Block-diagonal covariance selection for high-dimensional Gaussian graphical models

Emilie Devijver, Mélina Gallopin

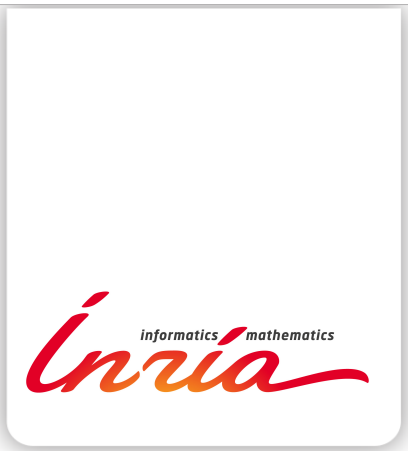
Définition

Block-diagonal covariance selection for high-dimensional Gaussian graphical models

Émilie Devijver, Mélina Gallopin

RESEARCH REPORT
N° ??
Novembre 2015
Project-Teams Select

ISSN 0249-6399 ISRN INRIA/RR--??--FR+ENG



Block-diagonal covariance selection for high-dimensional Gaussian graphical models

Émilie Devijver*, Mélina Gallopin†‡§

Project-Teams Select

Research Report n° ?? — Novembre 2015 — 25 pages

Abstract: Gaussian graphical models are widely utilized to infer and visualize networks of dependencies between continuous variables. However, inferring the graph is difficult when the sample size is small compared to the number of variables. To reduce the number of parameters to estimate in the model, we propose a non-asymptotic model selection procedure supported by strong theoretical guarantees based on an oracle inequality and a minimax lower bound. The covariance matrix of the model is approximated by a block-diagonal matrix. The structure of this matrix is detected by thresholding the sample covariance matrix, where the threshold is selected using the slope heuristic. Based on the block-diagonal structure of the covariance matrix, the estimation problem is divided into several independent problems: subsequently, the network of dependencies between variables is inferred using the graphical lasso algorithm in each block. The performance of the procedure is illustrated on simulated data. An application to a real gene expression dataset with a limited sample size is also presented: the dimension reduction allows attention to be objectively focused on interactions among smaller subsets of genes, leading to a more parsimonious and interpretable modular network.

Key-words: Network inference, graphical lasso, variable selection, non-asymptotic model selection, slope heuristic.

* Department of Mathematics and Leuven Statistics Research Center (LStat), KU Leuven, Leuven, Belgium
† Laboratoire MAP5, Université Paris Descartes and CNRS, Sorbonne Paris Cité
‡ Laboratoire de Mathématiques, UMR 8628, Bâtiment 425, Université Paris-Sud, F-91405, Orsay, France
§ INRA, UMR 1313 Génétique animale et biologie intégrative, 78352 Jouy-en-Josas, France

RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing
Campus de l'École Polytechnique
91120 Palaiseau

Block-diagonal covariance selection for high-dimensional Gaussian graphical models

Résumé :

Les modèles graphiques gaussiens permettent dâinférer et de visualiser les dépendances entre des variables. Ces modèles étant difficiles à estimer en très grande dimension, nous proposons une procédure non-asymptotique pour réduire la dimension du problème d'inférence. Cette procédure est justifiée par des résultats théoriques comportant une inégalité oracle et une borne minimax. Dans un premier temps, nous approchons la matrice de covariance par une matrice diagonale par blocs. Pour détecter la structure de cette matrice, nous seuillons la matrice de covariance empirique, le seuil étant choisi à lâaide de lâheuristique de pente. Le problème dâestimation est ainsi décomposé en plusieurs sous-problèmes indépendants : par la suite, nous estimons les dépendances entre les variables d'un même bloc à lâaide de l'algorithme du graphical lasso. Nous illustrons cette méthode sur des données simulées. Une application à un jeu de données réelles ayant un faible nombre d'échantillons est également présentée: la réduction de dimension permet de focaliser l'attention sur un nombre plus réduit de gènes et conduit à un réseau modulaire plus parcimonieux et interprétable.

Mots-clés : Inférence de réseaux, graphical lasso, sélection de variables, sélection de modèle non-asymptotic, heuristique de pente.

1 Introduction

Graphical models (Whittaker, 1990) have become a popular tool for representing conditional dependencies among variables using a graph. For Gaussian graphical models (GGMs), the edges of the corresponding graph are the non-zero coefficients of the inverse covariance matrix. Popular methods to estimate this matrix have been proposed in high-dimensional contexts (Meinshausen and Bühlmann, 2006, Banerjee et al., 2008). The graphical lasso introduced by Friedman et al. (2008) performs the estimation of the inverse covariance matrix based on an ℓ_1 penalized log-likelihood. GGMs have many potential applications for the reconstruction of networks of dependencies between variables from real omics data (Krumsiek et al., 2011, Akbani et al., 2014). Implementing and improving network reconstruction using graphical models is an area of active methodological developments (Ambroise et al., 2009, Guo et al., 2011, Allen and Liu, 2013, Tan et al., 2015).

However, these network reconstruction methods often perform poorly in so-called *ultra high-dimensional contexts* (Giraud, 2008, Verzelen, 2012), when the number of observations is much smaller than the number of variables. A small sample size is a common situation in various applications, such as in systems biology where the cost of the sequencing technologies may limit the number of available observations (Frazee et al., 2011). In practice, the network reconstruction problem is facilitated by restricting the analysis to a subset of variables, based on external knowledge and prior studies of the data (Ambroise et al., 2009, Yin and Li, 2011). When no external knowledge is available, only the most variable features are typically kept in the analysis (Guo et al., 2011, Allen and Liu, 2013). Choosing the appropriate subset of variables to focus on is a key step in reducing the model dimension and the number of parameters to estimate, but no procedure is clearly established to perform this selection in high-dimensional settings.

In the context of graphical lasso estimation, Mazumder and Hastie (2012) and Witten et al. (2011) have noticed that the block-diagonal structure of the graphical lasso solution is totally determined by the block-diagonal structure of the thresholded empirical covariance matrix. The graphical lasso estimation for a given level of regularization λ can be decomposed into two steps: first, the absolute value of the sample covariance matrix is thresholded at λ to detect subsets of connected variables; then the graphical lasso problem is divided into subproblems and solved in each subset independently using the same regularization parameter λ. This decomposition is of great interest to reduce the number of parameters to estimate for a fixed level of regularization. It has been exploited for large-scale problems (Zhao et al., 2012) and for joint graphical lasso estimations (Danaher et al., 2014). Tan et al. (2015) provided an adaptation of this two-step decomposition: the block-diagonal structure of the covariance matrix is detected using a hierarchical clustering of variables based on the sample covariance matrix. A leave-one-out algorithm recasts the unsupervised clustering into a supervised one and selects the partition of variables giving the smallest mean square error. Tan et al. (2015) also comment on the asymptotic properties of this algorithm. However, for high-dimensional problems, methods are needed to detect the best block structure of the covariance matrix (*i.e.* the value of the thresholding parameter λ) to divide the GGM estimation into several subproblems.

In this paper, we propose a non-asymptotic procedure to detect the block-diagonal structure of the covariance matrix. Pavlenko et al. (2012) provided a method to detect this structure for high-dimensional supervised classification that is supported by asymptotic guarantees. Hyodo et al. (2015) proposed tests to perform this detection and derived consistency for their method when the number of variables and the sample size tend to infinity. In our procedure, we recast the detection problem into a model selection problem and choose the best model among a collection of multivariate distributions with block-diagonal covariance matrices. This method is based on the slope heuristic developed by Birgé and Massart (2007), and is easy to implement in practice (Baudry et al., 2012). Unlike other methods to detect the appropriate block-diagonal covariance matrix (Pavlenko et al., 2012, Tan et al., 2015, Hyodo et al., 2015), our procedure is non-asymptotic and offers strong theoretical guarantees when the number of observations is limited, which is of great interest for many real applications. We prove that our estimator is approximately

minimax.

The paper is organized as follows. In Section 2, after providing basic notations and definitions, the non-asymptotic method to detect the block-diagonal structure of the GGM is presented. Section 3 details theoretical results supporting our model selection criterion. In particular, an oracle inequality upper bounds the risk between the true model and the model selected among the model collection, and a minimax lower bound guarantees that the non-asymptotic procedure has an optimal rate of convergence. Section 4 investigates the numerical performance of our method in a simulation study. Section 5 illustrates our procedure on a real gene expression RNA-seq dataset with a limited sample size. After a short discussion, all proofs are provided in Section 7.

2 A method to detect block-diagonal covariance structure

Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ be a sample in $\mathbb{R}^p$ from a multivariate normal distribution with density $\phi_p(0, \Sigma)$ where $\Sigma_{j,j} = 1$ for all $j \in \{1, \ldots, p\}$. Let S be the empirical covariance matrix associated with this sample. Our goal is to detect the optimal block-diagonal structure of the covariance matrix Σ, *i.e.* the optimal partition of variables into blocks. Let $B = (B_1, \ldots, B_K)$ be the partition of variables into K blocks where K is the number of blocks, B_k the subset of variables in block k, and p_k the number of variables in block k. We denote by Σ_B the corresponding block-diagonal covariance matrix where each block on the diagonal is denoted Σ^k for $k \in \{1, \ldots, K\}$. We denote by $f_B = \phi_p(0, \Sigma_B)$ the density of the multivariate distribution. The set of densities with block-diagonal covariance matrix with structure B is:

$$F_B = \{f_B \text{ with } \Sigma_B \in S_B\} \tag{1}$$

$$S_B = \left\{ \Sigma_B \in \mathbb{S}_p^{++}(\mathbb{R}) \middle| \Sigma_B = P_\sigma \begin{pmatrix} \Sigma^1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma^K \end{pmatrix} P_\sigma^{-1}, P_\sigma \text{ a permutation matrix}, \right.$$
$$\left. \Sigma^k \in \mathbb{S}_{p_k}^{++}(\mathbb{R}) \text{ for } k \in \{1, \ldots, K\} \right\}. \tag{2}$$

The dimension of the model F_B is $D_B = \sum_{k=1}^K p_k(p_k - 1)/2$. We denote by $\hat{f}_B$ the maximum likelihood estimator of the model F_B where parameters in each block k are estimated using the sample covariance matrix of the dataset restricted to variables in block k: $\hat{\Sigma}^k = S_{|k}$.

We consider $\mathscr{B}$ the set of all possible partitions of variables. In theory, we would like to consider the following model collection $\mathscr{F} = (F_B)_{B \in \mathscr{B}}$. However, the set $\mathscr{B}$ is large: there are $\sum_{k=1}^p \text{Stir}(p,k)$ possible partitions where $\text{Stir}(p,k)$ denotes the Stirling number of the second kind. An exhaustive exploration of the set $\mathscr{B}$ is not possible even for a moderate number of variables p. We restrict our attention to the sub-collection:

$$\mathscr{B}^\Lambda = (B_\lambda)_{\lambda \in \Lambda} \tag{3}$$

of $\mathscr{B}$ where B_λ is the partition of variables corresponding to the block-diagonal structure of the matrix $E_\lambda = (\mathbf{1}_{\{|S_{j,j'}| > \lambda\}})_{jj'}$, the thresholded absolute value of the sample covariance matrix. Recall that Mazumder and Hastie (2012) have proved that this method is equivalent to the graphical Lasso for detecting the block structure. Note that the data is scaled so that the set of thresholds $\Lambda \subset [0,1]$ covers all possible partitions derived from E_λ.

Once we have constructed the model collection $\mathscr{F}_\Lambda = (F_B)_{B \in \mathscr{B}^\Lambda}$, we select the optimal model among this collection, i.e. the optimal partition of variables into blocks. In our context, the number of obser-

vations n is limited. For this reason, we consider a non-asymptotic model selection based on the slope heuristic, developed by Birgé and Massart (2007). This heuristic leads to the following criterion:

$$\hat{B} = \underset{B \in \mathscr{B}^{\Lambda}}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(\hat{f}_B(\mathbf{y}_i)) + \operatorname{pen}(B) \right\}, \tag{4}$$
$$\operatorname{pen}(B) = \kappa D_B,$$

where $\hat{f}_B$ is the maximum likelihood estimator of the model F_B, D_B the dimension of F_B and κ a coefficient to calibrate.

Baudry et al. (2012) have provided practical tools to implement the slope heuristic developed by Birgé and Massart (2007). One calibration method is the *Slope Heuristic Dimension Jump* (SHDJ): the optimal coefficient κ_{opt} is approximated by twice the minimal coefficient κ_{min}, where κ_{min} corresponds to the largest dimension jump on the graph representing the model dimension as a function of the coefficient κ. Another method is the *Slope Heuristic Robust Regression* (SHRR): the coefficient κ_{opt} is approximated by twice κ_{min}, where κ_{min} corresponds to the slope of a robust regression performed between the log-likehood and the model dimension for complex models. The two methods are derived from the same heuristic and they offer two different visual checks of the adequacy of the model selection procedure to the data. They should select the same model. Note that the detection of the optimal B is easy to implement in practice and does not rely on heavy computation such as cross-validation techniques.

Once we have detected the optimal block-diagonal structure of the GGM, network inference is performed independently in each block using the graphical lasso introduced by Friedman et al. (2008). Our procedure has been implemented in a R package `shock` available on github (`https://github.com/Gallopin/shock`).

3 Theoretical results for non-asymptotic model selection

Model selection based on the slope heuristic with calibration of the κ coefficients by dimension jump (SHDJ) or robust regression (SHRR) have been proven to be effective in a variety of practical situations. For example, Rau et al. (2015) select the number of components in Poisson mixture models on RNA-seq gene expression data using the slope heuristic. Bouveyron et al. (2015) select the number of components in discriminative functional mixture models on data describing bike sharing systems using the slope heuristic. However, they did not provide any theoretical justification for their procedures.

In contrast, we do provide theoretical justification for our criterion based on an oracle inequality. Lebarbier (2005) have provided theoretical justification based on an oracle inequality for model selection in multiple change point detection, and Maugis and Michel (2011) for variable selection in mixture models. However, few papers provide a minimax lower bound, which we do have. Remark that the theoretical justification of the slope heuristic has encountered several technical difficulties. The existence of minimal penalties is proved in heteroscedastic regression with fixed design (Birgé and Massart (2007) and Baraud et al. (2009)), and for homoscedastic regression with fixed design (Arlot and Massart (2009)).

For our block-diagonal structure detection procedure, we prove an oracle inequality for a penalty proportional to the dimension (up to a logarithm term) and a lower bound of the risk between the true model and the model selected among the model collection. This ensures that the selected model is close to the oracle, the best one in estimation among our collection. Both inequalities guarantee that our model selection procedure has an optimal rate of convergence, which is a strong theoretical result. Note that these results are non-asymptotical, which means that they hold for a fixed sample size n.

To state the theorem, we recall the definition of the Hellinger distance between two densities f and g

defined on $\mathbb{R}^p$,

$$d_H^2(f,g) = \frac{1}{2}\int_{\mathbb{R}^p}(\sqrt{f(x)}-\sqrt{g(x)})^2 dx = 1 - \int_{\mathbb{R}^p}\sqrt{f(x)g(x)}dx,$$

and the Kullback-Leibler divergence between two densities f and g defined on $\mathbb{R}^p$,

$$KL(f,g) = \int_{\mathbb{R}^p}\log\left(\frac{f(x)}{g(x)}\right)f(x)dx.$$

In order to properly define the penalty term used in equation (4) to select the best partition of variables B, we work with the following model collection:

$$\mathscr{F}^{\text{bound}} = (F_B^{\text{bound}})_{B\in\mathscr{B}} \tag{5}$$

$$F_B^{\text{bound}} = \left\{\phi(0,\Sigma_B) \in F_B, \Sigma_B \in S_B^{\text{bound}}\right\} \tag{6}$$

$$\begin{aligned} S_B^{\text{bound}} = \{\Sigma_B \in S_B | e_m \leq \min(\Sigma_B) \leq \max(\Sigma_B) \leq e_M, \\ \lambda_m \leq \Lambda_{\min}(\Sigma_B) \leq \Lambda_{\max}(\Sigma_B) \leq \lambda_M\}, \end{aligned}$$

where $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ are the smallest and the largest eigenvalues of the matrix A.

The model collection (5) is defined such that covariance matrices have bounded coefficients, which is useful for constructing a discretization of this space. If the matrix has bounded coefficients, we can prove that it has bounded eigenvalues. Nevertheless, to simplify the reading, we denote by λ_m and λ_M bounds on eigenvalues. In the following, we denote by $Adj(\Sigma)$ the adjacency matrix associated to the covariance matrix Σ.

Theorem 3.1 *Let $\mathbf{y} = (\mathbf{y}_1,\ldots,\mathbf{y}_n)$ be the observations, arising from a density f^*. Consider the model collection $\mathscr{F}^{bound}$ defined in* (5)*. Suppose that there exists an absolute constant $\kappa' > 0$ such that for every partition B in the set of all possible partitions of variables $\mathscr{B}$,*

$$\text{pen}(B) \geq \kappa'\frac{D_B}{n}\left[2c^2 + \rho\log\left(\frac{1}{D_B(\frac{D_B}{n}c^2\wedge 1)}\right) + (1\vee\tau)\frac{p}{D_B}\log\left(\frac{0.792p}{\log(p+1)}\right)\right],$$

where c is an absolute constant, depending only on the model collection. Let $\hat{f}_B$ be the maximum likelihood estimator, $\mathscr{B}^\Lambda \subset \mathscr{B}$ as defined in (3)*, and $\hat{B}$ selected as follows:*

$$\hat{B} = \underset{B\in\mathscr{B}^\Lambda}{\text{argmin}}\left\{-\frac{1}{n}\sum_{i=1}^n \log(\hat{f}_B(\mathbf{y}_i)) + \text{pen}(B)\right\}.$$

Then, $\hat{f}_{\hat{B}}$ satisfies:

$$\mathbb{E}(d_H^2(f^*,\hat{f}_{\hat{B}})) \leq C\mathbb{E}\left(\inf_{B\in\mathscr{B}^\Lambda}\left(\inf_{t\in F_B^{bound}} KL(f^*,t) + \text{pen}(B)\right) + (1\vee\tau)\frac{1}{n}\right) \tag{7}$$

for some absolute constant C.

This non-asymptotic result is consistent with the point of view adopted in this work where the number of observations n is limited. The proof is presented in Appendix 7.2. This theorem is deduced from an adaptation for a random sub-collection of the whole model collection of a general model selection theorem for maximum likelihood estimator developed by Massart (2007). This adaptation is proved in Appendix 7.2.1. To apply our theorem, the main assumptions to satisfy are the control of the bracketing entropy of each model in the whole model collection and the construction of weights for each model to

control the model collection complexity. Remark that the control of the bracketing entropy is a classical tool to bound the Hellinger risk of the maximum likelihood estimator, and has already been done for Gaussian densities in Maugis and Michel (2011) and Genovese and Wasserman (2000).

Theorem 3.1 provides a lower bound for the penalty, which ensures a good model selection by penalized criterion: the model selected is as good as possible among the model collection. The only assumption made to state Theorem 3.1 is a classical one: we work with bounded parameters for each model as detailed in (5). Every constant involved in (7) depends on those bounds. Even if the bounds are not tractable in practice, this assumption is plausible. To guarantee a good model selection procedure, we need to assume that the true density of the data is not too far from the constructed model collection. Since a covariance matrix can always be considered to be a block-diagonal matrix, with possibly a single block, the block-diagonal covariance matrix assumption is not a strong one.

To complete this analysis, we provide a minimax lower bound for the risk between the true model and the model selected among the model collection. For the lower bound of the risk, some results have been previously obtained by Bickel and Levina (2008) and Cai et al. (2010). To obtain our lower bound, we use the lemma developed in Birgé (2005) in conjunction with a discretization of the model collection space, already constructed for the oracle inequality.

Theorem 3.2 *Let* $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ *be the observations, coming from a density* f^*. *Consider the model collection* $\mathscr{F}^{bound}$ *defined in* (5)*, and* D_B *the dimension of the model* F_B^{bound} *for each* $B \in \mathscr{B}$. *Let* $\hat{f}_B$ *being the maximum likelihood estimator for the model indexed by B. Then, for all* $B \in \mathscr{B}$, *there exists absolute constants* $C_1 > 0$ *and* $C_2 > 0$ *such that:*

$$\inf_{\hat{f}_B} \sup_{f \in F_B^{bound}} \mathbb{E}(d_H^2(\hat{f}_B, f)) \geq C_1 \frac{D_B}{n}(1 + \log\left(\frac{C_2}{D_B^2}\right)). \tag{8}$$

This theorem is proved in Appendix 7.3. Again, this result does not rely on strong assumptions, and the constants involved are explicit. It is also a non-asymptotic result.

This minimax lower bound obviously shows that since the estimator satisfies to (7) it is simultaneously approximately minimax on each set F_B^{bound} for every $B \in \mathscr{B}$. Theorem 3.2 and Theorem 3.1 lead to the use of the slope heuristic with a penalty proportional to the dimension to select a model among the collection.

Nevertheless, as typically the case, constants are higher in theory than needed (and not always tractable), and we prefer to compute constants from the dataset in practice using the `capushe` package developed in Baudry et al. (2012).

4 Simulation study

We simulate n observations from a $p-$multivariate normal distribution with a null mean and a block-diagonal covariance matrix Σ_B as defined in Section 2. We fix the number of variables $p = 100$ and the sample size $n = 70$. For the partition on variable $B^\star$: we vary the number of blocks among $K^\star \in \{1, 15\}$. For each block indexed by k, we design the Σ^k matrix as done in Giraud et al. (2012): $\Sigma^k = TT^t + D$ where T is a random lower triangular matrix with values drawn from a uniform distribution between -1 and 1, and D is a diagonal matrix designed to prevent Σ^k from having eigenvalues that are too small.

4.1 Block-diagonal covariance structure detection

First, we investigate the ability to recover the simulated partition of variables $B^\star$ using the hierarchical clustering from Tan et al. (2015), the non-asymptotic model selection based on the slope heuristic dimension jump (SHDJ) and the slope heuristic robust regression (SHRR). Selected partitions for each method are compared with the simulated partition $B^\star$ using the Adjusted Rand Index (Hubert and Arabie, 1985).

Block-diagonal covariance matrix Σ with $K^\star = 15$ blocks

We fix the design of the block-diagonal covariance matrix Σ with $K^\star = 15$ blocks of approximately equal sizes. Illustrations of the calibration of the penalty coefficient κ are presented in Figure 1. Both calibration methods yield the same results.

In addition, we compare the partition selection methods with an average linkage hierarchical clustering with $K = 15$ as proposed in the cluster graphical lasso (Tan et al., 2015). Figure 2 displays the ARI computed over 100 replicated datasets with $p = 100$ variables, $n = 70$ samples and $K = 15$ blocks. Despite the fact that the partition with the hierarchical clustering takes as an input parameter the true number of clusters ($K = 15$), the ARI for the hierarchical clustering is lower than the ARI for the two slope heuristic based methods (SHRR and SHDJ) which do not need to specify the number of clusters K in advance.

Full covariance matrix Σ with $K^\star = 1$ block

We simulate $n = 70$ observations from a multivariate normal distribution with a null mean and full covariance matrix Σ. The corresponding network of conditional dependencies is almost a clique. As anticipated, solving the graphical lasso problem in this context is too ambitious, as proved by Verzelen (2012): inferring the true network requires the estimation of $D = 4950$ parameters, with only $n \times p = 700$ data points. Illustrations of the calibration of the κ coefficient are displayed in Figure 3. In contrast with Figure 1, the biggest dimension jump in the graph representing the model dimension as a function of the coefficient κ is not clear. Moreover, the partition selected by dimension jump and robust regression are not equivalent. In this context, no relevant block-diagonal structure is detected.

4.2 Downstream network inference performance

To illustrate the potential advantages of prior block-diagonal covariance structure detection, we compare several strategies for network inference on data simulated under a multivariate normal distribution with a null mean and a block-diagonal covariance matrix Σ with $K^\star = 15$ blocks of approximate equal sizes.

To perform network inference, we use the graphical lasso algorithm proposed in Friedman et al. (2008) and implemented in the R package `glasso`, version 1.7. We compare the following strategies:

1. Glasso: We perform network inference using the graphical lasso on all variables, with regularization parameter ρ chosen using the following BIC^{net} criterion:

$$\text{BIC}^{\text{net}}(\rho) = \frac{n}{2}\left(\log\det\hat{\Theta}^{(\rho)} - \text{trace}\left(S\hat{\Theta}^{(\rho)}\right)\right) - \frac{\log(n)}{2}\text{df}\hat{\Theta}^{(\rho)}; \tag{9}$$

where $\hat{\Theta}^{(\rho)}$ the solution of the graphical lasso with regularization parameter ρ, S is the sample covariance matrix, and df the degrees of freedom.

2. CGL: We perform network inference using the cluster graphical lasso proposed in Tan et al. (2015). First, the partition of variables is detected using an average linkage hierarchical clustering with $K = 15$ clusters. Note that we set the number of clusters to the true number $K^\star$. Subsequently, the regularization parameters in each graphical lasso problem $\rho_1, \ldots, \rho_{K^\star}$ are chosen from Corollary 3 of Tan et al. (2015): the inferred network in each block must be as sparse as possible while still remaining a single connected component.

3. Inference on partitions based on model selection: First, we detect the partition using the two variants of our non asymptotic model selection (SHRR ou SHDJ).

(a) SHRR: The partition $\hat{B}_{\text{SHRR}}$ is selected using the *Slope Heuristic Robust Regression*.

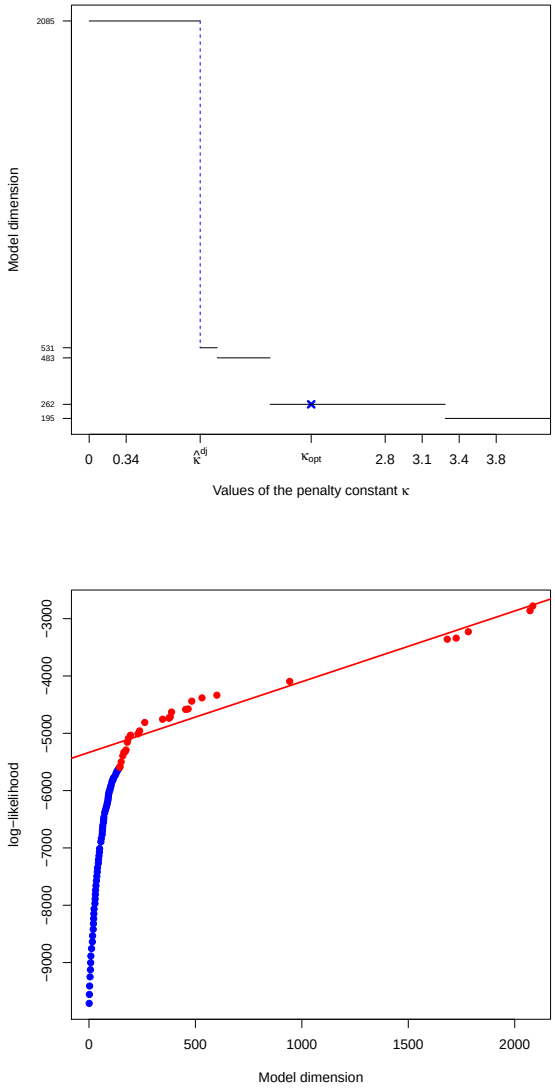


Figure 1: Calibration of the κ coefficient on a dataset simulated under a multivariate normal distribution with a block-diagonal covariance matrix Σ_B with $K^\star = 15$ blocks, $p = 100$, $n = 70$. Calibration by dimension jump (left): the dimension of the model is represented as a function of the κ coefficient. Based on the slope heuristic, the largest jump (dotted line) corresponds to the minimal coefficient κ_{min}. The optimal penalty (cross) is twice the minimal penalty. Calibration by robust regression (right): the log-likehood of the model is represented as a function of the model dimension. Based on the slope heuristic, the slope of the regression (line) between the log-likehood and the model dimension for complex models corresponds to the minimal coefficient κ_{min}. The optimal penalty is twice the minimal penalty.

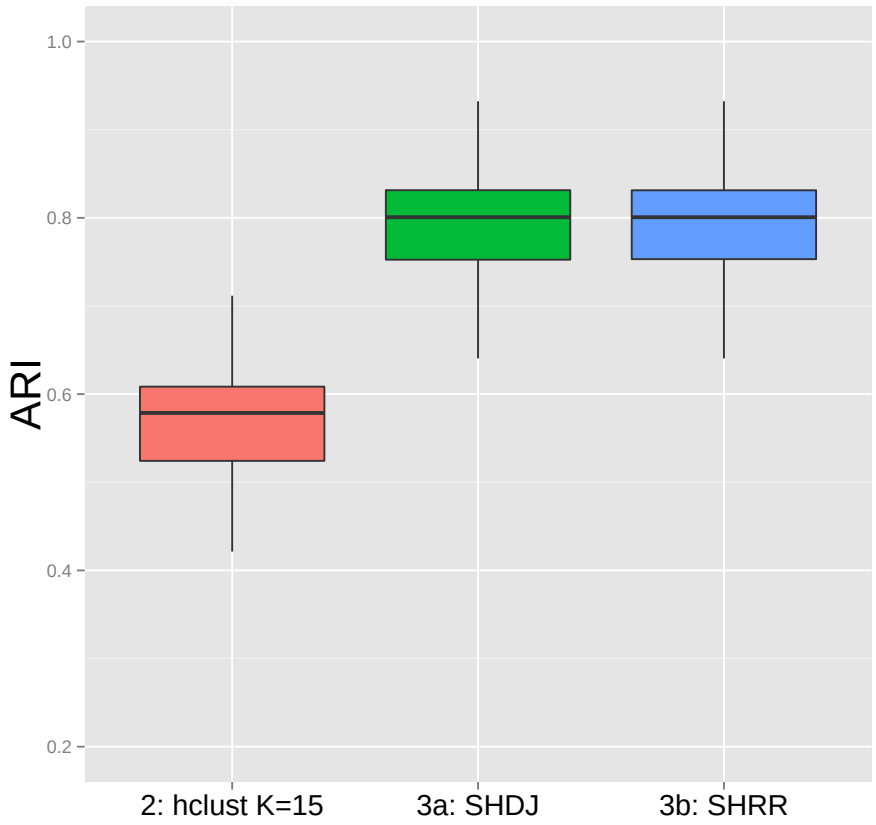


Figure 2: ARI between the simulated partition and the partitions selected by slope heuristic dimension jump (SHDJ), slope heuristic robust regression (SHRR) and by average hierarchical clustering with $K = 15$ clusters. The ARI are computed over 100 replicated datasets simulated under a multivariate normal distribution with block-diagonal covariance matrix with $K = 15$ blocks, $p = 100$ variables and $n = 70$ observations.

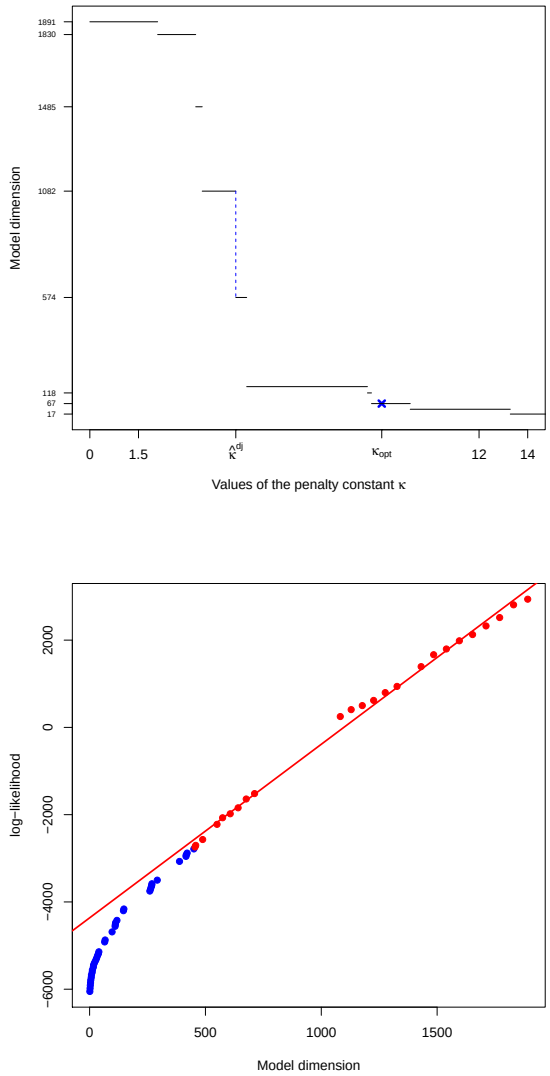


Figure 3: Calibration of the κ coefficient on a dataset simulated under a multivariate normal distribution with a full covariance matrix with one $K^\star = 1$ and $p = 100$, $n = 70$. Calibration by robust regression (left) and by dimension jump (right): in this extreme setting, no clear linear tendency (line) between the log-likelihood and the model complexity for complex models is observed and the largest jump (dotted line) is unclear.

(b) SHDJ: The partition $\hat{B}_{\text{SHDJ}}$ is detected using the *Slope Heuristic Dimension Jump*.

Subsequently, the regularization parameters in each graphical lasso problem $\rho_1, \ldots, \rho_{\hat{K}}$ are chosen using the BIC^{net} criterion:

$$\text{BIC}^{\text{net}}(\rho_k) = \frac{n}{2}\left(\log\det\hat{\Theta}^{(\rho_k)} - \text{trace}\left(S_{|k}\hat{\Theta}^{(\rho_k)}\right)\right) - \frac{\log(n)}{2}\text{df}\hat{\Theta}^{(\rho_k)}, \tag{10}$$

where $\hat{\Theta}^{(\rho_k)}$ is the solution of the graphical lasso problem restricted to the variables in block k, $S_{|k}$ is the sample covariance matrix on variables belonging to the block k and df the corresponding degrees of freedom.

4. Inference on the true partition of variables (truePart): First, we set the partition of variables to the true partition $B^\star$. Then, the regularization parameters in each graphical lasso problem $\rho_1, \ldots, \rho_{K^\star}$ are chosen using the BIC^{net} criterion (10).

We compare the performance of the five methods using the sensitivity ($SENS = TP/(TP+FN)$), the specificity ($SPEC = TN/(TN+FP)$) and the False Discovery Rate (FDR) ($FDR = FP/(TP+FP)$) where TN, TP, FN, FP are respectively the number of true negative, true positive, false negative, false positive dependencies detected. A network inference procedure is a compromise between sensitivity and specificity: we are looking for a high sensitivity, which measures the proportion of dependencies (presence of edges) that are correctly identified, and a high specificity, which measures the proportion of independencies (absence of edges) that are correctly identified. The False Discovery Rate is the proportion of dependencies wrongly detected.

As expected, the true partition strategy (truePart) performs the best: based on the true partition of variables, the network inference problem is easier because we solve problems of smaller dimension. The proposed strategies, based on the SHRR and SHDJ partitions, improve network inference compared to a simple graphical lasso on the set of all variables (glasso) or compared to the cluster graphical lasso (CGL), as illustrated in Figure 4.

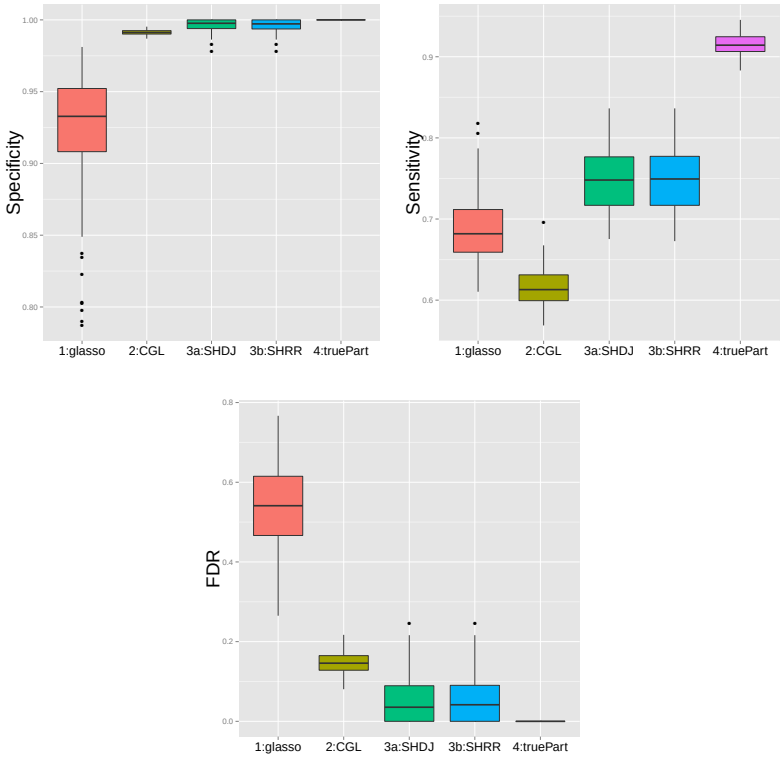


Figure 4: Performance of network inference methods (glasso: graphical lasso on the set of all variables, CGL: cluster graphical lasso, BIC: network inference based on the partition of variables $\hat{B}_{\text{BIC}}$, SSHR: network inference based on the partition of variables $\hat{B}_{\text{SHRR}}$, SHDJ: network inference based on the partition of variables $\hat{B}_{\text{SHDJ}}$ and truePart: network inference based on the partition of variables $B^\star$) measured by the sensitivity (SENS), the specificity (SPEC) and the False Discovery Rate (FDR) of the inferred graph over 100 replicated datasets simulated under a $p-$multivariate normal distribution with a null mean $\mathbf{0}$ and a block-diagonal covariance matrix $\Sigma_{B^\star}$ with $p = 100$, $K = 15$, $n = 70$ and clusters of approximately equal sizes.

5 Real data analysis

Pickrell *et al.* analyzed transcriptome expression variation from 69 lymphoblastoid cell lines derived from unrelated Nigerian individuals (Pickrell, 2010). The expression of 52580 genes across 69 observations was measured using RNA-seq. The data is extracted from the Recount database (Frazee et al., 2011). After filtering weakly expressed genes using the `HTSFilter` package (Rau et al., 2013), we identified the 200 most variable genes among the 9191 remaining genes, and restrict our attention to this set of genes for the following network inference analysis.

First, we select the partition $\hat{B}$ using model selection as described in equation (4). The log-likelihood increases with the number of parameters to be estimated in the model as displayed in Figure 5. We notice a linear tendency in the relationship between the log-likelihood and the model dimension for complex models (points corresponding to a model dimension higher than 500). This suggests that the use of the slope heuristic is appropriate for selecting a partition $\hat{B}$. The model selected by SHDJ and by SHRR described in Section 2 are the same. The number of blocks detected is $\hat{K}_{\mathrm{SH}} = 150$ and the corresponding model dimension is $D_{\hat{B}_{\mathrm{SH}}} = 283$. The partition $\hat{B}_{\mathrm{SH}}$ yields 4 blocks of size $18, 13, 8$ and 5, 4 blocks of size 3, 2 blocks of size 2 and 140 blocks of size 1. The partition selected by the Slope Heuristic offers a drastic reduction of the number of parameters to infer, as compared with the graphical lasso performed on the full set of variables, which corresponds to a total of $D = 19900$ parameters to estimate.

The networks within each cluster of variables are inferred using the graphical lasso algorithm of Friedman (Friedman et al., 2008) implemented in the `glasso` package, version 1.7. The regularization parameter for the graphical lasso on the set of all variables is chosen using the $\mathrm{BIC}^{\mathrm{net}}$ criterion (9). The model inferred based on partition $\hat{B}_{\mathrm{SH}}$ is more parsimonious and easier to interpret than the model inferred on the full set of variables. An illustration of inferred networks in the four largest connected components of the partition $\hat{B}_{\mathrm{SH}}$ are displayed on Figure 6. These four networks might be good candidates for further study.

6 Discussion

In this paper, we have proposed a non-asymptotic procedure to detect a block diagonal structure for covariance matrices in GGMs. It substantially reduces the number of parameters to estimate in the model. Although GGMs are widely used in practice, limited sample sizes typically force the user to restrict the number of variables to be included in the model. Usually, this restriction is performed manually, for instance, based on prior knowledge on the role of variables. Here, we propose an automatic procedure to select relevant subsets of variables based on the data. Therefore, our procedure is of great practical interest to estimate parameters in GGMs when the sample size is much smaller than the number of parameters to estimate.

The methodology we propose is easy to implement in practice and fast to compute. The calibration of the κ coefficient by robust regression and dimension jump encounters no particular difficulty. Moreover, graphical representation of the log-likelihood or the model dimension can indicate if the block diagonal assumption is incorrect.

Our method uses a model selection criterion to detect a block diagonal structure. We propose a non-asymptotic approach supported by strong theoretical results. Indeed, we obtain an oracle inequality, which ensures that the model we select by a penalized criterion is close to the oracle, *i.e.* the best model among our family. Moreover, we obtain a minimax lower bound of the risk between the true model and the model selected among the model collection, which ensures that the estimator is not overly penalized.

Finally, results on real data attest that the slope heuristic adapts the model selection to the bias induced by the model collection, and selects a sparse and interpretable model.

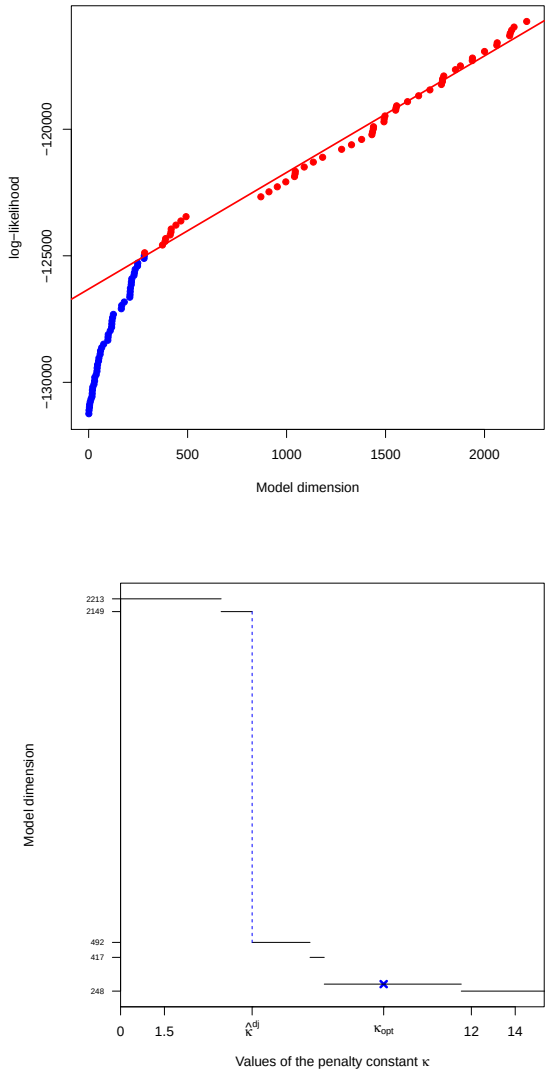


Figure 5: Calibration of the κ coefficient on the 200 most variable genes extracted from the Pickrell (2010) dataset. Calibration by robust regression (left): the log-likehood of the model is represented as a function of the model dimension. Based on the slope heuristic, the slope of the regression (line) between the log-likehood and the model dimension for complex models corresponds to the minimal coefficient κ_{min}. The optimal penalty is twice the minimal penalty. Calibration by dimension jump (right): the dimension of the model is represented as a function of the κ coefficient. Based on the slope heuristic, the largest jump corresponds to the minimal coefficient κ_{min}. In both cases, the optimal penalty is twice the minimal penalty.

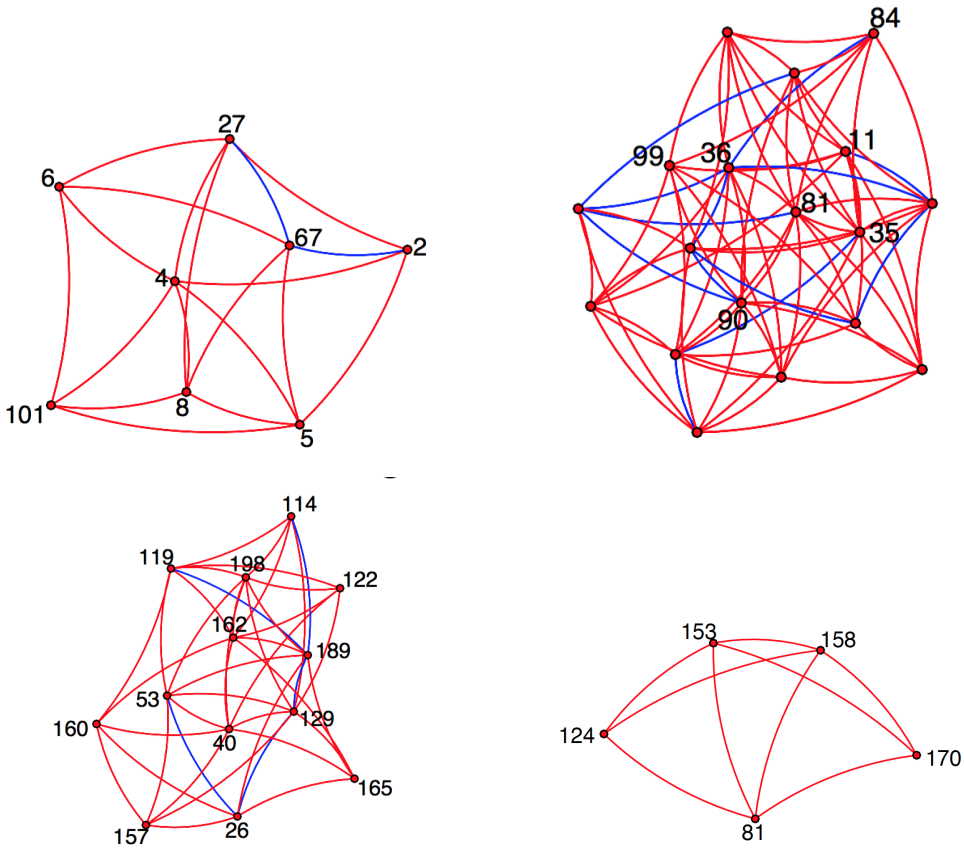


Figure 6: Networks inferred on the four largest components detected by slope heuristic. Regularization parameters in each set of variables are chosen using the BIC^{net} criterion (10). Numbers indicate gene labels.

7 Appendix

In this Appendix, we detail the proof of Theorems 3.1 and 3.2. First, we describe a discretization of the model collection used, which is useful in the two proofs. Then, in Section 7.2, we prove Theorem 3.1. We first generalize a model selection theorem for MLE, introduced by Massart, to random model selection. Subsequently, we prove that our model collection satisfies all the assumptions of this Theorem, and deduce the oracle inequality. In Section 7.3, we prove Theorem 3.2 using Birgé's Lemma with the discretization of the model collection obtained in Section 7.1.

7.1 Model collection and discretization

7.1.1 Discretization for the adjacency matrices

Let $B = (B_1, \ldots, B_K) \in \mathscr{B}$. For a given matrix $\Sigma_B \in S_B^{\text{bound}}$, we may identify a corresponding adjacency matrix A_B. This matrix of size p^2 could be summarized by the vector of concatenated upper triangular vectors. Then, we may construct a discrete space for $\{0,1\}^{p(p-1)/2}$ which is in bijection with

$$\mathscr{A}_B^{\text{bound}} = \left\{ A_B \in \mathbb{S}_p(\{0,1\}) | \exists \Sigma_B \in S_B^{\text{bound}} \text{ s.t. } Adj(\Sigma_B) = A_B \right\}.$$

Let focus first on $\{0,1\}^{p(p-1)/2}$.

Lemma 7.1 *Let $\{0,1\}^{p(p-1)/2}$ be equipped with Hamming distance δ. Let $\{0,1\}_B^{p(p-1)/2}$ be the subset of $\{0,1\}^{p(p-1)/2}$ of vectors for which the corresponding graph has structure B.*

For every $\alpha \in (0,1)$, let $\beta \in (0,1)$ such that $D_B \leq \alpha\beta p(p-1)/2$. There exists some subset $\mathscr{R}(\alpha)$ of $\{0,1\}_B^{p(p-1)/2}$ with the following properties

$$\delta(r, r') > 2(1-\alpha)D_B \text{ for every } (r, r') \in \mathscr{R}(\alpha)^2 \text{ with } r \neq r' \tag{11}$$

$$\log|\mathscr{R}(\alpha)| \geq \rho D_B \log\frac{p(p-1)}{2D_B} + \kappa K(1 - \log(K)) \tag{12}$$

where $\rho = -\alpha(-\log(\beta) + \beta - 1)/\log(\alpha\beta)$ and $D_B = \sum_{1 \leq k \leq K} p_k(p_k - 1)/2$.

Proof. Let $\mathscr{R}$ be a maximal subset of $\{0,1\}_B^{p(p-1)/2}$ satisfying property (11). Then the closed balls with radius ε whose belong to $\mathscr{R}$ cover $\{0,1\}_B^{p(p-1)/2}$. We remark that $x \mapsto P_\sigma x P_\sigma^{-1}$ is a group action, isometric and transitive on $\{0,1\}_B^{p(p-1)/2}$.

Hence,

$$|\{0,1\}_B^{p(p-1)/2}| \leq \sum_{x \in \mathscr{R}} |B_{\{0,1\}_B^{p(p-1)/2}}(x, \varepsilon)| = |\mathscr{R}||B_{\{0,1\}_B^{p(p-1)/2}}(x^0, \varepsilon)|$$

for every $x^0 \in \mathscr{R}$, where $B_A(x,r) = \{y \in A | \delta(x,y) \leq r\}$.

Our proof is similar to the proof of Lemma 4.10 in Massart (2007). Consider:

$$[\{0,1\}^{p(p-1)/2}]_D = \left\{ x \in \{0,1\}^{p(p-1)/2} | \delta(0,x) = D \right\}.$$

Let $\alpha \in (0,1), \beta \in (0,1)$ such that $D \leq \alpha\beta p(p-1)/2$. According to Massart (2007), we know that

$$|B_{[\{0,1\}^{p(p-1)/2}]_D}(x^0, 2(1-\alpha)D)| \leq \frac{\exp\left(-\rho D \log(p(p-1)/2D)\right)}{\binom{p(p-1)/2}{D}}$$

with $\rho = -\alpha(-\log(\beta) + \beta - 1)/\log(\alpha\beta)$.

Nevertheless, as $\{0,1\}_B^{p(p-1)/2} \subset [\{0,1\}^{p(p-1)/2}]_{D_B}$, for $D_B = \sum_{k=1}^K p_k(p_k-1)/2$,

$$|\{0,1\}_B^{p(p-1)/2}| \leq |\mathscr{R}| \frac{\exp\left(-\rho D_B \log(p(p-1)/2D_B)\right)}{\binom{p(p-1)/2}{D_B}}$$

As $\{0,1\}_B^{p(p-1)/2}$ corresponds to the stabilizer of x^0,

$$|\{0,1\}_B^{p(p-1)/2}| \geq \frac{p!}{p_1!\dots p_K!K!}.$$

Note that we divide by $K!$ because there are at worst K clusters with the same size.

As

$$\frac{p!}{p_1!\dots p_K!} \geq 1 \qquad \text{and} \qquad \binom{p(p-1)/2}{D_B} \geq 1,$$

$$|\mathscr{R}| \geq \frac{1}{K!} \exp\left(\rho D_B \log(p(p-1)/2D_B)\right).$$

Using Stirling's approximation, we obtain

$$\log(|\mathscr{R}|) \geq \kappa K(1-\log(K)) + \rho D_B \log\left(\frac{p(p-1)}{2D_B}\right).$$

7.1.2 Discretization for the set of covariance matrices

Corollary 1 *Let $\alpha \in (0,1)$ and $\beta \in (0,1)$ such that $D_B \leq \alpha\beta p(p-1)/2$. Let $\mathscr{R}(\alpha)$ as constructed in Lemma 7.1, and its equivalent $\mathscr{A}_B^{disc}(\alpha)$ for adjacency matrices. Let $\varepsilon > 0$. Let*

$$S_B^{disc}(\varepsilon,\alpha) = \left\{\Sigma \in \mathbb{S}_p^{++}(\mathbb{R}) | Adj(\Sigma) \in \mathscr{A}_B^{disc}(\alpha), \Sigma_{i,j} = \sigma_{i,j}\varepsilon, \sigma_{i,j} \in \left[\frac{e_m}{\varepsilon}, \frac{e_M}{\varepsilon}\right] \cap \mathbb{Z}\right\}.$$

Then,

$$||\Sigma - \Sigma'||_2^2 \geq 2(1-\alpha)D_B \wedge \varepsilon \text{ for every } (\Sigma,\Sigma') \in (S_B^{disc}(\varepsilon,\alpha))^2 \text{ with } \Sigma \neq \Sigma'$$
$$\log|S_B^{disc}(\varepsilon,\alpha)| \geq \rho D_B \log\left(\left\lfloor \frac{e_M - e_m}{\varepsilon} \right\rfloor \frac{p(p-1)}{2D_B}\right) + \kappa K(1-\log(K)).$$

Proof. Let $(\Sigma,\Sigma') \in (S_B^{\text{disc}}(\varepsilon,\alpha))^2$ with $\Sigma \neq \Sigma'$. If Σ and Σ' are close, either they have the same adjacency matrix and they differ only on a coefficient or they differ in their adjacency matrices. In the first case, $||\Sigma - \Sigma'||_2^2 \geq \varepsilon$. In the second case, $||\Sigma - \Sigma'||_2^2 \geq 2(1-\alpha)D_B$. Then,

$$||\Sigma - \Sigma'||_2^2 \geq 2(1-\alpha)D_B \wedge \varepsilon,$$

this minimum depending on α and ε.

7.2 Oracle inequality: proof of Theorem 3.1

First, we state the general theorem we use to get the oracle inequality, and its proof. Then, we deduce the oracle inequality by proving that our model collection satisfies all the assumptions.

7.2.1 Model selection theorem for MLE among a random subcollection

We denote by $\mathscr{H}_{[.]}(\varepsilon, S, d_H)$ the bracketing entropy of the set S with ε-brackets according to the Hellinger distance d_H.

Theorem 7.2 *Let f^* be an unknown density to be estimated from a sample of size n $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$. Consider $\{F_m\}_{m\in\mathscr{M}}$ some at most countable deterministic model collection. Let $\{w_m\}_{m\in\mathscr{M}}$ be some family of nonnegative numbers such that*

$$\sum_{m\in\mathscr{M}} \exp(-w_m) \leq \Omega < \infty. \tag{13}$$

We assume that for every $m \in \mathscr{M}$, $\sqrt{\mathscr{H}_{[.]}(\varepsilon, F_m, d_H)}$ is integrable in 0.

Moreover, for all $m \in \mathscr{M}$, we assume that there exists ψ_m on $\mathbb{R}_+$ such that ψ_m is nondecreasing, $\xi \mapsto \psi_m(\xi)/\xi$ is nonincreasing on $(0, +\infty)$, and for all $\xi \in \mathbb{R}^+$, for all $u \in F_m$, denoting by $F_m(u,\xi) = \{t \in F_m, d_H(t,u) \leq \xi\}$,

$$\int_0^\xi \sqrt{\mathscr{H}_{[.]}(\varepsilon, F_m(u,\xi), d_H)}\, d\varepsilon \leq \psi_m(\xi). \tag{14}$$

Let ξ_m such that $\psi_m(\xi_m) = \sqrt{n}\xi_m^2$.

Introduce $\{F_m\}_{m\in\tilde{\mathscr{M}}}$ some random subcollection of $\{F_m\}_{m\in\mathscr{M}}$. Let $\tau > 0$, and for all $m \in \mathscr{M}$, let $f_m \in F_m$ such that

$$KL(f^*, f_m) \leq 2 \inf_{t\in F_m} KL(f^*, t);$$
$$f_m \geq \exp(-\tau) f^*. \tag{15}$$

Let $\eta \geq 0$ and consider the collection of η-maximum likelihood estimators $\{\hat{f}_m\}_{m\in\tilde{\mathscr{M}}}$. Let pen : $\mathscr{M} \to \mathbb{R}^+$. Suppose that there exists an absolute constant $\kappa > 0$ such that for all $m \in \mathscr{M}$,

$$\mathrm{pen}(m) \geq \kappa \left(\xi_m^2 + (1 \vee \tau) w_m/n\right).$$

Let $\eta' \geq 0$. Then, $\hat{f}_{\hat{m}}$, with $\hat{m} \in \tilde{\mathscr{M}}$ such that

$$-\frac{1}{n}\sum_{i=1}^n \log(\hat{f}_{\hat{m}}(\mathbf{y}_i)) + \mathrm{pen}(\hat{m}) \leq \inf_{m\in\tilde{\mathscr{M}}} \left\{ -\frac{1}{n}\sum_{i=1}^n \log(\hat{f}_m(\mathbf{y}_i)) + \mathrm{pen}(m) \right\} + \eta'$$

satisfies

$$\mathbb{E}(d_H^2(f, \hat{f}_{\hat{m}})) \leq C \left(\inf_{m\in\tilde{\mathscr{M}}} \inf_{t\in F_m} KL(f,t) + \mathrm{pen}(m) \right) + (1 \vee \tau)\frac{\Omega^2}{n} + \eta + \eta'$$

for some absolute positive constant C.

This theorem is a generalization of Theorem 7.11 in Massart (2007) to a random model subcollection of the whole collection. As the proof is adapted from the proof of this theorem, we detail here only differences and we refer the interested reader to Massart (2007).

We denote by γ_n the empirical process and by $\bar{\gamma}_n$ the centered empirical process. Following the proof of the Massart's theorem, easy computations lead to

$$2KL\left(f, \frac{f + \hat{f}_{m'}}{2}\right) \leq KL(f, f_m) + \mathrm{pen}(m) - \mathrm{pen}(m') + 2(\bar{\gamma}_n(g_m) - \bar{\gamma}_n(\hat{s}_{m'}))$$

where

$$g_m = -\frac{1}{2}\log\left(\frac{f_m}{f}\right) \qquad \text{and} \qquad \hat{s}_m = -\log\left(\frac{f+\hat{f}_m}{2f}\right)$$

for $m \in \tilde{\mathscr{M}}$ and $m' \in \tilde{\mathscr{M}}(m) = \left\{m' \in \tilde{\mathscr{M}}, \gamma_n(\hat{f}_{m'}) + \text{pen}(m') \leq \gamma_n(\hat{f}_m) + \text{pen}(m)\right\}$.

To bound $\bar{\gamma}_n(\hat{s}_{m'})$, we use Massart's arguments. The main difference stands in the control of $\bar{\gamma}_n(g_m)$. As $\tilde{\mathscr{M}} \subset \mathscr{M}$ is random, $\mathbb{E}(\bar{\gamma}_n(g_m)) \neq 0$. Nevertheless, thanks to the Bernstein inequality, which we may use thanks to the inequality in (15), we obtain, for all $u > 0$, with probability smaller than $\exp(-u)$,

$$\nu_n(g_m) \leq \sqrt{\frac{1}{n}\alpha_\tau(1 \vee \tau)KL(f, f_m)u} + \frac{\tau}{2n}u,$$

where α_τ is a constant depending on τ. Then, choosing $u = w_m$ for all $m \in \mathscr{M}$, where w_m is defined in (13), some fastidious but straightforward computations similar to those of Massart's lead to Theorem 7.2.

We remark that this is a theoretically easy extension, but quite useful in practice, *e.g.* for controlling large model collections.

7.2.2 Bracketing entropy

Let $B \in \mathscr{B}$. Let $f \in F_B^{\text{bound}}$: $f = \Phi(0, \Sigma_B)$. Let $\varepsilon > 0$ and $\alpha > 0$. According to Corollary 1, there exists $S \in S_B^{\text{disc}}(\varepsilon, \alpha)$ such that:

$$||\Sigma_B - S||_2^2 \leq 2(1-\alpha)D_B \wedge \varepsilon.$$

If we take $\alpha = 1 - \varepsilon/2D_B$, we obtain $||\Sigma_B - S||_2^2 \leq \varepsilon$.

Then we consider:

$$u(x) = (1+2\delta)^\gamma \phi(x|0, (1+\delta)S)$$
$$l(x) = (1+2\delta)^{-\gamma} \phi(x|0, (1+\delta)^{-1}S)$$

According to the Proposition 4 in Maugis and Michel (2011), if $\delta = \beta/\sqrt{3}\gamma$ and if $\varepsilon = \lambda_m \beta/(3\sqrt{3}p^2)$, the set $\{l, u\}$ is a β-bracket set over F_B^{bound}.

If we denote by $\mathscr{N}_{[.]}(\beta, F_B^{\text{bound}}, d_H)$ the minimal number of brackets $[l, u]$ such that $d_h(l, u) \leq \varepsilon$ which are necessary to recover F_B^{bound} and $\mathscr{H}_{[.]}(\beta, F_B^{\text{bound}}, d_H)$ the logarithm of this number, which corresponds to the brackting entropy, we obtain from Corollary 1 that

$$\mathscr{N}_{[.]}(\beta, F_B^{\text{bound}}, d_H) \leq \kappa \left(\frac{3\sqrt{3}p^2(e_M - e_m)p(p-1)}{\lambda_m 2D_B\beta}\right)^{\rho D_B} K(1 - \log(K))$$
$$\mathscr{H}_{[.]}(\beta, F_B^{\text{bound}}, d_H) \leq D_B\left(\log C + \rho \log\left(\frac{1}{D_B\varepsilon}\right)\right).$$

with $C = \kappa K(1-\log(K))\frac{3\sqrt{3}p^2(e_M - e_m)p(p-1)}{2\lambda_m}$.

We then construct ψ_B satisfying Equation (14).

For all $\xi > 0$,

$$\int_0^\xi \sqrt{\mathscr{H}_{[.]}(\beta, F_B^{\text{bound}}, d_H)}d\beta \leq \xi\sqrt{D_B \log C} + \sqrt{D_B\rho}\int_0^\xi \sqrt{\log\left(\frac{1}{D_B\beta}\right)}d\beta.$$

According to Maugis and Michel (2011),

$$\int_0^{\xi} \sqrt{\log\left(\frac{1}{\beta}\right)} d\beta \leq \int_0^{\xi \wedge 1} \sqrt{\log\left(\frac{1}{\beta}\right)} d\beta \leq (\xi \wedge 1)\left(\sqrt{\pi} + \sqrt{\log\left(\frac{1}{\xi \wedge 1}\right)}\right).$$

Then, denoting by $c = \sqrt{\log C} + \sqrt{\pi}$, we can define ψ_B by

$$\psi_B(\xi) = \sqrt{D_B}\xi\left(c + \sqrt{\rho \log \frac{1}{D_B}} + \sqrt{\rho \log \frac{1}{\xi \wedge 1}}\right).$$

As we want ξ_B such that $\psi_B(\xi_B) = \sqrt{n}\xi_B^2$, we could take:

$$\xi_B^2 \leq \frac{D_B}{n}\left[2c^2 + \rho \log\left(\frac{1}{D_B(\frac{D_B}{n}c^2 \wedge 1)}\right)\right].$$

7.2.3 Construction of the weights

We need to control the Bell number, which is the cardinal of $\mathscr{B}$. For this, we use a result of Berend and Tassa (2010), which guarantees that

$$|\mathscr{B}| \leq \left(\frac{0.792p}{\log(p+1)}\right)^p$$

for $p \in \mathbb{N}$.

Lemma 7.3 *Let* $w_B = p \log\left(\frac{0.792p}{\log(p+1)}\right)$. *Then,* $\sum_{B \in \mathscr{B}} \exp(-w_B) \leq 1$.

7.3 Lower bound for the minimax risk: Proof of Theorem 3.2

Fix $B \in \mathscr{B}$.

First case: $p(p-1)/2 \geq 4D_B$

Let $\alpha = 3/4$, $\beta = 1/3$, and $\varepsilon = D_B/2$. Let $S_B^{\text{disc}}(D_B/2, 3/4)$ the discrete space constructed in Corollary 1, and

$$F_B(r) = \left\{rS, S \in S_B^{\text{disc}}\left(\frac{D_B}{2}, \frac{3}{4}\right)\right\}$$

for $r > 0$.

Let $f^* = \phi(0, \Sigma^*)$ be the true density. Let $\hat{f}$ be the considered estimator. We define $\tilde{f} = \underset{f \in F_B(r)}{\operatorname{argmin}} \left\{d_H(\hat{f}, f)\right\}$.

First, we have:

$$d_H(f, \tilde{f}) \leq d_H(f, \hat{f}) + d_H(\hat{f}, \tilde{f}) \leq 2d_H(\hat{f}, f). \tag{16}$$

Secondly, we have:

$$d_H(\tilde{f}, f)^2 \geq 1_{f \neq \tilde{f}} \min_{f' \neq f} d_H(f, f')^2$$

$$\mathbb{E}(d_H(\tilde{f}, f)^2) \geq P(f \neq \tilde{f}) \min_{f' \neq f} d_H(f, f')^2. \tag{17}$$

Then, by combining (16) and (17) we obtain:

$$\max_{f\in F_B(r)} \mathbb{E}(d_H^2(\hat{f},f)) \geq \frac{1}{4}\max_{f\in F_B(r)}\left[P_f(f\neq\tilde{f})\min_{f'\neq f} d_H^2(f,f')\right]. \tag{18}$$

We need to design a lower bound for

$$\max_{f\in F_B(r)} P_f(f\neq\tilde{f}).$$

For this purpose, we use the Birgé lemma.

Lemma 7.4 *Let $(P_f)_{f\in\mathscr{F}}$ a probability family, and $(A_f)_{f\in\mathscr{F}}$ some event pairwise disjoints. Let $a_0 = P_0(A_0)$ and $a = \min_{f\in\mathscr{F}} P_f(A_f)$. Then,*

$$\min_{f\in\mathscr{F}} P_f(A_f) \leq \frac{2e}{2e+1} \vee \frac{\max_{f\in\mathscr{F}} KL(P_f,P_0)}{\log(1+card(\mathscr{F}))}.$$

Then, if use Birgé's Lemma to control $\max_{f\in F_B(r)} P_f(f\neq\tilde{f})$ in (18), we obtain

$$\max_{f\in F_B(r)} E(d_H^2(\hat{f},f)) \geq \frac{1}{4(2e+1)}\frac{1}{4+p\log(e_M/e_m)}\frac{1}{2}\frac{\lambda_m p^3}{e_m^2}D_B r^2 \tag{19}$$

if the following inequality is satisfied

$$\max_{f_1,f_2\in F_B(r)} (nKL(f_1,f_2)) \leq \frac{2e}{2e+1}\log(1+\text{card}F_B(r)). \tag{20}$$

The inequality (20) is satisfied if the inequality (21) is fulfilled, with

$$\frac{n}{2}p^3\frac{\lambda_m}{e_m^2}D_B r^2 \leq \frac{2e}{2e+1}\left(\rho D_B\log\left(\frac{p(p-1)(e_M-e_m)}{D_B^2}\right)+\kappa K(1-\log(K))\right) \tag{21}$$

Then, we can replace this condition in (19) and we obtain

$$\max_{f\in F_B(r)} \mathbb{E}(d_H^2(\hat{f}_B,f)) \geq C\frac{D_B}{n}\left(1+\log\frac{C_2}{D_B^2}\right)$$

with

$$C = \frac{2e}{4(2e+1)^2}\frac{1}{4+p\log(e_M/e_m)}\rho$$

and with $0.233 \leq \rho \leq 0.234$, and $C_2 = p(p-1)(e_M-e_m)$.

Second case: $p(p-1)/2 \leq 4D_B$

We can use the Varshamov-Gilbert lemma (see for example Massart (2007), Lemma 4.7) to discretize the covariance space, and construct $\tilde{F}_B(r)$ as previously. Then, Birgé's Lemma involves

$$\sup_{f\in\tilde{F}_B(r)} E(d_H^2(\hat{f},f) \geq \frac{1}{4(2e+1)}\frac{1}{4+p\log(e_M/e_m)}\frac{1}{2}\frac{\lambda_m p^3}{e_m^2}D_B r^2$$

if

$$\frac{n}{2}p^3\frac{\lambda_m}{e_m^2}D_B r^2 \leq \frac{2e}{2e+1}\rho\frac{D_B}{2}.$$

Then, we obtain the following bound:

$$\sup_{f\in\tilde{F}_B(r)} E(d_H^2(\hat{f},f) \geq \frac{1}{4(2e+1)}\frac{1}{4+p\log(e_M/e_m)}(D_B r^2 \wedge \frac{2e}{2e+1}\rho\frac{D_B}{n})$$

Conclusion
As $F_B(r) \subset F_B^{\text{bound}}$, and $\tilde{F}_B(r) \subset F_B^{\text{bound}}$, choosing $r = (1+\log(C_2/D_B^2))^{1/2}$, we get that

$$\max_{f\in F_B^{\text{bound}}} \mathbb{E}(d_H^2(\hat{f}_B,f)) \geq C\frac{D_B}{n}\left(1+\log\frac{C_2}{D_B^2}\right)$$

with

$$C = \frac{2e}{4(2e+1)^2}\frac{1}{4+p\log(e_M/e_m)}\rho$$

and with $0.233 \leq \rho \leq 0.234$, and $C_2 = p(p-1)(e_M - e_m)$.

References

Akbani, R., Ng, P. K. S., Werner, H. M., Shahmoradgoli, M., Zhang, F., Ju, Z., Liu, W., Yang, J.-Y., Yoshihara, K., Li, J., et al. (2014). A pan-cancer proteomic perspective on the cancer genome atlas. *Nature communications*, 5.

Allen, G. I. and Liu, Z. (2013). A local Poisson graphical model for inferring networks from sequencing data. *IEEE Transactions on NanoBioscience*, 12(3):189–198.

Ambroise, C., Chiquet, J., and Matias, C. (2009). Inferring sparse Gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3:205–238.

Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279.

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

Baraud, Y., Giraud, C., and Huet, S. (2009). Gaussian model selection with an unknown variance. *The Annals of Statistics*, 37(2):630–672.

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470.

Berend, D. and Tassa, T. (2010). Improved bounds on bell numbers and on moments of sums of random variables. *Probability and Mathematical Statistics*, 30(2):185–205.

Bickel, P. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Birgé, L. (2005). A new lower bound for multiple hypothesis testing. *Information Theory, IEEE Transactions*, 51(4):1611–1615.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory & Related Fields*, 138(1-2).

Bouveyron, C., Côme, E., and Jacques, J. (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, in press.

Cai, T., Zhang, C.-H., and Zhou, H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144.

Danaher, P., Wang, P., and Witten, D. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397.

Frazee, A. C., Langmead, B., and Leek, J. T. (2011). ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12(449).

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Genovese, C. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *The Annals of Statistics*, 28(4):1105–1127.

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

Giraud, C., Huet, S., and Verzelen, N. (2012). Graph selection with GGMselect. *Statistical Applications in Genetics and Molecular Biology*, 11(3).

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Hyodo, M., Shutoh, N., Nishiyama, T., and Pavlenko, T. (2015). Testing block-diagonal covariance structure for high-dimensional data. *Statistica Neerlandica*, 69(4):460–482.

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology*, 5(1):21.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal Processing*, 85(4):717 – 736.

Massart, P. (2007). *Concentration inequalities and model selection*. Lecture Notes in Mathematics. Springer, 33, 2003, Saint-Flour, Cantal.

Maugis, C. and Michel, B. (2011). A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM. Probability and Statistics.*, 15:41–68.

Mazumder, R. and Hastie, T. (2012). Exact covariance thresholding into connected components for large-scale Graphical Lasso. *Journal of Machine Learning Research*, 13:781–794.

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

Pavlenko, T., Björkström, A., and Tillander, A. (2012). Covariance structure approximation via glasso in high dimensional supervised classification. *Journal of Applied Statistics*, 39(8):1643–1666.

Pickrell, J. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772.

Rau, A., Gallopin, M., Celeux, G., and Jaffrézic, F. (2013). Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics*, 29(17):2146–2152.

Rau, A., Maugis-Rabusseau, C., Martin-Magniette, M.-L., and Celeux, G. (2015). Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics*, 31(9):1420–1427.

Tan, K., Witten, D., and Shojaie, A. (2015). The Cluster Graphical Lasso for improved estimation of Gaussian graphical models. *Computational Statistics & Data Analysis*, 85:23–36.

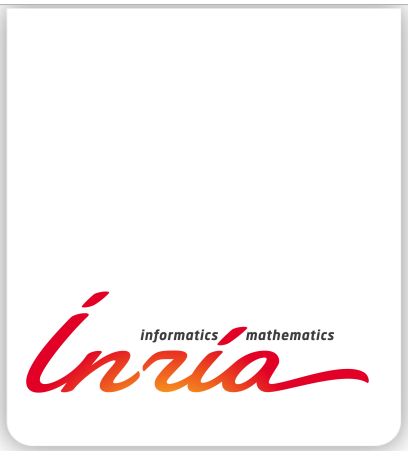
Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley Publishing.

Witten, D. M., Friedman, J. H., and Simon, N. (2011). New insights and faster computations for the Graphical Lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900.

Yin, J. and Li, H. (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics*, 5(4):2630–2650.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *The Journal of Machine Learning Research*, 13(1):1059–1062.



RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing
Campus de l'École Polytechnique
91120 Palaiseau

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399