Estimation of Gaussian graphs by model selection

Christophe GIRAUD[1]

Rapport technique 2007-3, 25 pp.

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

[1] christophe.giraud@jouy.inra.fr

Abstract

We investigate in this paper the estimation of Gaussian graphs by model selection from a non-asymptotic point of view. We start from a n-sample of a Gaussian law $\mathbb{P}_C$ in $\mathbb{R}^p$ and focus on the disadvantageous case where n is smaller than p. To estimate the graph of conditional dependences of $\mathbb{P}_C$, we introduce a collection of candidate graphs and then select one of them by minimizing a penalized empirical risk. Our main result assess the performance of the procedure in a non-asymptotic setting. We pay a special attention to the maximal degree D of the graphs that we can handle, which turns to be roughly $n/(2\log p)$.

Key words. Gaussian graphical model - Random matrices - Model selection - Penalized empirical risk

2000 Mathematics Subject Classification. 62G08, 15A52, 62J05

Estimation of Gaussian Graphs by Model Selection

Christophe Giraud

First version 02/10/2007, second version 17/04/2008

1 Introduction

Let us consider a Gaussian law $\mathbb{P}_C$ in $\mathbb{R}^p$ with mean 0 and positive definite covariance matrix C. We write θ for the matrix of the regression coefficients associated to the law $\mathbb{P}_C$, more precisely $\theta = \left[\theta_i^{(j)}\right]_{i,j=1,\dots,p}$ is the $p \times p$ matrix such that $\theta_j^{(j)} = 0$ for $j = 1, \dots, p$ and

$$\mathbb{E}\left[\mathrm{X}^{(j)} \,\middle|\, \mathrm{X}^{(k)},\ k \neq j\right] = \sum_{k \neq j} \theta_k^{(j)} \mathrm{X}^{(k)}, \quad j \in \{1, \dots, p\}, \quad \text{a.s.}$$

for any random vector $\mathrm{X} = \left(\mathrm{X}^{(1)}, \dots, \mathrm{X}^{(p)}\right)^T$ of law $\mathbb{P}_C$. Our aim is to estimate the matrix θ by model selection from a n-sample $X_1, \dots, X_n$ i.i.d. with law $\mathbb{P}_C$. We will focus on the disadvantageous case where the sample size n is smaller than the dimension p.

We call henceforth *shape of* θ, the set of the couples of integers (i, j) such that $\theta_i^{(j)} \neq 0$. The shape of θ is usually represented by a graph $\mathbf{g}$ with p labeled vertices $\{1, \dots, p\}$, by setting an edge between the vertices i and j when $\theta_i^{(j)} \neq 0$. This graph is well-defined since $\theta_i^{(j)} = 0$ if and only if $\theta_j^{(i)} = 0$; the latter property may be seen e.g. on the formula $\theta_i^{(j)} = -(C^{-1})_{i,j}/(C^{-1})_{j,j}$ for all $i \neq j$. The graph $\mathbf{g}$ is of interest for the statistician since it depicts the conditional dependences of the variables $X^{(j)}$s. Actually, there is an edge between i and j if and only if $X^{(i)}$ *is not independent of* $X^{(j)}$ *conditionally on the other variables.* The objective in Gaussian graphs estimation is usually to detect the graph $\mathbf{g}$. Even if the purpose of our procedure is to estimate θ and not $\mathbf{g}$, we propose to estimate $\mathbf{g}$ by the way as follows. We associate to our estimator $\hat{\theta}$ of θ, the graph $\hat{\mathbf{g}}$ where we set an edge between the vertices i and j when $\hat{\theta}_i^{(j)}$ is non-zero.

Estimation of Gaussian graphs with $n \ll p$ is a current active field of research motivated by applications in postgenomic. Biotechnological developments (microarrays, 2D-electrophoresis, etc) enable to produce a huge amount of proteomic and transcriptomic data. One of the challenge in postgenomic is to infer from these data the regulation network of a family of genes (or

proteins). The task is challenging for the statistician 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 (typically 4000) and the sample size n is no more than a few tens. The Gaussian graphical modeling appears to be a valuable tool for this issue, see the papers of Kishino and Waddell [14], Dobra *et al* [9], Wu and Ye [20]. The gene expression levels in the microarray are modeled by a Gaussian law $\mathbb{P}_C$ and the regulation network of the genes is then depicted by the graph $\mathbf{g}$ of the conditional dependences.

Various procedures have been proposed to perform graph estimation when $p > n$. Many are based on multiple testing, see for instance the papers of Schäfer and Strimmer [16], Drton and Perlman [8, 10] or Wille and Bühlmann [19]. We also mention the work of Verzelen and Villers [17] for testing in a non-asymptotic framework whether there are (or not) missing edges in a given graph. Recently, several authors advocate to take advantage of the nice computational properties of the l^1-penalization to either estimate the graph $\mathbf{g}$ or the concentration matrix C^{-1}. Meinshausen and Bühlmann [15] propose to learn the graph $\mathbf{g}$ by regressing with the Lasso each variable against the others. Huang *et al.* [13] or Yuan and Lin [21] (see also Banerjee *et al.* [1] and Friedman *et al.* [11]) suggest in turn to rather estimate C^{-1} by minimizing the log-likelihood for the concentration matrix penalized by the l^1-norm. The performance of these algorithms are mostly unknown: the few theoretical results are only valid under restrictive conditions on the covariance matrix and for large n (asymptotic setting). In addition to these few theoretical results, Villers *et al.* [18] propose a numerical investigation of the validity domain of some of the above mentioned procedures.

Our aim in this work is to investigate Gaussian graph estimation by model selection from a non-asymptotic point of view. We propose a procedure to estimate θ and assess its performance in a non-asymptotic setting. Then, we discuss on the maximum degree of the graphs that we can accurately estimate and explore the performance of our estimation procedure in a small numerical study.

We will use the Mean Square Error of Prediction (MSEP) as a criterion to assess the quality of our procedure. To define this quantity, we introduce a few notations. For any $k, q \in \mathbb{N}$, we write $\|\cdot\|_{k\times q}$ for the Frobenius norm in $\mathbb{R}^{k\times q}$, namely $\|A\|^2_{k\times q} = \mathrm{Trace}\,(A^T A)$, for any $A \in \mathbb{R}^{k\times q}$. The MSEP of the estimator $\hat\theta$ is then

$$\mathrm{MSEP}(\hat\theta) = \mathbb{E}\left[\|C^{1/2}(\hat\theta - \theta)\|^2_{p\times p}\right] = \mathbb{E}\left[\|\mathrm{X}^T_{new}(\hat\theta - \theta)\|^2_{1\times p}\right],$$

where $C^{1/2}$ is the positive square root of C and X_{new} is a random vector, independent of $\hat\theta$, with distribution $\mathbb{P}_C$. We underline that the MSEP focus on the quality of the estimation of θ and not of $\mathbf{g}$. In particular, we do not aim to estimate at best the "true" graph $\mathbf{g}$, but rather to estimate at best the regression matrix θ. We choose this point of view for two reasons. First, we do not believe that the matrix θ is exactly sparse in practice, in the sense that $\theta_i^{(j)} = 0$ for most of the $i, j \in \{1, \ldots, p\}$. Rather, we want to handle cases where the matrix θ is only approximately

sparse, which means that there exists a sparse matrix θ^* which is a good approximation of θ. In this case, the shape $\mathbf{g}$ of θ may not be sparse at all, it can even be the complete graph. Our goal, is then not to estimate $\mathbf{g}$ but rather to capture the main conditional dependences given by the shape $\mathbf{g}^*$ of θ^*. The second reason for considering the MSEP as a quality criterion for our procedure, is that we want to quantify the fact that we do not want to miss the important conditional dependences, but we do not worry too much missing a weak one. In other words, even in the case where the shape $\mathbf{g}$ of θ is sparse, we are interesting in finding the main edges of $\mathbf{g}$ (corresponding to strong conditional dependences) and we do not really care of missing a "weak" edge which is overwhelmed by the noise. The MSEP is a possible way to take this issue into account.

To estimate θ, we will first introduce a collection $\mathcal{M}$ of graphs, which are our candidates for describing the shape $\mathbf{g}$ of θ. If we have no prior information on $\mathbf{g}$, a possible choice for $\mathcal{M}$ is the set of all the graph with degree[1] less than some fixed integer D. Then, we associate to each graph $m \in \mathcal{M}$, an estimator $\hat{\theta}_m$ of θ by minimizing an empirical version of the MSEP with the constraint that the shape of $\hat{\theta}_m$ is given by m, see Section 2 for the details. Finally, we select one of the candidate graph $\hat{m}$ by minimizing a penalized empirical MSEP and set $\hat{\theta} = \hat{\theta}_{\hat{m}}$. Our main result states the following risk bound for a truncation $\tilde{\theta}$ of $\hat{\theta}$.

Theorem 1 *Assume that $p > n \geq 3$. If for some $\eta < 1$ we have*

$$\text{degree}(m) \leq \eta \, \frac{n}{2\left(1.1 + \sqrt{\log p}\right)^2}, \quad \textit{for all } m \in \mathcal{M}, \tag{1}$$

then there exists some constant C_η, depending on η only, such that

$$\text{MSEP}(\tilde{\theta}) \leq C_\eta \log(p) \times \left(\min_{m \in \mathcal{M}} \left\{ \text{MSEP}(\hat{\theta}_m) \right\} \vee \frac{1}{n} \|C^{1/2}(I - \theta)\|^2 \right) + R_n, \tag{2}$$

where the residual term R_n is of order a $p^2 n^{-4 \log n}$.

In other words, if the candidate graphs have a degree small compared to $n/(2 \log p)$, then the MSEP of $\tilde{\theta}$ nearly achieves, up to a $\log(p)$ factor, the minimal MSEP of the collection of estimators $\{\hat{\theta}_m,\ m \in \mathcal{M}\}$. In particular, if $\mathbf{g} \in \mathcal{M}$, the MSEP of $\tilde{\theta}$ is upper-bounded by $\log(p)$ times the MSEP of $\hat{\theta}_{\mathbf{g}}$, which in turn is roughly upper bounded by $\deg(\mathbf{g}) \times \|C^{1/2}(I - \theta)\|^2 \log(p)/n$. The additional term $n^{-1}\|C^{1/2}(I - \theta)\|^2$ in (2) can be interpreted as a minimal variance for the estimation of θ. This minimal variance is due to the inability of the procedure to detect with probability one whether an isolated vertex of $\mathbf{g}$ is isolated or not. We mention that when each vertex of the graph $\mathbf{g}$ is connected to at least one other vertex, this variance term $n^{-1}\|C^{1/2}(I - \theta)\|^2$ remains smaller than the MSEP of $\hat{\theta}_{\mathbf{g}}$.

[1] the degree of a graph corresponds to the maximum number of edges incident to a vertex.

It is of practical interest to know if the Condition (1) on the degree of the graphs can be avoided. This point is discussed in Section 3.1, where we emphasize that it is hopeless to try to estimate accurately graphs with a degree D large compared to $n/(1+\log(p/n))$. We also prove that the size of the penalty involved in the selection procedure is minimal in some sense.

The remaining of the paper is organized as follows. After introducing a few notations, we describe the estimation procedure in Section 2 and state a more precise version of Theorem 1 in Section 3. Section 4 is devoted to a small numerical study and Section 6 to the proofs.

A few notations

Before describing our estimation procedure, we introduce a few notations about graphs we shall use all along the paper.

a. Graphs

The set of the graphs with p vertices labeled by $\{1,\dots,p\}$ is in bijection with the set $\mathcal{G}$ of all the subset g of $\{1,\dots,p\}^2$ fulfilling

- $(j,j)\notin g$ for all $j\in\{1,\dots,p\}$,
- $(i,j)\in g \ \Rightarrow (j,i)\in g$ for all $i,j\in\{1,\dots,p\}$.

Indeed, to any $g\in\mathcal{G}$ we can associate a graph with p vertices labeled by $\{1,\dots,p\}$ by setting an edge between the vertices i and j if and only if $(i,j)\in g$. For simplicity, we call henceforth "graph" any element g of $\mathcal{G}$.

For a graph $g\in\mathcal{G}$ and an integer $j\in\{1,\dots,p\}$, we set $g_j=\{i:\ (i,j)\in g\}$ and denote by $|g_j|$ the cardinality of g_j. Finally, we defined the degree of g by $\deg(g)=\max\{|g_j|:\ j=1,\dots,p\}$.

b. Directed graphs

As before, we will represent the set of the directed graph with p vertices labeled by $\{1,\dots,p\}$ by the set $\mathcal{G}^+$ of all the subset g of $\{1,\dots,p\}^2$ fulfilling "$(j,j)\notin g$ for all $j\in\{1,\dots,p\}$". More precisely, we associate to $g\in\mathcal{G}^+$ the directed graph with p vertices labeled by $\{1,\dots,p\}$ and with directed edges from i to j if and only if $(i,j)\in g$.

We note that $\mathcal{G}\subset\mathcal{G}^+$ and we extend to $g\in\mathcal{G}^+$ the above definitions of g_j, $|g_j|$, and $\deg(g)$.

2 Estimation procedure

In this section, we explain our procedure to estimate θ. We first introduce a collection of graphs and models, then we associate to each model an estimator and finally we give a procedure to

select one of them.

2.1 Collection of graphs and models

Our estimation procedure starts by the choice of either a collection $\mathcal{M} \subset \mathcal{G}$ of graphs or a collection $\mathcal{M} \subset \mathcal{G}^+$ of directed graphs which are our candidates to describe the shape of θ. Among the possible choices for $\mathcal{M}$ we mention four of them:

1. the set $\mathcal{M}_D^{\#} \subset \mathcal{G}$ of all the graph with at most D edges,
2. the set $\mathcal{M}_D^{\deg} \subset \mathcal{G}$ of all the graph with degree less than D,
3. the set $\mathcal{M}_D^{\#,+} \subset \mathcal{G}^+$ of all the directed graph with at most D directed edges,
4. the set $\mathcal{M}_D^{\deg,+} \subset \mathcal{G}^+$ of all the directed graph with degree less than D.

We call degree of $\mathcal{M}$ the integer $D_{\mathcal{M}} = \max\{\deg(m)\, m \in \mathcal{M}\}$ and note that the above collections of graphs have a degree bounded by D.

To the collection of (directed) graphs $\mathcal{M}$, we associate the following collection $\{\Theta_m,\ m \in \mathcal{M}\}$ of models to estimate θ. The model Θ_m is the linear space of those matrices in $\mathbb{R}^{p\times p}$ whose shape is given by the graph m, namely

$$\Theta_m = \left\{ A \in \mathbb{R}^{p\times p} :\ (i,j) \notin m \Rightarrow A_i^{(j)} = 0 \right\}.$$

As mentioned before, we known that $\theta_i^{(j)} = 0$ if and only if $\theta_j^{(i)} = 0$, so it seems irrelevant to (possibly) introduce directed graphs instead of graphs. Nevertheless, we must keep in mind that our aim is to estimate θ at best in terms of the MSEP. In some cases, the results can be improved when using directed graphs instead of graphs, typically when for some $i, j \in \{1, \dots, p\}$ the variance of $\theta_i^{(j)} \mathrm{X}^{(i)}$ is large compared to the conditional variance $\mathrm{Var}(\mathrm{X}^{(j)} | \mathrm{X}^{(k)},\ k \neq j)$, where as the variance of $\theta_j^{(i)} \mathrm{X}^{(j)}$ is small compared to $\mathrm{Var}(\mathrm{X}^{(i)} | \mathrm{X}^{(k)},\ k \neq i)$. Finally, we note the following inclusions for the families of models mentioned above

$$\bigcup_{m \in \mathcal{M}_D^{\#,+}} \Theta_m \ \subset\ \bigcup_{m \in \mathcal{M}_D^{\#}} \Theta_m \ \subset\ \bigcup_{m \in \mathcal{M}_D^{\deg}} \Theta_m \ \subset\ \bigcup_{m \in \mathcal{M}_D^{\deg,+}} \Theta_m .$$

2.2 Collection of estimators

We assume henceforth that $3 \leq n < p$ and that the degree $D_{\mathcal{M}}$ of $\mathcal{M}$ is upper bounded by some integer $D \leq n-2$. We start with n observations $X_1, \dots, X_n$ i.i.d. with law $\mathbb{P}_C$ and we denote

by X the $n \times p$ matrix $X = [X_1, \ldots, X_n]^T$. In the following, we write $A^{(1)}, \ldots, A^{(p)}$ for the p columns of a matrix $A \in \mathbb{R}^{k \times p}$.

We remind the reader that $\|C^{1/2}(I-\theta)\|^2 = \inf_{A \in \Theta} \|C^{1/2}(I-A)\|^2$, where Θ is the space of $p \times p$ matrices with 0 on the diagonal. An empirical version of $\|C^{1/2}(I-A)\|^2$ is $n^{-1}\|X(I-A)\|^2_{n\times p}$, which can also be viewed as an empirical version of the "risk" $\|C^{1/2}(A-\theta)\|^2$, since by Pythagorean theorem $\|C^{1/2}(A-\theta)\|^2 = \|C^{1/2}(I-A)\|^2 - \|C^{1/2}(I-\theta)\|^2$, for all $A \in \Theta$.

In this direction, we associate to any $m \in \mathcal{M}$, an estimator $\hat{\theta}_m$ of θ by minimizing on Θ_m this empirical risk

$$\|X(I-\hat{\theta}_m)\|^2_{n\times p} = \min_{A \in \Theta_m} \|X(I-A)\|^2_{n\times p}. \tag{3}$$

We note that the $p \times p$ matrix $\hat{\theta}_m$ then fulfills the equalities

$$X\hat{\theta}^{(j)}_m = \mathrm{Proj}_{X\Theta^{(j)}_m}\left(X^{(j)}\right), \quad \text{for } j = 1, \ldots, p,$$

where $\Theta^{(j)}_m$ is the linear space $\Theta^{(j)}_m = \left\{\theta^{(j)} : \theta \in \Theta_m\right\} \subset \mathbb{R}^p$ and $\mathrm{Proj}_{X\Theta^{(j)}_m}$ is the orthogonal projector onto $X\Theta^{(j)}_m$ in $\mathbb{R}^n$ (for the usual scalar product). Hence, since the covariance matrix C is positive definite and D is less than n, the minimizer of (3) is unique a.s.

2.3 Selection procedure

To estimate θ, we will select one of the estimator $\hat{\theta}_m$ by minimizing some penalized version of the empirical risk $\|X(I-\hat{\theta}_m)\|^2/n$. More precisely, we set $\hat{\theta} = \hat{\theta}_{\hat{m}}$ where $\hat{m}$ is any minimizer on $\mathcal{M}$ of the criterion

$$\mathrm{Crit}(m) = \sum_{j=1}^{p} \left[\|X^{(j)} - X\hat{\theta}^{(j)}_m\|^2 \times \left(1 + \frac{\mathrm{pen}(|m_j|)}{n - |m_j|}\right)\right], \tag{4}$$

with the penalty function $\mathrm{pen} : \mathbb{N} \to \mathbb{R}^+$ of the form of the penalties introduced in Baraud *et al.* [4]. To compute this penalty, we define for any integers d and N the Dkhi function by

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

where $F_{d,N}$ denotes a Fisher random variable with d and N degrees of freedom. The function $x \mapsto \mathsf{Dkhi}(d, N, x)$ is decreasing and we write $\mathsf{EDkhi}[d, N, x]$ for its inverse, see [4] Section 6.1 for details. Then, we fix some constant $K > 1$ and set

$$\mathrm{pen}(d) = K\,\frac{n-d}{n-d-1}\,\mathsf{EDkhi}\left[d+1, n-d-1, \left(C^d_{p-1}(d+1)^2\right)^{-1}\right]. \tag{5}$$

Size of the penalty

The size of the penalty $\text{pen}(d)$ is roughly $2Kd\log p$ for large values of p. Indeed, we will work in the sequel with collections of models, such that

$$D_{\mathcal{M}} \leq \eta \, \frac{n}{2\left(1.1+\sqrt{\log p}\right)^2}, \quad \text{for some } \eta < 1,$$

and then, we approximately have for large values of p and n

$$\text{pen}(d) \lesssim K\left(1+e^{\eta}\sqrt{2\log p}\right)^2 (d+1), \quad d \in \{0,\dots,D_{\mathcal{M}}\},$$

see Proposition 4 in Baraud *et al.* [4] for an exact bound. In Section 3.2, we show that the size of this penalty is minimal in some sense.

Choice of the tuning parameter K

Increasing the value of K decreases the size of the graph $\hat{m}$ that is selected. The choice $K=2$ gives o good control of the MSEP of $\hat{\theta}$, both theoretically and numerically (see Section 3 and 4). If we want that the rate of false discovery of edges remains smaller than 5%, the choice $K=3$ may also be appropriated.

Computational cost

The computational cost of the selection procedure appears to be very high. For example, if $\mathcal{M} = \mathcal{M}_D^{\text{deg},+}$ the computational complexity of the procedure increases as $p^{(D+1)}$ with the dimension p. In a future work [12], we will propose a modified version of this procedure, which presents a much smaller complexity.

3 The main result

Next theorem is a more precise version of Theorem 1. It gives an upper-bound on the MSEP of a slight variation $\tilde{\theta}$ of $\hat{\theta}$, defined by

$$\tilde{\theta}^{(j)} = \hat{\theta}^{(j)} \, \mathbf{1}_{\left\{\|\hat{\theta}^{(j)}\| \leq \sqrt{p}\, T_n\right\}}, \text{ for all } j \in \{1,\dots,p\}, \quad \text{with } \; T_n = n^{2\log n}. \tag{6}$$

We note that $\hat{\theta}$ and $\tilde{\theta}$ coincide in practice since the threshold level T_n increases very fast with n, e.g. $T_{20} \approx 6.10^7$.

In the sequel, we write $\sigma_j^2 = \left(C^{-1}_{j,j}\right)^{-1} = \mathrm{Var}(\mathrm{X}^{(j)} \mid \mathrm{X}^{(k)}, k \neq j)$ and define θ_m by

$$\|C^{1/2}(\theta - \theta_m)\|^2 = \min_{A_m \in \Theta_m} \|C^{1/2}(\theta - A_m)\|^2.$$

Theorem 2 *Assume that $D_{\mathcal{M}} = \max\{\deg(m),\, m \in \mathcal{M}\}$ fulfills the condition*

$$1 \leq D_{\mathcal{M}} \leq \eta \, \frac{n}{2\left(1.1 + \sqrt{\log p}\right)^2}, \quad \textit{for some } \eta < 1. \tag{7}$$

Then, the MSEP of the estimator $\tilde{\theta}$ defined by (6) is upper bounded by

$$\begin{aligned}
&\mathbb{E}\left[\|C^{1/2}(\tilde{\theta} - \theta)\|^2\right] \\
&\quad \leq c(K,\eta) \min_{m \in \mathcal{M}} \left\{ \|C^{1/2}(\theta - \theta_m)\|^2 \left(1 + \frac{\mathrm{pen}(\deg(m))}{n - \deg(m)}\right) + \frac{1}{n}\sum_{j=1}^{p} (\mathrm{pen}(|m_j|) + K \log n)\sigma_j^2 \right\} \\
&\qquad\qquad + R_n(\eta, C)
\end{aligned} \tag{8}$$

where K is the constant appearing in (5), $c(K,\eta) = \frac{K}{(K-1)(1-\sqrt{\eta})^4}$ and the residual term $R_n(\eta, C)$ (made explicit in the proof) is of order a $p^2 n^{-4\log n}$.

The proof of this theorem is delayed to Section 6.3 and we explain in Section 6.2 how to derive Theorem 1 from Theorem 2. Below, we discuss on the necessity of Condition (7) on the degree of the graphs and on the size of the penalty.

3.1 Is Condition (7) optimal or avoidable?

Condition (7) requires that $D_{\mathcal{M}}$ remains small compared to $n/(2\log p)$. We may wonder if this condition is necessary, or if we can hope to handle graphs with larger degree D. A glance at the proof of Theorem 2 shows that Condition (7) can be replaced by the weaker condition $\left(\sqrt{D_{\mathcal{M}}+1} + \sqrt{2\log C_{p-1}^{D_{\mathcal{M}}} + 1/(4C_{p-1}^{D_{\mathcal{M}}})}\right)^2 \leq \eta n$. Using the classical bound $C_{p-1}^{D} \leq (ep/D)^D$, we obtain that the latter condition is satisfied when

$$D_{\mathcal{M}} \leq \frac{\eta}{3} \times \frac{n}{2.1 + \log\frac{p}{D_{\mathcal{M}}}}, \tag{9}$$

so we can replace Condition (7) by Condition (9) in Theorem 2. Let us check now that we cannot improve (up to a multiplicative constant) upon (9).

Pythagoran equality gives $\|C^{1/2}(\theta-\hat{\theta})\|^2 = \|C^{1/2}(I-\hat{\theta})\|^2 - \|C^{1/2}(I-\theta)\|^2$, so there is no hope to control the size of $\|C^{1/2}(\theta-\hat{\theta})\|^2$ if we do not have for some $\delta \in (0,1)$ the inequalities

$$(1-\delta)\|C^{1/2}(I-A)\|_{p\times p} \leq \frac{1}{\sqrt{n}}\|X(I-A)\|_{n\times p} \leq (1+\delta)\|C^{1/2}(I-A)\|_{p\times p} \quad \text{for all } A \in \bigcup_{m\in\mathcal{M}} \Theta_m \tag{10}$$

with large probability. Under Condition (7) or (9), Lemma 1 Section 6 ensures that these inequalities hold for any $\delta > \sqrt{\eta}$ with probability $1 - 2\exp\left(-n(\delta-\sqrt{\eta})^2/2\right)$. We emphasize next that in the simple case where $C = I$, there exists a constant $c(\delta) > 0$ (depending on δ only) such that the Inequalities (10) cannot hold if $\mathcal{M}_D^{\#} \subset \mathcal{M}$ or $\mathcal{M}_D^{\#,+} \subset \mathcal{M}$ with

$$D \geq c(\delta)\frac{n}{1+\log\frac{p}{n}}.$$

Indeed, when $C = I$ and $\mathcal{M}_D^{\#} \subset \mathcal{M}$ (or $\mathcal{M}_D^{\#,+} \subset \mathcal{M}$), the Inequalities (10) enforces that $n^{-1/2}X$ satisfies the so-called δ-Restricted Isometry Property of order D introduced by Candès and Tao [5], namely

$$(1-\delta)\|\beta\|_{p\times 1} \leq \|n^{-1/2}X\beta\|_{p\times p} \leq (1+\delta)\|\beta\|_{p\times 1}$$

for all β in $\mathbb{R}^p$ with at most D non-zero components. Barabiuk *et al.* [2] (see also Cohen *et al.* [6]) have noticed that there exists some constant $c(\delta) > 0$ (depending on δ only) such that no $n \times p$ matrix can fulfill the δ-Restricted Isometry Property of order D if $D \geq c(\delta)n/(1+\log(p/n))$. In particular, the matrix X cannot satisfies the Inequalities (10) when $\mathcal{M}_D^{\#} \subset \mathcal{M}$ (or $\mathcal{M}_D^{\#,+} \subset \mathcal{M}$) with $D \geq c(\delta)n/(1+\log(p/n))$.

3.2 Can we choose a smaller penalty?

As mentioned before, under Condition (7) the penalty $\mathrm{pen}(d)$ given by (5) is approximately upper bounded by $K\left(1+e^{\eta}\sqrt{2\log p}\right)^2(d+1)$. Similarly to Theorem 1 in Baraud *et al.* [4], a slight variation of the proof of Theorem 2 enables to justify the use of a penalty of the form $\mathrm{pen}(d) = 2Kd\log(p-1)$ with $K > 1$ as long as $D_{\mathcal{M}}$ remains small (the condition on $D_{\mathcal{M}}$ is then much stronger than Condition (7)). We underline in this section, that it is not recommended to choose a smaller penalty. Indeed, next proposition shows that choosing a penalty of the form $\mathrm{pen}(d) = 2(1-\gamma)d\log(p-1)$ for some $\gamma \in (0,1)$ leads to a strong overfitting in the simple case where $\theta = 0$, which corresponds to $C = I$.

Proposition 1 *Consider three integers $1 \leq D < n < p$ such that $p \geq e^{2/(1-\gamma)}+1$ and $\mathcal{M}_D^{\#} \subset \mathcal{M}$ or $\mathcal{M}_D^{\#,+} \subset \mathcal{M}$. Assume that $\mathrm{pen}(d) = 2(1-\gamma)d\log(p-1)$ for some $\gamma \in (0,1)$ and $\theta = 0$. Then, there exists some constant $c(\gamma)$ made explicit in the proof, such that when $\hat{m}$ is selected according to (4)*

$$\mathbb{P}\left(|\hat{m}| \geq \frac{c(\gamma)\min(n, p^{\gamma/4})}{(\log p)^{3/2}} \wedge \lfloor \gamma D/8 \rfloor\right) \geq 1 - 3(p-1)^{-1} - 2e^{-\gamma^2 n/8^3}.$$

In addition, in the case where $\mathcal{M} = \mathcal{M}_D^{\deg,+}$*, we have*

$$\mathbb{P}\left(|\hat{m}_j| \geq \frac{c(\gamma)\min(n, p^{\gamma/4})}{(\log p)^{3/2}} \wedge \lfloor \gamma D/8 \rfloor\right) \geq 1 - 3(p-1)^{-1} - 2e^{-\gamma^2 n/8^3} \quad \textit{for all } j \in \{1, \ldots, p\}.$$

4 Numerical study

In this section, we carry out a small simulation study to evaluate the performance of our procedure. Our study concerns the behaviour of the estimator $\hat{\theta}$ when the sparsity decreases (Section 4.2) or when the number of covariates p increases (Section 4.3). In this direction, we fix the sample size n to 15 (a typical value in post-genomics) and run simulations for different values of p and for different sparsity levels. For comparison, we include the procedure "or" of Meinshausen and Bühlmann [15]. This choice is based on the numerical study of Villers *et al.* [18], where this procedure achieves a good trade-off between the power and the FDR. We write henceforth "MB" to refer to this procedure.

4.1 Simulation scheme

The graphs $\mathbf{g}$ are sampled according to the Erdös-Rényi model: starting from a graph with p vertices and no edges, we set edges between each couple of vertices at random with probability q (independently of the others). Then, we associate to a graph $\mathbf{g}$ a positive-definite matrix K with shape given by $\mathbf{g}$ as follows. For each $(i,j) \in \mathbf{g}$, we draw $K_{i,j} = K_{j,i}$ from the uniform distribution in $[-1,1]$ and set the elements on the diagonal of K in such a way that K is diagonal dominant, and thus positive definite. Finally, we normalize K to have ones on the diagonal and set $C = K^{-1}$.

For each value of p and q we sample 20 graphs and covariance matrices C. Then, for each covariance matrix C, we generate 200 independent samples $(X_1, \ldots, X_{15})$ of size 15 with law $\mathbb{P}_C$. For each sample, we estimate θ with our procedure and the procedure of Meinshausen and Bühlmann. For our procedure, we set $\mathcal{M} = \mathcal{M}_4^{\deg}$ and $K = 2$ or 2.5. For Meinshausen and Bühlmann's estimator $\hat{\theta}_{\mathrm{MB}}$ we set λ according to (9) in [15] with $\alpha = 5\%$, as recommended by the authors.

On the basis of the 20*200 simulations we evaluate the risk ratio

$$\text{r.Risk} = \frac{\mathrm{MSEP}(\hat{\theta})}{\min_m \mathrm{MSEP}(\hat{\theta}_m)} \ ,$$

as well as the power and the FDR for the detection of the edges of the graph $\mathbf{g}$. The calculations are made with R `www.r-project.org/`.

	$q = 10\%$			$q = 30\%$			$q = 33\%$		
Estimator	r.Risk	Power	FDR	r.Risk	Power	FDR	r.Risk	Power	FDR
$K = 2$	2.3	82%	4.9%	4.3	23%	6.8%	4.4	13%	5.6%
$K = 2.5$	2.5	81%	4.4%	4.9	20%	5.4%	4.9	10%	4.1%
MB	3.3	81%	3.7%	6.9	14%	2.9%	6.4	3.8%	1.1%

Table 1: Our procedure with $K = 2$, $K = 2.5$ and MB procedure: Risk ratio (r.Risk), Power and FDR when $n = 15$, $p = 10$ and $q = 10\%$, 30% and 33%.

	$p = 15$			$p = 20$			$p = 40$		
Estimator	r.Risk	Power	FDR	r.Risk	Power	FDR	r.Risk	Power	FDR
$K = 2$	3.6	74%	6.6%	3.7	69%	6%	5.4	68%	5.4 %
$K = 2.5$	4.3	72%	6%	4.4	68%	5.3%	6.5	67%	4.7%
MB	17	60%	4%	160	20%	4.8%	340	0.0%	0.0%

Table 2: Our procedure with $K = 2$, $K = 2.5$ and MB procedure: Risk ratio (r.Risk), Power and FDR when $n = 15$, $s = 1$ and $p = 15$, 20 and 40.

4.2 Decreasing the sparsity

To investigate the behaviour of the procedure when the sparsity decreases, we fix $(n, p) = (15, 10)$ and consider the three graph-density levels $q = 10\%$, $q = 30\%$ and $q = 33\%$. The results are reported in Table 1.

When $q = 10\%$ the procedures have a good performance. They detect on average more than 80% of the edges with a FDR lower than 5% and a risk ratio around 2.5. We note that MB has a slightly larger risk ratio than our procedure, but also a slightly smaller FDR.

When q increases above 30% the performances of the procedures declines abruptly. They detect less than 25% of the edges on average and the risk ratio increases above 4. When $q = 30\%$ or $q = 33\%$ our procedure is more powerful than MB, with a risk ratio 33% smaller.

4.3 Increasing the number of covariates

In this section, we focus on the quality of the estimation of θ and $\mathbf{g}$ when the number of covariates p increases. We thus fix the sample size n to 15 and the sparsity index $s := pq$ to 1. This last index corresponds to the mean degree of a vertex in the Erdös-Rényi model. Then, we run

simulations for three values of p, namely $p = 15$, $p = 20$ and $p = 40$. The results are reported in Table 2.

When the number p of covariates increases, the risk ratios of the procedures increase and their power decrease. Nevertheless, the performance of our procedure remains good, with a risk ratio between 3.6 and 6.5, a power close to 70% and a FDR around $5.6 \pm 1\%$. In contrast, the performances of MB decrease abruptly when p increases. For values of p larger or equal to 22 (not shown), MB procedure does not detect any edge anymore. This phenomenon was already noticed in Villers *et al.* [18].

5 Conclusion

In this paper, we propose to estimate the matrix of regression coefficients θ by minimizing some penalized empirical risk. The resulting estimator has some nice theoretical and practical properties. From a theoretical point of view, Theorem 1 ensures that the MSEP of the estimator can be upper-bounded in terms of the minimum of the MSEP of the $\{\hat{\theta}_m,\ m \in \mathcal{M}\}$ in a non-asymptotic setting and with no condition on the covariance matrix C. From a more practical point of view, the simulations of the previous section exhibit a good behaviour of the estimator. The power and the risk of our procedure are better than those of the procedure of Meinshausen and Bühlmann, especially when p increases. The counterpart of this better power is a slightly higher FDR of our procedure compared to that of Meinshausen and Bühlmann. If the FDR should be reduced, we recommand to set the tuning parameter K to a larger value, e.g. $K = 3$.

The main drawback of our procedure is its computational cost and in practice it cannot be used when p is larger than 50. In a future work [12], we propose a modification of the procedure that enables to handle much larger values of p.

Finally, we emphasize that our procedure can only estimate accurately graphs with a degree smaller than $n/(2 \log p)$ and as explained in Section 3.1, we cannot improve (up to a constant) on this condition.

6 Proofs

6.1 A concentration inequality

Lemma 1 *Consider three integers $1 \leq d \leq n \leq p$, a collection $V_1, \ldots, V_N$ of d-dimensional linear subspaces of $\mathbb{R}^p$ and a $n \times p$ matrix Z whose coefficients are i.i.d. with standard gaussian distribution. We set $\|\cdot\|_n = \|\cdot\|_{n\times 1}/\sqrt{n}$ and*

$$\lambda_d^*(Z) = \inf_{v \in V_1 \cup \ldots \cup V_N} \frac{\|Zv\|_n}{\|v\|_{p\times 1}}.$$

Then, for any $x \geq 0$

$$\mathbb{P}\left(\lambda_d^*(Z) \leq 1 - \frac{\sqrt{d} + \sqrt{2\log N} + \delta_N + x}{\sqrt{n}}\right) \leq \mathbb{P}\left(\mathcal{N} \geq x\right) \leq e^{-x^2/2}, \tag{11}$$

where $\mathcal{N}$ *has a standard Gaussian distribution and* $\delta_N = \left(N\sqrt{8\log N}\right)^{-1}$.
Similarly, for any $x \geq 0$

$$\mathbb{P}\left(\sup_{v \in V_1 \cup \ldots \cup V_N} \frac{\|Zv\|_n}{\|v\|_{p\times 1}} \geq 1 + \frac{\sqrt{d} + \sqrt{2\log N} + \delta_N + x}{\sqrt{n}}\right) \leq \mathbb{P}\left(\mathcal{N} \geq x\right) \leq e^{-x^2/2}. \tag{12}$$

Proof. The map $Z \to (\sqrt{n}\,\lambda_d^*(Z))$ is 1-Lipschitz, therefore the Gaussian concentration inequality enforces that

$$\mathbb{P}\left(\lambda_d^*(Z) \leq \mathbb{E}\left(\lambda_d^*(Z)\right) - x/\sqrt{n}\right) \leq \mathbb{P}\left(\mathcal{N} \geq x\right) \leq e^{-x^2/2}.$$

To get (11), we need to bound $\mathbb{E}\left(\lambda_d^*(Z)\right)$ from below. For $i = 1, \ldots, N$, we set

$$\lambda_i(Z) = \inf_{v \in V_i} \frac{\|Zv\|_n}{\|v\|}.$$

We get from [7] the bound

$$\mathbb{P}\left(\lambda_i(Z) \leq 1 - \sqrt{\frac{d}{n}} - \frac{x}{\sqrt{n}}\right) \leq \mathbb{P}(\mathcal{N} \geq x)$$

hence, there exists some standard Gaussian random variables $\mathcal{N}_i$ such that

$$\lambda_i(Z) \geq 1 - \sqrt{d/n} - \left(\mathcal{N}_i\right)_+ / \sqrt{n},$$

where $(x)_+$ denotes the positive part of x. Starting from Jensen inequality, we have for any $\lambda > 0$

$$\begin{aligned}
\mathbb{E}\left(\max_{i=1,\ldots,N} (\mathcal{N}_i)_+\right) &\leq \frac{1}{\lambda}\log \mathbb{E}\left(e^{\lambda \max_{i=1,\ldots,N}(\mathcal{N}_i)_+}\right) \\
&\leq \frac{1}{\lambda}\log\left(\sum_{i=1}^N \mathbb{E}\left(e^{\lambda(\mathcal{N}_i)_+}\right)\right) \\
&\leq \frac{1}{\lambda}\log N + \frac{1}{\lambda}\log\left(e^{\lambda^2/2} + 1/2\right) \\
&\leq \frac{\log N}{\lambda} + \frac{\lambda}{2} + \frac{e^{-\lambda^2/2}}{2\lambda}.
\end{aligned}$$

Setting $\lambda = \sqrt{2\log N}$, we finally get

$$\mathbb{E}\left(\lambda_d^*(Z)\right) = \mathbb{E}\left(\min_{i=1,\ldots,N} \lambda_i(Z)\right) \geq 1 - \frac{\sqrt{d} + \sqrt{2\log N} + \delta_N}{\sqrt{n}}$$

This concludes the proof of (11) and the proof of (12) is similar.

6.2 Proof of Theorem 1

Theorem 1 is a direct consequence of Theorem 2 and of the three following facts.

1. The equality $\sum_{j=1}^{p} \sigma_j^2 = \|C^{1/2}(I-\theta)\|^2$ holds.
2. Proposition 4 in Baraud *et al.* [4] ensures that when $D_{\mathcal{M}}$ fulfills Condition (7), there exists a constant $C(K,\eta)$ depending on K and η only, such that
$$\frac{\mathrm{pen}(d)}{n-d} \leq C(K,\eta) \quad \text{for all } d \leq D_{\mathcal{M}}.$$
3. When $D_{\mathcal{M}}$ fulfills (7) the MSEP of the estimator $\hat{\theta}_m$ is bounded from below by
$$\mathbb{E}\left(\|C^{1/2}(\theta-\hat{\theta}_m)\|^2\right) \geq \|C^{1/2}(\theta-\theta_m)\|^2 + \frac{1}{\left(1+\sqrt{\eta/(2\log p)}\right)^2} \sum_{j=1}^{p} |m_j| \frac{\sigma_j^2}{n}.$$
The latter inequality follows directly from Lemma 1.

Finally, to give an idea of the size of $C(K,\eta)$, we mention the following approximate bound (for n and p large)
$$C(K,\eta) = \frac{\mathrm{pen}(D_{\mathcal{M}})}{n-D_{\mathcal{M}}} \lesssim \frac{K\left(1+e^{\eta}\sqrt{2\log p}\right)^2}{n-D_{\mathcal{M}}} \times \eta \, \frac{n}{2\left(1.1+\sqrt{\log p}\right)^2} \asymp K\eta\, e^{2\eta}.$$

6.3 Proof of Theorem 2

The proof is split into two parts. First, we bound from above $\mathbb{E}\left[\|C^{1/2}(\tilde{\theta}-\theta)\|^2\right]$ by $\left(1-\sqrt{\eta}\right)^{-4}\mathbb{E}\left[\|X(\hat{\theta}-\theta)\|_n^2\right] + R_n$. Then, we bound this last term by the right hand side of (8).
To keep formulaes short, we write henceforth D for $D_{\mathcal{M}}$.

a. From $\mathbb{E}\left[\|C^{1/2}(\tilde{\theta}-\theta)\|^2\right]$ to $\mathbb{E}\left[\|X(\hat{\theta}-\theta)\|_n^2\right]$.

We set $\|\cdot\|_n = \|\cdot\|_{n\times 1}/\sqrt{n}$, $\lambda_0 = \left(1-\sqrt{\eta}\right)^2$,
$$\lambda_j^1 = \frac{\|X\theta^{(j)}\|_n}{\|C^{1/2}\theta^{(j)}\|} \quad \text{and} \quad \lambda_j^* = \inf\left\{\frac{\|XC^{-1/2}v\|_n}{\|v\|} : v \in \bigcup_{m\in\mathcal{M}_{j,D}^*} V_m\right\}$$

where $V_m = C^{1/2} < \theta^{(j)} > + C^{1/2}\Theta_m^{(j)}$ and $\mathcal{M}_{j,D}^*$ is the set of those subsets m of $\{1,\ldots,j-1,j+1,\ldots,p\}\times\{j\}$ with cardinality D. Then, for any $j = 1,\ldots,p$

$$\begin{aligned}
\mathbb{E}\left[\|C^{1/2}(\tilde{\theta}^{(j)} - \theta^{(j)})\|^2\right] &= \mathbb{E}\left[\|C^{1/2}(\hat{\theta}^{(j)} - \theta^{(j)})\|^2 \mathbf{1}_{\left\{\lambda_j^* \geq \lambda_0,\ \hat{\theta}^{(j)} = \tilde{\theta}^{(j)}\right\}}\right] \\
&\quad + \mathbb{E}\left[\|C^{1/2}\theta^{(j)}\|^2 \mathbf{1}_{\left\{\lambda_j^* \geq \lambda_0,\ \tilde{\theta}^{(j)} = 0,\ \lambda_j^1 \leq 3/2\right\}}\right] \\
&\quad + \mathbb{E}\left[\|C^{1/2}\theta^{(j)}\|^2 \mathbf{1}_{\left\{\lambda_j^* \geq \lambda_0,\ \tilde{\theta}^{(j)} = 0,\ \lambda_j^1 > 3/2\right\}}\right] \\
&\quad + \mathbb{E}\left[\|C^{1/2}(\tilde{\theta}^{(j)} - \theta^{(j)})\|^2 \mathbf{1}_{\left\{\lambda_j^* < \lambda_0\right\}}\right] \\
&= \mathbb{E}_1^{(j)} + \mathbb{E}_2^{(j)} + \mathbb{E}_3^{(j)} + \mathbb{E}_4^{(j)}.
\end{aligned}$$

We prove in the next paragraphs that $\sum_{j=1}^p \mathbb{E}_1^{(j)} \leq \lambda_0^{-2}\mathbb{E}\left[\|X(\hat{\theta} - \theta)\|_n^2\right]$ and that the residual term $R_n(\eta, C) = \sum_{j=1}^p (\mathbb{E}_2^{(j)} + \mathbb{E}_3^{(j)} + \mathbb{E}_4^{(j)})$ is of order a $p^2 T_n^{-2}$. The proofs of these bounds bear the same flavor as the proof of Theorem 1 in Baraud [3].

Upper bound on $\mathbb{E}_1^{(j)}$. Since

$$C^{1/2}(\hat{\theta}^{(j)} - \theta^{(j)}) \in \bigcup_{m \in \mathcal{M}_{j,D}^*} V_m,$$

we have

$$\|C^{1/2}(\hat{\theta}^{(j)} - \theta^{(j)})\|^2 \mathbf{1}_{\left\{\lambda_j^* \geq \lambda_0\right\}} \leq \lambda_0^{-2}\,\|X(\hat{\theta}^{(j)} - \theta^{(j)})\|_n^2$$

and therefore

$$\mathbb{E}_1^{(j)} \leq \lambda_0^{-2}\,\mathbb{E}\left[\|X(\hat{\theta}^{(j)} - \theta^{(j)})\|_n^2\right]. \tag{13}$$

Upper bound on $\mathbb{E}_2^{(j)}$. All we need is to bound $\mathbb{P}\left(\lambda_j^* \geq \lambda_0,\ \tilde{\theta}^{(j)} = 0,\ \lambda_j^1 \leq 3/2\right)$ from above. Writing λ^- for the smallest eigenvalue of C, we have on the event $\left\{\lambda_j^* \geq \lambda_0\right\}$

$$\|\hat{\theta}^{(j)}\| \leq \frac{\|C^{1/2}\hat{\theta}^{(j)}\|}{\sqrt{\lambda^-}} \leq \frac{\|X\hat{\theta}^{(j)}\|_n}{\lambda_0\sqrt{\lambda^-}}.$$

Besides, for any $m \in \mathcal{M}$,

$$X\hat{\theta}_m^{(j)} = \mathrm{Proj}_{X\Theta_m^{(j)}}\left(X\theta^{(j)} + \sigma_j\varepsilon^{(j)}\right)$$

with $\varepsilon^{(j)}$ distributed as a standard Gaussian random variable in $\mathbb{R}^n$. Therefore, on the event $\left\{\lambda_j^* \geq \lambda_0,\ \tilde{\theta}^{(j)} = 0,\ \lambda_j^1 \leq 3/2\right\}$ we have

$$\begin{aligned}
\|\hat{\theta}^{(j)}\| &\leq \frac{\|X\theta^{(j)}\|_n + \sigma_j\|\varepsilon^{(j)}\|_n}{\lambda_0\sqrt{\lambda^-}} \\
&\leq \frac{1.5\,\|C^{1/2}\theta^{(j)}\| + \sigma_j\|\varepsilon^{(j)}\|_n}{\lambda_0\sqrt{\lambda^-}}.
\end{aligned}$$

As a consequence,

$$\begin{aligned}
&\mathbb{P}\left(\lambda_j^* \geq \lambda_0,\ \tilde{\theta}^{(j)} = 0,\ \lambda_j^1 \leq 3/2\right) \\
&\quad\leq\ \mathbb{P}\left(\frac{1.5\,\|C^{1/2}\theta^{(j)}\| + \sigma_j\|\varepsilon^{(j)}\|_n}{\lambda_0\sqrt{\lambda^-}} > T_n\sqrt{p}\right) \\
&\quad\leq\ \begin{cases} 1 & \text{when } 3\,\|C^{1/2}\theta^{(j)}\| > \lambda_0\sqrt{p\lambda^-}\,T_n \\ \mathbb{P}\left(2\sigma_j\|\varepsilon^{(j)}\|_n > \lambda_0\sqrt{p\lambda^-}\,T_n\right) & \text{else,} \end{cases} \\
&\quad\leq\ \begin{cases} 9\,\|C^{1/2}\theta^{(j)}\|^2/(\lambda_0^2\lambda^- pT_n^2) & \text{when } 3\,\|C^{1/2}\theta^{(j)}\| > \lambda_0\sqrt{p\lambda^-}\,T_n \\ 4\sigma_j^2/(\lambda_0^2\lambda^- pT_n^2) & \text{else.} \end{cases}
\end{aligned}$$

Finally,

$$\mathbb{E}_2^{(j)} \leq \|C^{1/2}\theta^{(j)}\|^2\,\frac{9\,\|C^{1/2}\theta^{(j)}\|^2 + 4\sigma_j^2}{\lambda_0^2\lambda^- pT_n^2}. \tag{14}$$

Upper bound on $\mathbb{E}_3^{(j)}$. We note that $n\left(\lambda_j^1\right)^2$ follows a χ^2 distribution, with n degrees of freedom. Markov inequality then yields the bound

$$\mathbb{P}\left(\lambda_j^1 > 3/2\right) \leq \exp\left(-\frac{n}{2}\left(9/4 - 1 - \log(9/4)\right)\right) \leq \exp(-n/5).$$

As a consequence, we have

$$\mathbb{E}_3^{(j)} \leq \|C^{1/2}\theta^{(j)}\|^2\exp(-n/5). \tag{15}$$

Upper bound on $\mathbb{E}_4^{(j)}$. Writing λ^+ for the largest eigenvalue of the covariance matrix C, we have

$$\begin{aligned}
\mathbb{E}_4^{(j)} &\leq 2\mathbb{E}\left[\left(\|C^{1/2}\theta^{(j)}\|^2 + \|C^{1/2}\hat{\theta}^{(j)}\|^2\right)\mathbf{1}_{\{\lambda_j^* < \lambda_0\}}\right] \\
&\leq 2\left(\|C^{1/2}\theta^{(j)}\|^2 + \lambda^+ pT_n^2\right)\mathbb{P}\left(\lambda_j^* < \lambda_0\right).
\end{aligned}$$

The random variable $Z = XC^{-1/2}$ is $n \times p$ matrix whose coefficients are i.i.d. and have the standard Gaussian distribution. The condition (7) enforces the bound

$$\frac{\sqrt{D+1} + \sqrt{2\log|\mathcal{M}_{j,D}^*|} + \delta_{|\mathcal{M}_{j,D}^*|}}{\sqrt{n}} \leq \sqrt{\eta},$$

so Lemma 1 ensures that

$$\mathbb{P}\left(\lambda_j^* < \lambda_0\right) \leq \exp\left(-n(1-\sqrt{\eta})\eta/2\right)$$

and finally

$$\mathbb{E}_4^{(j)} \leq 2\left(\|C^{1/2}\theta^{(j)}\|^2 + \lambda^+ p T_n^2\right)\exp\left(-n(1-\sqrt{\eta})\eta/2\right). \tag{16}$$

Conclusion. Putting together the bounds (13) to (16), we obtain

$$\mathbb{E}\left[\|C^{1/2}(\tilde{\theta}-\theta)\|^2\right] = \sum_{j=1}^{p}\mathbb{E}\left[\|C^{1/2}(\tilde{\theta}-\theta)\|^2\right] \leq \lambda_0^{-2}\,\mathbb{E}\left[\|X(\hat{\theta}-\theta)\|_n^2\right] + R_n(\eta, C) \tag{17}$$

with $R_n(\eta, C) = \sum_{j=1}^{p}(\mathbb{E}_2^{(j)} + \mathbb{E}_3^{(j)} + \mathbb{E}_4^{(j)})$ of order a $p^2 T_n^{-2} = p^2 n^{-4\log n}$.

b. Upper bound on $\mathbb{E}\left[\|X(\hat{\theta}-\theta)\|_n^2\right]$. Let m^* be an arbitrary index in $\mathcal{M}$. Starting from the inequality

$$\sum_{j=1}^{p}\left(\|X^{(j)} - X\hat{\theta}_{\hat{m}}^{(j)}\|^2 \times \left(1 + \frac{\text{pen}(|\hat{m}_j|)}{n-|\hat{m}_j|}\right)\right) \leq \sum_{j=1}^{p}\left(\|X^{(j)} - X\hat{\theta}_{m^*}^{(j)}\|^2 \times \left(1 + \frac{\text{pen}(|m_j^*|)}{n-|m_j^*|}\right)\right)$$

and following the same lines as in the proof of Theorem 2 in Baraud *et al.* [4] we obtain for any $K > 1$

$$\frac{K-1}{K}\sum_{j=1}^{p}\|X(\hat{\theta}^{(j)} - \theta^{(j)})\|_n^2$$
$$\leq \sum_{j=1}^{p}\left[\|X(\theta^{(j)} - \bar{\theta}_{m^*}^{(j)})\|_n^2 + R_{m^*}^{(j)} + \frac{\sigma_j^2}{n}\left(KU_{\hat{m}_j}^{(j)} - \text{pen}(|\hat{m}_j|)\frac{V_{\hat{m}_j}^{(j)}}{n-|\hat{m}_j|}\right)\right],$$

where for any $m \in \mathcal{M}$ and $j \in \{1, \ldots, p\}$

$$X\bar{\theta}_m^{(j)} = \text{Proj}_{X\Theta_m^{(j)}}(X\theta^{(j)}), \quad \mathbb{E}\left(R_m^{(j)} \mid X^{(k)}, k \neq j\right) \leq \text{pen}(|m_j|)\left[\frac{\|X(\theta^{(j)} - \bar{\theta}_m^{(j)})\|_n^2}{n-|m_j|} + \frac{\sigma_j^2}{n}\right] \text{ a.s.}$$

and the two random variables $U_{m_j}^{(j)}$ and $V_{m_j}^{(j)}$ are independent with a $\chi^2(|m_j|+1)$ and a $\chi^2(n-|m_j|-1)$ distribution respectively. Combining this bound with Lemma 6 in Baraud *et al.* [4], we get

$$\frac{K-1}{K}\mathbb{E}\left[\|X(\hat{\theta}-\theta)\|_n^2\right]$$
$$\leq \mathbb{E}\left[\|X(\theta - \bar{\theta}_{m^*})\|_n^2\right] + \sum_{j=1}^{p}\text{pen}(|m_j^*|)\left[\frac{\mathbb{E}\left[\|X(\theta^{(j)} - \bar{\theta}_{m^*}^{(j)})\|_n^2\right]}{n-|m_j^*|} + \frac{\sigma_j^2}{n}\right]$$
$$+ K\sum_{j=1}^{p}\frac{\sigma_j^2}{n}\sum_{m_j \in \mathcal{M}_j}(|m_j|+1)\mathsf{Dkhi}\left(|m_j|+1, n-|m_j|-1, \frac{(n-|m_j|-1)\text{pen}(|m_j|)}{K(n-|m_j|)}\right),$$

where $\mathcal{M}_j = \{m_j,\ m \in \mathcal{M}\}$. The choice (5) of the penalty ensures that the last term is upper bounded by $K\sum_{j=1}^p \sigma_j^2 \log(n)/n$. We also note that $\|X(\theta^{(j)} - \bar{\theta}^{(j)}_{m^*})\|_n^2 \leq \|X(\theta^{(j)} - \theta^{(j)}_{m^*})\|_n^2$ for all $j \in \{1,\dots,p\}$ since $X\bar{\theta}^{(j)}_{m^*} = \mathrm{Proj}_{X\Theta^{(j)}_{m^*}}(X\theta^{(j)})$. Combining this inequality with $\mathbb{E}\left[\|X(\theta^{(j)} - \theta^{(j)}_{m^*})\|_n^2\right] = \|C^{1/2}(\theta^{(j)} - \theta^{(j)}_{m^*})\|^2$, we obtain

$$\begin{aligned}
&\frac{K-1}{K}\mathbb{E}\left[\|X(\hat{\theta}-\theta)\|_n^2\right] \\
&\leq \|C^{1/2}(\theta-\theta_{m^*})\|^2 + \sum_{j=1}^p \mathrm{pen}(|m_j^*|)\left[\frac{\|C^{1/2}(\theta^{(j)}-\theta^{(j)}_{m^*})\|^2}{n-|m_j^*|} + \frac{\sigma_j^2}{n}\right] + K\sum_{j=1}^p \frac{\sigma_j^2}{n}\log n \\
&\leq \|C^{1/2}(\theta-\theta_{m^*})\|^2\left(1+\frac{\mathrm{pen}(D)}{n-D}\right) + \sum_{j=1}^p (\mathrm{pen}(|m_j|) + K\log n)\frac{\sigma_j^2}{n} \qquad (18)
\end{aligned}$$

c. Conclusion. The bound (18) is true for any m^*, so combined with (17) it gives (8).

6.4 Proof of Proposition 1.

The proof of Proposition 1 is based on the following Lemma.

Let us consider a $n \times p$ random matrix Z whose coefficients $Z_i^{(j)}$ are i.i.d. with standard Gaussian distribution and a random variable ε independant of Z, with standard Gaussian law in $\mathbb{R}^n$.

To any subset s of $\{1,\dots,p\}$ we associate the linear space $V_s = \mathrm{span}\{e_j,\ j \in s\} \subset \mathbb{R}^p$, where $\{e_1,\dots,e_p\}$ is the canonical basis of $\mathbb{R}^p$. We write $Z\hat{\theta}_s = \mathrm{Proj}_{ZV_s}(\varepsilon)$, we denote by $\hat{s}_d$ the set of cardinality d such that

$$\|Z\hat{\theta}_{\hat{s}_d}\|^2 = \max_{|s|=d}\|Z\hat{\theta}_s\|^2. \qquad (19)$$

and we define

$$\mathrm{Crit}'(s) = \|\varepsilon - Z\hat{\theta}_s\|^2\left(1+\frac{\mathrm{pen}(|s|)}{n-|s|}\right).$$

Lemma 2 *Assume that $p \geq e^{2/(1-\gamma)}$ and $\mathrm{pen}(d) = 2(1-\gamma)d\log p$. We write $D_{n,p}$ for the largest integer smaller than*

$$5D/6, \quad \frac{p^{\gamma/4}}{(4\log p)^{3/2}} \quad \text{and} \quad \frac{\gamma^2 n}{512(1.1+\sqrt{\log p})^2}.$$

Then, the probability to have

$$\mathrm{Crit}'(s) > \mathrm{Crit}'(\hat{s}_{D_{n,p}}) \text{ for all } s \text{ with cardinality smaller than } \gamma D_{n,p}/6$$

is bounded from below by $1 - 3p^{-1} - 2\exp(-n\gamma^2/512)$.

The proof of this lemma is technical and in a first time we only give a sketch of it. For the details, we refer to Section 6.5.

Sketch of the proof of Lemma 2. We have

$$\begin{aligned}\|Z\hat{\theta}_s\|^2 &= \|\varepsilon\|^2 - \inf_{\hat{\alpha}\in V_s} \|\varepsilon - Z\hat{\alpha}\|^2 \\ &= \sup_{\hat{\alpha}\in V_s} \left[2 < \varepsilon, Z\hat{\alpha} > -\|Z\hat{\alpha}\|^2\right].\end{aligned}$$

According to Lemma 1, when $|s|$ is small compare to $n/\log p$, we have $\|Z\hat{\alpha}\|^2 \approx n\|\hat{\alpha}\|^2$ with large probability and then

$$\|Z\hat{\theta}_s\|^2 \approx \sup_{\hat{\alpha}\in V_s} \left[2 < Z^T\varepsilon, \hat{\alpha} > -n\|\hat{\alpha}\|^2\right] = \frac{1}{n}\,\|\mathrm{Proj}_{V_s}(Z^T\varepsilon)\|^2.$$

Now, $Z^T\varepsilon = \|\varepsilon\|Y$ with Y independent of ε and with $\mathcal{N}(0, I_p)$ distribution, so

$$\|Z\hat{\theta}_s\|^2 \approx \frac{\|\varepsilon\|^2}{n}\,\|\mathrm{Proj}_{V_s}Y\|^2.$$

Since $\max_{|s|=d} \|\mathrm{Proj}_{V_s}Y\|^2 \approx 2d\log p$ with large probability, we have $\|Z\hat{\theta}_{\hat{s}_d}\|^2 \approx 2d\log p \times \|\varepsilon\|^2/n$ and then

$$\min_{|s|=d} \mathrm{Crit}'(s) = \mathrm{Crit}'(\hat{s}_d) \approx \|\varepsilon\|^2\left(1 - \frac{2\gamma d\log p}{n}\right).$$

Therefore, with large probability we have $\mathrm{Crit}'(s) > \mathrm{Crit}'(\hat{s}_{D_{n,p}})$ for all s with cardinality less than $\gamma D_{n,p}/6$.

Proof of Proposition 1. We start with the case $\mathcal{M}_D^{\#,+} \subset \mathcal{M}$. When $|\hat{m}| \le \gamma D_{n,p-1}/6$, we have in particular $|\hat{m}_1| \le \gamma D_{n,p-1}/6$. We build $\tilde{m}$ from $\hat{m}$ by replacing $\hat{m}_1$ by a set $\tilde{m}_1 \subset \{1\}\times\{2,\dots,p\}$ which maximizes $\|X\hat{\theta}^{(1)}_{\tilde{m}}\|^2$ among all the subset $\tilde{m}_1$ of $\{1\}\times\{2,\dots,p\}$ with cardinality $D_{n,p-1}$. It follows from Lemma 2 (with p replaced by $p-1$) that the probability to have $\mathrm{Crit}(\hat{m}) \le \mathrm{Crit}(\tilde{m})$ is bounded from above by $3(p-1)^{-1} + 2\exp(-n\gamma^2/512)$. Since $\tilde{m} \in \mathcal{M}_D^{\#,+}$, the first part of Proposition 1 follows. When $\mathcal{M}_D^{\#} \subset \mathcal{M}$, the proof is similar.

When $\mathcal{M}_D^{\deg,+} \subset \mathcal{M}$, the same argument shows that for any $j \in \{1,\dots,p\}$ the probability to have $|\hat{m}_j| \le \gamma D_{n,p-1}/6$ is bounded from above by $3(p-1)^{-1} + 2\exp(-n\gamma^2/512)$.

6.5 Proof of Lemma 2

We write D for $D_{n,p}$ and Ω_0 for the event

$$\Omega_0 = \left\{ \begin{array}{ll} \|Z\hat{\theta}_{\hat{s}_D}\|^2 \ge 2D(1-\gamma/2)\|\varepsilon\|_n^2\log p & \text{and} \\ \|Z\hat{\theta}_s\|^2 \le 2\,|s|\,(2+\gamma)\|\varepsilon\|_n^2\log p, & \text{for all } s \text{ with } |s| \le D \end{array} \right\}.$$

We will prove first that on the event Ω_0 we have $\mathrm{Crit}'(s) > \mathrm{Crit}'(\hat{s}_{D_{n,p}})$ for any s with cardinality less than $\gamma D_{n,p}/6$ and then we will prove that Ω_0 has a probability bounded from below by $1 - 3p^{-1} - 2\exp(-n\gamma^2/512)$.

We write $\Delta(s) = \mathrm{Crit}'(\hat{s}_D) - \mathrm{Crit}'(s)$. Since we are interested in the sign of $\Delta(s)$, we will still write $\Delta(s)$ for any positive constant times $\Delta(s)$. We have on Ω_0

$$\frac{\Delta(s)}{\|\varepsilon\|^2} \leq \left(1 - \frac{2\log p}{n}(1-\gamma/2)D\right)\left(1 + \frac{\mathrm{pen}(D)}{n-D}\right) - \left(1 - \frac{2\log p}{n}(2+\gamma)|s|\right)\left(1 + \frac{\mathrm{pen}(|s|)}{n-|s|}\right).$$

We note that $\mathrm{pen}(|s|)/(n-|s|) \leq \mathrm{pen}(D)/(n-D)$. Multiplying by $n/(2\log p)$ we obtain

$$\begin{aligned}
\Delta(s) &\leq (1-\gamma)D\left(1 + \frac{D - 2(1-\gamma/2)D\log p}{n-D}\right) - (1-\gamma/2)D \\
&\quad -(1-\gamma)|s| + (2+\gamma)|s| + (2+\gamma)|s|\frac{\mathrm{pen}(D)}{n-D} \\
&\leq (1-\gamma)D\left(1 + \frac{D - 2(1-\gamma/2)D\log p + 2(2+\gamma)|s|\log p}{n-D}\right) - (1-\gamma/2)D + (1+2\gamma)|s|.
\end{aligned}$$

When $p \geq e^{2/(1-\gamma)}$ and $|s| \leq \gamma D/6$ the first term on the right hand side is bounded from above by $(1-\gamma)D$, then since $\gamma < 1$

$$\Delta(s) \leq (1+2\gamma)\gamma D/6 - \gamma D/2 < 0.$$

We will now bound $\mathbb{P}(\Omega_0^c)$ from above. We write $Y = Z^T\varepsilon/\|\varepsilon\|$ (with the convention that $Y = 0$ when $\varepsilon = 0$) and

$$\begin{aligned}
\Omega_1 &= \left\{\frac{2}{2+\gamma} \leq \frac{\|Z\hat{\alpha}\|_n^2}{\|\hat{\alpha}\|^2} \leq (1-\gamma/2)^{-1/2}, \quad \text{for all } \hat{\alpha} \in \bigcup_{|s|=D} V_s\right\}, \\
\Omega_2 &= \left\{\max_{|s|=D} \|\mathrm{Proj}_{V_s} Y\|^2 \geq 2(1-\gamma/2)^{1/2} D\log p\right\}, \\
\Omega_3 &= \left\{\max_{i=1,\ldots,p} Y_i^2 \leq 4\log p\right\}.
\end{aligned}$$

We first prove that $\Omega_1 \cap \Omega_2 \cap \Omega_3 \subset \Omega_0$. Indeed, we have on $\Omega_1 \cap \Omega_2$

$$\begin{aligned}
\|Z\hat{\theta}_{\hat{s}_D}\|^2 &= \max_{|s|=D} \sup_{\hat{\alpha} \in V_s} \left[2 < \varepsilon, Z\hat{\alpha} > - \|Z\hat{\alpha}\|^2\right] \\
&\geq \max_{|s|=D} \sup_{\hat{\alpha} \in V_s} \left[2 < Z^T\varepsilon, \hat{\alpha} > - n(1-\gamma/2)^{-1/2}\|\hat{\alpha}\|^2\right] \\
&\geq \frac{(1-\gamma/2)^{1/2}\|\varepsilon\|^2}{n} \max_{|s|=D} \|\mathrm{Proj}_{V_s} Y\|^2 \\
&\geq 2D(1-\gamma/2)\|\varepsilon\|_n^2 \log p.
\end{aligned}$$

Similarly, on Ω_1 we have $\|Z\hat{\theta}_s\|^2 \leq \|\varepsilon\|_n^2 \|\mathrm{Proj}_{V_s} Y\|^2 \times (2+\gamma)/2$ for all s with cardinality less than D. Since $\|\mathrm{Proj}_{V_s} Y\|^2 \leq |s| \max_{i=1,\ldots,p}(Y_i^2)$, we have on $\Omega_1 \cap \Omega_3$

$$\|Z\hat{\theta}_s\|^2 \leq 2(2+\gamma)|s| \, \|\varepsilon\|_n^2 \log p,$$

for all s with cardinality less than D and then $\Omega_1 \cap \Omega_2 \cap \Omega_3 \subset \Omega_0$.
To conclude, we bound $\mathbb{P}(\Omega_i^c)$ from above, for $i = 1, 2, 3$. First, we have

$$\mathbb{P}(\Omega_3^c) = \mathbb{P}\left(\max_{i=1,\ldots,p} Y_i^2 > 4 \log p\right) \leq 2p \, \mathbb{P}(Y_1 \geq 2\sqrt{\log(p)}) \leq 2p^{-1}.$$

To bound $\mathbb{P}(\Omega_1^c)$, we note that $(1-\gamma/2)^{-1/4} \geq 1 + \gamma/8$ and $\sqrt{2/(2+\gamma)} \leq 1 - \gamma/8$ for any $0 < \gamma < 1$, so Lemma 1 ensures that $\mathbb{P}(\Omega_1^c) \leq 2e^{-n\gamma^2/512}$. Finally, to bound $\mathbb{P}(\Omega_2^c)$, we sort the Y_i^2 in decreasing order $Y_{(1)}^2 > Y_{(2)}^2 > \ldots > Y_{(p)}^2$ and note that

$$\max_{|s|=D} \|\mathrm{Proj}_{V_s} Y\|^2 \geq D Y_{(D)}^2.$$

Furthermore, we have

$$\begin{aligned}
\mathbb{P}\left(Y_{(D)}^2 \leq 2(1-\gamma/2)^{1/2} \log p\right) &\leq \binom{D-1}{p} \mathbb{P}\left(Y_1^2 \leq 2(1-\gamma/2)^{1/2} \log p\right)^{p-D+1} \\
&\leq p^{D-1} \left(1 - \frac{p\sqrt{1-\gamma/2}}{4(1-\gamma/2)^{1/4}\sqrt{2\log p}}\right)^{p-D+1},
\end{aligned}$$

where the last inequality follows from $p \geq e^{2/(1-\gamma)}$ and Inequality (60) in Baraud *et al.* [4]. Finally, we obtain

$$\mathbb{P}\left(Y_{(D)}^2 \leq 2(1-\gamma/2)^{1/2} \log p\right) \leq p^{-1} \exp\left(D \log p - \frac{(p-D+1)p\sqrt{1-\gamma/2}}{4(1-\gamma/2)^{1/4}\sqrt{2\log p}}\right) \leq p^{-1},$$

where the last inequality comes from $D \leq p^{\gamma/4}/(4\log p)^{3/2}$. To conclude $\mathbb{P}(\Omega_2^c) \leq p^{-1}$ and $\mathbb{P}(\Omega_0^c) \leq 3p^{-1} + 2\exp(-n\gamma^2/512)$.

References

[1] O. Banerjee, L.E. Ghaoui and A. d'Aspremont. *Model selection through sparse maximum likelihood estimation.* To appear, J. Machine Learning Research **101** (2007).

[2] R. Baraniuk, M. Davenport, R. De Vore and M. Wakin. *A simple proof of the restricted isometry property for random matrices.* To appear in Constructive Approximation (2007)

[3] Y. Baraud. *Model selection for regression on a random design.* ESAIM Probab. Statist. **6** (2002), 127–146 (electronic).

[4] Y. Baraud, C. Giraud and S. Huet. *Gaussian model selection with unknown variance.* To appear in the Annals of Statistics. http://arxiv.org/abs/math/0701250v1

[5] E. Candès and T. Tao. *Decoding by linear programing.* IEEE Trans. Inf. Theory **51** (2005) no. 12, 4203–4215.

[6] A. Cohen, W. Dahmen and R. De Vore. *Compressed sensing and the best k-term approximation.* Preprint (2006) http://www.math.sc.edu/ devore/publications/CDDSensing_6.pdf

[7] K.R. Davidson and S.J. Szarek. *Local operator theory, random matrices and Banach spaces.* Handbook in Banach Spaces Vol I, ed. W. B. Johnson, J. Lindenstrauss, Elsevier (2001), 317–366.

[8] M. Drton and M. Perlman. *A sinful approach to Gaussian graphical model selection.* Tech. Rep. 457 (2004), Dept. of Statistics, University of Washington, Seattle. http://www.stat.washington.edu/www/research/reports/2004/tr457.pdf

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

[10] M. Drton and M. Perlman. *Multiple testing and error control in Gaussian Graphical model selection.* To appear in Statistical Science (2007).

[11] J. Friedman, T. Hastie, R. Tibshirani. *Sparse inverse covariance estimation with the lasso.* Preprint (2007). http://www-stat.stanford.edu/ tibs/ftp/graph.pdf

[12] C. Giraud, S. Huet and N. Verzelen. In preparation.

[13] J.Z. Huang, N. Liu, M. Pourahmadi and L. Liu. *Covariance matrix selection and estimation via penalised normal likelihood.* Biometrika **93** no 1, (2006), 85–98

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

[15] N. Meinshausen and P. Bühlmann. *High dimensional graphs and variable selection with the lasso.* Annals of Statistics **34** (2006), 1436–1462.

[16] J. Schäfer and K. Strimmer. *An empirical bayes approach to inferring large-scale gene association networks.* Bioinformatics **21** (2005), 754–764.

[17] N. Verzelen and F. Villers. *Test of neighborhood for Gaussian graphical models.* Preprint (2007).

[18] F. Villers, B. Schaeffer, C. Bertin, and S. Huet. *Assessing the validity domains of graphical Gaussian models in order to infer relationships among components of complex biological systems.* Technical Report, INRA (2008).

[19] A. Wille and P. Bühlmann. *Low-order conditional independence graphs for inferring genetic networks.* Stat. Appl. Genet. Mol. Biol. **5** (2006).

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

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