Estimation of Gaussian graphs by model selection

Christophe Giraud

ESTIMATION OF GAUSSIAN GRAPHS BY MODEL SELECTION

CHRISTOPHE GIRAUD

Abstract. Our aim in this paper is to investigate Gaussian graph estimation from a theoretical and non-asymptotic point of view. We start from a n-sample of a Gaussian law $\mathbb{P}_C$ in $\mathbb{R}^p$ and we focus on the disadvantageous case where n is smaller than p. To estimate the graph of conditional dependences of $\mathbb{P}_C$, we propose to 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 $n/(2\log p)$.

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 $X_1, \dots, X_n$ of the law $\mathbb{P}_C$. We will focus on the disadvantageous case where the sample size n is smaller than the dimension p.

The shape of θ is usually represented by a graph $\mathcal{G}$ with p 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 objective in Gaussian graphs estimation is usually to detect the graph $\mathcal{G}$. Even if the purpose of our procedure is to estimate θ and not $\mathcal{G}$, we propose to estimate $\mathcal{G}$ by the way as follows. We associate to our estimator $\hat{\theta}$ of θ, the graph $\hat{\mathcal{G}}$ where we set an edge between the vertices i and j when either $\hat{\theta}_i^{(j)}$ or $\hat{\theta}_j^{(i)}$ is non-zero.

Estimation of Gaussian graphs with $n \ll p$ is currently an active field of research motivated by applications to microarray analysis. The challenge is to infer from a small sample of microarrays

Date: Le 2 octobre 2007.

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

Key words and phrases. Gaussian graphs - Random matrices - Model selection - Penalized criterion.

the regulation network of a large family of genes. A possible way is to model the gene expression levels in the microarray by a (high-dimensional) Gaussian law (as proposed in Kishino and Waddell [11]) and then to detect the underlying graph $\mathcal{G}$ of conditional dependences from the sample. This modeling has the nice property to be simple, but one of its drawback is that it does not model retroaction loops, which is a common feature in biological processes.

Various procedure have been proposed to perform graph estimation. Many are based on multiple testing, see for instance the papers of Schäfer and Strimmer [13], Drton and Perlman [7, 8] or Wille and Bühlmann [15]. We also mention the work of Verzelen and Villers [14] 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 $\mathcal{G}$ or the concentration matrix C^{-1}. Meinshausen and Bühlmann [12] propose to learn the graph $\mathcal{G}$ by regressing with Lasso each variable against the others. Huang *et al.* [10] or Yuan and Lin [16] (see also Banerjee *et al.* [1] and Friedman *et al.* [9]) suggest in turn to rather estimate C^{-1} by minimizing the log-likelihood for the concentration matrix penalized by the l^1-norm.

Our aim in this work is to investigate Gaussian graph estimation from a theoretical and non-asymptotic point of view. First, we propose a procedure to estimate θ and asses its performance in a non-asymptotic setting, which is important for biological data where typical values of n are a few tens. Then, we discuss on the maximum degree of the graphs that we can accurately estimate. Our work is theoretic and from a practical point of view the procedure we propose suffers of a very high computational cost. In a future work, we will suggest various strategies to reduce this cost and we will compare its performance to the performance of other procedures in a numerical study.

To estimate the matrix θ, we introduce a collection of candidate graphs (possibly directed graphs) on $\{1,\dots,p\}$. To each graph, we associate an estimator of θ by minimizing some empirical risk over the space of the matrices in $\mathbb{R}^{p\times p}$ with shape given by the graph at hand, see Section 2 for the details. Then, we select one of these estimators by minimizing a penalized criterion. We evaluate the performance of the procedure through the Mean Square Error of Prediction (MSEP) of the resulting estimator $\hat\theta$. To define this quantity, we introduce a few notations. For any $k,q\in\mathbb{N}$, we denote by $A^{(1)},\dots,A^{(q)}$ the q columns of a matrix $A\in\mathbb{R}^{k\times q}$. We also write $\|\cdot\|_{k\times q}$ for the Euclidean norm in $\mathbb{R}^{k\times q}$, namely

$$\|A\|_{k\times q}^2 = \sum_{i=1}^k\sum_{j=1}^q \left(A_i^{(j)}\right)^2 = \sum_{j=1}^q \|A^{(j)}\|_{k\times 1}^2, \quad \text{for any } A\in\mathbb{R}^{k\times q}.$$

The MSEP of the estimator $\hat\theta$ is then

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

where $C^{1/2}$ is the positive square root of C and X_{new} is a random vector, independent of the sample $X_1,\dots,X_n$, with distribution $\mathbb{P}_C$. Our main result roughly states that if the candidate

graphs have a degree small compared to $n/(2\log p)$, then a slight variation of the estimator $\hat{\theta}$ performs almost as well as the best of the estimators in the collection in terms of the MSEP. Next, we emphasize that it is hopeless to try to estimate accurately graphs with degree D large compared to $n/(1+\log(p/n))$. To conclude, we 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. We describe the estimation procedure in Section 2, we state our main results in Section 3 and Section 4 is devoted to the proofs.

2. Estimation procedure

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

2.1. Collection of models.

We write $\Delta = \{(j,j) : \ j = 1,\dots,p\}$ and $\mathcal{M}^*$ for the set of all the subset of $\{1,\dots,p\}^2 \setminus \Delta$. The set $\mathcal{M}^*$ is in bijection with the set of directed graphs on $\{1,\dots,p\}$. Indeed, to a set m in $\mathcal{M}^*$ we can associate a directed graph on $\{1,\dots,p\}$ by setting for each $(i,j) \in m$ a directed edge from i to j. To any m in $\mathcal{M}^*$, we associate the linear space Θ_m (we call henceforth *model*) of those matrices θ in $\mathbb{R}^{p\times p}$ such that $\theta_i^{(j)} = 0$ when $(i,j) \notin m$. To estimate θ, we will consider a collection of models $\{\Theta_m,\ m \in \mathcal{M}\}$ indexed by some suitable subset $\mathcal{M}$ of $\mathcal{M}^*$.

We give below some possible choices for the set $\mathcal{M}$. In the sequel, $\mathcal{M}^{*,s}$ denotes the set of those m in $\mathcal{M}^*$ that are symmetric with respect to Δ, which means that $(i,j) \in m$ if and only if $(j,i) \in m$. The set $\mathcal{M}^{*,s}$ is in bijection with the set of graphs on $\{1,\dots,p\}$: to a set m in $\mathcal{M}^{*,s}$ we associate the graph with an edge between i and j if and only if $(i,j) \in m$. 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 $\mathrm{Var}(\mathrm{X}^{(i)}|\mathrm{X}^{(k)},\ k \neq i)$.

Henceforth, we write $m_j = \{i : \ (i,j) \in m\}$ for any $j \in \{1,\dots,p\}$ and $m \in \mathcal{M}^*$. We also denote by $|m_j|$ the cardinality of m_j and call degree of m (or of the graph associated to m) the integer $\deg(m) = \max\{|m_j| : \ j = 1,\dots,p\}$.

Some collections of interest.

$\mathcal{M}_D^c$: We write $\mathcal{M}_D^c$ (respectively $\mathcal{M}_D^{c,s}$) for those m in $\mathcal{M}^*$ (resp. $\mathcal{M}^{*,s}$) such that $|m_j| \leq D$ for all $j \in \{1,\dots,p\}$. We note that the set $\mathcal{M}_D^c$ (resp. $\mathcal{M}_D^{c,s}$) is in bijection with the set of directed graphs (resp. graphs) with p vertices and degree at most D.

$\mathcal{M}_D^b$: We write $\mathcal{M}_D^b \subset \mathcal{M}_D^c$ (respectively $\mathcal{M}_D^{b,s}$) for the set of the m in $\mathcal{M}^*$ (resp. $\mathcal{M}^{*,s}$) with cardinality at most D (resp. $2D$). The set $\mathcal{M}_D^b$ (resp. $\mathcal{M}_D^{b,s}$) is in bijection with the set of directed graphs (resp. graphs) with p vertices and at most D edges.

$\mathcal{M}_{D,q}$: For $1 \leq D \leq q$ we write $\mathcal{M}_{D,q} \subset \mathcal{M}_D^c$ (respectively $\mathcal{M}_{D,q}^s$) for the set of the m in $\mathcal{M}^*$ (resp. $\mathcal{M}^{*,s}$) with cardinality at most q (resp. $2q$) and such that $|m_j| \leq D$ for all $j \in \{1,\dots,p\}$. As before, the set $\mathcal{M}_{D,q}$ (resp. $\mathcal{M}_{D,q}^s$) is in bijection with the set of directed graphs (resp. graphs) with p vertices, degree bounded by D and at most q edges.

We note that all the graphs in the above collections have a degree bounded by D.

2.2. **Collection of estimators.** We start with n observations $X_1,\dots,X_n$ i.i.d. with law $\mathbb{P}_C$. We denote by X the $n \times p$ matrix $X = [X_1,\dots,X_n]^T$ and we remind that $X^{(1)},\dots,X^{(p)}$ stand for the p columns of the matrix X.

We assume henceforth that $3 \leq n < p$ and that $\mathcal{M} \subset \mathcal{M}_D^c$ for some $D \in \{1,\dots,n-2\}$. To any $m \in \mathcal{M}$ we associate an estimator $\hat\theta_m$ of θ by minimizing the squares

$$\|X(\hat\theta_m - I)\|_{n\times p}^2 = \min_{\hat\theta \in \Theta_m} \|X(\hat\theta - I)\|_{n\times p}^2. \tag{1}$$

We note that the $p \times p$ matrix $\hat\theta_m$ then fulfills the equalities

$$X\hat\theta_m^{(j)} = \mathrm{Proj}_{X\Theta_m^{(j)}}\left(X^{(j)}\right), \quad \text{for } j = 1,\dots,p,$$

where $\Theta_m^{(j)}$ is the linear space $\Theta_m^{(j)} = \left\{\theta^{(j)} :\ \theta \in \Theta_m\right\} \subset \mathbb{R}^p$ and $\mathrm{Proj}_{X\Theta_m^{(j)}}$ is the orthogonal projector onto $X\Theta_m^{(j)}$ 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 (1) is unique a.s.

2.3. **Selection procedure.** We estimate θ by $\hat\theta = \hat\theta_{\hat m}$ where $\hat m$ is any minimizer on $\mathcal{M}$ of the criterion

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

with the penalty function $\mathrm{pen} : \mathbb{N} \to \mathbb{R}^+$ computed as follows. As in Baraud *et al.* [3], 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 [3] Section 6.1 for details. Then, we fix some constant $K > 1$ and set

$$\mathrm{pen}(d) = K\,\frac{n-d}{n-d-1}\,\mathsf{EDkhi}\left[d+1, n-d-1, \left(C_{p-1}^d (d+1)(D_{\mathcal{M}}+1)\right)^{-1}\right], \tag{3}$$

where $D_{\mathcal{M}} = \max\{\deg(m) : \ m \in \mathcal{M}\}$.

Size of the penalty. The size of the penalty $\mathrm{pen}(d)$ 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

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

see Proposition 4 in Baraud *et al.* [3] for an exact bound. In Section 3.2, we show that the size of this penalty is minimal in some sense.

Computational cost. The computational cost of the selection procedure appears to be very high. For example, if $\mathcal{M} = \mathcal{M}_D^c$ it increases as $p^{(D+1)}$ with the dimension p. In a future work, we will propose various strategies to reduce the cardinality of the set $\mathcal{M}$ over which the minimization (2) occurs.

3. Main results

We are not able to derive an upper bound for the MSEP of the estimator $\hat{\theta}$ (except when $\mathcal{M} = \mathcal{M}_D^c$ for some D). Nevertheless, next theorem bounds from above the MSEP of a slight variation $\tilde{\theta}$ of $\hat{\theta}$, defined by

$$\tilde{\theta}^{(j)} = \hat{\theta}^{(j)} \mathbf{1}_{\left\{\|\hat{\theta}^{(j)}\| \leq \sqrt{p}\, T_n\right\}}, \text{ for all } j \in \{1, \ldots, p\}, \quad \text{with } \ T_n = n^{2\log n}. \tag{4}$$

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 $\sigma_j^2 = \left(C_{j,j}^{-1}\right)^{-1} = \mathrm{Var}(\mathrm{X}^{(j)} \mid \mathrm{X}^{(k)}, k \neq j)$ and define θ_m by

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

Theorem 1. *Assume that* $D_{\mathcal{M}} = \max\{\deg(m) : \ m \in \mathcal{M}\}$ *fulfills the condition*

$$1 \leq D_{\mathcal{M}} \leq \eta \, \frac{n}{2\left(1.1 + \sqrt{\log p}\right)^2}, \quad \text{for some } \eta < 1. \tag{5}$$

Then, the MSEP of the estimator $\tilde{\theta}$ *defined by (4) is upper bounded by*

$$\begin{aligned} &\mathbb{E}\left[\|C^{1/2}(\tilde{\theta} - \theta)\|^2\right] \\ &\quad \leq c(K,\eta) \ \min_{m \in \mathcal{M}} \left\{ \|C^{1/2}(\theta - \theta_m)\|^2 \left(1 + \frac{\mathrm{pen}(D_{\mathcal{M}})}{n - D_{\mathcal{M}}}\right) + \frac{1}{n}\sum_{j=1}^{p} (\mathrm{pen}(|m_j|) + K)\sigma_j^2 \right\} \\ &\qquad\qquad + R_n(\eta, C, \theta) \end{aligned} \tag{6}$$

where K *is the constant appearing in (3),* $c(K,\eta) = \frac{K}{(K-1)(1-\sqrt{\eta})^4}$ *and the residual term* $R_n(\eta, C, \theta)$ *(made explicit in the proof) is of order a* $p^2 n^{-4\log n}$.

The proof of this theorem is delayed to Section 4.2. Below, we discuss the bound (6). We first emphasize that the factor $\mathrm{pen}(D_{\mathcal{M}})/(n - D_{\mathcal{M}})$ behaves like a constant under condition (5) and then we explain how (6) enables to compare the MSEP of $\tilde{\theta}$ with the minimum over $\mathcal{M}$ of the MSEP of $\hat{\theta}_m$.

Proposition 4 in Baraud *et al.* [3] ensures that when $D_{\mathcal{M}}$ fulfills Condition (5) we can bound $\mathrm{pen}(D_{\mathcal{M}})/(n - D_{\mathcal{M}})$ by some constant depending on K and η only. Indeed, under Condition (5) we approximately have for large values of n and p

$$\frac{\mathrm{pen}(D_{\mathcal{M}})}{n - D_{\mathcal{M}}} \lesssim \frac{K\left(1 + e^{\eta}\sqrt{2\log p}\right)^2}{n - D_{\mathcal{M}}} \times \eta \, \frac{n}{2\left(1.1 + \sqrt{\log p}\right)^2} \asymp K\eta \, e^{2\eta}.$$

When $D_{\mathcal{M}}$ fulfills (5) the MSEP of the estimator $\hat{\theta}_m$ is bounded from below by

$$\mathbb{E}\left(\|C^{1/2}(\theta - \hat{\theta}_m)\|^2\right) \geq \|C^{1/2}(\theta - \theta_m)\|^2 + \frac{1}{\left(1 + \sqrt{\eta/(2\log p)}\right)^2} \ \sum_{j=1}^{p} |m_j| \, \frac{\sigma_j^2}{n} \, .$$

Besides, we have $\sum_{j=1}^{p} \sigma_j^2 = \|C^{1/2}(I - \theta)\|^2$, so there exists some constant c' depending on K and η only, such that under Condition (5) we have

$$\mathbb{E}\left(\|C^{1/2}(\theta - \tilde{\theta})\|^2\right) \leq c' \left[\log(p) \inf_{m \in \mathcal{M}} \mathbb{E}\left(\|C^{1/2}(\theta - \hat{\theta}_m)\|^2\right) \vee \frac{\|C^{1/2}(I - \theta)\|^2}{n}\right] + R_n(\eta, C, \theta).$$

The MSEP of $\tilde{\theta}$ thus nearly achieves, up to a $\log(p)$ factor, the minimal MSEP of the collection of estimators $\left\{\hat{\theta}_m, \ m \in \mathcal{M}\right\}$.

3.1. **Is Condition (5) optimal?** Condition (5) 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 1 shows that Condition (5) 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

$$D_{\mathcal{M}} \leq \frac{\eta}{6} \times \frac{n}{1+\log\frac{p}{D_{\mathcal{M}}}}, \tag{7}$$

so we can replace Condition (5) by Condition (7) in Theorem 1. Let us check now that we cannot improve (up to a multiplicative constant) upon (7).

Pythagora's theorem 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

$$(1-\delta)\|C^{1/2}(I-\alpha)\|_{p\times p} \leq \frac{1}{\sqrt{n}}\|X(I-\alpha)\|_{n\times p} \leq (1+\delta)\|C^{1/2}(I-\alpha)\|_{p\times p} \quad \text{for all } \alpha \in \bigcup_{m\in\mathcal{M}} \Theta_m \tag{8}$$

with large probability. Under Condition (5) or (7), Lemma 1 Section 4 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 (8) cannot hold if $\mathcal{M}_D^b \subset \mathcal{M}$ with

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

Indeed, when $C=I$ and $\mathcal{M}_D^b \subset \mathcal{M}$, the Inequalities (8) enforces that $n^{-1/2}X$ satisfies the so-called δ-Restricted Isometry Property of order D introduced by Candès and Tao [4], 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.* [5]) 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 (8) when $\mathcal{M}_D^b \subset \mathcal{M}$ with $D \geq c(\delta)n/(1+\log(p/n))$.

3.2. **Can we choose a smaller penalty?** As mentioned before, under Condition (5) the penalty $\mathrm{pen}(d)$ given by (3) 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.* [3], a slight variation of the proof of Theorem 1 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 (5)). 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 $1 \le D < n < p$ such that $p \ge e^{2/(1-\gamma)} + 1$ and $\mathcal{M}_D^b \subset \mathcal{M}$. 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 (2)*

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

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

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

4. Proofs

4.1. A concentration inequality.

Lemma 1. *Consider three integers $1 \le d \le n \le p$, a collection $V_1, \dots, V_N$ 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*

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

Then, for any $x \ge 0$

$$\mathbb{P}\left(\lambda_d^*(Z) \le 1 - \frac{\sqrt{d} + \sqrt{2\log N} + \delta_N + x}{\sqrt{n}} \right) \le \mathbb{P}\left(\mathcal{N} \ge x\right) \le e^{-x^2/2}, \tag{9}$$

where $\mathcal{N}$ has a standard Gaussian distribution and $\delta_N = \left(N\sqrt{8\log N}\right)^{-1}$.

Similarly, for any $x \ge 0$

$$\mathbb{P}\left(\sup_{v \in V_1 \cup \dots \cup V_N} \frac{\|Zv\|_n}{\|v\|_{p\times 1}} \ge 1 + \frac{\sqrt{d} + \sqrt{2\log N} + \delta_N + x}{\sqrt{n}} \right) \le \mathbb{P}\left(\mathcal{N} \ge x\right) \le e^{-x^2/2}. \tag{10}$$

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

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

To get (9), we need to bound $\mathbb{E}\left(\lambda_d^*(Z)\right)$ from below. For $i = 1, \dots, N$, we set

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

We get from [6] 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

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

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

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

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

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

This concludes the proof of (9) and the proof of (10) is similar.

4.2. **Proof of Theorem 1.** 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 $\{1,\dots,j-1,j+1,\dots,p\}\times\{j\}$ with cardinality D. Then, for any $j = 1,\dots,p$

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

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

$$\mathbb{E}_1^{(j)} \leq \lambda_0^{-2}\, \mathbb{E}\left[\|X(\hat\theta^{(j)} - \theta^{(j)})\|_n^2\right]. \tag{11}$$

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,

$$\mathbb{E}_2^{(j)} \leq \|C^{1/2}\theta^{(j)}\|^2\,\frac{9\,\|C^{1/2}\theta^{(j)}\|^2 + 4\sigma_j^2}{\lambda_0^2\lambda^- pT_n^2}. \tag{12}$$

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

$$\mathbb{E}_3^{(j)} \leq \|C^{1/2}\theta^{(j)}\|^2\exp(-n/5). \tag{13}$$

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

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

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 (5) 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

$$\mathbb{E}_4^{(j)} \leq 2\left(\|C^{1/2}\theta^{(j)}\|^2 + \lambda^+ pT_n^2\right)\exp\left(-n(1-\sqrt{\eta})\eta/2\right). \tag{14}$$

Conclusion. Putting together the bounds (11) to (14), we obtain

$$\mathbb{E}\left[\|C^{1/2}(\tilde\theta-\theta)\|^2\right]=\sum_{j=1}^p\mathbb{E}\left[\|C^{1/2}(\tilde\theta-\theta)\|^2\right]\le\lambda_0^{-2}\,\mathbb{E}\left[\|X(\hat\theta-\theta)\|_n^2\right]+R_n(\eta,C,\theta) \tag{15}$$

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

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

$$\sum_{j=1}^p\left(\|X^{(j)}-X\hat\theta_{\hat m}^{(j)}\|^2\times\left(1+\frac{\mathrm{pen}(|\hat m_j|)}{n-|\hat m_j|}\right)\right)\le\sum_{j=1}^p\left(\|X^{(j)}-X\hat\theta_{m^*}^{(j)}\|^2\times\left(1+\frac{\mathrm{pen}(|m_j^*|)}{n-|m_j^*|}\right)\right)$$

and following the same lines as in the proof of Theorem 2 in Baraud *et al.* [3] we obtain for any $K>1$

$$\begin{aligned}\frac{K-1}{K}&\sum_{j=1}^p\|X(\hat\theta^{(j)}-\theta^{(j)})\|_n^2\\&\le\sum_{j=1}^p\left[\|X(\theta^{(j)}-\bar\theta_{m^*}^{(j)})\|_n^2+R_{m^*}^{(j)}+\frac{\sigma_j^2}{n}\left(KU_{\hat m_j}^{(j)}-\mathrm{pen}(|\hat m_j|)\frac{V_{\hat m_j}^{(j)}}{n-|\hat m_j|}\right)\right],\end{aligned}$$

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

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

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.* [3], we get

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

where $\mathcal{M}_j=\{m_j,\ m\in\mathcal{M}\}$. The choice (3) of the penalty ensures that the last term is equal to $K\sum_{j=1}^p\sigma_j^2/n$. We also note that $\|X(\theta^{(j)}-\bar\theta_{m^*}^{(j)})\|_n^2\le\|X(\theta^{(j)}-\theta_{m^*}^{(j)})\|_n^2$ 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

$$\begin{aligned}
&\frac{K-1}{K}\mathbb{E}\left[\|X(\hat{\theta}-\theta)\|_n^2\right] \\
&\quad\leq \|C^{1/2}(\theta-\theta_{m^*})\|^2 + \sum_{j=1}^{p}\mathrm{pen}(|m_j^*|)\left[\frac{\|C^{1/2}(\theta^{(j)}-\theta^{(j)}_{m^*})\|^2}{n-|m_j^*|}+\frac{\sigma_j^2}{n}\right] + K\sum_{j=1}^{p}\frac{\sigma_j^2}{n} \\
&\quad\leq \|C^{1/2}(\theta-\theta_{m^*})\|^2\left(1+\frac{\mathrm{pen}(D)}{n-D}\right) + \sum_{j=1}^{p}(\mathrm{pen}(|m_j|)+K)\frac{\sigma_j^2}{n} \qquad (16)
\end{aligned}$$

c. Conclusion. The bound (16) is true for any m^*, so combined with (15) it gives (6).

4.3. **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 $Z\hat{\theta}_s = \mathrm{Proj}_{ZV_s}(\varepsilon)$, $\hat{s}_d$ for the set of cardinality d such that

$$\|Z\hat{\theta}_{\hat{s}_d}\|^2 = \max_{|s|=d}\|Z\hat{\theta}_s\|^2. \qquad (17)$$

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 we only give here a sketch of it. For the details, we refer to Section 4.4.

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 $\mathcal{M}_D^b \subset \mathcal{M}$. 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 $\tilde m \in \mathcal{M}_D^b$, the first part of Proposition 1 follows.

When $\mathcal{M} = \mathcal{M}_D^c$, 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 $3(p-1)^{-1} + 2\exp(-n\gamma^2/512)$.

4.4. **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

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

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

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

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 $\|\mathrm{Proj}_{V_s}Y\|^2 \leq |s|\max_{i=1,\dots,p}(Y_i^2)$, 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

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

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.* [3]. 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, C. Giraud and S. Huet. *Gaussian model selection with unknown variance.* Preprint (2007). http://arxiv.org/abs/math/0701250v1

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

[5] 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

[6] 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.

[7] 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

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

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

[10] 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

[11] H. Kishino and P.J. Waddell. *Correspondence analysis of genes and tissue types and finding genetic links from microarray data.* Genome Informatics **11** (2000), 8395.

[12] N. Meinshausen and P. Bühlmann. *High dimensional graphs and variable selection with the lasso.* Annals of Statistics **34** (2006), 1436–1462.
[13] J. Schäfer and K. Strimmer. *An empirical bayes approach to inferring large-scale gene association networks.* Bioinformatics **21** (2005), 754–764.
[14] N. Verzelen and F. Villers. *Test of neighborhood for Gaussian graphical models.* Preprint (2007).
[15] A. Wille and P. Bühlmann. *Low-order conditional independence graphs for inferring genetic networks.* Stat. Appl. Genet. Mol. Biol. **5** (2006).
[16] M. Yuan and Y. Lin *Model selection and estimation in the Gaussian graphical model.* Biometrika **94** (2007), 19–35.

Université de Nice Sophia-Antipolis, Laboratoire J-A Dieudonné, Parc Valrose, 06108 Nice cedex 02

INRA, MIA 78352 Jouy-en-Josas Cedex

E-mail address: cgiraud@math.unice.fr