Graph selection with GGMselect

Christophe Giraud, Sylvie Huet and Nicolas Verzelen

Abstract: Applications on inference of biological networks have raised a strong interest in the problem of graph estimation in high-dimensional Gaussian graphical models. To handle this problem, we propose a two-stage procedure which first builds a family of candidate graphs from the data, and then selects one graph among this family according to a dedicated criterion. This estimation procedure is shown to be consistent in a high-dimensional setting, and its risk is controlled by a non-asymptotic oracle-like inequality. The procedure is tested on a real data set concerning gene expression data, and its performances are assessed on the basis of a large numerical study. The procedure is implemented in the R-package GGMselect available on the CRAN.

AMS 2000 subject classifications: Primary 62G08; secondary 62J05.
Keywords and phrases: Gaussian graphical model, Model selection, Penalized empirical risk.

1. Introduction

Biotechnological developments in proteomics or transcriptomics enable to produce a huge amount of data. One of the challenges for the statistician is to infer from these data the regulation network of a family of genes (or proteins). The task is difficult due to the very high-dimensional nature of the data and the small sample size. For example, microarrays measure the expression levels of a few thousand genes and the sample size n is no more than a few tens. When no additional information is available, the Gaussian graphical modeling, denoted GGM, has been proposed as a tool to handle this issue, see e.g. [19, 11, 32]. Graphical modeling is based on the notion of conditional dependency. The principle underlying the GGM approach is the following : the existence of a regulation dependence between two genes corresponds to the existence of a conditional dependence between their gene expression levels. The conditional dependences between the gene expression levels are represented by a graph G, where each node represents a gene and where an edge is set between two nodes a and b if there exists a conditional dependence between their gene expression levels. According to the GGM principle, this graph G coincides with the gene regulation network.

Let us describe more precisely the GGM setting. The gene expression levels $(X_1, ..., X_p)$ of p genes are modeled by a centered Gaussian law with covariance matrix Σ, denoted $\mathbb{P}_\Sigma$. This law $\mathbb{P}_\Sigma$ is a so-called graphical model according to a graph G, if for any genes a and b that are not neighbours in G, the variables X_a and X_b are independent conditionally on the remaining variables. Roughly speaking, if genes a and b are not neighbours in G, the variables X_a and X_b are uncorrelated when the values of the remaining variables are fixed. There exists a unique graph G_Σ which is minimal for the inclusion and such that $\mathbb{P}_\Sigma$ is a graphical model according to G_Σ. An edge between a and b in G_Σ therefore represents the existence

of a conditional dependence between the variables X_a and X_b. As mentioned before, G_Σ is a graph of interest since it shall coincide with the gene regulation network. Our aim in this paper is to estimate this graph from microarrays data which are assumed to be a n-sample of the law $\mathbb{P}_\Sigma$. We will pay a special attention to the case where $n < p$ and we assume in the following that Σ is non-singular.

The problem of graph estimation in Gaussian graphical model when the sample size n is smaller (or much smaller) than the number p of variables is a current active field of research in statistics. Many estimation procedures have been proposed recently to perform graph estimation in Gaussian graphical model when $n < p$. A first class of procedures is based on multiple testing on empirical partial covariance. If G_Σ denotes the (minimal) graph of the law $\mathbb{P}_\Sigma$, there is an edge in G_Σ between a and b, if and only if the conditional covariance of X_a and X_b given all the other variables is non-zero. When $n < p$, the empirical version of the latter conditional covariance cannot be computed, so several papers suggest to use instead the empirical conditional covariance of X_a and X_b given $\{X_s,\ s \in S\}$ for some subsets S of $\{1, \dots, p\} \setminus \{a, b\}$ with cardinality less than $n-2$. A multiple testing procedure is then applied to detect if the conditional covariance $\mathrm{cov}(X_a, X_b | X_s,\ s \in S)$ is non-zero. Wille and Bühlmann [30] restrict to the sets S of cardinality less or equal to one, Castelo and Roverato [6] consider the sets S with cardinality at most q (for some fixed q) and Spirtes et al. [27] (see also [18]) propose a procedure which avoid an exhaustive search over all S. A second class of procedures relies on the fact that the entries $\Omega_{a,b}$ of the inverse covariance matrix $\Omega = \Sigma^{-1}$ are non-zero if and only if there is an edge between a and b in G_Σ. Several papers then suggest to perform a sparse estimation of Ω in order to estimate the graph G_Σ, see [17, 33, 2, 14, 13]. They propose to maximize the log-likelihood of Ω under l^1 constraints to enforce sparsity and they design optimization algorithms to perform this maximization. A third class of procedures uses the fact that the coefficients $\theta_{a,b}$ of the regression of X_a on $\{X_b,\ b \neq a\}$ are non-zeros if and only if there is an edge between a and b in G_Σ. Meinshausen and Bühlmann [22] and Rocha et al. [24] perform regressions with l^1 constraints, whereas Giraud [15] (see also [28]) proposes an exhaustive search over the set of sparse graphs to obtain a sparse estimate of the matrix θ and then detect the graph G_Σ. Finally, a series of papers (e.g. [31, 10, 26]) investigate a Bayesian approach to estimate the graph.

In this paper, we propose a new estimation scheme which combines the good properties of different procedures. On the one hand, the procedures based on the empirical covariance or on l^1 regularisation share some nice computational properties and they can handle several hundred of variables $X_1, \dots, X_p$. Nevertheless, the theoretical results assessing their statistical accuracy are either of asymptotic nature or rely on strong assumptions on the covariance [20, 25]. Moreover, their performance heavily depends on one (or several) tuning parameter, which is usually not dimensionless and whose optimal value is unknown. To cope with this issue, many authors propose to apply cross-validation or the BIC criterion. However, the BIC criterion often overfits in a high dimensional setting (see [3] and the simulations Section 4) and cross-validation offers little theoretical warranty. On the other hand, the method proposed by Giraud [15] has a good statistical accu-

racy and strong theoretical results have been established, but its computational complexity is huge and it cannot be performed when the number p of variables is larger than a few tens.

Our strategy here is to build a data-driven family of candidate graphs using several fast above-mentioned procedures and then to apply the selection criterion presented in [15] to select one graph among them. We show that this criterion can be used both

(i) to choose the tuning parameter(s) of any estimation procedure with no need of any additional knowledge. As such, our criterion is an alternative to the BIC criterion and cross-validation.
(ii) and to compare several graphs produced by various estimation procedures, including graphs built from a priori knowledge.

Our estimation procedure can handle several hundred of variables $X_1, \ldots, X_p$ and presents good statistical properties. It is proved to be consistent in a high-dimensional setting. Furthermore, its risk is controlled by a non-asymptotic oracle-like inequality. This means that the risk of our estimator is almost as small as if we knew in advance the best graph in the data-driven family. A more formalized definition of the oracle property is given in Section 3.1. In contrast to other results in the literature [15, 22], this oracle inequality allows to deal with a data-driven collection of graphs. In addition, we propose families of candidate graphs which work well in practice as shown on simulated examples. Finally, the procedure is implemented in the *R*-package *GGMselect* available on the *Comprehensive R Archive Network*. `http://cran.r-project.org/`

The remaining of the paper is organized as follows. We describe the estimation procedure in the next section and state some theoretical results on its statistical accuracy in Section 3. In Section 4, we carry out some numerical experiments in order to assess the performances of our procedure. In Section 5, we test our method on a real data set concerning gene expression data provided in [16] and already analyzed in [1]. Section 6 is devoted to the proofs. Details on the collections of graphs are postponed to Section 7.

Notations. To estimate the graph G_Σ, we will start from a n-sample $X^{(1)}, \ldots, X^{(n)}$ of the law $\mathbb{P}_\Sigma$. We denote by $\mathbf{X}$ the $n \times p$ matrix whose rows are given by the vectors $X^{(i)}$, namely $\mathbf{X}_{i,a} = X_a^{(i)}$ for $i = 1, \ldots, n$ and $a = 1, \ldots, p$. We write $\mathbf{X}_a$ for the a^{th} column of $\mathbf{X}$. We also set $\Gamma = \{1, \ldots, p\}$ and for any graph G with nodes indexed by Γ, we write $d_a(G)$ for the degree of the node a in the graph G (which is the number of edges incident to a) and $\deg(G) = \max_{a \in \Gamma} d_a(G)$ for the degree of G. Moreover, the notation $a \overset{G}{\sim} b$ means that the nodes a and b are neighbours in the graph G. Finally, we write Θ for the set of $p \times p$ matrices with

0 on the diagonal, $\|\cdot\|_{q\times p}$ for the Frobenius norm on $q\times p$ matrices

$$\|A\|^2_{q\times p} = \mathrm{Tr}(A^TA) = \sum_{i=1}^{q}\sum_{j=1}^{p} A^2_{i,j}\ ,$$

$\|\cdot\|_n$ for the Euclidean norm on $\mathbb{R}^n$ divided by $\sqrt{n}$, and for any $\beta\in\mathbb{R}^p$ we define $\mathrm{supp}(\beta)$ as the set of the labels $a\in\Gamma$ such that $\beta_a\neq 0$.

2. Estimation procedure

GGMselect is a two-stage estimation procedure which first builds a data-driven family $\widehat{\mathcal{G}}$ of candidate graphs and then applies a selection procedure to pick one graph among these. We present the selection procedure in the next paragraph and then describe different possible choices for the family of candidate graphs $\widehat{\mathcal{G}}$.

2.1. Selection procedure

We assume here that we have at hand a family $\widehat{\mathcal{G}}$ of candidate graphs, which all have a degree smaller than $n-2$. To select a graph $\widehat{G}$ among the family $\widehat{\mathcal{G}}$, we use the selection criterion introduced in [15]. We write θ for the $p\times p$ matrix such that

$$\mathbb{E}_\Sigma\left[X_a\middle|X_b,\ b\neq a\right] = \sum_{b\neq a}\theta_{a,b}X_b \quad\text{and}\quad \theta_{a,a}=0 \quad \text{for all } a\in\{1,\dots,p\}\,.$$

The matrix θ minimizes $\|\Sigma^{1/2}(I-\theta')\|_{p\times p}$ over the set Θ of $p\times p$ matrices θ' with 0 on the diagonal. Since $\mathbf{X}^T\mathbf{X}/n$ is an empirical version of Σ, an empirical version of $\|\Sigma^{1/2}(I-\theta)\|_{p\times p}$ is $\|\mathbf{X}(I-\theta)\|_{n\times p}$ divided by $\sqrt{n}$. Therefore, for any graph G in $\widehat{\mathcal{G}}$, we associate an estimator $\widehat{\theta}_G$ of θ by setting

$$\widehat{\theta}_G = \mathrm{argmin}\left\{\|\mathbf{X}(I-\theta')\|_{n\times p} : \theta'\in\Theta_G\right\}, \tag{1}$$

where Θ_G is the set of $p\times p$ matrices θ' such that $\theta'_{a,b}$ is non-zero if and only if there is an edge between a and b in G.

Finally, we select a graph $\widehat{G}$ in $\widehat{\mathcal{G}}$ by taking any minimizer over $\widehat{\mathcal{G}}$ of the criterion

$$\mathrm{Crit}(G) = \sum_{a=1}^{p}\left[\|\mathbf{X}_a - \mathbf{X}[\widehat{\theta}_G]_a\|^2_n\left(1+\frac{\mathrm{pen}[d_a(G)]}{n-d_a(G)}\right)\right]\ , \tag{2}$$

where $d_a(G)$ is the degree of the node a in the graph G and the penalty function $\mathrm{pen}:\mathbb{N}\to\mathbb{R}^+$ is of the form of the penalties introduced in Baraud et al. [3] for the fixed design regression model. To compute this penalty, we define for any integers d and N the DKhi function by

$$\mathrm{DKhi}(d,N,x) = \mathbb{P}\left(F_{d+2,N}\geq\frac{x}{d+2}\right) - \frac{x}{d}\,\mathbb{P}\left(F_{d,N+2}\geq\frac{N+2}{Nd}x\right),\ x>0\,,$$

where $F_{d,N}$ denotes a Fisher random variable with d and N degrees of freedom. The function $x \mapsto \mathrm{DKhi}(d, N, x)$ is decreasing and we write $\mathrm{EDKhi}[d, N, x]$ for its inverse, see [3] Sect. 6.1 for more details. Then, we fix some constant $K > 1$ and set

$$\mathrm{pen}(d) = K \frac{n-d}{n-d-1} \mathrm{EDKhi}\left[d+1, n-d-1, \left(\binom{p-1}{d}(d+1)^2\right)^{-1}\right]. \quad (3)$$

When d remains small compared to n, the penalty function increases approximately linearly with d. Actually, when $d \leq \gamma\, n/\left(2\left(1.1+\sqrt{\log p}\right)^2\right)$ for some $\gamma < 1$, we approximately have for large values of p and n

$$\mathrm{pen}(d) \lesssim K\left(1+e^{\gamma}\sqrt{2\log p}\right)^2 (d+1),$$

see Proposition 4 in [3] for an exact bound.

The selection procedure depends on a dimensionless tuning parameter K. A larger value for K yields a procedure more conservative. In theory (and in practice) K has to be larger than one. In our simulations, we set $K = 2.5$.

2.2. Family $\widehat{\mathcal{G}}$ of candidate graphs

The computational complexity of the minimization of the criterion (2) over the family $\widehat{\mathcal{G}}$ is linear with respect to its size. In particular, minimizing (2) over all the graphs with degree smaller than some integer D, as proposed in [15], is intractable when p is larger than a few tens. To overcome this issue, we propose to build a much smaller (data-driven) family $\widehat{\mathcal{G}}$ of candidate graphs, with the help of various fast algorithms dedicated to graph estimation.

Since the procedure applies for any family $\widehat{\mathcal{G}}$, GGMselect allows to select the tuning parameter(s) of any graph estimation procedure and also to compare any collection of estimation procedures. Nevertheless, we advise in practice to choose one of the four families of candidate graphs $\widehat{\mathcal{G}}_{\mathrm{EW}}$, $\widehat{\mathcal{G}}_{\mathrm{C01}}$, $\widehat{\mathcal{G}}_{\mathrm{LA}}$, and $\widehat{\mathcal{G}}_{\mathrm{QE}}$ presented below, or the union of them. These families have been chosen on the basis of theoretical results and simulation studies.

In the following, we explain how to tune and compare graph estimation procedures with GGMselect. Afterwards, we describe the four above-mentioned families, provide algorithms to compute them efficiently, and discuss their computational complexity and their size. Each family depends on an integer D, smaller than $n-2$, which corresponds to the maximal degree of the graphs in this family.

2.2.1. Tuning a procedure and comparing several ones

Suppose we are given an estimation procedure P depending on a tuning parameter $\lambda > 0$ whose optimal value is unknown or depends on unknown quantities. Let us denote by $\widehat{\mathcal{G}}_P$ the collection of graphs estimated using this procedure P :

$$\widehat{\mathcal{G}}_P = \left\{\widehat{G}_P(\lambda)\ ,\ \lambda > 0 \text{ and } \deg(G_P(\lambda)) \leq D\right\}. \quad (4)$$

We propose to choose λ by minimizing the criterion (2) over the collection $\widehat{\mathcal{G}}_P$. Thus, we get an estimated graph $\widehat{G}_P = \widehat{G}_P(\widehat{\lambda}_P)$. Theorems 3.2 and 3.3 in Section 3 state that GGMselect almost selects the best graph among this collection.

Assume now that we have at hand a collection $\mathcal{P}$ of estimation procedures which possibly depends on tuning parameters. For any procedure $P \in \mathcal{P}$, we compute the collection $\widehat{\mathcal{G}}_P$ either defined by (4) if P depends on tuning parameters or by $\{\widehat{G}_P\}$ if not. Then, we propose to select a procedure $\hat{P}$ and a graph $\widehat{G}_{\widehat{P}}$ by minimizing the criterion (2) over the collection

$$\widehat{\mathcal{G}}_{\mathcal{P}} = \left\{ \widehat{\mathcal{G}}_P \ , \ P \in \mathcal{P} \right\} \ . \tag{5}$$

Again, Theorems 3.2 and 3.3 in Section 3 ensure that GGMselect almost selects the best graph among the collection $\widehat{\mathcal{G}}_{\mathcal{P}}$.

Next, we briefly describe the four families of candidate graphs $\widehat{\mathcal{G}}_{\mathrm{C01}}$, $\widehat{\mathcal{G}}_{\mathrm{LA}}$, $\widehat{\mathcal{G}}_{\mathrm{EW}}$ and $\widehat{\mathcal{G}}_{\mathrm{QE}}$ that we advise to use, the details being postponed to Section 7. Except $\widehat{\mathcal{G}}_{\mathrm{QE}}$, all the other families are built from estimators in the literature that depend on an unknown tuning parameter. On the one hand, GGMselect allows to tune these procedures. On the other hand, GGMselect allows to select an estimator by combining these different procedures.

2.2.2. C01 family $\widehat{\mathcal{G}}_{\mathrm{C01}}$

The family $\widehat{\mathcal{G}}_{\mathrm{C01}}$ derives from the estimation procedure proposed by Wille and Bühlmann [30] and is based on the 0-1 *conditional independence graph* G_{01}. This graph is defined as follows. For each pair of nodes (a, b), we write $R_{a,b|\emptyset}$ for the correlation between the variables X_a and X_b and $R_{a,b|c}$ for the correlation of X_a and X_b conditionally on X_c. Then, there is an edge between a and b in G_{01}, if and only if $R_{a,b|\emptyset} \neq 0$ and $R_{a,b|c} \neq 0$ for all $c \in \Gamma \setminus \{a, b\}$, viz

$$a \overset{G_{01}}{\sim} b \iff \min\left\{ |R_{a,b|c}|, \ c \in \{\emptyset\} \cup \Gamma \setminus \{a, b\} \right\} > 0 \, . \tag{6}$$

Although the 0-1 conditional independence graph G_{01} does not usually coincide with the graph G_Σ, there is a close connection between both graphs in some cases (see Wille and Bühlmann). Given a number $0 < \alpha < 1$, Wille and Bühlmann propose to estimate G_{01} by a graph $\widehat{G}_{01,\alpha}$ built from a collection of likelihood ratio test level of α. The graph $\widehat{G}_{01,\alpha}$ becomes more connected when α increases. We define the family $\widehat{\mathcal{G}}_{\mathrm{C01}}$ as the set of graphs $\widehat{G}_{01,\alpha}$ with all levels α small enough to ensure that $\deg(\widehat{G}_{01,\alpha}) \leq D$.

Complexity. The computation of $\widehat{\mathcal{G}}_{\mathrm{C01}}$ goes very fast since its complexity is of order np^3 (see Section 7). The size of the family $\widehat{\mathcal{G}}_{\mathrm{C01}}$ is smaller than pD. Computational times for some examples are given in Section 4.1.

2.2.3. Lasso-And family $\widehat{\mathcal{G}}_{\mathrm{LA}}$

The Lasso-And family $\widehat{\mathcal{G}}_{\mathrm{LA}}$ derives from the estimation procedure proposed by Meinshausen and Bühlmann [22] and is based on the LARS-lasso algorithm [12]. For any $\lambda > 0$, we define the $p \times p$ matrix $\widehat{\theta}^{\lambda}$ by

$$\widehat{\theta}^{\lambda} = \operatorname{argmin} \left\{ \|\mathbf{X} - \mathbf{X}\theta'\|_{n\times p}^2 + \lambda \|\theta'\|_1 : \ \theta' \in \Theta \right\}, \tag{7}$$

where Θ is the set of $p \times p$ matrices with 0 on the diagonal and $\|\theta'\|_1 = \sum_{a\neq b} |\theta'_{a,b}|$. Then, we define the graph $\widehat{G}^{\lambda}_{\mathrm{and}}$ by setting an edge between a and b if both $\widehat{\theta}^{\lambda}_{a,b}$ and $\widehat{\theta}^{\lambda}_{b,a}$ are non-zero. This graph $\widehat{G}^{\lambda}_{\mathrm{and}}$ is exactly the estimator (7) introduced in [22]. The size of $\widehat{G}^{\lambda}_{\mathrm{and}}$ has a tendency to increase when the tuning parameter λ decreases. Hence, we define the family $\widehat{\mathcal{G}}_{\mathrm{LA}}$ as the set of graphs $\widehat{G}^{\lambda}_{\mathrm{and}}$ with all λ large enough to ensure that $\deg(\widehat{G}^{\lambda}_{\mathrm{and}}) \leq D$.

Complexity. The complexity of the LARS-lasso algorithm is unknown in general. Nevertheless, according to Efron et al. [12] the algorithm requires $O(np(n \wedge p))$ operations in most cases. Hence, the whole complexity of the LA algorithm is generally of the order $p^2 n(n \wedge p)$ (see Section 7). Finally, the size of the family $\widehat{\mathcal{G}}_{\mathrm{LA}}$ cannot be bounded uniformly, but it remains smaller than pD in practice.

2.2.4. Adaptive lasso family $\widehat{\mathcal{G}}_{\mathrm{EW}}$

The family $\widehat{\mathcal{G}}_{\mathrm{EW}}$ is a modified version of $\widehat{\mathcal{G}}_{\mathrm{LA}}$ inspired by the adaptive lasso [34]. The major difference between $\widehat{\mathcal{G}}_{\mathrm{EW}}$ and $\widehat{\mathcal{G}}_{\mathrm{LA}}$ lies in the replacement of the l^1 norm $\|\theta'\|_1$ in (7) by $\|\theta'/\widehat{\theta}^{\mathrm{init}}\|_1$, where $\widehat{\theta}^{\mathrm{init}}$ is a preliminary estimator of θ and $\theta'/\widehat{\theta}^{\mathrm{init}}$ stands for the matrix with entries $(\theta'/\widehat{\theta}^{\mathrm{init}})_{a,b} = \theta'_{a,b}/\widehat{\theta}^{\mathrm{init}}_{a,b}$. Zou suggests to take for $\widehat{\theta}^{\mathrm{init}}$ a ridge estimator. Here, we propose to use instead the Exponential Weights estimator $\widehat{\theta}^{EW}$ of Dalalyan and Tsybakov [8, 9]. The choice of this estimator appears more natural to us since it is designed for the sparse setting and enjoys nice theoretical properties. Moreover, we have observed on some simulated examples, that the adaptive lasso with the Exponential Weights initial estimator performs much better than the adaptive lasso with the ridge initial estimator.

Given $\lambda > 0$, $\widehat{\theta}^{\mathrm{EW},\lambda}$ is the adaptive lasso estimator of θ with initial estimator $\widehat{\theta}^{\mathrm{EW}}$. We define the graph $\widehat{G}^{\mathrm{EW},\lambda}_{\mathrm{or}}$ by setting an edge between a and b if either $\widehat{\theta}^{EW,\lambda}_{b,a}$ or $\widehat{\theta}^{EW,\lambda}_{a,b}$ is non-zero. Finally, the family $\widehat{\mathcal{G}}_{\mathrm{EW}}$ is the set of graphs $\widehat{G}^{\mathrm{EW},\lambda}_{\mathrm{or}}$ with λ large enough to ensure that $\deg(\widehat{G}^{\mathrm{EW},\lambda}_{\mathrm{or}}) \leq D$.

Complexity. The complexity of the estimation $\widehat{\theta}^{\mathrm{EW}}$ depends on the choices of the tuning parameters for the Exponential Weights estimator (see Section 7). Some examples are given in Section 4.1. The complexity of the other computations is the same as for the LA-algorithm and is of the order $p^2 n(n \wedge p)$ in practice. Finally, as for $\widehat{\mathcal{G}}_{\mathrm{LA}}$, we do not know a general bound for the size of $\widehat{\mathcal{G}}_{\mathrm{EW}}$, but it remains smaller than pD in practice.

2.2.5. Quasi-exhaustive family $\widehat{\mathcal{G}}_{\mathrm{QE}}$

Roughly, the idea is to break down the minimization of the criterion (2) over all the graphs of degree at most D into p independent problems. For each node $a \in \Gamma$, we estimate the neighborhood of a by

$$\widehat{\mathrm{ne}}(a) = \operatorname{argmin}\left\{ \|\mathbf{X}_a - \mathrm{Proj}_{V_S}(\mathbf{X}_a)\|_n^2 \left(1 + \frac{\mathrm{pen}(|S|)}{n - |S|}\right) : \ S \subset \Gamma \setminus \{a\} \text{ and } |S| \leq D \right\},$$

where pen is the penalty function (3) and Proj_{V_S} denotes the orthogonal projection from $\mathbb{R}^n$ onto $V_S = \{\mathbf{X}\beta : \beta \in \mathbb{R}^p \text{ and } \mathrm{supp}(\beta) = S\}$. We know from [28] that $\widehat{\mathrm{ne}}(a)$ is a good estimator of the true neighborhood of a, from a non-asymptotic point of view. We then build two nested graphs $\widehat{G}_{K,\mathrm{and}}$ and $\widehat{G}_{K,\mathrm{or}}$ in a similar way as in [22]. Namely, there is an edge between a and b in $\widehat{G}_{K,\mathrm{and}}$ if $a \in \widehat{\mathrm{ne}}(b)$ and $b \in \widehat{\mathrm{ne}}(a)$ and there is an edge between a and b in $\widehat{G}_{K,\mathrm{or}}$ if $a \in \widehat{\mathrm{ne}}(b)$ or $b \in \widehat{\mathrm{ne}}(a)$. The family $\widehat{\mathcal{G}}_{\mathrm{QE}}$ is defined as the collection of all the graphs that lie between $\widehat{G}_{K,\mathrm{and}}$ and $\widehat{G}_{K,\mathrm{or}}$

$$\widehat{\mathcal{G}}_{\mathrm{QE}} = \left\{ G, \ \widehat{G}_{K,\mathrm{and}} \subset G \subset \widehat{G}_{K,\mathrm{or}} \text{ and } \deg(G) \leq D \right\}.$$

It is likely that the graph $\widehat{G}_{\mathrm{exhaustive}}$ which minimizes (2) over all the graphs of degree at most D belongs to the family $\widehat{\mathcal{G}}_{\mathrm{QE}}$. In such a case, the minimizer $\widehat{G}_{\mathrm{QE}}$ of the criterion (2) over $\widehat{\mathcal{G}}_{\mathrm{QE}}$ coincides with the estimator $\widehat{G}_{\mathrm{exhaustive}}$.

Complexity. The complexity of the computation of the collections $\widehat{\mathrm{ne}}(a)$ is much smaller than the complexity of the computation of $\widehat{G}_{\mathrm{exhaustive}}$. Nevertheless, it still remains of order $np^{D+1}D^3$ and the size of the family $\widehat{\mathcal{G}}_{\mathrm{QE}}$ can be of order $2^{pD/2}$ in the worst cases. However, for sparse graphs G_Σ, the graphs $\widehat{G}_{K,\mathrm{and}}$ and $\widehat{G}_{K,\mathrm{or}}$ are quite similar in practice, which makes the size of $\widehat{\mathcal{G}}_{\mathrm{QE}}$ much smaller. The procedure then remains tractable for p and D reasonably small.

3. Theoretical results

In order to assess the performance of our selection procedure, we state in this section two kinds of theoretical results: a non-asymptotic oracle-like inequality concerning the estimation of θ and a consistency result for the estimation of G_Σ.

3.1. A non-asymptotic oracle-like inequality

We associate to the graph $\widehat{G}$ selected by the procedure of Section 2, the estimator $\tilde{\theta} = \widehat{\theta}_{\widehat{G}}$ of the matrix θ, where $\widehat{\theta}_G$ is given by (1) for any graph $G \in \widehat{\mathcal{G}}$. The quality of the estimation of θ is quantified by the MSEP of $\tilde{\theta}$ defined by

$$\mathrm{MSEP}(\tilde{\theta}) = \mathbb{E}\left[\|\Sigma^{1/2}(\tilde{\theta} - \theta)\|_{p\times p}^2 \right].$$

We refer to the introduction of [15] for a discussion on the relevance of the use of the MSEP of $\tilde{\theta}$ to assess the quality of the estimator $\widehat{G}$. In the sequel, I stands for the identity matrix of size p.

First, we can compare the MSEP of $\tilde{\theta}$ to the MSEP of $\widehat{\theta}_{G_\Sigma}$ when the minimal graph G_Σ belongs to $\widehat{\mathcal{G}}$ with large probability. Roughly speaking, the MSEP of $\tilde{\theta}$ is in this case smaller (up to a $\log p$ factor) than the MSEP of $\widehat{\theta}_{G_\Sigma}$. This means that $\tilde{\theta}$ performs almost as well as if we knew the true graph G_Σ in advance.

Proposition 3.1. *Assume that $n \geq 9$. Let $\widehat{\mathcal{G}}$ be any (data-driven) family of graphs with maximal degree $D_{\widehat{\mathcal{G}}} = \max\{\deg(G),\ G \in \widehat{\mathcal{G}}\}$ fulfilling*

$$1 \leq D_{\widehat{\mathcal{G}}} \leq \gamma \, \frac{n}{2(1.1 + \sqrt{\log p})^2}\ , \qquad \text{for some } \gamma < 1\,. \tag{8}$$

If the minimal graph G_Σ belongs to the family $\widehat{\mathcal{G}}$ with large probability

$$\mathbb{P}\left(G_\Sigma \in \widehat{\mathcal{G}}\right) \geq 1 - \alpha \exp(-\beta n^\delta), \qquad \text{for some } \alpha, \beta, \delta > 0 \tag{9}$$

then, the MSEP *of the estimator $\widetilde{\theta}$ is upper bounded by*

$$\mathrm{MSEP}(\widetilde{\theta}) \leq L_{K,\gamma} \log(p) \left(\mathrm{MSEP}(\widehat{\theta}_{G_\Sigma}) \vee \frac{\mathrm{MSEP}(I)}{n} \right) + R_n\,. \tag{10}$$

where $L_{K,\gamma}$ is a positive constant depending on K and γ only and the residual term $R_n = R_n(\Sigma, \gamma, \alpha, \beta, \delta)$ is of order $n^3 tr(\Sigma)[e^{-n(\sqrt{\gamma}-\gamma)^2/4} + \sqrt{\alpha} e^{-\frac{\beta}{2} n^\delta}]$.

Observe that the residual term R_n goes to 0 exponentially fast with respect to n. If we forget the term $n^{-1}\mathrm{MSEP}(I)$, then the risk bound (10) essentially states that the estimator $\tilde{\theta}$ performs almost as well as if we knew the graph G_Σ in advance.

Let us now compare the additional term $n^{-1}\mathrm{MSEP}(I)$ appearing in (10) with the risk $\mathrm{MSEP}(\widehat{\theta}_{G_\Sigma})$. The additional term $n^{-1}\mathrm{MSEP}(I)$ is equal to $n^{-1}\sum_a \sigma_a^2$, where σ_a^2 stands for the conditional variance of X_a given the remaining variables. Hence, this quantity is usually smaller than the risk $\mathrm{MSEP}(\widehat{\theta}_{G_\Sigma})$ which is a variance term of order $n^{-1}\sum_a d_a(G_\Sigma)\sigma_a^2$. Nevertheless, when the true graph G_Σ is empty and the collection $\widehat{\mathcal{G}}$ contains the empty graph, the additional term $n^{-1}\mathrm{MSEP}(I)$ is dominant and the estimator $\tilde{\theta}$ is not optimal. Such a drawback is actually unavoidable in model selection when the target is too close to zero (see Sect.2.3.3 of [5] for a discussion). Assumption (8) is discussed after Theorem 3.2.

In Proposition 3.1, we state that $\tilde{\theta}$ performs almost as well as $\widehat{\theta}_{G_\Sigma}$. Nevertheless, the risk of the estimator $\widehat{\theta}_{G_\Sigma}$ can be quite large, especially when the graph G_Σ contains a lot of edges. For an arbitrary graph G, the risk $\mathrm{MSEP}(\widehat{\theta}_G)$ is the sum of the bias and the variance terms. If we consider a sparser graph G, the estimator $\widehat{\theta}_G$ is biased but its variance is smaller, so its risk $\mathrm{MSEP}(\widehat{\theta}_G)$ can be smaller. The estimator $\widehat{\theta}_{G^*}$ which minimizes the MSEP over the collection of estimators $(\widehat{\theta}_G)_{G \in \widehat{\mathcal{G}}}$ is called the *oracle*. Observe that the graph G^* is unknown since it is related to the unknown matrix θ. One goal of model selection is to select an estimator $\tilde{\theta}$ which performs almost as well as the oracle estimator. Such a result is stronger than Proposition 3.1. We state it in the next theorem by providing a so-called *oracle inequality* (Eq. (11)).

Theorem 3.2. *Assume that $n \geq 9$. Let $\widehat{\mathcal{G}}$ be any (data-driven) family of graphs with maximal degree $D_{\widehat{\mathcal{G}}} = \max\{\deg(G),\ G \in \widehat{\mathcal{G}}\}$ fulfilling (8). Then, the* MSEP *of the estimator $\widetilde{\theta}$ is upper bounded by*

$$\mathrm{MSEP}(\widetilde{\theta}) \leq L_{K,\gamma} \log(p) \left(\mathbb{E}\left[\inf_{G \in \widehat{\mathcal{G}}} \left(\mathrm{MSEP}(\widehat{\theta}_G) \right) \right] \vee \frac{\mathrm{MSEP}(I)}{n} \right) + R_n \,. \tag{11}$$

where $L_{K,\gamma}$ is a positive constant depending on K and γ only and the residual term $R_n = R_n(\Sigma, \gamma)$ (made explicit in the proof) is of order $n^3 tr(\Sigma) e^{-n(\sqrt{\gamma}-\gamma)^2/4}$.

If we forget the term $n^{-1}\mathrm{MSEP}(I)$ in (11), Theorem 3.2 states that under Condition (8) the MSEP of $\tilde{\theta}$ nearly achieves, up to a $\log(p)$ factor, the average minimal MSEP of the family of estimators $\{\widehat{\theta}_G,\ G \in \widehat{\mathcal{G}}\}$. Hence, $\widetilde{\theta}$ performs almost as well as the oracle up to a $\log p$ factor. This logarithmic factor is proved to be unavoidable from a minimax point of view (see [28] Sect. 4.2).

Let us compare the risk bound (11) with Theorem 1 of Giraud [15]. This theorem claims that the procedure nearly selects the best graph among a *fixed* collection of graphs. In contrast, our collection of graphs $\widehat{\mathcal{G}}$ is not fixed a priori and depends on the data $\mathbf{X}$. Here, we prove that the graph $\widehat{G}$ is nearly the best (in terms of MSEP) among the random collection $\widehat{G}$. As a simple example, let us consider the procedure GGMselect with the Lasso-And family $\widehat{\mathcal{G}}_{LA}$. Theorem 3.2 tells us that the selected graph $\widehat{G}$ nearly achieves the smallest MSEP among the collection of Lasso-And graph estimators $\{\widehat{G}^{\lambda}_{\mathrm{and}}\}_{\lambda>0}$. In other words, GGMselect nearly selects the best tuning parameter of the Lasso-And procedure.

The condition (8) roughly states that we restrict ourselves to graphs whose maximal degree is smaller than $n/(2\log(p))$. For the related problem of random design regression, it is proved in [29] that theoretical limitations are occurring when the size of the support of the parameter is larger than $n/(2\log(p))$. In this so-called ultra-high dimensional setting, it is not possible to obtain an oracle bound of the form (11) and it is shown that recovering the support of the parameter is almost impossible. In short, estimating a graph whose maximal degree is larger than $n/(2\log(p))$ is nearly impossible.

3.2. Consistency of the selection procedure

The next theorem states, under mild assumptions, a consistency result for our selection procedure in a high-dimensional setting. In the spirit of the results of Meinshausen and Bühlmann [22], we consider the case where the number of variables p increase with the sample size n.

We make the following assumptions:

(H.1) $p_n \geq n$.

(H.2) $\deg(G_{\Sigma_n}) \leq \dfrac{n^s}{\log p_n} \wedge \dfrac{n}{\log^2 p_n}$ for some $s < 1$.

(H.3) $\min\limits_{a \neq b,\ b \in \mathrm{ne}_{G_{\Sigma_n}}(a)} \theta^2_{a,b} \min\limits_{a \neq b} \dfrac{\mathrm{Var}(X_a | X_{-a})}{\mathrm{Var}(X_b | X_{-b})} \geq n^{s'-1}$ for some $s' > s$.

Theorem 3.3. *Assume that the family $\widehat{\mathcal{G}}$ of candidate graphs contains the true graph with probability going to 1 and* **(H.1)**, **(H.2)**, **(H.3)** *are fulfilled. Then, the estimation procedure GGMselect with $K > \left[3 \vee \frac{2.5}{(1-s)}\right]$ and*

$$D_{\widehat{\mathcal{G}}} = \max\{\deg(G),\ G \in \widehat{\mathcal{G}}\} \quad \leq \quad \frac{n}{\log^2 p_n}$$

is consistent. More precisely, there exist some universal constant L and some integer $n_0 = n_0\,[K, s, s']$ not depending on the true graph G_{Σ_n} nor on the covariance Σ_n such that

$$\mathbb{P}\left[\widehat{G} = G_{\Sigma_n}\right] \geq 1 - Lp_n^{-1/2} - \mathbb{P}\left[G_{\Sigma_n} \notin \widehat{\mathcal{G}}\right], \qquad \textit{for any } n \geq n_0\,.$$

Let us discuss the assumptions of the theorem and their similarity with some of the hypotheses made in [22]. The Assumption **(H.2)** is met if p_n grows polynomially with respect to n and the degree of the true graph does not grow faster than n^κ with $\kappa < s$ (which corresponds to Assumptions 1 and 2 in [22]). We mention that **(H.2)** is not satisfied when p_n grows exponentially with n unless G_{Σ_n} is empty. It is actually impossible to consistently estimate a non-empty graph if p_n is of order $\exp(n)$, see [29].

The Assumption **(H.3)** ensures that the conditional variances as well as the non-zero terms $\theta_{a,b}$ are large enough so that the edges can be detected. To compare with [22], Assumption **(H.3)** is met as soon as Assumption 2 and 5 in [22] are satisfied. In addition, we underline that we make no assumption on the l^1-norm of the prediction coefficients or on the signs of $\theta_{a,b}$ (Assumptions 4 and 6 in [22]).

Finally, we do not claim that the condition $K > [2.5/(1-s) \vee 3]$ is minimal to obtain consistency. It seems from simulation experiments that smaller choices of K also provide good estimations.

4. Numerical study

It is essential to investigate the performance of statistical procedures on data. Since we do not know the actual underlying graph of conditional dependences on real data sets, we mainly opt for a numerical study with simulated data. Our aims in this study are to evaluate the feasibility of the GGMselect procedure and to compare its performances with those of recent graph-selection procedures.

Simulating the data. The matrix $\mathbf{X}$ is composed of n i.i.d. rows with Gaussian $\mathcal{N}_p(0, \Omega^{-1})$ distribution where the inverse covariance matrix Ω is constructed according to the following procedure. We set $\Omega = BB^T + D$, where B is a random sparse lower triangular matrix and D is a diagonal matrix with random entries of order 10^{-3}. The latter matrix D prevents Ω from having too small eigenvalues. To generate B we split $\{1, \ldots, p\}$ into three consecutive sets I_1, I_2, I_3 of approximately equal size, and choose two real numbers η_{int} and η_{ext} between 0 and 1. For any a, b such that $1 \leq a < b \leq p$, we set $B_{a,b} = 0$ with probability $1 - \eta_{\text{int}}$ if a and b are in the same set, and we set $B_{a,b} = 0$ with probability $1 - \eta_{\text{ext}}$ if a and b

belong to two different sets. Then, the lower diagonal values that have not been set to 0 are drawn according to a uniform law on $[-1, 1]$ and the diagonal values are drawn according to a uniform law on $[0, \varepsilon]$. Finally, we rescale Ω in order to have 1 on the diagonal of $\Sigma = \Omega^{-1}$. This matrix Σ defines a graph $G = G_\Sigma$ and a matrix θ defined as in Section 2.1. The sparsity of the graph is measured via a sparsity index noted I_s, defined as the average number of edges per nodes in the graph.

In our simulation study we set $\eta = \eta_{\text{int}} = 5\eta_{\text{ext}}$, and $\varepsilon = 0.1$. We evaluate the value of η corresponding to a desired value of the sparsity index I_s by simulation. I_s equals the desired value. Choosing I_s small, we get sparse graphs whose edges distribution is not uniform, see Figure 1.

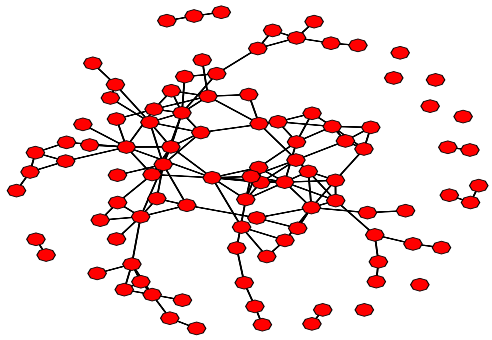

FIGURE 1. *One simulated graph G with $p = 30$ and $I_s = 3$. The degree $\deg(G)$ of the graph equals 8.*

GGMselect: choice of graphs families. Our procedure is applied for the families of graphs presented in Section 2.2. The methods are respectively denoted `CO1`, `LA`, `EW` and `QE`.

The family $\widehat{\mathcal{G}}_{\text{EW}}$ is based on the calculation of exponential weight estimators $\widehat{\theta}^{\text{EW}}$. This calculation depends on parameters, denoted $\alpha, \beta, \sigma, \tau$ in [9], that defined the aggregation procedure, and on parameters, denoted h and T in [9], used in the Langevin Monte-Carlo algorithm (see Section 7 for details). We chose these parameters as follows. The matrix $\mathbf{X}$ being scaled such that the norm of each column equals 1, we took $\sigma = 1/\sqrt{n}$, and we set $\alpha = 0$, $\beta = 2/n$, $\tau = 1/\sqrt{n(p-1)}$ and $h = 10^{-3}$, $T = 200$. Using these parameters values we did not encountered convergence problems in our simulation study.

As it was already mentioned in Section 2.2, the size of the family $\widehat{\mathcal{G}}_{\text{QE}}$ may be very large leading to memory size problems in the computational process. In that case, as soon as a memory size problem is encountered, the research between $\widehat{G}_{K,\text{and}}$ and $\widehat{G}_{K,\text{or}}$ is stopped and prolonged by a stepwise procedure.

Our procedure depends on two parameters: K occurring in the penalty function

(see Equation 3) and D the maximum degree of the graph. We choose $K = 2.5$ in all simulation experiments. In practice, we want to choose D as large as possible. From theoretical results in Section 3 and in [29], we know that we can take D as large as $\lfloor n/(2\log(p))\rfloor$, and that it is nearly impossible to perform estimation of a graph when the maximal degree is larger than $n/(2\log(p))$. We then set $D = \lfloor n/(2\log(p))\rfloor$ except for QE whose algorithmic complexity increases exponentially with D.

All these methods are implemented in R-2.7.2 in the package `GGMselect`.

4.1. CPU times

We assess the practical feasibility of the methods we propose from the point of view of the memory size and computer time. To this aim, we simulate graphs with $p = 30, 100, 200, 300, 500$ nodes, sparsity $I_s = 3$ and $n = 50$. The simulation were run on a Bi-Pro Xeon quad core 2.66 GHz with 24 Go RAM. The computer time being strongly dependent on the simulated graph we calculate the mean of computer times over $N_G = 100$ simulated graphs. For each of these graphs, one matrix $\mathbf{X}$ is simulated. The results are given in Table 1. The maximum degree D of the estimated graph was set to $\lfloor n/2\log(p)\rfloor$, except for the `QE` method where $D = 3$ and 5. The maximum allowed memory size is exceeded for the `QE` method when $D = 5$ and $p \geq 100$, and when $D = 3$ for $p \geq 300$. The `LA` and `CO1` methods are running very fast. The computing time for the `EW` method increases quickly with p: in this simulation study, it is roughly proportional to $\exp\left(\sqrt{p}/2\right)$, see Figure 2. This order of magnitude is obviously dependent on the choice of the parameters occurring in the Langevin Monte-Carlo algorithm for calculating $\widehat{\theta}^{EW}$.

	$D = \lfloor n/(2\log(p))\rfloor$			$D = 3$		$D = 5$	
p	`EW`	`LA`	`CO1`	`QE`		`QE`	
30	7.1	0.46	0.04	16	$[1.9, 1366]$	146	$[125, 975]$
100	111	3.11	0.13	1956	$[240, 5628]$	>ams	
200	853	8.0	0.68	4240	$[4008, 5178]$	>ams	
300	4277	15.5	2.27	>ams		>ams	
500	158550	43	9.7	>ams		>ams	

TABLE 1

Means and ranges (in square brackets) of computing times in seconds calculated over $N_G = 100$ simulated graphs. For `EW`, `LA` *and* `CO1` *there is nearly no variability in the computing times.* >ams *means that the maximum allowed memory size was exceeded.*

4.2. Methods comparison

We compare our methods with the following ones:

- the 0-1 conditional independence approach proposed in [30], with the decision rule based on the adjusted p-values following the Benjamini-Hochberg procedure taking $\alpha = 5\%$.
- the lasso approach, with the two variants `and` and `or` proposed in [22], taking $\alpha = 5\%$.

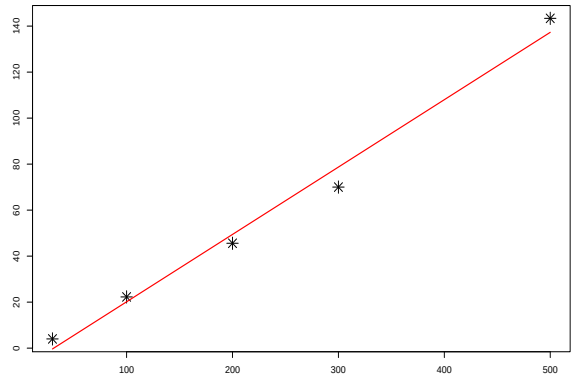

FIGURE 2. *Graphic of* $\log^2(CPU\ time)$ *versus* p *for the* `EW` *method.*

- the adaptive glasso method proposed in [13]. It works in two steps. First, the matrix Ω is estimated using the glasso method. Then the glasso procedure is run again using weights in the penalty that depend on the previous estimate of Ω, see Equation (2.5) in [13]. At each step the regularization parameter is calculated by K-fold cross-validation.

These methods will be denoted as `WB`, `MB.and`, `MB.or` and `Aglasso`. They were implemented in R-2.7.2 using the packages `lars` for the `MB` methods and the package `glasso` for the last one.

Assessing the performances of the methods. We assess the performances of the investigated methods on the basis of $N_G \times N_X$ runs where N_G is the number of simulated graphs and N_X the number of matrices $\mathbf{X}$ simulated for each of these graphs. We compare each simulated graph G with the estimated graphs $\widehat{G}$ by counting edges that are correctly identified as present or absent, and those that are wrongly identified. We thus estimate the false discovery rate (or FDR) defined as the expected proportion of wrongly detected edges among edges detected as present, and the power defined as the expected proportion of rightly detected edges among edges present in the graph.

The statistical procedures designed to select graphs have one or several parameters that must be tuned. The quality of the final estimation is then affected as well by the intrinsic ability of the procedure to select an accurate graph, as by the parameter tuning. First, we investigate the first issue by varying the values of the tuning parameters and plotting *power versus FDR curves*. We choose $p = 100$, $n = 50$ and $I_s = 3$. Then, taking the point of view of a typical user, we compare the different procedures with the tuning parameter recommended in the literature. We investigate the effect of n by choosing $n = 30, 50, 100, 150$, keeping $p = 100$. We also evaluate the effect of graph sparsity taking $I_s = 1, 2, 3, 4, 5$, $p = 30$ to keep the computer time under reasonable values, and $n = 30$. Finally, we compare our

criterion defined by Equations (2) and (3) to a BIC-type criterion which selects a graph by minimizing with respect to $G \in \widehat{\mathcal{G}}$,

$$\mathrm{Crit}_{\mathrm{BIC}}(G) = \sum_{a=1}^{p} \exp\left(\log\left\{\|\mathbf{X}_a - \mathbf{X}[\widehat{\theta}_G]_a\|_n^2\right\} + d_a \frac{\log(p)}{n}\right) .$$

We base this last simulation study on empty graphs with $p = 1000$ and $n = 100$, in order to evaluate in practice, the tendancy of BIC to overfit in a high dimensional setting.

4.2.1. Power versus FDR curves when $p = 100$

The number of nodes p and the number of observations n being fixed to $p = 100$, $n = 50$, for each of the $N_G = 20$ simulated graphs, we estimated the FDR, the power and the MSEP on the basis of $N_X = 20$ simulations. These calculations are done for different values of the tuning parameter. The means over the N_G graphs are shown at Figure 3. The standard errors of the means over the N_G graphs are smaller than 0.0057 for the FDR, and 0.018 for the power.

Choice of the family of candidate graphs in our procedure. The `QE` method presents good performances: the FDR stays small and the power is high. Though it was performed with $D = 3$, while `EW`, `LA` and `CO1` were performed with $D = 5$, it works the best. The `EW` method is more powerful than `LA` and `CO1` if one accepts a FDR greater than 2.5%.

Comparison with the other methods. The procedures `LA` and `CO1` behave similarly to `WB` method. The `MB.or` method presents higher values of the power when the FDR is larger than 5%. The `MB.and` keeps down the FDR but lacks power. The `Aglasso` method behaves completely in a different way: the curve stays under the others as long as the FDR is smaller than 20%. When the regularization parameter is chosen by 5-fold cross-validation, the power equals 59% at the price of a very large FDR equal to 90% (not shown). In the following we do not consider anymore the adaptive glasso method, and focus on methods that have a good control of the FDR.

Results when p is very large face to n. Keeping $n = 50$, and taking $p = 500$, we estimated the FDR and the power for all methods except the `EW` method for which the computing time is too large for carrying out a simulation study. The results are given at Figure 4. The method `QE` was performed with $D = 2$, while the `LA` and `CO1` were performed with $D = 5$. As expected, for all methods, the power is lower for $p = 500$ than for $p = 100$. The between procedures comparison stay the same.

4.2.2. Effect of the number of observations n

Keeping $p = 100$ and $I_s = 3$, the variations of the FDR and power values versus the number of observations, are shown in Figure 5. The `QE` method is applied with

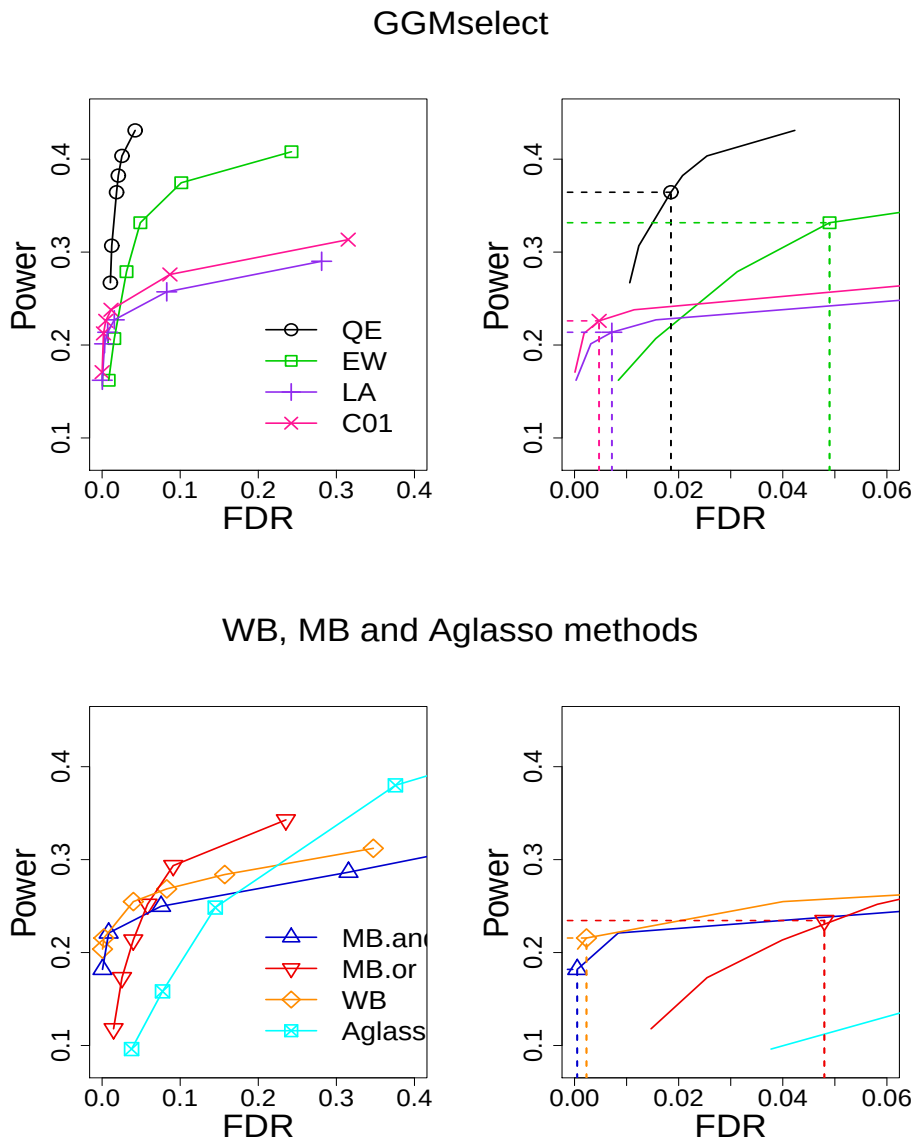

FIGURE 3. *Graphics of power versus FDR for the case* $p = 100$, $n = 50$ *and* $I_s = 3$. *The marks on the graphics on the left correspond to different values of the tuning parameter. The curves for small FDR values are magnified on the graphics on the right. The FDR and power values corresponding to the tuning parameter recommended in the literature are superimposed on the curves (dashed lines) :* $K = 2.5$ *for* `GGMselect`, $\alpha = 5\%$ *for* `WB` *and* `MB` *methods. For* `Aglasso`, *with* λ *chosen by 5-fold cross-validation, the FDR equals 0.90 and the power equals 0.59 (not shown).*

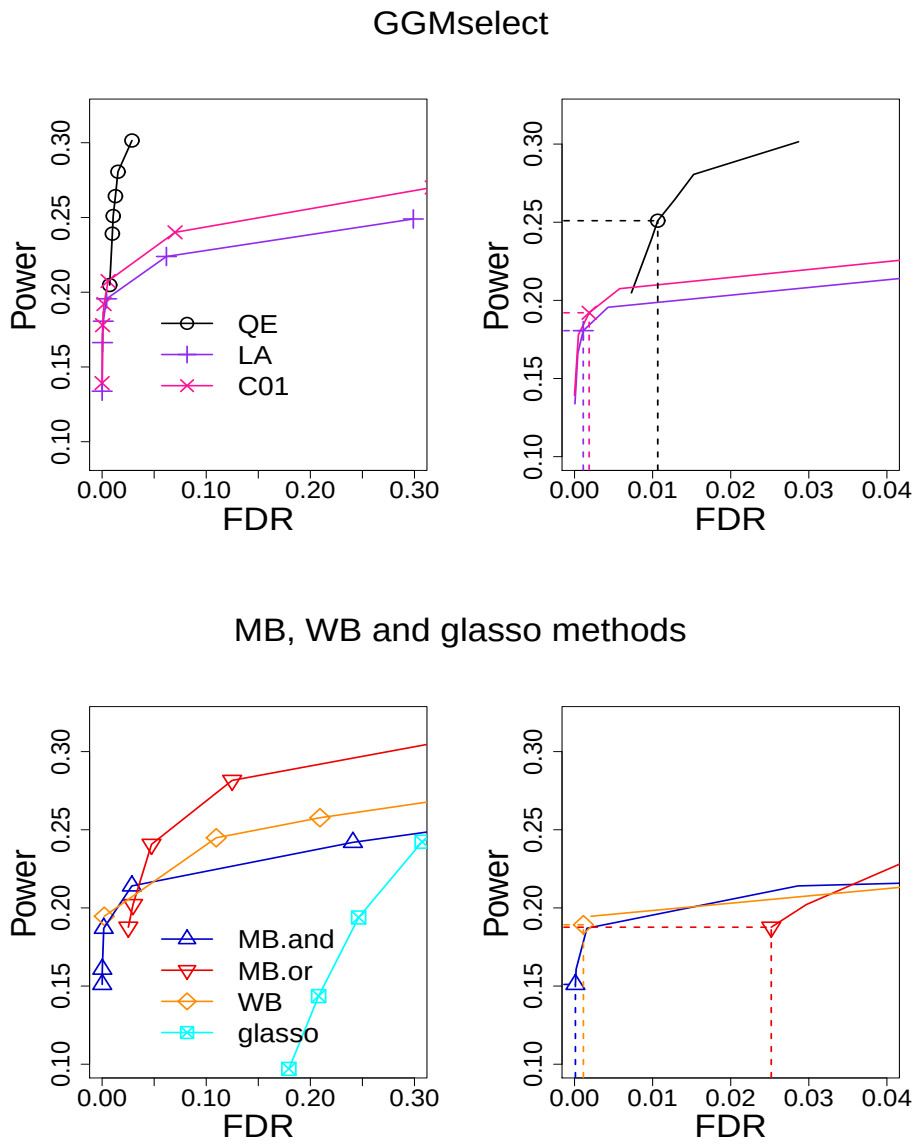

FIGURE 4. *Graphics of power versus FDR for the case* $p = 500$, $n = 50$ *and* $I_s = 3$. *The marks on the graphics on the left correspond to different values of the tuning parameter. The curves for small FDR values are magnified on the graphics on the right. The FDR and power values corresponding to the tuning parameter recommended in the literature are superimposed on the curves (dashed lines) :* $K = 2.5$ *for* `GGMselect`, $\alpha = 5\%$ *for* `WB` *and* `MB` *methods.*

$D = 3$ while `EW`, `LA` and `C01` are applied with $D = 5$. For all methods the power increases with n while the FDR decreases for `EW` and increases for `MB.or`, `LA` and `C01`. `QE` and `EW` are the most powerful. When n is small, the `QE` method stays more powerful than `EW` in spite of a smaller D.

4.2.3. Effect of graph sparsity

We have seen that when p is large, the GGMselect procedures using the graphs families `QE` and `EW` are powerful and have a good control of the FDR. Nevertheless, the simulated graphs were sparse, $I_s = 3$, and it may be worthwhile testing how the methods perform when the graph sparsity varies. Because the performances depend strongly on the simulated graph, the FDR and power are estimated on the basis of a large number of simulations: the number of simulated graphs N_G equals 50 and the number of simulated matrices $\mathbf{X}$ for each graph, N_X equals 50. In order to keep reasonable computing times, we choose $p = 30$. The results are shown in Figure 6. The standard errors of the means over the N_G graphs are smaller than 0.0055 for the FDR, and 0.025 for the power.

For all methods the power decreases when I_s increases. The FDR values are slightly increasing with I_s for the `EW` and `MB.or` methods. The superiority of `QE` over the others is clear. `EW` is more powerful then `LA`, `C01`, `MB` and `WB` methods but its FDR is greater.

4.2.4. GGMselect : mixing the graphs families

Our procedure allows to mix several graphs families. It may happen that some graphs, or type of graphs, are known to be good candidates for modelling the observed data set. In that case, they can be considered in the procedure, and thus compete with $\widehat{\mathcal{G}}_{\mathrm{EW}}$ or $\widehat{\mathcal{G}}_{\mathrm{QE}}$. This can be done with the function `selectMyFam` of the package `GGMselect`.

Considering the results of our simulation study, we could ask if mixing $\widehat{\mathcal{G}}_{\mathrm{LA}}$ or $\widehat{\mathcal{G}}_{\mathrm{C01}}$ with $\widehat{\mathcal{G}}_{\mathrm{EW}}$ would not give a better control of the FDR than `EW` while keeping high values of the power. To answer this question we carried out simulation studies taking $\widehat{\mathcal{G}}_{\mathtt{mix}} = \widehat{\mathcal{G}}_{\mathrm{C01}} \cup \widehat{\mathcal{G}}_{\mathrm{LA}} \cup \widehat{\mathcal{G}}_{\mathrm{EW}}$ as the family of graphs. In all considered cases for p, n, I_s, the FDR and power values based on $\widehat{\mathcal{G}}_{\mathtt{mix}}$ are similar to those based on $\widehat{\mathcal{G}}_{\mathrm{EW}}$. This result can be explained by studying the behavior of the MSEP estimated by averaging the quantities $\|\Sigma^{1/2}(\widehat{\theta}_{\widehat{G}} - \theta)\|^2$ over the $N_G \times N_X$ runs. The results are given at Figure 7. One can see that the smallest values of the MSEP are obtained for `QE`, then `EW`. Moreover, the MSEP decreases when the power increases, while it does not show any particular tendency when the FDR varies. Considering these tendencies together with the fact that our procedure aims at minimizing the MSEP, we can understand why we do not improve the performances of `EW` by considering $\widehat{\mathcal{G}}_{\mathtt{mix}}$.

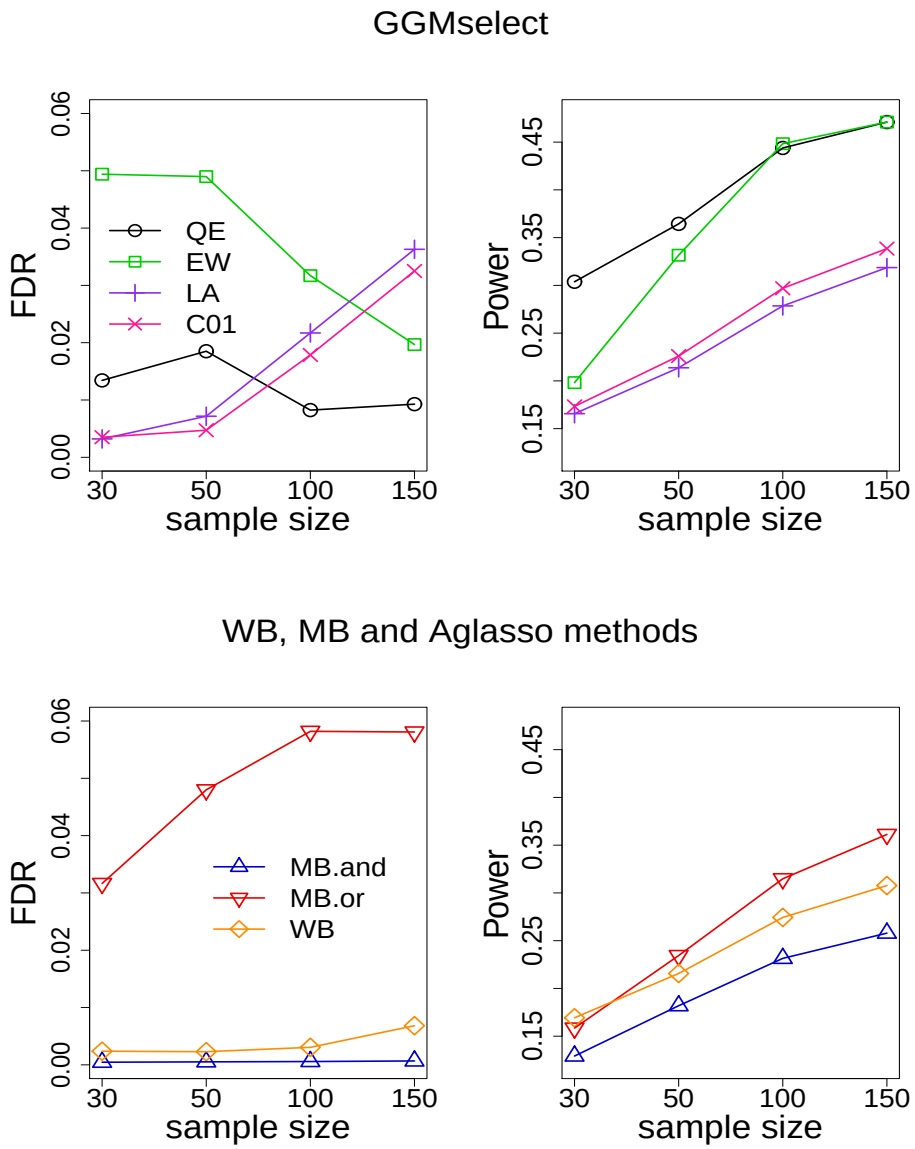

FIGURE 5. *FDR and power estimated values as a function of n for $p = 100$ and $I_s = 3$. The results are calculated on the basis of $N_G = 20$ simulated graphs and $N_X = 20$ runs of matrices $\mathbf{X}$ for each simulated graph. Our procedures were carried out with $K = 2.5$. The value of D was equal to 3 for the* `QE` *method and 5 for the others. For the procedures* `MB.or`, `MB.and` *and* `WB` *the tuning parameter α was taken equal to 5%.*

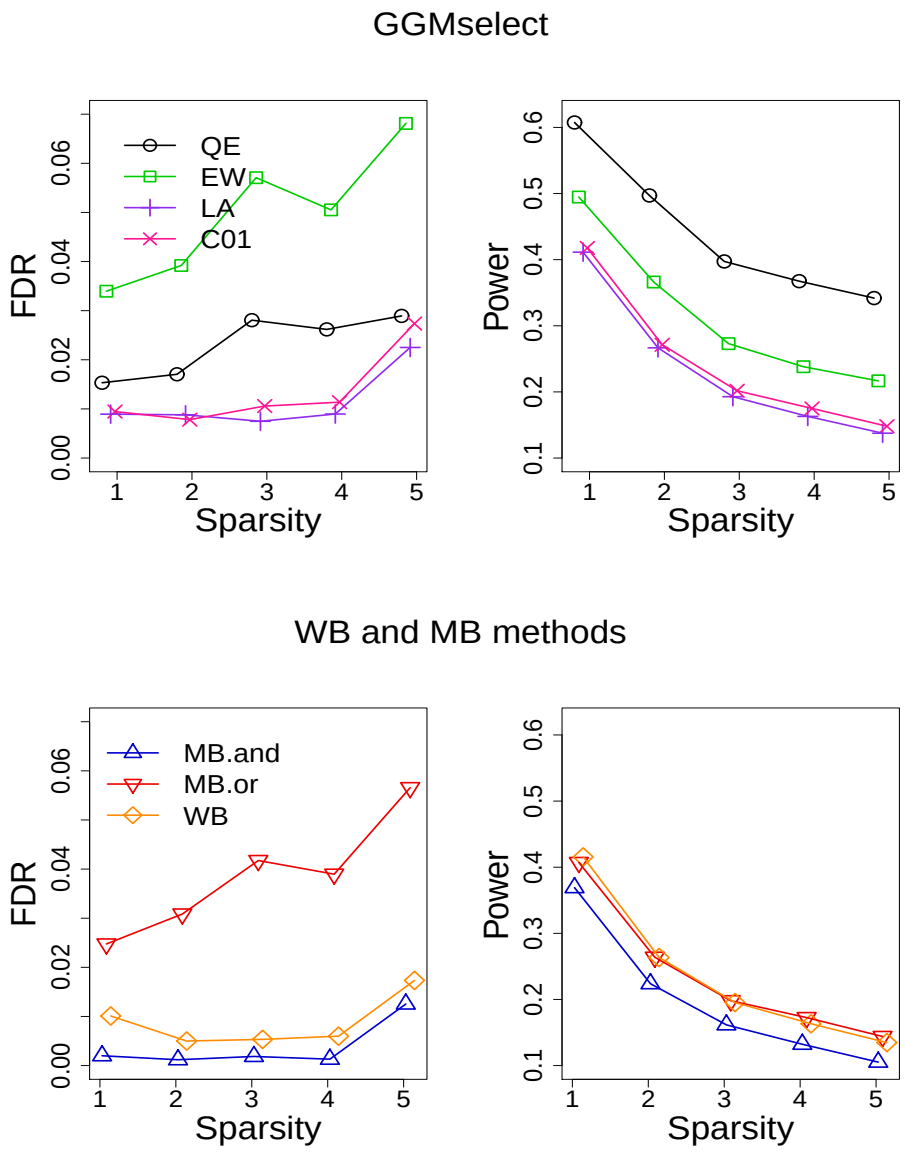

FIGURE 6. *Graphs of FDR and power estimated values versus the graph sparsity I_s, for $p = 30$ and $n = 30$. The results are calculated on the basis of $N_G = 50$ simulated graphs and $N_X = 50$ runs of matrices $\mathbf{X}$ for each simulated graph. Our procedures were carried out with $K = 2.5$ and $D = 5$. For the procedures* `MB.or`, `MB.and` *and* `WB` *the tuning parameter α was taken equal to 5%.*

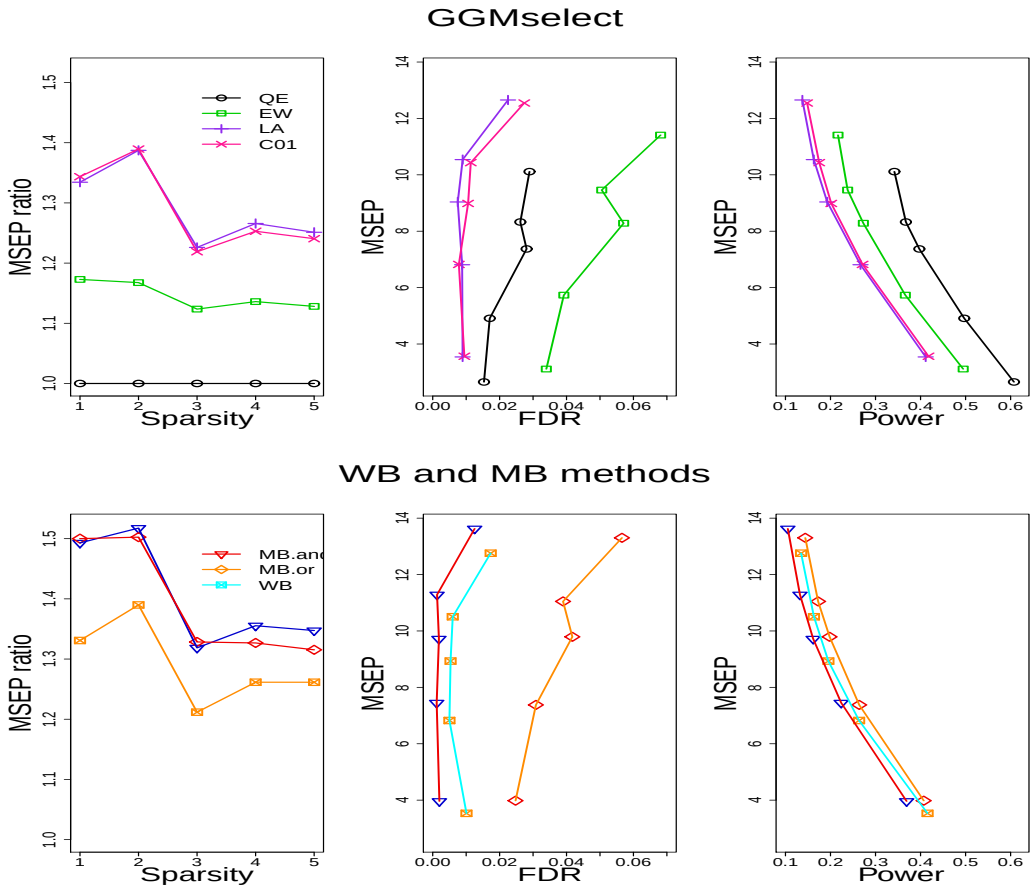

FIGURE 7. *Values of the MSEP for the simulation results given at Figure 6. The first graphic on the left presents the ratio of the MSEP over the MSEP of the* `QE` *method. The two others present the MSEP versus the FDR and the power.*

4.2.5. Comparison with the BIC criteria

As it was shown in the fixed design regression model [3], the BIC criterion overfits in a high-dimensional setting. To compare BIC with our procedure for estimating an empty graph, we simulate N_X matrices $\mathbf{X}$ composed of $n = 100$ i.i.d. rows distributed as $\mathcal{N}_p(0, I_p)$, with $p = 1000$. We consider the collection of graphs given by the procedure LA, with $D = 7$, and choose among this collection using our criterion, and the BIC criterion. The mean of the number of false positive, calculated on the basis of 100 simulations, equals 0 for our procedure, and equals 1077 when applying the BIC procedure. This confirms that BIC should not be used for such problems.

4.3. Summary

We recommend to use the `QE` method if the calculation of $\widehat{G}_{K,\mathrm{and}}$ and $\widehat{G}_{K,\mathrm{or}}$ is possible. Next, working out the family $\widehat{\mathcal{G}}_{\mathrm{QE}}$ can always be done using some suitable algorithms if necessary (as a stepwise procedure for example). When p is large, `QE` can be used for small values of D ($D = 3$ or even $D = 2$). It may perform better than all the others when n is small. The procedure based on $\widehat{\mathcal{G}}_{\mathrm{EW}}$ can be used for large p: the gain in power over `LA`, `CO1`, `MB` and `WB` methods is significant, but the FDR is slightly greater. The `LA` and `CO1` methods are running very quickly, keep the FDR under control and are slightly more powerful than `WB` and `MB.and`.

5. Breast cancer data

We test our procedure on a gene expression data set provided in Hess et al. [16]. The data set concerns 133 patients with breast cancer treated with chemotherapy. The patient response to chemotherapy can be classified into two groups according to a pathologic complete response (PCR) or residual disease (NotPCR). Natowicz et al. [23] selected 26 genes having a high predictive value for this response. We propose to estimate possible regulation dependencies between these 26 genes, for each group of patients : 34 patients achieved PCR, while 99 did not.

This data set was already considered by Ambroise et al. [1] who proposed a method to infer a Gaussian Graphical Model taking into account some hidden structure on the nodes. They simultaneously infer the nodes groups and the graph using an l^1 penalized likelihood criterion. Their method is performed in an iterative EM-like algorithm, names SIMoNe, available in an R-package [7].

We apply our procedure for choosing among the graphs coming from the families `QE`, `LA`, `CO1`, `EW` and from the family of graphs proposed by the SIMoNe algorithm.

We only present results for the group of patients not achieving PCR. The chosen graph presents 14 edges. The minimum value of the criteria equals 686.64 and is achieved for the `QE` family. Let us assess the stability of the results between the different methods, and the stability when the constant K in our procedure is varying.

Stability between the different methods. Let us have a look at graphs that minimize the criteria for each family considered. The results are given at

Figure 8. Firstly, let us note that for this data set, the SIMoNe algorithm gives results similar to the `LA` method. Secondly we remark that some characteristics of the *best* graph are shared by the others, as for example the path between KIA1467, GAMT, E2F3, MELK and RRM2. This allows to be confident in that motif. All methods allocate edges between genes ZNF552, FLJ10916, JMJD2B, BECNI, PDGFRA, but the motifs connecting these genes differ between methods. This suggests that these genes are probably linked, but the estimation of the motif is not completely secure. If the motif is an hub centered in JMJD2B, as it is shown by the `QE` method, then instability in estimating this motif is not surprising: it is more difficult to estimate the neighbours of highly connected nodes.

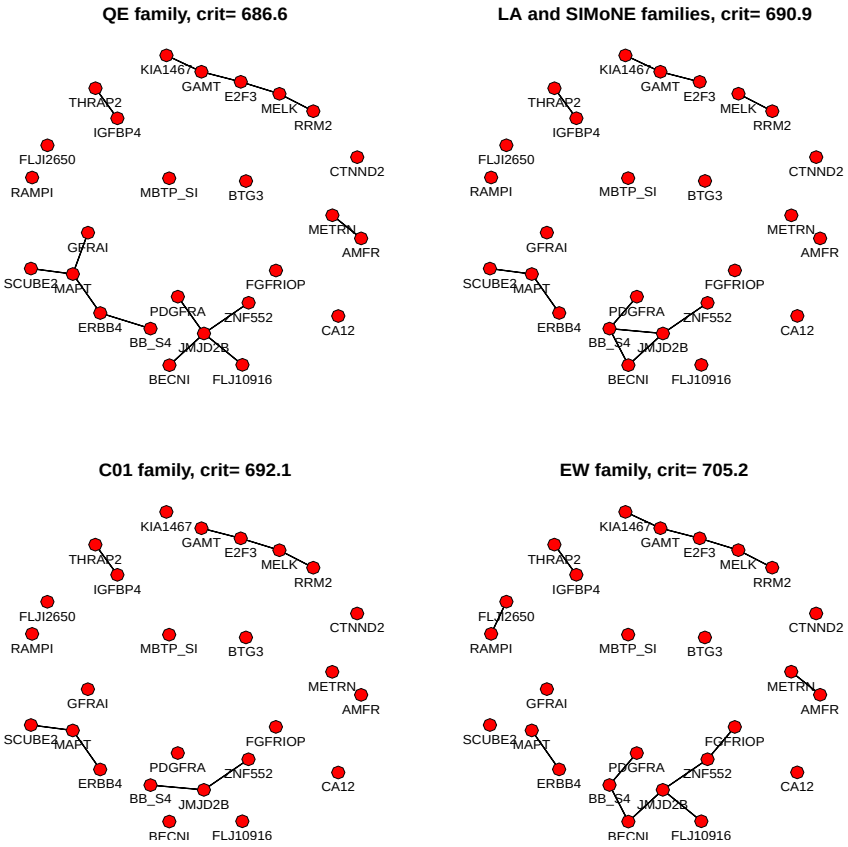

FIGURE 8. *For each considered family, criterion (2) value and estimated graph for the group of patients with residual disease. The graph minimizing the criterion (2) is given by the* `QE` *method.*

Stability when the constant K varies. The estimated graphs based on the `LA` family when K varies from 1.5 to 3 are presented at Figure 9. Increasing K to 3 leads to delete two edges. The difference between $K = 2$ and $K = 2.5$ is more important: some of the edges that were detected by the `QE` method with $K = 2.5$

are detected by the LA method with $K = 2$. There is no difference between the estimated graphs using $K = 1.5$ or $K = 2$. This suggests that all potentially detectable edges with the LA method are detected with $K = 2$.

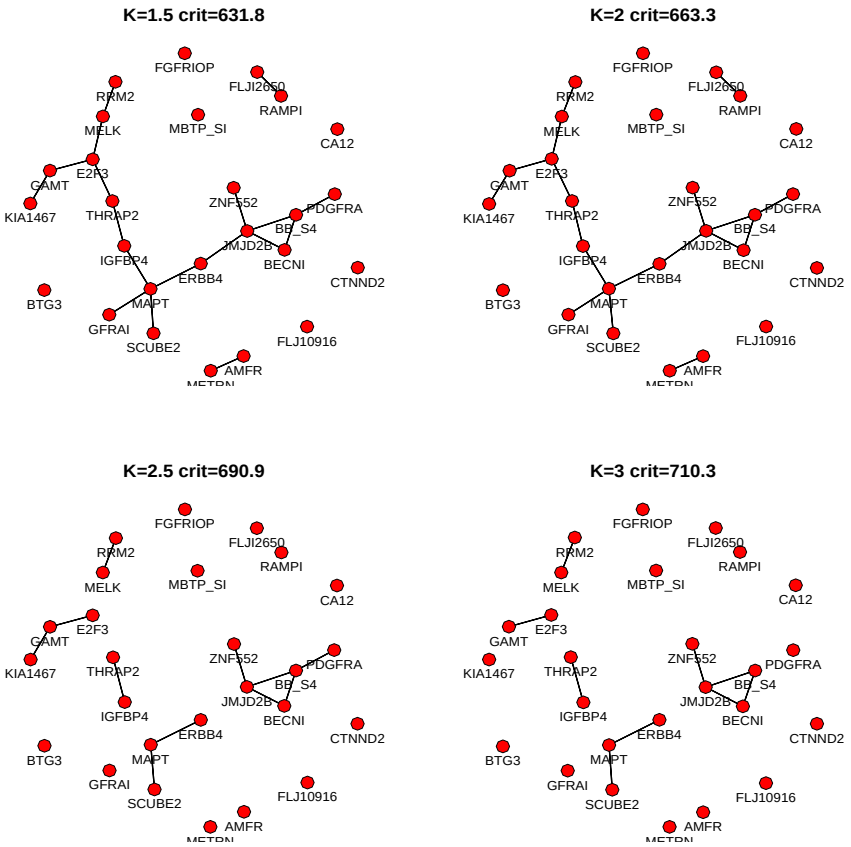

FIGURE 9. *For the family* LA*, criteria value and estimated graph when the constant K varies.*

6. Proofs

In the sequel, L, L_1, L_2,... denote universal constants that may vary from line to line. The notation $L(.)$ specifies the dependency on some quantities.

6.1. Proof of Theorem 3.2

We write $\mathcal{G}_D$ for the family of all the graph with nodes in Γ and degree less than D. We remind the reader that for any graph $G \in \mathcal{G}_D$ we have noted Θ_G the space of $p \times p$ matrices θ such that $\theta_{a,b}$ is non zero if and only if there is en edge between a and b in G. We also set $\bar{\Theta}_{D_{\max}} = \cup_{G \in \mathcal{G}_{D_{\max}}} \Theta_G$. norm $\|\cdot\|_{q \times p}$ on $q \times p$ matrices.

We set $\lambda = (1-\sqrt{\gamma})^2$ and introduce the event

$$\mathbb{B} = \left\{ \lambda \|\Sigma^{1/2} A\|_{p\times p} \leq \frac{1}{\sqrt{n}} \|\mathbf{X}A\|_{n\times p} \leq \lambda^{-1} \|\Sigma^{1/2} A\|_{p\times p}, \text{ for all } A \in \theta + \bar{\Theta}_{D_{\max}} \right\}.$$

On this event we can control the l^2-loss of $\tilde{\theta}$ by the empirical loss since

$$\|\Sigma^{1/2}(\tilde{\theta}-\theta)\|^2_{p\times p} \mathbf{1}_{\mathbb{B}} \leq \frac{\lambda^{-2}}{n} \|\mathbf{X}(\tilde{\theta}-\theta)\|^2_{n\times p} \mathbf{1}_{\mathbb{B}}\,. \tag{12}$$

Moreover, according to Lemma 1 in [15], we have $\mathbb{P}(\mathbb{B}^c) \leq 2e^{-n(\sqrt{\gamma}-\gamma)^2/2}$ when Condition (8) is met. To bound the risk of the procedure, we consider apart the events $\mathbb{B}$ and $\mathbb{B}^c$.

6.1.1. Bound on $\mathbb{E}\left[\|\Sigma^{1/2}(\tilde{\theta}-\theta)\|^2_{p\times p}\mathbf{1}_{\mathbb{B}}\right]$

We have $\mathbf{X} = \mathbf{X}\theta + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a $n \times p$ matrix distributed as follows: for each $a \in \Gamma$, the column $\boldsymbol{\epsilon}_a$ is independent of $\mathbf{X}_{-a}$ and is distributed according to the Gaussian law $\mathcal{N}(0, \sigma_a^2 I_n)$, with $\sigma_a^2 = 1/\Omega_{a,a}$. For any $G \in \mathcal{G}_D$, we write henceforth θ^G for the orthogonal projection of θ on Θ_G according to the Euclidean norm $\|\Sigma^{1/2}\cdot\|_{p\times p}$ on $\mathbb{R}^{p\times p}$. Similarly, we write $\bar{\theta}^G$ for the orthogonal projection of θ on Θ_G according to the (random) Euclidean norm $\|\mathbf{X}\cdot\|_{n\times p}$ on $\mathbb{R}^{p\times p}$. For any $G \in \mathcal{G}_D$, we write $d_a(G)$ for the degree of the node a in G and introduce the positive quantity

$$R(G) = \sum_{a=1}^{p} \left(1 + \frac{\mathrm{pen}(d_a(G))}{n-d_a(G)}\right) \left(\|\mathbf{X}(\theta_a - \bar{\theta}_a^G)\|^2 + 2|<\mathbf{X}\theta_a - \mathbf{X}\bar{\theta}_a^G, \boldsymbol{\epsilon}_a>|\right) + \sum_{a=1}^{p} \frac{\mathrm{pen}(d_a(G))}{n-d_a(G)} \|\boldsymbol{\epsilon}_a\|^2,$$

where $\|.\|$ and $<.,.>$ denote the canonical norm and scalar product on $\mathbb{R}^n$. Following the same lines as in the beginning of the proof of Theorem 2 in [3], we get for any G^* in $\widehat{\mathcal{G}}$

$$\frac{K-1}{K} \|\mathbf{X}(\tilde{\theta}-\theta)\|^2_{n\times p} \mathbf{1}_{\mathbb{B}} \leq R(G^*)\mathbf{1}_{\mathbb{B}} + \Delta(\widehat{G})\mathbf{1}_{\mathbb{B}} \tag{13}$$

with

$$\Delta(G) = \sum_{a=1}^{p} \sigma_a^2 \left(K U_{\mathrm{ne}_G(a)} - \frac{\mathrm{pen}(d_a(G))}{n-d_a(G)} V_{\mathrm{ne}_G(a)} \right)_+$$

where $U_{\mathrm{ne}_G(a)}$ and $V_{\mathrm{ne}_G(a)}$ are two independent χ^2 random variables with $d_a(G)+1$ and $n-d_a(G)-1$ degrees of freedom.

We note that under Condition (8) there exists some constant $c(\gamma)$ depending on γ only, such that

$$\mathrm{pen}(d) \leq c(\gamma) K (d+1) \log(p), \quad \text{for all } d \in \{0, \ldots, D_{\max}\},$$

see Proposition 4 in [3]. In particular, we have for any $G \in \mathcal{G}_D$

$$\frac{\mathrm{pen}(d_a(G))}{n - d_a(G)} \leq \frac{c(\gamma)K(D_{\max}+1)\log(p)}{n/2} \leq 4K\gamma c(\gamma) = L_{\gamma,K}.$$

Using this bound together with

$$|2 < \mathbf{X}\theta - \mathbf{X}\bar{\theta}_a^G, \boldsymbol{\epsilon}_a > | \leq \|\mathbf{X}(\theta_a - \bar{\theta}_a^G)\|^2 + \sigma_a^2 \xi_{a,G}^2,$$

where for any $G \in \mathcal{G}$ and $a \in \{1, \ldots, p\}$, the random variable

$$\xi_{a,G} =< \mathbf{X}(\theta_a - \bar{\theta}_a^G), \boldsymbol{\epsilon} > /(\sigma_a \|\mathbf{X}(\theta_a - \bar{\theta}_a^G)\|)$$

is standard Gaussian, we obtain

$$\begin{aligned} R(G) &\leq (1 + L_{\gamma,K}) \sum_{a=1}^{p} \left(2\|\mathbf{X}(\theta_a - \bar{\theta}_a^G)\|^2 + \sigma_a^2 \xi_{a,G}^2\right) + \frac{\mathrm{pen}(d_a(G))}{n - d_a(G)} \|\boldsymbol{\epsilon}_a\|^2 \\ &\leq 2(1 + L_{\gamma,K})\|\mathbf{X}(\theta - \bar{\theta}^G)\|_{n\times p}^2 + (4 + L_{\gamma,K}) \sum_{a=1}^{p} \mathrm{pen}(d_a(G))\sigma_a^2 + r(\mathcal{G}_D) \end{aligned}$$

where $r(\mathcal{G}_D)$ equals

$$\sum_{a=1}^{p} \sigma_a^2 \left((1 + L_{\gamma,K}) \sum_{G \in \mathcal{G}} \left[\xi_{a,G}^2 - \mathrm{pen}(d_a(G))\right]_+ + L_{\gamma,K} \left[\|\boldsymbol{\epsilon}_a\|^2/\sigma_a^2 - 3n/2\right]_+ \right).$$

Furthermore, we have $\|\mathbf{X}(\theta - \bar{\theta}^G)\|_{n\times p} \leq \|\mathbf{X}(\theta - \theta^G)\|_{n\times p}$ and on the event $\mathbb{B}$ we also have $\|\mathbf{X}(\theta - \theta^G)\|_{n\times p}^2 \leq n\lambda^{-2}\|\Sigma^{1/2}(\theta - \theta^G)\|_{p\times p}^2$ so that on $\mathbb{B}$

$$R(G) \leq L'_{\gamma,K} \left(n\lambda^{-2}\|\Sigma^{1/2}(\theta - \theta^G)\|_{p\times p}^2 + \sum_{a=1}^{p} \mathrm{pen}(d_a(G))\sigma_a^2 \right) + r(\mathcal{G}_D),$$

with $L'_{\gamma,K} = \max(2 + 2L_{\gamma,K}, 4 + L_{\gamma,K})$. Putting this bound together with (12) and(13), we obtain

$$\begin{aligned} \|\Sigma^{1/2}(\tilde{\theta} - \theta)\|_{p\times p}^2 \mathbf{1}_{\mathbb{B}} &\leq \frac{K}{n\lambda^2(K-1)} \left(\inf_{G^* \in \widehat{\mathcal{G}}} R(G^*) + \Delta(\widehat{G}) \right) \mathbf{1}_{\mathbb{B}} \\ &\leq L''_{\gamma,K} \inf_{G^* \in \widehat{\mathcal{G}}} \left(\|\Sigma^{1/2}(\theta - \theta^{G^*})\|_{p\times p}^2 + \sum_{a=1}^{p} \mathrm{pen}(d_a(G^*)) \frac{\sigma_a^2}{n} \right) \\ &\quad + L''_{\gamma,K} n^{-1} \left(r(\mathcal{G}_D) + \Delta(\widehat{G}) \right). \end{aligned}$$

We note that

$$n^{-1}\mathbb{E}(r(\mathcal{G}_D)) \leq \sum_{a=1}^{p} \frac{\sigma_a^2}{n}(1 + L_{\gamma,K})(3 + \log(p))$$

and we get from the proof of Theorem 1 in [15] that

$$n^{-1}\mathbb{E}(\Delta(\widehat{G})) \leq n^{-1}\mathbb{E}\left(\sup_{G \in \mathcal{G}_D} \Delta(G) \right) \leq K \sum_{a=1}^{p} \frac{\sigma_a^2}{n}(1 + \log(p)).$$

Since $\mathrm{pen}(d) \leq c(\gamma)K(d+1)\log(p)$, the latter bounds enforce the existence of constants $L_{\gamma,K}$ and $L'_{\gamma,K}$ depending on γ and K only, such that

$$\begin{aligned}
&\mathbb{E}\left[\|\Sigma^{1/2}(\tilde{\theta}-\theta)\|^2_{p\times p}\mathbf{1}_{\mathbb{B}}\right] \\
&\quad\leq\ L_{\gamma,K}\,\mathbb{E}\left[\inf_{G^*\in\widehat{\mathcal{G}}}\left(\|\Sigma^{1/2}(\theta-\theta^{G^*})\|^2_{p\times p}+\sum_{a=1}^{p}\left(\log(p)\vee\mathrm{pen}[d_a(G^*)]\right)\frac{\sigma_a^2}{n}\right)\right] \\
&\quad\leq\ L'_{\gamma,K}\,\log(p)\left(\mathbb{E}\left[\inf_{G^*\in\widehat{\mathcal{G}}}\mathrm{MSEP}(\widehat{\theta}_{G^*})\right]\vee\sum_{a=1}^{p}\frac{\sigma_a^2}{n}\right).
\end{aligned}$$

Finally, we note that $\sum_{a=1}^{p}\sigma_a^2/n=\mathrm{MSEP}(I)$.

6.1.2. *Bound on* $\mathbb{E}\left[\|\Sigma^{1/2}(\tilde{\theta}-\theta)\|^2_{p\times p}\mathbf{1}_{\mathbb{B}^c}\right]$

We now prove the bound $\mathbb{E}\left[\|\Sigma^{1/2}(\tilde{\theta}-\theta)\|^2_{p\times p}\mathbf{1}_{\mathbb{B}^c}\right]\leq Ln^3\mathrm{tr}(\Sigma)\sqrt{\mathbb{P}(\mathbb{B}^c)}$. We have

$$\mathbb{E}\left[\|\Sigma^{1/2}(\tilde{\theta}-\theta)\|^2_{p\times p}\mathbf{1}_{\mathbb{B}^c}\right]=\sum_{a=1}^{p}\mathbb{E}\left[\|\Sigma^{1/2}(\tilde{\theta}_a-\theta_a)\|^2\mathbf{1}_{\mathbb{B}^c}\right]$$

and we will upper bound each of the p terms in this sum. Let a be any node in Γ. Given a graph G, the vector $[\widehat{\theta}_G]_a$ depends on G only through the neighborhood $\mathrm{ne}_G(a)$ of a in G. Henceforth, we write $\widehat{\theta}_{\mathrm{ne}_{\widehat{G}}(a)}$ for $\tilde{\theta}_a$ in order to emphasize this dependency. By definition $\widehat{\theta}_{\mathrm{ne}_{\widehat{G}}(a)}$ is the least-squares estimator of θ_a with support included in $\mathrm{ne}_{\widehat{G}}(a)$. Let us apply the same arguments as in the proof of Lemma 7.12 in [28]. By Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[\|\Sigma^{1/2}(\tilde{\theta}_a-\theta_a)\|^2\mathbf{1}_{\mathbb{B}^c}\right]\leq\sqrt{\mathbb{P}(\mathbb{B}^c)}\,\sqrt{\mathbb{E}\left[\|\Sigma^{1/2}(\widehat{\theta}_{\mathrm{ne}_{\widehat{G}}(a)}-\theta_a)\|^4\right]}. \tag{14}$$

Let $\mathcal{N}_D(a)$ be the set made of all the subsets of $\Gamma\setminus\{a\}$ whose size are smaller than $\gamma n/[2(1.1+\sqrt{\log(p)})^2]$. By Condition (8), it holds that the estimated neighborhood $\mathrm{ne}_{\widehat{G}}(a)$ belongs to $\mathcal{N}_D(a)$, so Hölder inequality gives

$$\begin{aligned}
\mathbb{E}\left[\|\Sigma^{1/2}(\widehat{\theta}_{\mathrm{ne}_{\widehat{G}}(a)}-\theta_a)\|^4\right]&=\sum_{\mathrm{ne}(a)\in\mathcal{N}_D(a)}\mathbb{E}\left[\mathbf{1}_{\mathrm{ne}_{\widehat{G}}(a)=\mathrm{ne}(a)}\|\Sigma^{1/2}(\widehat{\theta}_{\mathrm{ne}(a)}-\theta_a)\|^4\right] \\
&\leq\sum_{\mathrm{ne}(a)\in\mathcal{N}_D(a)}\mathbb{P}\left[\mathrm{ne}_{\widehat{G}}(a)=\mathrm{ne}(a)\right]^{1/u}\mathbb{E}\left[\|\Sigma^{1/2}(\widehat{\theta}_{\mathrm{ne}(a)}-\theta_a)\|^{4v}\right]^{1/v} \\
&\leq\sum_{\mathrm{ne}(a)\in\mathcal{N}_D(a)}\mathbb{P}\left[\mathrm{ne}_{\widehat{G}}(a)=\mathrm{ne}(a)\right]^{1/u}\sup_{\mathrm{ne}(a)\in\mathcal{N}_D(a)}\mathbb{E}\left[\|\Sigma^{1/2}(\widehat{\theta}_{\mathrm{ne}(a)}-\theta_a)\|^{4v}\right]^{1/v},
\end{aligned}$$

where $v = \lfloor \frac{n}{8} \rfloor$, and $u = \frac{v}{v-1}$ (we remind the reader that n is larger than 8). In particular, we have the crude bound

$$\sqrt{\mathbb{E}\left[\|\Sigma^{1/2}(\widehat{\theta}_{\mathrm{ne}_{\widehat{G}}(a)} - \theta_a)\|^4\right]} \leq [\mathrm{Card}(\mathcal{N}_D(a))]^{1/2v} \sup_{\mathrm{ne}(a)\in\mathcal{N}_D(a)} \mathbb{E}\left[\|\Sigma^{1/2}(\widehat{\theta}_{\mathrm{ne}(a)} - \theta_a)\|^{4v}\right]^{1/2v},$$

since the sum is maximum when every $\mathbb{P}[\mathrm{ne}(a) = \mathrm{ne}_{\widehat{G}}(a)]$ equals $[\mathrm{Card}(\mathcal{N}_D(a))]^{-1}$. We first bound the term $[\mathrm{Card}(\mathcal{N}_D(a))]^{1/2v}$. The size of the largest subset in $\mathcal{N}_D(a)$ is smaller than $n/(2\log(p))$, so the cardinality of $\mathcal{N}_D(a)$ is smaller than $p^{D_{\widehat{g}}}$. Since n is larger than 8, we get

$$[\mathrm{Card}(\mathcal{N}_D(a))]^{1/2v} \leq \exp\left[\frac{n}{4\lfloor n/8 \rfloor}\right] \leq L\ ,$$

which ensures the bound

$$\sqrt{\mathbb{E}\left[\|\Sigma^{1/2}(\widehat{\theta}_{\mathrm{ne}_{\widehat{G}}(a)} - \theta_a)\|^4\right]} \leq L \sup_{\mathrm{ne}(a)\in\mathcal{N}_D(a)} \mathbb{E}\left[\|\Sigma^{1/2}(\widehat{\theta}_{\mathrm{ne}(a)} - \theta_a)\|^{4v}\right]^{1/2v}. \quad (15)$$

To conclude, we need to upper bound this supremum. Given a subset $\mathrm{ne}(a)$ in $\mathcal{N}_D(a)$, we define $\theta_{\mathrm{ne}(a)}$ as the vector in $\mathbb{R}^p$ such that $\Sigma^{1/2}\theta_{\mathrm{ne}(a)}$ is the orthogonal projection of $\Sigma^{1/2}\theta_a$ onto the linear span $\left\{\Sigma^{1/2}\beta : \mathrm{supp}(\beta) \subset \mathrm{ne}(a)\right\}$. Pythagorean inequality gives

$$\|\Sigma^{1/2}(\widehat{\theta}_{\mathrm{ne}(a)} - \theta_a)\|^2 = \|\Sigma^{1/2}(\theta_{\mathrm{ne}(a)} - \theta_a)\|^2 + \|\Sigma^{1/2}(\widehat{\theta}_{\mathrm{ne}(a)} - \theta_{\mathrm{ne}(a)})\|^2$$

and we obtain from Minkowski's inequality that

$$\mathbb{E}\left[\|\Sigma^{1/2}(\widehat{\theta}_{\mathrm{ne}(a)} - \theta_a)\|^{4v}\right]^{1/(2v)} \leq \|\Sigma^{1/2}(\theta_{\mathrm{ne}(a)} - \theta_a)\|^2 + \mathbb{E}\left[\|\Sigma^{1/2}(\widehat{\theta}_{\mathrm{ne}(a)} - \theta_{\mathrm{ne}(a)})\|^{4v}\right]^{1/(2v)}.$$

The first term is smaller than $\mathrm{Var}(X_a)$. In order to bound the second term, we use the following lemma which rephrases Proposition 7.8 in [28].

Lemma 6.1. *For any neighborhood* $\mathrm{ne}(a)$ *and any* $r > 2$ *such that* $n - |\mathrm{ne}(a)| - 2r + 1 > 0$,

$$\mathbb{E}\left[\|\Sigma^{1/2}(\widehat{\theta}_{\mathrm{ne}(a)} - \theta_{\mathrm{ne}(a)})\|^{2r}\right]^{1/r} \leq Lr|\mathrm{ne}(a)|n\,Var(X_a)\,.$$

Since v is smaller than $n/8$ and since $|\mathrm{ne}(a)|$ is smaller than $n/2$, it follows that for any model $\mathrm{ne}(a) \in \mathcal{N}_D(a)$, $n - |\mathrm{ne}(a)| - 4v + 1$ is positive and

$$\mathbb{E}\left[\|\Sigma^{1/2}(\widehat{\theta}_{\mathrm{ne}(a)} - \theta_a)\|^{4v}\right]^{1/(2v)} \leq \mathrm{Var}(X_a)\left[1 + Ln^2v\right] \leq Ln^3\Sigma_{a,a}\,.$$

Gathering this last upper bound with (14) and (15), we get that

$$\mathbb{E}\left[\|\Sigma^{1/2}(\tilde{\theta} - \theta)\|^2_{p\times p}\mathbf{1}_{\mathbb{B}^c}\right] \leq Ln^3\mathrm{tr}(\Sigma)\sqrt{\mathbb{P}(\mathbb{B}^c)}.$$

6.1.3. Conclusion

Finally, putting together the bound on $\mathbb{E}[\|\Sigma^{1/2}(\tilde{\theta}-\theta)\|^2\mathbf{1}_{\mathbb{B}}]$, the bound on $\mathbb{E}[\|\Sigma^{1/2}(\tilde{\theta}-\theta)\|^2\mathbf{1}_{\mathbb{B}^c}]$, and the bound $\mathbb{P}(\mathbb{B}^c) \leq 2pe^{-n(\sqrt{\gamma}-\gamma)^2/2}$, we obtain

$$\mathrm{MSEP}(\widetilde{\theta}) \leq L_{K,\gamma}\log(p)\left(\mathbb{E}\left[\inf_{G\in\widehat{\mathcal{G}}}\left(\mathrm{MSEP}(\widehat{\theta}_G)\right)\right] \vee \frac{\mathrm{MSEP}(I)}{n}\right) + R_n\,,$$

with $R_n \leq Ln^3\mathrm{tr}(\Sigma)e^{-n(\sqrt{\gamma}-\gamma)^2/4}$.

6.2. Proof of Proposition 3.1

The result is proved analogously except that we replace the event $\mathbb{B}$ by

$$\mathbb{B}' = \mathbb{B} \cup \left\{G_\Sigma \in \widehat{\mathcal{G}}\right\}\,.$$

Hence, the residual term now satisfies

$$\begin{aligned} R_n &\leq Ln^3\mathrm{tr}(\Sigma)\sqrt{\mathbb{P}(\mathbb{B}^c)} \\ &\leq Ln^3\mathrm{tr}(\Sigma)\left[e^{-n(\sqrt{\gamma}-\gamma)^2/4} + \sqrt{\alpha}e^{-\frac{\beta}{2}n^\delta}\right]\,. \end{aligned}$$

6.3. Proof of Theorem 3.3

In this proof, the notations $o(1)$, $O(1)$ respectively refer to sequences that converge to 0 or stay bounded when n goes to infinity. These sequences may depend on K, s, s' but *do not* depend on G_n, on the covariance Σ, or a particular subset $S \subset \Gamma$. The technical lemmas are postponed to Section 6.4. In the sequel, we omit the dependency of p and Σ on n for the sake of clarity. First, observe that the result is trivial if $n/\log(p)^2 < 1$, because the assumptions imply that G_Σ is the empty graph whereas the family $\widehat{\mathcal{G}}$ contains at most the empty graph. In the sequel, we assume that $n/\log(p)^2 \geq 1$.

Let us set $D_{\max} = n/\log(p)^2$. We shall prove that for some $L > 0$,

$$\mathbb{P}\left(\mathrm{Crit}(G_\Sigma) = \inf_{G',\ \deg(G')\leq D_{\max}} \mathrm{Crit}(G')\right) \geq 1 - Lp^{-1/2}\,, \tag{16}$$

for n larger than $n_0(K,s,s')$. Since $\widehat{G}$ minimizes the criterion Crit(.) on the family $\widehat{\mathcal{G}}$, this will imply the result of the theorem.

In fact, we shall prove a slightly stronger result than (16). Let a be a node in Γ and let $\mathrm{ne}(a)$ be a subset of $\Gamma \setminus \{a\}$. As defined in Section 6.1.2, $\widehat{\theta}_{\mathrm{ne}(a)}$ is the least-squares estimator of θ_a whose support is included in $\mathrm{ne}(a)$.

$$\widehat{\theta}_{\mathrm{ne}(a)} = \arg\inf_{\theta'_a,\ \mathrm{supp}(\theta'_a)\subset\mathrm{ne}(a)} \|\mathbf{X}_a - \mathbf{X}\theta'_a\|_n^2\,.$$

If G is a graph such that the neighborhood $\mathrm{ne}_G(a)$ equals $\mathrm{ne}(a)$, then $\widehat{\theta}_{\mathrm{ne}(a)} = [\widehat{\theta}_G]_a$. We then define the partial criterion $\mathrm{Crit}(a, \mathrm{ne}(a))$ by

$$\mathrm{Crit}(a, \mathrm{ne}(a)) = \|\mathbf{X}_a - \mathbf{X}\widehat{\theta}_{\mathrm{ne}(a)}\|_n^2 \left(1 + \frac{\mathrm{pen}(|\mathrm{ne}(a)|)}{n - |\mathrm{ne}(a)|}\right) .$$

Observe that for any graph G, $\mathrm{Crit}(G) = \sum_{a=1}^{p} \mathrm{Crit}(a, \mathrm{ne}_G(a))$. We note $\widehat{\mathrm{ne}}(a)$ the set that minimizes the criterion $\mathrm{Crit}(a, .)$ among all subsets of size smaller than $D_{\max}$.

$$\widehat{\mathrm{ne}}(a) = \arg \inf_{\mathrm{ne}(a) \in \mathcal{N}_{D_{\max}}(a)} \mathrm{Crit}(a, \mathrm{ne}(a)) .$$

If for all nodes $a \in \Gamma$, the selected set $\widehat{\mathrm{ne}}(a)$ equals $\mathrm{ne}_{G_\Sigma}(a)$, then G_Σ minimizes the criterion $\mathrm{Crit}(.)$ over all graphs of degree smaller than $D_{\max}$. Consequently, the property (16) is satisfied if for any node $a \in \Gamma$, it holds that

$$\mathbb{P}\left[\widehat{\mathrm{ne}}(a) = \mathrm{ne}_{G_\Sigma}(a)\right] \geq 1 - 7p_n^{-3/2} , \tag{17}$$

for n larger than some $n_0[K, s, s']$.

Let us fix some node $a \in \Gamma$. We prove the lower bound (17) in two steps:

1. With high probability, the estimated neighborhood $\widehat{\mathrm{ne}}(a)$ does not strictly contain the true one $\mathrm{ne}_{G_\Sigma}(a)$.

$$\mathbb{P}\left[\widehat{\mathrm{ne}}(a) \supsetneq \mathrm{ne}_{G_\Sigma}(a)\right] \leq p_n^{-3/2} , \tag{18}$$

   for n larger than some $n_0[K, s, s']$.
2. With high probability, the estimated neighborhood $\widehat{\mathrm{ne}}(a)$ contains the true one $\mathrm{ne}_{G_\Sigma}(a)$.

$$\mathbb{P}\left[\widehat{\mathrm{ne}}(a) \not\supseteq \mathrm{ne}_{G_\Sigma}(a)\right] \leq 6p_n^{-3/2} , \tag{19}$$

   for n larger than some $n_0[K, s, s']$.

The remaining part of the proof is deserved to (18) and (19).

Let us recall some notations and let us introduce some other ones. The component X_a decomposes as

$$X_a = X\theta_a + \epsilon_a ,$$

where ϵ_a follows a centered normal distribution with variance $\Omega_{a,a}^{-1} = \mathrm{Var}(X_a|X_{-a})$. The variables ϵ_a are independent of X_{-a}. Given a set $S \subset \Gamma$, Π_S stands for the projection of $\mathbb{R}^n$ into the space generated by $(\mathbf{X}_a)_{a \in S}$, whereas $\Pi_S^\perp$ denotes the projection along the space generated by $(\mathbf{X}_a)_{a \in S}$. The notation $\langle ., . \rangle_n$ refers to the empirical inner product associated with the norm $\|.\|_n$. For any neighborhood $\mathrm{ne}(a) \subset \Gamma \setminus \{a\}$ such that $|\mathrm{ne}(a)| \leq D_{\max}$, let us define $\Delta(\mathrm{ne}(a), \mathrm{ne}_{G_\Sigma}(a))$ by

$$\Delta(\mathrm{ne}(a), \mathrm{ne}_{G_\Sigma}(a)) = \mathrm{Crit}(a, \mathrm{ne}(a)) - \mathrm{Crit}(a, \mathrm{ne}_{G_\Sigma}(a)) .$$

6.3.1. Bound on $\mathbb{P}\left(\widehat{\mathrm{ne}}(a) \supsetneq \mathrm{ne}_{G_\Sigma}(a)\right)$

We shall upper bound the probability that $\Delta(\mathrm{ne}(a), \mathrm{ne}_{G_\Sigma}(a))$ is negative for at least one of the neighborhoods $\mathrm{ne}(a) \in \mathcal{N}_{D_{\max}}(a)$ such that $\mathrm{ne}(a)$ strictly contains $\mathrm{ne}_{G_\Sigma}(a)$. For such a set $\mathrm{ne}(a)$, $\Delta(\mathrm{ne}(a), \mathrm{ne}_{G_\Sigma}(a))$ decomposes as (see e.g. Lemma 7.1 in [28]).

$$\begin{aligned}
&\Delta(\mathrm{ne}(a), \mathrm{ne}_{G_\Sigma}(a)) \\
&\quad = \|\Pi^{\perp}_{\mathrm{ne}(a)}\boldsymbol{\epsilon}_a\|_n^2 \left[1 + \frac{\mathrm{pen}(|\mathrm{ne}(a)|)}{n - |\mathrm{ne}(a)|}\right] - \|\Pi^{\perp}_{\mathrm{ne}_{G_\Sigma}(a)}\boldsymbol{\epsilon}_a\|_n^2 \left[1 + \frac{\mathrm{pen}(|\mathrm{ne}_{G_\Sigma}(a)|)}{n - |\mathrm{ne}_{G_\Sigma}(a)|}\right] \\
&\quad = -\|\Pi_{\mathrm{ne}_{G_\Sigma}(a)^{\perp} \cap \mathrm{ne}(a)}\boldsymbol{\epsilon}_a\|_n^2 \left[1 + \frac{\mathrm{pen}(|\mathrm{ne}_{G_\Sigma}(a)|)}{n - |\mathrm{ne}_{G_\Sigma}(a)|}\right] \\
&\qquad + \|\Pi^{\perp}_{\mathrm{ne}(a)}\boldsymbol{\epsilon}_a\|_n^2 \left[\frac{\mathrm{pen}(|\mathrm{ne}(a)|)}{n - |\mathrm{ne}(a)|} - \frac{\mathrm{pen}(|\mathrm{ne}_{G_\Sigma}(a)|)}{n - |\mathrm{ne}_{G_\Sigma}(a)|}\right] .
\end{aligned}$$

Hence, $\Delta(m, \mathrm{ne}_{G_\Sigma}(a)) > 0$ if

$$\begin{aligned}
&\frac{\|\Pi_{\mathrm{ne}_{G_\Sigma}(a)^{\perp} \cap \mathrm{ne}(a)}\boldsymbol{\epsilon}_a\|_n^2 / (|\mathrm{ne}(a) \setminus \mathrm{ne}_{G_\Sigma}(a)|)}{\|\Pi^{\perp}_{\mathrm{ne}(a)}\boldsymbol{\epsilon}_a\|_n^2 / (n - |\mathrm{ne}(a)|)} \\
&\qquad < \frac{\mathrm{pen}(|\mathrm{ne}(a)|) - \mathrm{pen}(|\mathrm{ne}_{G_\Sigma}(a)|)}{|\mathrm{ne}(a) \setminus \mathrm{ne}_{G_\Sigma}(a)|} \left[1 + \frac{\mathrm{pen}(|\mathrm{ne}_{G_\Sigma}(a)|)}{n - |\mathrm{ne}_{G_\Sigma}(a)|}\right]^{-1} . \qquad (20)
\end{aligned}$$

To conclude, it remains to prove that the bound (20) holds with high probability. Let us call A_1 the right expression of (20) and let us derive a lower bound of A_1. Afterwards, we shall upper bound with high probability the left expression of (20).

Upper bound of A_1. We first upper bound the penalty function.

Lemma 6.2. *Let $d_1 \geq d_2$ be two positive integers such that $d_1 \leq e^{-2}(p-1)$. We have*

$$\mathrm{pen}(d_1) - \mathrm{pen}(d_2) \geq 2K(d_1 - d_2)\log\left(\frac{p - d_1}{d_1}\right) . \qquad (21)$$

A proof of this lemma is provided in Section 6.4. By Proposition 4 in [3], the penalty $\mathrm{pen}(|\mathrm{ne}_{G_\Sigma}(a)|)$ satisfies

$$\mathrm{pen}\left(|\mathrm{ne}_{G_\Sigma}(a)|\right) \leq LK\frac{|\mathrm{ne}_{G_\Sigma}(a)|}{n}\log\left(\frac{p-1}{|\mathrm{ne}_{G_\Sigma}(a)|}\right) ,$$

where L is some numerical constant. This last term converges towards 0 as n goes to infinity since $|\mathrm{ne}_{G_\Sigma}(a)| \leq (n^s/\log(p)) \wedge (n/\log(p)^2)$ (Assumption 2). Gathering this upper bound with Lemma 6.2, we get

$$A_1 \geq 2K\frac{\log\left(\frac{p-|\mathrm{ne}(a)|}{|\mathrm{ne}(a)|}\right)}{1 + \frac{\mathrm{pen}(|\mathrm{ne}_{G_\Sigma}(a)|)}{n-|\mathrm{ne}_{G_\Sigma}(a)|}} \geq 2K\log\left(\frac{p}{|\mathrm{ne}(a)|}\right)(1 - o(1)) . \qquad (22)$$

Lower bound of the left part of (20). The random variables involved in this expression follow a Fisher distribution with $|\mathrm{ne}(a)\backslash \mathrm{ne}_{G_\Sigma}(a)|$ and $n-|\mathrm{ne}(a)|$ degrees of freedom. To conclude, we only need to compare the quantile of such a variable with the bound (22). Let $u \in (0,1)$ and let $F^{-1}_{D,N}(u)$ denote the $1-u$ quantile of a Fisher random variable with D and N degrees of freedom. By Lemma 1 in [4], it holds that

$$\begin{aligned} DF^{-1}_{D,N}(u) &\leq D + 2\sqrt{D\left(1+2\frac{D}{N}\right)\log\left(\frac{1}{u}\right)} \\ &+ \left(1+2\frac{D}{N}\right)\frac{N}{2}\left[\exp\left(\frac{4}{N}\log\left(\frac{1}{u}\right)\right)-1\right] . \end{aligned}$$

Let us set u to

$$u = \left\{ p^{3/2} e^{|\mathrm{ne}(a)\backslash \mathrm{ne}_{G_\Sigma}(a)|} \binom{p-|\mathrm{ne}_{G_\Sigma}(a)|-1}{|\mathrm{ne}(a)\setminus \mathrm{ne}_{G_\Sigma}(a)|} \right\}^{-1} .$$

Since we consider the case $n/\log(p)^2 \geq 1$ and $p \geq n$, the term $4/(n-|\mathrm{ne}(a)|)\log(1/u)$ goes to 0 with n (uniformly w.r.t. $\mathrm{ne}(a)$).

$$\begin{aligned} A_2 = F^{-1}_{|\mathrm{ne}(a)\backslash \mathrm{ne}_{G_\Sigma}(a)|, n-|\mathrm{ne}(a)|}(u) &\leq 1 + 2\sqrt{\frac{1}{|\mathrm{ne}(a)\setminus \mathrm{ne}_{G_\Sigma}(a)|}(1+o(1))\log\left(\frac{1}{u}\right)} \\ &+ \frac{2}{|\mathrm{ne}(a)\setminus \mathrm{ne}_{G_\Sigma}(a)|}(1+o(1))\log\left(\frac{1}{u}\right) . \end{aligned}$$

The term $\log(1/u)/|\mathrm{ne}(a)\backslash \mathrm{ne}_{G_\Sigma}(a)|$ goes to infinity with n (uniformly w.r.t. $\mathrm{ne}(a)$). Hence, we get

$$A_2 \leq 1 + \frac{2}{|\mathrm{ne}(a)\setminus \mathrm{ne}_{G_\Sigma}(a)|}\log\left(\frac{1}{u}\right)(1+o(1)) .$$

Applying the classical inequality $\log\binom{l}{k} \leq k\log(el/k)$, we obtain

$$\begin{aligned} A_2 &\leq \left[3\frac{\log(p)}{|\mathrm{ne}(a)\setminus \mathrm{ne}_{G_\Sigma}(a)|} + 2\log\left(\frac{p}{|\mathrm{ne}(a)\setminus \mathrm{ne}_{G_\Sigma}(a)|}\right)\right](1+o(1)) \\ &\leq 5\log\left(\frac{p}{|\mathrm{ne}(a)\setminus \mathrm{ne}_{G_\Sigma}(a)|}\right)(1+o(1)) . \end{aligned} \tag{23}$$

Conclusion. Let us compare the lower bound (22) of A_1 with the upper bound (23) of A_2.

- Let us first assume that $|\mathrm{ne}(a)| \leq 2|\mathrm{ne}_{G_\Sigma}(a)|$. Then, we have

$$A_1 \geq 2K\log\left(\frac{p}{|\mathrm{ne}_{G_\Sigma}(a)|}\right)(1-o(1)) \geq 2K(1-s)\log(p)\,(1-o(1)) ,$$

since $|\mathrm{ne}_{G_\Sigma}(a)| \leq n^s/\log(p) \leq p^s$. In particular,

$$A_2 \leq 5\log\left(\frac{p}{|\mathrm{ne}(a)\setminus \mathrm{ne}_{G_\Sigma}(a)|}\right)(1+o(1)) < A_1,$$

for n large enough since we assume that $2K(1-s) > 5$.

- If $|\mathrm{ne}(a)| > 2|\mathrm{ne}_{G_\Sigma}(a)|$, we also have

$$A_2 \leq 5\log\left(\frac{p}{|\mathrm{ne}(a)|}\right)(1+o(1)) < A_1\ ,$$

for n large enough since we assume that $2K > 5$.

It follows from Ineq. (20) and the definition of A_1 and A_2 that

$$\mathbb{P}\left[\Delta(\mathrm{ne}(a), \mathrm{ne}_{G_\Sigma}(a)) < 0\right] \leq \left\{p^{3/2} e^{|\mathrm{ne}(a)\setminus \mathrm{ne}_{G_\Sigma}(a)|}\binom{p-|\mathrm{ne}_{G_\Sigma}(a)|}{|\mathrm{ne}(a)\setminus \mathrm{ne}_{G_\Sigma}(a)|}\right\}^{-1}\ ,$$

for n larger than some positive constant that may depend on K, s, but does *not* depend on $\mathrm{ne}(a)$. Applying this bound to any neighborhood $\mathrm{ne}(a)$ that strictly contains $\mathrm{ne}_{G_\Sigma}(a)$ yields Statement (18):

$$\mathbb{P}\left[\widehat{\mathrm{ne}}(a) \supsetneq \mathrm{ne}_{G_\Sigma}(a)\right] \leq p^{-3/2}\ ,$$

for n large enough.

6.3.2. Bound on $\mathbb{P}\left(\widehat{\mathrm{ne}}(a) \not\supseteq \mathrm{ne}_{G_\Sigma}(a)\right)$

Again, we shall prove that $\Delta[\mathrm{ne}(a), \mathrm{ne}_{G_\Sigma}(a)]$ is positive for $\mathrm{ne}(a) \not\supseteq \mathrm{ne}_{G_\Sigma}(a)$ with overwhelming probability. We recall that $\theta_{\mathrm{ne}(a)}$ is the vector in $\mathbb{R}^p$ such that $\Sigma^{1/2}\theta_{\mathrm{ne}(a)}$ is the orthogonal projection of $\Sigma^{1/2}\theta_a$ onto the linear span $\left\{\Sigma^{1/2}\beta\ : \mathrm{supp}(\beta) \subset \mathrm{ne}(a)\right\}$. Moreover, $\|\Sigma^{1/2}(\theta_a - \theta_{\mathrm{ne}(a)})\|^2 = \mathrm{Var}(X_a|X_{\mathrm{ne}(a)}) - \mathrm{Var}(X_a|X_{-a})$ (see e.g. Lemma 7.1 in [28]).

Then, $\Delta(\mathrm{ne}(a), \mathrm{ne}_{G_\Sigma}(a))$ decomposes as

$$\begin{aligned}\Delta(\mathrm{ne}(a), \mathrm{ne}_{G_\Sigma}(a)) = &\left\|\Pi^\perp_{\mathrm{ne}(a)}\left[\boldsymbol{\epsilon}_a + \mathbf{X}(\theta_a - \theta_{\mathrm{ne}(a)})\right]\right\|_n^2\left[1+\frac{\mathrm{pen}(|\mathrm{ne}(a)|)}{n-|\mathrm{ne}(a)|}\right]\\ &- \left\|\Pi^\perp_{\mathrm{ne}_{G_\Sigma}(a)}\boldsymbol{\epsilon}_a\right\|_n^2\left[1+\frac{\mathrm{pen}(|\mathrm{ne}_{G_\Sigma}(a)|)}{n-|\mathrm{ne}_{G_\Sigma}(a)|}\right]\ .\end{aligned}$$

Let $\kappa = 6/7$ and let us define

$$E_{\mathrm{ne}(a)} = \kappa^{-1}\left\langle \frac{\Pi^\perp_{\mathrm{ne}(a)}\mathbf{X}(\theta-\theta_{\mathrm{ne}(a)})}{\|\Pi^\perp_{\mathrm{ne}(a)}\mathbf{X}(\theta-\theta_{\mathrm{ne}(a)})\|_n}, \Pi^\perp_{\mathrm{ne}(a)}\boldsymbol{\epsilon}_a\right\rangle_n^2 + \|\Pi_{\mathrm{ne}(a)}\boldsymbol{\epsilon}_a\|_n^2\ .$$

We recall that $\langle .,.\rangle_n$ is the inner product associated to the norm $\|.\|_n$. The quantity $\Delta(\mathrm{ne}(a), \mathrm{ne}_{G_\Sigma}(a))$ is positive if

$$\begin{aligned}(1-\kappa)\|\Pi^\perp_{\mathrm{ne}(a)}\mathbf{X}(\theta-\theta_{\mathrm{ne}(a)})\|_n^2 > E_{\mathrm{ne}(a)}&\left[1+\frac{\mathrm{pen}(|\mathrm{ne}(a)|)}{n-|\mathrm{ne}(a)|}\right]\\ +\quad \|\boldsymbol{\epsilon}_a\|_n^2&\left[\frac{\mathrm{pen}(|\mathrm{ne}_{G_\Sigma}(a)|)}{n-|\mathrm{ne}_{G_\Sigma}(a)|} - \frac{\mathrm{pen}(|\mathrm{ne}(a)|)}{n-|\mathrm{ne}(a)|}\right]\ .\end{aligned} \tag{24}$$

We respectively call A_3 and A_4 the right and the left terms of the inequality. We shall control their deviations in order to prove that (24) holds with high probability.

Upper Bound of A_3. On an event $\mathbb{A}$ of probability larger than $1-2p^{-3/2}$, the random variable $\|\boldsymbol{\epsilon}_a\|_n^2$ satisfies (see Lemma 1 in [21]).

$$1-2\sqrt{\frac{3\log(p)}{2n}} \leq \frac{\|\boldsymbol{\epsilon}_a\|_n^2}{\mathrm{Var}(X_a|X_{-a})} \leq 1+2\sqrt{\frac{3\log(p)}{2n}}+3\frac{\log(p)}{n} .$$

Let us bound the other random variables involved in (24). As explained in the proof of Th.3.1 in [28], the random variables $\|\Pi^{\perp}_{\mathrm{ne}(a)}\mathbf{X}(\theta-\theta_{\mathrm{ne}(a)})\|_n^2$ and $E_{\mathrm{ne}(a)}$ follow distributions of linear combinations of χ^2 random variables. We apply again Lemma 1 in [21] . On a event $\mathbb{A}_{\mathrm{ne}(a)}$ of probability larger than $1-2p^{-3/2}e^{-|\mathrm{ne}(a)|}\binom{p-1}{|\mathrm{ne}(a)|}^{-1}$, it holds that

$$\frac{\|\Pi^{\perp}_{\mathrm{ne}(a)}\mathbf{X}(\theta-\theta_{\mathrm{ne}(a)})\|_n^2}{\mathrm{Var}(X_a|X_{\mathrm{ne}(a)})-\mathrm{Var}(X_a|X_{-a})} \geq 1-\frac{|\mathrm{ne}(a)|}{n} - 2\sqrt{\frac{\frac{3}{2}\log(p)+|\mathrm{ne}(a)|\left[2+\log\left(p-1\right)\right]}{n}}$$

and

$$\begin{aligned}\frac{E_{\mathrm{ne}(a)}}{\mathrm{Var}(X_a|X_{-a})} &\leq \frac{|\mathrm{ne}(a)|+\kappa^{-1}}{n} \\ &+\frac{2}{n}\sqrt{(|\mathrm{ne}(a)|+\kappa^{-2})\left[|\mathrm{ne}(a)|\left(2+\log\left(\frac{p-1}{|\mathrm{ne}(a)|}\right)\right)+\frac{3}{2}\log(p)\right]} \\ &+\frac{2\kappa^{-1}}{n}\left[|\mathrm{ne}(a)|\left(2+\log\left(\frac{p-1}{|\mathrm{ne}(a)|}\right)\right)+\frac{3}{2}\log(p)\right] .\end{aligned}$$

We derive that

$$\begin{aligned}\frac{E_{\mathrm{ne}(a)}}{\mathrm{Var}(X_a|X_{-a})} &\leq \frac{2\kappa^{-1}}{n}\left[|\mathrm{ne}(a)|\log\left(\frac{p-1}{|\mathrm{ne}(a)|}\right)+\frac{3}{2}\log(p)\right](1+o(1)) \\ &+ \frac{\sqrt{6|\mathrm{ne}(a)|\log(p)}}{n}+\frac{\kappa^{-1}}{n} .\end{aligned}$$

- **CASE 1**: $\mathrm{ne}(a)$ **is non empty**.

$$\frac{E_{\mathrm{ne}(a)}}{\mathrm{Var}(X_a|X_{-a})} \leq \kappa^{-1}\frac{2|\mathrm{ne}(a)|\log\left(\frac{p-1}{|\mathrm{ne}(a)|}\right)+3\log(p)}{n}(1+o(1)) .$$

Let us upper bound the terms involving $\mathrm{pen}(|\mathrm{ne}(a)|)$ in (24) on the event

$\mathbb{A} \cap \mathbb{A}_{\mathrm{ne}(a)}$.

$$\left\{ E_{\mathrm{ne}(a)} \left[1 + \frac{\mathrm{pen}(|\mathrm{ne}(a)|)}{n - |\mathrm{ne}(a)|} \right] - \|\boldsymbol{\epsilon}_a\|_n^2 \frac{\mathrm{pen}(|\mathrm{ne}(a)|)}{n - |\mathrm{ne}(a)|} \right\} / \mathrm{Var}(X_a | X_{-a})$$
$$\leq \frac{\kappa^{-1}}{n} \left(2|\mathrm{ne}(a)| \log\left(\frac{p-1}{|\mathrm{ne}(a)|} \right) + 3\log(p) \right) (1 + o(1))$$
$$- \frac{2K}{n} |\mathrm{ne}(a)| \log\left(\frac{p-1}{|\mathrm{ne}(a)|} \right) (1 + o(1)) \ .$$

This last quantity is negative for n large enough since $K \geq 3$.

- **CASE 2**: $\mathrm{ne}(a)$ **is empty**. We get the upper bound

$$E_{\mathrm{ne}(a)} \left[1 + \frac{\mathrm{pen}(|\mathrm{ne}(a)|)}{n - |\mathrm{ne}(a)|} \right] - \|\boldsymbol{\epsilon}_a\|_n^2 \frac{\mathrm{pen}(|\mathrm{ne}(a)|)}{n - |\mathrm{ne}(a)|}$$
$$\leq \frac{\kappa^{-1} + 3\log(p)}{n} \mathrm{Var}(X_a | X_{-a})$$
$$\leq (3 + \kappa^{-1}) n^{s-1} \mathrm{Var}(X_a | X_{-a}) \ .$$

Indeed, $\log(p)$ has to be smaller than n^s. If this is not the case, then $\mathrm{ne}_{G_\Sigma}(a)$ should be empty and $\mathrm{ne}(a)$ cannot satisfy $\mathrm{ne}_{G_\Sigma}(a) \nsubseteq \mathrm{ne}(a)$.

We conclude that on the event $\mathbb{A} \cap \mathbb{A}_{\mathrm{ne}(a)}$,

$$A_3 \leq (3 + \kappa^{-1}) n^{s-1} \mathrm{Var}(X_a | X_{-a}) + \|\boldsymbol{\epsilon}_a\|_n^2 \frac{\mathrm{pen}(|\mathrm{ne}_{G_\Sigma}(a)|)}{n - |\mathrm{ne}_{G_\Sigma}(a)|} \ ,$$

for n large enough. Let us upper bound the penalty term as done in the upper bound of A_1.

$$\mathrm{pen}(\mathrm{ne}_{G_\Sigma}(a)) \leq LK \frac{|\mathrm{ne}_{G_\Sigma}(a)|}{n} \log\left(\frac{p-1}{|\mathrm{ne}_{G_\Sigma}(a)|} \right) \ .$$

Since $|\mathrm{ne}_{G_\Sigma}(a)|$ is assumed to be smaller than $\frac{n^s}{\log(p)}$, the term A_3 is upper bounded as follows

$$A_3 \leq (K+1) n^{s-1} \mathrm{Var}(X_a | X_{-a}) O(1) \ . \tag{25}$$

for n large enough.

Lower Bound of A_4. Let us lower bound the left term A_4 in (24) on the event $\mathbb{A} \cap \mathbb{A}_{\mathrm{ne}(a)}$.

$$A_4 \geq (1 - o(1))(1 - \kappa) \left[\mathrm{Var}(X_a | X_{\mathrm{ne}(A)}) - \mathrm{Var}(X_a | X_{-a}) \right]$$
$$\geq (1 - o(1))(1 - \kappa) \min_{b \in \Gamma \setminus \{a\}} (\theta_{a,b})^2 \min_{b,c \in \Gamma \setminus \{a\}} \frac{\mathrm{Var}(X_b | X_{-b})}{\mathrm{Var}(X_c | X_{-c})} \mathrm{Var}(X_a | X_{-a})$$
$$\geq (1 - \kappa)(1 - o(1)) n^{s'-1} \mathrm{Var}(X_a | X_{-a}) \ .$$

Thanks to the last bound and (25) and since s' is larger than s, $A_3 < A_4$ on the event $\mathbb{A} \cap \mathbb{A}_{\mathrm{ne}(a)}$ and for n large enough (not depending on $\mathrm{ne}(a)$). Hence, for n large

enough the inequality (24) holds simultaneously for all neighborhoods $\mathrm{ne}(a)$ such that $\mathrm{ne}_{G_\Sigma}(a) \not\subseteq \mathrm{ne}(a)$ with probability larger than $1 - 2p^{-3/2} - 2(e/(e-1))p^{-3/2}$. We conclude that

$$\mathbb{P}\left(\widehat{\mathrm{ne}}(a) \not\supseteq \mathrm{ne}_{G_\Sigma}(a)\right) \leq 6p^{-3/2} ,$$

for n large enough.

6.4. Lemmas

Let us prove the following lemmas.

Lemma 6.3. *For any positive integer $d \leq e^{-2}(p-1)$,*

$$\mathrm{EDKhi}\left[d+1, n-d-1, \left[\binom{p-1}{d}(d+1)^2\right]^{-1}\right] \geq d+1 .$$

Lemma 6.4. *For any positive number x and any positive integers d and N, $\mathrm{EDKhi}(d, N, x)$ is an increasing function with respect to d and a decreasing function with respect to N.*

Lemma 6.5. *For any integer $d \geq 2$, the function*

$$\frac{\mathbb{E}\left[(X_d - x\frac{X_N}{N})_+\right]}{\mathbb{E}\left[(X_2 - x)_+\right]}$$

is increasing with respect to x as soon as $x \geq d$.

6.4.1. Proof of Lemma 6.2

Let us write $L_1 = \log\left(\binom{p-1}{d_1}\right)$ and $L_2 = \log\left(\binom{p-1}{d_2}\right)$. Lemma 6.4 ensures that

$$\mathrm{EDKhi}\left(d_1+1, n-d_1-1, e^{-L_1}\right) \geq \mathrm{EDKhi}\left(d_2+1, n-d_2-1, e^{-L_1}\right) . \quad (26)$$

Let $x_1 \geq x_2$ be two positive numbers larger than some integer d_2+1. By Lemma 6.5, it holds that

$$\frac{\mathrm{DKhi}(d_2+1, n-d_2-1, x_1)}{\mathrm{DKhi}(d_2+1, n-d_2-1, x_2)} \geq \frac{\mathbb{E}\left[(X_2 - x_1)_+\right]}{\mathbb{E}\left[(X_2 - x_2)_+\right]} = e^{-(x_1-x_2)/2} .$$

By Lemma 6.3, $\mathrm{EDKhi}(d_2+1, n-d_2-1, e^{-L_2})$ is larger than d_2+1. Setting $x_1 = \mathrm{EDKhi}(d_2+1, n-d_2-1, e^{-L_1})$ and $x_2 = \mathrm{EDKhi}(d_2+1, n-d_2-1, e^{-L_2})$, we obtain

$$\mathrm{EDKhi}(d_2+1, n-d_2-1, e^{-L_1}) - \mathrm{EDKhi}(d_2+1, n-d_2-1, e^{-L_2}) \geq 2(L_1 - L_2) , \quad (27)$$

for $d_2 \geq 1$. Gathering the bounds (26), (27) with the definition (3) of the penalty enables to conclude

$$\mathrm{pen}(d_1) - \mathrm{pen}(d_2) \geq 2K(d_1 - d_2)\log\left(\frac{p-d_1}{d_1}\right) .$$

6.4.2. Proof of Lemma 6.3

We write henceforth X_d and X'_N for two independent χ^2 variables with d and N degrees of freedom. By Jensen inequality, we get

$$\begin{aligned} d \times \mathrm{DKhi}(d,N,x) &= \mathbb{E}\left[\left(X_d - x\,\frac{X'_N}{N}\right)_+\right] \\ &\geq \mathbb{E}\left[(X_d - x)_+\right] \geq \mathbb{E}\left[(X_2 - x)_+\right] = 2e^{-x/2}. \end{aligned}$$

for any $x > 0$ and any $d \geq 2$. Setting $x = \mathrm{EDKhi}(d, N, e^{-L})$ with $L \geq 0$, we obtain

$$\mathrm{EDKhi}(d, N, e^{-L}) \geq 2L - 2\log(d), \quad \text{for } d \geq 2.$$

$$\mathrm{EDKhi}\left[d+1, n-d-1, \left[\binom{p-1}{d}(d+1)^2\right]^{-1}\right] \;\geq\; 2\log\binom{p-1}{d},$$

which is larger than $2d\log[(p-1)/(ed)]$. This allows to conclude.

6.4.3. Proof of Lemma 6.4

By definition (3) of the function EDKhi, we only have to prove that $\mathrm{DKhi}(d, N, x)$ is increasing with respect to d and decreasing with respect to n.

Conditioning on X_N (resp. X_d) it suffices to prove the two following facts:
FACT 1: Let d be a positive integer. For any positive number x,

$$d\mathbb{E}\left[(X_{d+1} - x)_+\right] \geq (d+1)\mathbb{E}\left[(X_d - x)_+\right] .$$

FACT 2: Let N be a positive integer. For any positive numbers x and x',

$$\mathbb{E}\left[\left(x' - x\frac{X_N}{N}\right)_+\right] \geq \mathbb{E}\left[\left(x' - x\frac{X_{N+1}}{N+1}\right)_+\right] .$$

Proof of FACT 1. Let $(Z_1, \ldots, Z_{d+1})$ be $d+1$ independent χ^2 random variables with 1 degree of freedom. Let $Y = \sum_{i=1}^{d+1} Z_i$ and for any $i \in \{1, \ldots d+1\}$, let $Y^{(i)}$ be the sum $Y^{(i)} = \sum_{j\neq i} Z_j$. The variable Y follows a χ^2 distribution with $d+1$ degrees of freedom, while the variables $Y^{(i)}$ follow χ^2 distribution with d degrees of freedom. It holds that

$$d\,(Y - x)_+ \geq \sum_{i=1}^{d+1}\left(Y^{(i)} - x\right)_+ . \tag{28}$$

Indeed, if all the variables $Y^{(i)}$ are larger than x, one observes that $d\,(Y-x)_+ = d(\sum_{i=1}^{d+1} Z_i - dx)$ while the second term equals $d\sum_{i=1}^{d+1} Z_i - d(d+1)x$. If some of the variables $Y^{(i)}$ are smaller than x, it is sufficient to note that the variables $Y^{(i)}$

are smaller than Y. We prove FACT 1 by integrating the inequality (28).

Proof of FACT 2. It is sufficient to prove that for any positive number x,

$$\mathbb{E}\left[\left(x-\frac{X_N}{N}\right)_+\right] \geq \mathbb{E}\left[\left(x-\frac{X_{N+1}}{N+1}\right)_+\right] .$$

Observe that $\mathbb{E}\left[\left(x-\frac{X_N}{N}\right)_+\right] = (x-1)+\mathbb{E}\left[\left(\frac{X_N}{N}-x\right)_+\right]$. Hence, it remains to prove that

$$(N+1)\mathbb{E}\left[(X_N-Nx)_+\right] \geq N\mathbb{E}\left[(X_{N+1}-(N+1)x)_+\right] . \tag{29}$$

As in the proof of FACT 1, let $(Z_1,\dots,Z_{d+1})$ be $d+1$ independent χ^2 random variables with 1 degree of freedom. Let $Y=\sum_{i=1}^{d+1} Z_i$ and for any $i\in\{1,\dots d+1\}$, let $Y^{(i)}$ be the sum $Y^{(i)}=\sum_{j\neq i} Z_i$. It holds that

$$\sum_{i=1}^{N+1}\left(Y^{(i)}-Nx\right)_+ \geq N\left(Y-(N+1)x\right)_+ . \tag{30}$$

This bound is trivial if $Y\leq (N+1)x$. If Y is larger than $(N+1)x$, then the second term equals $(N+1)\sum_{i=1}^{N+1}(Y^{(i)}-Nx)$, which is clearly smaller than the first term. Integrating the bound (30) enables to prove (29) and then FACT 2.

6.4.4. Proof of Lemma 6.5

We show that the derivate of the function $\mathbb{E}\left[(X_d-x\frac{X_N}{N})_+\right]/\mathbb{E}\left[(X_2-x)_+\right]$ in non-negative for any $x\geq d$. Thus, we have to prove the following inequality:

$$\frac{\mathbb{E}\left[\left(X_d-x\frac{X_N}{N}\right)_+\right]}{\mathbb{E}\left[\frac{X_N}{N}\mathbf{1}_{X_d\geq x\frac{X_N}{N}}\right]} \geq \frac{\mathbb{E}\left[(X_2-x)_+\right]}{\mathbb{P}(X_2\geq x)} = 2 .$$

Hence, we aim at proving that the function

$$\Psi(x)=\mathbb{E}\left[\left(X_d-x\frac{X_N}{N}\right)_+\right]-2\mathbb{E}\left[\frac{X_N}{N}\mathbf{1}_{X_d\geq x\frac{X_N}{N}}\right]$$

is positive. Observe that $\Psi(x)$ converges to 0 when x goes to infinity. Let us respectively note $f_{X_d}(t)$ and $f_{\frac{X_N}{N}}(t)$ the densities of X_d and X_N/N.

$$\Psi'(x)=\int_{t=0}^{\infty} t\left[2tf_{X_d}(xt)-\int_{u=xt}^{\infty} f_{X_d}(u)du\right] f_{\frac{X_N}{N}}(t)dt .$$

Integrating by part the density of a χ^2 distribution, we get the lower bound

$$\int_{u=xt}^{\infty} f_{X_d}(u)du \geq \frac{(1/2)^{d/2}}{\Gamma(d/2)}2(xt)^{d/2-1}e^{-xt/2} .$$

Finally, we upper bound $\Psi'(x)$.

$$\begin{aligned}
\Psi'(x) &\leq \frac{(1/2)^{d/2-1}}{\Gamma(d/2)} \int_{t=0}^{\infty} t(xt)^{d/2-1} e^{-xt/2}(t-1) f_{\frac{X_N}{N}}(t) dt \\
&\leq \frac{(1/2)^{(N+d)/2-1}}{\Gamma(d/2)\Gamma(N/2)} N^{N/2} x^{d/2-1} \int_{t=0}^{\infty} t^{d/2}(t-1) t^{N/2-1} e^{-(x+N)t/2} dt \\
&\leq \frac{2N^{N/2}x^{d/2-1}}{\Gamma(d/2)\Gamma(N/2)(x+N)^{(d+N)/2}} \int_{t=0}^{\infty} t^{(d+N)/2-1} \left(\frac{2t}{x+N} - 1\right) e^{-t} dt \\
&\leq \frac{2N^{N/2}x^{d/2-1}}{\Gamma(d/2)\Gamma(N/2)(x+N)^{(d+N)/2}} \left[\frac{2\Gamma\left(\frac{d+N}{2}+1\right)}{x+N} - \Gamma\left(\frac{d+N}{2}\right)\right] \\
&\leq \frac{2N^{N/2}x^{d/2-1}\Gamma\left(\frac{d+N}{2}\right)}{\Gamma(d/2)\Gamma(N/2)(x+N)^{(d+N)/2}} \left[\frac{d+N}{x+N} - 1\right] \leq 0 \ ,
\end{aligned}$$

since $x \geq d$. Hence, Ψ is decreasing to 0 for x larger than d and it is therefore non-negative.

7. Details for the family $\widehat{\mathcal{G}}$ of candidate graphs

7.1. CO1 family $\widehat{\mathcal{G}}_{\mathbf{C01}}$

The following construction of the family $\widehat{\mathcal{G}}_{01}$ derives from the estimation procedure of Wille and Bühlmann [30]. We write $P(a,b|c)$ for the p-value of the likelihood ratio test of the hypothesis "$R_{a,b|c} = 0$" and set

$$P_{\max}(a,b) = \max\left\{P(a,b|c),\ c \in \{\emptyset\} \cup \Gamma \setminus \{a,b\}\right\} \ .$$

For any $\alpha > 0$, the graph $\widehat{G}_{01,\alpha}$ is defined by

$$a \overset{\widehat{G}_{01,\alpha}}{\sim} b \iff P_{\max}(a,b) \leq \alpha$$

and the family $\widehat{\mathcal{G}}_{\mathrm{C01}}$ is the family of nested graphs

$$\widehat{\mathcal{G}}_{\mathrm{C01}} = \left\{\widehat{G}_{01,\alpha},\ \alpha > 0 \text{ and } \deg(\widehat{G}_{01,\alpha}) \leq D\right\}.$$

C01 Algorithm

1. Compute the $p(p-1)/2$ values $P_{\max}(a,b)$.
2. Order them.
3. Extract from these values the nested graphs $\left\{\widehat{G}_{01,\alpha} : \alpha > 0\right\}$.
4. Stop when the degree becomes larger than D.

7.2. Lasso-And family $\widehat{\mathcal{G}}_{\mathbf{LA}}$

From a computational point of view, the family $\widehat{\mathcal{G}}_{\mathrm{LA}}$ can be efficiently computed with the LARS-lasso algorithm. The optimization problem (7) is broken into the p independent minimization problems

$$\widehat{\theta}_a^{\lambda} = \operatorname{argmin}\left\{\|\mathbf{X}_a - \mathbf{X}v\|^2 + \lambda\|v\|_1 :\ v \in \mathbb{R}^p \text{ and } v_a = 0\right\}, \text{ for any } a \in \Gamma, \quad (31)$$

with $\|v\|_1 = \sum_{b=1}^p |v_b|$. When λ decreases, the support of $\widehat{\theta}_a^\lambda$ is piecewise constant and the LARS-lasso algorithm provides the sequences $(\lambda_a^l)_{l\geq 1}$ of the values of λ where the support of $\widehat{\theta}_a^\lambda$ changes, as well as the sequence of the supports $\left(\mathrm{supp}(\theta^{\lambda_a^l})\right)_{l\geq 1}$. Then, we gather these p sequences as described in the algorithm below.

Given $\lambda > 0$, we define the graph $\widehat{G}_{\mathrm{and}}^\lambda$ by

$$a \overset{\widehat{G}_{\mathrm{and}}^\lambda}{\sim} b \iff \widehat{\theta}_{a,b}^\lambda \neq 0 \text{ and } \widehat{\theta}_{b,a}^\lambda \neq 0\,.$$

Finally, we define the family $\widehat{\mathcal{G}}_{\mathrm{LA}}$ as the set of graphs $\widehat{G}_{\mathrm{and}}^\lambda$ with λ large enough to ensure that $\deg(\widehat{G}_{\mathrm{and}}^\lambda) \leq D$, viz

$$\widehat{\mathcal{G}}_{\mathrm{LA}} = \left\{\widehat{G}_{\mathrm{and}}^\lambda\ , \lambda > \widehat{\lambda}_{\mathrm{and},D}\right\}, \quad \text{where} \quad \widehat{\lambda}_{\mathrm{and},D} = \sup\left\{\lambda,\ \deg(\widehat{G}_{\mathrm{and}}^\lambda) > D\right\}.$$

LA Algorithm

1. Compute with LARS-lasso the $\left(\lambda_a^l, \mathrm{supp}(\widehat{\theta}^{\lambda_a^l})\right)_{l\geq 1}$ for all $a \in \Gamma$.
2. Order the sequence $\left\{\lambda_a^l : a \in \Gamma,\ l \geq 1\right\}$.
3. Compute $\widehat{G}_{\mathrm{and}}^{\lambda_a^l}$ for all $\lambda_a^l > \widehat{\lambda}_{\mathrm{and},D}$.

7.3. Adaptive lasso family $\widehat{\mathcal{G}}_{\mathrm{EW}}$

To build the family $\widehat{\mathcal{G}}_{\mathrm{EW}}$ we start by computing the Exponential Weight estimator $\widehat{\theta}^{EW}$. For each $a \in \Gamma$, we set $H_a = \{v \in \mathbb{R}^p : v_a = 0\}$ and

$$\widehat{\theta}_a^{EW} = \int_{H_a} v\, e^{-\beta\|\mathbf{X}_a - \mathbf{X}v\|_n^2} \prod_j \left(1 + (v_j/\tau)^2\right)^{-\alpha} \frac{dv}{\mathcal{Z}_a}\,, \tag{32}$$

with $\mathcal{Z}_a = \int_{H_a} e^{-\beta\|\mathbf{X}_a - \mathbf{X}v\|_n^2} \prod_j \left(1 + (v_j/\tau)^2\right)^{-\alpha} dv$ and $\alpha, \beta, \tau > 0$. We note that $\widehat{\theta}_a^{EW}$ with $\beta = n/(2\sigma_a^2)$ and $\sigma_a^2 = \mathrm{var}(X_a \,|\, X_{-a})$ is simply the Bayesian estimator of θ_a with prior distribution $d\pi(v) \propto \prod_j \left(1 + (v_j/\tau)^2\right)^{-\alpha} dv$ on H_a. In the Gaussian setting, Dalalyan and Tsybakov [8] give a sharp and assumption-free sparse inequality for $\widehat{\theta}_a^{EW}$ with $\beta \leq n/(4\sigma_a^2)$, see Corollary 4 in Dalalyan and Tsybakov.

The construction of $\widehat{\mathcal{G}}_{\mathrm{EW}}$ is now similar to the construction of $\widehat{\mathcal{G}}_{\mathrm{LA}}$. For any $\lambda > 0$ we set

$$\widehat{\theta}^{EW,\lambda} = \mathrm{argmin}\left\{\|\mathbf{X} - \mathbf{X}\theta'\|_{n\times p}^2 + \lambda\|\theta'/\widehat{\theta}^{EW}\|_1 : \theta' \in \Theta\right\}\,, \tag{33}$$

and we define the graph $\widehat{G}_{\mathrm{or}}^{\mathrm{EW},\lambda}$ by setting an edge between a and b if either $\widehat{\theta}_{b,a}^{EW,\lambda}$ or $\widehat{\theta}_{a,b}^{EW,\lambda}$ is non-zero:

$$a \overset{\widehat{G}_{\mathrm{or}}^{\mathrm{EW},\lambda}}{\sim} b \iff \widehat{\theta}_{a,b}^{\mathrm{EW},\lambda} \neq 0 \text{ or } \widehat{\theta}_{b,a}^{\mathrm{EW},\lambda} \neq 0\,.$$

Finally, the family $\widehat{\mathcal{G}}_{\mathrm{EW}}$ is given by

$$\widehat{\mathcal{G}}_{\mathrm{EW}} = \left\{ \widehat{G}_{\mathrm{or}}^{\mathrm{EW},\lambda},\ \lambda > \widehat{\lambda}_{\mathrm{or},d}^{EW} \right\}, \quad \text{where} \quad \widehat{\lambda}_{\mathrm{or},D}^{\mathrm{EW}} = \sup \left\{ \lambda,\ \deg(\widehat{G}_{\mathrm{or}}^{\mathrm{EW},\lambda}) > D \right\}.$$

The Exponential Weight estimator $\widehat{\theta}^{EW}$ can be computed with a Langevin Monte-Carlo algorithm. We refer to [9] for the details. Once $\widehat{\theta}^{EW}$ is computed, the family $\widehat{\mathcal{G}}_{\mathrm{EW}}$ is obtained as before with the help of the LARS-lasso algorithm.

As for the family $\widehat{\mathcal{G}}_{\mathrm{LA}}$, the collection $\widehat{\mathcal{G}}_{\mathrm{EW}}$ is computed efficiently by breaking down the criterion (33) into p independent minimization problems. When λ decreases, the support of $\widehat{\theta}_a^{\mathrm{EW},\lambda}$ is piecewise constant and the LARS-lasso algorithm provides the sequences $(\lambda_a^{\mathrm{EW},l})_{l\geq 1}$ of the values of λ where the support of $\widehat{\theta}_a^{\mathrm{EW},\lambda}$ changes. Then, we gather these p sequences as described in the algorithm below.

EW Algorithm

1. Compute $\widehat{\theta}^{EW}$ with a Langevin Monte-Carlo algorithm.
2. Compute with LARS-lasso the $\left(\lambda_a^{\mathrm{EW},l}, \mathrm{supp}(\widehat{\theta}^{\lambda_a^{\mathrm{EW},l}})\right)_{l\geq 1}$ for all $a \in \Gamma$.
3. Order the sequence $\left\{\lambda_a^{\mathrm{EW},l} : a \in \Gamma,\ l \geq 1\right\}$.
4. Compute $\widehat{G}_{\mathrm{or}}^{\mathrm{EW},\lambda_a^l}$ for all $\lambda_a^{\mathrm{EW},l} > \widehat{\lambda}_{\mathrm{or},D}^{\mathrm{EW}}$.

7.4. *Quasi-exhaustive family* $\widehat{\mathcal{G}}_{\mathrm{QE}}$

QE Algorithm

1. Compute $\widehat{\mathrm{ne}}(a)$ for all $a \in \Gamma$.
2. Compute the graphs $\widehat{G}_{K,\mathrm{and}}$ and $\widehat{G}_{K,\mathrm{or}}$.
3. Work out the family $\widehat{\mathcal{G}}_{\mathrm{QE}}$.

References

[1] Christophe Ambroise, Julien Chiquet, and Catherine Matias. Inferring sparse Gaussian graphical models with latent structure. *Electron. J. Stat.*, 3:205–238, 2009.

[2] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516, 2008.

[3] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with an unknown variance. *Ann. Statist.*, 37(2):630–672, 2009.

[4] Y. Baraud, S. Huet, and B. Laurent. Adaptive tests of linear hypotheses by model selection. *Ann. Statist.*, 31(1):225–251, 2003.

[5] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[6] R. Castelo and A. Roverato. A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *J. Mach. Learn. Res.*, 7:2621–2650, 2006.

[7] Julien Chiquet, Alexander Smith, Gilles Grasseau, Catherine Matias, and Christophe Ambroise. SIMoNe: Statistical Inference for MOdular NEtworks. *Bioinformatics*, 25(3):417–418, FEB 1 2009.

[8] A. Dalayan and A. Tsybakov. Aggregation by exponential weighting, sharp oracle inequalities and sparsity. *Machine Learning*, 72(1-2):39– 61, 2008.

[9] A. Dalayan and A. Tsybakov. Sparse regression learning by aggregation and langevin monte-carlo, 2009. `arXiv:0903.1223`.

[10] P. Dellaportas, P. Giudici, and G. Roberts. Bayesian inference for nondecomposable graphical Gaussian models. *Sankhyā*, 65(1):43–55, 2003.

[11] A. Dobra, C. Hans, B. Jones, J. Nevins, G. Yao, and M. West. Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.*, 90(1):196–212, 2004.

[12] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. With discussion, and a rejoinder by the authors.

[13] J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive lasso and scad penalties. *Ann. Appl. Stat*, 3(2):521–541, 2009.

[14] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the lasso. *Biostatistics*, 3:432–441, 2008.

[15] C. Giraud. Estimation of Gaussian graphs by model selection. *Electron. J. Stat.*, 2:542–563, 2008.

[16] Kenneth R. Hess, K. Anderson, W. F. Symmans, V. Valero, N. Ibrahim, J. A. Mejia, D. Booser, R. L. Theriault, A. U. Buzdar, P. J. Dempsey, R. Rouzier, N. Sneige, J. S. Ross, T. Vidaurre, H. L. Gomez, G. N. Hortobagyi, and L. Pusztai. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncoloy*, 24(26):4236–4244, SEP 10 2006.

[17] J. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

[18] M. Kalisch and P. Bühlmann. Robustification of the pc-algorithm for directed acyclic graphs. *J. Comput. Graph. Statist.*, 17(4):773–789, 2008.

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

[20] C. Lam and J. Fan. Sparsistency and Rates of Convergence in Large Covariance Matrices Estimation. *Ann. Statist.*, 37(6B):4254–4278, 2009.

[21] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

[22] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

[23] Rene Natowicz, Roberto Incitti, Euler Guimaraes Horta, Benoit Charles, Philippe Guinot, Kai Yan, Charles Coutant, Fabrice Andre, Lajos Pusztai, and Roman Rouzier. Prediction of the outcome of preoperative chemother-

apy in breast cancer using DNA probes that provide information on both complete and incomplete responses. *BMC Bioinformatics*, 9, MAR 15 2008.

[24] G. Rocha, P. Zhao, and B. Yu. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice). Technical Report 759, Statistics Department, UC Berkeley, 2008.

[25] A. Rothman, P. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electron. J. Stat.*, 2:494–515, 2008.

[26] J.G. Scott and C. M. Carvalho. Feature-inclusion stochastic search for gaussian graphical models. *J. Comp. Graph. Statist.*, 17:790–808, 2009.

[27] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2000. with additional material by D. Heckerman, C. Meek, G.F. Cooper and T. Richardon, A Bradford Book.

[28] N. Verzelen. High-dimensional gaussian model selection on a gaussian design. *Ann. Inst. H. Poincaré Probab. Statist.*, 46(2):480–524, 2010.

[29] N. Verzelen. Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons., 2010. `arXiv:1008.0526`.

[30] A. Wille and P. Bühlmann. Low-order conditional independence graphs for inferring genetic networks. *Stat. Appl. Genet. Mol. Biol.*, 5:Art. 1, 34 pp. (electronic), 2006.

[31] F. Wong, C. K. Carter, and R. Kohn. Efficient estimation of covariance selection models. *Biometrika*, 90(4):809–830, 2003.

[32] W. Wu and Y. Ye. Exploring gene causal interactions using an enhanced constraint-based method. *Pattern Recognition*, 39(12):2349–2449, 2006.

[33] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[34] H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.