An introduction to graph analysis and modeling

Network Inference with Sparse Graphical Models

Julien Chiquet

September, 2019

https://github.com/jchiquet/CourseNetworkLondon

Imperial College London

INRA SCIENCE & IMPACT

Statistical analysis of Networks

Different questions

Understanding the network topology

- Data = observed network
- Questions: central nodes? cluster structure? small-world property?

Inferring/Reconstructing the network

- Data = repeated signal observed at each node
- Questions: which nodes are connected?

Each to be combined with

covariates, time, heterogeneous data set, missing data, ...

Reconstruction and analysis of biological networks

E. coli regulatory network

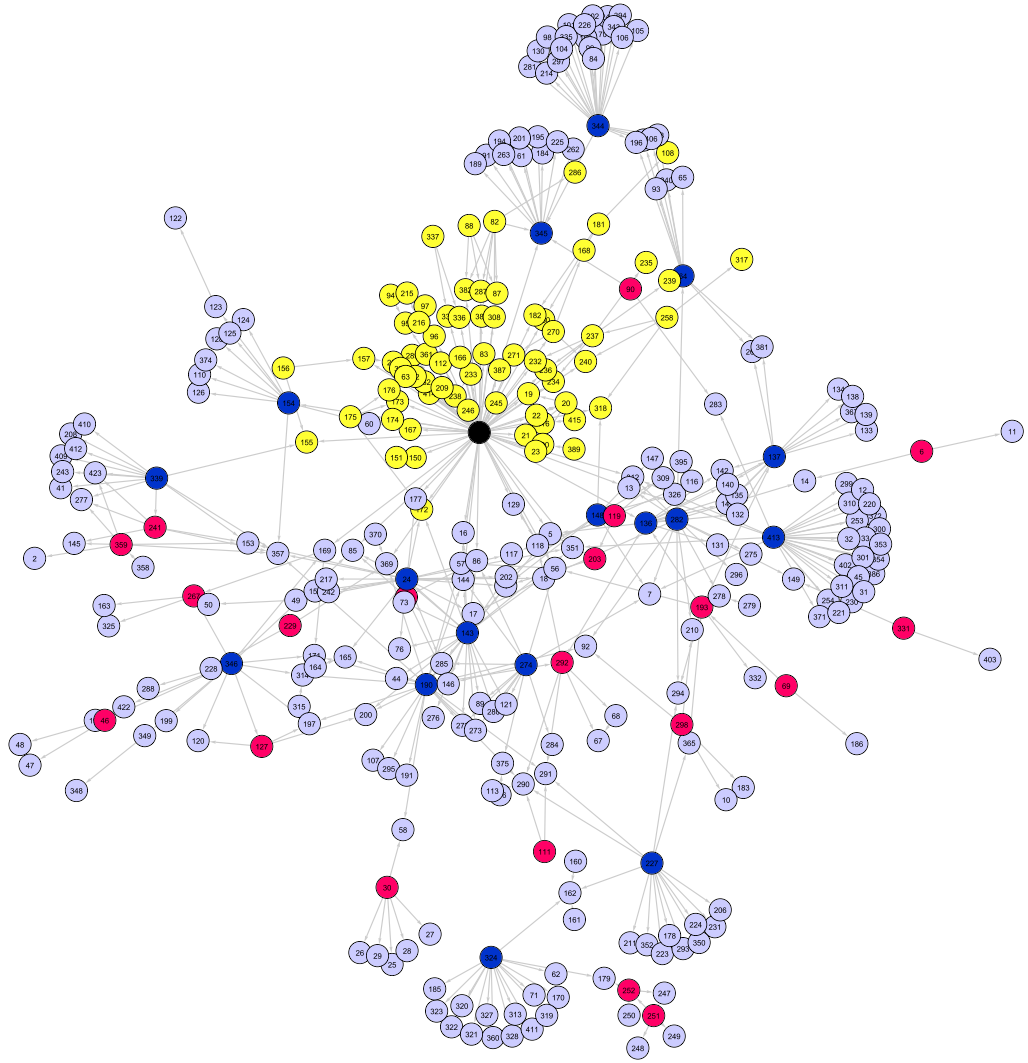
Target network

Relations between genes and their products

- highly structured
- always incomplete

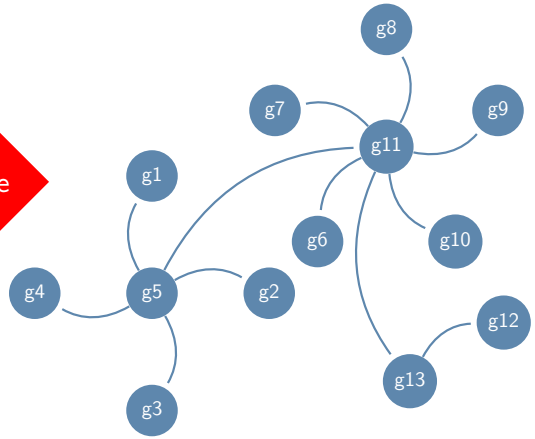
Data and method

- transcriptomic data
- Inference: sparse Gaussian graphical model
- Analysis: Stochastic Block Model



A challenging problem

Inference



$\approx$ 10s/1,000s assays

$\approx$ 1,000s/1,000,000s features

Model point of view

1. Nodes (genes, OTUS, ...)
   - fixed variables
2. Edges (biological interactions)
   - use (partial) correlations or others fancy statistical concepts
3. Data (intensities, counts)
   - a tidy $n \times p$ dat matrix

$\rightsquigarrow$ Quantities and goals well defined

Data point of view: non classical statistics

- (Ultra) High dimensionality ($n < p$, $n \ll p$)
- Heterogeneous data

Biological point of view: not well defined goals and questions

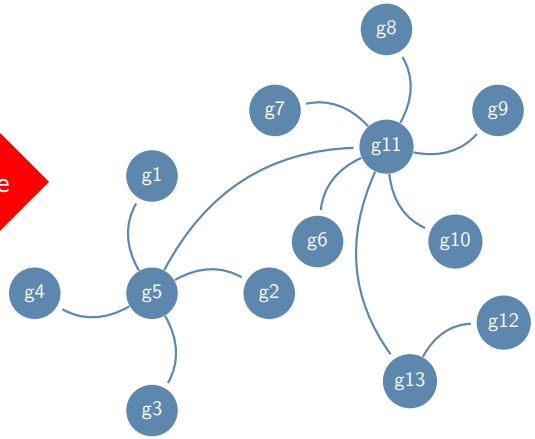
- What interaction? Direct? Indirect? Causal?
- Whole network? Subnetwork? Groups of key actors?
- structured data, mixed data

A challenging problem

≈ 10s/1,000s assays

≈ 1,000s/1,000,000s features

Inference



Model point of view

1. Nodes (genes, OTUS, ...)
   - fixed variables
2. Edges (biological interactions)
   - use (partial) correlations or others fancy statistical concepts
3. Data (intensities, counts)
   - a tidy $n \times p$ dat matrix

$\rightsquigarrow$ Quantities and goals well defined

Data point of view: non classical statistics

- (Ultra) High dimensionality ($n < p$, $n \lll p$)
- Heterogeneous data

Biological point of view: not well defined goals and questions

- What interaction? Direct? Indirect? Causal?
- Whole network? Subnetwork? Groups of key actors?
- structured data, mixed data

Outline

Outline

Correlation networks

Correlation (association network)

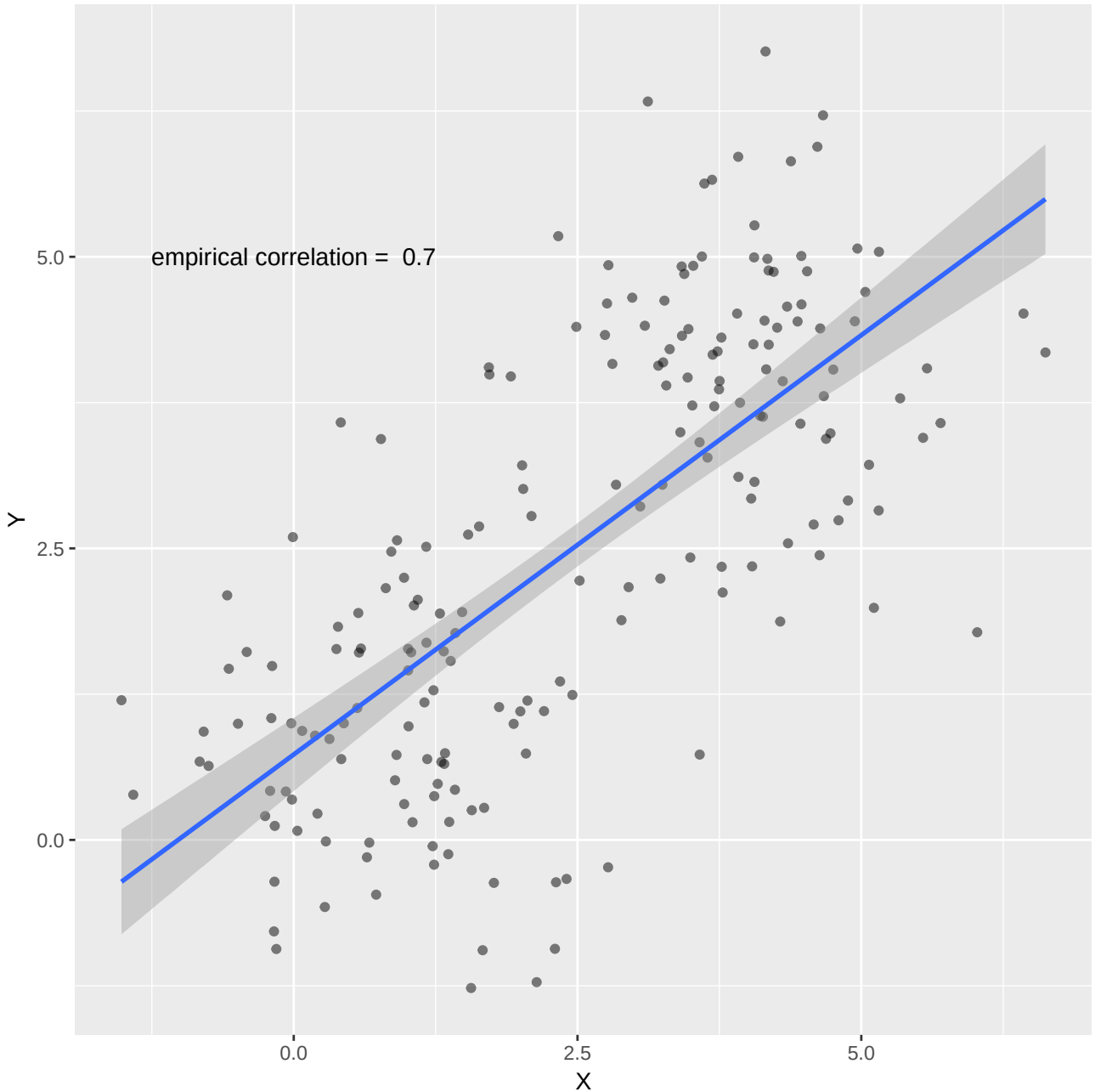
Similar expression profile $\rightsquigarrow$ high-correlation

1. Compute the correlation matrix (Pearson, Spearman, ...)
2. Predict an edge between two actors if their absolute correlation is above a given threshold

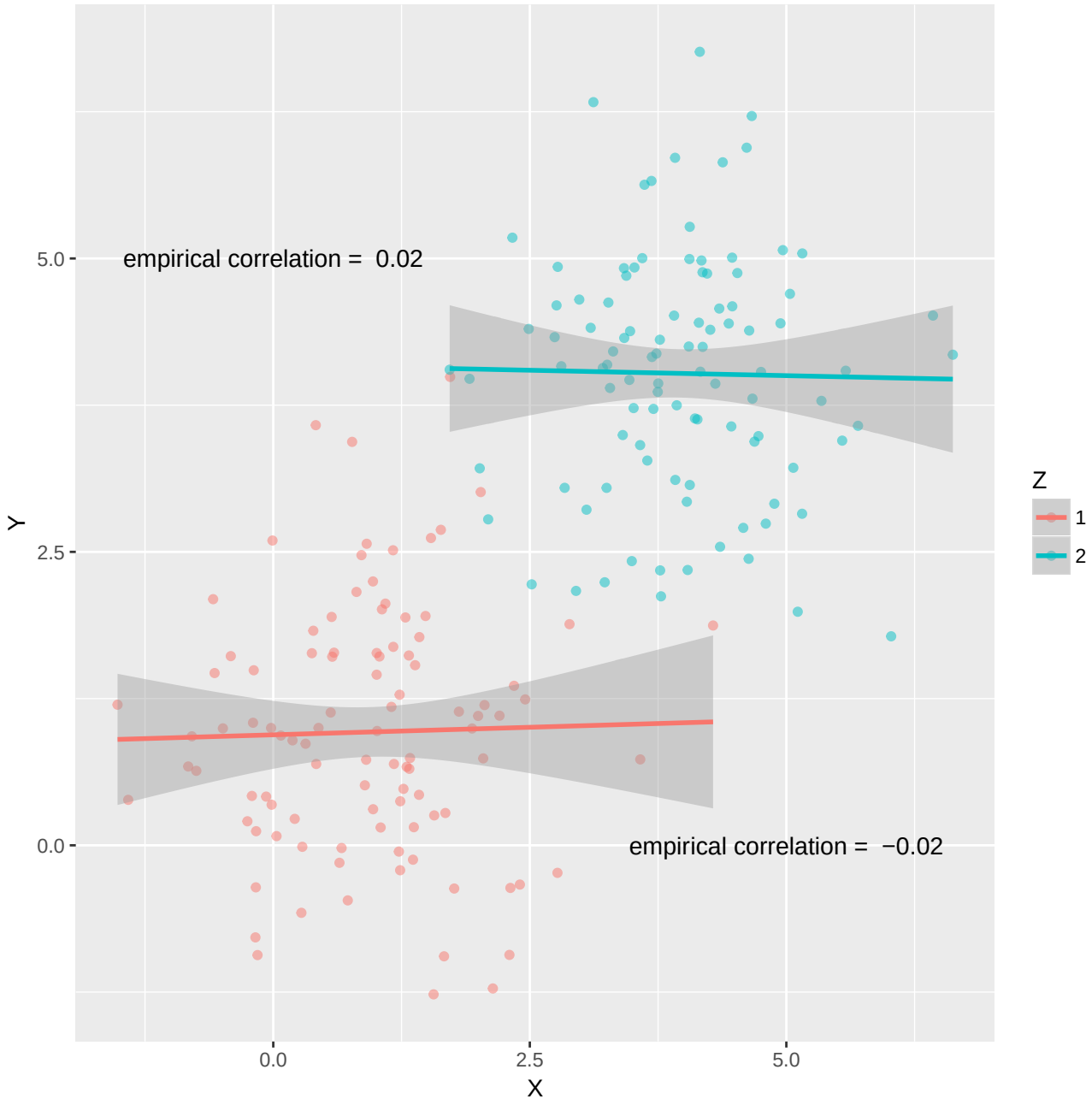
Questions

- How to set up the threshold?
- If we target actors with similar profiles, why not clustering?
- Information is drowned (all actors are correlated ...)

Limits of correlation for network reconstruction



Limits of correlation for network reconstruction



Graphical models

Definition

A graphical model gives a graphical (intuitive) representation of the dependence structure of a probability distribution, by linking

1. a random vector (or a set of random variables.) $X = \{X_1, \ldots, X_p\}$ with distribution $\mathbb{P}$,

2. a graph $\mathcal{G} = (\mathcal{P}, \mathcal{E})$ where
   - $\mathcal{P} = \{1, \ldots, p\}$ is the set of nodes associated to each variable,
   - $\mathcal{E}$ is a set of edges describing the dependence relationship of $X \sim \mathbb{P}$.

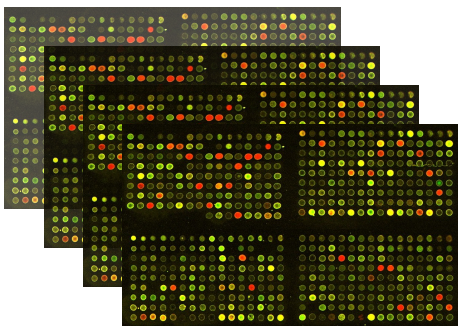
Conditional independence graph

It is the undirected graph $\mathcal{G} = \{\mathcal{P}, \mathcal{E}\}$ where

$$(i, j) \notin \mathcal{E} \Leftrightarrow X_i \perp\!\!\!\perp X_j | \mathcal{P} \backslash \{i, j\}.$$

The Gaussian case

The data



Inference

$$\mathbf{X} = \begin{pmatrix} x_1^1 & x_1^2 & x_1^3 & \dots & x_1^p \\ \vdots & & & & \\ x_n^1 & x_n^2 & x_1^2 & \dots & x_n^p \end{pmatrix}$$

Assuming $f_X(\mathbf{X})$ multivariate Gaussian

Greatly simplifies the inference:

- $\rightsquigarrow$ naturally links independence and conditional independence to the covariance and partial covariance,
- $\rightsquigarrow$ gives a straightforward interpretation to the graphical modeling previously considered.

Why Gaussianity helps?

Case of 2 variables or size-2 random vector

Let X, Y be two real random variables.

Definitions

$$\mathrm{cov}(X,Y) = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\Big] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

$$\rho_{XY} = \mathrm{cor}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathbb{V}(X) \cdot \mathbb{V}(Y)}}.$$

Proposition

- $\mathrm{cov}(X,X) = \mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$,
- $\mathrm{cov}(X+Y,Z) = \mathrm{cov}(X,Z) + \mathrm{cov}(X,Z)$,
- $\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + \mathrm{cov}(X,Y)$.
- $X \perp\!\!\!\perp Y \Rightarrow \mathrm{cov}(X,Y) = 0$.
- $X \perp\!\!\!\perp Y \Leftrightarrow \mathrm{cov}(X,Y) = 0$ *when* X, Y *are Gaussian.*

The bivariate Gaussian distribution

The Covariance Matrix

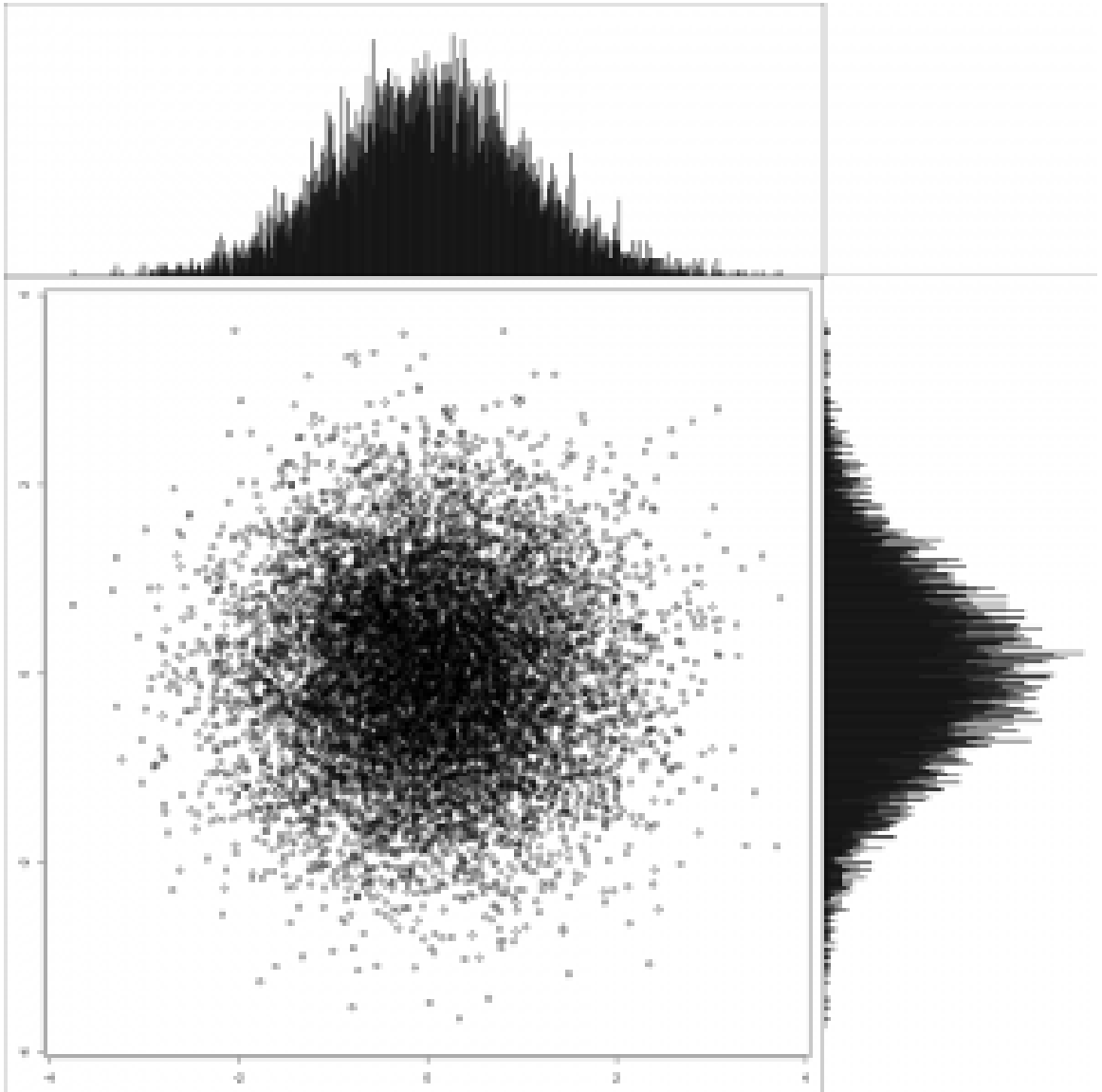
Let

$$X \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}),$$

with unit variance and $\rho_{XY} = 0$

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The shape of the 2-D distribution evolves accordingly.



The bivariate Gaussian distribution

The Covariance Matrix

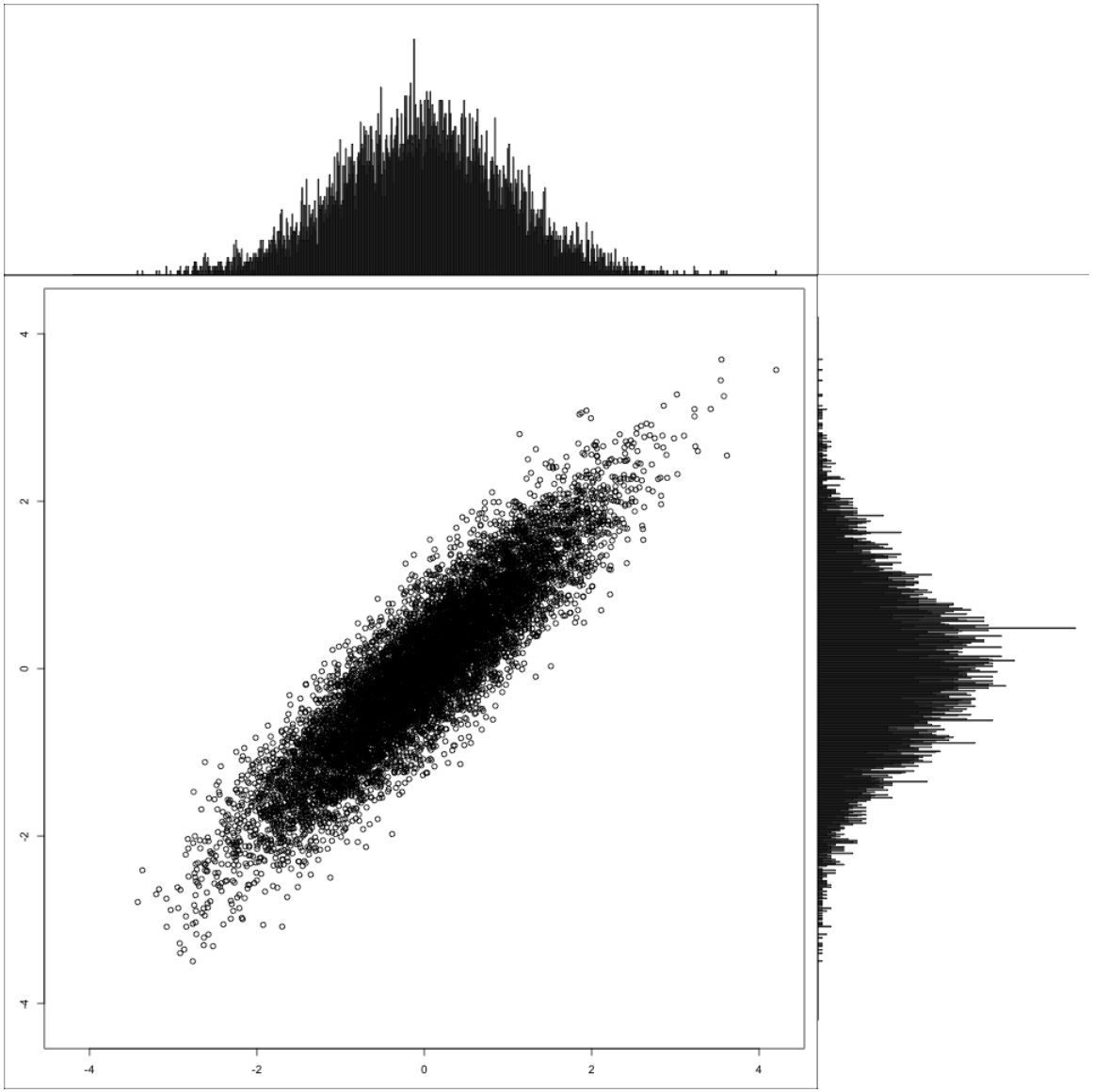
Let

$$X \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}),$$

with unit variance and $\rho_{XY} = 0.9$

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}.$$

The shape of the 2-D distribution evolves accordingly.



Generalization: multivariate Gaussian vector

Now need partial covariance and partial correlation

Let X, Y, Z be real random variables.

Definitions

$$\mathrm{cov}(X, Y|Z) = \mathrm{cov}(X, Y) - \mathrm{cov}(X, Z)\mathrm{cov}(Y, Z)/\mathbb{V}(Z).$$

$$\rho_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2}}.$$

$\rightsquigarrow$ Give the interaction between X and Y once removed the effect of Z.

Proposition

When X, Y, Z are jointly Gaussian, then

$$\mathrm{cov}(X, Y|Z) = 0 \Leftrightarrow \mathrm{cor}(X, Y|Z) = 0 \Leftrightarrow X \perp\!\!\!\perp Y|Z.$$

Gaussian Graphical Model: canonical settings

Biological experiments in comparable Gaussian conditions

Profiles of a set $\mathcal{P} = \{1, \ldots, p\}$ of genes is described by $X \in \mathbb{R}^p$ such as

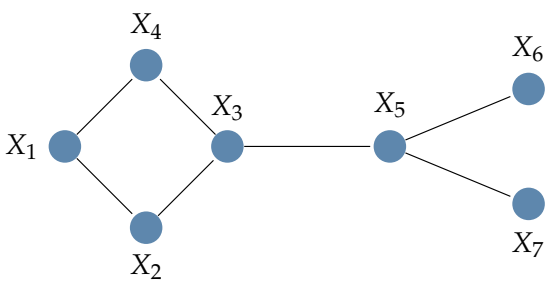
1. $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ the precision matrix.
2. a sample $(X^1, \ldots, X^n)$ of exp. stacked in an $n \times p$ data matrix $\mathbf{X}$.

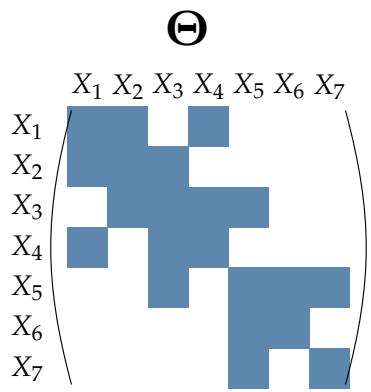
Conditional independence structure

$$(i,j) \notin \mathcal{E} \Leftrightarrow X_i \perp\!\!\!\perp X_j | X_{\setminus\{i,j\}} \Leftrightarrow \Theta_{ij} = 0.$$

Graphical interpretation

$$\mathcal{G} = (\mathcal{P}, \mathcal{E})$$





$\rightsquigarrow$ "Covariance" selection

Gaussian Graphical Model and Linear Regression

Linear regression viewpoint

Gene expression X_i is linearly explained by the other genes':

$$X_i | X_{\setminus i} = -\sum_{j \neq i} \frac{\Theta_{ij}}{\Theta_{ii}} X_j + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \Omega_{ii}^{-1}), \quad \varepsilon_i \perp X$$

Conditional on its neighborhood, other profiles do not give additional insights

$$X_i | X_{\setminus i} = \sum_{j \in \text{neighbors}(i)} \beta_j X_j + \varepsilon_i \quad \text{with } \beta_j = -\frac{\Theta_{ij}}{\Theta_{ii}}.$$

$\rightsquigarrow$ "Neighborhood" selection

Gaussian Graphical Model and AR process (1)

Time course data

Time course- data experiment can be represented as a multivariate vector $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$, generated through a first order vector autoregressive process $VAR(1)$:

$$X^t = \mathbf{\Theta} X^{t-1} + \mathbf{b} + \boldsymbol{\varepsilon}^t, \quad t \in [1, n]$$

where $\boldsymbol{\varepsilon}^t$ is a white noise to ensure the Markov property and $X^0 \sim \mathcal{N}(0, \mathbf{\Sigma}^0)$.

Consequence: a Gaussian Graphical Model

- Each $X^t | X^{t-1} \sim \mathcal{N}(\theta X^{t-1}, \mathbf{\Sigma})$,
- or, equivalently, $X_j^t | X^{t-1} \sim \mathcal{N}(\mathbf{\Theta}_j X^{t-1}, \mathbf{\Sigma})$

where $\mathbf{\Sigma}$ is known and $\mathbf{\Theta}_j$ is the jth row of $\mathbf{\Theta}$.

Gaussian Graphical Model and AR process (2)

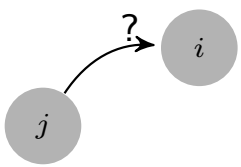
Interpretation as a GGM

The VAR(1) as a covariance selection model

$$\theta_{ij} = \frac{\text{cov}\left(X_i^t, X_j^{t-1} | X_{\mathcal{P}\setminus j}^{t-1}\right)}{\text{var}\left(X_j^{t-1} | X_{\mathcal{P}\setminus j}^{t-1}\right)},$$

Graphical Interpretation

$\rightsquigarrow$ The matrix $\mathbf{\Theta} = (\theta_{ij})_{i,j\in\mathcal{P}}$ encodes the network $\mathcal{G}$ we are looking for.



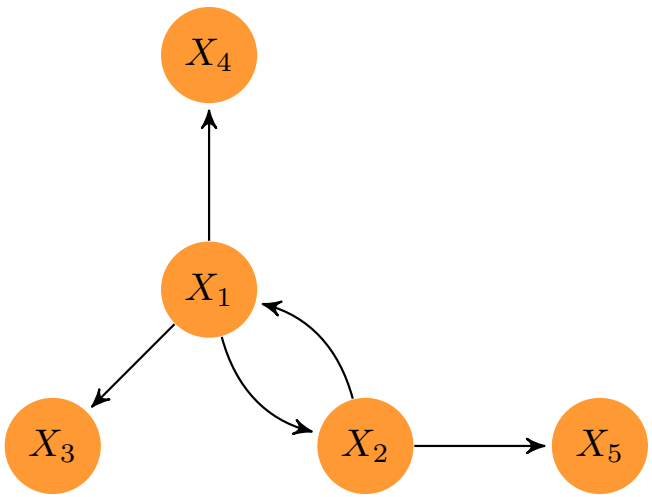
if and only if

conditional dependency between X_j^{t-1} and X_i^t
or
non-null partial correlation between X_j^{t-1} and X_i^t
$\Updownarrow$
$\theta_{ij} \neq 0$

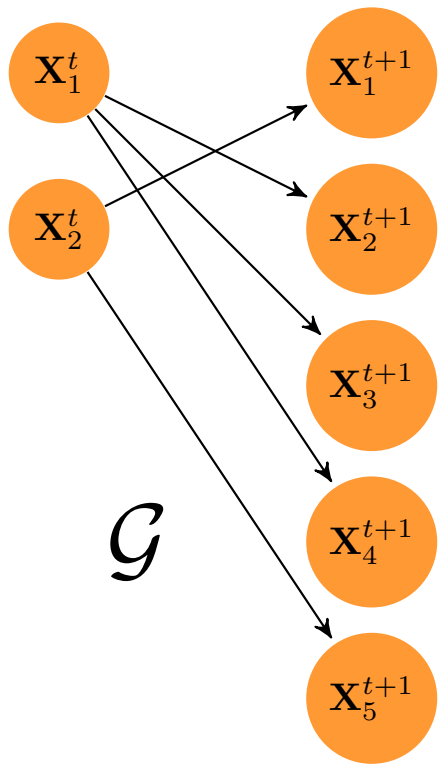
Gaussian Graphical Model and AR process (3)

Graphical interpretation

1. Follow-up of one single experiment/individual;
2. Close enough time-points to ensure
   - dependency between consecutive measurements;
   - homogeneity of the Markov process.



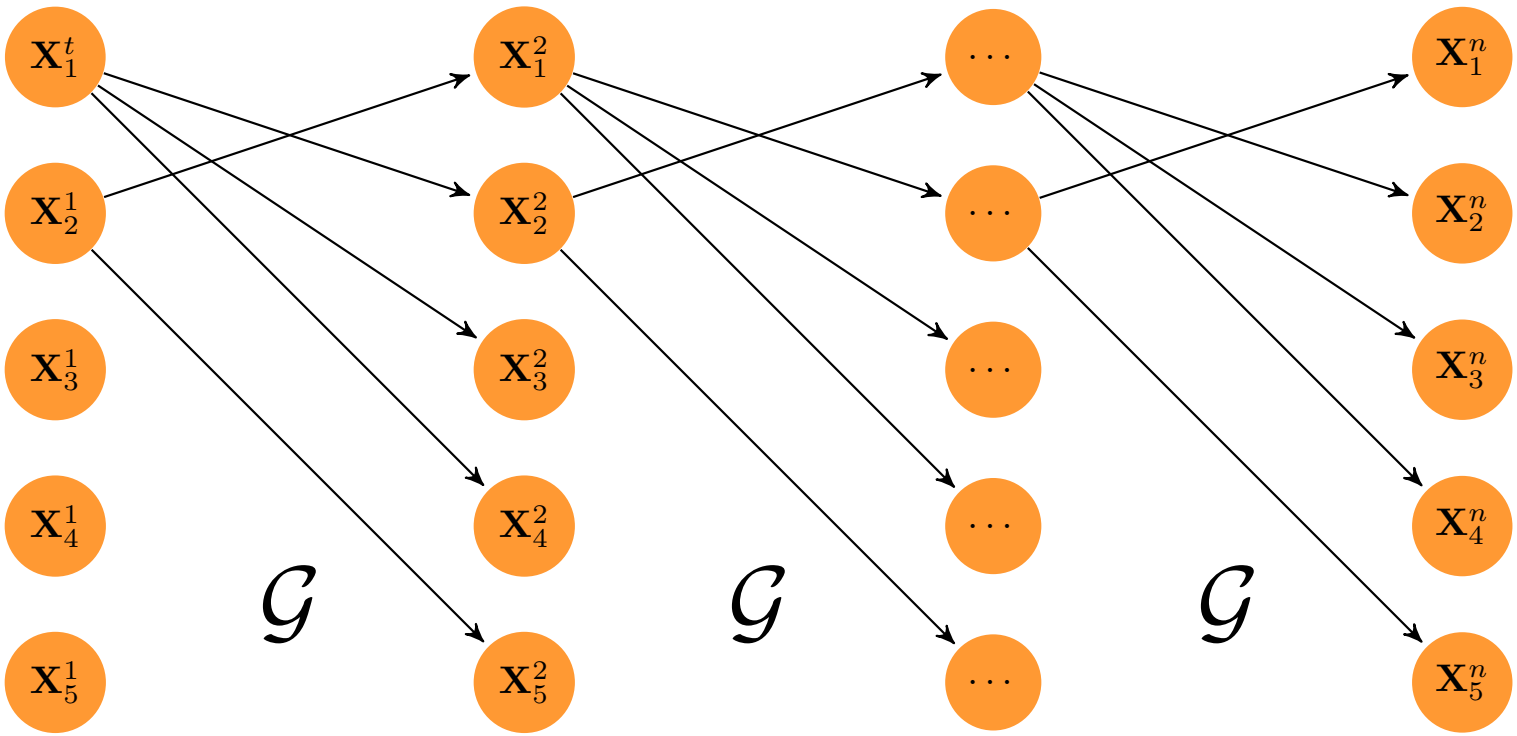
stands for



Gaussian Graphical Model and AR process (3)

Graphical interpretation

1. Follow-up of one single experiment/individual;
2. Close enough time-points to ensure
   - dependency between consecutive measurements;
   - homogeneity of the Markov process.



Outline

Some families of methods for network reconstruction

Test-based methods

- Tests the nullity of each entries
- Combinatorial problem when $p > 30$ ...

Sparsity-inducing regularization methods

- induce sparsity with the ℓ_1-norm penalization
- Use results from convex optimization
- Versatile and computationally efficient

Bayesian methods

- Compute the posterior probability of each edge
- Usually more computationally demanding
- For special graphs, computation gets easier

Inference: maximum likelihood estimator

The natural approach for parametric statistics

Let X be a random vector with distribution defined by $f_X(x; \mathbf{\Theta})$, where $\mathbf{\Theta}$ are the model parameters.

Maximum likelihood estimator

$$\hat{\mathbf{\Theta}} = \arg\max_{\mathbf{\Theta}} \ell(\mathbf{\Theta}; \mathbf{X})$$

where ℓ is the log likelihood, a function of the parameters:

$$\ell(\mathbf{\Theta}; \mathbf{X}) = \log \prod_{i=1}^{n} f_X(\mathbf{x}_i; \mathbf{\Theta}),$$

where $\mathbf{x}_i$ is the ith row of $\mathbf{X}$.

Remarks

- This a convex optimization problem,
- We just need to detect non zero coefficients in $\mathbf{\Theta}$

The multivariate Gaussian log-likelihood

Let $\mathbf{S} = n^{-1}\mathbf{X}^\intercal\mathbf{X}$ be the empirical variance-covariance matrix: $\mathbf{S}$ is a sufficient statistic of $\mathbf{\Theta}$.

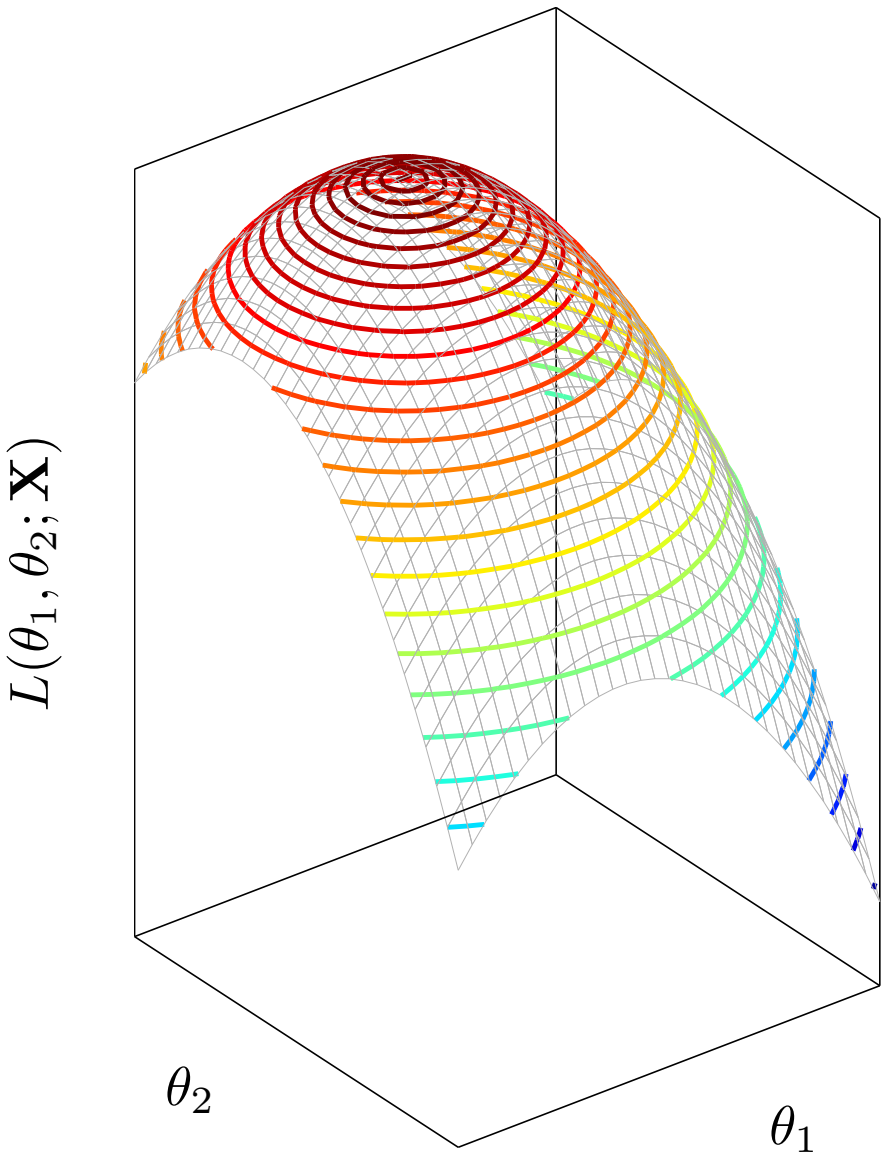
The log-likelihood

$$\ell(\mathbf{\Theta}; \mathbf{S}) = \frac{n}{2}\log\det(\mathbf{\Theta}) - \frac{n}{2}\mathrm{Trace}(\mathbf{S}\mathbf{\Theta}) + \frac{n}{2}\log(2\pi).$$

- $\rightsquigarrow$ The MLE $= \mathbf{S}^{-1}$ of $\mathbf{\Theta}$ is not defined for $n < p$ and never sparse.
- $\rightsquigarrow$ The need for regularization is huge.

A Geometric View of Shrinkage

Constrained Optimization



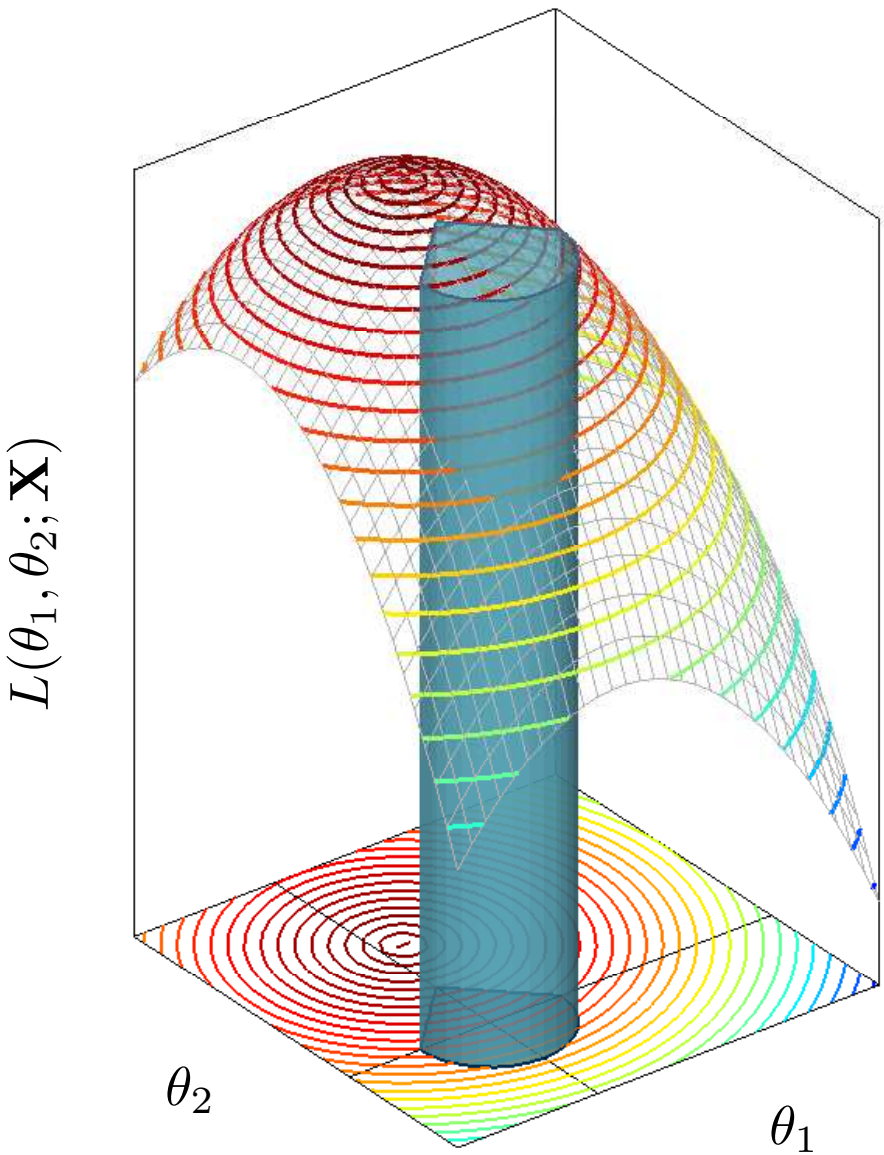
We basically want to solve a problem of the form

$$\underset{\theta_1,\theta_2}{\text{maximize}}\, \ell(\theta_1, \theta_2; \mathbf{X})$$

where ℓ is typically a concave likelihood function.

A Geometric View of Shrinkage

Constrained Optimization



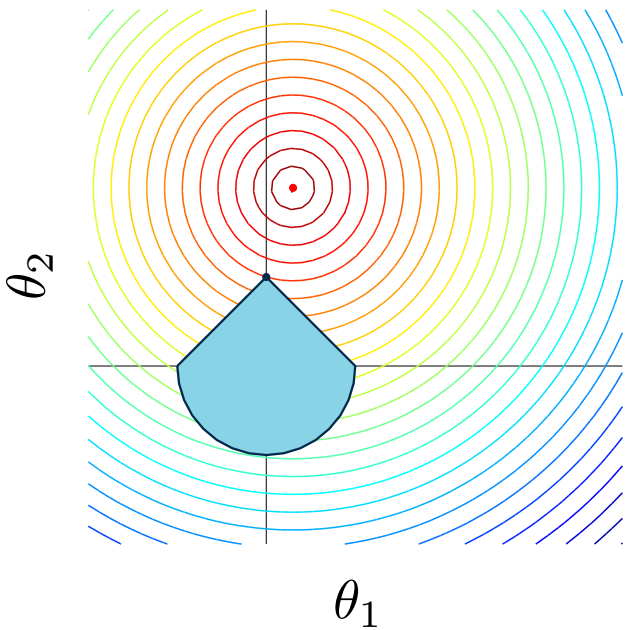
$$\begin{cases} \underset{\theta_1,\theta_2}{\text{maximize}} & \ell(\theta_1, \theta_2; \mathbf{X}) \\ \text{s.t.} & \Omega(\theta_1, \theta_2) \leq c \end{cases},$$

where Ω defines a domain that *constrains* $\boldsymbol{\beta}$.

How shall we define Ω ?

A Geometric View of Shrinkage

Constrained Optimization



$$\begin{cases} \underset{\theta_1,\theta_2}{\text{maximize}} & \ell(\theta_1, \theta_2; \mathbf{X}) \\ \text{s.t.} & \Omega(\theta_1, \theta_2) \leq c \end{cases},$$

where Ω defines a domain that *constrains* $\boldsymbol{\beta}$.

How shall we define Ω ?

The Lasso

Least Absolute Shrinkage and Selection Operator

Idea

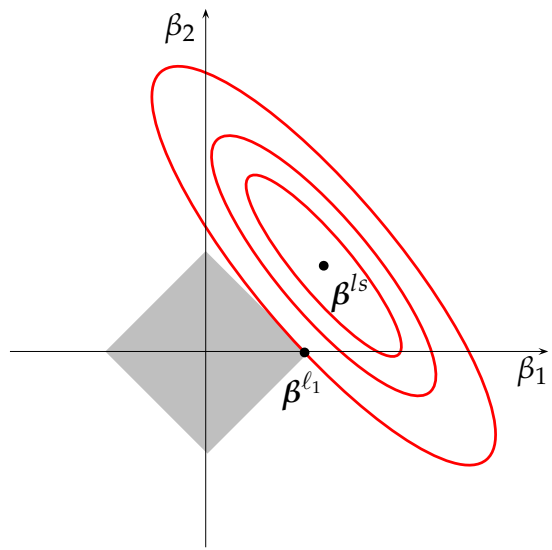
Suggest an admissible set that induces sparsity (force several entries to exactly zero in $\hat{\boldsymbol{\beta}}$).

Lasso as a regularization problem

The Lasso estimate of $\boldsymbol{\beta}$ is the solution to

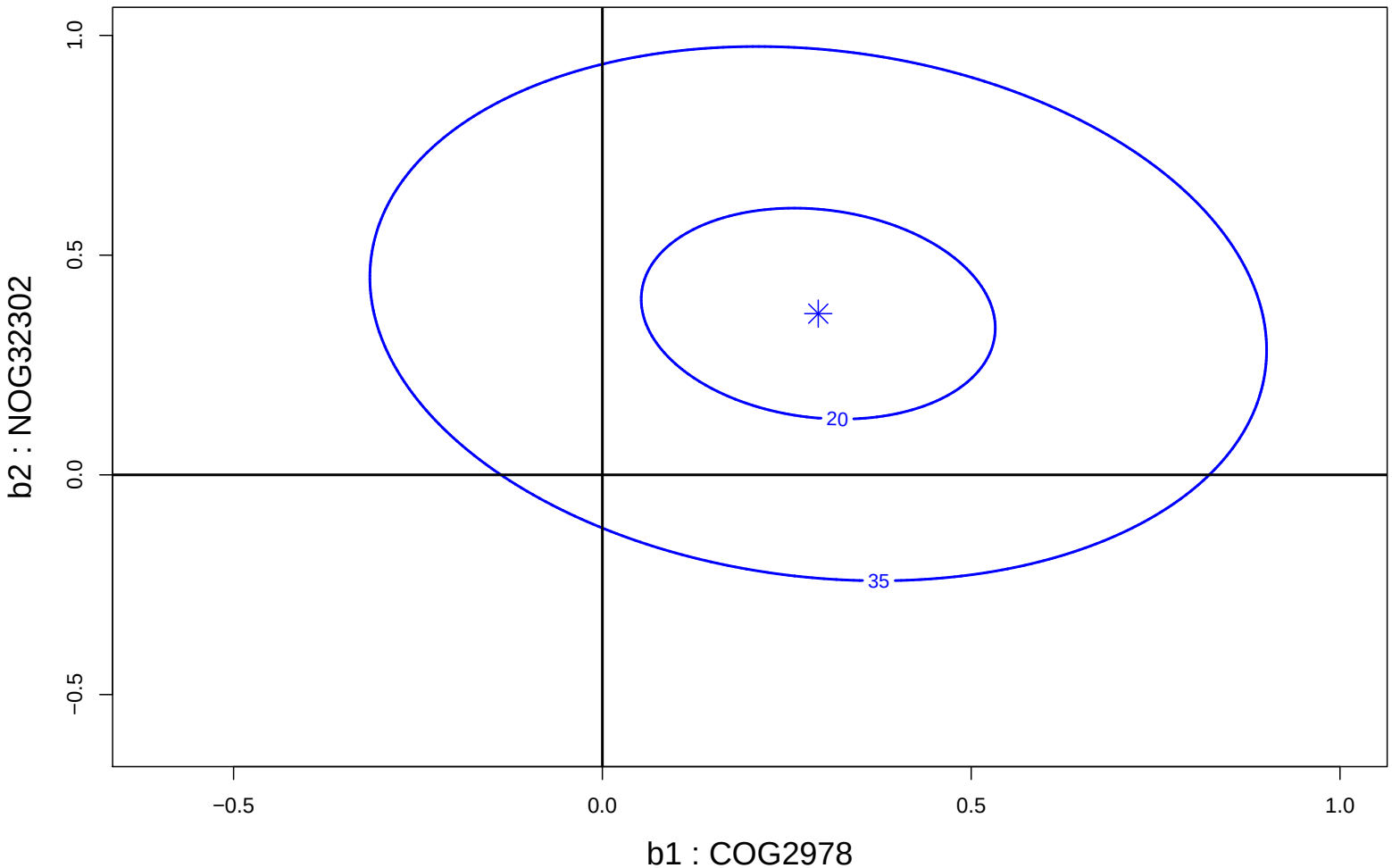
$$\hat{\boldsymbol{\theta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\theta}} -\ell(\boldsymbol{\theta}), \quad \text{s.t.} \sum_{j=1}^{p} |\theta_j| \leq s,$$

where s is a shrinkage factor.



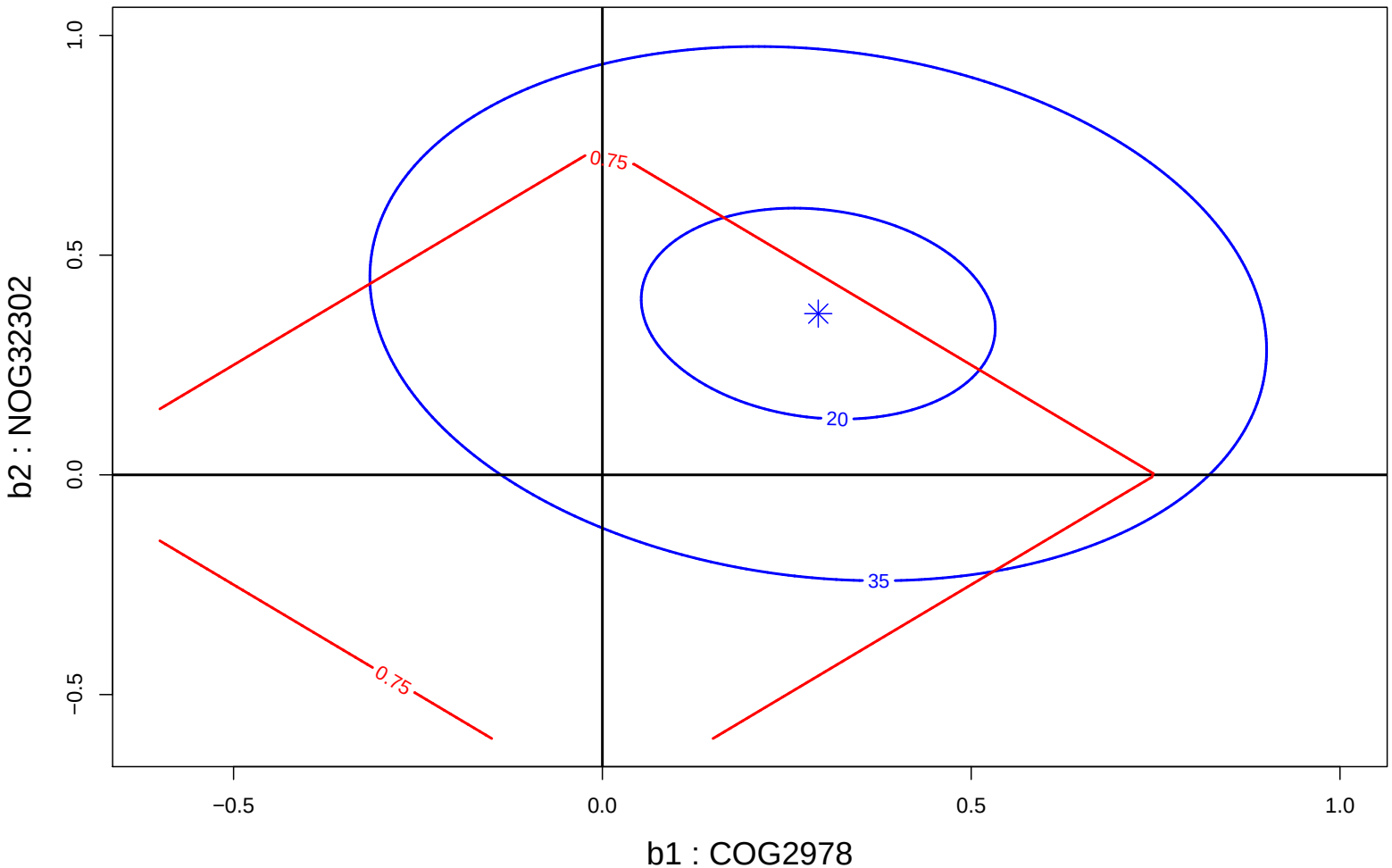
Insights: 2-dimensional example with the square loss

$$\sum_{i=1}^{n}(y_i - x_i^1\theta_1 - x_i^2\theta_2)^2, \qquad \text{no constraints}$$



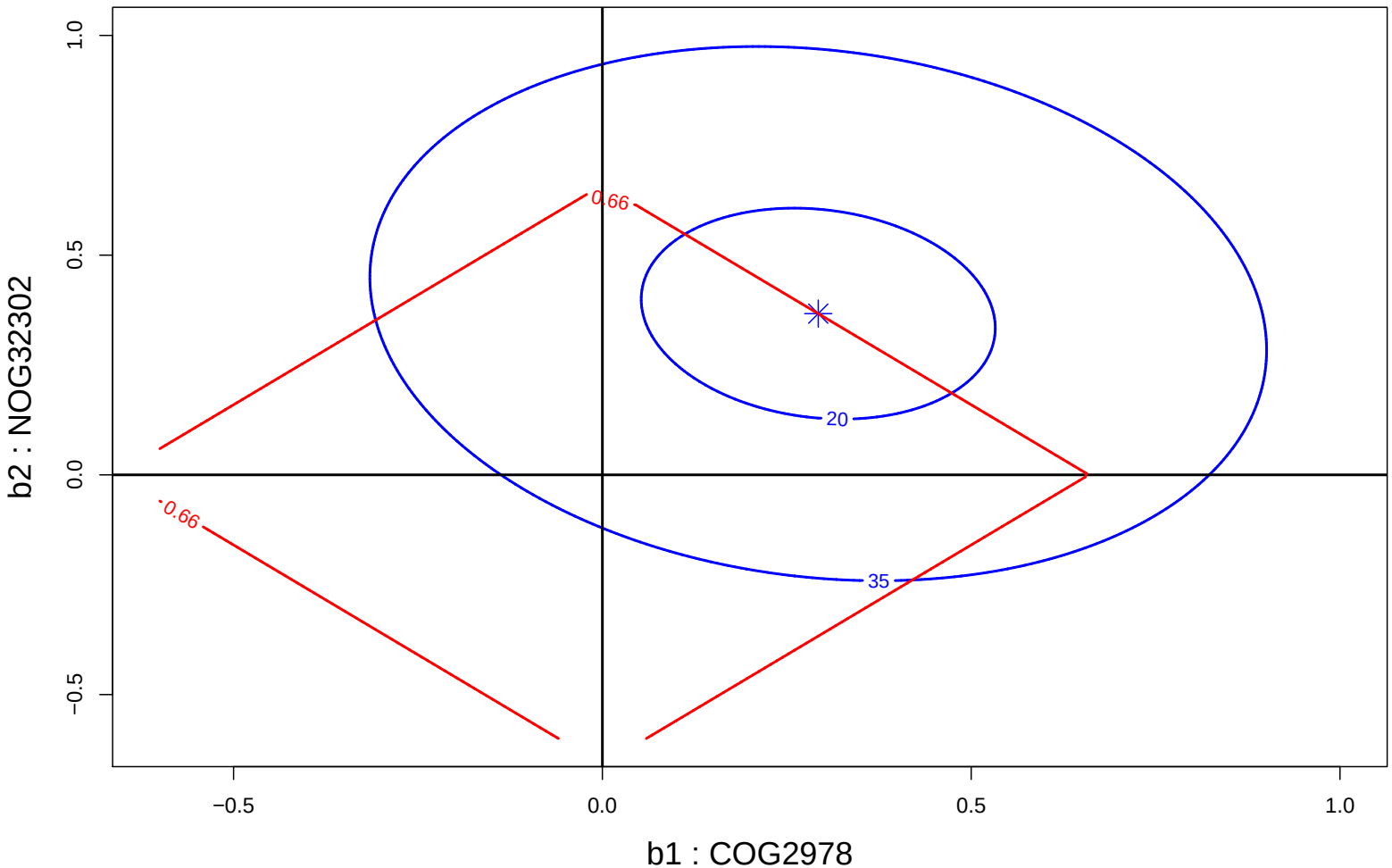
Insights: 2-dimensional example with the square loss

$$\sum_{i=1}^{n}(y_i - x_i^1\theta_1 - x_i^2\theta_2)^2, \qquad \text{s.t. } |\theta_1| + |\theta_2| < 0.75$$



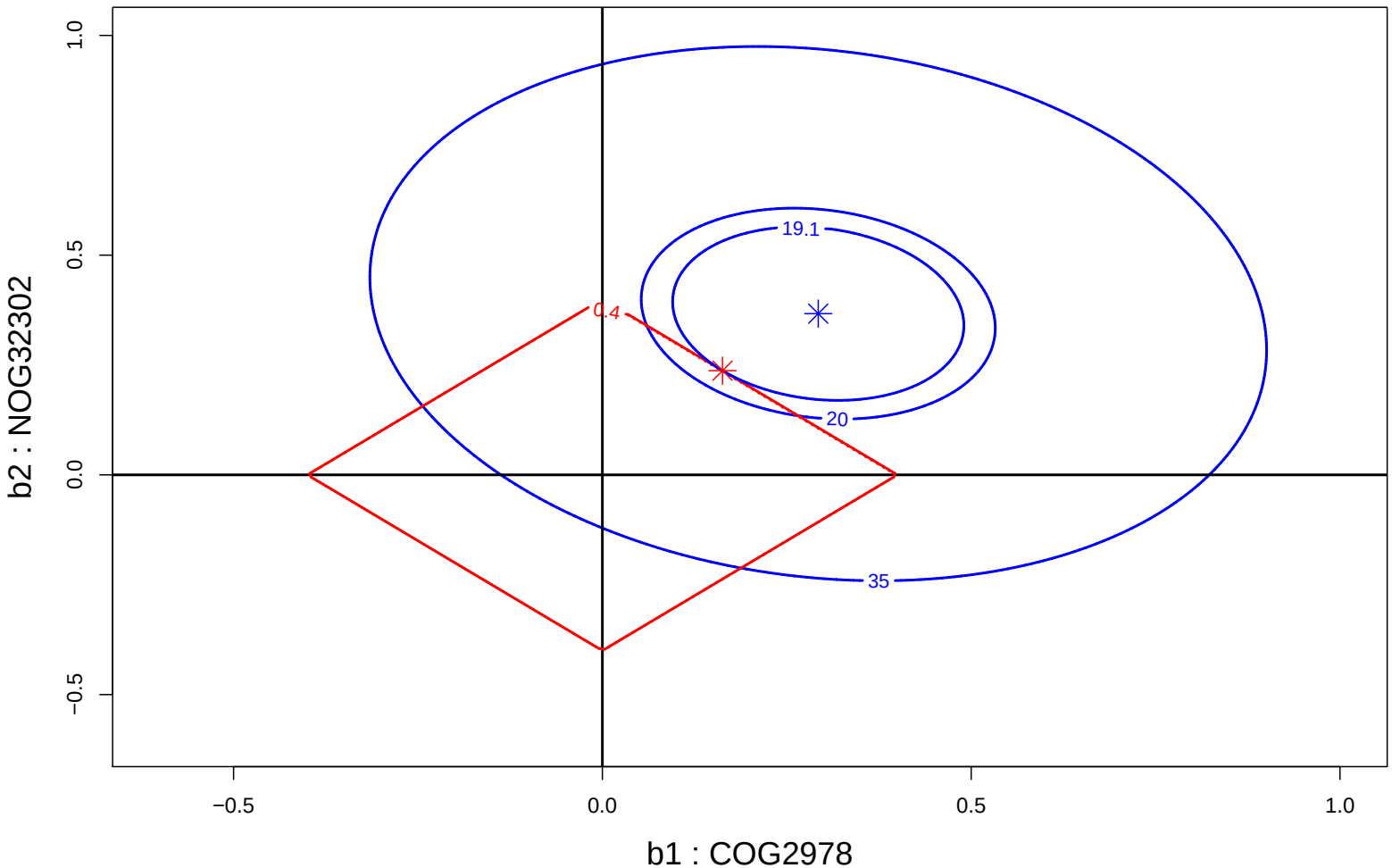
Insights: 2-dimensional example with the square loss

$$\sum_{i=1}^{n}(y_i - x_i^1\theta_1 - x_i^2\theta_2)^2, \qquad \text{s.t. } |\theta_1| + |\theta_2| < 0.66$$



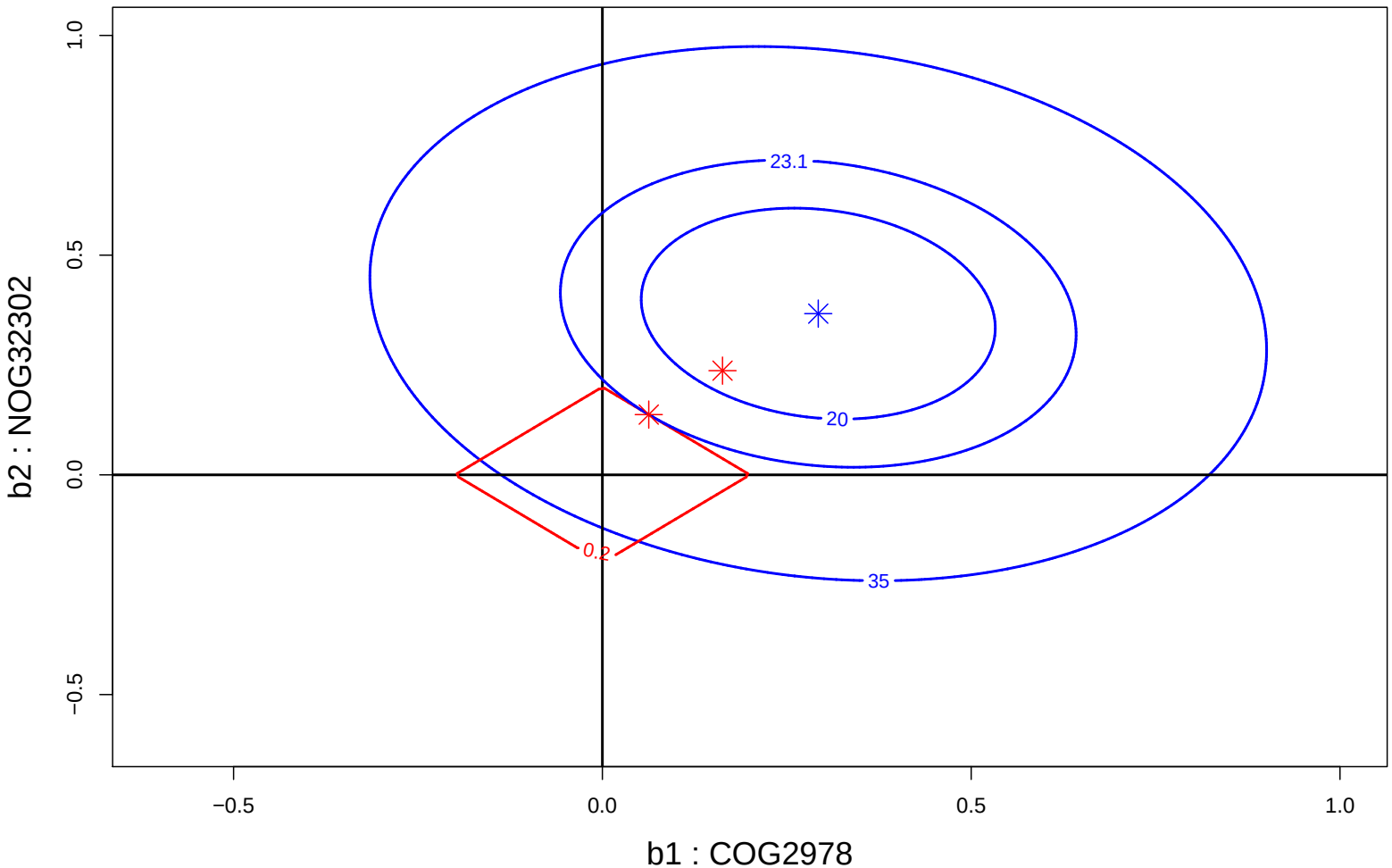
Insights: 2-dimensional example with the square loss

$$\sum_{i=1}^{n}(y_i - x_i^1\theta_1 - x_i^2\theta_2)^2, \qquad \text{s.t. } |\theta_1| + |\theta_2| < 0.4$$



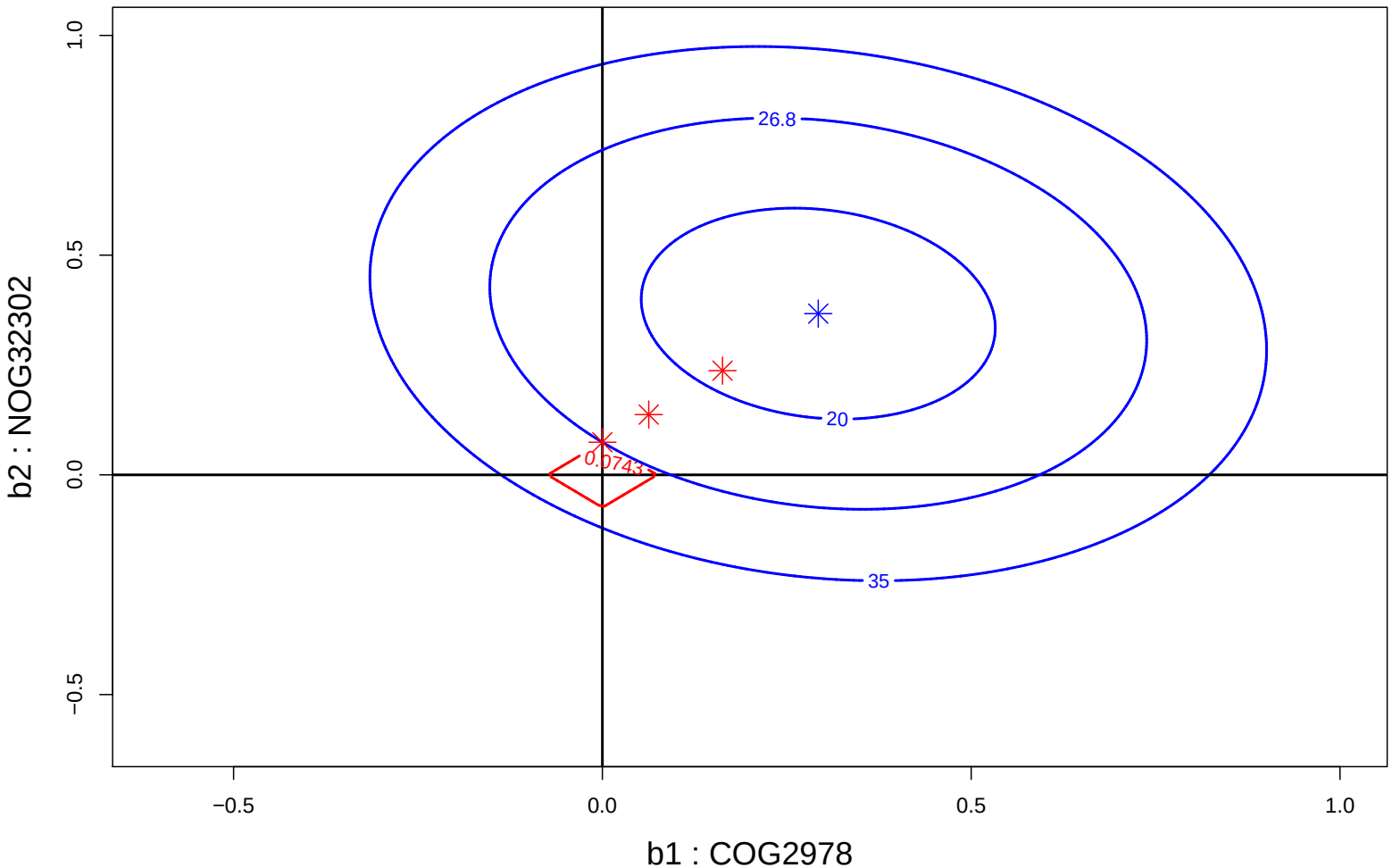
Insights: 2-dimensional example with the square loss

$$\sum_{i=1}^{n}(y_i - x_i^1\theta_1 - x_i^2\theta_2)^2, \qquad \text{s.t. } |\theta_1| + |\theta_2| < 0.2$$



Insights: 2-dimensional example with the square loss

$$\sum_{i=1}^{n}(y_i - x_i^1\theta_1 - x_i^2\theta_2)^2, \qquad \text{s.t. } |\theta_1| + |\theta_2| < 0.0743$$



Application to GGM: the "Graphical-Lasso"

A penalized likelihood approach

$$\hat{\mathbf{\Theta}}_\lambda = \arg\max_{\mathbf{\Theta} \in \mathbb{S}_+} \ell(\mathbf{\Theta}; \mathbf{X}) - \lambda \|\mathbf{\Theta}\|_{\ell_1}$$

where

- ℓ is the model log-likelihood,
- $\|\cdot\|_{\ell_1}$ is a penalty function tuned by $\lambda > 0$.
  1. *regularization* (needed when $n \ll p$),
  2. *selection* (sparsity induced by the ℓ_1-norm),
- solved in `R`-packages **glasso**, **quic**, **huge** ($\mathcal{O}(p^3)$)

Application to GGM: "Neighborhood selection"

A close cousin, thank to the relationship between Gaussian vector and linear regression

Remember that

$$X_i | X_{\setminus i} = \sum_{j \in \text{neighbors}(i)} \beta_j X_j + \varepsilon_i \quad \text{with } \beta_j = -\frac{\Theta_{ij}}{\Theta_{ii}}.$$

A penalized least-square approach

Let $\mathbf{X}_i$ be the ith column of the data matrix (i.e data associated to variable (gene) i), and $\mathbf{X}_{\setminus i}$ deprived of colmun i. We select the neighbors of variable i by solving

$$\widehat{\boldsymbol{\beta}}^{(i)} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{p-1}} \frac{1}{n} \left\| \mathbf{X}_i - \mathbf{X}_{\setminus i} \boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1$$

- not symmetric, not positive-definite
- \+ p Lasso solved with Lars-like algorithms ($\mathcal{O}(npd)$ for d neighbors).

Practical implications of theoretical results

Selection consistency (Ravikumar, Wainwright, 2009-2012)

Denote $d = \max_{j \in \mathcal{P}}(\text{degree}_\text{j})$. Consistency for an appropriate λ and

- $n \approx \mathcal{O}(d^2 \log(p))$ for the graphical Lasso and Clime.
- $n \approx \mathcal{O}(d \log(p))$ for neighborhood selection (sharp).

(Irrepresentability) conditions are not strictly comparable...

Ultra high-dimension phenomenon (Verzelen, 2011)

Minimax risk for sparse regression with d-sparse models: useless when

$$\frac{d \log(p/d)}{n} \geq 1/2, \qquad (\text{e.g.}, n = 50, p = 200, d \geq 8).$$

Good news! when n is small, we don't need to solve huge problems because they can't but fail.

Practical implications of theoretical results

Selection consistency (Ravikumar, Wainwright, 2009-2012)

Denote $d = \max_{j \in \mathcal{P}}(\text{degree}_\text{j})$. Consistency for an appropriate λ and

- $n \approx \mathcal{O}(d^2 \log(p))$ for the graphical Lasso and Clime.
- $n \approx \mathcal{O}(d \log(p))$ for neighborhood selection (sharp).

(Irrepresentability) conditions are not strictly comparable...

Ultra high-dimension phenomenon (Verzelen, 2011)

Minimax risk for sparse regression with d-sparse models: useless when

$$\frac{d \log(p/d)}{n} \geq 1/2, \qquad (\text{e.g.}, n = 50, p = 200, d \geq 8).$$

Good news! when n is small, we don't need to solve huge problems because they can't but fail.

Model selection

Cross-validation

Optimal in terms of prediction, not in terms of selection

Information based criteria

- GGMSelect (Girault *et al*, '12) selects among a family of candidates.
- Adapt IC to sparse high dimensional problems, e.g.

$$\text{EBIC}_\gamma(\widehat{\mathbf{\Theta}}_\lambda) = -2\text{loglik}(\widehat{\mathbf{\Theta}}_\lambda; \mathbf{X}) + |\mathcal{E}_\lambda|(\log(n) + 4\gamma \log(p)),$$

Resampling/subsampling

Keep edges frequently selected on an range of λ after sub-samplings

- Stability Selection (Meinshausen and Bühlman, 2010, Bach 2008)
- Stability approach to Regularization Selection (StaRS) (Liu, 2010).

Concluding remark about GGM

Sparse GGM

+ very solid statistical and computational framework
+ competitive to other inference methods (DREAM 5 benchmark, 2012)
− performances remain questionable on real data, as for other methods

$\rightsquigarrow$ Network inference is a very difficult problem
$\rightsquigarrow$ Some biological questions can be answered without network inference

Extensions motivated by biological data

Strengthen the inference by

- accounting for biological features
  1. structure of the network (organization of biological mechanisms)
  2. sample heterogeneity (structure of the population)
  3. horizontal integration (use multiple data and platforms)
  4. Deal with covariates
- accounting for data features
  1. What if some important actor is missing?
  2. Extend to non strictly normal distribution
  3. Deal with a large number of actors

$\rightsquigarrow$ How? Essentially by crafting the regularization according to our prior knowledge

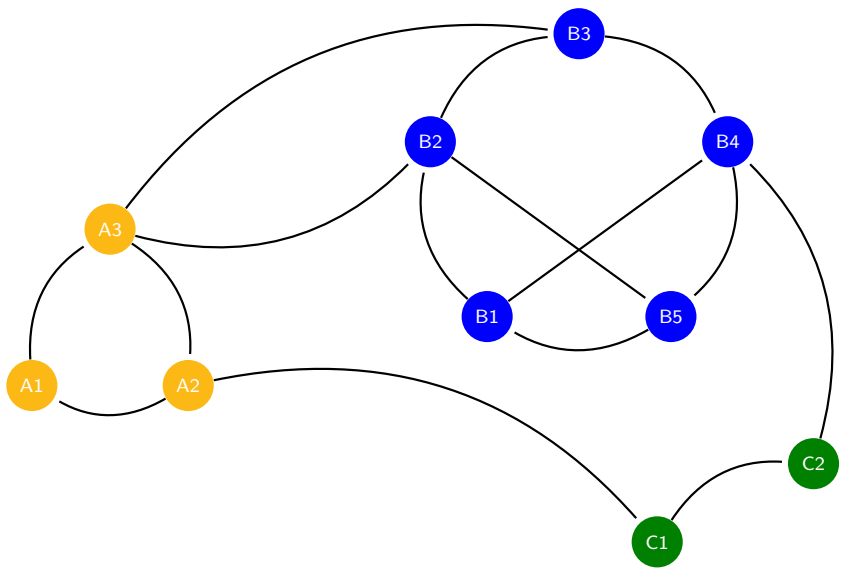
Outline

Handling with the data structure and scarcity

By introducing some prior

Priors should be biologically grounded

1. no too many genes effectively interact: sparsity,
2. networks are organized: latent clustering.



Structured regularization

SIMoNe: Statistical Inference for MOdular NEtworks

$$\arg\max_{\mathbf{\Theta},\mathbf{Z}} \ell(\mathbf{\Theta};\mathbf{Y}) - \lambda \, \|\mathbf{P}_{\mathbf{Z}} \star \mathbf{\Theta}\|_{\ell_1},$$

where $\mathbf{P}_{\mathbf{Z}}$ is a matrix of weights depending on a underlying latent structure $\mathbf{Z}$ (depicted through a stochastic block model).

$\rightsquigarrow$ Cluster-driven inference via an EM-like strategy.


Ambroise, Chiquet, Matias. Inferring sparse GGM with latent structure, EJS, 2009.


Marlin, Schmidt, Murphy: similar Bayesian work UCI 2010.


Wong et al., close update: *Adaptive Graphical Lasso*, 2014.


Chiquet et al., SIMoNe R-package *(needs updates. . .)*, Note Bioinformatics, 2009.

How to come up with a latent clustering?

Biological expertise

- Build $\mathbf{Z}$ from prior biological information
  - transcription factors vs. regulatees,
  - number of potential binding sites,
  - KEGG pathways, ...
- Build the weight matrix from $\mathbf{Z}$.

Inference: Stochastic Bloc Model

- Spread the nodes into $\mathcal{Q}$ classes;
- Connexion probabilities depend upon node classes:

$$\mathbb{P}(i \leftrightarrow j | i \in \text{class } q, j \in \text{class } \ell) = \pi_{q\ell}.$$

- Build $P_{\mathbf{Z}} \propto 1 - \pi_{q\ell}$.

Illustration on breast Cancer

Prediction of the outcome of preoperative chemotherapy

Hess *et al.*
Journal. of Clinical Oncology, 2006.

Data set

133 patients classified as

1. pathologic complete response,
2. residual disease,

according to a signature of 26 genes (small network).

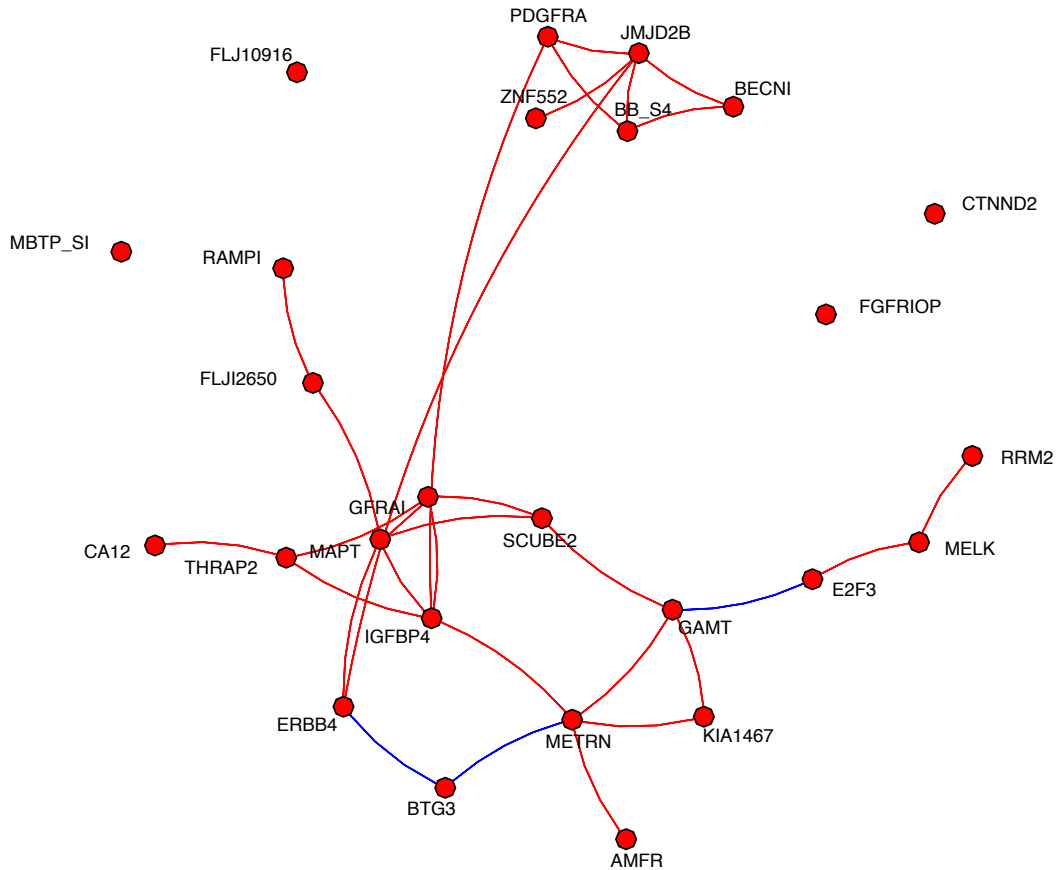


Figure: Pooling the data, Neighborhood Selection

Illustration on breast Cancer

Prediction of the outcome of preoperative chemotherapy

Hess *et al.*
Journal. of Clinical Oncology, 2006.

Data set

133 patients classified as

1. pathologic complete response,
2. residual disease,

according to a signature of 26 genes (small network).

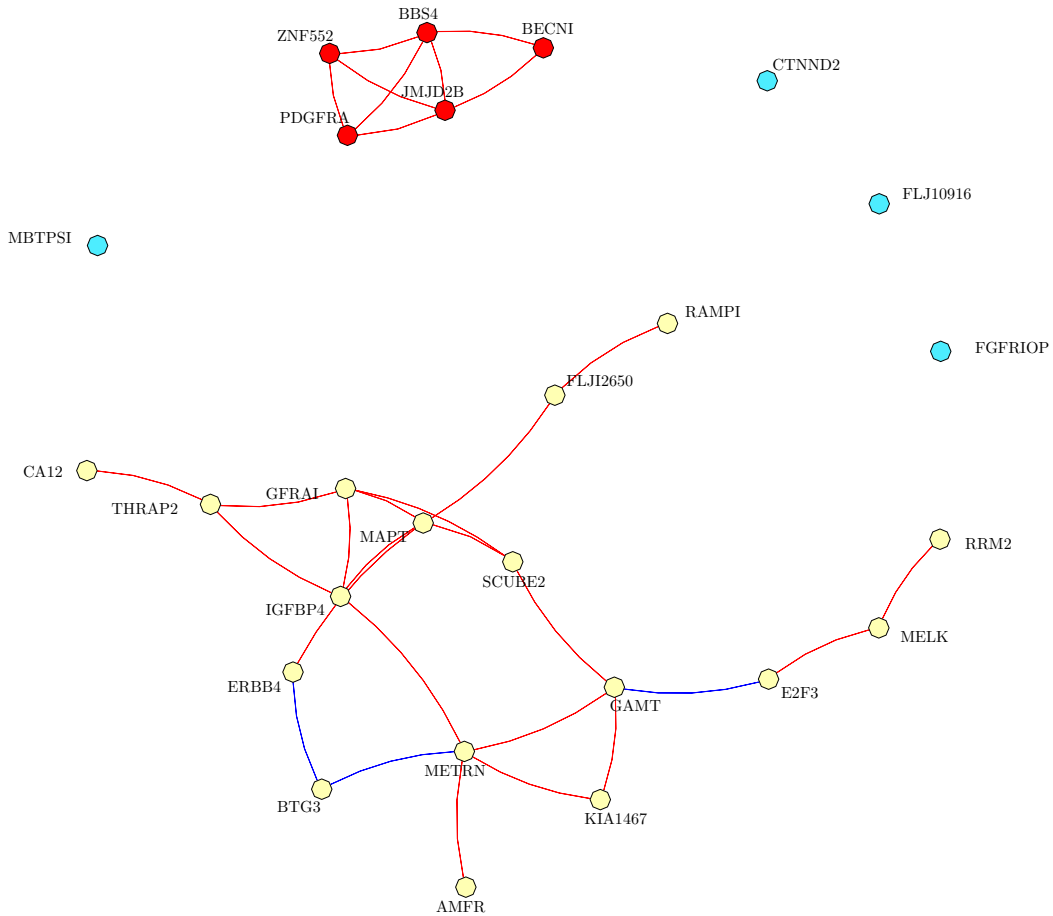


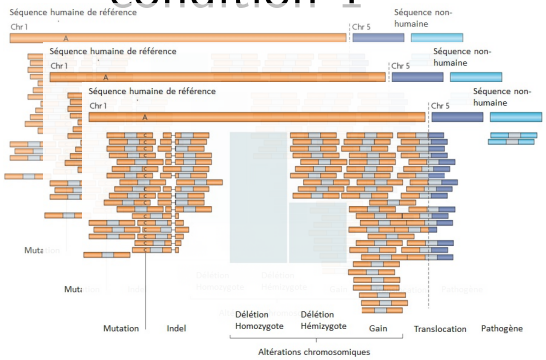
Figure: Pooling the data, SIMoNE with clustering

Outline

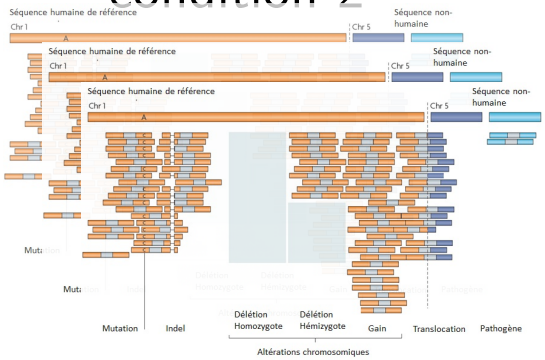
Handling scarcity and heterogeneity of data

Merge several experimental conditions

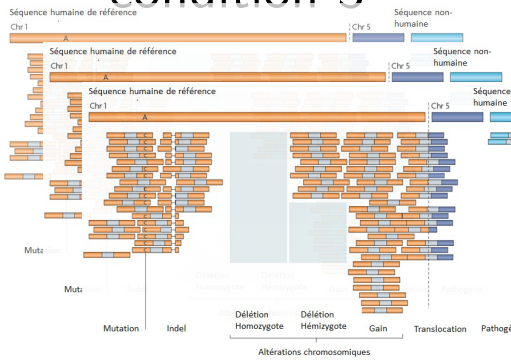
condition 1



condition 2



condition 3



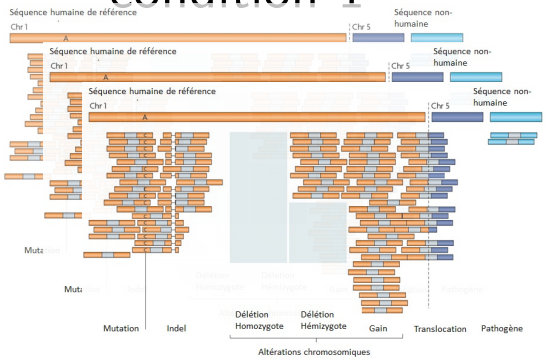
Multiple inference of GGM

$$\arg\max_{\mathbf{\Theta}^{(c)}, c=1\ldots,C} \sum_{c=1}^{C} \ell(\mathbf{\Theta}^{(c)}; \mathbf{S}^{(c)}) - \lambda \, \mathrm{pen}_{\ell_1}(\mathbf{\Theta}^{(c)}).$$

Handling scarcity and heterogeneity of data

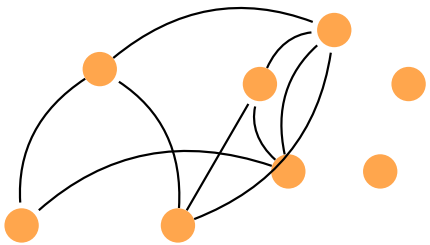
Inferring each graph independently does not help

condition 1

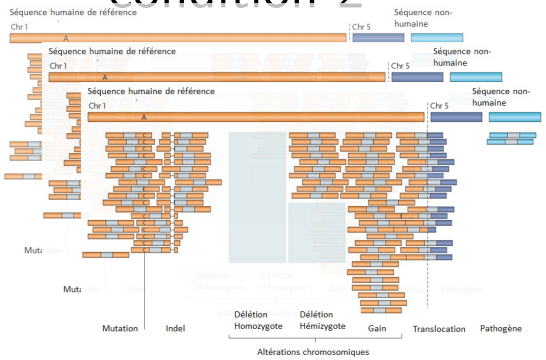


$(Y_1^{(1)}, \ldots, Y_{n_1}^{(1)})$

inference

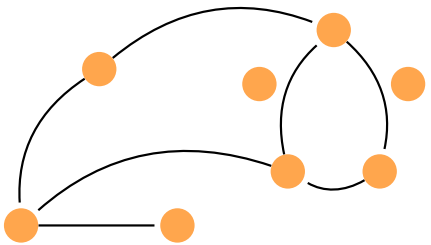


condition 2

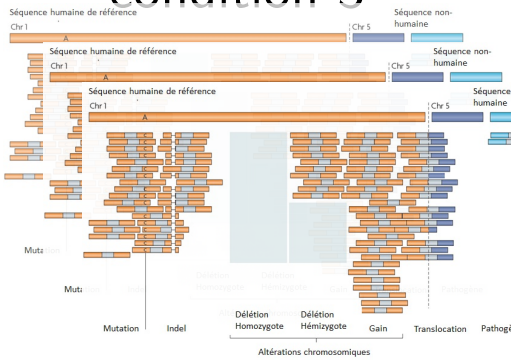


$(Y_1^{(2)}, \ldots, Y_{n_2}^{(2)})$

inference

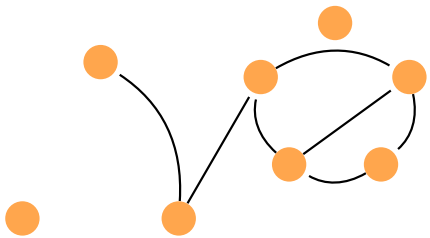


condition 3



$(Y_1^{(3)}, \ldots, Y_{n_3}^{(3)})$

inference

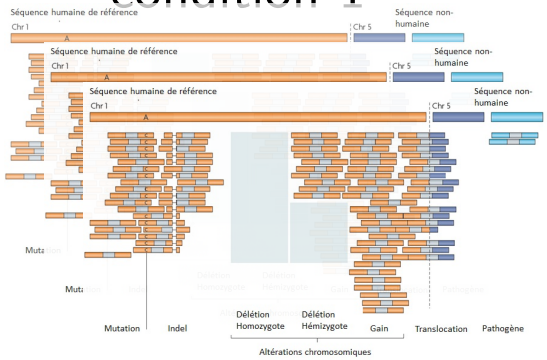


Multiple inference of GGM

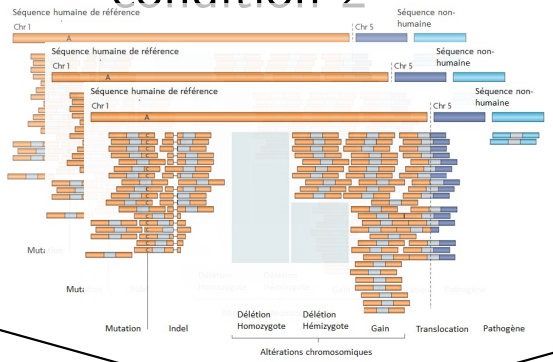
Handling scarcity and heterogeneity of data

By pooling all the available data (like we just have with Hess' data set)

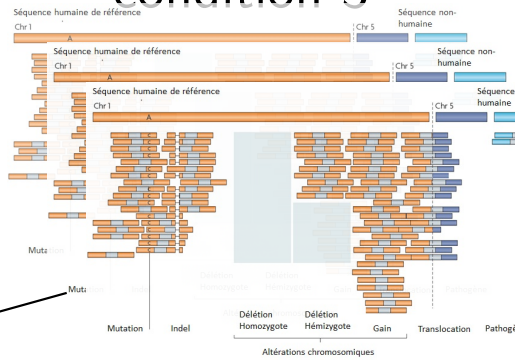
condition 1



condition 2

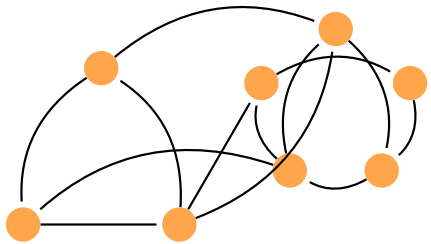


condition 3



$$(Y_1, \ldots, Y_n),\ n = n_1 + n_2 + n_3.$$

inference

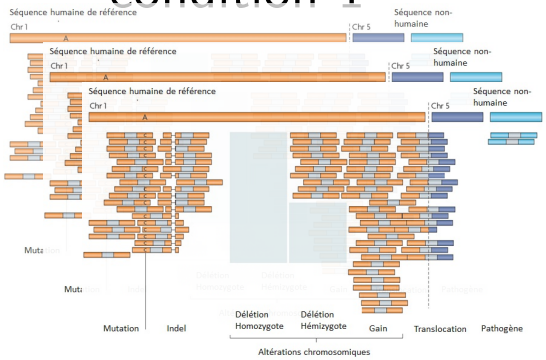


Multiple inference of GGM

Handling scarcity and heterogeneity of data

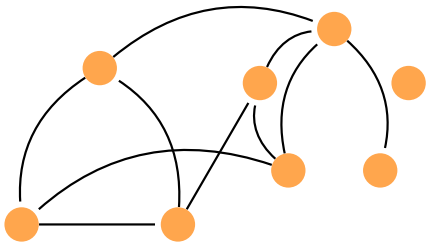
By breaking the separability

condition 1

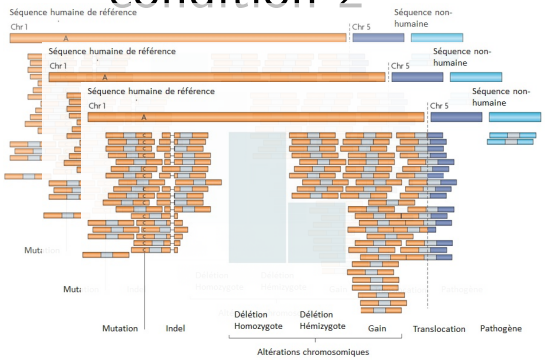


$(Y_1^{(1)}, \ldots, Y_{n_1}^{(1)})$

inference

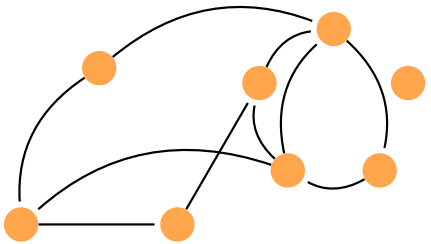


condition 2

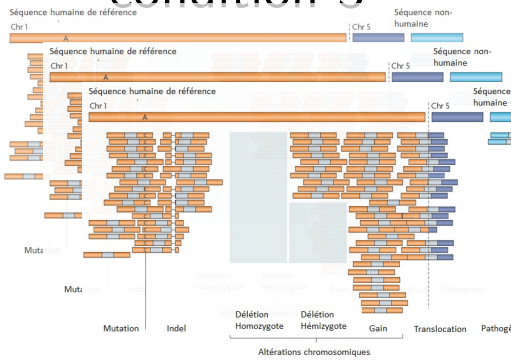


$(Y_1^{(2)}, \ldots, Y_{n_2}^{(2)})$

inference

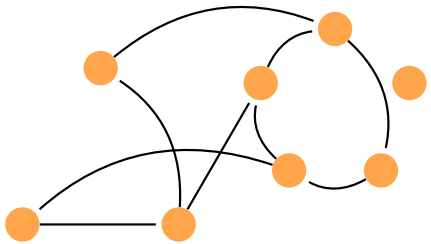


condition 3



$(Y_1^{(3)}, \ldots, Y_{n_3}^{(3)})$

inference

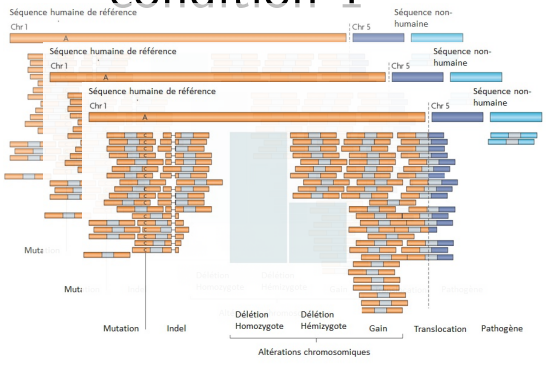


Multiple inference of GGM

Handling scarcity and heterogeneity of data

By breaking the separability

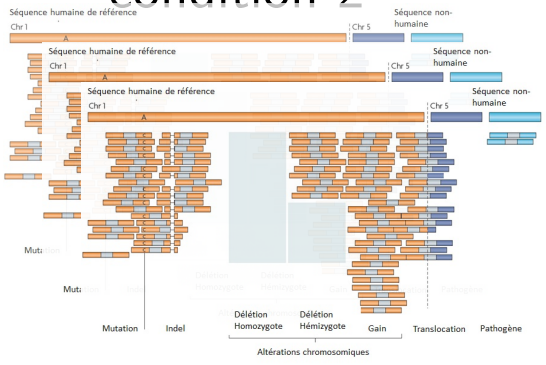
condition 1



$(Y_1^{(1)}, \ldots, Y_{n_1}^{(1)})$

inference

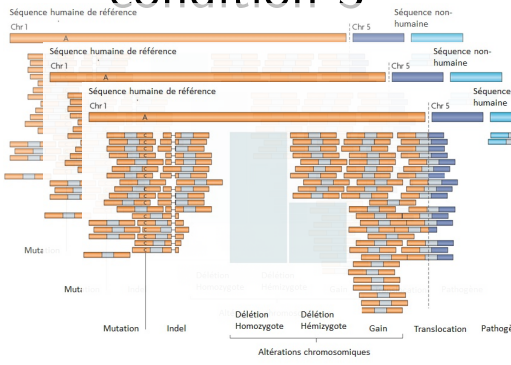
condition 2



$(Y_1^{(2)}, \ldots, Y_{n_2}^{(2)})$

inference

condition 3



$(Y_1^{(3)}, \ldots, Y_{n_3}^{(3)})$

inference

Multiple inference of GGM

$$\underset{\mathbf{\Theta}^{(c)}, c=1\ldots,C}{\arg\max} \sum_{c=1}^{C} \ell(\mathbf{\Theta}^{(c)}; \mathbf{S}^{(c)}) - \lambda \, \mathrm{pen}_{\ell_1}(\mathbf{\Theta}^{(c)}).$$

A multitask approach

Chiquet, Grandvalet, Ambroise, Statistics and Computing 2010/11

Break the separability

Joint the optimization problem by either modifying

$$\underset{\mathbf{\Theta}^{(c)}, c=1\ldots,C}{\arg\max} \sum_{c=1}^{C} \tilde{\ell}(\mathbf{\Theta}^{(c)}; \tilde{\mathbf{S}}^{(c)}) - \lambda \ \mathrm{pen}_{\ell_1}(\mathbf{\Theta}^{(c)}).$$

1. the fitting term
2. the regularization term

A multitask approach

Chiquet, Grandvalet, Ambroise, Statistics and Computing 2010/11

Break the separability

Joint the optimization problem by either modifying

$$\underset{\mathbf{\Theta}^{(c)}, c=1\ldots,C}{\arg\max} \sum_{c=1}^{C} \tilde{\ell}(\mathbf{\Theta}^{(c)}; \tilde{\mathbf{S}}^{(c)}) - \lambda \, \mathrm{pen}_{\ell_1}(\mathbf{\Theta}^{(c)}).$$

1. the fitting term
2. the regularization term

Intertwined-Lasso

- $\overline{\mathbf{S}} = \frac{1}{n} \sum_{t=1}^{T} n_t \mathbf{S}^{(t)}$ is the "pooled-tasks" covariance matrix.
- $\widetilde{\mathbf{S}}^{(t)} = \alpha \mathbf{S}^{(t)} + (1-\alpha)\overline{\mathbf{S}}$ is a mixture between specific and pooled covariance matrices.

A multitask approach

Chiquet, Grandvalet, Ambroise, Statistics and Computing 2010/11

Break the separability

Joint the optimization problem by either modifying

$$\underset{\mathbf{\Theta}^{(c)}, c=1\ldots,C}{\arg\max} \sum_{c=1}^{C} \tilde{\ell}(\mathbf{\Theta}^{(c)}; \tilde{\mathbf{S}}^{(c)}) - \lambda \, \mathrm{pen}_{\ell_1}(\mathbf{\Theta}^{(c)}).$$

1. the fitting term
2. the regularization term

Sparsity with grouping effect

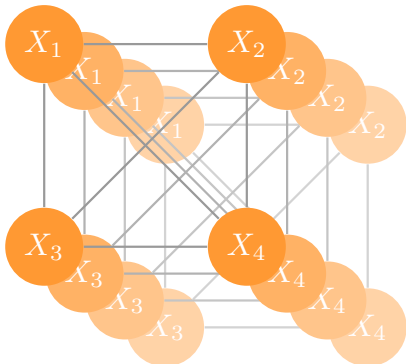
- Group-Lasso (Yuan and Lin 2006, Grandvalet and Canu, 1998),
- Cooperative-Lasso (Chiquet et al, AoAS, 2012),

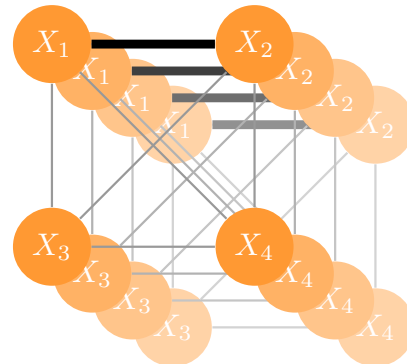
Grouping effects induced

Potential groups

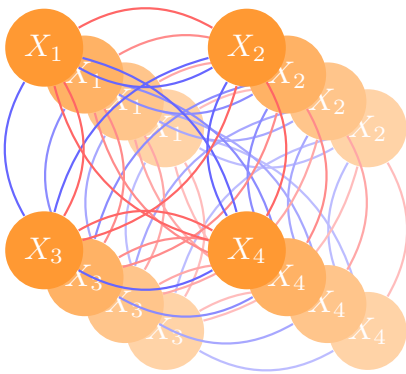
Group(s) induced by edges $(1, 2)$

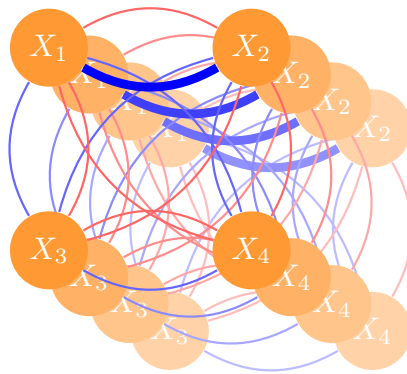
Group-Lasso

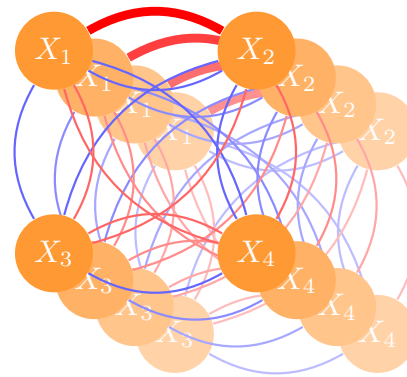




Cooperative-Lasso







Revisiting the Hess *et al.* data set

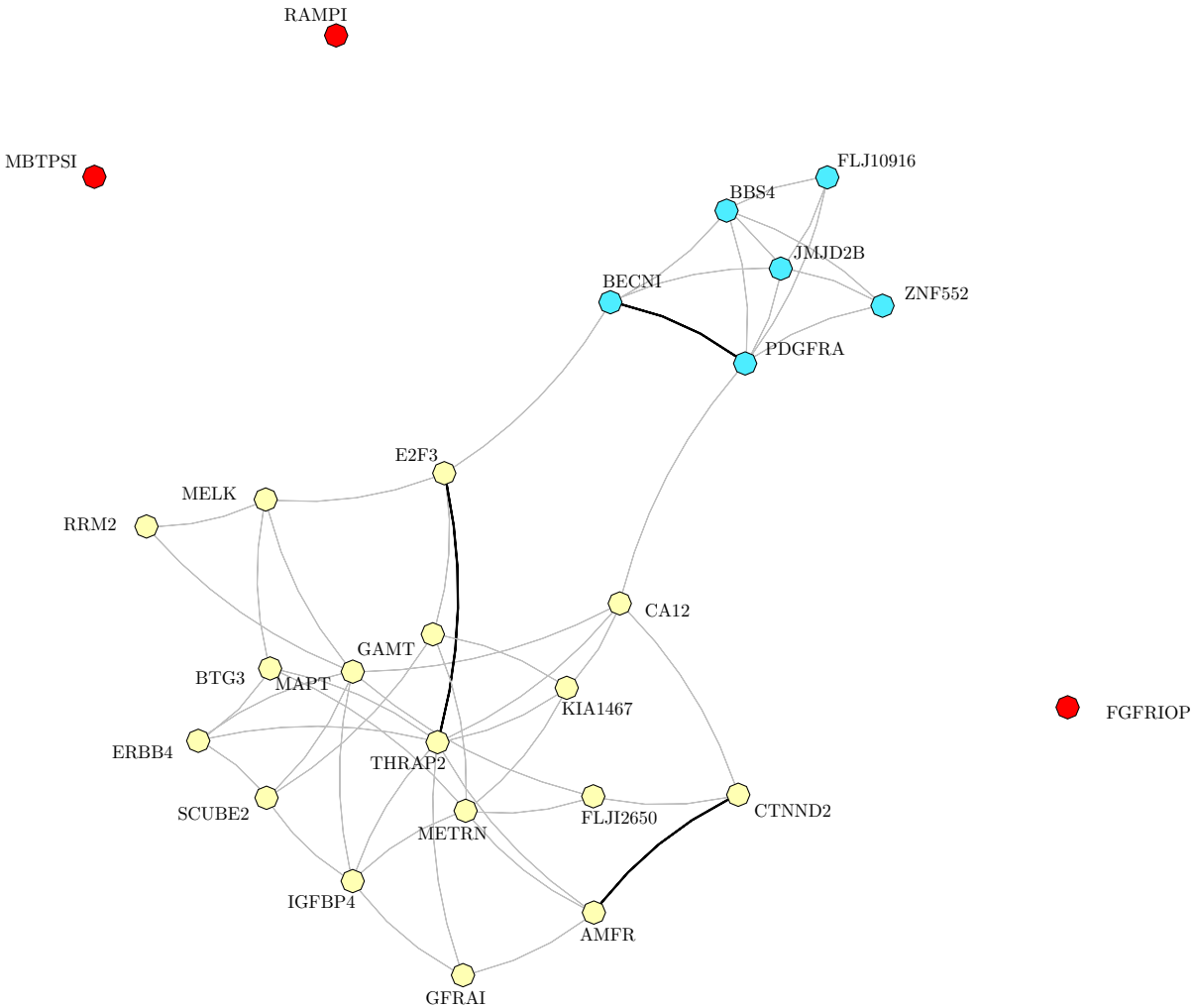
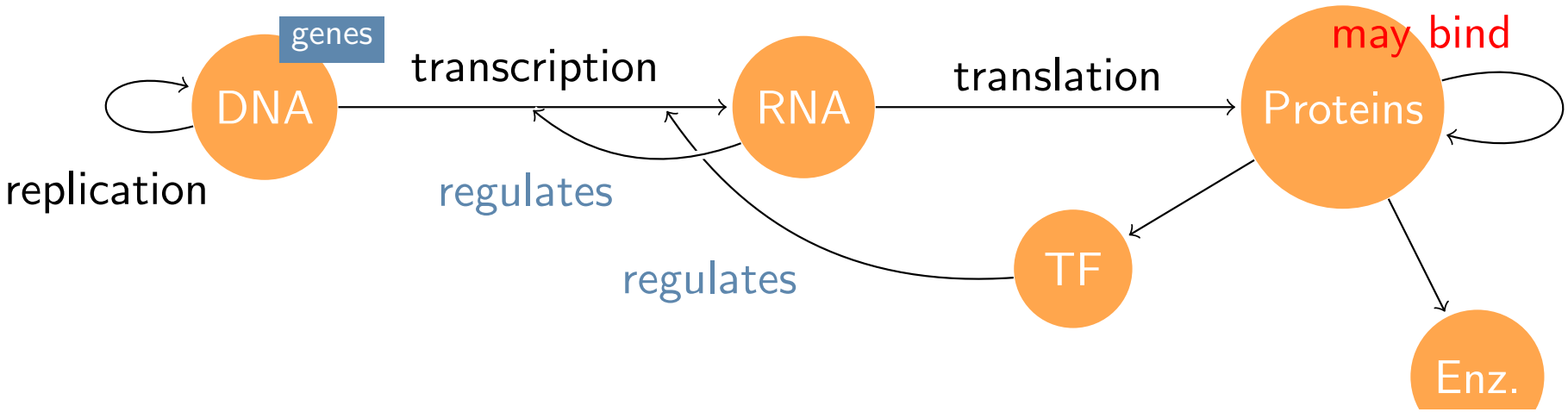


Figure: Cooperative-Lasso applied on the two sets of patients (PCR/noPCR). Bold edges are different in the finally selection graph.

Outline

Why Multi-attribute Networks?

Joint work with E. Kolaczyk (Boston) and C. Ambroise (Évry)



Data integration

- Omic technologies can profile cells at different levels: DNA, RNA, protein, chromosomal, and functional.
- multiple molecular profiles combined on the same set of biological samples can be *synergistic*.

Multiattribute GGM

Consider e.g. some p genes of interest and the $K = 2$ omic experiments

1. X_{i1} is the expression profile of gene i (transcriptomic data),
2. X_{i2} is the corresponding protein concentration (proteomic data).

Define a block-wise precision matrix

- $X = (X_1, \ldots, X_p)^T \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ in $\mathbb{R}^{pK}$,
- $X_i = (X_{i1}, \ldots, X_{iK})^\intercal \in \mathbb{R}^K$.

$$\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = \begin{bmatrix} \mathbf{\Omega}_{11} & & \mathbf{\Omega}_{1p} \\ & \ddots & \\ \mathbf{\Omega}_{p1} & & \mathbf{\Omega}_{pp} \end{bmatrix}, \qquad \mathbf{\Omega}_{ij} \in \mathcal{M}_{K,K}, \ \forall (i,j) \in \mathcal{P}^2.$$

Graphical Interpretation

Define $\mathcal{G} = (\mathcal{P}, \mathcal{E})$ as the multivariate analogue of the *conditional graph*:

$$(i,j) \in \mathcal{E} \Leftrightarrow \mathbf{\Omega}_{ij} \neq \mathbf{0}_{KK}.$$

Multiattribute GGM as multivariate regression

Multivariate analysis view point

Straightforward algebra and we have

$$X_j \,|\, X_{\setminus j} = x \sim \mathcal{N}(-\mathbf{\Omega}_{jj}^{-1}\mathbf{\Omega}_{j\setminus j}x, \mathbf{\Omega}_{ii}^{-1}) \ .$$

or equivalently, letting $\mathbf{B}_j^T = -\mathbf{\Omega}_{jj}^{-1}\mathbf{\Omega}_{i\setminus j}$,

$$X_j \,|\, X_{\setminus j} = \mathbf{B}_j^T X_{\setminus j} + \boldsymbol{\varepsilon}_j \quad \boldsymbol{\varepsilon}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}_{ii}^{-1}), \quad \boldsymbol{\varepsilon}_j \perp X.$$

Remembering the univariate case?

$$X_j | X_{\setminus j} = - \sum_{k \in \text{neighbors}(j)} \frac{\mathbf{\Omega}_{jk}}{\mathbf{\Omega}_{jj}} X_j + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}(0, \mathbf{\Omega}_{jj}^{-1}), \quad \varepsilon_j \perp X.$$

Multivariate neighborhood selection

The penalized multivariate regression approach

For each node /gene, recover its neighborhood by solving

$$\arg \min_{\mathbf{B}_j \in \mathcal{M}_{(p-1)K,K}} \frac{1}{2n} \left\| \mathbf{X}_j - \mathbf{X}_{\setminus j} \mathbf{B}_j \right\|_F^2 + \lambda \, \text{Pen}(\mathbf{B}_j),$$

Choice of penalty

Group-based penalty to activate the set of attributes simultaneously on a given link:

$$\text{Pen}(\mathbf{B}_j) = \sum_{k \neq j} \|\mathbf{B}_j^{(k)}\| \quad , \quad \mathbf{B}_j^{(k)} \in \mathcal{M}_{KK}$$

- $\|M\| = \|M\|_F = \left(\sum_{i,j} M_{ij}^2 \right)^{1/2}$, the Frobenius norm,
- $\|M\| = \|M\|_\infty = \max_{i,j} |M_{ij}|$, the sup norm (shared magnitude),
- $\|M\| = \|M\|_\star = \sum \text{eig}(M)$, the nuclear norm (rank penalty).

Breast cancer data: application

Two cohorts with both proteomic and transcriptomic data

1. NCI-60: $n = 60$ diverse human cancer cell lines, $p = 91$
2. RATHER: $n = 100$ sample from patients with breast cancer, $p = 117$

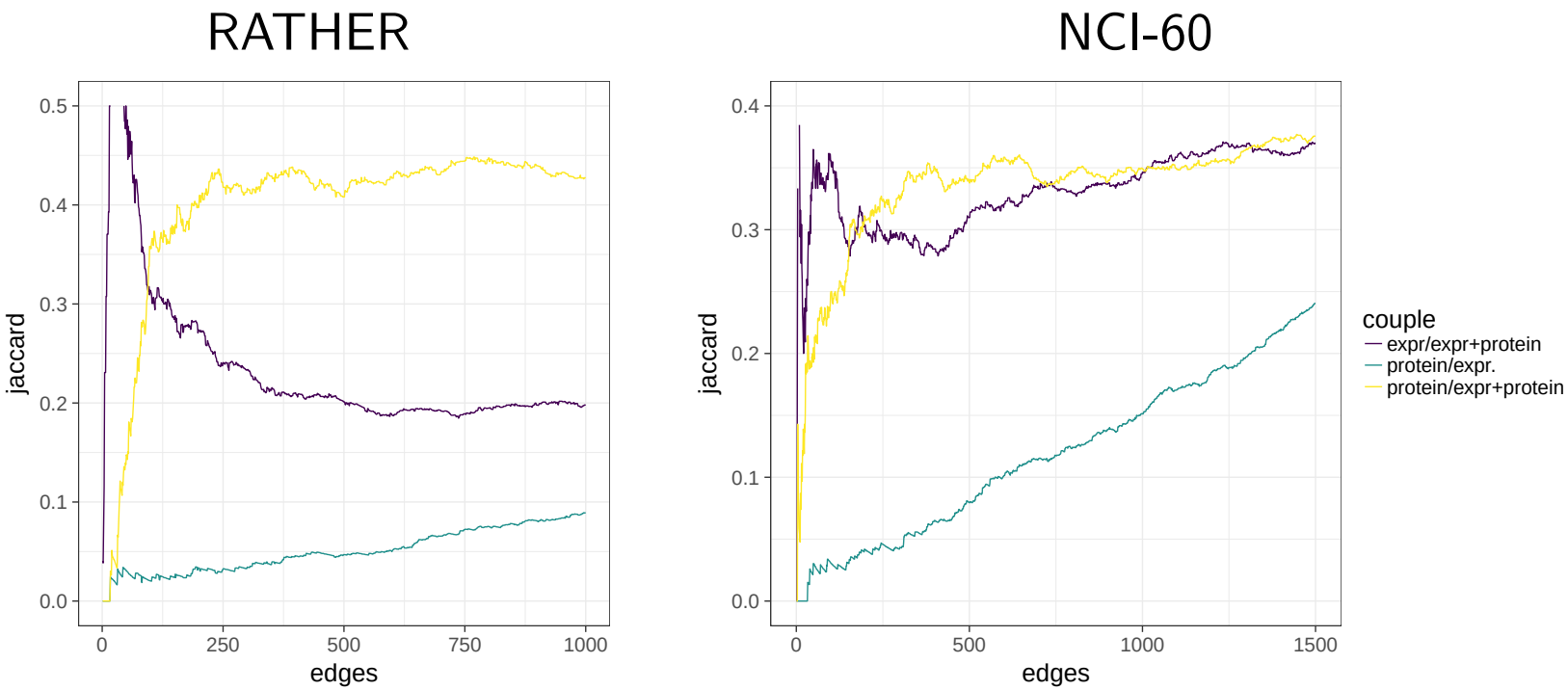


Figure: Jaccard's similarity index $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$ between uni-attribute and multiattribute networks, for RATHER and NCI60 data set: multiattribute networks share a high Jaccard index with both uni-attribute networks.

Inferred networks

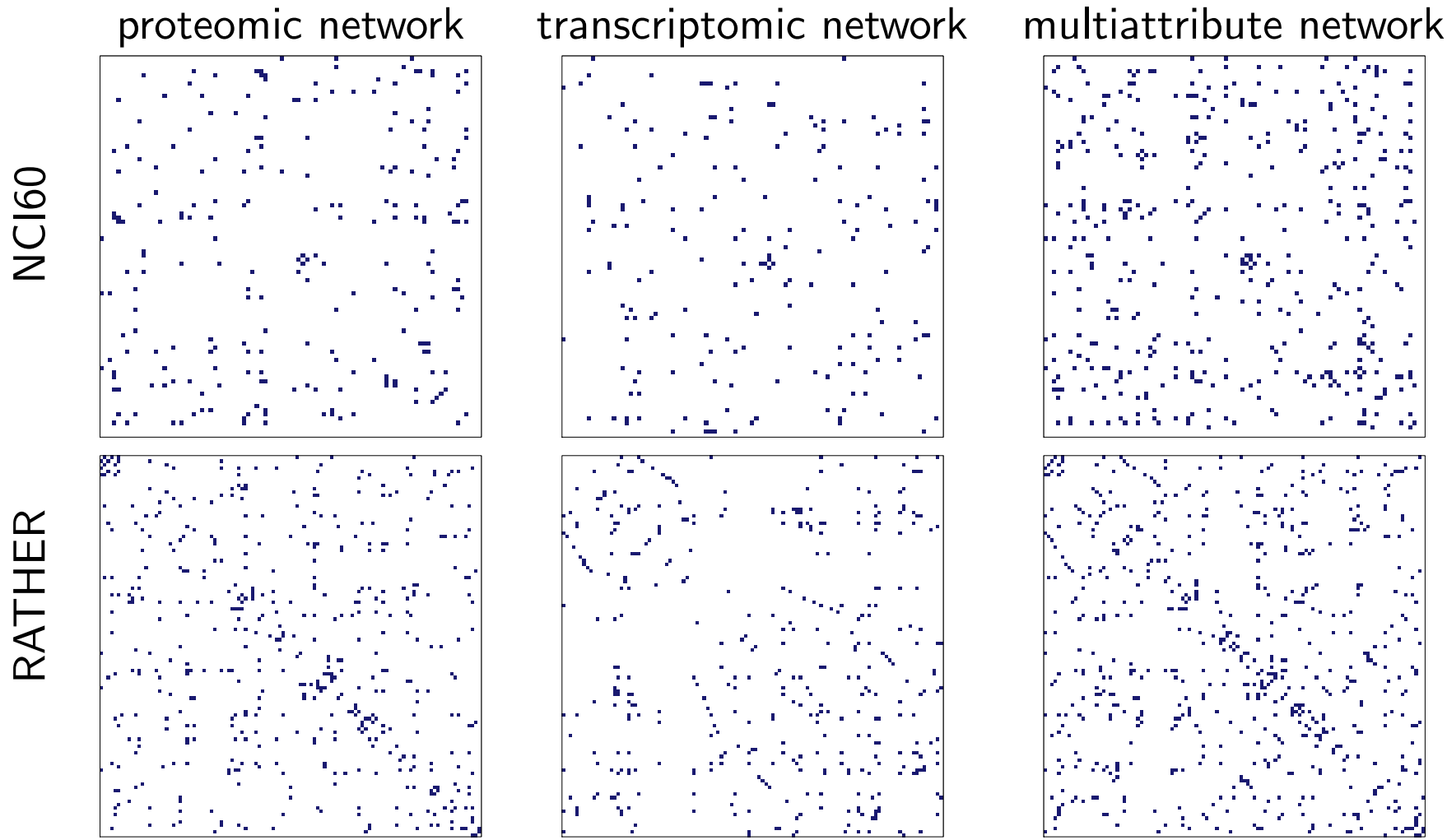


Figure: Uni-attribute and multiattribute networks inferred on both NCI60 and RATHER dataset. The number of neighbors of each entity is chosen by cross-validation. Multiattribute networks catch motif found in the uniattribute counterparts.

Outline

Motivations: oak powdery mildew pathobiome

Metabarcoding data from [JFS16]

- $n = 116$ leaves, $p = 114$ species (66 bacteria, 47 fungies + *E. alphitoides*)

```
##       f_1 f_2 f_3 f_4 E_alphitoides b_1045 b_109 b_1093
## A1.02  72   5 131   0             0      0     0      0
## A1.03 516  14 362   0             0      0     0      0
## A1.04 305  24 238   0             0      0     0      0
```

- $d = 8$ covariates (tree susceptibility, distance to trunk, orientation, . . .)

```
##         treeStatus orientation branch distToTrunk
## A1.02 intermediate          SW      1         202
## A1.03 intermediate          SW      1         175
## A1.04 intermediate          SW      1         168
```

- Sampling effort in each sample (bacteria $\neq$ fungi)

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,] 2488 2488 2488 2488 2488 8315 8315 8315
## [2,] 2054 2054 2054 2054 2054  662  662  662
## [3,] 2122 2122 2122 2122 2122  480  480  480
```

Problematic & Basic formalism

Data tables: $\mathbf{Y} = (Y_{ij}), n \times p$; $\mathbf{X} = (X_{ik}), n \times d$; $\mathbf{O} = (O_{ij}), n \times p$ where

- $Y_{ij} =$ abundance (read counts) of species (genes) j in sample i
- $X_{ik} =$ value of covariate k in sample i
- $O_{ij} =$ offset (sampling effort) for species j in sample i

Need for multivariate analysis to

- understand between-species/genes interactions
  $\rightsquigarrow$ 'network' inference (variable/covariance selection)
- correct for technical and confounding effects
  $\rightsquigarrow$ account for covariables and sampling effort

$\rightsquigarrow$ need a generic framework to model dependences between count variables

Models for multivariate count data

If we were in a Gaussian world, the general linear model would be appropriate

For each sample $i = 1, \ldots, n$, it explains

- the abundances of the p species ($\mathbf{Y}_i$)
- by the values of the d covariates $\mathbf{X}_i$ and the p offsets $\mathbf{O}_i$

$$\mathbf{Y}_i = \underbrace{\mathbf{X}_i\mathbf{B}}_{\text{account for covariates}} + \underbrace{\mathbf{O}_i}_{\text{account for sampling effort}} + \boldsymbol{\varepsilon}_i,\ \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}_p, \underbrace{\boldsymbol{\Sigma}}_{\text{dependence between species}})$$

+ null covariance $\Leftrightarrow$ independence $\rightsquigarrow$ uncorrelated species do not interact

But we are not, and there is no generic model for multivariate counts

- Data transformation ($\log$, $\sqrt{}$) : quick and dirty
- Non-Gaussian multivariate distributions: do not scale to data dimension yet
- Latent variable models: interaction occur in a latent (unobserved) layer

Models for multivariate count data

If we were in a Gaussian world, the general linear model would be appropriate

For each sample $i = 1, \ldots, n$, it explains

- the abundances of the p species ($\mathbf{Y}_i$)
- by the values of the d covariates $\mathbf{X}_i$ and the p offsets $\mathbf{O}_i$

$$\mathbf{Y}_i = \underbrace{\mathbf{X}_i\mathbf{B}}_{\text{account for covariates}} + \underbrace{\mathbf{O}_i}_{\text{account for sampling effort}} + \boldsymbol{\varepsilon}_i,\ \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}_p, \underbrace{\boldsymbol{\Sigma}}_{\text{dependence between species}})$$

+ ~~null covariance $\Leftrightarrow$ independence $\rightsquigarrow$ uncorrelated species do not interact~~

But we are not, and there is no generic model for multivariate counts

- Data transformation ($\log$, $\sqrt{}$) : quick and dirty
- Non-Gaussian multivariate distributions: do not scale to data dimension yet
- Latent variable models: interaction occur in a latent (unobserved) layer

Poisson-log normal (PLN) distribution

A latent Gaussian model

Originally proposed by Atchisson [AiH89]

$$\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$$

$$\mathbf{Y}_i \,|\, \mathbf{Z}_i \sim \mathcal{P}(\exp\{\mathbf{O}_i + \mathbf{X}_i^\intercal \mathbf{B} + \mathbf{Z}_i\})$$

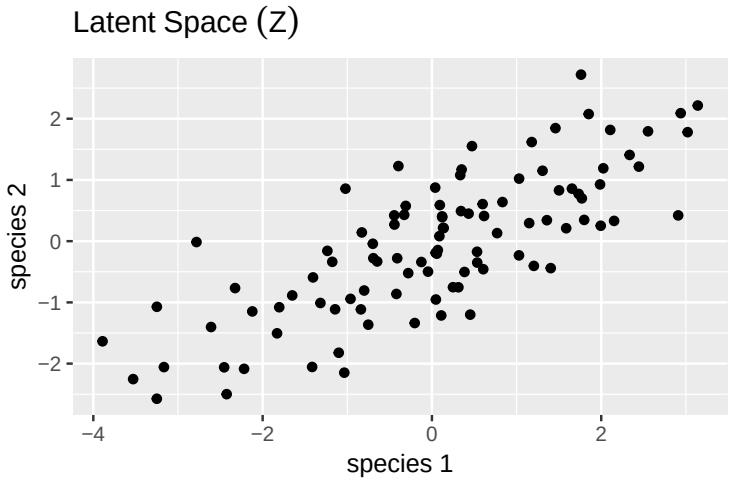
Interpretation

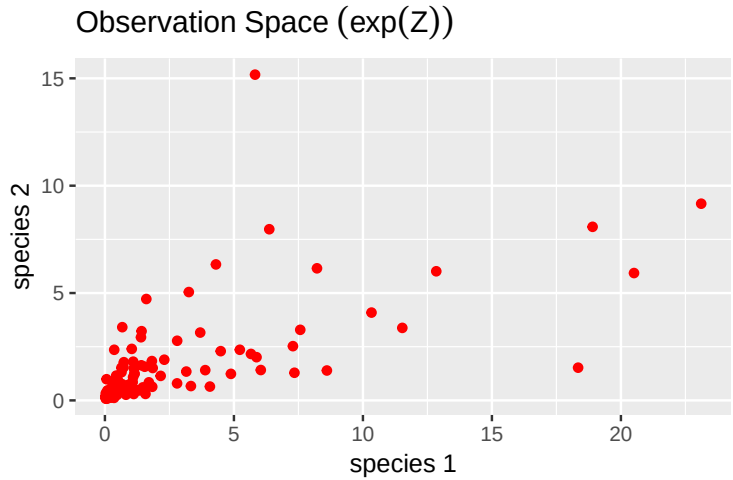
- Dependency structure encoded in the latent space (i.e. in $\mathbf{\Sigma}$)
- Additional effects are fixed
- Conditional Poisson distribution $=$ noise model

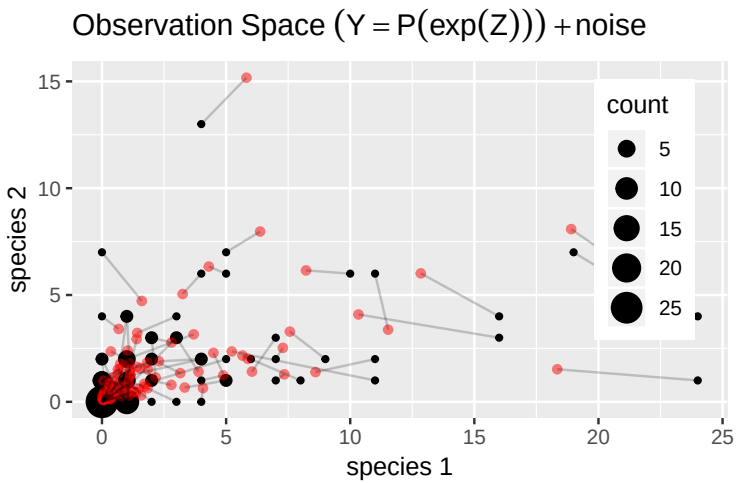
Properties

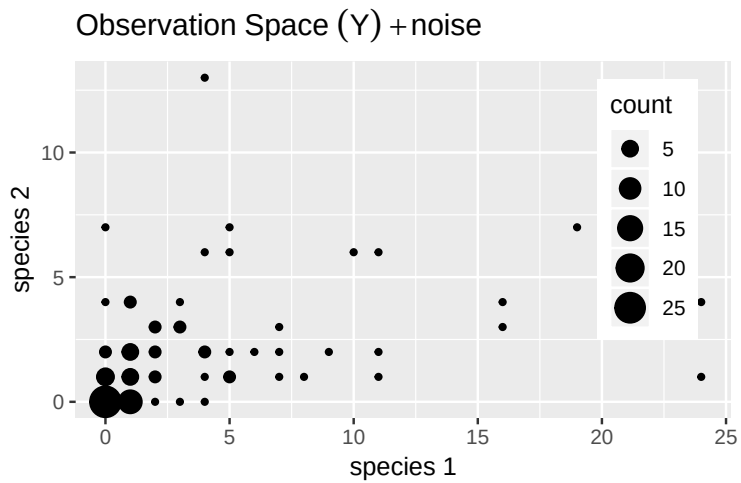
+ over-dispersion
+ covariance with arbitrary signs
- maximum likelihood via EM algorithm is limited to a couple of variables

Geometrical view









Our contributions

Algorithm/Numerical

A variational approach coupled with convex optimization techniques suited to higher dimensional data sets.
PLNmodels R/C++-package: `https://github.com/jchiquet/PLNmodels`

Extensions for multivariate analysis

Idea: put some additional constraint on the residual variance.

- Network Inference
  $\rightsquigarrow$ select direct interaction in $\boldsymbol{\Sigma}^{-1}$ via sparsity constraints
- *Principal component analysis*
  constraint the rank of $\boldsymbol{\Sigma}$ (most important effect in the variance)

Challenge: a variant of the variational algorithm is required for each model

PLN-network: unravel important interactions

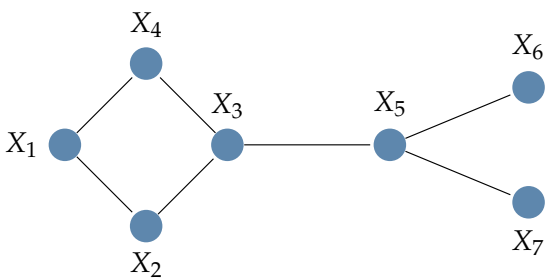
Variable selection of direct effects.

$$\mathbf{Z}_i \text{ iid} \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}), \qquad \|\boldsymbol{\Sigma}^{-1}\|_1 \leq c$$
$$\mathbf{Y}_i \,|\, \mathbf{Z}_i \sim \mathcal{P}(\exp\{\mathbf{O}_i + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\})$$

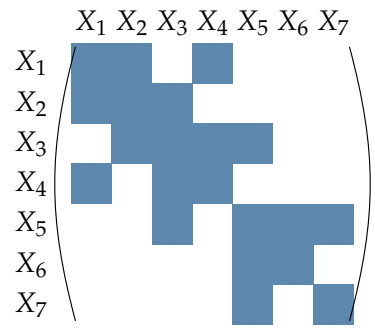
Interpretation: conditional independence structure.

$$(i,j) \notin \mathcal{E} \Leftrightarrow Z_i \perp\!\!\!\perp Z_j | Z_{\setminus\{i,j\}} \Leftrightarrow \boldsymbol{\Sigma}^{-1}_{ij} = 0.$$

$\mathcal{G} = (\mathcal{P}, \mathcal{E})$



$\boldsymbol{\Sigma}^{-1}$



PLN-network: find a sparse reconstruction of the latent inverse covariance
Iterate over variational estimator and Graphical-Lasso [BDE08,YL08,FHT07] in the latent layer

Networks of partial correlations for oak mildew pathobiome

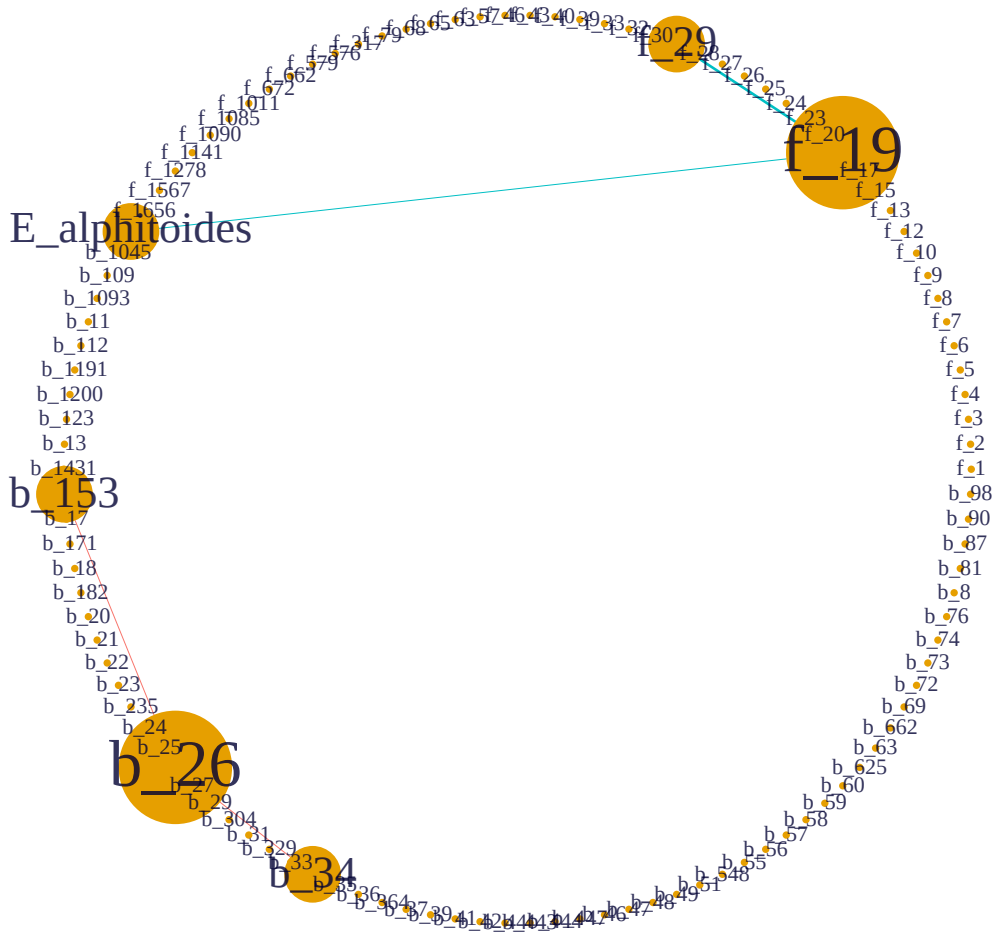
```
# Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models <- PLNnetwork(formula, penalties = 10^seq(log10(2), log10(0.6), len = 30))
```

Networks of partial correlations for oak mildew pathobiome

```
# Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models <- PLNnetwork(formula, penalties = 10^seq(log10(2), log10(0.6), len = 30))
```

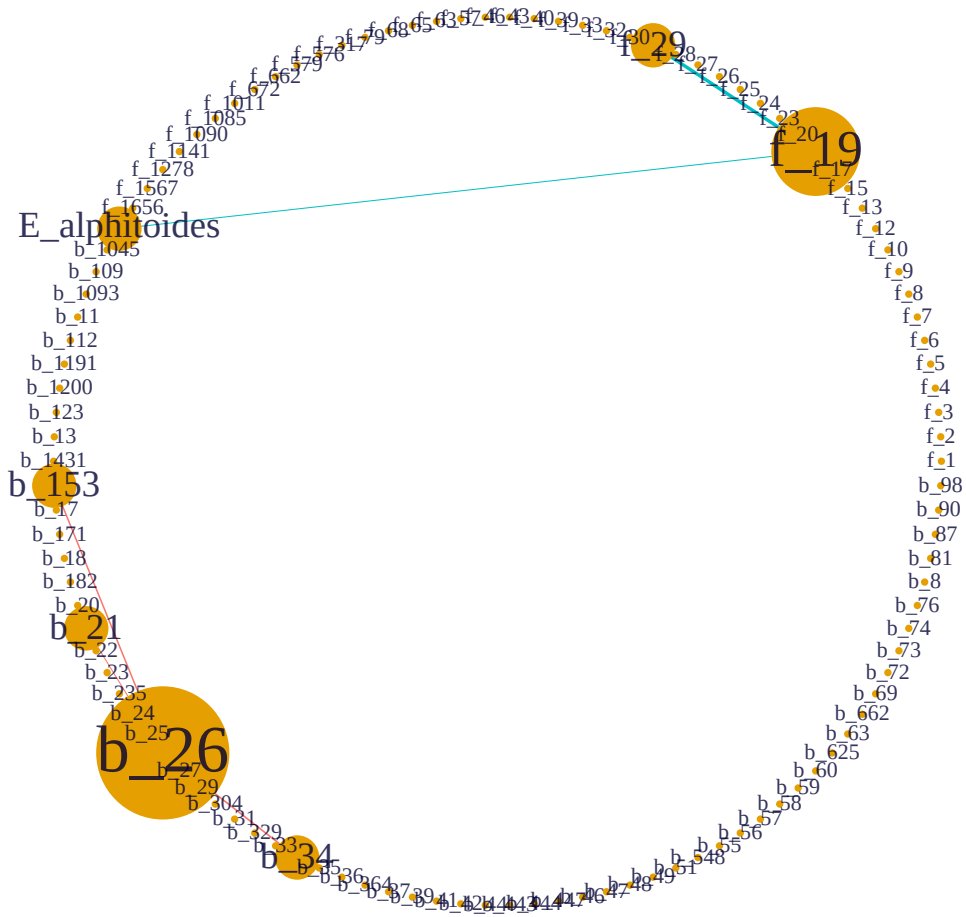
Networks of partial correlations for oak mildew pathobiome

```
# Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models <- PLNnetwork(formula, penalties = 10^seq(log10(2), log10(0.6), len = 30))
```



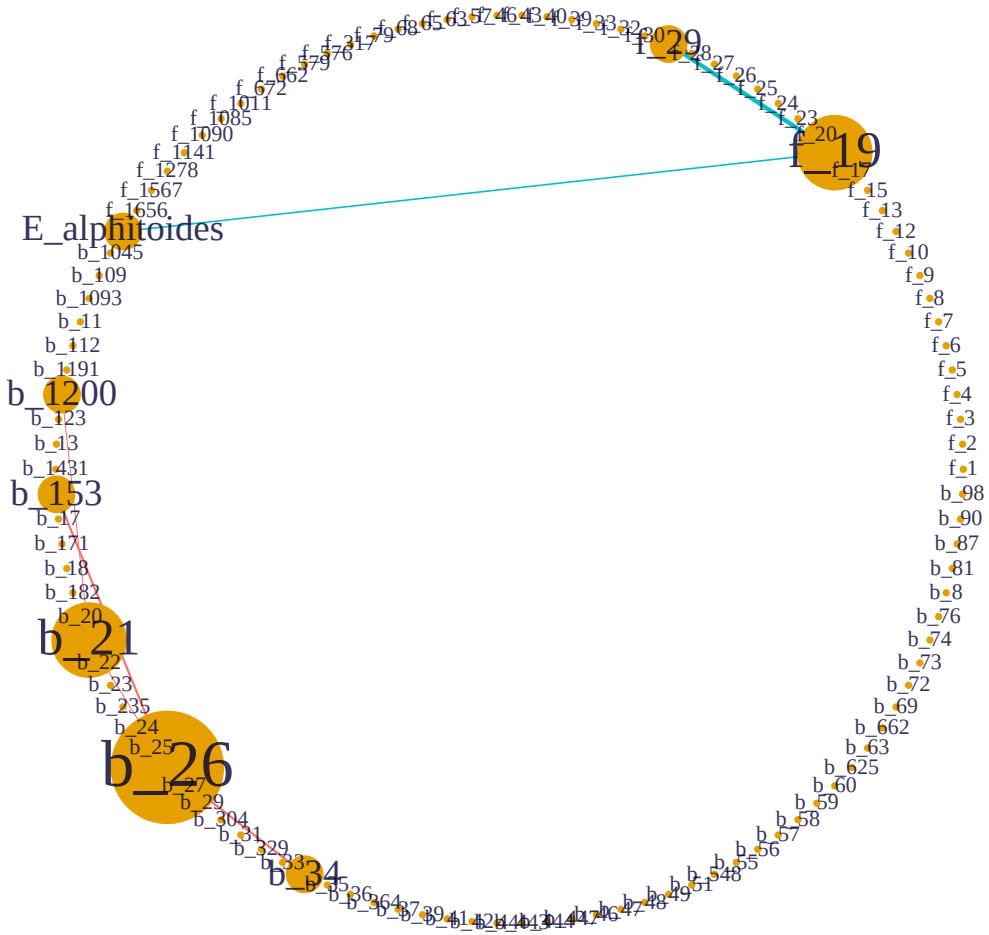
Networks of partial correlations for oak mildew pathobiome

```
# Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models <- PLNnetwork(formula, penalties = 10^seq(log10(2), log10(0.6), len = 30))
```



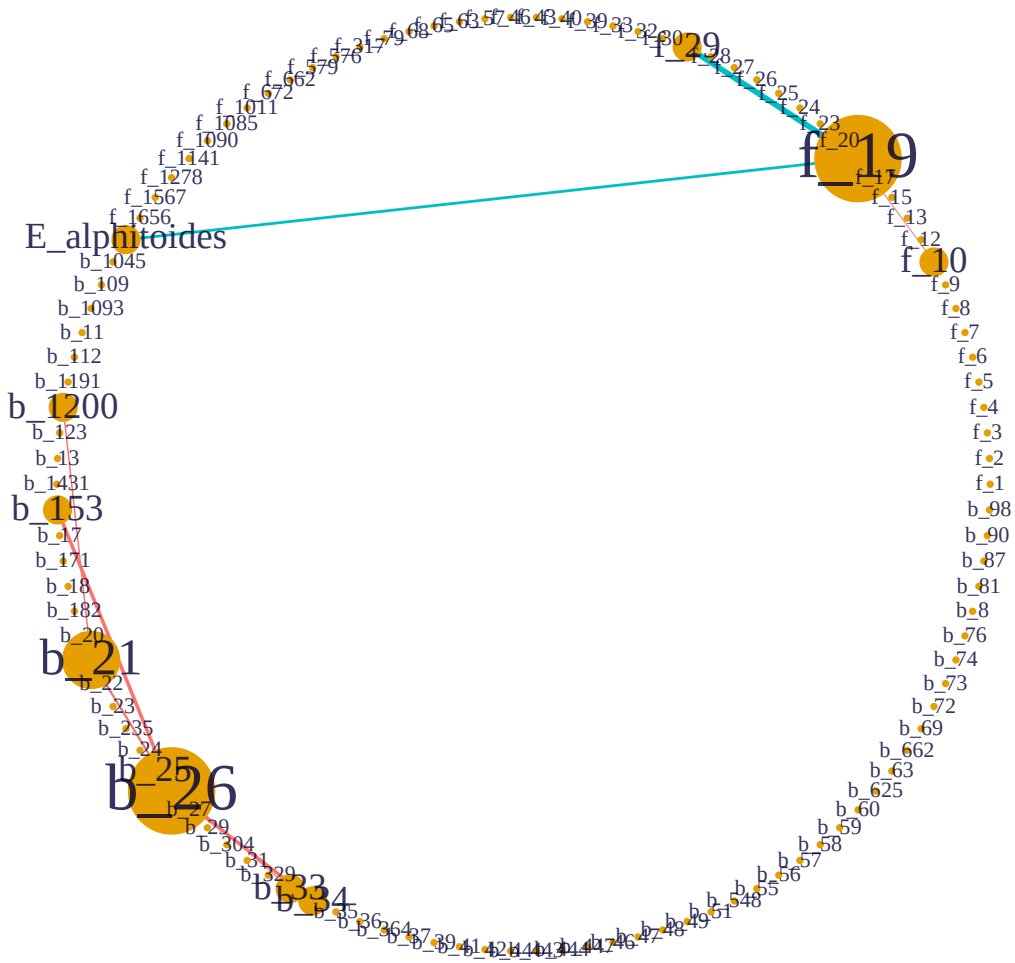
Networks of partial correlations for oak mildew pathobiome

```
# Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models <- PLNnetwork(formula, penalties = 10^seq(log10(2), log10(0.6), len = 30))
```



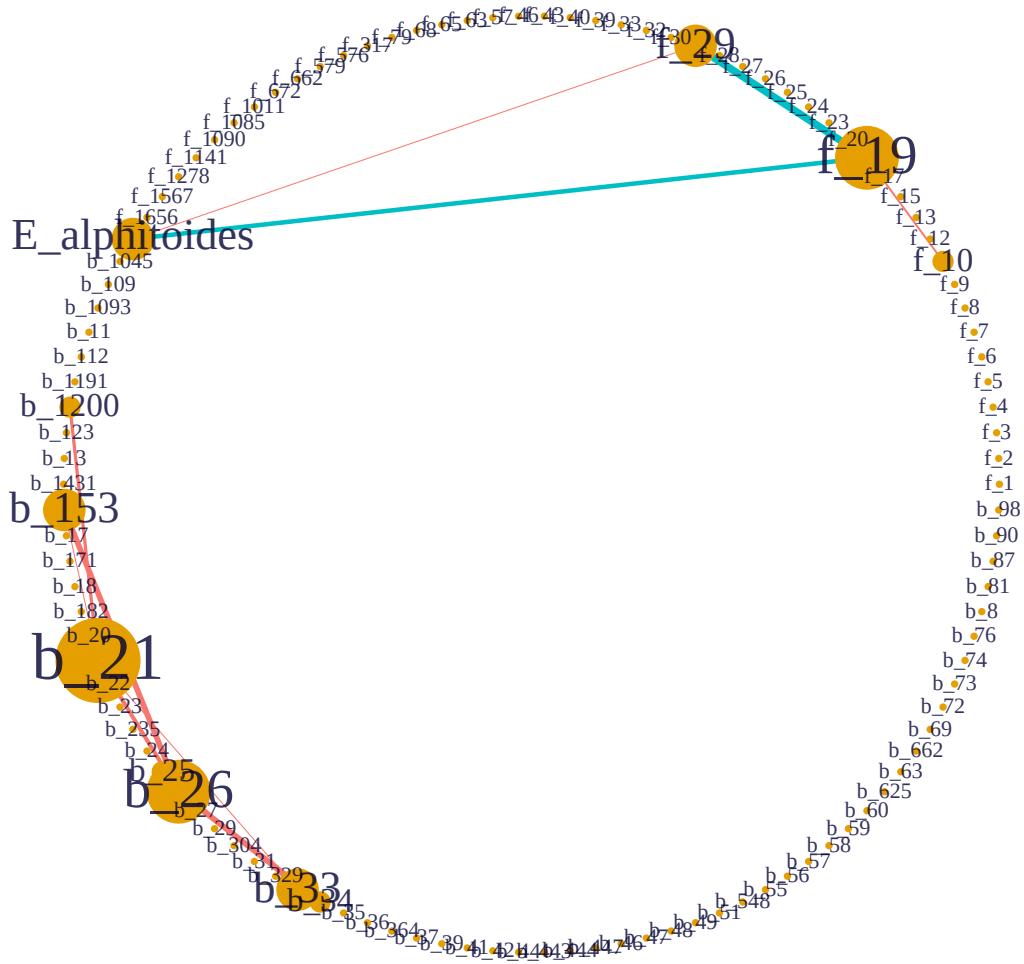
Networks of partial correlations for oak mildew pathobiome

```
# Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models <- PLNnetwork(formula, penalties = 10^seq(log10(2), log10(0.6), len = 30))
```



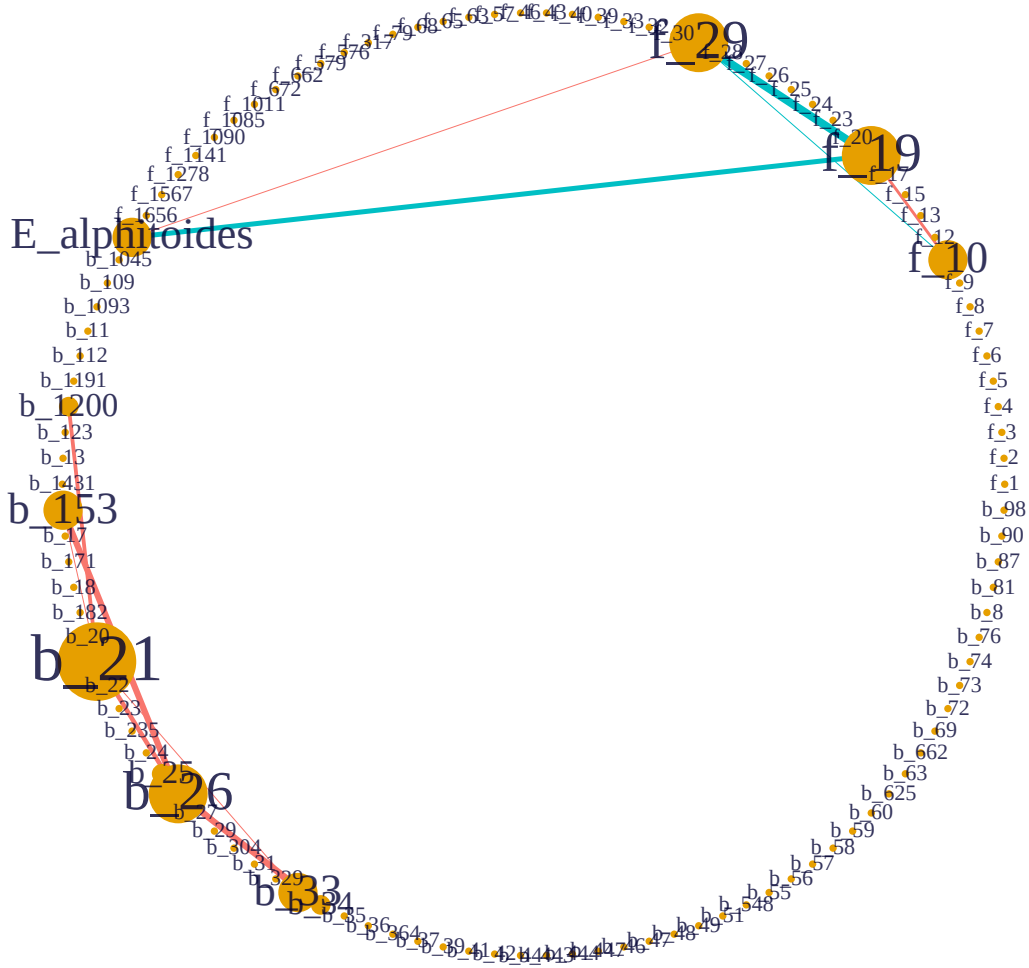
Networks of partial correlations for oak mildew pathobiome

```
# Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models <- PLNnetwork(formula, penalties = 10^seq(log10(2), log10(0.6), len = 30))
```



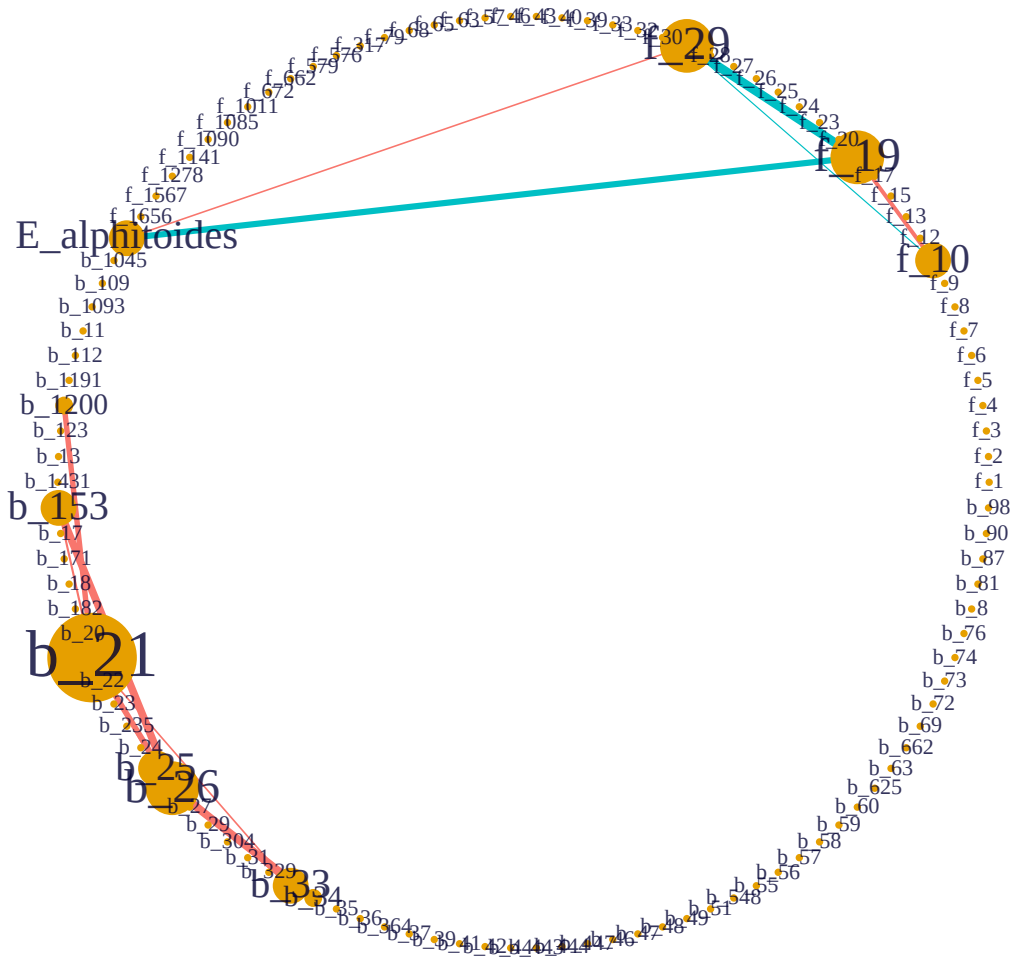
Networks of partial correlations for oak mildew pathobiome

```
# Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models <- PLNnetwork(formula, penalties = 10^seq(log10(2), log10(0.6), len = 30))
```



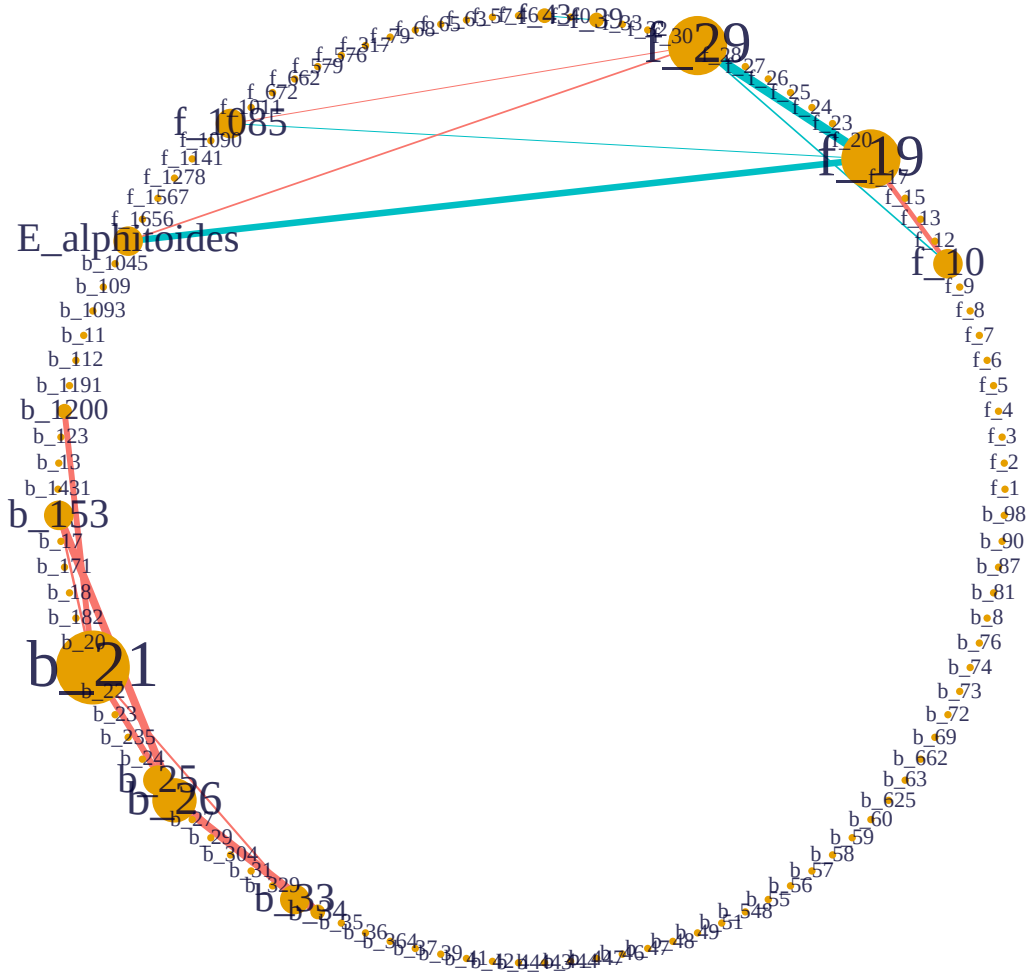
Networks of partial correlations for oak mildew pathobiome

```
# Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models <- PLNnetwork(formula, penalties = 10^seq(log10(2), log10(0.6), len = 30))
```



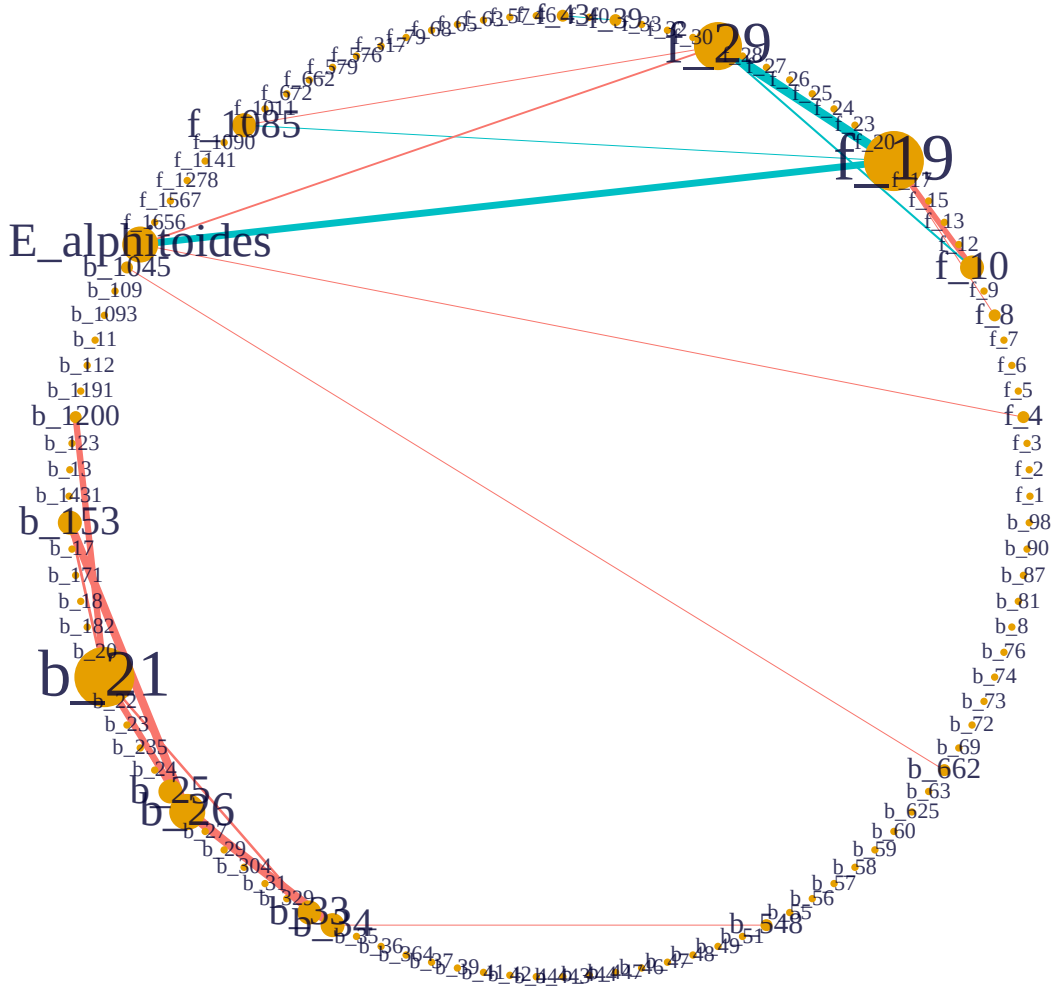
Networks of partial correlations for oak mildew pathobiome

```
# Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models <- PLNnetwork(formula, penalties = 10^seq(log10(2), log10(0.6), len = 30))
```



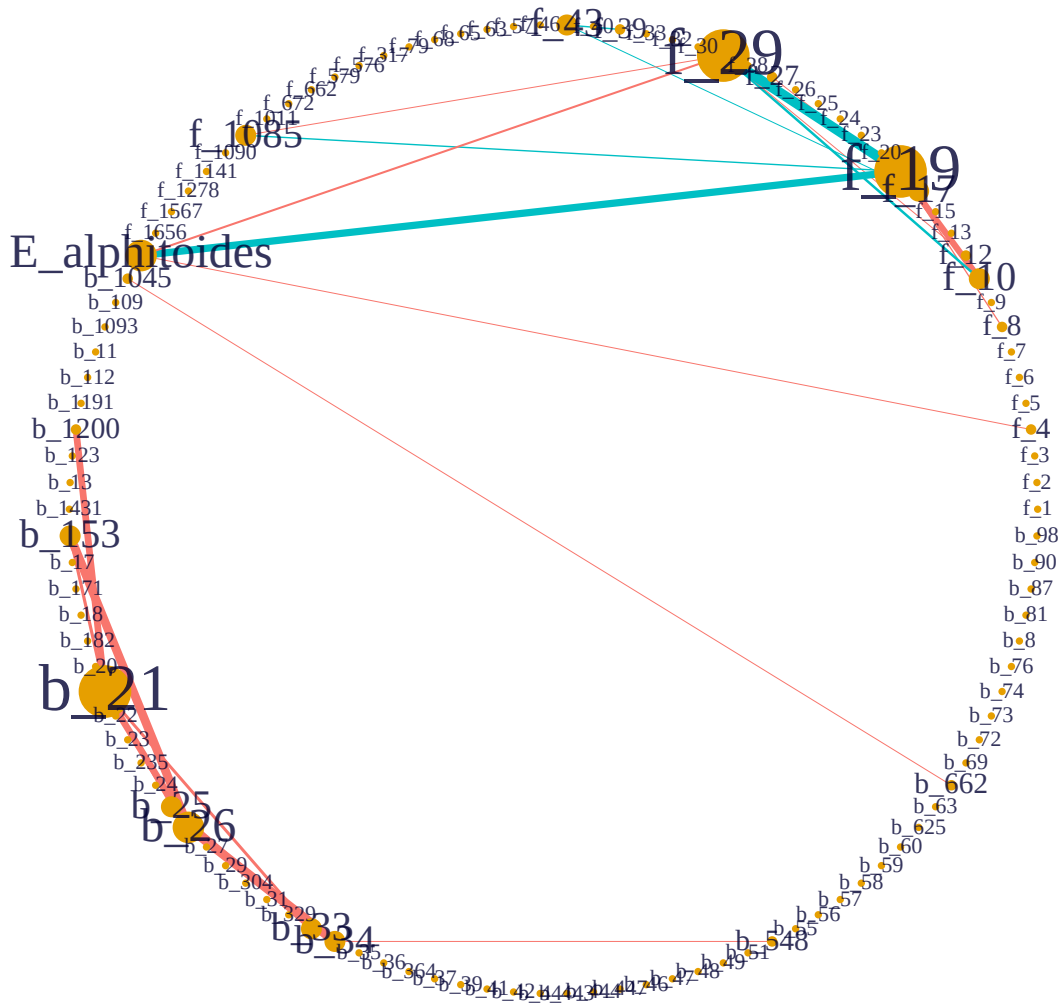
Networks of partial correlations for oak mildew pathobiome

```
# Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models <- PLNnetwork(formula, penalties = 10^seq(log10(2), log10(0.6), len = 30))
```



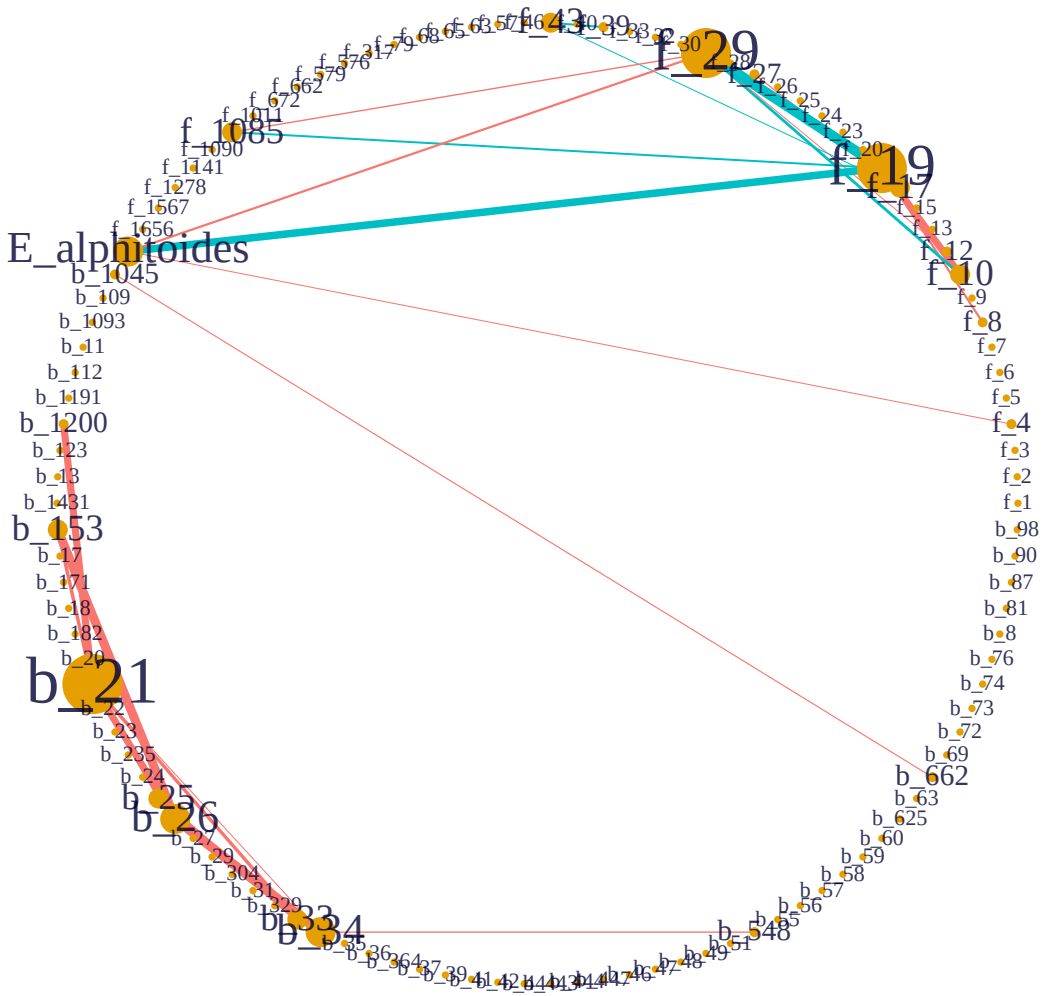
Networks of partial correlations for oak mildew pathobiome

```
# Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models <- PLNnetwork(formula, penalties = 10^seq(log10(2), log10(0.6), len = 30))
```



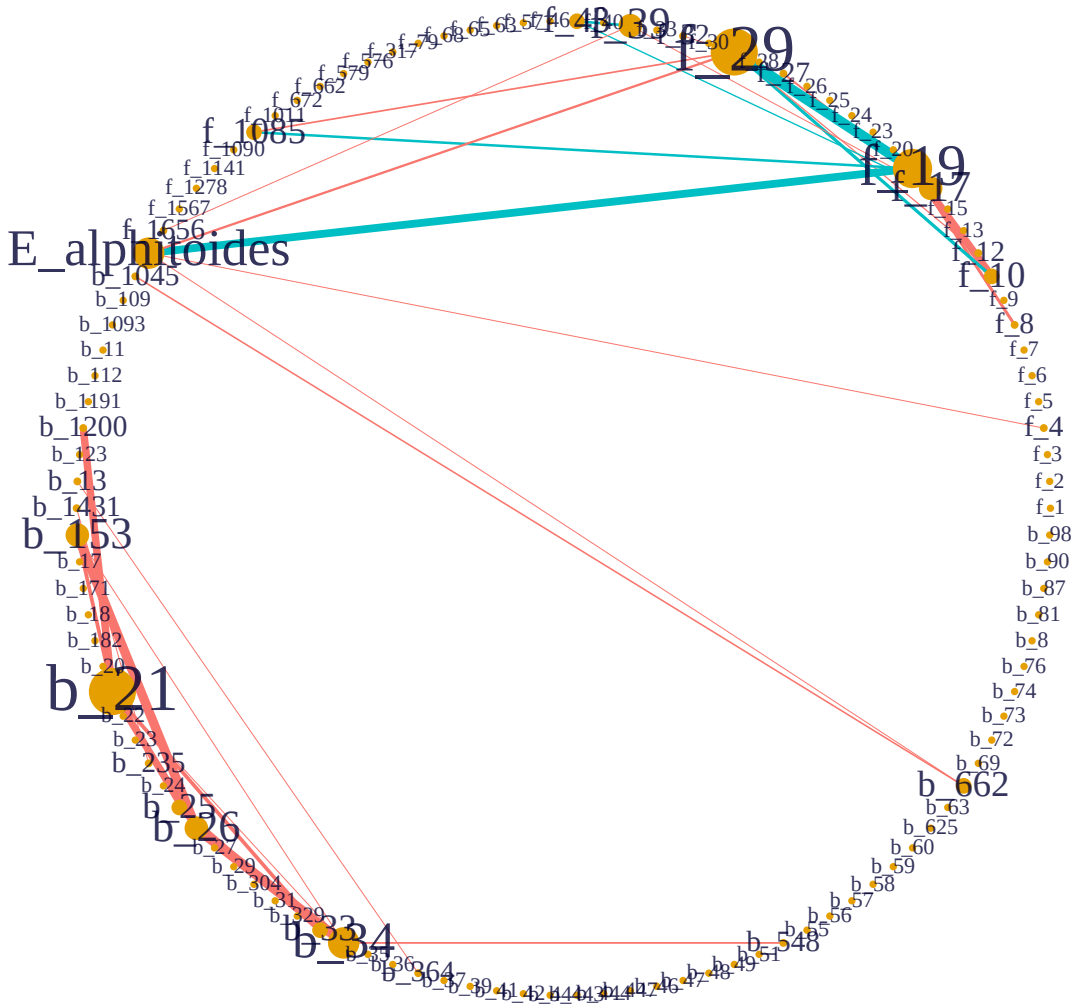
Networks of partial correlations for oak mildew pathobiome

```
# Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models <- PLNnetwork(formula, penalties = 10^seq(log10(2), log10(0.6), len = 30))
```



Networks of partial correlations for oak mildew pathobiome

```
# Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models <- PLNnetwork(formula, penalties = 10^seq(log10(2), log10(0.6), len = 30))
```



Networks of partial correlations for oak mildew pathobiome

```
# Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models <- PLNnetwork(formula, penalties = 10^seq(log10(2), log10(0.6), len = 30))
```