Inferring sparse Gaussian graphical models with latent structure

by

Christophe Ambroise, Julien Chiquet and Catherine Matias

Research Report No. 21
October, 2008

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

Inferring sparse Gaussian graphical models with latent structure

Christophe Ambroise, Julien Chiquet and Catherine Matias

Laboratoire Statistique et Génome
523, place des Terrasses de l'Agora
91000 Évry, FRANCE
e-mail: christophe.ambroise*;* julien.chiquet*;* catherine.matias@genopole.cnrs.fr
url: http://stat.genopole.cnrs.fr

Abstract: Our concern is selecting the concentration matrix's nonzero coefficients for a sparse Gaussian graphical model in a high-dimensional setting. This corresponds to estimating the graph of conditional dependencies between the variables. We describe a novel framework taking into account a latent structure on the concentration matrix. This latent structure is used to drive a penalty matrix and thus to recover a graphical model with a constrained topology. Our method uses an ℓ_1 penalized likelihood criterion. Inference of the graph of conditional dependencies between the variates and of the hidden variables is performed simultaneously in an iterative EM-like algorithm. The performances of our method is illustrated on synthetic as well as real data, the latter concerning breast cancer.

AMS 2000 subject classifications: Primary 62H20, 62J07; secondary 62H30.
Keywords and phrases: Gaussian graphical model, Mixture model, ℓ_1-penalization, Model selection, Variational inference, EM algorithm.

Contents

1. Introduction

Estimating the concentration matrix (namely the inverse of the covariance matrix) of a Gaussian vector in a sparse, high-dimensional setting has received much attention recently. Graphical models provide a convenient setting for modelling multivariate dependence patterns. In this framework, an undirected graph is matched to the Gaussian random vector, where each vertex corresponds to one coordinate of the vector, and an edge is not present between two vertices if the corresponding random variables are independent, conditional on the remaining variables. Now, conditional independence between two coordinates of the Gaussian random vector corresponds exactly to a zero entry in the concentration matrix. Thus, detecting nonzero elements in the concentration matrix is equivalent to reconstructing the Gaussian graphical model (GGM, see e.g. Lauritzen 1996).

We focus here on the crucial problem of selecting the concentration matrix's nonzero coefficients. In other words, we focus on variable selection rather than estimation. Application areas include gene-regulation graph inference in Biology (using gene expression microarray data), as well as spectroscopy, climate studies, functional magnetic resonance imaging, etc. We provide a very novel approach driving the graph selection according to an unobserved modular structure on the vertices.

The idea of covariance selection first appeared in the work of Dempster (1972). In the so-called "large p, small n" setting (namely when the number of observations is smaller than the dimension of the observed response), the need for covariance selection is huge, as the empirical covariance matrix is no longer regular.

In Drton and Perlman (2007), a classification of the different methods for model selection/estimation in GGMs into three group types is suggested: constraint-based methods, performing statistical tests; Bayesian approaches; and score-based methods, maximizing a model-based criterion. The multiple testing problem has been taken into account in Drton and Perlman (2007; 2008). The authors perform GGM covariance selection by multiple testing of hypotheses about vanishing partial correlation coefficients. Such procedures may also be implemented using the PC-algorithm (Kalisch and Bühlmann 2007). Starting from a complete, undirected graph, the PC-algorithm deletes edges recursively, according to conditional independence decisions. However, the statistical procedure in Drton and Perlman (2007) relies on asymptotic considerations, a regime never attained in real situations.

Another attempt in this vein is to consider limited-order partial correlations (Wille and Bühlmann 2006, Castelo and Roverato 2006). In Wille and Bühlmann (2006) the authors consider only zero and first-order conditional dependencies. They argue that for sparse graphical models, these low-order dependencies still reflect reasonably well the full-order conditional dependency structure. Moreover, these dependencies may be well estimated even with a small number of observations. In Castelo and Roverato (2006), the authors introduce a *non-rejection rate* to reduce the multiple testing and computational problems to which these approaches give rise.

A Bayesian framework is proposed in Dobra et al. (2004), Jones et al. (2005)

and their method was applied for evaluating patterns of association in large-scale gene expression data. The approach is based on *dependency networks*, namely a collection of conditional distributions $\{\mathbb{P}(X_i|X_{\setminus i})\}$ (where $X_{\setminus i}$ stands for the set of all variables but X_i). However, such conditional distributions will not in general result in a coherent joint distribution. This point is further discussed below concerning a similar problem appearing in Meinshausen and Bühlmann (2006). Moreover, constructing priors on the set of concentration matrices is not a trivial task (it mainly relies on Wishart priors). The use of MCMC procedures limits the range of applications to moderate-sized networks.

Before focusing on score-based methods, let us first introduce regularization procedures. In the context of linear regression, the Lasso (least absolute shrinkage and selection operator) technique was introduced by Tibshirani (1996). This procedure performs model selection and parameter estimation at the same time. The idea is that ordinary least-squares criterion may be improved in a sparse context, using an ℓ_1-norm penalty. The ℓ_1-norm penalty shrinks the estimates to zero while preserving the convexity of the optimization problem. Note that the ℓ_1-norm penalization is also known as 'basis pursuit' in signal processing (Chen et al. 2001).

It is well known that if the ultimate goal is parameter estimation, model selection and estimation should be done in a single step. Performing the model selection prior to parameter estimation in the selected model will, in fact, result in a non-robust procedure. However, our primary focus here is on model selection, as we want to infer sparse networks. We therefore concentrate on model selection rather than on estimation performances.

The Lars algorithm (Efron et al. 2004) is one of the most popular techniques for solving the Lasso problem. It gives the path of solutions obtained when varying the penalty parameter (the penalty parameter is used as a scaling factor of the ℓ_1-norm penalty). The larger the penalty parameter, the sparser the Lasso solution.

Using convex optimization techniques (see for instance Minoux 1986), the Lasso problem may be stated as a primal problem, whose dual formulation may be solved more easily. This approach is taken in Osborne et al. (2000b). The authors obtain an iterative algorithm, the "homotopy method" (Osborne et al. 2000a). Other very efficient approaches are based on focusing on each coordinate iteratively. Indeed, for each coordinate, the Lasso problem is solved very simply (assuming the other coordinates are fixed) by soft-thresholding (Donoho and Johnstone 1995). Thus, different 'coordinate optimization' procedures have been proposed in the literature. Following the work of Fu (1998), a cyclic procedure is proposed in Friedman et al. (2007), where optimization with respect to each coordinate is done iteratively; whereas Wu and Lange (2008) propose a greedy approach, computing the solution for each coordinate and choosing that which provides the largest decrease in a surrogate objective function. Note that these approaches rely on the underlying assumption that the predictors for the regression problem are uncorrelated.

Let us now come back to covariance (or concentration) matrix inference in GGMs, using maximization of a model-based criterion.

Meinshausen and Bühlmann (2006) were the first authors to apply Lasso techniques for inferring a covariance matrix in a GGM. Their approach is to solve

p different Lasso regression problems, where p is the dimension of the observed vector. The main drawback of such a procedure is that a symmetrization step is required to obtain the final network. It might, for instance, be the case that the estimator of the regression coefficient for X_i on X_j is zero, whereas the estimator for X_j on X_i is not zero. Meinshausen and Bühlmann propose to use either an "AND" or an "OR" final step procedure to recover an undirected correlation graph. However, these two procedures might result in different estimates and there is no way of choosing between them. Moreover, as previously stated, a set of conditional distributions does not necessarily cohere into a joint distribution. Using a set of possibly non-coherent conditional distributions corresponds to a pseudo-likelihood approach. This aspect was not underlined in Meinshausen and Bühlmann (2006), and we clarify this point in Section 4.

Subsequently, two other articles, Banerjee et al. (2008) and Yuan and Lin (2007) independently provided an improvement of the initial work of Meinshausen and Bühlmann (2006). In both works, the problem is seen as a penalized maximum-likelihood (PML) problem. Instead of considering p different regression problems, these two articles focus on the likelihood of the Gaussian vector, penalizing the entries of the concentration matrix with an ℓ_1-norm penalty. They explain how the PML estimation may be solved as a "Lasso-like" problem. The major issue with PML strategies in the context of the concentration matrix estimation of GGMs is to obtain a positive definite estimate. However, the approach for solving the problem in Yuan and Lin (2007) is not suited to high-dimensional settings, in contrast to the approach proposed in Banerjee et al. (2008). In Yuan and Lin (2007), a non-negative garrote-type estimator is used, and asymptotic properties (as n tends to infinity while p is held fixed) are given. In Banerjee et al. (2008), two different algorithms are proposed for solving problems in a high-dimensional setting. The first approach relies on a block-coordinate descent algorithm. The second is a semi-definite programming algorithm, based on Nesterov's method, which is computationally intensive.

The next improvement in this vein comes with Friedman et al. (2008). Relying on coordinate descent techniques, previously described in Friedman et al. (2007), the authors revisit Banerjee et al.'s first approach and propose an efficient algorithm to solve the PML estimation problem, under the positive definite constraint. In fact, they use the block coordinate descent approach proposed by Banerjee et al. (2008) and combine it with a second coordinate descent method. Our method will make use of this approach.

To conclude this part, we remark that a completely different shrinkage estimate was proposed by Schäfer and Strimmer (2005) in the same context of large-scale covariance matrix estimation. The approach consists in using a weighted average of two different estimators, the first being unconstrained (thus having small bias but large variance), the second being low-dimensional (and thus exhibiting small variance but large bias).

Now let us motivate the use of hidden structures in networks. Modularity is a property observed in real (biological) networks (see for instance Ihmels et al. 2002). Heterogeneity in the node behaviors is an important property of these data. For example, so-called 'hubs' are highly-connected nodes, showing a different behavior from the rest of the graph nodes. An interesting model capturing these network features is a mixture model for random graphs (see for instance Daudin et al. 2008). This model has been rediscovered many times in the literature, and a non-exhaustive bibliography should include Frank and Harary (1982), Snijders and Nowicki (1997), Nowicki and Snijders (2001), Tallberg (2005), Daudin et al. (2008), Mariadassou and Robin (2007), Zanghi et al. (2008). To state it simply, this model assumes that each node belongs to some unobserved group. Conditional on the node groups, the (weighted) edges are independent and identically-distributed (i.i.d.) random variables, whose distribution depends on the groups of the nodes to be connected. As we are interested in GGMs, weighted edges correspond to entries of the concentration matrix.

In this work, we aim at estimating a hidden structure, namely node groups, while discovering the network. This hidden structure should help us in choosing *adaptive* penalty parameters. Indeed, we wish to penalize the elements of the concentration matrix, according to the unobserved clusters to which the nodes belong. For instance, if two nodes belong to the same unobserved group, we wish to lower the penalty parameter acting on the corresponding entry in the concentration matrix. Conversely, if we increase the penalty parameters on the entries corresponding to nodes belonging to different groups, we shrink the estimated coefficient to zero. Our approach is completely new and improves inference of sparse modular networks.

Another adaptive Lasso procedure is given in Zou (2006), whose idea is to lower the bias of the large coefficients by adapting the penalty parameter of each coefficient so that it automatically scales with the inferred value. It is known that the non-adaptive Lasso procedure may result in inconsistent parameter estimation. An illustration of the conflict between optimal prediction and consistent variable selection for the Lasso procedure is given in Meinshausen and Bühlmann (2006). They proved that the optimal penalty parameter for prediction gives inconsistent variable selection results, motivating the use of another penalty parameter to ensure the control of the probability of falsely connecting two or more distinct connectivity components of the graph. Like them, we also focus on optimal selection rather than on optimal prediction. The adaptivity of our procedure is not used for lowering the bias of large coefficients, but instead for constraining the prediction to fit the underlying structure of the graph.

Model. Let us now briefly describe the general approach of our work. The model will be presented in detail in Section 2. Let $X = (X_1, \dots, X_p)^\intercal$ be a Gaussian random vector in $\mathbb{R}^p$, with zero mean and positive definite covariance matrix $\mathbf{\Sigma}$, namely $X \sim \mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma})$. We observe independent and identically-distributed (i.i.d) vectors $(X^1, \dots, X^n)$ with the same distribution as X. The matrix $\mathbf{K} = \mathbf{\Sigma}^{-1}$ is the concentration matrix of the model. Let $\mathbf{S}$ be the empirical covariance matrix. The log-likelihood of the observations is given by

$$\mathcal{L}(\mathbf{K}) = \frac{n}{2}\log\det(\mathbf{K}) - \frac{1}{2}\sum_{k=1}^{n}(X^k)^\intercal \mathbf{K} X^k + c = \frac{n}{2}\log\det(\mathbf{K}) - \frac{n}{2}\mathrm{Tr}(\mathbf{SK}) + c$$

where c is a constant term.

The ℓ_1-penalized estimator proposed by Banerjee et al. (2008) is given by

$$\widehat{\mathbf{K}} = \arg\max_{\mathbf{K}\succ 0}\ \log\det(\mathbf{K}) - \mathrm{Tr}(\mathbf{SK}) - \rho\|\mathbf{K}\|_{\ell_1}, \tag{1}$$

where $\mathbf{K} \succ 0$ stands for positive definiteness, $\rho > 0$ is a penalty parameter and $\|\mathbf{K}\|_{\ell_1} = \sum_{ij} |K_{ij}|$.

A natural generalization of this approach is to have different penalty parameters for different entries K_{ij}. Namely,

$$\log\det(\mathbf{K}) - \operatorname{Tr}(\mathbf{SK}) - \|\boldsymbol{\rho}(\mathbf{K})\|_{\ell_1},$$

where $\boldsymbol{\rho}(\mathbf{K}) = (\rho_{ij}(K_{ij}))_{i,j\in\mathcal{P}}$ is a matrix of penalty functions acting on each entry. As a general rule, using as many penalty functions as there are entries in the concentration matrix to be estimated is not meaningful.

Here, we propose to take into account a hidden structure on the correlations between the coordinates random variables X_k. Thus, we consider latent i.i.d. random variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_p$ with values in a finite set $\{1, \ldots, Q\}$. Each variable $\mathbf{Z}_i$ describes the *state* of X_i, and we wish to adapt the penalty function ρ_{ij} with respect to the states of X_i, X_j. More precisely, we wish to use a criterion of the form

$$\log\det(\mathbf{K}) - \operatorname{Tr}(\mathbf{SK}) - \|\boldsymbol{\rho}_{\mathbf{Z}}(\mathbf{K})\|_{\ell_1},$$

where $\boldsymbol{\rho}_{\mathbf{Z}}(\mathbf{K}) = (\rho_{\mathbf{Z}_i\mathbf{Z}_j}(K_{ij}))_{i,j\in\mathcal{P}}$ is a matrix of random penalty functions whose entries depend on the latent structure $\mathbf{Z} = \mathbf{Z}_1, \ldots, \mathbf{Z}_p$. However, the hidden structure is not supposed to be known, thus we cannot rely on the previous criteria. Intuitively, following the principle of Expectation-Maximization (EM) algorithm of Dempster et al. (1977), the idea will be to replace the unobserved value $\boldsymbol{\rho}_{\mathbf{Z}}(\mathbf{K})$ with its conditional expectation $\mathbb{E}(\boldsymbol{\rho}_{\mathbf{Z}}(\mathbf{K})|X^1, \ldots, X^n; \mathbf{K}^{(m)})$ under some model with parameter $\mathbf{K}^{(m)}$, and iterate the following steps

(E) Compute $\mathbb{E}(\boldsymbol{\rho}_{\mathbf{Z}}(\mathbf{K})|X^1, \ldots, X^n; \mathbf{K}^{(m)})$
(M) Update $\mathbf{K}^{(m+1)} = \operatorname{argmax}_{\mathbf{K}\succ 0} \mathbb{E}(\boldsymbol{\rho}_{\mathbf{Z}}(\mathbf{K})|X^1, \ldots, X^n; \mathbf{K}^{(m)})$.

One of our aims is to provide a very simple framework for such an analysis.

Note that the ℓ_1-norm used here acts on diagonal elements of the matrix $\mathbf{K}$. It is counter-intuitive to penalize diagonal elements of the concentration matrix, as these do not reflect sparsity in the correlation structure. However, from a technical point of view, this strategy ensures that the procedure will select a positive definite estimator (see Remark 2). This point was not emphasized in the previous procedures using ℓ_1 penalized likelihood of GGMs.

Road-map. In Section 2 we present the model and the penalized maximum-likelihood criterion on which we base our inference procedure, described in Section 3. This procedure relies on a variational EM algorithm, combined with a Lasso-like procedure. Section 4 explains how Meinshausen and Bühlmann's approach may be interpreted as a penalized pseudo-likelihood method. Finally, Section 5 illustrates the performance of the method on synthetic data, for which an R–package, `SIMoNe` (Statistical Inference for Modular Network), can be downloaded from the first author's website. We also test our algorithm on a real data set provided by Hess et al. (2006) and concerning $n = 133$ patients with breast cancer treated using chemotherapy. According to Hess et al. (2006) and Natowicz et al. (2008), the patient response to the treatment can be classified either as a pathologic complete response (pCR), or as a residual disease (not-pCR). The prediction of the patient response is achieved accurately by studying the expression levels of a limited number of genes ($p = 26$). Our algorithm is applied on each class of patients (pcR and not-pCR). Two distinct gene-regulatory networks are thus inferred, showing a very different structure according to the selected class of patients.

2. A latent structure model for network inference

In this section we present a framework for modelling heterogeneity among dependencies between the variables. To this end, let us first recall classical notations from Gaussian Graphical Models (see Lauritzen 1996, for elementary results about GGMs).

2.1. *Gaussian graphical models: general settings*

Let $\mathcal{P} = \{1, \ldots, p\}$ be a set of fixed vertices, $X = (X_1, \ldots, X_p)^\intercal$ a random vector describing a signal over this set and a sample $(X^1, \ldots, X^n)$ of size n with the same distribution as X.

The vector X is assumed to be Gaussian with positive definite covariance matrix $\mathbf{\Sigma} = (\Sigma_{ij})_{(i,j)\in\mathcal{P}^2}$. No loss of generality is involved when centering X, so we may assume that $X \sim \mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma})$. GGMs are based on a classical result, originally emphasized by Dempster (1972), claiming that any couple of entries (X_i, X_j) with $i \neq j$ are independent conditional on all other variables indexed by $\mathcal{P}\backslash\{i,j\}$, if and only if the entry $(\mathbf{\Sigma}^{-1})_{ij}$ is zero. The inverse of the covariance matrix $\mathbf{K} = (K_{ij})_{(i,j)\in\mathcal{P}^2} = \mathbf{\Sigma}^{-1}$, known as the concentration matrix, thus describes the conditional independence structure of X. Moreover, each entry $K_{ij}, i \neq j$ is directly linked to the partial correlation coefficient $r_{ij|\mathcal{P}\backslash\{i,j\}}$ between variables X_i and X_j. In fact, we have $r_{ij|\mathcal{P}\backslash\{i,j\}} = -K_{ij}/\sqrt{K_{ii}K_{jj}}$, and also $K_{ii} = \mathrm{Var}(X_i|X_{\mathcal{P}\backslash i})^{-1}$. Hence, after a simple rescaling, the matrix $\mathbf{K}$ can be interpreted as the adjacency matrix of an undirected weighted graph $\mathcal{G}$ representing the partial correlation structure between variables $X_1, \ldots, X_p$. This graph has no self-loop, with a random set of edges composed by all pairs (i, j) such that $K_{ij} \neq 0$. Note that we are seeking only pairs of vertices (i, j) such that $i < j$, since there is no self-loop, and since $K_{ij} = K_{ji}$. Inferring nonzero entries of $\mathbf{K}$ is equivalent to inferring $\mathcal{G}$, and is therefore a highly relevant issue in this framework.

2.2. *Providing the network with a latent structure*

Let us now extend the modeling by providing the network with an internal latent structure.

The model proposed in Daudin et al. (2008) attempts a better fit of data, as it places the network $\mathcal{G}$ in the mixture framework, in order to take account of the heterogeneity among vertices. The same general mixture model is adopted here: vertices of $\mathcal{P}$ are distributed among a set $\mathcal{Q} = \{1, \ldots, Q\}$ of hidden clusters that model the latent structure of the network. For any vertex i, the indicator variable Z_{iq} is equal to 1 if $i \in q$ and 0 otherwise, hence describing which cluster the vertex i belongs to. A vertex is assumed to belong to one cluster only, thus the random vector $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iQ})$ obviously follows a multinomial distribution. Namely,

$$\mathbf{Z}_i \sim \mathcal{M}(1, \boldsymbol{\alpha}), \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q)$ is a vector of cluster proportions, so that $\sum_q \alpha_q = 1$.

The concentration matrix structure. We shall now extend the clustering of vertices from $\mathcal{P}$ to the concentration matrix $\mathbf{K}$. Accordingly, both the existence and the weight of edges, described by the off-diagonal elements of $\mathbf{K}$, will depend on the cluster each vertex belongs to. Conditional on the events $i \in q$ and $j \in \ell$ where q, ℓ are clusters chosen from $\mathcal{Q}$, each K_{ij} $(i \neq j)$ is a random variable whose probability density function is denoted by $f_{q\ell}$, that is,

$$K_{ij} | \{Z_{iq}Z_{j\ell} = 1\} \sim f_{q\ell}(\cdot), \quad i \neq j. \tag{3}$$

It will be remarked that in this formulation the variables K_{ij} are assumed to be independent, conditional on the clusters the vertices belong to. Moreover, we are considering only undirected graphs, so we may assume that $f_{q\ell} = f_{\ell q}$. For technical reasons (see Remark 2), we also assume a prior distribution on diagonal elements of $\mathbf{K}$, namely

$$K_{ii} \sim f_0(\cdot).$$

Our suggestion is to adopt Laplace distributions; hence

$$\forall x \in \mathbb{R}, \quad f_{q\ell}(x) = \frac{1}{2\lambda_{q\ell}} \exp\left\{ -\frac{|x|}{\lambda_{q\ell}} \right\}, \quad \text{and} \quad f_0(x) = \frac{1}{2\lambda_0} \exp\left\{ -\frac{|x|}{\lambda_0} \right\}, \tag{4}$$

where $\lambda_{q\ell}, \lambda_0 > 0$ are scaling parameters and $\lambda_{q\ell} = \lambda_{\ell q}$. Below, the parameter λ_0 will be fixed and not estimated.

The reason for choosing a Laplace distribution is that it is reminiscent of the ℓ_1-norm, itself linked to Lasso-techniques for which appropriate tools are available. In fact, when considering the general penalized least-square problem, the penalty term can be seen as a log-prior density on the vector of parameters. In the case of Lasso, the prior distribution corresponding to the ℓ_1-norm is actually the Laplace distribution (see, e.g. Hastie et al. 2001).

The affiliation model. The affiliation model is a special case of network structure (to be investigated below), where there are many different clusters, but where the focus is restricted to two types of edges: edges between nodes of the same cluster, and edges between nodes from different clusters. In the affiliation model the densities $f_{q\ell}$ in (4) are of only two kinds; that is, for all $q, \ell \in \mathcal{Q}$, let

$$f_{q\ell} = \begin{cases} f_{qq} = f_{\text{in}}(\cdot; \lambda_{\text{in}}) & \text{if } q = \ell, \quad \text{the \textit{intra-cluster} density of edges,} \\ f_{q\ell} = f_{\text{out}}(\cdot; \lambda_{\text{out}}) & \text{if } q \neq \ell, \quad \text{the \textit{inter-cluster} density of edges.} \end{cases} \tag{5}$$

2.3. *The complete likelihood*

Having described the modeling of the network, we now focus on the inference issue.

We denote as $\mathbf{X}$ the $n \times p$ matrix that contains the data-set $\{X^1, X^2, \dots, X^n\}$ row-wisely organized, i.e., $(X^k)^\intercal$ is the kth row of $\mathbf{X}$. Furthermore, we denote as $\mathbf{Z} = \{Z_{iq}\}_{i \in \mathcal{P}, q \in \mathcal{Q}}$ the set of all latent indicator variables for vertices. For the sake of simplicity, the number of clusters Q and the parameters $\boldsymbol{\alpha} = (\alpha_q)_{q \in \mathcal{Q}}$ and $\boldsymbol{\lambda} = \{\lambda_{q\ell}\}_{q,\ell \in \mathcal{Q}}$ are assumed to be known for the moment.

The data experiments $\mathbf{X}$ are the only observations available, and from these we should like to be able to infer the graph $\mathcal{G}$ of conditional dependencies or, equivalently, nonzero entries of $\mathbf{K}$. As the matrix $\mathbf{K}$ has been given a prior distribution, our aim is to maximize the posterior probability of $\mathbf{K}$, given the data $\mathbf{X}$, or equivalently, the logarithm of the joint distribution. The estimate is thus defined as follows:

$$\widehat{\mathbf{K}} = \arg\max_{\mathbf{K}\succ 0} \mathbb{P}(\mathbf{K}|\mathbf{X}) = \arg\max_{\mathbf{K}\succ 0}\ \log \mathbb{P}(\mathbf{X},\mathbf{K}),$$

where $\mathbf{K}\succ 0$ stands for positive-definiteness.

To solve this problem, we place ourselves in the classical complete-data framework. The distribution of $\mathbf{K}$ is only known conditionally on the latent structure described by $\mathbf{Z}$. We denote as $\mathcal{Z}$ the set of all possible clusterings over nodes from $\mathcal{P}$. The marginalization over the latent clusters $\mathbf{Z}$ leads to

$$\widehat{\mathbf{K}} = \arg\max_{\mathbf{K}\succ 0}\ \log \sum_{\mathbf{Z}\in\mathcal{Z}} \mathcal{L}_c(\mathbf{X},\mathbf{K},\mathbf{Z}),$$

where the so-called complete-data likelihood $\mathcal{L}_c(\mathbf{X},\mathbf{K},\mathbf{Z}) = \mathbb{P}(\mathbf{X},\mathbf{K},\mathbf{Z})$ is the function we shall develop using an EM-like strategy hereafter. For this purpose, a closed form of $\mathcal{L}_c$ is required.

Proposition 1. *The following relation holds for the complete-data likelihood $\mathcal{L}_c$.*

$$\begin{aligned}\log\mathcal{L}_c(\mathbf{X},\mathbf{K},\mathbf{Z}) = \frac{n}{2}\left(\log\det(\mathbf{K}) - \mathrm{Tr}(\mathbf{SK})\right) - \|\boldsymbol{\rho}_{\mathbf{Z}}(\mathbf{K})\|_{\ell_1} \\ - \sum_{\substack{i,j\in\mathcal{P}, i\neq j\\ q,\ell\in\mathcal{Q}}} Z_{iq}Z_{j\ell}\log(2\lambda_{q\ell}) + \sum_{i\in\mathcal{P},q\in\mathcal{Q}} Z_{iq}\log\alpha_q + c, \quad (6)\end{aligned}$$

where $\mathbf{S} = n^{-1}(\mathbf{X}-\bar{\mathbf{X}})^\intercal(\mathbf{X}-\bar{\mathbf{X}})$ is the empirical covariance matrix, c is a constant term and $\boldsymbol{\rho}_{\mathbf{Z}}(\mathbf{K}) = \left(\rho_{\mathbf{Z}_i\mathbf{Z}_j}(K_{ij})\right)_{i,j\in\mathcal{P}}$ is defined by

$$\rho_{\mathbf{Z}_i\mathbf{Z}_j}(K_{ij}) = \begin{cases} \displaystyle\sum_{q,\ell\in\mathcal{Q}} Z_{iq}Z_{j\ell}\frac{|K_{ij}|}{\lambda_{q\ell}} & \text{if } i\neq j, \\ \displaystyle\frac{|K_{ii}|}{\lambda_0} & \text{otherwise.}\end{cases} \quad (7)$$

Proof. Using the Bayes rule, $\mathcal{L}_c$ divides into three terms:

$$\log\mathcal{L}_c(\mathbf{X},\mathbf{K},\mathbf{Z}) = \log\mathbb{P}(\mathbf{X},\mathbf{K},\mathbf{Z}) = \log\mathbb{P}(\mathbf{X}|\mathbf{K}) + \log\mathbb{P}(\mathbf{K}|\mathbf{Z}) + \log\mathbb{P}(\mathbf{Z}),$$

where we make use of the fact that $\log\mathbb{P}(\mathbf{X}|\mathbf{K},\mathbf{Z}) = \log\mathbb{P}(\mathbf{X}|\mathbf{K})$.

The first term is the likelihood associated with a size-n sample of a multivariate Gaussian distribution, since $X\sim\mathcal{N}(\mathbf{0}_p,\boldsymbol{\Sigma})$. Routine computations lead to

$$\log\mathbb{P}(\mathbf{X}|\mathbf{K}) = \frac{n}{2}\log\det(\mathbf{K}) - \frac{n}{2}\mathrm{Tr}(\mathbf{SK}) - \frac{np}{2}\log(2\pi).$$

As regards the second term, using the expression (4), we have

$$\begin{aligned}\log \mathbb{P}(\mathbf{K}|\mathbf{Z}) &= \sum_{\substack{i,j\in\mathcal{P},i\neq j\\ q,\ell\in\mathcal{Q}}} Z_{iq}Z_{j\ell}\log f_{q\ell}(K_{ij}) + \sum_{i\in\mathcal{P}}\log f_0(K_{ii})\\ &= -\sum_{\substack{i,j\in\mathcal{P},i\neq j\\ q,\ell\in\mathcal{Q}}} Z_{iq}Z_{j\ell}\left(\frac{|K_{ij}|}{\lambda_{q\ell}} + \log(2\lambda_{q\ell})\right) - \sum_{i\in\mathcal{P}}\frac{|K_{ii}|}{\lambda_0} - p\log(2\lambda_0).\end{aligned}$$

From (2), we have $\log\mathbb{P}(\mathbf{Z}) = \sum_{i,q} Z_{iq}\log\alpha_q$, and the result follows. □

3. Inference strategy by alternate optimization

In the classical EM framework developed by Dempster et al. (1977), where $\mathbf{X}$ is the available data, inferring the unknown parameters $\mathbf{K}$ spread over a latent structure $\mathbf{Z}$ would make use of the following conditional expectation:

$$\begin{aligned}&Q\left(\mathbf{K}|\mathbf{K}^{(m)}\right) = \mathbb{E}\left\{\log\mathcal{L}_c(\mathbf{X},\mathbf{K},\mathbf{Z})\middle|\mathbf{X};\mathbf{K}^{(m)}\right\}\\ &= \sum_{\mathbf{Z}\in\mathcal{Z}}\mathbb{P}\left(\mathbf{Z}\middle|\mathbf{X},\mathbf{K}^{(m)}\right)\log\mathcal{L}_c(\mathbf{X},\mathbf{K},\mathbf{Z}) = \sum_{\mathbf{Z}\in\mathcal{Z}}\mathbb{P}\left(\mathbf{Z}\middle|\mathbf{K}^{(m)}\right)\log\mathcal{L}_c(\mathbf{X},\mathbf{K},\mathbf{Z}),\end{aligned} \tag{8}$$

where $\mathbf{K}^{(m)}$ is the estimation of $\mathbf{K}$ from the previous step of the algorithm.

The usual EM strategy would be to alternate an E-step computing the conditional expectation (8) with an M-step maximizing this quantity over the parameter of interest $\mathbf{K}$. Unfortunately, no closed form of $Q\left(\mathbf{K}|\mathbf{K}^{(m)}\right)$ can be formulated in the present case. The technical difficulty lies in the complex dependency structure contained in the model. Indeed, $\mathbb{P}(\mathbf{Z}|\mathbf{K})$ cannot be factorized, as argued in Daudin et al. (2008). This makes the direct calculation of $Q\left(\mathbf{K}|\mathbf{K}^{(m)}\right)$ impossible. To tackle this problem we use a variational approach (see, e.g., Jaakkola 2000, for elementary results on variational methods). In this framework, the conditional distribution of the latent variables $\mathbb{P}(\mathbf{Z}|\mathbf{K}^{(m)})$ is approximated by a more convenient distribution denoted by $R_m(\mathbf{Z})$, which is chosen carefully in order to be tractable. Hence, our EM-like algorithm deals with the following approximation of the conditional expectation (8)

$$\mathbb{E}_{R_m}\left\{\log\mathcal{L}_c(\mathbf{X},\mathbf{K},\mathbf{Z})\right\} = \sum_{\mathbf{Z}\in\mathcal{Z}} R_m(\mathbf{Z})\log\mathcal{L}_c(\mathbf{X},\mathbf{K},\mathbf{Z}). \tag{9}$$

In the following section we develop a variational argument in order to choose an approximation $R_m(\mathbf{Z})$ of $\mathbb{P}(\mathbf{Z}|\mathbf{K}^{(m)})$. This enables us to compute the conditional expectation (9) and proceed to the maximization step.

3.1. Variational estimation of the latent structure (E-step)

In this part, $\mathbf{K}$ is assumed to be known, and we are looking for an approximate distribution $R(\cdot)$ of the latent variables. The variational approach consists in maximizing a lower bound $\mathcal{J}$ of the log-likelihood $\log\mathbb{P}(\mathbf{X},\mathbf{K})$, defined as follows:

$$\mathcal{J}\left(\mathbf{X},\mathbf{K},R(\mathbf{Z})\right) = \log\mathbb{P}(\mathbf{X},\mathbf{K}) - \mathrm{D}_{KL}\left\{R(\mathbf{Z})\|\mathbb{P}(\mathbf{Z}|\mathbf{K})\right\} \tag{10}$$

where D_{KL} is the Küllback-Leibler divergence. This measures the difference between the probability distribution $\mathbb{P}(\cdot|\mathbf{K})$ in the underlying model and its approximation $R(\cdot)$. An intuitively straightforward choice for $R(\cdot)$ is a completely factorized distribution (see Mariadassou and Robin 2007, Zanghi et al. 2008)

$$R_{\boldsymbol{\tau}}(\mathbf{Z}) = \prod_{i\in\mathcal{P}} h_{\boldsymbol{\tau}_i}(\mathbf{Z}_i), \tag{11}$$

where $h_{\boldsymbol{\tau}_i}$ is the density of the multinomial probability distribution $\mathcal{M}(1;\boldsymbol{\tau}_i)$, and $\boldsymbol{\tau}_i = (\tau_{i1},\dots,\tau_{iQ})$ is a random vector containing the variational parameters to optimize. The complete set of parameters $\boldsymbol{\tau} = \{\tau_{iq}\}_{i\in\mathcal{P},q\in\mathcal{Q}}$ is what we are seeking to obtain via the variational inference. In the case in hand the variational approach intuitively operates as follows: each τ_{iq} must be seen as an approximation of the probability that vertex i belongs to cluster q, conditional on the data, that is, τ_{iq} estimates $\mathbb{P}(Z_{iq}=1|\mathbf{K})$, under the constraint $\sum_q \tau_{iq} = 1$. In the ideal case where $\mathbb{P}(\mathbf{Z}|\mathbf{K})$ can be factorized as $\prod_i \mathbb{P}(\mathbf{Z}_i|\mathbf{K})$ and the parameters τ_{iq} are chosen as $\tau_{iq} = \mathbb{P}(Z_{iq}=1|\mathbf{K})$, the Küllback-Leibler divergence is null and the bound $\mathcal{J}$ reaches the log-likelihood.

The following proposition gives the form of the lower bound $\mathcal{J}$ to be maximized in order to estimate $\boldsymbol{\tau}$.

Proposition 2. *Let us assume that $R_{\boldsymbol{\tau}}$ can be factorized as in (11), and let us denote $\mathcal{J}_{\boldsymbol{\tau}}(\mathbf{X},\mathbf{K}) := \mathcal{J}(\mathbf{X},\mathbf{K},R_{\boldsymbol{\tau}}(\mathbf{Z}))$. Then $\mathcal{J}_{\boldsymbol{\tau}}$ satisfies the following expression*

$$\mathcal{J}_{\boldsymbol{\tau}}(\mathbf{X},\mathbf{K}) = c - \sum_{\substack{i\in\mathcal{P}\\ q\in\mathcal{Q}}} \tau_{iq}\log\tau_{iq} + \sum_{\substack{i\in\mathcal{P}\\ q\in\mathcal{Q}}} \tau_{iq}\log\alpha_q \\ - \|\boldsymbol{\rho}_{\boldsymbol{\tau}}(\mathbf{K})\|_{\ell_1} - \sum_{\substack{i,j\in\mathcal{P},i\neq j\\ q,\ell\in\mathcal{Q}}} \tau_{iq}\tau_{j\ell}\log 2\lambda_{q\ell}, \tag{12}$$

where c does not depend on $\boldsymbol{\tau}$ and $\boldsymbol{\rho}_{\boldsymbol{\tau}}(\mathbf{K}) = (\rho_{\boldsymbol{\tau}_i\boldsymbol{\tau}_j}(K_{ij}))_{i,j\in\mathcal{P}^2}$ is defined similarly as (7), replacing Z_{iq} by τ_{iq}.

Proof. Starting from (10), classical results on variational methods show that

$$\mathcal{J}_{\boldsymbol{\tau}}(\mathbf{X},\mathbf{K}) = \widehat{Q}_{\boldsymbol{\tau}}(\mathbf{K}) + \mathcal{H}(R_{\boldsymbol{\tau}}(\mathbf{Z})),$$

where $\mathcal{H}(R_{\boldsymbol{\tau}}(\cdot))$ is the entropy of the distribution $R_{\boldsymbol{\tau}}(\cdot)$ and $\widehat{Q}_{\boldsymbol{\tau}}(\mathbf{K})$ is the approximation of the complete log-likelihood conditional expectation, computed under the distribution $R_{\boldsymbol{\tau}}$. Namely,

$$\widehat{Q}_{\boldsymbol{\tau}}(\mathbf{K}) = \mathbb{E}_{R_{\boldsymbol{\tau}}}\{\log\mathcal{L}_c(\mathbf{X},\mathbf{K},\mathbf{Z})\} = \sum_{\mathbf{Z}\in\mathcal{Z}} R_{\boldsymbol{\tau}}(\mathbf{Z})\log\mathcal{L}_c(\mathbf{X},\mathbf{K},\mathbf{Z}). \tag{13}$$

In the special case of factorized distribution (11), the entropy is

$$\mathcal{H}(R_{\boldsymbol{\tau}}(\mathbf{Z})) = \sum_{i\in\mathcal{P}} \mathcal{H}(h_{\boldsymbol{\tau}_i}(\mathbf{Z}_i)) = -\sum_{i\in\mathcal{P},q\in\mathcal{Q}} \tau_{iq}\log\tau_{iq}.$$

Moreover,

$$\widehat{Q}_{\boldsymbol{\tau}}(\mathbf{K}) = \log\mathbb{P}(\mathbf{X}|\mathbf{K}) + \mathbb{E}_{R_{\boldsymbol{\tau}}}[\log\mathbb{P}(\mathbf{K}|\mathbf{Z})] + \mathbb{E}_{R_{\boldsymbol{\tau}}}[\log\mathbb{P}(\mathbf{Z})].$$

Equation (12) follows via Proposition 1, by using that $\mathbb{E}_{R_{\boldsymbol{\tau}}}(Z_{iq}) = \tau_{iq}$ and $\mathbb{E}_{R_{\boldsymbol{\tau}}}(Z_{iq}Z_{j\ell}) = \tau_{iq}\tau_{j\ell}$. □

The optimal approximate distribution $R_{\boldsymbol{\tau}}$ is then derived by direct maximization of $\mathcal{J}_{\boldsymbol{\tau}}$. The following proposition gives the estimate $\widehat{\boldsymbol{\tau}}$ that solves the problem.

Proposition 3. *Let $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ be known. The following fixed-point relationship holds for the optimal variational parameters $\widehat{\boldsymbol{\tau}} = \arg\max_{\boldsymbol{\tau}} \mathcal{J}_{\boldsymbol{\tau}}$*

$$\widehat{\tau}_{iq} \propto \alpha_q \prod_{\substack{j\in\mathcal{P}\backslash\{i\}\\ \ell\in\mathcal{Q}}} \left(\frac{1}{2\lambda_{q\ell}} \exp\left\{-\frac{|K_{ij}|}{\lambda_{q\ell}}\right\}\right)^{\widehat{\tau}_{j\ell}}, \tag{14}$$

where $\propto$ means that there is a scaling factor such that for any $i \in \mathcal{P}$, we have $\sum_q \widehat{\tau}_{iq} = 1$.

Proof. This is just an adaptation to the Laplace case of Mariadassou and Robin (2007, Proposition 3). □

The initial value of $\boldsymbol{\tau}$ is chosen using a classification algorithm such as spectral clustering (see for instance Ng et al. 2002). As a consequence, the initial values for τ_{iq} lie in $\{0, 1\}$. We then use an iterative procedure setting $\widehat{\boldsymbol{\tau}}^{(m+1)} = g(\widehat{\boldsymbol{\tau}}^{(m)})$, where g is the function (implicitly defined above) for which $\widehat{\boldsymbol{\tau}}$ is a fixed point. Note that we cannot ensure uniqueness of the fixed point for g, nor convergence of this iterative procedure. In practice, we can always use a maximal number of iterations, and if convergence has not occurred, we keep the initial value of $\boldsymbol{\tau}$ given by the clustering method. In appendix A.2 we explain that at least in the affiliation model (5), if the current values $K_{ij}^{(m)}$ of the precision matrix are small enough, and if the penalty parameters λ_{in}^{-1} and $\lambda_{\text{out}}^{-1}$ are well-chosen, then uniqueness of the fixed point is ensured. However, such a result does not hold in the general case, which is one of the drawbacks of of the variational approach in this context.

Estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$. The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ have been previously considered as known to keep the statement as clear as possible.

Two different strategies may be used with respect to these parameters. The first approach is to fix their values. Fixing the value of $\boldsymbol{\alpha}$ comes down to choosing *a priori* the proportions of the groups, which is quite a common strategy in mixture models. As for the choice of $\boldsymbol{\lambda}$, this is equivalent to choosing the penalty parameter in the classical LASSO. Concerning general parameters $\boldsymbol{\lambda}$, a number of values need to be determined, which might be a problem. However in the particular affiliation model (5), only 2 parameters have to be fixed: a parameter λ_{in} that corresponds to a light penalty, since many intra-cluster edges are expected, and another parameter λ_{out} that fits with a heavier penalty, since we do not expect many inter-cluster edges. This is typically the kind of strategy that will be used for numerical applications (see Section 5). More generally, the matrix penalty can be tuned to obtain a desired quantity of inferred edges, or to constrain the topology of the graph, e.g. graphs with hubs.

The second strategy is to make use of the current inferred graph to estimate the parameters. The basic idea is to include this estimation in the variational method. Unfortunately, the maximization of $\mathcal{J}_{\boldsymbol{\tau}}$ given in equation (12) with respect to $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$ at the same time is not possible. To tackle this problem, we use an alternate strategy. The parameter $\boldsymbol{\tau}$ is computed with the fixed-point relationship (14) for fixed values of $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$. Then we maximize $\mathcal{J}_{\boldsymbol{\tau}}$ with respect to $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$, once $R_{\boldsymbol{\tau}}$ is fixed (that is, once $\boldsymbol{\tau}$ is fixed), as in the following proposition. We successively iterate these two steps until stabilization.

Proposition 4. *For fixed values of $\boldsymbol{\tau}$, the parameters $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\lambda}}$ maximizing $\mathcal{J}_{\boldsymbol{\tau}}$ are given by*

$$\forall q, \ell \in \mathcal{Q}, \ \hat{\alpha}_q = \frac{1}{p} \sum_{i \in \mathcal{P}} \tau_{iq} \ \textit{and} \ \hat{\lambda}_{q\ell} = \frac{\sum_{i \neq j} \tau_{iq}\tau_{j\ell}|K_{ij}|}{\sum_{i \neq j} \tau_{iq}\tau_{j\ell}}.$$

Proof. Once terms that do not depend on the parameters of interest have been removed from $\mathcal{J}_{\boldsymbol{\tau}}$, the problem becomes

$$\hat{\alpha}_q = \operatorname*{argmax}_{\alpha_q} \sum_i \tau_{iq} \log \alpha_q \text{ and } \hat{\lambda}_{q\ell} = \operatorname*{argmax}_{\lambda_{q\ell}} - \sum_{i \neq j} \tau_{iq}\tau_{j\ell} \left(\frac{|K_{ij}|}{\lambda_{q\ell}} + \log 2\lambda_{q\ell} \right).$$

Null-differentiation with respect to α_q (under the constraint $\sum_q \alpha_q = 1$) and $\lambda_{q\ell}$ leads straightforwardly to the result. □

3.2. *A Lasso-like method to estimate the concentration matrix (the M-step)*

Now that we are able to compute the approximate conditional expectation $\widehat{Q}_{\boldsymbol{\tau}}(\mathbf{K})$ defined by (13), we wish to infer the concentration matrix $\mathbf{K}$, assuming $\boldsymbol{\tau}$ is known. This is the aim of the M-step of our EM–like strategy, that deals with the maximization problem $\arg\max_{\mathbf{K} \succ 0} \widehat{Q}_{\boldsymbol{\tau}}(\mathbf{K})$.

Using Proposition 1 and the equality $\mathbb{E}_{R_{\boldsymbol{\tau}}}(Z_{iq}Z_{j\ell}) = \tau_{iq}\tau_{j\ell}$, it is a simple matter to rewrite the problem as follows

$$\widehat{\mathbf{K}} = \operatorname*{argmax}_{\mathbf{K} \succ 0} \left\{ \frac{n}{2} \left(\log \det(\mathbf{K}) - \operatorname{Tr}(\mathbf{SK}) \right) - \|\boldsymbol{\rho}_{\boldsymbol{\tau}}(\mathbf{K})\|_{\ell_1} \right\}. \tag{15}$$

Hence, our M–step can be seen as a penalized maximum likelihood estimation problem, exactly like in Friedman et al. (2008), Banerjee et al. (2008). The likelihood considered here is $\mathbb{P}(\mathbf{X}|\mathbf{K})$, that is, the likelihood which corresponds to the n realizations of the Gaussian vector X for a given concentration matrix $\mathbf{K}$. The difference of our approach lies in the complexity of the penalty term, and in slight discrepancies as regards some constant factors.

Remark 1. Since we are using a penalty term $1/\lambda_0$ on matrix $\mathbf{K}$'s diagonal elements, the solution to (15) satisfies

$$\forall i \in \mathcal{P}, \quad \widehat{K}_{ii}^{-1} = S_{ii} + 2/(n\lambda_0), \tag{16}$$

when $\lambda_0^{-1} < n|S_{ii}|/2$ for any $i \in \mathcal{P}$. Indeed, the sub-gradient equation is $n/2(K_{ii}^{-1} - S_{ii}) + \operatorname{sgn}(K_{ii})/\lambda_0 = 0$, and $K_{ii} \geq 0$ since it is the inverse of a conditional variance.

Let us now look at the solution of the M-step: the following proposition gives an equivalent formulation of (15) that is more likely to be solved. The result draws its inspiration from Banerjee et al. (2008).

Proposition 5. *The maximization problem* (15) *over the concentration matrix* $\mathbf{K}$ *is equivalent to the following, dealing with the covariance matrix* $\mathbf{\Sigma}$

$$\widehat{\mathbf{\Sigma}} = \operatorname*{argmax}_{\|(\mathbf{\Sigma}-\mathbf{S})\cdot/\mathbf{P}_{\boldsymbol{\tau}}\|_\infty \leq 1} \log\det(\mathbf{\Sigma}), \tag{17}$$

where $\cdot/$ *is the term-by-term division and*

$$\mathbf{P}_{\boldsymbol{\tau}} = (P_{\boldsymbol{\tau}_i\boldsymbol{\tau}_j})_{i,j\in\mathcal{P}} \quad \textit{with} \quad P_{\boldsymbol{\tau}_i\boldsymbol{\tau}_j} = \begin{cases} 2n^{-1}\sum_{q,\ell}\tau_{iq}\tau_{j\ell}\lambda_{q\ell}^{-1} & i \neq j, \\ 2(n\lambda_0)^{-1} & i = j. \end{cases}$$

Remark 2. By penalizing the diagonal terms of the concentration matrix $\mathbf{K}$ in the initial problem, the set of matrices $\mathbf{\Sigma}$ over which we maximize our criterion contains, for instance, the matrix $\mathbf{S}+2/(n\lambda_0)I$, (where I stands for the identity matrix). Thus, provided that the value of penalty parameter $1/\lambda_0$ is set sufficiently high, this set contains positive definite matrices. This ensures that our estimator is always invertible. Obviously, when $\mathbf{S}$ is invertible, which is usually true for n greater or equal than p, penalizing the diagonal terms becomes futile. In this case $1/\lambda_0$ is set to zero.

Proof. The penalty term in (15) can be written as follows

$$\|\boldsymbol{\rho}_{\boldsymbol{\tau}}(\mathbf{K})\|_{\ell_1} = \sum_{q,\ell\in\mathcal{Q}} \sum_{\substack{i,j\in\mathcal{P}\\ i\neq j}} \frac{|K_{ij}|}{\lambda_{q\ell}}\tau_{iq}\tau_{j\ell} + \sum_{i\in\mathcal{P}}\frac{|K_{ii}|}{\lambda_0} = \sum_{q,\ell\in\mathcal{Q}} \|\mathbf{T}_{q\ell}\star\mathbf{K}\|_{\ell_1},$$

where $\star$ is the term-by-term product. The set $\{\mathbf{T}_{q\ell}\}_{q,\ell\in\mathcal{Q}}$ contains $p\times p$ symmetric matrices, defined, for each couple (q,ℓ), by

$$\mathbf{T}_{q\ell} = (T_{q\ell;ij})_{i,j\in\mathcal{P}} \quad \text{with} \quad \forall i\neq j, \quad T_{q\ell;ij} = \frac{\tau_{iq}\tau_{j\ell}}{\lambda_{q\ell}} \quad \text{and} \quad T_{q\ell;ii} = \frac{1}{\lambda_0 Q^2}.$$

Let us now use the fact that $\|\mathbf{A}\|_{\ell_1} = \max_{\|\mathbf{U}\|_\infty\leq 1}\operatorname{Tr}(\mathbf{AU})$, for a given matrix $\mathbf{A}$. The optimization problem (15) can now be written as

$$\max_{\mathbf{K}\succ 0}\min_{\{\mathbf{U}_{q\ell}:\|\mathbf{U}_{q\ell}\|_\infty\leq 1\}} \left\{\frac{n}{2}\log\det\mathbf{K} - \operatorname{Tr}\left(\frac{n}{2}\mathbf{SK} + \sum_{q,\ell\in\mathcal{Q}}(\mathbf{T}_{q\ell}\star\mathbf{K})\,\mathbf{U}_{q\ell}\right)\right\},$$

since the trace operator is linear. The dual version of the above expression is obtained by swapping max and min. The maximization is solved by differentiating with respect to $\mathbf{K}$. To do this, we recall that in our specific case the matrices $\mathbf{T}$ are symmetrical, and thus $\operatorname{Tr}((\mathbf{T}\star\mathbf{K})\mathbf{U}) = \operatorname{Tr}(\mathbf{K}(\mathbf{T}\star\mathbf{U}))$. Then, applying the usual rules for the derivative of the trace operator, null-differentiation with respect to $\mathbf{K}$ yields

$$\mathbf{\Sigma} := \mathbf{K}^{-1} = \mathbf{S} + \frac{2}{n}\sum_{q,\ell\in\mathcal{Q}}(\mathbf{U}_{q\ell}\star\mathbf{T}_{q\ell}). \tag{18}$$

The dual problem therefore becomes

$$\min_{\{\mathbf{U}_{q\ell}:\|\mathbf{U}_{q\ell}\|_\infty\leq 1\}}\left\{-\frac{n}{2}\log\det(\mathbf{\Sigma}) - \frac{np}{2}\right\},$$

or in other words,

$$\max_{\{\mathbf{U}_{q\ell}:\|\mathbf{U}_{q\ell}\|_\infty\leq 1\}} \log\det(\boldsymbol{\Sigma}).$$

Finally, we need to write the constraint as a function of $\boldsymbol{\Sigma}$ rather than the set $\{\mathbf{U}_{q\ell}\}$. In fact, we simply need to show that

$$\{\mathbf{U}_{q\ell}; \forall q,\ell\in\mathcal{Q}, \|\mathbf{U}_{q\ell}\|_\infty \leq 1\} = \left\{\boldsymbol{\Sigma}; \left\|(\boldsymbol{\Sigma}-\mathbf{S})\cdot/\mathbf{P}_{\boldsymbol{\tau}}\right\|_\infty \leq 1\right\},$$

which is straightforward (see Appendix A.1 for details). □

To solve (17) and thus obtain the estimate $\widehat{\boldsymbol{\Sigma}}$, we successively use two coordinate descent methods. The first corresponds to a block-wise strategy suggested by Banerjee et al.. The second one is used to solve the resulting Lasso problem and was suggested by Friedman et al. (2007).

Let us first explain the block-wise strategy. For this purpose, we introduce the following notation for $\widehat{\boldsymbol{\Sigma}}$, $\mathbf{S}$ and the penalty matrix $\mathbf{P}_{\boldsymbol{\tau}}$

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix}\widehat{\boldsymbol{\Sigma}}_{11} & \widehat{\boldsymbol{\sigma}}_{12}\\ \widehat{\boldsymbol{\sigma}}_{12}^\intercal & \widehat{\Sigma}_{22}\end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix}\mathbf{S}_{11} & \mathbf{s}_{12}\\ \mathbf{s}_{12}^\intercal & S_{22}\end{bmatrix}, \quad \mathbf{P}_{\boldsymbol{\tau}} = \begin{bmatrix}\mathbf{P}_{11} & \mathbf{p}_{12}\\ \mathbf{p}_{12}^\intercal & P_{22}\end{bmatrix}, \tag{19}$$

where $\widehat{\boldsymbol{\Sigma}}_{11}$, $\mathbf{S}_{11}$ and $\mathbf{P}_{11}$ are $(p-1)\times(p-1)$ matrices, $\widehat{\boldsymbol{\sigma}}_{12}$, $\mathbf{s}_{12}$ and $\mathbf{p}_{12}$ are $(p-1)$ length column vectors and $\widehat{\Sigma}_{22}$, S_{22} and P_{22} are real numbers. We have already remarked (Remark 1) that the solution to (17) satisfies $\widehat{\Sigma}_{22} = S_{22} + 2/(n\lambda_0)$. Moreover, using Schür complement, the vector $\widehat{\boldsymbol{\sigma}}_{12}$ satisfies

$$\widehat{\boldsymbol{\sigma}}_{12} = \operatorname*{argmin}_{\{\mathbf{y}:\|(\mathbf{y}-\mathbf{s}_{12})\cdot/\mathbf{p}_{12}\|_\infty\leq 1\}} \left\{\mathbf{y}^\intercal\widehat{\boldsymbol{\Sigma}}_{11}^{-1}\mathbf{y}\right\}. \tag{20}$$

We have $\det(\widehat{\boldsymbol{\Sigma}}) = \det(\widehat{\boldsymbol{\Sigma}}_{11})(\widehat{\Sigma}_{22} - \widehat{\boldsymbol{\sigma}}_{12}^\intercal\widehat{\boldsymbol{\Sigma}}_{11}^{-1}\widehat{\boldsymbol{\sigma}}_{12})$. The full matrix $\widehat{\boldsymbol{\Sigma}}$ is approximated in the following way: first, if required when p is greater than n, we initialize the procedure with $S+2/(n\lambda_0)I$, where $\lambda_0>0$ is chosen so as to make $S+2/(n\lambda_0)I$ invertible; secondly, we permute the columns (and thus the rows) of $\widehat{\boldsymbol{\Sigma}}$ and iteratively solve problems like (20) until convergence of the procedure. This convergence is ensured by the following lemma.

Lemma 1. *The procedure which starts with a positive definite matrix and iteratively updates the columns and rows of this matrix according to the solutions of* (20) *converges to the solution* $\widehat{\boldsymbol{\Sigma}}$ *of* (17).

Proof. The proof relies on Banerjee et al. (2008, Theorem 3) and Tseng (2001, Theorem 4.1). Convergence of block-coordinate descent methods is a well-documented topic in convex optimization literature. Here, we have to bear in mind that using ℓ_1-norm penalty leads to non-differentiable functions. Thus, we rely on a result by Tseng (2001, Theorem 4.1), which in our case ensures the convergence of the procedure, provided there is at most one solution to each minimization problem (20). This point is proved in Banerjee et al. (2008, Theorem 3). □

Then, starting from a result given in Banerjee et al. (2008), an interpretation of (20) as an ℓ_1–penalized problem is given in Friedman et al. (2008). This ℓ_1–penalized problem is reminiscent of the Lasso and may thus be solved using a coordinate descent strategy (Friedman et al. 2007). The following proposition enunciates a result similar to those obtained in Banerjee et al. (2008, equation (6)) and Friedman et al. (2008, equation (2.4)), although with a more general penalty term and a factor $\frac{1}{2}$ that differs. Since none of these articles gives an explicit proof for this result, it is fitting that we provide our own proof here.

Proposition 6. *Solving* (20) *is equivalent to solving the dual problem*

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \left\| \frac{1}{2} \widehat{\boldsymbol{\Sigma}}_{11}^{1/2} \boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{11}^{-1/2} \mathbf{s}_{12} \right\|_2^2 + \| \mathbf{p}_{12} \star \boldsymbol{\beta} \|_{\ell_1} , \tag{21}$$

where solution $\widehat{\boldsymbol{\sigma}}_{12}$ *to* (20) *and* $\widehat{\boldsymbol{\beta}}$ *to* (21) *are linked through*

$$\widehat{\boldsymbol{\sigma}}_{12} = \widehat{\boldsymbol{\Sigma}}_{11} \widehat{\boldsymbol{\beta}} / 2. \tag{22}$$

Proof. Problem (20) can be written as follows, by splitting the constraint:

$$\begin{cases} & \min_{\mathbf{y}} \mathbf{y}^\intercal \widehat{\boldsymbol{\Sigma}}_{11}^{-1} \mathbf{y} \\ \text{subject to} & -(\mathbf{p}_{12})_i \leq y_i - (\mathbf{s}_{12})_i - (\mathbf{p}_{12})_i \leq 0, \quad \forall i = 1, \dots, p-1, \\ \text{or} & -(\mathbf{p}_{12})_i \leq -y_i + (\mathbf{s}_{12})_i - (\mathbf{p}_{12})_i \leq 0, \quad \forall i = 1, \dots, p-1. \end{cases}$$

Let us introduce L the so-called Lagrangian, with vectors of Lagrange coefficients denoted by $\boldsymbol{\beta}^1 = (\beta_i^1)_{i \leq p-1}, \boldsymbol{\beta}^2 = (\beta_i^2)_{i \leq p-1}$ with nonnegative entries. Also, let $\boldsymbol{\beta} = \boldsymbol{\beta}^2 - \boldsymbol{\beta}^1$. The Lagrange version of the above problem is

$$\min_{\mathbf{y}} \left\{ \mathbf{y}^\intercal \widehat{\boldsymbol{\Sigma}}_{11}^{-1} \mathbf{y} + \max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) \right\} , \tag{23}$$

where, in the present case, L is given by

$$L(\boldsymbol{\beta}) = \sum_i \beta_i^1 \left(y_i - (\mathbf{s}_{12})_i - (\mathbf{p}_{12})_i \right) + \sum_i \beta_i^2 \left(-y_i + (\mathbf{s}_{12})_i - (\mathbf{p}_{12})_i \right) ,$$

The coefficients β_i^1 and β_i^2 maximizing $L(\boldsymbol{\beta})$ are null when the constraints are satisfied, and for each index i, at least one coefficient among $\{\beta_i^1, \beta_i^2\}$ is zero. Then

$$\|\boldsymbol{\beta}\|_{\ell_1} = \sum_i |\beta_i| = \sum_i \left(\beta_i^1 + \beta_i^2 \right) .$$

Meanwhile, consider the dual problem of (23), swapping min and max: the solution that minimizes the dual problem with respect to $\mathbf{y}$ satisfies the null-gradient hypothesis. We obtain $2\widehat{\boldsymbol{\Sigma}}_{11}^{-1} \mathbf{y} - \boldsymbol{\beta} = 0$, that is $\mathbf{y} = \frac{1}{2} \widehat{\boldsymbol{\Sigma}}_{11} \boldsymbol{\beta}$ (which proves equation (22)). Introducing this result in the dual of (23), we get

$$\max_{\boldsymbol{\beta}} -\frac{1}{4} \boldsymbol{\beta}^\intercal \widehat{\boldsymbol{\Sigma}}_{11} \boldsymbol{\beta} + \mathbf{s}_{12}^\intercal \boldsymbol{\beta} - \sum_i \left(\beta_i^1 + \beta_i^2 \right) (\mathbf{p}_{12})_i,$$

also equivalent to

$$\min_{\boldsymbol{\beta}} \frac{1}{4} \boldsymbol{\beta}^\intercal \widehat{\boldsymbol{\Sigma}}_{11} \boldsymbol{\beta} - \mathbf{s}_{12}^\intercal \boldsymbol{\beta} + \| \mathbf{p}_{12} \star \boldsymbol{\beta} \|_{\ell_1} .$$

Expressing this quantity by using the Euclidean norm achieves the proof. □

Hence, the column $\widehat{\boldsymbol{\sigma}}_{12}$ of the estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}$ is computed by solving the Lasso problem (21) using another coordinate descent method.

Lemma 2. *The solution to (21) is computed by updating the* j*th coordinate of* $\widehat{\boldsymbol{\beta}}$ *via*

$$\widehat{\beta}_j = 2S\left((\mathbf{s}_{12})_j - \frac{1}{2}\sum_{k\neq j} (\widehat{\mathbf{\Sigma}}_{11})_{jk}\widehat{\beta}_k \; ; \; (\mathbf{p}_{12})_j \right) / (\widehat{\mathbf{\Sigma}}_{11})_{jj}, \tag{24}$$

where $S(x;\rho) = \operatorname{sgn}(x)(|x|-\rho)_+$ *is the soft-thresholding operator.*

Moreover, the procedure which iteratively updates the entries of vector $\widehat{\boldsymbol{\sigma}}_{12} = \widehat{\mathbf{\Sigma}}_{11}\widehat{\boldsymbol{\beta}}/2$ *according to the solutions* $\widehat{\boldsymbol{\beta}}$ *of (24) converges to the solution of (20).*

Proof. The proof of this lemma is postponed to Appendix A.3. □

Finally, the estimate of the matrix of concentration $\mathbf{K}$ is recovered by inverting $\widehat{\mathbf{\Sigma}}$, which can be done at low computational cost (see appendix A.4 for details). Hence, we solve the initial maximization problem (15) that defines the M-step of our algorithm.

Implementation of the full EM algorithm is outlined in Algorithm 1.

Algorithm 1: The full EM–like algorithm

```
while $\widehat{Q}_{\boldsymbol{\tau}}(\widehat{\mathbf{K}}^{(m)})$ has not stabilized do
    //THE E-STEP: LATENT STRUCTURE INFERENCE
    if $m = 1$ then
        // First pass
        Apply spectral clustering on the empirical covariance $\mathbf{S}$ to initialize $\widehat{\boldsymbol{\tau}}$
    else
        Compute $\widehat{\boldsymbol{\tau}}$ with the fixed-point relationship (14), using $\widehat{\mathbf{K}}^{(m-1)}$

    //THE M-STEP: NETWORK INFERENCE
    Construct the penalty matrix $\mathbf{P}$ according to $\widehat{\boldsymbol{\tau}}$
    while $\widehat{\mathbf{\Sigma}}^{(m)}$ has not stabilized do
        for each column of $\widehat{\mathbf{\Sigma}}^{(m)}$ do
            Compute $\widehat{\boldsymbol{\sigma}}_{12}$ by solving the LASSO–like problem with path-wise coordinate
            optimization

    Compute $\widehat{\mathbf{K}}^{(m)}$ by block inversion of $\widehat{\mathbf{\Sigma}}^{(m)}$

    $m \leftarrow m + 1$
```

3.3. *Choice of penalty parameters*

As previously stated, the penalty parameters $\boldsymbol{\lambda}$ may be estimated in the E-step of the algorithm (see subsection 3.1). However, this choice is not necessarily optimal for the estimation of $\mathbf{K}$, and other choices might in practice lead to a better solution. A good strategy is to keep the estimated value of $\boldsymbol{\lambda}$ in the E-step that leads to the estimation of $\boldsymbol{\tau}$, and to impose another value of $\boldsymbol{\lambda}$ during the M-step. In this part, we indicate a possible choice for the penalty parameters to use in the M-step, ensuring a small error on the connectivity components of the estimated graph.

Let us first introduce some notation. For any node $i \in \mathcal{P}$, let C_i denote the connectivity component of node i in the true underlying conditional dependency graph, and $\widehat{C}_i$ the corresponding component resulting from the estimate $\widehat{\mathbf{K}}$ of this graph structure. The following proposition is based on Meinshausen and Bühlmann (2006, Theorem 2) and Banerjee et al. (2008, Theorem 2).

Proposition 7. *Fix some $\varepsilon > 0$ and choose the penalty parameters $\boldsymbol{\lambda}$ such that, for all $q, \ell \in \mathcal{Q}$,*

$$2p^2 F_{n-2}\left(\frac{2}{n\lambda_{q\ell}}\left(\max_{i\neq j} S_{ii}S_{jj} - \frac{1}{\lambda_{q\ell}^2}\right)^{-1/2}(n-2)^{1/2}\right) \leq \varepsilon, \qquad (25)$$

where $1 - F_{n-2}$ is the c.d.f. of Student's t-distribution with $n-2$ degrees of freedom. Then

$$\mathbb{P}(\exists k, \widehat{C}_k \not\subseteq C_k) \leq \varepsilon. \qquad (26)$$

Proof. Here we simply indicate the main differences between the proof of Banerjee et al. (2008, Theorem 2) and what is valid in our context. Note that according to (15), the estimator $\widehat{\mathbf{K}}$ must satisfy the following sub-gradient equation

$$\forall i \neq j, \quad \frac{n}{2}\left(\widehat{K}_{ij}^{-1} - S_{ij}\right) - \left(\sum_{q,\ell} \frac{Z_{iq}Z_{j\ell}}{\lambda_{q\ell}}\right)\nu_{ij} = 0$$

where $\nu_{ij} \in \operatorname{sgn}(\widehat{K}_{ij})$. Following the proof of Banerjee et al. (2008, Theorem 2), we easily get

$$\mathbb{P}(\exists k, \widehat{C}_k \not\subseteq C_k) \leq p^2 \max_{i\in\mathcal{P}, j\notin C_i} \mathbb{P}\left(\frac{n}{2}|S_{ij}| \geq \sum_{q,\ell}\frac{Z_{iq}Z_{j\ell}}{\lambda_{q\ell}}\right).$$

Performing some computations involving the correlation between variables X_i and X_j, we also obtain

$$\mathbb{P}(\exists k, \widehat{C}_k \not\subseteq C_k) \leq 2p^2 \max_{q,\ell\in\mathcal{Q}} F_{n-2}\left(\frac{2(n-2)^{1/2}}{n\lambda_{q\ell}}\left(\max_{i\in\mathcal{P}, j\notin C_i} S_{ii}S_{jj} - \frac{1}{\lambda_{q\ell}^2}\right)^{-1/2}\right),$$

which entails the conclusion. □

Remark 3. Following Banerjee et al. (2008), note that in order to ensure (25), it is enough to choose the penalty parameter $\boldsymbol{\lambda}$ such that, for all $q, \ell \in \mathcal{Q}$,

$$\lambda_{q\ell}(\varepsilon) \geq \frac{2}{n}\left(n - 2 + t_{n-2}^2\left(\frac{\varepsilon}{2p^2}\right)\right)^{1/2}\left(\max_{i\neq j} S_{ii}S_{jj}\right)^{-1/2} t_{n-2}\left(\frac{\varepsilon}{2p^2}\right)^{-1},$$

where $t_{n-2}(u)$ is the $(1-u)$-quantile of Student's t-distribution with $(n-2)$ degrees of freedom, i.e. $F_{n-2}(t_{n-2}(u)) = u$.

Remark 4. Inequality (25) does not take into account that different penalty parameters are used for different hidden classes $q, \ell \in \mathcal{Q}$. An adaptation of the preceding strategy is to use current values $\mathbf{Z}^{(m)}$ obtained from the probabilities $\boldsymbol{\tau}^{(m)}$ of the hidden classes and to choose the current penalty parameters $\boldsymbol{\lambda}^{(m)}$ accordingly. More precisely, let us set, for instance

$$\forall i \in \mathcal{P}, \quad Z_{iq}^{(m)} = \begin{cases} 1 & \text{if } q = \operatorname{argmax}_\ell \tau_{i\ell}^{(m)} \\ 0 & \text{otherwise.} \end{cases}$$

Then, when

$$2p^2 F_{n-2}\left(\frac{2}{n\lambda_{q\ell}^{(m)}}\left(\max_{\substack{i\neq j\\ Z_{iq}^{(m)}Z_{j\ell}^{(m)}=1}} S_{ii}S_{jj}-\frac{1}{(\lambda_{q\ell}^{(m)})^2}\right)^{-1/2}(n-2)^{1/2}\right)\leq\varepsilon, \quad (27)$$

for all $q,\ell\in\mathcal{Q}$, the current estimate $\widehat{\mathbf{K}}^{(m)}$ of the dependency graph will approximately satisfy (26). Moreover, in order to ensure (27), it is enough to choose, for all $q,\ell\in\mathcal{Q}$,

$$\lambda_{q\ell}^{(m)}(\varepsilon)\geq\frac{2}{n}\left(n-2+t_{n-2}^2\left(\frac{\varepsilon}{2p^2}\right)\right)^{1/2}\left(\max_{\substack{i\neq j\\ Z_{iq}^{(m)}Z_{j\ell}^{(m)}=1}} S_{ii}S_{jj}\right)^{-1/2} t_{n-2}\left(\frac{\varepsilon}{2p^2}\right)^{-1}. \quad (28)$$

Typically, the kind of values obtained with (28) will lead to large penalties and, consequently, to *very* sparse graphs: practically, more informative networks can be obtained by replacing the term $\varepsilon/2p^2$ in (28) by greater values. In any cases, (28) should be seen as a starting value.

4. Link with Meinshausen and Bühlmann's approach

We should also like to fill the gap between, on the one hand solving (15) and, on the other, the approach proposed in Meinshausen and Bühlmann (2006), where p independent penalized regression problems are solved using the Lasso. In fact, we shall show that Meinshausen and Bühlmann's approach is equivalent to maximizing the penalized *pseudo* log-likelihood corresponding to the size-n sample of the multivariate Gaussian vector X on the set of non symmetric matrices. Let us denote as $\widetilde{\mathcal{L}}$ this pseudo-likelihood, defined by

$$\log\widetilde{\mathcal{L}}(\mathbf{X};\mathbf{K})=\sum_{i\in\mathcal{P}}\left(\sum_{k=1}^{n}\log\mathbb{P}(X_i^k|X_{\mathcal{P}\backslash i}^k;\mathbf{K}_i)\right),$$

where $X_{\mathcal{P}\backslash i}^k$ is the kth realization of the Gaussian vector X, once the ith coordinate has been removed. In this section, the ℓ_1-norm of matrices is restricted to off-diagonal elements only, that is, $\|\mathbf{A}\|_{\ell_1}=\sum_{i\neq j}|A_{ij}|$.

Proposition 8. *Consider the solution $\widehat{\mathbf{K}}^{pseudo}$ to the penalized pseudo-likelihood problem*

$$\widehat{\mathbf{K}}^{pseudo}=\underset{\{K_{ij},i\neq j\}}{\operatorname{argmax}}\ \log\widetilde{\mathcal{L}}(\mathbf{X};\mathbf{K})-\|\mathbf{P}\star\mathbf{K}\|_{\ell_1}, \quad (29)$$

(whose diagonal is fixed) and the solution $\widehat{\mathbf{K}}^{MB}$ given in Meinshausen and Bühlmann (2006) to the p different regression problems, using the matrix penalty $2\mathbf{P}/n$. The two solutions have exactly the same null entries.

Proof. Denote by $\mathbf{K}_{\backslash i\backslash i}$ and $\mathbf{S}_{\backslash i\backslash i}$, respectively, the matrices $\mathbf{K}$ and $\mathbf{S}$ once their ith row and ith column have been removed. Moreover, $\mathbf{K}_{i\backslash i}$ and $\mathbf{S}_{i\backslash i}$ are the ith rows of the matrices with the ith term removed. After some routine computations, and using classical results for Gaussian multivariate vectors (see Appendix A.5), it can be shown that

$$\log \widetilde{\mathcal{L}}(\mathbf{X};\mathbf{K}) = \frac{n}{2}\sum_{i\in\mathcal{P}}\left(\log K_{ii} - K_{ii}S_{ii} - 2\mathbf{S}_{i\setminus i}\mathbf{K}_{i\setminus i} - \frac{1}{K_{ii}}\mathbf{K}_{i\setminus i}\mathbf{S}_{\setminus i\setminus i}\mathbf{K}^{\intercal}_{i\setminus i}\right) + c, \tag{30}$$

where c does not depend on $\mathbf{K}$. Thus, if we forget the symmetry constraint on $\mathbf{K}$, maximizing the pseudo-likelihood (30) with respect to the non-diagonal entries of $\mathbf{K}$ is equivalent to p independent maximization problems with respect to each column $\mathbf{K}^{\intercal}_{i\setminus i}$. Consider, for instance, the last column of $\mathbf{K}$, that is, for $i = p$, and the relative term in (30). This term can be written as

$$\begin{aligned} &-\frac{n}{2K_{22}}\left(2K_{22}\mathbf{s}^{\intercal}_{12}\mathbf{K}^{\intercal}_{i\setminus i} + \mathbf{K}^{\intercal}_{i\setminus i}\mathbf{S}_{11}\mathbf{K}_{i\setminus i}\right) \\ &\qquad\qquad = -\frac{n}{2K_{22}}\left\|\mathbf{S}_{11}^{1/2}\mathbf{K}^{\intercal}_{i\setminus i} + K_{22}\mathbf{S}_{11}^{-1/2}\mathbf{s}_{12}\right\|_2^2 + c', \end{aligned}$$

where we use the block-wise notation defined above (19). The term C' does not depend on $\mathbf{K}_{i\setminus i}$, which is the current column of the concentration matrix to infer. Namely, $c' = -K_{22}^2\mathbf{s}^{\intercal}_{12}\mathbf{S}_{11}^{-1}\mathbf{s}_{12}$.

Consider now the penalized version of the log-likelihood (29): we wish to solve p penalized problems of minimization as defined above, which can be written as follows

$$\min_{\boldsymbol{\beta}} \left\|\mathbf{S}_{11}^{1/2}\boldsymbol{\beta} + K_{22}\mathbf{S}_{11}^{-1/2}\mathbf{s}_{12}\right\|_2^2 + \frac{2K_{22}}{n}\left\|\mathbf{p}_{12}\star\boldsymbol{\beta}\right\|_{\ell_1}. \tag{31}$$

Meinshausen and Bühlmann wish to solve p Lasso-problems, for instance for the last variable p,

$$\min_{\boldsymbol{\alpha}} \frac{1}{n}\left\|\mathbf{X}_p - \mathbf{X}_{\setminus p}\boldsymbol{\alpha}\right\|_2^2 + \left\|2n^{-1}\mathbf{p}_{12}\star\boldsymbol{\alpha}\right\|_{\ell_1}, \tag{32}$$

where $\mathbf{X}_p$ is the pth column of $\mathbf{X}$ and $\mathbf{X}_{\setminus p}$ is the matrix of data the pth column has been removed (note that we adapted the penalization term corresponding to the framework developed here).

The minimum is reached in (31) for null-differentiation, and we get

$$2\mathbf{S}_{11}\boldsymbol{\beta} + 2K_{22}\mathbf{s}^{\intercal}_{12} + \frac{2K_{22}}{n}\mathbf{p}_{12}\star\nu = 0,$$

where $\nu \in \operatorname{sign}(\boldsymbol{\beta})$. The same for (32), and we get

$$\frac{2}{n}\mathbf{X}^{\intercal}_{\setminus p}\mathbf{X}_{\setminus p}\boldsymbol{\alpha} - \frac{2}{n}\mathbf{X}^{\intercal}_{p}\mathbf{X}_{\setminus p} + 2n^{-1}\mathbf{p}_{12}\star\gamma = 0,$$

where $\gamma \in \operatorname{sign}(\boldsymbol{\alpha})$. Now, just note that $n^{-1}\mathbf{X}^{\intercal}_{\setminus p}\mathbf{X}_{\setminus p} = \mathbf{S}_{11}$ and $n^{-1}\mathbf{X}^{\intercal}_{p}\mathbf{X}_{\setminus p} = \mathbf{s}^{\intercal}_{12}$, and problems (31) and (32) are equivalent, provided that $\boldsymbol{\alpha} = -\boldsymbol{\beta}/K_{22}$.

Thus, the columns of the concentration matrix (with a removed diagonal term) inferred from the penalized maximum pseudo-likelihood problem (29), and those inferred with Meinshausen and Bühlmann's approach, share exactly the same null-entries, that is, the same network of conditional dependencies. □

5. Numerical experiments

In this section we present numerical experiments on both synthetic data, to investigate how well the proposed selection procedure behaves, and real data, to demonstrate the practical use of GGM covariance selection with latent structure. In the remainder of this section we focus on an affiliation model (5), the choice of the penalty being made in line with Section 3.3. More precisely, we fix the ratio $\lambda_{\text{in}}/\lambda_{\text{out}} = 1.2$ and either let the value $1/\lambda_{\text{in}}$ vary when considering precision/recall curves for synthetic data, or fix this parameter according to (28) when dealing with real data.

5.1. Synthetic data

We perform numerical experiments to assess the performance of our approach (SIMoNe, Statistical Inference for Modular Network) and compare it to already existing methods for GGM covariance selection: GLasso (Friedman et al. 2008) and GeneNet (Schäfer and Strimmer 2005).

Data synthesis in our framework requires the simulation of a structured sparse inverse covariance matrix. To this aim, we first simulate a graph with an affiliation structure. We consider a simple binary affiliation model where two types of edges exist: edges between nodes of the same class and edges between nodes of different classes. The binary incidence matrix of the graph is transformed by randomly flipping the sign of some elements in order to simulate both positively and negatively correlated variables. Positive definiteness of this matrix is ensured by adding a large enough constant to the diagonal. The matrix is then further normalized to have a diagonal of ones. A Gaussian sample of size n with zero mean and the above covariance matrix is then simulated 50 times. The results we present below are averaged over the 50 samples. At the end of this section we discuss the performances of our method when there is no latent structure on the data.

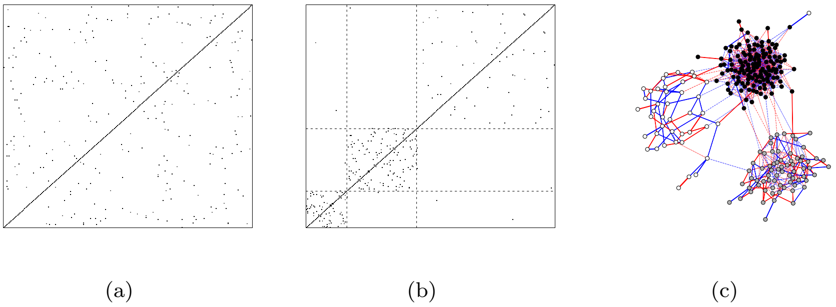

(a) (b) (c)

FIG 1. *Simulation of the structured sparse concentration matrix. Adjacency matrix without (a) and with (b) rows and columns reorganized according the affiliation structure and corresponding graph (c).*

We simulate sparse graphs with $p = 200$ and n from 100 to 2000 ($n/p \in \{1/2, 2, 3, 6, 10\}$). We use a probability of intra-cluster connection of 0.125, a probability of inter-cluster connection of 0.0025, $Q = 3$ groups and equal group proportions $\alpha_i = 1/3$. With these settings, the theoretical expected number of edges is about 862 and the total number of potential edges is 19900. A sample graph is given in Figure 1. The running times of `GLasso` and `SIMoNe` are of the same order. For the settings described above the running time varies from a few seconds to a few minutes, according to the penalty parameter.

We focus the experiments on the ability to recover existing edges of the network, that is the nonzero entries of the concentration matrix. This is a binary decision problem where the compared algorithms are considered as classifiers. The decision made by a binary classifier can be summarized using four numbers: True Positives (TP), False Positive (FP), True Negatives (TN) and False Negatives (FN). We have chosen to draw precision/recall curves to display this information and compare how well the methods perform (Figure 2).

Precision ($TP/(TP + FP)$) is the ratio of the number of true nonzero elements to the total number of nonzero elements in the estimated concentration matrix $\widehat{\mathbf{K}}$. Recall that ($TP/(TP + FN)$) is the ratio of true nonzero elements in $\widehat{\mathbf{K}}$ to all nonzero entries of the real concentration matrix $\mathbf{K}$. In a sparse context where the number of actual positives ($TP + FN$) is small compared to the number of actual negatives ($FP + TN$), precision/recall curves give a more informative picture of an algorithm's performance than classical Receiver Operator Characteristic (ROC) curves. Indeed, ROC curves plot the False Positive Rate ($FPR = FP/(FP + TN)$) against the True Positive Rate ($TPR = TP/(TP + FN)$). When the number of total positives is small compared to the number of total negatives, small variations of FP and TP will result in small variations of FPR and large variations of TPR, which is not relevant for comparing performances. In a statistical framework, the recall is equivalent to the power and the precision is equivalent to one minus the False Discovery Proportion.

Additionally to the `GLasso` (Friedman et al. 2008) and `GeneNet` (Schäfer and Strimmer 2005) we consider two other procedures:

- When n is greater than p, a straightforward way to obtain an estimate of the inverse covariance matrix is to invert the empirical covariance matrix. Although this approach is unlikely to perform well in a selection context (since it is designed for estimation purposes), it is worth comparing it to its competitors in order to assess the scale of improvement. We call this procedure `InvCor`.
- When the latent structure $\mathbf{Z}$ of the concentration matrix is known, our method can be applied without its E-step and produce a relevant selection of the nonzero entries of the concentration matrix. This approach represents the upper limit of our method, since it makes use of an usually unavailable source of information. This procedure is denoted `perfect SIMoNe`.

  In some problems the latent structure of the graph is partially known and this information can be used in the E-step to improve the estimation of the latent structure. For example, when inferring gene regulation networks, a subset of identified genes may be known to belong to the same functional module.

The approach of Meinshausen and Bühlmann (2006) was also tested. The principle of this approach, and the performances obtained are close to those of

GLasso, but it was always slightly outperformed. We have therefore decided, for the sake of brevity, to report only the four previously described procedures.

For the methods based on penalization (GLasso, SIMoNe and Perfect SIMoNe), the precision/recall curves are plotted by varying the penalty parameter (namely $1/\lambda_{\text{in}}$ in our case). The penalty parameter varies from close to zero to a maximum value which forces all off-diagonal elements of $\widehat{\mathbf{K}}$ to be null (see Appendix A.6). The GeneNet and InvCor methods are plotted by sorting the elements of $\widehat{\mathbf{K}}$ according to their absolute values, and choosing different thresholds to find nonzero entries.

Even when n is really greater than p (Figures 2 (a-b)) Invcor is always dominated by the other methods from a selection point of view. This simple check shows that even in a favorable context with abundant data, penalization procedures improve the selection of nonzero entries of the concentration matrix, in comparison with methods based on estimation of these entries.

Although GeneNet and GLasso can provide different results on a given run, both methods perform similarly on average (50 runs for our experiment). The only parameter we change in this experimental setting is the n/p ratio.

Perfect SIMoNe's curves dominate all other curves for any n/p ratio. This clearly shows that the knowledge of the structure provides a valuable information for selecting the nonzero entries of the concentration matrix. When the structure is hidden, the main problem of our approach is then to find a reliable estimate of this structure from the initial data.

Perfect SIMoNe and SIMoNe perform equivalently when $n = 10p$ and when the ratio n/p decreases, Perfect SIMoNe tends to outperform SIMoNe more clearly. This means that SIMoNe is able to recover the latent structure when there is enough data, but does not find a substantial structure when n drops below p.

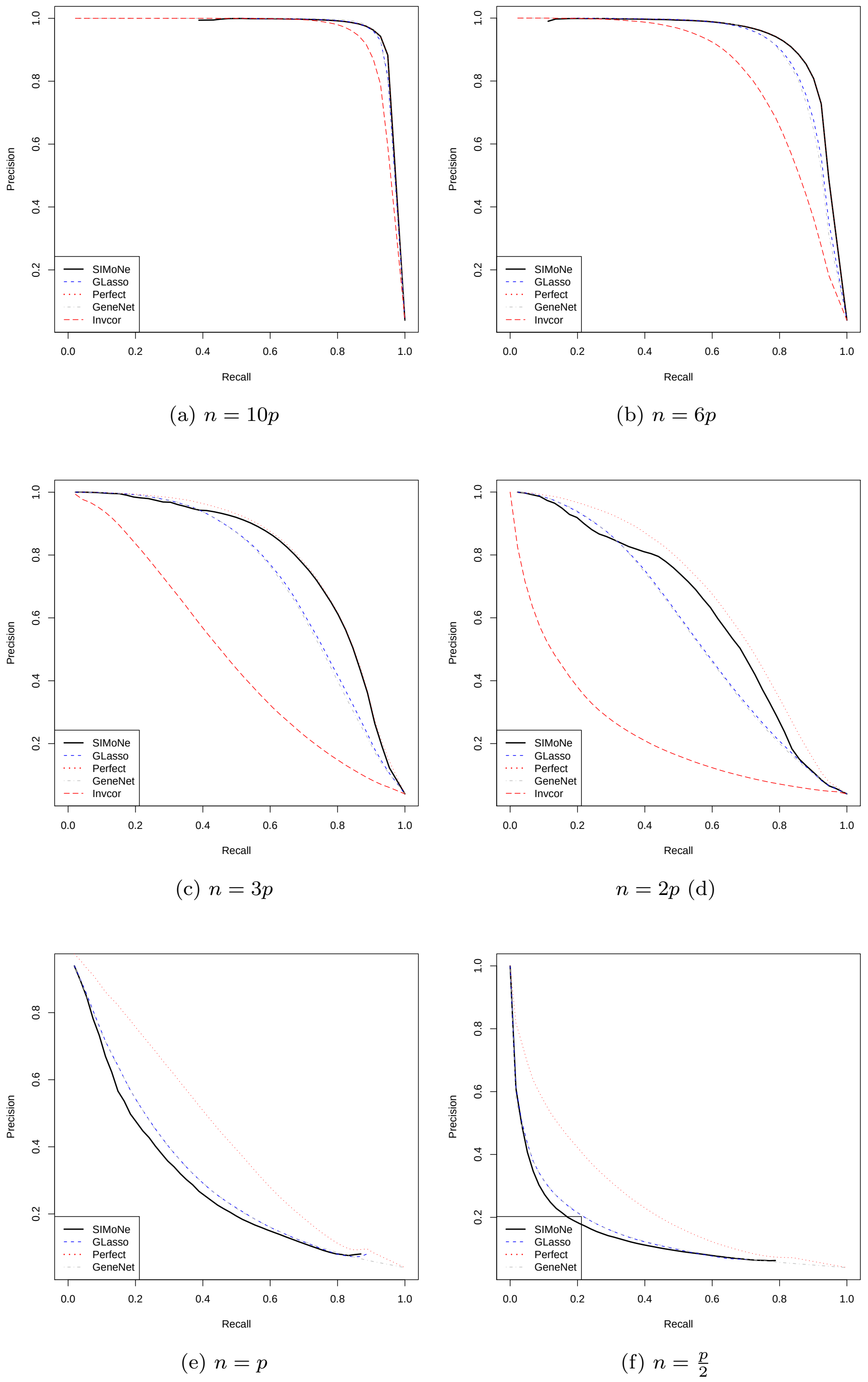
When $p > n$, the empirical covariance matrix ceases to be invertible. Thus, Figures 2 (e-f) do not display the InvCor results. Although it is possible to show that both GLasso and SIMoNe increase the number of inferred true nonzero elements with the number of iterations in all settings, precision/recall curves show the relative poor performances for all tested algorithms when $p \geq n$.

Notice that when $p > n$, the estimated latent structure is not reliable. Nevertheless, the performance of SIMoNe remains comparable to that of GLasso. We can therefore see that assuming the existence of a latent structure when there is none does not impair the selection of nonzero entries of the matrix $\mathbf{K}$.

5.2. Breast Cancer data

We tested our algorithm on a gene expression data set provided by Hess et al. (2006) and concerning 133 patients with stage $I - III$ breast cancer. The patients were treated with chemotherapy prior to surgery. Patient response to the treatment is classified as either a pathologic complete response (pCR) or a residual disease (not-pCR). Hess et al. (2006) and Natowicz et al. (2008) developed and tested a multigene predictor for treatment response on this data set. They focused on a set of 26 genes having a high predictive value (see Table 1). We thus consider a total of $n = 133$ cases containing $p = 26$ gene expression levels.

When dealing with gene regulatory networks, we typically observe n independent microarray experiments, each giving the expression levels of the same



(a) $n = 10p$

(b) $n = 6p$

(c) $n = 3p$

$n = 2p$ (d)

(e) $n = p$

(f) $n = \frac{p}{2}$

FIG 2. *Precision/recall curves comparing the performance of GeneNet, GLasso, SIMoNe and perfect SIMoNe, when inferring the structure of a simulated graph with* $p = 200$ *variables.*

p genes. If the same experimental conditions are used for all microarrays, these may be considered as a sample of the same experiment. In the application in question, cases from the pCR class (34 cases) and from the not-pCR class (99 cases) clearly do not have the same distribution. We apply our algorithm on each class of patients. Two distinct gene regulatory networks are thus inferred.

Figure 3 plots the resulting networks obtained for three different penalizations. The penalization parameters were heuristically chosen from the number of expected nonzero entries. We used $Q = 2$ latent clusters, and it is interesting to note that when assuming more than two clusters, the algorithm systematically produces exactly two non-empty clusters.

The inferred networks exhibit very different structures according to the class of patients. This in itself is interesting and suggests that gene regulation differs with respect to the presence or absence of a pCR.

The network obtained with not-pCR cases displays a two-star pattern. Each star connects to a unique gene, either SCUBE2 or IGFBP4. Almost all the most significant connections imply SCUBE2. This star pattern suggests that further studies of this particular gene would be of interest for understanding residual disease.

The network estimated with the pCR cases has a different two-cluster structure. In particular, it groups IGFBP4 and SCUBE2 in the same cluster with a direct significant link. This again indicates a completely different relationship between the genes in pCR versus non-pCR.

Gene symbol	Gene name
MAPT	Microtubule-associated protein
BBS4	Bardet-Biedl syndrome 4
THRAP2	Thyroid hormone receptor associated protein 2
MBTP-S1	Hypothetical protein
PDGFRA	Human clone 23,948 mRNA sequence
ZNF552	Zinc finger protein 552
RAMP1	Receptor (calcitonin) activity modifying protein 1
BECN1	Beclin 1 (coiled-coil, myosin-likeBCL2 interacting protein)
BTG3	BTG family, member 3
SCUBE2	Signal peptide, CUB domain,EGF-like 2
MELK	Maternal embryonic leucine zipper kinase
AMFR	Autocrine motility factor receptor
CTNND2	Catenin, delta 2
GAMT	Guanidinoacetate N-methyl transferase
CA12	Carbonic anhydrase XII
FGFR1OP	FGFR1 oncogene partner
KIAA1467	KIAA1467 protein
MTRN	Meteorin, glial cell differentiation regulator
FLJ10916	Hypothetical protein FLJ10916
E2F3	E2F transcription factor 3
ERBB4	V-erb-a erythroblastic leukemiaviral oncogene homolog 4(avian)
JMJD2B	Jumonji domain containing 2B
RRM2	Ribonucleotide reductase M2polypeptide
FLJ12650	Hypothetical protein FLJ12650
GFRA1	GDNF family receptor 1
IGFBP4	Insulin-like growth factor binding protein 4

Table 1

The key genes that composed the inferred networks.

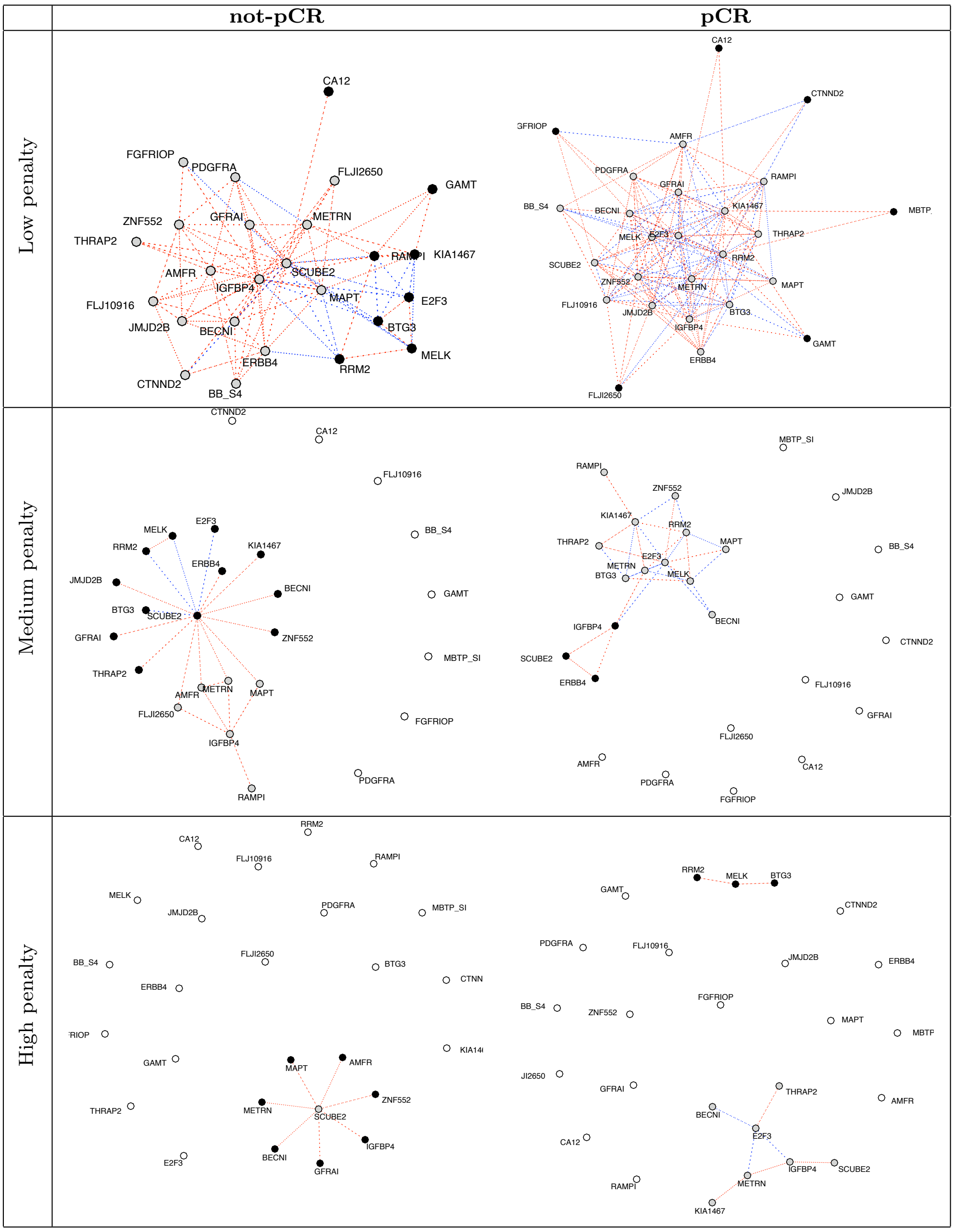


FIG 3. *Inferred graphs for three different penalization's levels.*

References

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

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

Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159 (electronic), 2001. ISSN 0036-1445. Reprinted from SIAM J. Sci. Comput. **20** (1998), no. 1, 33–61 (electronic) [MR1639094 (99h:94013)].

J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Stat. Comput.*, 18(2):173–183, 2008.

A. P. Dempster. Covariance selection. *Biometrics, Special Multivariate Issue*, 28:157–175, 1972.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1): 1–38, 1977. ISSN 0035-9246. With discussion.

Adrian Dobra, Chris Hans, Beatrix Jones, Joseph R. Nevins, Guang Yao, and Mike West. Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.*, 90(1):196–212, 2004.

David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224, 1995.

Mathias Drton and Michael D. Perlman. Multiple testing and error control in gaussian graphical model selection. *Statist. Sci.*, 22:430, 2007.

Mathias Drton and Michael D. Perlman. A SINful approach to gaussian graphical model selection. *J. Statist. Plann. Inference*, 138(4):1179–1200, 2008.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. With discussion, and a rejoinder by the authors.

Ove Frank and Frank Harary. Cluster inference by using transitivity indices in empirical graphs. *J. Amer. Statist. Assoc.*, 77(380):835–840, 1982. ISSN 0162-1459.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

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

W.J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer, 2001.

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

Jan Ihmels, Gilgi Friedlander, Sven Bergmann, Ofer Sarig, Yaniv Ziv, and Naama Barkai. Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, pages 370–377, July 2002.

Jaakkola. *Advanced mean field methods: theory and practice.* MIT Press, 2000.

Beatrix Jones, Carlos Carvalho, Adrian Dobra, Chris Hans, Chris Carter, and Mike West. Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci.*, 20(4):388–400, 2005.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *J. Mach. Learn. Res.*, 8:613–636, Mar 2007.

Steffen L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996. ISBN 0-19-852219-3. Oxford Science Publications.

M. Mariadassou and S. Robin. Uncovering latent structure in valued graphs: a variational approach. Technical Report 10, Statistics for Systems Biology, 2007.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

M. Minoux. *Mathematical programming. Theory and algorithms.* John Wiley and Sons, 1986.

R. Natowicz, R. Incitti, E.G. Horta, B. Charles, P. Guinot, K. Yan, C. Coutant, F. André, and R. Pusztai, L. Rouzier. Prediction of the outcome of a preoperative chemotherapy in breast cancer using dna probes that provide information on both complete and incomplete response. *BMC Bioinformatics*, 9 (149), 2008.

A.Y. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS 14*, 2002.

Krzysztof Nowicki and Tom A. B. Snijders. Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.*, 96(455):1077–1087, 2001. ISSN 0162-1459.

M. R. Osborne, Brett Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.*, 20(3):389–403, 2000a.

Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. On the LASSO and its dual. *J. Comput. Graph. Statist.*, 9(2):319–337, 2000b.

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

Tom A. B. Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification*, 14(1):75–100, 1997. ISSN 0176-4268.

C. Tallberg. A Bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology*, 29(1):1–23, 2005.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, 2001.

Anja Wille and Peter Bühlmann. Low-order conditional independence graphs for inferring genetic networks. *Statistical Applications in Genetics and Molecular Biology*, 5(1), 2006.

Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.*, 2(1):224–244, 2008.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

H. Zanghi, F. Picard, V. Miele, and C. Ambroise. Strategies for online inference of network mixture. Technical Report 14, Statistics for Systems Biology, 2008.

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

Appendix A: Appendix section

A.1. Proof of the equivalence between the constraints

When $\|\mathbf{U}_{q\ell}\|_\infty \leq 1$, we have for each couple $(i,j) \in \mathcal{P}^2$,

$$\left|(\mathbf{\Sigma} - \mathbf{S})_{ij}\right| = \frac{2}{n}\sum_{q,\ell} |(\mathbf{U}_{q\ell})_{ij} \cdot (\mathbf{T}_{q\ell})_{ij}| \leq \frac{2}{n}\sum_{q,\ell} T_{q\ell;ij} = P_{\boldsymbol{\tau}_i\boldsymbol{\tau}_j}.$$

Thus $\|\mathbf{U}_{q\ell}\|_\infty \leq 1 \Rightarrow \|(\mathbf{\Sigma} - \mathbf{S}) \cdot / \mathbf{P}_{\boldsymbol{\tau}}\|_\infty \leq 1$.

On the other hand, assume that $\|(\mathbf{\Sigma} - \mathbf{S}) \cdot / \mathbf{P}_{\boldsymbol{\tau}}\|_\infty \leq 1$, that is, for all $i, j \in \mathcal{P}$, we have

$$-P_{\boldsymbol{\tau}_i\boldsymbol{\tau}_j} \leq (\mathbf{\Sigma} - \mathbf{S})_{ij} \leq P_{\boldsymbol{\tau}_i\boldsymbol{\tau}_j}.$$

This also means that there exists some $\delta_{ij} \in [0,1]$ such that

$$(\mathbf{\Sigma} - \mathbf{S})_{ij} = \delta_{ij} P_{\boldsymbol{\tau}_i\boldsymbol{\tau}_j} + (1 - \delta_{ij})(-P_{\boldsymbol{\tau}_i\boldsymbol{\tau}_j}) = \frac{2}{n}\sum_{q,\ell}(2\delta_{ij} - 1) T_{q\ell;ij}.$$

We choose $\mathbf{U}_{ql}$ such that $(\mathbf{U}_{ql})_{ij} = (2\delta_{ij} - 1)$ for all $q, \ell \in \mathcal{Q}$. Then, since $\delta_{ij} \in [0,1]$, we have

$$-1 \leq (\mathbf{U}_{q\ell})_{ij} \leq 1, \qquad \forall i, j \in \mathcal{P},$$

which proves that $\|(\mathbf{\Sigma} - \mathbf{S}) \cdot / \mathbf{P}_{\boldsymbol{\tau}}\|_\infty \leq 1 \Rightarrow \|\mathbf{U}_{q\ell}\|_\infty \leq 1$.

A.2. Fixed-point study

Let us first introduce some notation. For any $i, j \in \mathcal{P}$ and any $q, \ell \in \mathcal{Q}$, consider the random variables

$$L_{ijq\ell} = \frac{|K_{ij}|}{\lambda_{q\ell}} + \log 2\lambda_{q\ell}.$$

Let $u : \mathbb{R}^{pQ} \to \mathbb{R}^{pQ}$ be defined by its coordinate functions $u = (u_{iq})_{i\in\mathcal{P}, q\in\mathcal{Q}}$ in the following way

$$\begin{aligned}
&\forall a = (a_{iq})_{i\in\mathcal{P},q\in\mathcal{Q}} \in \mathbb{R}^{pQ},\\
&\qquad u_{iq}(a) = \alpha_q \exp\Big\{ -\sum_{j\neq i}\sum_{\ell} a_{j\ell} L_{ijq\ell}\Big\}\\
&\qquad\qquad = \alpha_q \exp\Big\{ -\sum_{j\neq i}\sum_{\ell} a_{j\ell}\left(\frac{|K_{ij}|}{\lambda_{q\ell}} + \log 2\lambda_{q\ell}\right)\Big\},
\end{aligned}$$

and let $g = (g_{iq})_{i\in\mathcal{P},q\in\mathcal{Q}} : \mathbb{R}^{pQ} \to \mathbb{R}^{pQ}$ satisfy

$$\forall a \in \mathbb{R}^{pQ}, \quad g_{iq}(a) = \frac{u_{iq}(a)}{\sum_\ell u_{i\ell}(a)}.$$

According to Proposition 3, the optimal parameter $\widehat{\boldsymbol{\tau}}$ is a fixed-point of g.

Now, let

$$\Theta = \Big\{a = (a_{iq})_{i\in\mathcal{P},q\in\mathcal{Q}} \in \mathbb{R}^{pQ}; \forall i \in \mathcal{P}, q \in \mathcal{Q}, a_{iq} \in [0,1] \text{ and } \sum_q a_{iq} = 1\Big\}.$$

We wish to study the fixed-points of g in Θ. First, let us note that as Θ is a compact state space and as the function g satisfies $g : \Theta \to \Theta$ and is continuous, the existence of a fixed-point of g follows from Brouwer's Theorem.

We now restrict our attention to a smaller set than the whole state space Θ. For any $\varepsilon > 0$, let

$$\Theta_\varepsilon = \Big\{a \in \Theta, \forall i \in \mathcal{P}, q \in \mathcal{Q}, a_{iq} \in [\varepsilon, 1-\varepsilon]\Big\}.$$

Note that we do not claim that $g : \Theta_\varepsilon \to \Theta_\varepsilon$. However, the existence of a fixed-point of g is ensured in Θ and if we assume $\alpha_q > 0$ for any $q \in \mathcal{Q}$ (which is a reasonable assumption if the number of classes Q is not too large), it can easily be seen that any fixed-point satisfies $a_{iq} > 0$, for any $i \in \mathcal{P}$ and any $q \in \mathcal{Q}$. Thus for sufficiently small $\varepsilon > 0$, the fixed-points of g belong to Θ_ε.

In order to study the behaviour of g in the vicinity of a fixed-point, we need to look at some kind of contraction property for g. To this end we introduce a distance d on Θ_ε that will make use of the form of the state space Θ_ε. For all $a, b \in \Theta_\varepsilon$, denote by $a_i = (a_{iq})_{q\in\mathcal{Q}} \in \mathbb{R}^Q$ and $b_i = (b_{iq})_{q\in\mathcal{Q}} \in \mathbb{R}^Q$. Moreover, let

$$d(a,b) = \max_{i\in\mathcal{P}} d_0(a_i, b_i) = \max_{i\in\mathcal{P}} \log\left(\frac{\max_{q\in\mathcal{Q}} a_{iq}/b_{iq}}{\min_{q\in\mathcal{Q}} a_{iq}/b_{iq}}\right) = \max_{i\in\mathcal{P}} \max_{q,\ell\in\mathcal{Q}} \log\left(\frac{a_{iq}b_{i\ell}}{b_{iq}a_{i\ell}}\right).$$

It is well known that d_0 is a distance in $[\varepsilon, 1-\varepsilon]^Q$, and it is easy to check that the resulting d is also a distance in Θ_ε.

Now, fix $a, b \in \mathbb{R}^{pQ}$ and consider the distance $d(g(a), g(b))$. It is easily checked that

$$d(g(a), g(b)) = \max_{i\in\mathcal{P}} d_0(g_i(a), g_i(b)) = \max_{i\in\mathcal{P}} d_0(u_i(a), u_i(b)) = \max_{i\in\mathcal{P}} d_0(\bar{u}_i(a), \bar{u}_i(b)),$$

where $\bar{u} = (\bar{u}_i)_{i\in\mathcal{P}} = (\bar{u}_{iq})_{i\in\mathcal{P},q\in\mathcal{Q}}$ is defined in the following way

$$\forall a = (a_{iq})_{i\in\mathcal{P},q\in\mathcal{Q}} \in \mathbb{R}^{pQ},$$
$$\bar{u}_{iq}(a) = \exp\Big\{\sum_{j\neq i}\sum_\ell a_{j\ell} L_{ijq\ell}\Big\} = \exp\Big\{\sum_{j\neq i}\sum_\ell a_{j\ell}\left(\frac{|K_{ij}|}{\lambda_{q\ell}} + \log 2\lambda_{q\ell}\right)\Big\}.$$

In the following, fix $\varepsilon > 0$ and $a, b \in \Theta_\varepsilon$ and denote by

$$\forall i \in \mathcal{P}, \quad c_1^i = \min_{q\in\mathcal{Q}} \frac{a_{iq}}{b_{iq}}, \quad c_2^i = \max_{q\in\mathcal{Q}} \frac{a_{iq}}{b_{iq}}.$$

With these notations, we have

$$d(a,b) = \max_{i\in\mathcal{P}} d_0(a_i, b_i) = \max_{i\in\mathcal{P}} \log\left(\frac{c_2^i}{c_1^i}\right). \tag{33}$$

We only consider the affiliation model described in (5). Thus, there are only two different values for $\lambda_{q\ell}$, namely λ_{in} and λ_{out} for intra and extra cluster connectivity.

Lemma 3. *If for any $i, j \in \mathcal{P}, i \neq j$ and any $\lambda \in \{\lambda_{in}, \lambda_{out}\}$, we have*

$$0 < \frac{|K_{ij}|}{\lambda} + \log 2\lambda < \frac{\varepsilon}{2(p-1)(1+\varepsilon)} \text{ almost surely,} \tag{34}$$

then the function g satisfies a contraction property on Θ_ε.

Before proving the lemma, let us explain the consequences of this result. Consider the function h_K defined on $(0, +\infty)$ by

$$h_K(\lambda) = \frac{|K|}{\lambda} + \log 2\lambda.$$

This function first decreases from $+\infty$ to the value $1 + \log 2|K|$ on the interval $(0, |K|)$ and then increases from $1 + \log 2|K|$ to $+\infty$ on $(|K|, +\infty)$.

At any step of the algorithm, if the current values $K_{ij}^{(m)}$ of the concentration matrix are small enough, namely smaller than $1/(2e) \simeq 0.184$ then the functions $h_{K_{ij}^{(m)}}$ take all the values between $1 + \log 2|K| < 0$ and $+\infty$. Thus, there is room for choosing $\lambda_{\text{in}}, \lambda_{\text{out}}$ such that (34) is satisfied. In such a case, the fixed-point we are looking for is unique and the iterative procedure setting $\widehat{\tau}^{(s+1)} = g(\widehat{\tau}^{(s)})$ converges.

Proof. Using that for any $j \in \mathcal{P}$ and any $\ell \in \mathcal{Q}$, we have $c_1^j b_{j\ell} \leq a_{j\ell} \leq c_2^j b_{j\ell}$ and $L_{ijq\ell} > 0$, we get

$$\exp\left(\sum_{j\neq i} c_1^j \sum_\ell b_{j\ell} L_{ijq\ell}\right) \leq \bar{u}_{iq}(a) \leq \exp\left(\sum_{j\neq i} c_2^j \sum_\ell b_{j\ell} L_{ijq\ell}\right).$$

Thus, it follows

$$\exp\left(\sum_{j\neq i}(c_1^j - 1) \sum_\ell b_{j\ell} L_{ijq\ell}\right) \leq \frac{\bar{u}_{iq}(a)}{\bar{u}_{iq}(b)} \leq \exp\left(\sum_{j\neq i}(c_2^j - 1) \sum_\ell b_{j\ell} L_{ijq\ell}\right). \tag{35}$$

In the case of the affiliation model, for fixed $i, j \in \mathcal{P}$ and $q \in \mathcal{Q}$, the set of random variables $\{L_{ijq\ell}\}_{\ell\in\mathcal{Q}}$ is reduced to only two random values, namely

$$L_{ij}^{\text{in}} = \frac{|K_{ij}|}{\lambda_{\text{in}}} + \log 2\lambda_{\text{in}}, \quad L_{ij}^{\text{out}} = \frac{|K_{ij}|}{\lambda_{\text{out}}} + \log 2\lambda_{\text{out}}.$$

For the sake of simplicity, we assume $Q = 2$ groups (our arguments may be easily generalized to 3 groups or more). Now, denoting $L_{ij}^{\max} = \max(L_{ij}^{\text{in}}, L_{ij}^{\text{out}})$ and $L_{ij}^{\min} = \min(L_{ij}^{\text{in}}, L_{ij}^{\text{out}})$, it can easily be seen that (for $\varepsilon < 1/2$),

$$\begin{aligned}
\sup_{b\in\Theta_\varepsilon} \sum_\ell b_{j\ell} L_{ijq\ell} &= (1-\varepsilon) L_{ij}^{\max} + \varepsilon L_{ij}^{\min} \\
\inf_{b\in\Theta_\varepsilon} \sum_\ell b_{j\ell} L_{ijq\ell} &= (1-\varepsilon) L_{ij}^{\min} + \varepsilon L_{ij}^{\max},
\end{aligned}$$

almost surely. Note that if we have $Q \geq 3$ groups, explicit bounds can also be obtained (their expression is only slightly more complicated). Coming back to (35), we get

$$\exp\left(\sum_{j\neq i}(c_1^j-1)\{(1-\varepsilon)L_{ij}^{\min}+\varepsilon L_{ij}^{\max}\}\right) \\ \leq \frac{\bar u_{iq}(a)}{\bar u_{iq}(b)} \leq \exp\left(\sum_{j\neq i}(c_2^j-1)\{(1-\varepsilon)L_{ij}^{\max}+\varepsilon L_{ij}^{\min}\}\right).$$

This leads to

$$\begin{aligned} d_0(\bar u_i(a),\bar u_i(b)) &= \log\frac{\max_{q\in\mathcal Q}\bar u_{iq}(a)/\bar u_{iq}(b)}{\min_{q\in\mathcal Q}\bar u_{iq}(a)/\bar u_{iq}(b)} \\ &\leq \sum_{j\neq i}(c_2^j-1)\{(1-\varepsilon)L_{ij}^{\max}+\varepsilon L_{ij}^{\min}\} - \sum_{j\neq i}(c_1^j-1)\{(1-\varepsilon)L_{ij}^{\min}+\varepsilon L_{ij}^{\max}\} \\ &\leq \sum_{j\neq i} L_{ij}^{\max}\{c_2^j-1-\varepsilon(c_2^j+c_1^j-2)\} + L_{ij}^{\min}\{1-c_1^j+\varepsilon(c_2^j+c_1^j-2)\}. \end{aligned}$$

Finally, recall that $d(g(a),g(b)) = \max_i d_0(\bar u_i(a),\bar u_i(b))$, leading to

$$d(g(a),g(b)) \leq \max_{i\in\mathcal P}\left\{\Big(c_2^i-1-\varepsilon(c_2^i+c_1^i-2)\Big)\vee\Big(1-c_1^i+\varepsilon(c_2^i+c_1^i-2)\Big)\right\} \\ \times \max_{i\in\mathcal P}\sum_{j\neq i}(L_{ij}^{\max}+L_{ij}^{\min}).$$

Now, using the inverse triangle inequality, and the fact that $c_1^i \leq 1 \leq c_2^i$, we get for any $i \in \mathcal P$,

$$|c_2^i+c_1^i-2| = \big||c_2^i-1|-|1-c_1^i|\big| \leq |c_2^i-c_1^i| = c_2^i-c_1^i.$$

Moreover, we have $0 \leq c_2^i-1 \leq c_2^i-c_1^i$ and $0 \leq 1-c_1^i \leq c_2^i-c_1^i$. This leads to

$$\begin{aligned} d(g(a),g(b)) &\leq (1+\varepsilon)\max_{i\in\mathcal P}(c_2^i-c_1^i)\times\max_{i\in\mathcal P}\sum_{j\neq i}(L_{ij}^{\max}+L_{ij}^{\min}) \\ &\leq (1+\varepsilon)\max_{i\in\mathcal P}(c_2^i-c_1^i)\times 2(p-1)\max_{j\neq i}L_{ij}^{\max}. \end{aligned} \tag{36}$$

Since a and b belong to Θ_ε, we get that $c_1^i, c_2^i \in [\varepsilon,\varepsilon^{-1}]$ and thus

$$c_2^i-c_1^i = \exp(\log c_2^i)-\exp(\log c_1^i) \leq \frac{1}{\varepsilon}\log\left(\frac{c_2^i}{c_1^i}\right).$$

In particular, recalling (33), we have

$$0 \leq \max_{i\in\mathcal P} c_2^i - c_1^i \leq \frac{1}{\varepsilon}d(a,b).$$

Coming back to (36), we get

$$d(g(a),g(b)) \leq (1+\varepsilon^{-1})2(p-1)\Big(\max_{j\neq i}L_{ij}^{\max}\Big)d(a,b). \tag{37}$$

Now, under assumption (34) the multiplicative random factor $(1+\varepsilon^{-1})2(p-1)\max_{j\neq i}L_{ij}^{\max}$ is strictly smaller than 1. □

*A.3. Proof of Lemma 2 (*Lasso *with pathwise coordinate optimization)*

The following is partly based on Friedman et al. (2007). There are various algorithms for solving the Lasso problem. When there is just one predictor, the Lasso solution is simply given by soft-thresholding (Donoho and Johnstone 1995). The approach used here is based on iterative soft-thresholding with a "partial residual" as a response variable.

The usual formulation of the Lasso problem is the minimization with respect to $\boldsymbol{\beta}$ of the quantity

$$\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \rho\|\boldsymbol{\beta}\|_{\ell_1}, \tag{38}$$

where $(y_i)_{i=1,\dots,n}$ is a vector of response and $(x_{ij})_{i=1,\dots,n;j=1,\dots,p}$ a matrix of predictors such that $\sum_i x_{ij} = 0$, with no loss of generality. Using a coordinate-descent approach, we simply write the problem (38) in the form

$$\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j\right)^2 + \rho\sum_{k\neq j}|\beta_k| + \rho|\beta_j|$$

and minimizing this function with respect to β_j will lead to the solution

$$\beta_j(\rho) = S\left(\sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}), \rho\right) N_j^{-2},$$

where $\tilde{y}_i^{(j)} = \sum_{k\neq j} x_{ik}\beta_k(\rho)$, the normalizing term N_j^2 satisfies $N_j^2 = \sum_{i=1}^n x_{ij}^2$ and the function $S(x,\rho) = \operatorname{sgn}(x)(|x|-\rho)_+$ is the soft-thresholding operator.

This leads to an iterative procedure, repeated on each coordinate of $\boldsymbol{\beta}$ until stabilization of the full vector. Note that as each coordinate-wise solution is unique, results from Tseng (2001, Theorem 4.1) imply that the procedure converges.

Now, we want to apply this approach to solve the problem (21), which can be written

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\left\|\frac{1}{\sqrt{2}}\widehat{\boldsymbol{\Sigma}}_{11}^{1/2}\boldsymbol{\beta} - \sqrt{2}\widehat{\boldsymbol{\Sigma}}_{11}^{-1/2}\mathbf{s}_{12}\right\|_2^2 + \|\mathbf{p}_{12}\star\boldsymbol{\beta}\|_{\ell_1}. \tag{39}$$

From the previous lines, the solution for jth entry of $\boldsymbol{\beta}$ is

$$\beta_j(\mathbf{p}_{12}) = S\left(\sum_i (\widehat{\boldsymbol{\Sigma}}_{11}^{1/2})_{ij}\left((\widehat{\boldsymbol{\Sigma}}_{11}^{-1/2}\mathbf{s}_{12})_i - \frac{1}{2}\sum_{k\neq j}(\widehat{\boldsymbol{\Sigma}}_{11}^{1/2})_{ik}\beta_k(\mathbf{p}_{12})\right), (\mathbf{p}_{12})_j\right) N_j^{-2}.$$

Then, using the symmetry of the matrices, it is easy to see that

$$\begin{gathered}
\textstyle\sum_i (\widehat{\boldsymbol{\Sigma}}_{11}^{1/2})_{ij}(\widehat{\boldsymbol{\Sigma}}_{11}^{-1/2}\mathbf{s}_{12})_i = \sum_\ell (\widehat{\boldsymbol{\Sigma}}_{11}^{1/2}\widehat{\boldsymbol{\Sigma}}_{11}^{-1/2})_{j\ell}(\mathbf{s}_{12})_\ell = (\mathbf{s}_{12})_j, \\
\textstyle\sum_i (\widehat{\boldsymbol{\Sigma}}_{11}^{1/2})_{ij}\sum_{k\neq j}(\widehat{\boldsymbol{\Sigma}}_{11}^{1/2})_{ik}\beta_k(\mathbf{p}_{12}) = \sum_{k\neq j}(\widehat{\boldsymbol{\Sigma}}_{11})_{jk}\beta_k(\mathbf{p}_{12}), \\
N_j^2 = \textstyle\sum_i \left(\frac{(\widehat{\boldsymbol{\Sigma}}_{11}^{1/2})_{ij}}{\sqrt{2}}\right)^2 = (\widehat{\boldsymbol{\Sigma}}_{11}/2)_{jj}.
\end{gathered}$$

Finally, the solution to (21) is computed by updating the jth coordinate of $\boldsymbol{\beta}$ via

$$\beta_j(\mathbf{p}_{12}) = 2S\left((\mathbf{s}_{12})_j - \frac{1}{2}\sum_{k\neq j}(\widehat{\boldsymbol{\Sigma}}_{11})_{jk}\beta_k(\mathbf{p}_{12}); (\mathbf{p}_{12})_j\right)/(\widehat{\boldsymbol{\Sigma}}_{11})_{jj},$$

and permuting the rows of $\widehat{\boldsymbol{\Sigma}}$ until convergence.

A.4. Reconstruction of the concentration matrix

At the end of the block-wise resolution algorithm, a solution $\widehat{\boldsymbol{\Sigma}}$ is available. In order to recover $\widehat{\mathbf{K}}$, we simply use the fact that $\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{K}} = I$. Block-wisely, we get

$$\begin{aligned}
\widehat{\mathbf{K}}_{12} &= -\widehat{\boldsymbol{\Sigma}}_{11}^{-1}\widehat{\boldsymbol{\sigma}}_{12}K_{22} = -K_{22}\widehat{\boldsymbol{\beta}}/2,\\
\widehat{K}_{22} &= 1/(\widehat{\boldsymbol{\sigma}}_{12} - \widehat{\boldsymbol{\sigma}}_{12}^{\intercal}\widehat{\boldsymbol{\Sigma}}_{11}^{-1}\widehat{\boldsymbol{\sigma}}_{12}) = 1/(\widehat{\boldsymbol{\sigma}}_{12} - \widehat{\boldsymbol{\sigma}}_{12}^{\intercal}\widehat{\boldsymbol{\beta}}/2),
\end{aligned}$$

thanks to the fact that $\widehat{\boldsymbol{\sigma}}_{12} = \widehat{\boldsymbol{\Sigma}}_{11}\widehat{\boldsymbol{\beta}}/2$.

To perform this inversion, note that we need to stock the successive solutions $\widehat{\boldsymbol{\beta}}$ of the penalized regressions along the algorithm.

A.5. Pseudo-likelihood of a Gaussian vector

It is well known that the distribution of X_i^k conditional on the remaining variables $X_{\setminus i}^k$ is Gaussian with parameters (μ_i^k, σ_i) given by

$$\mu_i^k = \boldsymbol{\Sigma}_{i\setminus i}\boldsymbol{\Sigma}_{\setminus i\setminus i}^{-1}X_{\setminus i}^k, \qquad \sigma_i = \Sigma_{ii} - \boldsymbol{\Sigma}_{i\setminus i}\boldsymbol{\Sigma}_{\setminus i\setminus i}^{-1}\boldsymbol{\Sigma}_{i\setminus i}^{\intercal}. \tag{40}$$

Denoting $\boldsymbol{m}_i = (\mu_i^1, \dots, \mu_i^n)^{\intercal}$, we get

$$\log\tilde{\mathcal{L}}(\mathbf{X}; \boldsymbol{K}) = -\frac{n}{2}\sum_{i=1}^{p}\log\sigma_i - \sum_{i=1}^{p}\frac{1}{2\sigma_i}(\mathbf{X}_i - \boldsymbol{m}_i)^{\intercal}(\mathbf{X}_i - \boldsymbol{m}_i) + c.$$

It is easy to see that $\boldsymbol{m}_i^{\intercal} = \boldsymbol{\Sigma}_{i\setminus i}\boldsymbol{\Sigma}_{\setminus i\setminus i}^{-1}\mathbf{X}_{\setminus i}^{\intercal}$. Then,

$$\begin{aligned}
\log\tilde{\mathcal{L}}(\mathbf{X}; \boldsymbol{K}) = &-\frac{n}{2}\sum_{i=1}^{p}\log\sigma_i\\
&-\sum_{i=1}^{p}\frac{1}{2\sigma_i}(\mathbf{X}_i^{\intercal}\mathbf{X}_i - 2\mathbf{X}_i^{\intercal}\mathbf{X}_{\setminus i}\boldsymbol{\Sigma}_{\setminus i\setminus i}^{-1}\boldsymbol{\Sigma}_{i\setminus i}^{\intercal} + \boldsymbol{\Sigma}_{i\setminus i}\boldsymbol{\Sigma}_{\setminus i\setminus i}^{-1}\mathbf{X}_{\setminus i}^{\intercal}\mathbf{X}_{\setminus i}\boldsymbol{\Sigma}_{\setminus i\setminus i}^{-1}\boldsymbol{\Sigma}_{i\setminus i}^{\intercal}) + c.
\end{aligned}$$

Note that we have $n^{-1}\mathbf{X}_i^{\intercal}\mathbf{X}_i = S_{ii}$, as well as $n^{-1}\mathbf{X}_i^{\intercal}\mathbf{X}_{\setminus i} = \mathbf{S}_{i\setminus i}$ and $n^{-1}\mathbf{X}_{\setminus i}^{\intercal}\mathbf{X}_{\setminus i} = \mathbf{S}_{\setminus i\setminus i}$. Thus,

$$\begin{aligned}
\log\tilde{\mathcal{L}}(\mathbf{X}; \boldsymbol{K}) = &-\frac{n}{2}\sum_{i=1}^{p}\log\sigma_i\\
&-n\sum_{i=1}^{p}\frac{1}{2\sigma_i}(S_{ii} - 2\mathbf{S}_{i\setminus i}\boldsymbol{\Sigma}_{\setminus i\setminus i}^{-1}\boldsymbol{\Sigma}_{i\setminus i}^{\intercal} + \boldsymbol{\Sigma}_{i\setminus i}\boldsymbol{\Sigma}_{\setminus i\setminus i}^{-1}\mathbf{S}_{\setminus i\setminus i}\boldsymbol{\Sigma}_{\setminus i\setminus i}^{-1}\boldsymbol{\Sigma}_{i\setminus i}^{\intercal}) + c. \qquad (41)
\end{aligned}$$

Recalling that $\mathbf{K} = \mathbf{\Sigma}^{-1}$, and by reordering the rows and columns of the matrices, as well as using a block-wise notation, this becomes

$$\begin{bmatrix} \Sigma_{ii} & \mathbf{\Sigma}_{i\setminus i} \\ \mathbf{\Sigma}^{\intercal}_{i\setminus i} & \mathbf{\Sigma}_{\setminus i\setminus i} \end{bmatrix} \times \begin{bmatrix} K_{ii} & \mathbf{K}_{i\setminus i} \\ \mathbf{K}^{\intercal}_{i\setminus i} & \mathbf{K}_{\setminus i\setminus i} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & I_{p-1} \end{bmatrix},$$

where I_{p-1} is the identity matrix with size $p-1$. In particular, this leads to the identity $\Sigma_{ii}K_{ii} = 1 - \mathbf{\Sigma}_{i\setminus i}\mathbf{K}^{\intercal}_{i\setminus i}$. Thus

$$\Sigma_{ii} = (1 - \mathbf{\Sigma}_{i\setminus i}\mathbf{K}^{\intercal}_{i\setminus i})/K_{ii}. \tag{42}$$

In the same way, we can easily get that $\mathbf{\Sigma}^{\intercal}_{i\setminus i}K_{ii} = -\mathbf{\Sigma}_{\setminus i\setminus i}\mathbf{K}^{\intercal}_{i\setminus i}$ and

$$\mathbf{\Sigma}^{-1}_{\setminus i\setminus i}\mathbf{\Sigma}^{\intercal}_{i\setminus i} = -\mathbf{K}^{\intercal}_{i\setminus i}/K_{ii}. \tag{43}$$

Using identities (42), (43) and (40), we obtain

$$\sigma_i = (1 - \mathbf{\Sigma}_{i\setminus i}\mathbf{K}^{\intercal}_{i\setminus i})/K_{ii} + \mathbf{\Sigma}_{i\setminus i}\mathbf{K}^{\intercal}_{i\setminus i}/K_{ii} = 1/K_{ii}. \tag{44}$$

Now, coming back to (41) and using the identities (42), (43) and (44), we finally obtain the desired result

$$\log \tilde{\mathcal{L}}(\mathbf{X};\mathbf{K}) = \frac{n}{2}\sum_{i=1}^{p}\log K_{ii} - n\sum_{i=1}^{p}\left(\frac{K_{ii}}{2}S_{ii} + \mathbf{S}_{i\setminus i}\mathbf{K}_{i\setminus i} + \frac{1}{2K_{ii}}\mathbf{K}_{i\setminus i}\mathbf{S}_{\setminus i\setminus i}\mathbf{K}^{\intercal}_{i\setminus i}\right) + c.$$

A.6. Penalization upper bound

The following lemma states that if the penalization parameters $\lambda_{q\ell}^{-1}$ and λ_0^{-1} are chosen large enough (according to the observations), then the penalized estimator obtained from the Lasso-like iteration step has null entries.

Lemma 4. *If for any $i, j \in \mathcal{P}$ we have*

$$\sum_{q,\ell}\frac{Z_{iq}Z_{j\ell}}{\lambda_{q\ell}} \geq \frac{n}{2}|S_{ij}|, \text{ when } i \neq j \quad \text{and} \quad \frac{1}{\lambda_0} \geq \frac{n}{2}|S_{ii}|, \tag{45}$$

then the solution $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{K}}^{-1}$ of problem (15) satisfies $\widehat{\mathbf{K}}^{-1} = 0$.

Proof. The sub-gradient equation arising from (15) gives

$$\forall i \neq j, \quad \frac{n}{2}\left(\widehat{K}^{-1}_{ij} - S_{ij}\right) - \left(\sum_{q,\ell}\frac{Z_{iq}Z_{j\ell}}{\lambda_{q\ell}}\right)\nu_{ij} = 0$$

$$\text{and } \forall i \in \mathcal{P}, \quad \frac{n}{2}\left(\widehat{K}^{-1}_{ii} - S_{ii}\right) - \frac{1}{\lambda_0}\nu_{ii} = 0,$$

where $\nu_{ij} \in \operatorname{sgn}(\widehat{K}_{ij})$ and thus $\nu_{ij} \in [-1, 1]$. In particular, we have

$$\forall i \neq j, \quad \frac{n}{2}\left|\widehat{K}^{-1}_{ij} - S_{ij}\right| \leq \left(\sum_{q,\ell}\frac{Z_{iq}Z_{j\ell}}{\lambda_{q\ell}}\right) \text{ and } \forall i \in \mathcal{P}, \ \frac{n}{2}\left|\widehat{K}^{-1}_{ii} - S_{ij}\right| \leq \frac{1}{\lambda_0}.$$

Now, if the set of penalty parameters satisfies the constraint (45), then the matrix $\mathbf{K}^{-1} = 0$ satisfies the sub-gradient equation. Thus, the conclusion comes from uniqueness of the solution to (15).

□