for all $b \in \Gamma^{-\{a\}}$. Meinshausen & Bühlmann (2006) proposed to detect the non zero coefficients of the regression of X_a on the variables X_b for $b \in \Gamma^{-\{a\}}$ on the basis of the n-sample $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$, using the LASSO method as a model selection procedure. Precisely, for a given smoothing parameter λ, the estimators of $\{\theta_{a,b}, b \in \Gamma^{-\{a\}}\}$ minimize the sum of squares penalized by the ℓ_1-norm of the parameters vector:

$$\sum_{i=1}^{n}\left(X_{ia} - \sum_{b\in\Gamma^{-\{a\}}}\theta_{a,b}X_{ib}\right)^2 + \lambda\sum_{b\in\Gamma^{-\{a\}}}|\theta_{a,b}|. \tag{4}$$

The solution to this minimization problem is given by a set of $(\widehat{\theta}_{a,b}, b \in \Gamma^{-\{a\}})$ that are either equal to zero or not. The set of nodes $b \in \Gamma^{-\{a\}}$

such that $\widehat{\theta}_{a,b}$ is non zero constitutes $\widehat{V}_a$, the estimated set of neighbors of the node a. Two estimated graphs may be deduced from all these $\widehat{V}_a$ for $a = 1, \ldots, p$, depending on whether we decide to put an edge between nodes a and b if both $\widehat{\theta}_{a,b}$ are $\widehat{\theta}_{b,a}$ are non zero or if one of these is non zero.

Meinshausen and Bühlmann proved that, under some conditions ensuring that the signal to noise ratio is not too small, the method is consistent, namely the probability for $\widehat{V}_a$ to be exactly equal to V_a tends to one. The asymptotic framework is similar to the one considered by Kalisch & Bühlmann (2007) : sparse graphs of high dimension. The smoothing parameter λ is assumed to decrease to zero at a rate smaller than $n^{-1/2}$.

For the sake of application they propose a choice of λ such that the probability to connect two distinct connectivity components of the graph is bounded by some α. Precisely

$$\lambda = 2\sqrt{\sum_{i=1}^{n} X_{ia}^2 \Phi^{-1}\left(1 - \alpha/2p^2\right)}. \tag{5}$$

This choice is based on the Bonferroni inequality and it is assumed that the variance of the variables X_a for $a = 1, \ldots, p$ are all equal to one.

2.3.2. *Model selection procedure.* Giraud (2008) considered the problem of estimating by a model selection procedure the non zero θ's occuring in the p regression models defined at Equation (3). The procedure starts with the choice of a collection of graphs with p nodes or equivalently to the choice of a collection of sets of edges, denoted $\{E_1, \ldots, E_L\}$, where L is the cardinality of the collection. For the sake of simplicity, we say that a $p \times p$ matrix $\Omega \sim E_\ell$ if $\Omega_{a,a} = 0$ and if $\Omega_{a,b} = 0$ is equivalent to $(a, b) \notin E_\ell$. For each set E_ℓ in the collection, the parameters θ are estimated by minimising the residual sums of squares: $\widehat{\theta}(\ell)$ is the $p \times p$ matrix that minimizes

$$SCR(\Omega) = \sum_{a\in\Gamma}\sum_{i=1}^{n}\left(X_{ia} - \sum_{b\in\Gamma}\Omega_{a,b}X_{ib}\right)^2$$

with respect to Ω such that $\Omega \sim E_\ell$. The choice of the best graph among $\{E_1, \ldots, E_L\}$ is done by selecting the estimator $\widehat{\theta}(\ell)$ that minimizes the following criteria:

$$\mathrm{Crit}(\ell) = \sum_{a\in\Gamma} q(K, \nu_a(\ell)) \sum_{i=1}^{n}\left(X_{ia} - \sum_{b\in\Gamma}\widehat{\theta}_{a,b}(\ell)X_{ib}\right)^2$$

where $\nu_a(\ell)$ is the number of neighbors of node a in the graph associated to E_ℓ, K is a constant greater than 1 and q is a penalty function given in Giraud (2008). We denote by $\widehat{\theta}$ this estimator.

The theoretical properties of the method are given in a non-asymptotic framework with $n < p$. The graph is assumed to be sparse in the following sense: the maximum number of neighbors over all the nodes in the graph, denoted D, must be smaller than a a quantity of the order $n/2(\log p)$. Under this assumption, it is proved that the Mean Square Error of Prediction of the estimator $\mathrm{MSEP}(\widehat{\theta})$ is bounded above, up to a $\log p$ factor, by a quantity closed to the minimum over ℓ of the Mean Square Error of Prediction of $\widehat{\theta}(\ell)$.

In practice a collection of graphs has to be chosen. For example, one can choose the set of all graphs with at most D edges. Obviously such a choice leads to very high computational cost for large values of p.

3. SIMULATIONS

3.1. **Methods of simulation.**

3.1.1. *Simulating a graph.* Graphs were simulated according to two different approaches.

The first approach is based on the Erdös-Rényi model, noted ER model, which assumes that edges are independent and occur with the same probability. Practically, we fix the number of nodes p and the percentages of edges η then we draw the number of edges according to a binomial distribution with parameters $p(p-1)/2$, η. Next we choose uniformly and independently the positions of the edges.

The second approach was proposed by Daudin et al. (2006) to take into account the topological features of biological networks such as connectivity degree or clustering coefficient. Their model called Erdos-Rényi Mixtures for Graphs, noted $ERMG$, supposes that nodes are spread into Q clusters with probabilities $\{p_1, \ldots, p_Q\}$, and that the connection probabilities of each cluster and between clusters are heterogenous. These connection probabilities constitute the connectivity matrix C. The parameters available from Daudin et al. (2006) study, correspond to a graph with 199 nodes. As we wanted to study the influence of p, we adapted those parameters to our simulations. However, we kept the same graph structure by taking a large weakly connected cluster, a small highly connected cluster and the same group connection structure. Thus we used the following parameter values

$$Q = 4, \qquad (p_1, \ldots, p_Q) = \begin{pmatrix} 0.07 & 0.1 & 0.18 & 0.65 \end{pmatrix} \tag{6}$$

$$C = \begin{pmatrix} 0.999 & 10^{-6} & 10^{-6} & 0.005 \\ 10^{-6} & 0.4 & 0.014 & 0.003 \\ 10^{-6} & 0.014 & 0.2065 & 0.011 \\ 0.005 & 0.003 & 0.011 & 0.013 \end{pmatrix}. \tag{7}$$

That leads to a mean percentage of edges η equals to 2.5%.

Whatever the approach, we finally obtain a matrix composed of 0 and 1, the values 1 indicating the edge positions in the corresponding graph. This matrix is denoted the incidence matrix.

3.1.2. *Simulating the data.* From the incidence matrix of a given graph, we simulated n observations as follows: first we generate a partial correlation matrix Π by replacing the values 1 indicating the edge positions in the incidence matrix, by values drawn from the uniform distribution between -1 and 1. Then we compute columm-wise sums of the absolute values and set the corresponding diagonal element equal to this sum plus a small constant. This ensures that the resulting matrix is diagonally dominant and thus positive definite. Next we standardize the matrix so that each diagonal entry equals to 1. Finally, we generate n independent samples from the multivariate normal distribution with mean zero, unit variance, and correlation structure associated to the partial correlation matrix Π.

3.2. **Simulation setup.** We simulated graphs and data for different values of p, η, n and we estimated graphs from these data using different methods. We review the methods and the way we carried them out. Then we present how we assessed their performances.

3.2.1. *Methods.* The methods for which we present simulation are the following:

- the $\widehat{\Pi}^{\texttt{bagged}}$ and $\widehat{\Pi}^{\texttt{shrinked}}$ methods, proposed by Schäfer & Strimmer (2005a,b) with the decision rule based on posterior probabilities. The threshold $1-\alpha$ is fixed at 0.95. Both methods are implemented in R software (GeneTS package, R-2.2.0; GeneNet package R-2.4.1).

- the `glasso` proposed by (Friedman et al., 2007) with $\alpha = 5\%$ in accordance with Banerjee et al. (2008). This method is implemented in R software (glasso package, R-2.4.1).

- the 0-1 conditional independence graph approach, proposed by Wille & Bühlmann (2006), with the decision rule based on the adjusted p-values following the Benjamini-Hochberg procedure taking $\alpha = 5\%$. We implemented the method in R-2.4.1.

- the PC-algorithm, as proposed by Kalisch & Bühlmann (2007) with $\alpha = 5\%$. This method is implemented in R software (pcalg package, R-2.6.1).

- the Lasso approach, with the two variants `and` and `or` proposed by Meinshausen & Bühlmann (2006) and $\alpha = 5\%$. This method is implemented in R software using the lars package R-2.4.1. A part of the algorithm is implemented in R according to the description given in Section 6.

- the model selection approach proposed by Giraud (2008) taking $K = 3$ in the penalty function as suggested by the author to better control the FDR. The method implemented in R-2.4.1 was kindly provided by the author. For saving computational time, the collection of graphs was a subset of the set of all graphs with at most 3 neighbors per node.

In the continuation of this document we will respectively denote these methods as `bagging`, `shrinkage`, `glasso`, `pcAlgo`, `WB`, `MB.and` and `MB.or`, `KGGM`.

3.2.2. *Assessing the performance of methods.* To assess the performance of the investigated methods we compared each simulated graph with the estimated graph by counting true positives TP (correctly identified edges), false positive FP (wrongly detected edges), true negatives TN (correctly identified zero-edges), and false negatives FN (not recognized edges). From those quantities we estimated the power and the false discovery rate FDR, which are defined by:

$$\text{power} = E\left(\frac{\text{TP}}{\text{TP} + \text{FN}}\right)$$

$$\text{FDR} = E\left(\frac{\text{FP}}{\text{TP} + \text{FP}}|(\text{TP} + \text{FP}) > 0\right).$$

The power and FDR values presented in this work, are the means over 2000 simulations (according to our preliminary results which showed that the stability of the FDR estimation was reached with 2000 simulations).

The performance of the methods were evaluated for several combinations of the parameters p, η and n, in regards to the problematic we wanted to investigate. Moreover, the parameter values were chosen in order to both make the computer time reasonable and extrapolate the results to biological fields.

The first problematic we have focused on, is the influence of the sample size. To this aim, we simulated random graphs fixing the number of nodes p equal to 30, η equal to 2.5% and varying the number of observations n in $\{15, 22, 30, 60\}$. Secondly, we investigated the sparsity assumption common to all methods taking η in $\{0\%, 2.5\%, 4\%, 5\%, 10\%\}$. Third, we were interested in the influence of the node number p. So, we increased p and chose n in order to keep the p/n rates similar to the p/n values used in the first considered point. For all of these three studies, graphs were simulated with the ER method.

The forth problematic we investigated concerns the influence of the graph structure. In this goal, we also simulated graphs with the $ERMG$ method fixing p equal to 30 and varying n in $\{15, 22, 30, 60\}$.

Finally, we focused on the method proposed by Wille and Bühlmann to evaluate the consequence of estimating the 0-1 graph instead of the concentration graph. For this purpose we fixed $p = 30$ and varied η from 0.025 to 0.2 and n from 60 to 1200.

3.3. **Results and discussion.**

3.3.1. *Comparing the methods.* As shown in Figure 1 methods behave very differently. Let us first discuss methods presenting high FDR values.

Comments on `shrinkage`*,* `glasso` *and* `pcAlgo` *methods.* The FDR values for these methods are very high for all considered values of n when $p = 30$ and $\eta = 2.5\%$ as it is shown in Figure 1a. The FDR does not vary with n and remains close to 47% and 30% for `glasso` and `pcAlgo` respectively, while it increases with n from 45% with $n = 15$ to 75% with $n = 60$ for `shrinkage`. When $\eta = 0$, the FDR is small for `shrinkage` and `glasso` methods while it equals 1 for `pcAlgo` (at least one rejected edge at each simulation, see Figure 1c). The high FDR values are associated with high power values. When the graph is sparse enough, say η smaller than 5%, the methods are powerfull (Figures 1b and d), particularly `glasso`: the power varies from 97% to 99% when $\eta = 2.5\%$ and n varies from 15 to 60. This result suggests that it may be of interest to look for better choices of the thresholding parameter α. This will be the object of section 3.3.2.

Comments on `bagging`*,* `WB`*,* `MB`*, and* `KGGM` *methods.* For these methods the FDR values never exceed 6% except with the `bagging` method for $n = 15$. The FDR values obtained with `MB.or` remain steady around 5.5% whereas the FDR values obtained with `MB.and` never exceed 1%. `KGGM` behaves similarly as `MB.or`, with slightly smaller FDR and power values. FDR from `bagging` reaches 18% when $n = 15$ then deeply declines below 3%. The power represented in Figure 1b gradually increases with the number of observations n except with the `bagging`

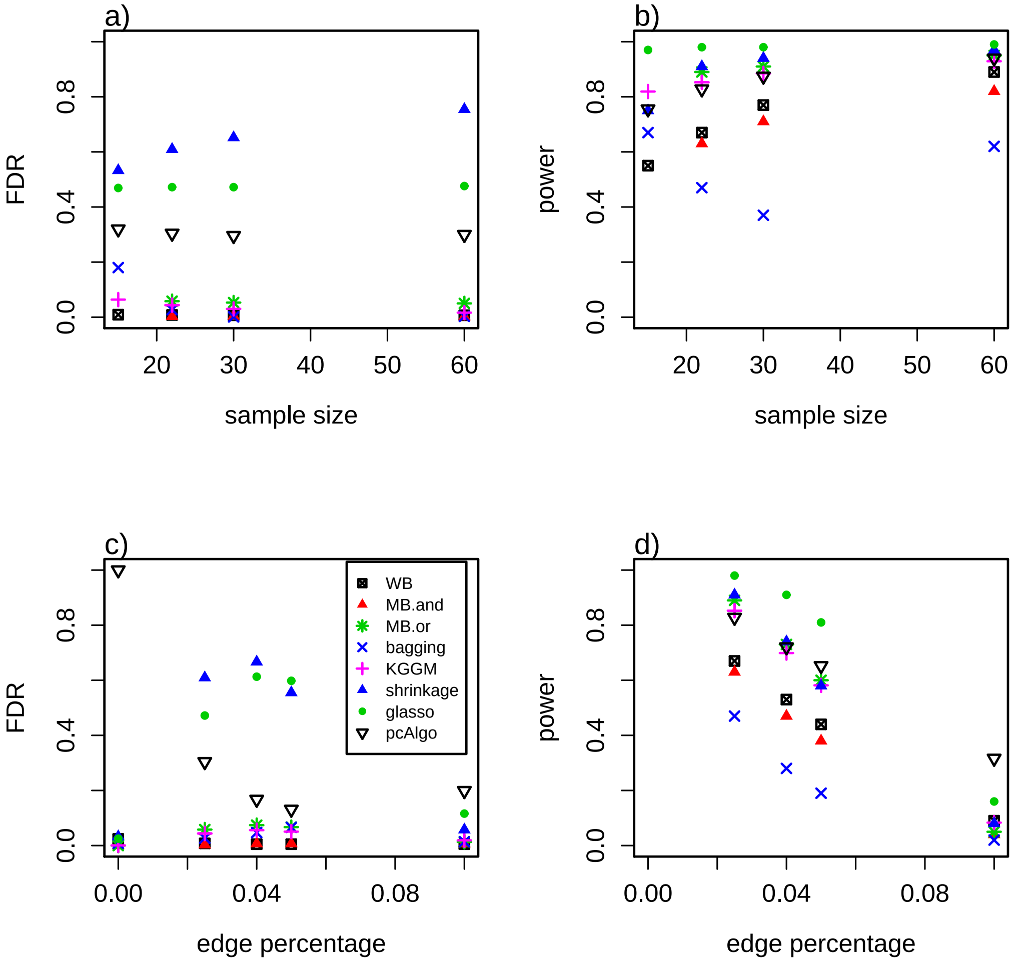

Figure 1. FDR and power of the different methods tested, in function of the sample size (plots a) and b), respectively) and the edge percentage (plots c) and d), respectively). Plots a) and b) were obtained with $\eta = 0.025$. Plots c) and d) were obtained with $n = 22$. All plots correspond to $p = 30$.

method which shows a drop for $n = p$. This phenomena was commented by the authors in Schäfer & Strimmer (2005a). Let us notice that `MB.or` and `MB.and` do not work when $n = 15$. This is due to the fact that when n, α, p satisfy Equation (11), no edge will be detected whatever the data.

The influence of the edge percentages η is shown in Figure 1c and 1d, for $n = 22$. When η in $\{2.5\%, 4\%, 5\%\}$, the FDR values, shown in Figure 1c, stay under 1% with `WB` and `MB.and` methods, around 5% for `KGGM` and exceed 5% with `bagging` and `MB.or` . For all methods the power dramatically falls as η increases and is close to 0 when η equals 10%, whatever the method used. Similar graphics were obtained for $n = 30$ and $n = 60$. When $\eta = 0$, the FDR values lie between 0 for the `MB` methods and 2.4% for `WB`.

Considering the reliability feature (low FDR), the results presented in Figure 1 reveal that the `MB.and` and `WB` methods perform quite well in all cases. Referring to the power, the `MB.or` and `KGGM` methods outperfom the others. The `MB.and` and `WB` methods are less powerfull with the advantage of producing smaller FDR values. The `bagging` appears as the less competitive method in terms of power. All methods similarly show a strong decrease of the power when η increases, in accordance with the sparsity assumption.

3.3.2. *Focus on the high FDRs.* Previously we have seen that the FDR values of the `shrinkage`, `glasso` and `pcAlgo` methods were very high. This behavior may be due to a bad choice of the thresholding parameter α occurring in each of these methods. Hence it may be worthwhile to verify if a more severe thresholding leads to reduce the FDR keeping at the same time a good power. Therefore, we estimated by simulation the power and the FDR for decreasing values of the thresholding parameter using $p = 30$, $\eta = 0.025$ and n varying from 15 to 60.

The curves of the power versus the FDR are shown at Figure 2 for `shrinkage` and `glasso` methods and $n = 22$. For the sake of comparison, we represent the corresponding curve for the `MB.or` method on the same graphic. This graphic shows that we cannot both reduce the FDR and keep a good power with the `shrinkage` and `glasso` methods. When the FDR equals 5%, the power of `MB.or`, `glasso` and `shrinkage` are respectively equal to 0.86, 0.47, 0.05. These values are obtained when α equals 1% for `MB.or` and α equals 10^{-12} for `shrinkage`. For `glasso` FDR values smaller than 0.45 could not be obtained by varying α. Indeed, when α equals 10^{-12}, λ given by Equation (1) equals 0.981, FDR equals 0.46 and power 0.95. Therefore we carried out the `glasso` procedure by varying the values of λ. We got FDR equals 0.05 with $\lambda = 0.9996$.

When η increases ($\eta = 4\%, 5\%, 10\%$) or when p is taken equal to 60, these two methods behave in the same way (results not shown). Therefore they cannot be used for estimating graphs if one wants to control the FDR to a value around 5%. Then we did not keep `shrinkage` and `glasso` methods for further studies.

For the `pcAlgo` algorithm, power decreases with α while the FDR is not a monotone function of α, as shown in Figure 3. Indeed the `pcAlgo` algorithm is a stepwise procedure and at each step only nodes for which the estimated neighborhood is large enough are involved in the next step. If α is too small not enough nodes are kept for the following step. So edges may appear between two nodes even though they are linked through a dropped node. One interesting feature of the variation of FDR versus α is that the FDR is minimum in $\alpha = 0.1\%$ whatever the values of n (Figure 3). Simulations (not shown) with $\eta = 4\%$ lead to the same result: the FDR is a convex function of α and is minimum for $\alpha = 0.1\%$ whatever the values of n. Therefore we tested again the performances of `pcAlgo` method using $\alpha = 0.1\%$ for different values of n and η. FDR decreases from 8.4% to 1.6% when n increases (Figure 4a). It is equal to 47% when $\eta = 0$, remains around 5% when η varies between 2.5% and 5%, and equals 8.8% when $\eta = 10\%$ (Figure 4b). Concerning the power, the `pcAlgo` behaves nearly as the `WB` method.

3.3.3. *Influence of the number of nodes.* In Section 3.3.1 we showed that all methods loose in power when η increases. We now investigate the influence of the number of nodes, p, on that loss of power. The graphic in Figure 5 represents the power in function of η, for different values of p, with $n = p$. Results are shown for `MB.and` procedure, the behavior of the other procedures being similar. In all cases the FDR is smaller than 1%. Figure 5 shows that whatever the value of p, the power decreases when η increases. However, the larger p, the faster the loss of power with η. Consequently, all methods are efficient for sparse graphs, and the edge percentage from which the methods fail depends on the number of nodes.

3.3.4. *Influence of the numbers of neighbors.* In Section 3.3.1 we showed that if the graph is highly connected, the methods are not powerful anymore. In this section we aim at understanding why, and we show in particular the behavior of the methods according to the number of neighbors of the nodes. We focus on the procedure proposed by Meinshausen and Bühlmann and we consider the experiment simulation for $p = 30$, $\eta = 0.025$ and $n = 30$. For each node of the 2000 simulated graphs we count the number of neighbors and the number of correctly detected neighbors. In Table 1 we present for i in $\{1, \ldots, 5\}$, the number n_i of nodes with i neighbors and the percentage $p_{i,j}$ of nodes for

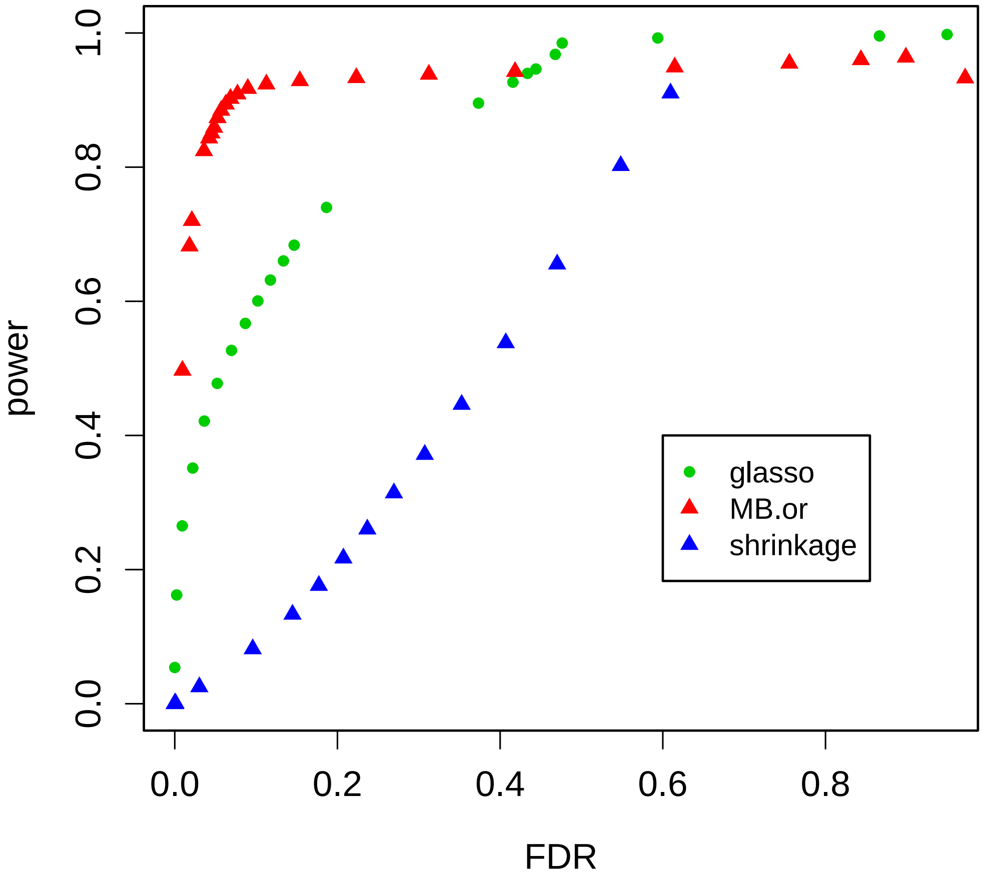

FIGURE 2. Power in function of FDR for the `glasso` and `shrinkage` methods. The curves for `MB.or` method is given as reference. Plots correspond to $p = 30$, $n = 22$ and $\eta = 0.025$.

which the method has correctly detected j neighbors exactly, for j in $\{0, \dots, i\}$.

The percentage $(p_{ii})_{i=1,\dots,5}$ of nodes for which the whole set of neighbors is correctly detected decreases when the number of neighbors i increases. In other words when a node has several neighbors, it often happens that at least one neighbor is not detected. This may explain

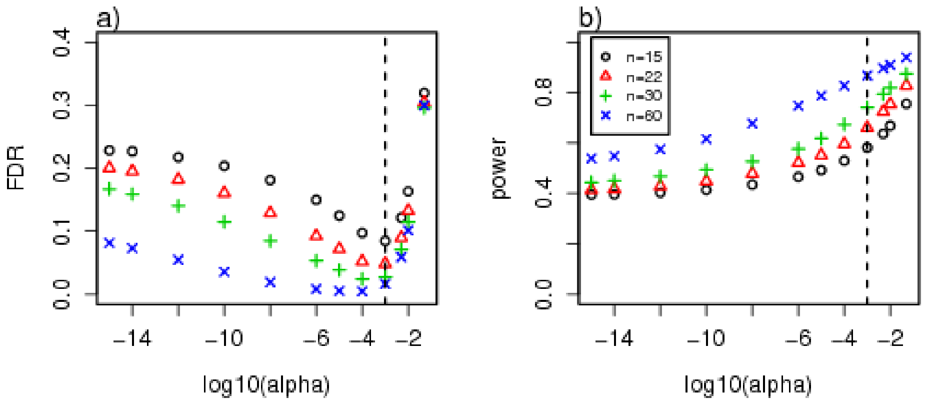

FIGURE 3. Influence of α on FDR and power (plots a) and b), respectively) for different values of n. The level $\alpha = 0.001$ is indicated by the dashed line. Plots correspond to $p = 30$ and $\eta = 0.025$.

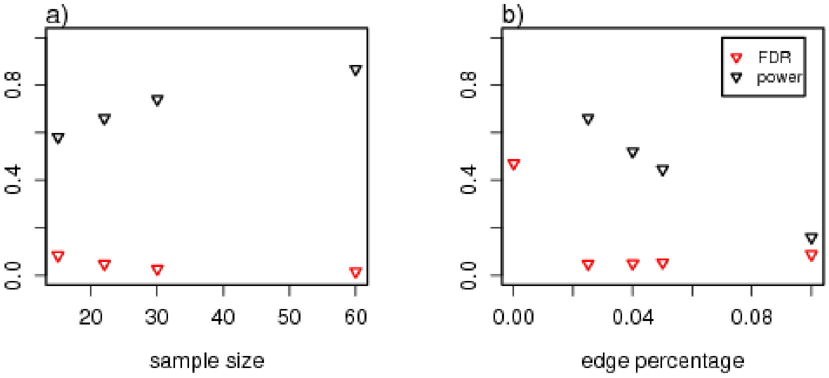

FIGURE 4. Performance of `pcAlgo` for $\alpha = 0.001$; a) FDR and power in function of sample size for $\eta = 0.025$; b) FDR and power in function of the edge percentage for $n = 22$. Plots correspond to $p = 30$.

the loss of power previously observed (see Section 3.3.1) when η increases, because the average number of neighbors increases with η.

Let us now compare the results obtained with `MB.and` and `MB.or` procedures. In Section 3.3.1 we showed that `MB.or` procedure is more

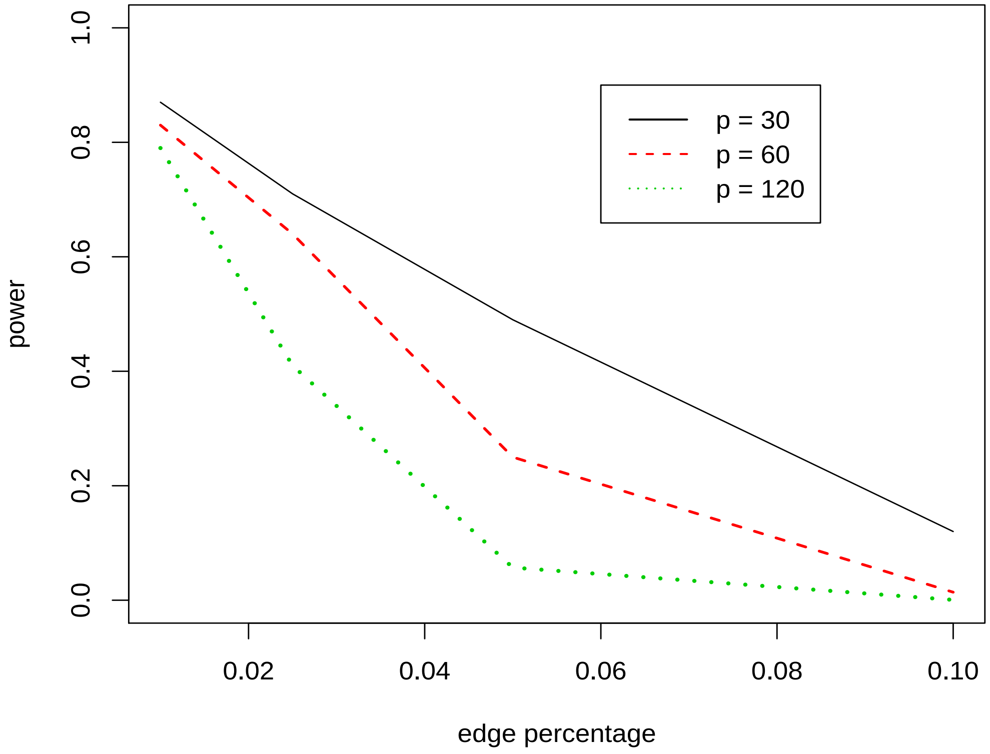

FIGURE 5. Power according to the edge percentage, for different values of p with $n = p$ and for Meinshausen and Bühlmann method using its *and* variant.

powerful and we recover in Table 1 that the percentages of nodes for which the whole set of neighbors is correctly detected are significantly larger with `MB.or` procedure than with `MB.and` procedure. Let us consider for example, as illustrated in Figure 6, a node a with two neighbors b and c such that a is the only neighbor of b and of c. As it has been noticed just before, the procedure of Meinshausen and Bühlmann will detect more easily that a is the neighbor of b and c than the nodes b and c are both neighbors of a. This is the reason why `MB.or` procedure is more powerful than `MB.and` procedure.

i	1	2	3	4	5
n_i	21398	7684	1788	287	38

percentage $p_{i,j}$ obtained with `MB.and`

j \ i	1	2	3	4	5
0	0.182	0.078	0.062	0.066	0.105
1	0.818	0.643	0.465	0.380	0.263
2		0.279	0.414	0.411	0.500
3			0.059	0.139	0.132
4				0.004	0.000

percentage $p_{i,j}$ obtained with `MB.or`

j \ i	1	2	3	4	5
0	0.022	0.016	0.030	0.049	0.079
1	0.978	0.267	0.097	0.063	0.105
2		0.717	0.370	0.195	0.079
3			0.503	0.387	0.132
4				0.306	0.447
5					0.158

Table 1. Number of nodes with i neighbors and percentages of nodes for which exactly j neighbors have been correctly detected by both methods of Meinshausen and Bühlmann with $n = 30$. Graphs are simulated according to the ER model with $p = 30$, $\eta = 0.025$.

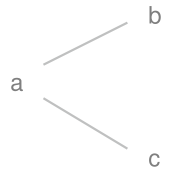

Figure 6. Node a with two neighbors b and c such that a is the only neighbor of b and of c.

3.3.5. *Influence of the graph structure.* In this section we present results when graphs are simulated according to the ERMG model described in Section 3.1.1. Our aim is to evaluate the influence of heteregeneous clusters in the graph. Results are shown in Figure 7 for $p = 30$

and for n taking the values 15, 22, 30 and 60 and for all methods chosen for their low FDR. The parameter α for the `pcAlgo` method was taken equal to 0.1% in accordance with results given at Section 3.3.2. For the parameters given in Equations (6) and (7) the percentage η of edges equals 2.5% which makes the results comparable with those of Figure 1a, 1b and 4.

Using the ERMG model for simulating graphs does not change the shapes of FDR and power curves. As in Figure 1a the FDR value obtained with `bagging` is high when $n = 15$ then deeply declines, and the power drops for $n = p$. Moreover we recover that the FDR values stay very low with `WB` and `MB.and` procedure, stay under 5% for `KGGM` and are larger with `MB.or` and `pcAlgo` procedures. Referring to the power, as in Figure 1b the `MB.or` and `KGGM` procedures outperform the others.

The main difference when graphs are simulated according to the ERMG models is that the power remains under 0.8 even for large n (Figure 7b) whereas it achieves 0.95 when ER model is used (Figure 1b). So, the methods are less powerful when graphs are simulated according to the ERMG model than according to the ER model. The next section shed light on this loss of power.

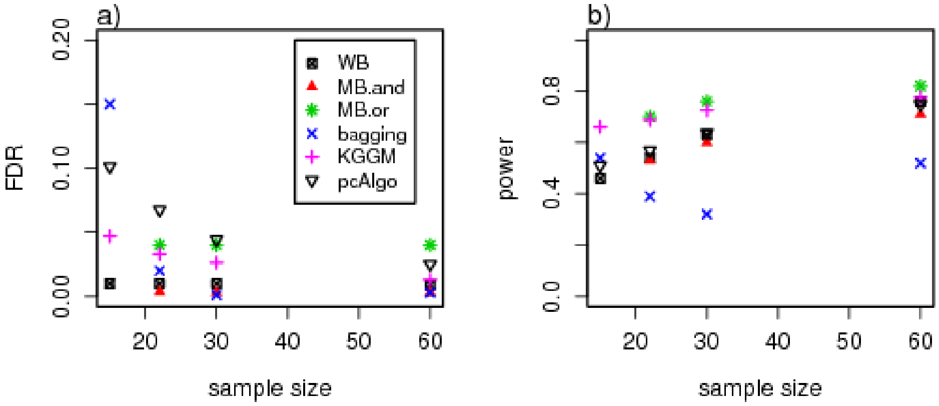

FIGURE 7. FDR (a)) and power (b)) obtained with the different methods tested, in function of the sample size. Graphs were simulated according to the ERMG model with $p = 30$.

3.3.6. *Influence of the neighborhood structure.* In this section we study why the methods are less powerful when graphs are simulated according to the ERMG model than according to the ER model and we underline in particular, the influence of the neighborhood structure.

We consider the same experiment study as in Section 3.3.4 except that the graphs are simulated according to the ERMG model. In Table 2, one can read for each i in $\{1, \dots, 6\}$, the number n_i of nodes with i neighbors and the percentage $p_{i,j}$ of nodes for which the method has correctly detected j neighbors, for j in $\{0, \dots, i\}$. Results are obtained with the procedure `MB.or`.

i	1	2	3	4	5	6
n_i	16326	6720	2603	941	332	63

percentage $p_{i,j}$ obtained with `MB.or`

i j	1	2	3	4	5	6
0	0.042	0.129	0.289	0.530	0.654	0.841
1	0.958	0.305	0.206	0.148	0.133	0.048
2		0.566	0.256	0.128	0.075	0.048
3			0.249	0.121	0.057	0.032
4				0.073	0.054	0.000
5					0.027	0.032
6						0.000

Table 2. Number of nodes with i neighbors and percentages of nodes for which j neighbors have been correctly detected by the `MB.or` procedure with $n = 30$. Graphs are simulated according to the ERMG model with $p = 30$.

Comparing Tables 1 and 2 shows that the number of nodes with 1 and 2 neighbors is smaller for the ERMG model than for the ER model, while the number of nodes with more than 3 neighbors is greater. It appears also that the percentages p_{1j} are similar in both tables for nodes with one neighbor. But when the number of neighbors i is larger than one, the percentage of nodes p_{ii} for which the whole set of neighbors is correctly detected are smaller in Table 2 than in Table 1. Moreover the main difference between Table 1 and 2 concerns the percentage of nodes for which no neighbors is detected: these percentages p_{i0} are very small in Table 1, but large in Table 2 and increases with the number of neighbours. In other words, for graphs simulated according to the EMRG model, detecting no neighbor often happens, especially for nodes with a large number of neighbors. This can be explained by the structure of the neighbors, which is more complex for graphs simulated according to the ERMG model. This point is illustrated above.

In the following we present the FDR and the power estimated into each cluster and between the clusters. We first simulate a graph $\mathcal{G}$ according to the ERMG model with the parameters defined in Section 3.1.1 in order to fix the number of nodes and the localisation of edges in each cluster and between clusters. We simulate a graph with $p = 120$ nodes to ensure that each cluster contains a minimal number of nodes. We denote by $(n_1, \dots, n_Q)$ the number of nodes in each cluster and by N_{edges} the matrix which specifies the number of edges within each cluster and between two clusters. For the simulated graph $\mathcal{G}$, these parameters are:

$$(n_1, \dots, n_Q) = \begin{pmatrix} 7 & 11 & 23 & 79 \end{pmatrix}$$

and

$$N_{\text{edges}} = \begin{pmatrix} 21 & 0 & 0 & 3 \\ 0 & 21 & 6 & 3 \\ 0 & 6 & 38 & 19 \\ 3 & 3 & 19 & 35 \end{pmatrix}.$$

We simulate 2000 data matrix as described in Section 3.1.2 from this graph $\mathcal{G}$ and we estimate the FDR and the power for detecting edges within and between clusters. The results obtained with the `MB.or` procedure and with $n = p$, are presented in the matrices FDR and *power* given at Equations (8) and (9). The component $(a, b)_{a \neq b}$ of the matrix FDR (respectively *power*) gives the estimated false discovery rate (respectively power) of edges between clusters a and b. When there is no edge between two clusters, estimating the power does not make any sense and we put Na. The elements on the diagonal give the estimated false discovery rate (respectively power) of edges within each cluster.

$$\text{FDR} = \begin{pmatrix} 0.000 & 0.001 & 0.016 & 0.008 \\ 0.001 & 0.005 & 0.004 & 0.012 \\ 0.016 & 0.004 & 0.006 & 0.014 \\ 0.008 & 0.012 & 0.014 & 0.021 \end{pmatrix} \tag{8}$$

$$\text{power} = \begin{pmatrix} 0.10 & \text{Na} & \text{Na} & 0.46 \\ \text{Na} & 0.26 & 0.22 & 0.61 \\ \text{Na} & 0.22 & 0.29 & 0.61 \\ 0.46 & 0.61 & 0.61 & 0.87 \end{pmatrix} \tag{9}$$

We can notice from Equation 8 that all estimated FDR values are small. Moreover, the estimated powers vary a lot according to the clusters. Indeed, in the first cluster which contains 21 edges among the $n_1(n_1 - 1)/2 = 21$ possible edges, the power is very small whereas in the fourth cluster which contains 35 edges among the $n_4(n_4 - 1)/2 = 3081$ possible edges, the power is large. The neighbors of the neighbors also influence the power. This can be observed by comparing the power for detecting edges between the second and third clusters, $\text{power}[2, 3] = 0.22$, with the power for detecting edges in the fourth cluster, $\text{power}[4, 4] = 0.87$. So, it appears that it is more difficult to detect edges between clusters 2 and 3 than within cluster 4, while in both cases the percentage of edges to detect is approximately equal to 0.01. This comes from the fact that clusters 2 and 3 are both highly connected. Therefore these two clusters involves nodes for which the structure of the neighbors is complex.

Because of these highly connected parts, the power estimated over the whole graph $\mathcal{G}$ is smaller than if the edges were distributed uniformly in the graph. Indeed, the FDR and the power estimated for the whole graph $\mathcal{G}$ equal respectively 0.016 and 0.44. For graphs simulated according to the ER model with $p = 120$ and $\eta = 0.025$, the average FDR and power estimated over 2000 simulations with the `MB.or` procedure and $n = 120$ equal respectively 0.009 and 0.50.

3.3.7. *Inferring a concentration graph using a 0-1 conditional independence graph.* If the gaussian distribution is *faithfull* for the concentration graph $\mathcal{G}$ (see Proposition 1 in Wille & Bühlmann (2006)), then all edges in $\mathcal{G}$ are edges in the 0-1 conditional independence graph denoted $\mathcal{G}_{\{0,1\}}$. A comparison between $\mathcal{G}$ and $\mathcal{G}_{\{0,1\}}$ is given at Table 3. For each concentration matrix whose values are simulated as described in Section 3.1.2, and for each pair (a, b), $1 \leq a < b \leq 1$, we calculated $\phi_{a,b}$ defined at Equation (2). It appears that, as it was already noticed by Wille and Bühlmann, the number of edges in $\mathcal{G}_{\{0,1\}}$ may be considerably larger than in $\mathcal{G}$. The power and FDR for estimating the graph $\mathcal{G}$ are reported on Figure 8, a) and b). It shows that the FDR increases with n and reaches its maximum for $\eta = 10\%$. This behaviour can be

	$\mathcal{G} \cap \mathcal{G}_{\{0,1\}}$			$\mathcal{G}_{\{0,1\}} \setminus \mathcal{G}$		
η	Number	mean	range	Number	mean	range
0.025	11	0.72	$[10^{-4}, 0.99]$	0.3	0.09	$[10^{-4}, 0.33]$
0.05	22	0.57	$[10^{-5}, 0.99]$	17	0.05	$[10^{-6}, 0.46]$
0.1	43	0.35	$[10^{-7}, 0.99]$	217	0.02	$[10^{-9}, 0.41]$
0.15	65	0.24	$[10^{-9}, 0.99]$	322	0.012	$[10^{-9}, 0.30]$
0.2	87	0.18	$[10^{-8}, 0.99]$	337	0.01	$[10^{-9}, 0.21]$
0.3	131	0.11		304	0.009	

Table 3. Comparison of $\mathcal{G}_{\{0,1\}}$ and $\mathcal{G}$ for $p = 30$ and several values of η. The column $\mathcal{G} \cap \mathcal{G}_{\{0,1\}}$ gives the mean (over 2000 simulations) number of edges that are both in $\mathcal{G}$ and $\mathcal{G}_{\{0,1\}}$, followed by the mean and range of the $\phi_{a,b}$'s corresponding to these edges. The column $\mathcal{G}_{\{0,1\}} \setminus \mathcal{G}$ gives the sames results for edges that are in $\mathcal{G}_{\{0,1\}}$ and not in $\mathcal{G}$. In all simulations the edges of $\mathcal{G}$ are edges of $\mathcal{G}_{\{0,1\}}$.

easily explained by looking at Figure 8, c) and d) where the FDR and the power for estimating $\mathcal{G}_{\{0,1\}}$ are reported. It shows that the FDR for estimating $\mathcal{G}_{\{0,1\}}$ stays very small and that the power increases with n, as expected. Unfortunatly the edges detected in $\mathcal{G}_{\{0,1\}}$ are not in $\mathcal{G}$, leading to increase the FDR for detecting edges in $\mathcal{G}$.

When η is small, say $\eta \leq 2.5\%$, the number of edges that are in $\mathcal{G}_{\{0,1\}}$ but not in $\mathcal{G}$ is very small, and then the FDR for detecting edges in $\mathcal{G}$ is not changed. But when η is large the FDR becomes very large, up to 20% for $\eta = 10\%$. Nevertheless when η is larger, the FDR decreases. This can be explained by the values of the $\phi_{a,b}$'s that are smaller when η increases as it is shown in table 3. Obviously the behaviour of the procedure proposed by Wille and Bühlmann shown in this simulation study, may depend on the way we simulate the concentration matrix. Nevertheless, we have to keep in mind that if the graph is highly connected, or if a part of it is highly connected, then, inferring a concentration graph on the basis of its approximation by a 0-1 conditional independence graph, may lead to detect edges wrongly.

4. APPLICATION TO BIOLOGICAL DATA

In this section, we apply the different methods to the multivariate flow cytometry data produced by Sachs et al. (2005). These data concern a human T cell signaling pathway whose deregulation may lead to carcenogenesis. Therefore, this pathway was extensively studied in the

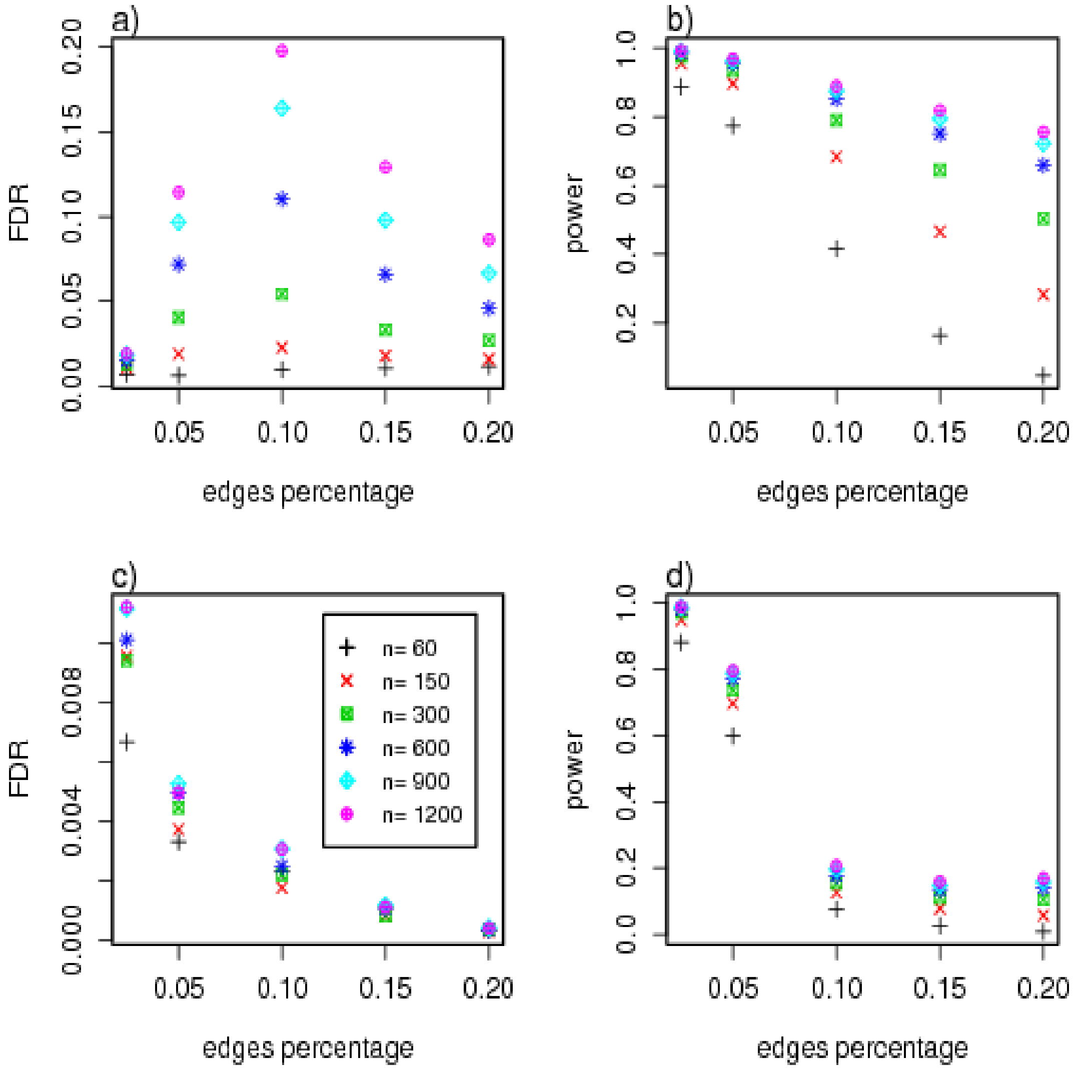

FIGURE 8. FDR and power for estimating $\mathcal{G}$ (plots a) and b)), and $\mathcal{G}_{\{0,1\}}$ (plots c) and d)), in function of η for $p = 30$ and different values of n.

literature and a network involving 11 proteins and 18 interactions was conventionally accepted (Sachs et al., 2005). This network we denoted $\mathcal{G}_{raf}$ is represented in Figure 9. Sachs *et al.* 's data consist of amounts of these 11 proteins, simultaneously measured from single cells under several disturbed conditions. In the sequel, we focus on one general disturbance (+ ICAM-2) that overall stimulates the cellular signaling network. In this condition the quantities of the 11 proteins were measured in 902 cells. Let denote D this data set constituted of $p = 11$

variables and $n = 902$ observations. A log-transformation of the data was made to fit the gaussian assumption better, and the vector of the n-observations for each protein were centered and normalized.

Contrary to most of postgenomic data, flow cytometry data provide a large sample of observations that allow us to measure the influence of the sample size on the power of the estimation methods. From this data set we first compare the networks inferred using the five methods retained for their low FDR. As such abundance of data is rarely available in postgenomic data, we secondly carry out a study to determine the influence of the observation number on the methods.

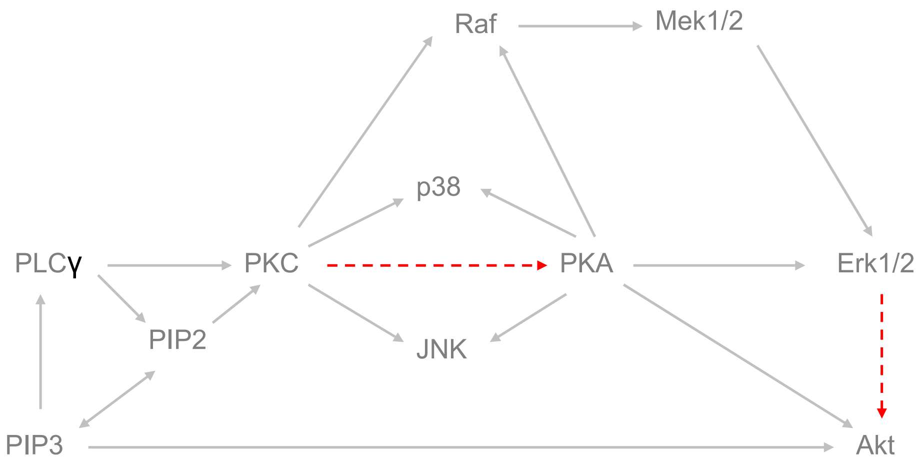

FIGURE 9. $\mathcal{G}_{raf}$. Classic signaling network of the human T cell pathway. The connections well-established in the literature are in grey and the connections cited at least once in the literature are represented by red dashed lines.

We represent the estimated graphs in Figure 10. The graphs inferred with the `bagging`, `WB` and `pcAlgo` methods are identical. This graph involving 10 edges is denoted $\mathcal{G}_1$. The `KGGM` method and the two variants `MB.or` and `MB.and` infer the same graph denoted $\mathcal{G}_2$. This graph involves 9 edges and is identical to $\mathcal{G}_1$ except for the edge between PKA and $Erk1/2$ which is missing. To assess the quality of the methods, we refer to the conventionally acccepted network shown in Figure 9. This network involves 18 connections among which 16 connections are well-established. As the data set D is obtained by considering only one disturbed condition we do not expect the methods to detect all the connections established in the literature. In fact, 10 connections are detected by three of the five methods. Among those connections,

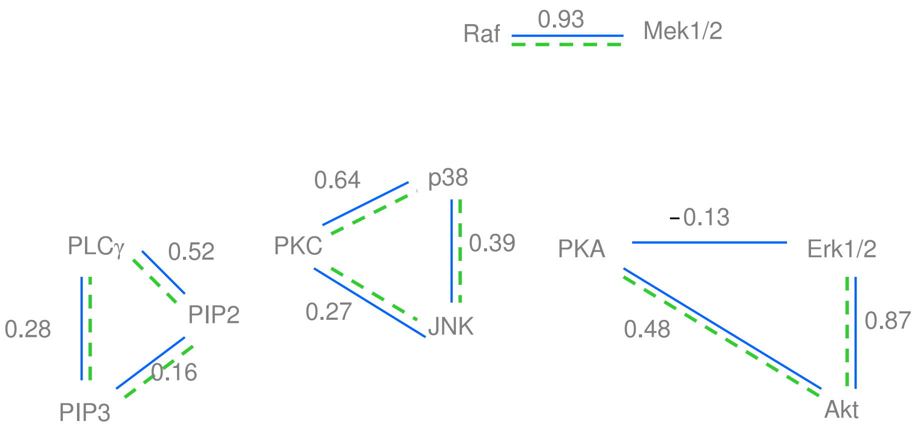

FIGURE 10. Inferred graphs. The graph $\mathcal{G}_1$ estimated with the `bagging`, `WB` and `pcAlgo` methods is represented in blue. The graph $\mathcal{G}_2$ estimated with the `KGGM`, `MB.or` and `MB.and` methods is in green dashed line. The values of the partial correlation matrix associated to the data set D are reported along each edge.

nine of them were well-established or cited at least once in the literature. The tenth one, between $p38$ and JNK, was detected by the five methods previously cited. Moreover the same ten connections were detected by Sachs et al. (2005) (Supplementary Material) applying a bayesian network analysis. Therefore, in the following, we assume that the graph $\mathcal{G}_1$ represents the conditional independence structure of the data set D.

We now investigate the influence of the observation number n on the power of the methods for estimating the graph $\mathcal{G}_1$. We choose n equal to 15, 30, 100, 200, and 300. For each value of n, 2000 $n-$samples are drawn from D without replacement. From each sample, we estimate graphs using the five methods and we compare each estimated graph with the graph $\mathcal{G}_1$. We compute the proportion of wrongly detected edges among the detected edges and the proportion of correctly identified edges among the edges of $\mathcal{G}_1$. The means of these quantities over the 2000 simulations are denoted FDR and power. Results are presented in Table 4. As expected, the power of all methods increases with the number of observations n. However, n has to be large in order to detect most of the edges. It comes from the fact that the graph $\mathcal{G}_1$ involves 11 proteins and 10 edges, which corresponds to a large percentage of edges (18%). In this study, we notice that the edges

$Raf - Mek1/2$ and $Erk1/2 - Akt$ are detected in most of the simulations even for small n and whatever the method; on the contrary the edge $PKA - Erk1/2$ is less often detected. It is in accordance with the values of the partial correlation matrix given in Figure 10: indeed the largest values of the partial correlation matrix correspond to the most often detected edges.

Let us now compare the methods according to the number of observations at our disposal. When n is small ($n = 15$), the `pcAlgo` and `KGGM` methods are the most powerful with a FDR around 1%. When n is moderate ($n = 30$ or $n = 100$), we advise to use the `MB.or` procedure, because the FDR is small and the benefit in power is large. When n is very large and referring to the power, all methods perform quite well. Nevertheless `KGGM` is slightly less powerfull. The FDR obtained with the `MB.and` procedure being null, this procedure is recommended.

FDR

n	`bagging`	`WB`	`MB.and`	`MB.or`	`pcAlgo`	`KGGM`
15	0.0227	0.0037	0.0007	0.0017	0.0086	0.0106
30	0.0159	0.0020	0.0011	0.0044	0.0030	0.0051
100	0.0117	0.0017	0.0001	0.0067	0.0018	0.0051
200	0.0098	0.0010	0.0000	0.0111	0.0011	0.0068
300	0.0056	0.0005	0.0000	0.0136	0.0005	0.0056

Power

n	`bagging`	`WB`	`MB.and`	`MB.or`	`pcAlgo`	`KGGM`
15	0.27	0.33	0.23	0.26	0.38	0.40
30	0.43	0.47	0.47	0.62	0.48	0.57
100	0.68	0.69	0.68	0.77	0.69	0.69
200	0.79	0.79	0.77	0.81	0.79	0.75
300	0.85	0.83	0.82	0.83	0.83	0.79

TABLE 4. FDR and power for estimating the graph $\mathcal{G}_1$. Results for the different methods and for different values of n.

5. CONCLUSION

In this work, we were interested in recent methods that infer direct links between entities, from experimental datasets. The results we obtained underline both common features and specificities of these methods regarding the parameters p, n and η of the application context. The most relevant points from our simulation study are the following:

- If one aims to control the FDR at a low level, `shrinkage` and `glasso` should not be used.

- `pcAlgo` gives a better control of the FDR when the parameter α is suitably chosen. However there is no simulation condition where it performs better than the `MB` methods for example.

- The `bagging` procedure is less powerfull than the others though the FDR is not better controled.

- The `WB` method has good performances, but we have to keep in mind that it aims at estimating an approximation of the concentration graph, and may lead to high FDR values when the 0-1 conditional independence graph differs from the concentration graph.

- `KGGM` performs well, particularly when n is small. However, this procedure cannot be carried out when p is large, say larger than 40.

- We recommend to use the `MB` procedure when it can be applied (n large enough so that Equation (11) is not satisfied). If one can accept a false discovery rate of the order 5%, then we recommend to use the variant `MB.or` which is more powerfull than the variant `MB.and`. This last one must be preferred when the false discovery rate has to be very small.

The structure of the graph should also be considered; if the edges are not uniformly distributed over the nodes as in the Erdös-Rényi model, then edges localized in highly connected parts of the graph or edges joining two highly connected parts may be difficult to detect.

In the end, methods inferring graphs do not behave equivalently faced to the graph and dataset structures. Consequently, we have to pay great attention to the validity domain of each method before carrying it out.

6. Appendix. Algorithm for the `MB` method

For each variable $a \in \Gamma$, let $(\widehat{\theta}_{a,b}(\lambda), b \in \Gamma^{-\{a\}})$ be the LASSO estimators of the parameters $\theta_{a,b}$ defined in Equation (4). In this section the algorithm used for detecting the $\widehat{\theta}_{a,b}$ that are non zero is described. The first step of the algorithm consists in using the LARS algorithm for ranking the variables $\boldsymbol{X}^{-\{a\}}$ according to the covariance structure of the matrix $\boldsymbol{X}$. Then, for the chosen value of λ, the non zero components of $(\widehat{\theta}_{a,b}(\lambda), b \in \Gamma^{-\{a\}})$ are detected. This second step is described below.

Let us define the following notations: for x a vector with q components, $\|x\|^2 = \sum_{l=1}^{q} x_l^2$, $\|x\|_\infty = \sup_{l=1,\ldots,q} |x_l|$. For the sake of simplicity, we set $\boldsymbol{Y} = \boldsymbol{X}^a$, $\boldsymbol{U} = \boldsymbol{X}^{-\{a\}}$, and we assume that $\boldsymbol{Y}$ and the columns of $\boldsymbol{U}$ are centered and scaled such that $\|\boldsymbol{Y}\|^2 = n$ and for all $b = 1, \ldots, q$ $(q = p-1)$, $\|\boldsymbol{U}^b\|^2 = n$. Let

$$\widehat{\beta}(\lambda) = \text{Arg} \min_{\beta \in \mathbb{R}^{p-1}} \|\boldsymbol{Y} - \boldsymbol{U}\beta\|^2 + \lambda \sum_{b=1}^{p-1} |\beta_b|. \tag{10}$$

We will use the following properties
Property A. $\widehat{\beta}(\lambda)$ is solution of Equation (10) if and only if there exists a $p-1$-vector v satisfying

- for all $b = 1, \ldots, p-1$, $v_b = \text{sign}(\widehat{\beta}_b(\lambda))$ if $\widehat{\beta}_b(\lambda) \neq 0$ and $v_b \in [-1, 1]$ if not
- $\lambda v = 2\boldsymbol{U}^T(\boldsymbol{Y} - \boldsymbol{U}\widehat{\beta}(\lambda))$.

Property B. Solving (10) is equivalent to solving the following constraint mimimization problem

$$\widehat{\beta}(t) = \text{Arg} \min_{\sum_{b=1}^{p-1} |\beta_b| \leq t} \|\boldsymbol{Y} - \boldsymbol{U}\beta\|^2 .$$

Therefore, for all λ, there exists t_λ such that $\widehat{\beta}(\lambda) = \widehat{\beta}(t_\lambda)$.
Property C. Let

$$C(t) = 2 \left\| \boldsymbol{U}^T \left(Y - \boldsymbol{U}\widehat{\beta}(t) \right) \right\|_\infty .$$

It can be shown that the function C satisfies the following properties: C is a decreasing function of t (see Efron et al. (2004), lemma 7), and $\lambda = C(t_\lambda)$.

From Property A it comes out that $\lambda \geq 2\|\boldsymbol{U}^T\boldsymbol{Y}\|_\infty$ is equivalent to $\widehat{\beta}(\lambda) = 0$. As $2\|\boldsymbol{U}^T\boldsymbol{Y}\|_\infty \leq 2n$, we get that $\widehat{\beta}(\lambda) = 0$ as soon as $\lambda \geq 2n$. Comparing this lower bound with the value λ given by Meinshausen and Bühlmann (see Equation (5)), it appears that the parameters will be estimated by zero whatever the data if

$$n \leq \left\{ \Phi^{-1}(1 - \alpha/2p^2) \right\}^2 . \tag{11}$$

Consider now the case where $\lambda < 2\|\boldsymbol{U}^T\boldsymbol{Y}\|_\infty$ and let us denote by $\mathcal{A}(t)$ the set of active parameters :

$$\mathcal{A}(t) = \left\{ b \in \{1, \ldots, p-1\}, \widehat{\beta}_b(t) \neq 0 \right\}.$$

When t increases $\mathcal{A}(t)$ becomes larger. The LARS algorithm gives the values of t, $t_0 = 0, t_1, t_2, \ldots$, for which $\mathcal{A}(t)$ gains a variable: for each $k = 0, 1, 2 \ldots$, for all $t \in]t_{k-1}, t_k]$, $\mathcal{A}(t)$ is constant and equals $\mathcal{A}(t_k)$.

Thanks to the third property it remains to find $k^* = \min\{k, C(t_k) < \lambda\}$. The non zero components of $(\widehat{\theta}_{a,b}(\lambda), b \in \Gamma^{-\{a\}})$ are then equal to $\mathcal{A}(t_{k^*})$.

References

Banerjee, O., Ghaoui, L., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Machine Learning Research*, *9*, 485–516.

Castelo, R., & Roverato, A. (2006). A robust procedure for gaussian graphical model search from microarray data with p larger than n. *Journal of Machine Learning Research*, *7*, 2621–2650.

Daudin, J. J., Picard, F., & Robin, S. (2006). A mixture model for random graphs. Tech. Rep. RR-5840, INRIA, Rapport de Recherche.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., & West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, *90*, 196–212.

Drton, M., & Perlman, M. (2007). Multiple testing and error control in gaussian graphical model selection. *Statistical Sciences, To appear*.

Efron, B., Hastie, T., Johnston, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*, 407–451.

Friedman, J., Hastie, T., & Tibshirani, R. (2007). Sparse inverse covariance estimation with the lasso. Tech. Rep. , , `http://www-stat.stanford.edu/tibs/ftp/graph.pdf`.

Giraud, C. (2008). Estimation of gaussian graphs by model selection. *Electronic Journal of Statistics*, *2*, 542–563.

Huang, J., Liu, N., Pourahmadi, M., & Liu, L. (2006). Covariance matrix selection abd estimation via penalised normal likelihood. *Biometrika*, *93*(1), 85–98.

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics*, *19*, 2271–2282.

Kalisch, M., & Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, *8*, 613–636.

Kishino, H., & Waddell, P. J. (2000). Correspondence analysis of genes and tissue types and finding genetic links from microarray data. *Genome Informatics*, *11*, 83–95.

Malouche, D., & Sevestre, S. (2007). Estimating high dimensional faithful gaussian graphical models : upc-algorithm. Tech. Rep. arXiv:0705.1613, Technical Report.

Meinshausen, N., & Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, *34*(3), 1436–1462.

Okamoto, S., Yamanishi, Y., Ehira, S., Kawashima, S., Tonomura, K., & Kanehisa, M. (2007). Prediction of nitrogen metabolism-related genes in anabaena by kernel-based network analysis. *Proteomics*, *7*(6), 900–909.

Sachs, K., Perez, O., D.Pe'er, Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, *308*, 523–529.

Schäfer, J., & Strimmer, K. (2005a). An empirical bayes approach to inferring large-scale gene association nerworks. *Bioinformatics*, *21*(6), 754–764.

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

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction and Search*. London: The MIT Press 2nd edition.

Werhli, A., & Husmeier, D. (2007). Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Statistical Applications in Genetics and Molecular Biology*, *6*.

Wille, A., & Bühlmann, P. (2006). Low-order conditional independence graphs for inferring genetic networks. *Statistical Applications in Genetics and Molecular Biology*, *5*, 1–34.

Wu, W., & Ye, Y. (2006). Exploring gene causal interactions using an enhanced constraint-based method. *Pattern Recognition*, *39*, 2439–2449.

Yellaboina, S., Goyal, K., & Mande, S. (2007). Inferring genome-wide functional linkages in e-coli by combining improved genome context methods: Comparison with high-throughput experimental data. *Genome Research*, *17*(4), 527–535.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, *94*( ), 19–35.