

Inference of Multiscale Gaussian Graphical Model

Do Edmond Sanou, Christophe Ambroise, Geneviève Robin

February 2021

Abstract

Gaussian Graphical Models (GGMs) are widely used for exploratory data analysis in various fields such as genomics, ecology, psychometry. In a high-dimensional setting, when the number of variables exceeds the number of observations by several orders of magnitude, the estimation of GGM is a difficult and unstable optimization problem. Clustering of variables or variable selection is often performed prior to GGM estimation. We propose a new method allowing to simultaneously infer a hierarchical clustering structure and the graphs describing the structure of independence at each level of the hierarchy. This method is based on solving a convex optimization problem combining a graphical lasso penalty with a fused type lasso penalty. Results on real and synthetic data are presented.

1 Introduction

Probabilistic graphical models (Lauritzen, 1996; Koller and Friedman, 2009) are widely used in high-dimensional data analysis to synthesize the interaction between variables. In many applications, such as genomics or image analysis, graphical models reduce the number of parameters by selecting the most relevant interactions between variables. Undirected Gaussian Graphical Models (GGMs) are a class of graphical models used in Gaussian settings. In the context of high-dimensional statistics, graphical models are generally assumed sparse, meaning that a small number of variables interact, compared to the total number of possible interactions. This assumption has been shown to provide both statistical and computational advantages by simplifying the structure of dependence between variables (Dempster, 1972) and allowing efficient algorithms (Meinshausen and Bühlmann, 2006). See, for instance, (Fan et al., 2016) for a review about sparse graphical models inference.

In GGMs, it is well known (see, e.g., Lauritzen (1996) that inferring the graphical model or equivalently the Conditional Independence Graph (CIG) boils down to inferring the support of the precision matrix Ω (the inverse of the variance-covariance matrix). To do so, several ℓ_1 penalized methods have been proposed in the literature to learn the CIG of GGMs. For instance, the neighborhood selection (Meinshausen and Bühlmann, 2006) (MB) based on a nodewise regression approach via the least absolute shrinkage and selection operator (Lasso,

Tibshirani (1996) is a popular method. Each variable is regressed on the others, taking advantage of the link between the so-obtained regression coefficients and partial correlations. More precisely, for all $1 \leq i \leq p$, the following problem is solved:

$$\hat{\beta}^i(\lambda) = \underset{\beta^i \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \frac{1}{n} \left\| \mathbf{X}^i - \mathbf{X}^{\setminus i} \beta^i \right\|_2^2 + \lambda \left\| \beta^i \right\|_1. \quad (1.1)$$

In Equation (1.1), λ is a non negative regularization parameter and $\mathbf{X}^{\setminus i}$ denotes the matrix \mathbf{X} deprived of column i . The MB method defined by the estimation problem (1.1) has generated a long line of work in the field of nodewise regression methods. For instance, Rocha et al. (2008), Ambroise et al. (2009) showed that nodewise regression could be seen as a pseudo-likelihood approximation, and Peng et al. (2009) extended the MB method to estimate sparse partial correlations using a single regression problem. Other inference methods similar to nodewise regression include a method based on Dantzig selector (Yuan, 2010) and the introduction of the Clime estimator (Cai et al., 2011).

Another family of sparse CIG inference methods directly estimates Ω via direct minimization of the ℓ_1 -penalized negative log-likelihood (Banerjee et al., 2008), without resorting to the auxiliary regression problem. This method, called the graphical lasso (Friedman et al., 2007), benefits from many optimization algorithms (Yuan and Lin, 2007; Rothman et al., 2008; Banerjee et al., 2008; Hsieh et al., 2014).

Such inference methods are widely used and enjoy many favorable theoretical and empirical properties, including robustness to high-dimensional problems. However, some limitations have been observed, particularly in the presence of strongly correlated variables. These limitations are caused by known impairments of Lasso-type regularization in this context (Bühlmann et al., 2012; Park et al., 2006). To overcome this, in addition to sparsity, several previous works attempt to estimate CIG by integrating clustering structures among variables for the sake of both statistical sanity and interpretability. A non-exhaustive list of works that integrate a clustering structure to speed up or improve the estimation procedure includes (Honorio et al., 2009; Ambroise et al., 2009; Mazumder and Hastie, 2012; Tan et al., 2013; Yao and Allen, 2019; Devijver and Gallopin, 2018).

The above methods exploit the group structure to simplify the graph inference problem and infer the CIG between single variables. Another question that has received less attention is the inference of the CIG between the groups of variables, i.e., between the meta-variables representative of the group structure. A recent work introducing inference of graphical models on multiple grouping levels is (Cheng et al., 2017). They proposed inferring the CIG of gene data on two levels corresponding to genes and pathways, respectively. Note that pathways are groups of functionally related genes known in advance. The inference is achieved by optimizing a penalized maximum likelihood that estimates a sparse network at both gene and group levels. Our work is also part of this dynamic. We introduce a penalty term allowing parsimonious networks to be built at different hierarchical clustering levels. The main difference with the

procedure of (Cheng et al., 2017) is that we do not require prior knowledge of the group structure, which makes the problem significantly more complex. In addition, our method has the advantage of proposing CIGs at more than two levels of granularity.

We introduce the Multiscale Graphical Lasso (MGLasso), a novel method to estimate simultaneously a hierarchical clustering structure, and graphical models depicting the conditional independence structure between clusters of variables at each level of the hierarchy. The procedure is based on a convex optimization problem with a hybrid penalty term combining a graphical Lasso and a group-fused Lasso penalty. In the spirit of convex hierarchical clustering, introduced by (Hocking et al., 2011; Lindsten et al., 2011), the hierarchy is obtained by spanning the entire regularization path. At each level of the hierarchy, variables in the same clusters are represented by a meta-variable; the new CIG is estimated between these new meta-variables, leading to a multiscale graphical model. Unlike (Yao and Allen, 2019), who introduced convex clustering in GGMs, our approach is expected to produce sparse estimations, thanks to an additional ℓ_1 penalty.

The remainder of this paper is organized as follows. In section 2, we formally introduce the Multiscale Graphical Lasso based on a convex estimation problem and an optimization algorithm based on the continuation of Nesterov’s smoothing technique (Hadj-Seleem et al., 2018). Section 4 presents numerical results on simulated and real data.

2 Multiscale Graphical Lasso

The proposed method aims at inferring a graphical Gaussian model while hierarchically grouping variables. It infers conditional independence between different groups of variables. The approach is based on neighborhood selection (Meinshausen and Bühlmann, 2006) and considers an additional fused-Lasso type penalty for clustering. In the spirit of hierarchical convex clustering, the hierarchical structure is recovered by spanning the regularization path.

Let $X = (X^1, \dots, X^p)^T \in \mathbb{R}^p$ be a p -dimensional Gaussian random vector, with mean vector μ and covariance matrix Σ . The conditional independence structure of X is characterized by a graph $G = (V, E)$, where $V = \{1, \dots, p\}$ is the set of variables and E the set of edges, uniquely determined by the support of the precision matrix $\Omega = \Sigma^{-1}$ (see, e.g., Dempster (1972)). In other words, for any two vertices $i, j \in V$, the edge (i, j) belongs to the set E if and only if $\Omega_{ij} \neq 0$, that is if and only if the i -th and j -th variables are conditionally independent given all the others i.e. $X^i \perp\!\!\!\perp X^j | X^{\setminus(i,j)}$ where $X^{\setminus(i,j)}$ is the set of all p variables deprived of variables i and j .

Considering the linear model $X^i = \sum_{j \neq i} \beta_j^i X^j + \epsilon_i$ where ϵ_i is a Gaussian centered random variable, we have $\beta_j^i = -\frac{\Omega_{ij}}{\Omega_{ii}}$. We define the regression matrix $\beta := [\beta^1, \dots, \beta^p]^T \in \mathbb{R}^{p \times (p-1)}$, whose rows are the regression vectors for each of the p regressions.

Let the $n \times p$ -dimensional matrix \mathbf{X} contain n independent observations of X . We propose to minimize the following criterion which combines Lasso and group-fused Lasso penalties:

$$J_{\lambda_1, \lambda_2}(\boldsymbol{\beta}; \mathbf{X}) = \frac{1}{2} \sum_{i=1}^p \left\| \mathbf{X}^i - \mathbf{X}^{\setminus i} \boldsymbol{\beta}^i \right\|_2^2 + \lambda_1 \sum_{i=1}^p \left\| \boldsymbol{\beta}^i \right\|_1 + \lambda_2 \sum_{i < j} \left\| \boldsymbol{\beta}^i - \tau_{ij}(\boldsymbol{\beta}^j) \right\|_2, \quad (2.1)$$

where τ_{ij} is a permutation exchanging the coefficients $\boldsymbol{\beta}_j^i$ and $\boldsymbol{\beta}_i^j$ and leaves other coefficients untouched, $\mathbf{X}^i \in \mathbb{R}^n$ denotes the i -th column of \mathbf{X} , $\boldsymbol{\beta}_i$ denotes the i -th row of $\boldsymbol{\beta}$, λ_1 and λ_2 are penalization parameters. Let us consider

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}} J_{\lambda_1, \lambda_2}(\boldsymbol{\beta}, \mathbf{X}).$$

The lasso penalty term encourages sparsity and the penalty term $\left\| \boldsymbol{\beta}^i - \tau_{ij}(\boldsymbol{\beta}^j) \right\|_2$ encourages to fuse regression vectors $\boldsymbol{\beta}^i$ and $\boldsymbol{\beta}^j$. These fusions uncover a clustering structure. The model is likely to cluster together variables that have the same conditional effects on the others. Variables X^i and X^j are assigned to the same cluster when $\boldsymbol{\beta}^i = \tau_{ij}(\boldsymbol{\beta}^j)$.

Let us illustrate by an example the effect of the proposed approach. If we consider a group of q variables whose regression vectors have at least q non-zero coefficients and further assume that for each pair of group variables i and j , $\left\| \boldsymbol{\beta}^i - \tau_{ij}(\boldsymbol{\beta}^j) \right\|_2 = 0$. After some permutations, we get a $q \times q$ block of non-zeros coefficient β_{ij} corresponding to the group in the $\boldsymbol{\beta}$ matrix, where $(i, j) \in \{1, \dots, q\}^2$. If we consider three different indices $i, j, k \in \{1, \dots, q\}^3$, it is straightforward to show that the six coefficients indexed by (i, j, k) are all equal. Thus the distance constraints between vectors $\boldsymbol{\beta}^i$ of a group forces equality of all regression coefficients in the group.

The greater the regularization weight λ_2 , the larger groups become. This is the core principle of the convex relaxation of hierarchical clustering introduced by Hocking et al. (2011). Hence, we can derive a hierarchical clustering structure by spanning the regularization path obtained by varying λ_2 while λ_1 is fixed. The addition of a fused-type term in graphical models inference has been studied previously by authors such as Honorio et al. (2009), Ganguly and Polonik (2014), Grechkin et al. (2015). However, these existing methods require prior knowledge of the neighborhood of each variable. On the contrary, our approach allows simultaneous inference of a multi-level graphical model and a hierarchical clustering of the variables.

In practice, if some information about the clustering structure is available, the problem can be generalized into:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^p \frac{1}{2} \left\| \mathbf{X}^i - \mathbf{X}^{\setminus i} \boldsymbol{\beta}^i \right\|_2^2 + \lambda_1 \sum_{i=1}^p \left\| \boldsymbol{\beta}^i \right\|_1 + \lambda_2 \sum_{i < j} w_{ij} \left\| \boldsymbol{\beta}^i - \tau_{ij}(\boldsymbol{\beta}^j) \right\|_2, \quad (2.2)$$

where w_{ij} is a positive weight encoding prior knowledge of the groups to which variables i and j belong to. In the remainder of the paper, we will consider $w_{ij} = 1$.

3 Numerical scheme

This section introduces a complete numerical scheme to apply MGLasso in practice, using a convex optimization algorithm and a model selection procedure. Section 3.1 reviews the principles of the Continuation with Nesterov smoothing in a shrinkage-thresholding algorithm (CONESTA, Hadj-Selem et al. (2018)), the optimization algorithm used in practice to solve MGLasso. Section 3.2 details a reformulation of MGLasso, which enables us to apply CONESTA. Finally, Section 3.3 presents the procedure used to select the regularization parameters.

3.1 Optimization via CONESTA algorithm

The optimization problem for Multiscale Graphical Lasso is convex but not straightforward to solve using classical algorithms because of the fused-lasso type penalty, which is non-separable and admits no closed-form solution for the proximal gradient. We rely on the Continuation with Nesterov smoothing in a shrinkage-thresholding algorithm (CONESTA, Hadj-Selem et al. (2018)), dedicated to high-dimensional regression problems with structured sparsity such as group structures.

The CONESTA solver, initially introduced for neuro-imaging problems, addresses a general class of convex optimization problems which includes group-wise penalties, admitting loss functions of the form:

$$f(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + \lambda_1 h(\boldsymbol{\theta}) + \lambda_2 s(\boldsymbol{\theta}),$$

where λ_1 and λ_2 are penalty weights, and $\boldsymbol{\theta} \in \mathbb{R}^d$ is a d -dimensional vector of parameters to optimize. In the original paper (Hadj-Selem et al., 2018), the function $g(\boldsymbol{\theta})$ is the sum of a least squares criterion and a ridge penalty, $h(\boldsymbol{\theta})$ is a penalty whose proximal operator is known in closed-form, and $s(\boldsymbol{\theta})$ is an $\ell_{1,2}$ penalty of the form

$$s(\boldsymbol{\theta}) = \sum_{\phi \in \Phi} \|\mathbf{D}_\phi \boldsymbol{\theta}_\phi\|_2.$$

In the definition of $s(\boldsymbol{\theta})$, $\Phi = \{\phi_1, \dots, \phi_{\text{Card}(\Phi)}\}$ is a set of subsets of indices, i.e., $\Phi_i \subset \{1, \dots, d\}$ for all $i \in \{1, \dots, \text{Card}(\Phi)\}$ and, for all $\phi \in \Phi$, $\boldsymbol{\theta}_\phi$ is the sub-vector of $\boldsymbol{\theta}$ defined by $\boldsymbol{\theta}_\phi = (\theta_i)_{i \in \phi}$. Finally, \mathbf{D}_ϕ are linear operators. The main ingredient of CONESTA is the approximation of the non-smooth $\ell_{2,1}$ -norm penalty with unknown proximal gradient, by a smooth function with known proximal gradient computed using Nesterov's smoothing. Given a smoothing parameter $\mu > 0$, let us define the smooth approximation

$$s_\mu(\boldsymbol{\theta}) = \max_{\boldsymbol{\alpha} \in \mathcal{K}} \left\{ \boldsymbol{\alpha}^T \mathbf{D} \boldsymbol{\theta} - \frac{\mu}{2} \|\boldsymbol{\alpha}\|_2^2 \right\},$$

where \mathcal{K} is the ℓ_2 unit ball. Note that $\lim_{\mu \rightarrow 0} s_\mu(\boldsymbol{\theta}) = s(\boldsymbol{\theta})$. An accelerated proximal gradient algorithm (FISTA, Beck and Teboulle (2009)) step can then be applied after computing the gradient of the smooth part of the approximated

criterion which is given by $g(\boldsymbol{\theta}) + s_\mu(\boldsymbol{\theta})$. At each iteration, the smoothing parameter μ is updated dynamically using the duality gap, and a new approximation is computed. The CONESTA algorithm enjoys a linear convergence rate, and was shown empirically to outperform other computational options for structured-sparsity problems such as ADMM and inexact FISTA in terms of convergence speed (Hadj-Selem et al., 2018).

3.2 Reformulation of MGLasso for CONESTA algorithm

To apply CONESTA, it is necessary to reformulate the MGLasso problem in order to comply with the form of loss function required by CONESTA. The objective of MGLasso can indeed be written as

$$\operatorname{argmin} \frac{1}{2} \|\mathbf{Y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}\|_2^2 + \lambda_1 \|\tilde{\boldsymbol{\beta}}\|_1 + \lambda_2 \sum_{i < j} \|\mathbf{A}_{ij}\tilde{\boldsymbol{\beta}}\|_2, \quad (3.1)$$

where $\mathbf{Y} = \operatorname{Vec}(\mathbf{X}) \in \mathbb{R}^{np}$, $\tilde{\boldsymbol{\beta}} = \operatorname{Vec}(\boldsymbol{\beta}) \in \mathbb{R}^{p(p-1)}$, $\tilde{\mathbf{X}}$ is a $\mathbb{R}^{[np] \times [p \times (p-1)]}$ block-diagonal matrix with $\mathbf{X}^{\setminus i}$ on the i -th block. The matrix \mathbf{A}_{ij} is a $p \times p(p-1)$ matrix defined by

$$\mathbf{A}_{ij}(k, l) = \begin{cases} 1, & \text{if } l = (i-1)p + k, \\ -1, & \text{if } l = (j-1)p + k, \\ 0, & \text{otherwise.} \end{cases}$$

Note that we introduce this notation for simplicity of exposition, but, in practice, the sparsity of the matrices \mathbf{A}_{ij} allows a more efficient implementation. Based on reformulation (3.1), we may apply CONESTA to solve the objective of MGLasso for fixed λ_1 and λ_2 . The procedure is applied, for fixed λ_1 , to a range of decreasing values of λ_2 to obtain a hierarchical clustering. The corresponding pseudo-code is given in the following algorithm where $(\mathbf{X}^i)^+$ denotes the pseudo-

inverse of \mathbf{X}^i and ϵ_{fuse} the threshold for merging clusters.

Algorithm 1: MGLasso algorithm

```

input : data  $\mathbf{X} \in \mathbb{R}^{n \times p}$ ,  $\lambda_1$ , starting value  $\lambda_2$ , step  $\kappa > 1$ 
clusters  $\leftarrow \{\{1\}, \dots, \{p\}\}$ 
 $\beta^i \leftarrow (\mathbf{X}^i)^+ \mathbf{X}^i$ 
while Card(clusters)  $\geq 2$  do
     $\beta \leftarrow CONESTA(\mathbf{X}, \mathbf{A}, \lambda_1, \lambda_2)$ 
    for  $i \leftarrow 1$  to  $p$  do
        for  $j \leftarrow i$  to  $p$  do
            if dist( $\beta^i, \beta^j$ )  $< \epsilon_{fuse}$  then
                clusters $i$  = clusters $j$ 
            end
        end
    end
    update: clusters
     $\lambda_2 \leftarrow \lambda_2 \times \kappa$ 
end
return :  $\beta \quad \forall (\lambda_1, \lambda_2)$ 

```

3.3 Model selection

A crucial question for practical applications is the definition of a rule to select the penalty parameters (λ_1, λ_2) . This selection problem operates at two levels: λ_1 controls the sparsity of the graphical model, and λ_2 controls the number of clusters in the optimal clustering partition. These two parameters are dealt with separately: the sparsity parameter λ_1 is chosen via model selection, while the clustering parameter λ_2 varies across a grid of values, in order to obtain graphs with different levels of granularity. The problem of model selection in graphical models is difficult in the high dimensional case where the number of samples is small compared to the number of variables, as classical AIC and BIC criteria tend to perform poorly (Liu et al., 2010). Alternative criteria have been proposed in the literature, such as cross-validation (Bien and Tibshirani, 2011), GGMSselect (Giraud et al., 2012), stability selection (Meinshausen and Bühlmann, 2010; Liu et al., 2010), Extended Bayesian Information Criterion (EBIC) (Foygel and Drton, 2010), and Rotation Information Criterion (Zhao et al., 2012).

In this paper, we focused on the StARS stability selection approach proposed by Liu et al. (2010). The method uses k subsamples of data to estimate the associated graphs for a given range of λ_1 values. For each value, a global instability of the graph edges is computed. The optimal value of λ_1 is chosen so as to minimize the instability, as follows. Let $\lambda_1^{(1)}, \dots, \lambda_1^{(K)}$ be a grid of sparsity regularization parameters, and S_1, \dots, S_N be N bootstrap samples obtained by sampling the rows of the data set \mathbf{X} . For each $k \in \{1, \dots, K\}$ and for each $j \in \{1, \dots, N\}$, we denote by $\mathcal{A}^{k,j}(\mathbf{X})$ the adjacency matrix of the estimated graph obtained by applying the inference algorithm to S_n with regularization

parameter $\lambda_1^{(k)}$. For each possible edge $(s, t) \in \{1, \dots, p\}^2$, the probability of edge appearance is estimated empirically by

$$\hat{\theta}_{st}^{(k)} = \frac{1}{N} \sum_{j=1}^N \mathcal{A}_{st}^{k,j}.$$

Define

$$\hat{\xi}_{st}(\Lambda) = 2\hat{\theta}_{st}(\Lambda) \left(1 - \hat{\theta}_{st}(\Lambda)\right)$$

the empirical instability of edge (s, t) (that is, twice the variance of the Bernoulli indicator of edge (s, t)). The instability level associated to $\lambda_1^{(k)}$ is given by

$$\hat{D}(\lambda_1^{(k)}) = \frac{\sum_{s < t} \hat{\xi}_{st}(\lambda_1^{(k)})}{\binom{p}{2}},$$

StARS selects the optimal penalty parameter as follows

$$\hat{\lambda} = \max_k \left\{ \lambda_1^{(k)} : \hat{D}(\lambda_1^{(k)}) \leq \beta, k \in \{1, \dots, K\} \right\},$$

where β is the threshold chosen for the instability level.

4 Simulation experiments

In this section, we conduct a simulation study to evaluate the performance of the MGLasso method, both in terms of clustering and support recovery. Receiver Operating Characteristic (ROC) curves are used to evaluate the adequacy of the inferred graphs with the reality for the MGLasso and GLasso methods in the Erdős-Renyi, Scale-free, and Stochastic Block Models frameworks. The Adjusted Rand indices are used to compare the partitions obtained with MGLasso, hierarchical agglomerative clustering, and K-means clustering in a stochastic block model framework.

4.1 Synthetic data models

We consider three different synthetic network models: the Stochastic Block Model (SBM, Fienberg and Wasserman (1981), the Erdős-Renyi model (Erdős et al., 1960) and the Scale-Free model (Newman et al., 2001). In each case, Gaussian data is generated by drawing n independent realizations of a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{p \times p}$ and $\Omega = \Sigma^{-1}$. The support of Ω , equivalent to the network adjacency matrix, is generated from the three different models. The difficulty level of the problem is controlled by varying the ratio $\frac{n}{p}$ with p fixed at 40: $\frac{n}{p} \in \{0.5, 1, 2\}$.

4.1.1 Stochastic Block-Model

We construct a block-diagonal precision matrix $\mathbf{\Omega}$ as follows. First, we generate the support of $\mathbf{\Omega}$ as shown in Figure 1, denoted by $\mathbf{A} \in \{0, 1\}^{p \times p}$. To do this, the variables are first partitioned into $K = 5$ hidden groups, noted C_1, \dots, C_K described by a latent random variable Z_i , such that $Z_i = k$ if $i \in C_k$. Z_i follows a multinomial distribution

$$P(Z_i = k) = \pi_k, \quad \forall k \in \{1, \dots, K\},$$

where $\pi = (\pi_1, \dots, \pi_K)$ is the vector of proportions of clusters whose sum is equal to one. The set of latent variables is noted $\mathbf{Z} = \{Z_1, \dots, Z_K\}$. Conditionally to \mathbf{Z} , A_{ij} follows a Bernoulli distribution such that

$$A_{ij}|Z_i = k, Z_j = l \sim \mathcal{B}(\alpha_{kl}), \quad \forall k, l \in \{1, \dots, K\},$$

where α_{kl} is the probability of inter-cluster connectivity, with $\alpha_{kl} = 0.01$ if $k \neq l$ and $\alpha_{ll} = 0.75$. For $k \in \{1, \dots, K\}$, we define $p_k = \sum_{i=1}^p \mathbf{1}_{\{Z_i=k\}}$. The precision matrix $\mathbf{\Omega}$ of the graph is then calculated as follows. We define $\Omega_{ij} = 0$ if $Z_i \neq Z_j$; otherwise, we define $\Omega_{ij} = A_{ij}\omega_{ij}$ where, for all $i \in \{1, \dots, p\}$ and for all $j \in \{1, \dots, p|Z_j = Z_i\}$, ω_{ij} is given by :

$$\begin{aligned} \omega_{ii} &:= \frac{1 + \rho(p_{Z_i} - 2)}{1 + \rho(p_{Z_i} - 2) - \rho^2(p_{Z_i} - 1)}; \\ \omega_{ij} &:= \frac{-\rho}{1 + \rho(p_{Z_i} - 2) - \rho^2(p_{Z_i} - 1)}. \end{aligned} \tag{4.1}$$

If α_{ll} were to be equal to one, this construction of $\mathbf{\Omega}$ would make it possible to control the level of correlation between the variables in each block to ρ . Introducing a more realistic scheme with $\alpha_{ll} = 0.75$ allows only to have an approximate control.

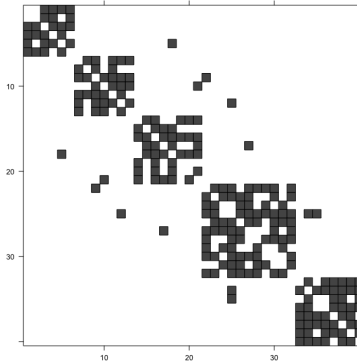


Figure 1: Adjacency matrix of a stochastic block model with 5 blocks.

4.1.2 Erdős-Renyi Model

The Erdős-Renyi model is a special case of the stochastic block model where $\alpha_{kl} = \alpha_{ll} = \alpha$ is constant. We set the density α of the graph to 0.1; see Figure 2 for an example of the graph resulting from this model.

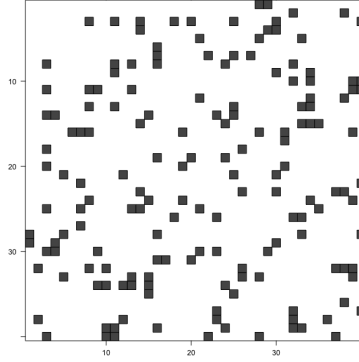


Figure 2: Adjacency matrix of an Erdős-Renyi model

4.1.3 Scale-free Model

The Scale-free Model generates networks whose degree distributions follow a power law. The graph starts with an initial chain graph of 2 nodes. Then, new nodes are added to the graph one by one. Each new node is connected to an existing node with a probability proportional to the degree of the existing node. We set the number of edges in the graph to 40. An example of scale-free graph is shown in Figure 3.

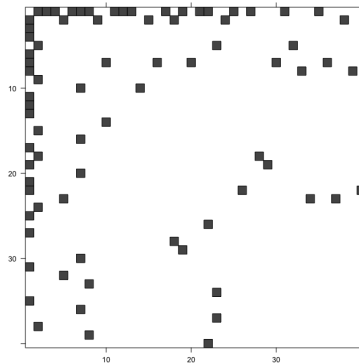


Figure 3: Adjacency matrix of a Scale-free model

4.2 Support recovery

We compare the network structure learning performance of our approach to that of GLasso in its neighborhood selection version using ROC curves. In both GLasso and MGLasso, the sparsity is controlled by a regularization parameter λ_1 ; however, MGLasso admits an additional regularization parameter, λ_2 , which controls the strength of convex clustering. To compare the two methods, in each ROC curve, we vary the parameter λ_1 while the parameter λ_2 (for MGLasso) is kept constant. We computed ROC curves for 4 different penalty levels for the λ_2 parameter; since GLasso does not depend on λ_2 , the GLasso ROC curves are replicated.

In a decision rule associated with a sparsity penalty level λ_1 , we recall the definition of the two following functions. The sensitivity, also called the true positive rate or recall, is given by :

$$\lambda_1 \mapsto \text{sensitivity}(\lambda_1) = \frac{TP(\lambda_1)}{TP(\lambda_1) + FN(\lambda_1)}.$$

Specificity, also called true negative rate or selectivity, is defined as follows:

$$\lambda_1 \mapsto \text{specificity}(\lambda_1) = \frac{TN(\lambda_1)}{TN(\lambda_1) + FP(\lambda_1)}.$$

The ROC curve with the parameter λ_1 represents $\text{sensitivity}(\lambda_1)$ as a function of $1 - \text{specificity}(\lambda_1)$ which is the false positive rate.

For each configuration (n, p fixed), we generate 50 replications and their associated ROC curves, which are then averaged. The average ROC curves for the three models are given in Figure 4, Figure 5 and Figure 6 by varying $\frac{n}{p} \in \{0.5, 1, 2\}$.

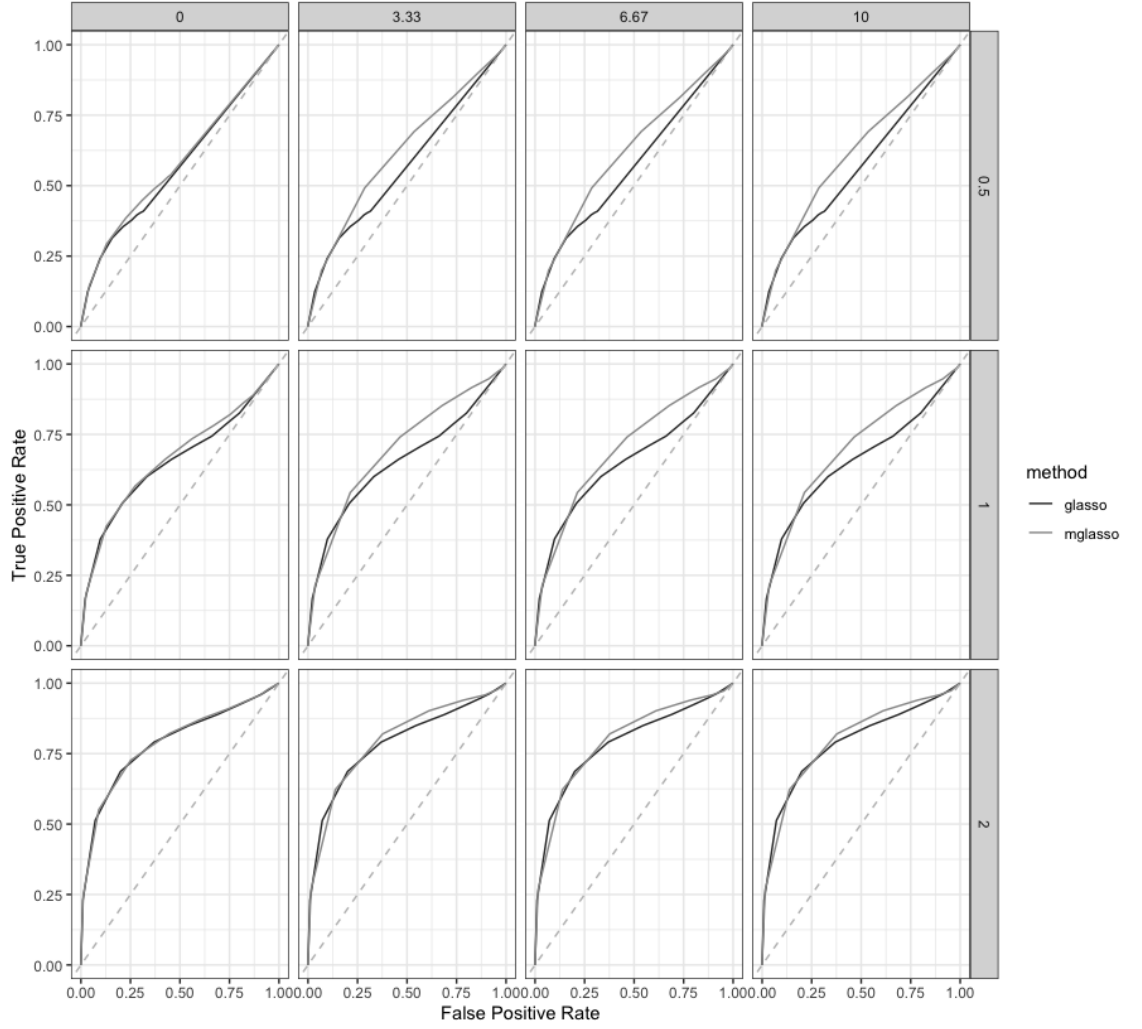


Figure 4: ROC curves for the Erdős-Renyi model comparing MGLasso and GLasso methods. The ratio $\frac{n}{p} \in \{0.5, 1, 2\}$ and the total variation penalty $\lambda_2 \in \{0, 3.33, 6.67, 10\}$

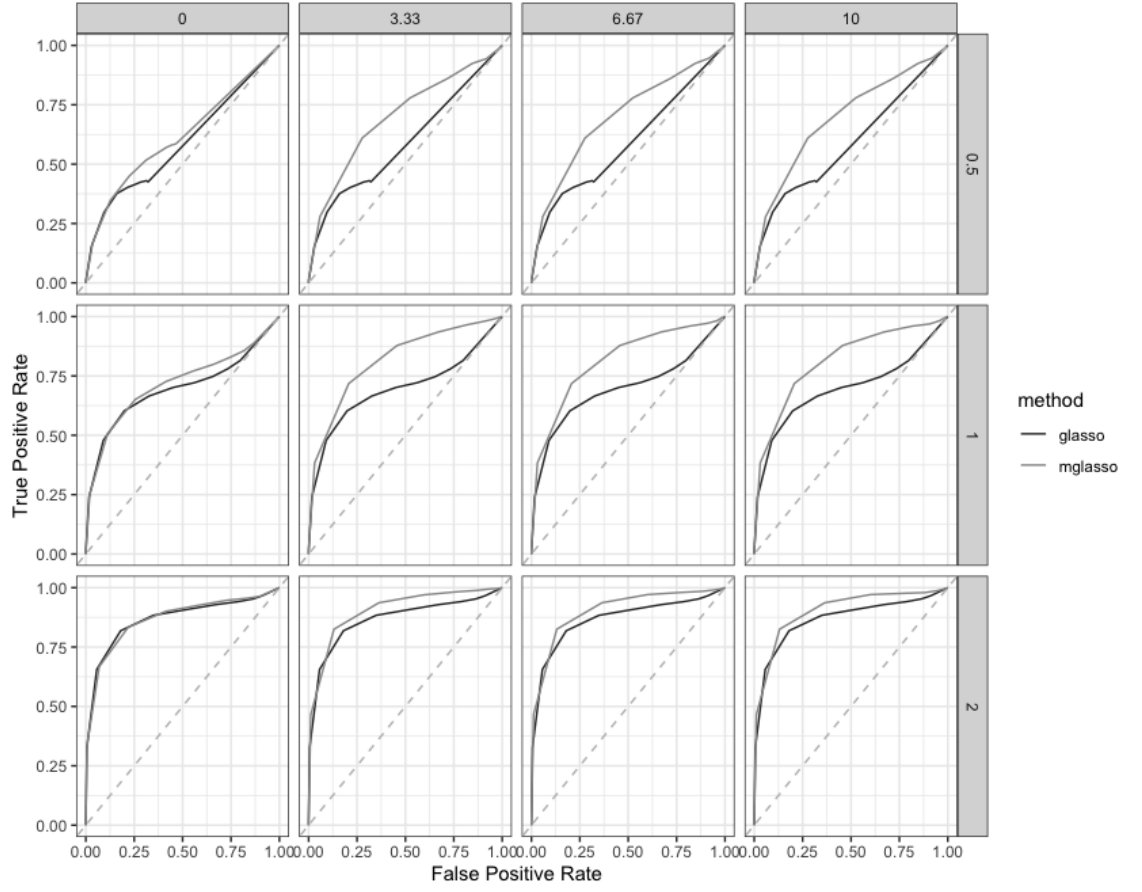


Figure 5: ROC curves for the Scale-free model comparing MGLasso and GLasso methods. The ratio $\frac{n}{p} \in \{0.5, 1, 2\}$ and the total variation penalty $\lambda_2 \in \{0, 3.33, 6.67, 10\}$

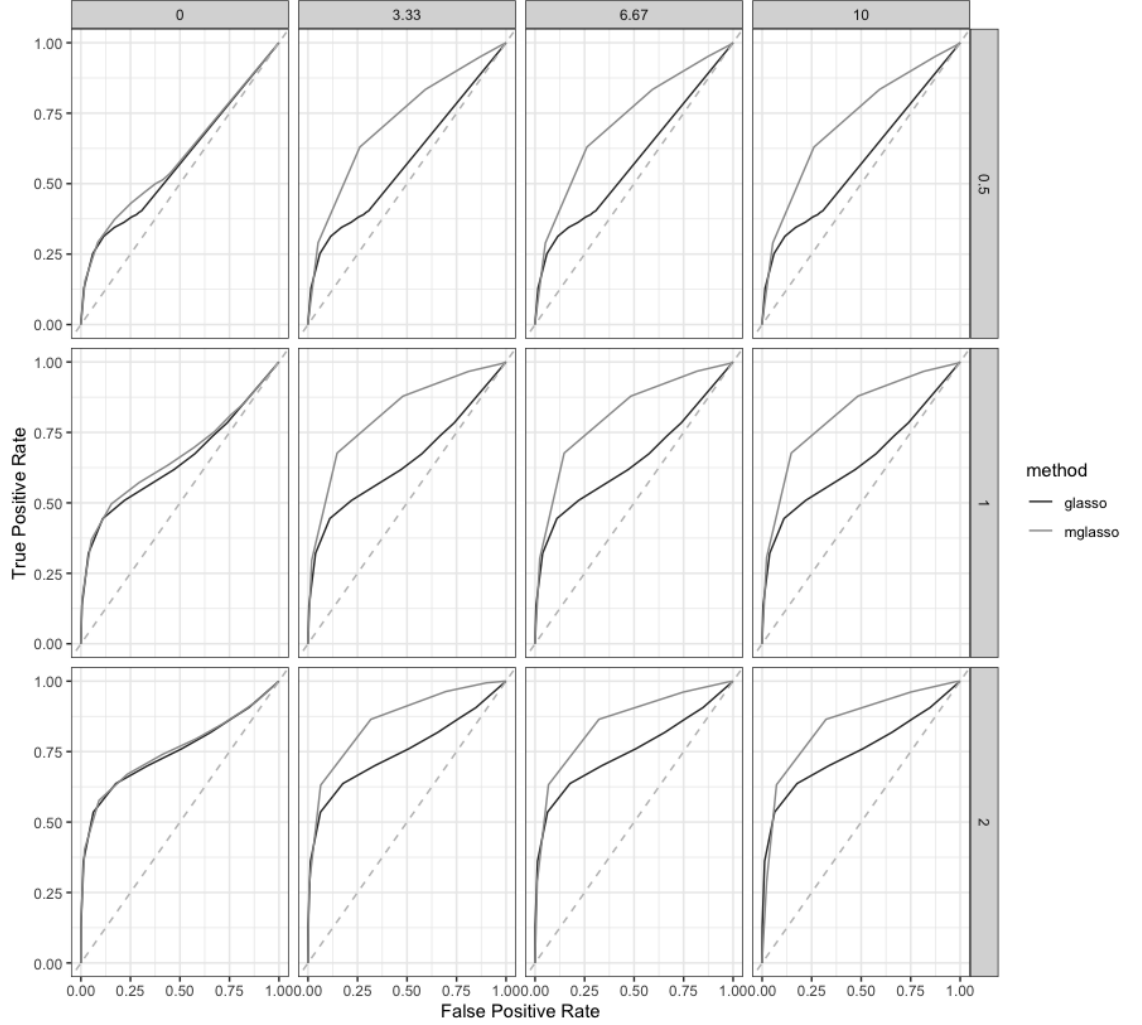


Figure 6: ROC curves for the Stochastic Block model comparing MGLasso and GLasso methods. The ratio $\frac{n}{p} \in \{0.5, 1, 2\}$ and the total variation penalty $\lambda_2 \in \{0, 3.33, 6.67, 10\}$

Based on these empirical results, we first observe that, in all the considered simulation models, MGLasso improves over GLasso in terms of support recovery in the high-dimensional setting where $p < n$. In addition, in the absence of a total variation penalty, i.e., $\lambda_2 = 0$, MGLasso performs no worse than GLasso in each of the 3 models. However, for $\lambda_2 > 0$, increasing penalty value does not seem to significantly improve the support recovery performances for the MGLasso, as we observe similar results for $\lambda_2 = 3.3, 6.6, 10$. Preliminary analyses show that, as λ_2 increases, the estimates of the regression vectors are shrunk towards 0.

This shrinkage effect of group-fused penalty terms was also observed in (Chu et al., 2021).

4.3 Clustering

In order to obtain clustering performance, we compared the partitions estimated by MGLasso, Hierarchical Agglomerative Clustering (HAC) with Ward's distance and K-means to the true partition in a stochastic block model framework. Euclidean distances between variables are used for HAC and K-means. The criterion used for the comparison is the adjusted Rand index. We studied the influence of the correlation level inside clusters on the clustering performances through two different parameters: $\rho \in \{0.1, 0.3\}$; the vector of cluster proportions is fixed at $\pi = (1/5, \dots, 1/5)$. We then simulate 100 Gaussian data sets for each simulation configuration (ρ , n/p fixed). The optimal sparsity penalty for MGLasso is chosen by the Stability Approach to Regularization Selection (StARS) method (Liu et al., 2010), and we vary the parameter λ_2 .

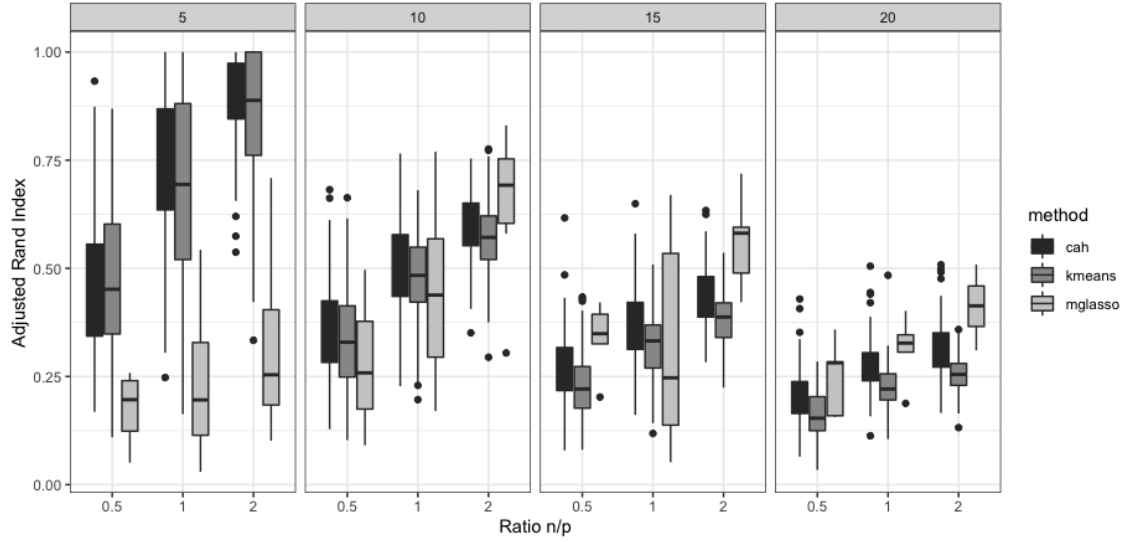


Figure 7: Adjusted Rand Indices for the HAC, k-means and MGLasso methods. Performances are observed for 4 different number of clusters in a high correlation context

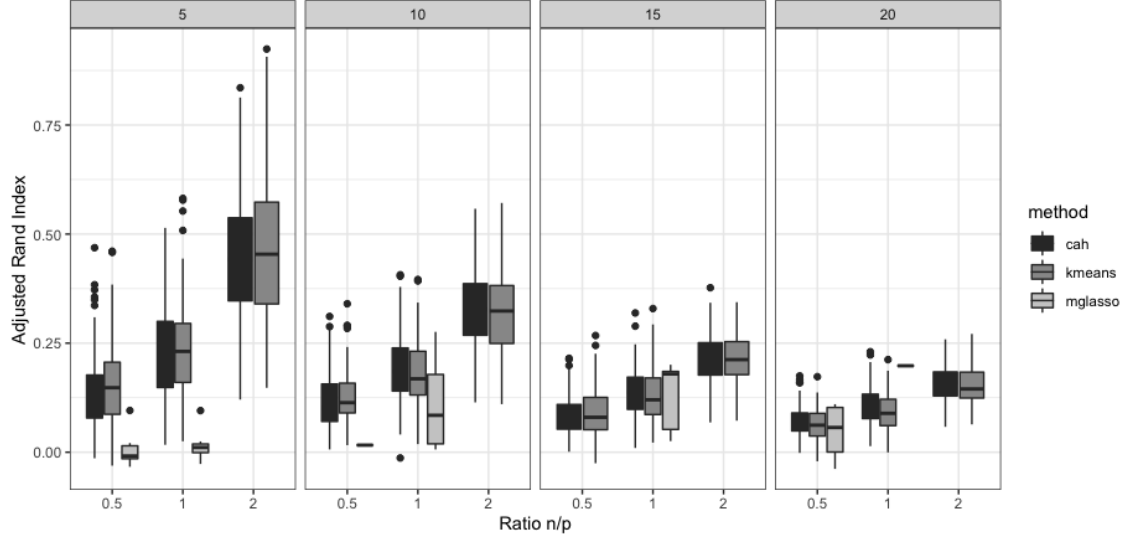


Figure 8: Adjusted Rand Indices for the HAC, k-means and MGLasso methods. Performances are observed for 4 different number of clusters in a low correlation context

The results shown in Figure 7 and Figure 8 demonstrate that, particularly at the lower to medium levels of the hierarchy (between 20 and 10 clusters), the hierarchical clustering structure uncovered by MGLasso is comparable to popular clustering methods used in practice. For the higher levels (5 clusters), the performances of MGLasso deteriorate. As expected, the three compared methods also deteriorate as the level of correlation inside clusters decreases.

5 Analysis of microbial associations data

We finally illustrate our new method of inferring the multiscale Gaussian graphical model, with an application to the analysis of microbial associations in the American Gut Project. The data used are count data that have been previously normalized by applying the log-centered ratio technique as used in (Kurtz et al., 2015). After some filtering steps (Kurtz et al., 2015) on the operational taxonomic units OTUs counts (removed if present in less than 37% of the samples) and the samples (removed if sequencing depth below 2700), the top OTUs are grouped in a dataset composed of $n_1 = 289$ for 127 OTUs. As a preliminary analysis, we perform a hierarchical agglomerative clustering (HAC) on the OTUs, which allows us to identify four significant groups. The correlation matrix of the dataset is given in Figure 9; variables have been rearranged according to the HAC partition.

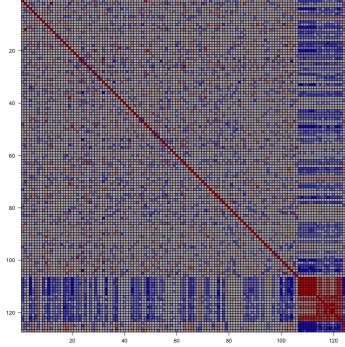
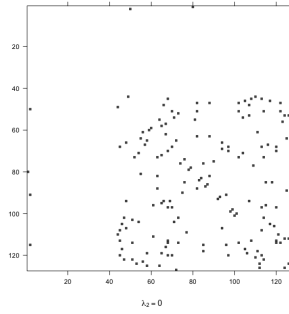


Figure 9: Empirical correlations in the gut data

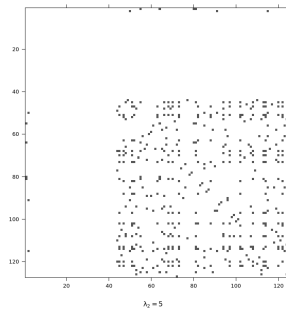
The average correlations within blocks of variables belonging to the same cluster are given below. We observe relatively high levels of correlation in small blocks, similar to the simulation models used to evaluate the performance of clustering in the section 4.

| Clusters | Mean correlation |
|----------|------------------|
| 1 | 0.0127 |
| 2 | 0.815 |
| 3 | 0.555 |
| 4 | 0.566 |

We apply MGLasso to the normalised counts to infer a graph and a clustering structure. The graphs obtained by MGLasso for $\lambda_2 = 0$ and for $\lambda_2 = 5$ (corresponding respectively 127 and 80 clusters) are given below. In each case, the parameter λ_1 is chosen by stability selection (see section 3.3).



(a) MGLasso with $\lambda_2 = 0$



(b) MGLasso with $\lambda_2 = 5$

Figure 10: Inferred graphs using MGLasso for different values of λ_2

The variables were reordered according to the clustering partition obtained from

the distances between the regression vectors. Increasing λ_2 reduces the number of clusters and leads to a shrinking effect on the estimates. The adjacency matrix of the neighborhood selection equivalent to setting λ_2 to 0 is given in Figure 10a (left). In Figure 10b (right), the deduced partition is composed of 80 clusters. A confusion matrix comparing the edges deduced by MGLasso with $\lambda_2 = 5$ and neighborhood selection is given below. Adding a total variation parameter increases the merging effect, resulting in a larger number of edges in the graph.

| | Neighborhood selection | | |
|--------------------------------|------------------------|-----------|-------|
| MGLasso ($\lambda_2 = 5$) | non-edges | non-edges | edges |
| | edges | 15678 | 0 |
| | | 288 | 163 |

6 Conclusion

We proposed a new technique that combines Gaussian Graphical Model inference and hierarchical clustering called MGLasso. The method proceeds via convex optimization and minimizes the neighborhood selection objective penalized by a hybrid regularization combining a sparsity-inducing norm and a convex clustering penalty. We developed a complete numerical scheme to apply MGLasso in practice, with an optimization algorithm based on CONESTA and a model selection procedure. Our simulations results over synthetic and real datasets showed that MGLasso can perform better than GLasso in network support recovery in the presence of groups of correlated variables, and we illustrated the method with the analysis of microbial associations data. The present work paves the way for future improvements: first, by incorporating prior knowledge through more flexible weighted regularization; second, by studying the theoretical properties of the method in terms of statistical guarantees for the MGLasso estimator.

References

- Ambroise, C., Chiquet, J., and Matias, C. (2009). Inferring sparse gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3(0):205–238.
- Banerjee, O., El Ghaoui, L., and d’Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. 9:485–516.
- Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2:183–202.
- Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820.

- Bühlmann, P., Rütimann, P., Van De Geer, S., and Zhang, C.-H. (2012). Correlated variables in regression: clustering and sparse estimation.
- Cai, T., Liu, W., and Luo, X. (2011). A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.
- Cheng, L., Shan, L., and Kim, I. (2017). Multilevel Gaussian graphical model for multilevel networks. *Journal of Statistical Planning and Inference*, 190:1–14.
- Chu, S., Jiang, H., Xue, Z., and Deng, X. (2021). Adaptive convex clustering of generalized linear models with application in purchase likelihood prediction. *Technometrics*, 63(2):171–183.
- Dempster, A. P. (1972). Covariance Selection. *Biometrics*, 28(1):157.
- Deijver, E. and Gallopin, M. (2018). Block-Diagonal Covariance Selection for High-Dimensional Gaussian Graphical Models. *Journal of the American Statistical Association*, 113(521):306–314.
- Erdős, P., Rényi, A., et al. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60.
- Fan, J., Liao, Y., and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32.
- Fienberg, S. E. and Wasserman, S. S. (1981). Categorical data analysis of single sociometric relations. *Sociological methodology*, 12:156–192.
- Foygel, R. and Drton, M. (2010). Extended bayesian information criteria for gaussian graphical models. *arXiv preprint arXiv:1011.6640*.
- Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso.
- Ganguly, A. and Polonik, W. (2014). Local neighborhood fusion in locally constant gaussian graphical models.
- Giraud, C., Huet, S., and Verzelen, N. (2012). Graph selection with ggmselect. *Statistical applications in genetics and molecular biology*, 11(3).
- Grechkin, M., Fazel, M., Witten, D., and Lee, S.-I. (2015). Pathway graphical lasso. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI’15, page 2617–2623. AAAI Press.
- Hadj-Selim, F., Lofstedt, T., Dohmatob, E., Frouin, V., Dubois, M., Guillemot, V., and Duchesnay, E. (2018). Continuation of Nesterov’s Smoothing for Regression with Structured Sparsity in High-Dimensional Neuroimaging. *IEEE Transactions on Medical Imaging*, 2018.

- Hocking, T., Vert, J.-P., Bach, F., and Joulin, A. (2011). Clusterpath: an algorithm for clustering using convex fusion penalties. In *ICML*.
- Honorio, J., Samaras, D., Paragios, N., Goldstein, R., and Ortiz, L. E. (2009). Sparse and locally constant gaussian graphical models. *Advances in Neural Information Processing Systems*, 22:745–753.
- Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., and Ravikumar, P. (2014). Quic: Quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research*, 15(83):2911–2947.
- Koller, D. and Friedman, N. (2009). Probabilistic Graphical Models: Principles. *Italica*, 51(3):327.
- Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLOS Computational Biology*, 11:1–25.
- Lauritzen, S. L. (1996). *Graphical models*. Clarendon Press.
- Lindsten, F., Ohlsson, H., and Ljung, L. (2011). Clustering using sum-of-norms regularization: With application to particle filter output computation. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pages 201–204.
- Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010*, pages 1–14.
- Mazumder, R. and Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, 6(none):2125 – 2149.
- Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462.
- Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.
- Newman, M. E., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2):026118.
- Park, M. Y., Hastie, T., and Tibshirani, R. (2006). Averaged gene expressions for regression. *Biostatistics*, 8(2):212–227.
- Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746. PMID: 19881892.
- Rocha, G. V., Zhao, P., and Yu, B. (2008). A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice).

- Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2(none):494 – 515.
- Tan, K. M., Witten, D., and Shojaie, A. (2013). The Cluster Graphical Lasso for improved estimation of Gaussian graphical models.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288.
- Yao, T. and Allen, G. I. (2019). Clustered gaussian graphical model via symmetric convex clustering. In *2019 IEEE Data Science Workshop (DSW)*, pages 76–82.
- Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(79):2261–2286.
- Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.
- Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in r. *The Journal of Machine Learning Research*, 13(1):1059–1062.