# Sparse Gaussian graphical models for biological network inference

## From gene expression to genomic network

Julien Chiquet

Bioinformatics Summer School in Port-Royal, July the 21st, 2016

UMR 518 AgroParisTech/INRA

`http://julien.cremeriefamily.info/bioinfo_ips2.html`

# Outline

Motivations

Network and data modeling

- Statistical dependence
- Gaussian Graphical models

Network Inference

- Inducing sparsity for edge selection
- Limitations and extensions of sparse GGM

# Outline

Motivations

Network and data modeling

Network Inference

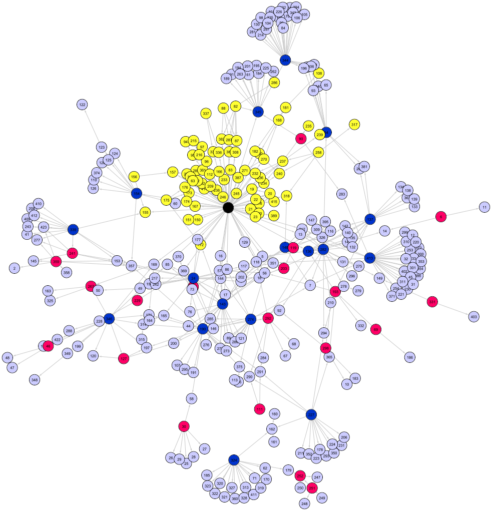
# Automatic reconstruction of biological networks (1)

## Regulatory networks

### E. coli regulatory network

Relationships between genes and their products

- highly structured
- always incomplete



# Automatic reconstruction of biological networks (2)

## Protein-Protein interaction networks

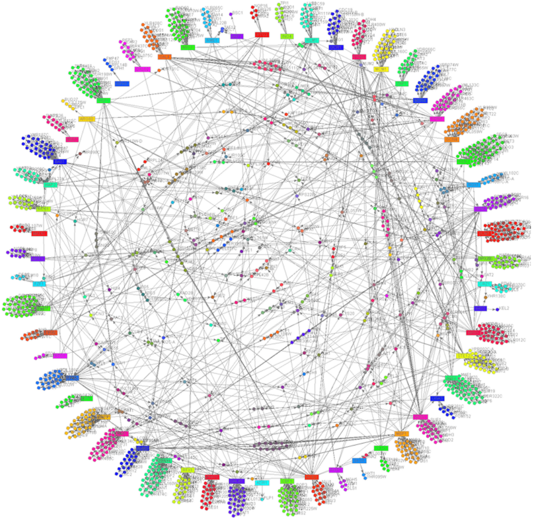


Figure: Yeast PPI network

# Automatic reconstruction of biological networks (3)

## Association networks

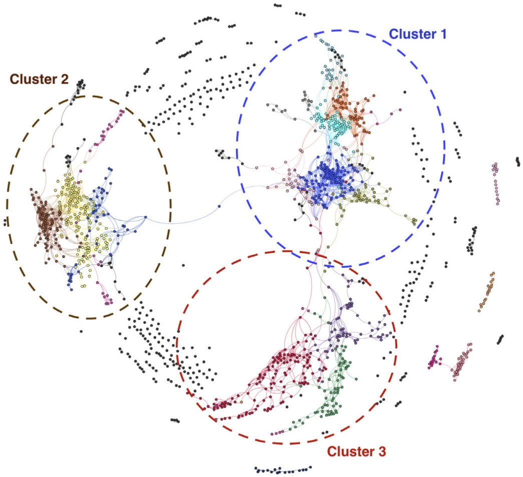
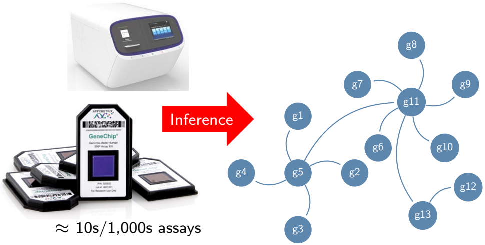


Figure: Co-occurence network between bacterial lineages of *Caulerpa*

# A challenging problem



≈ 10s/1,000s assays

≈ 1,000s/1,000,000s features

1. Nodes are fixed
   - restricted to a set of interest
2. Edges (interactions) are inferred
   - based upon statistical concepts

## Main statistical challenges

1. (Ultra) High dimensionality ($n < p$, $n \lll p$)
2. Heterogeneity/structure of the data

## Exploratory research

By pointing important actors (genes, OTU), it may assist the biologist in

1. formulating a hypothesis for further experiments,
2. unraveling main tendencies at play in complex systems.

# Outline

Motivations

Network and data modeling

- Statistical dependence
- Gaussian Graphical models

Network Inference

# Canonical model settings

## Biological microarrays in comparable conditions

### Notations

1. a set $\mathcal{P} = \{1, \dots, p\}$ of $p$ variables:
   these are typically the genes (could be proteins);
2. a sample $\mathcal{N} = \{1, \dots, n\}$ of individuals associated to the variables:
   these are typically the microarray (could be sequence counts).

### Basic statistical model

This can be view as

- a random vector $X$ in $\mathbb{R}^p$, whose $j$th entry is the $j$th variable,
- a $n$-size sample $(X^1, \dots, X^n)$, such as $X^i$ is the $i$th microarrays,
  - could be independent identically distributed copies (steady-state)
  - could be dependent in a certain way (time-course data)
- assume a parametric probability distribution for $X$ (Gaussian).

# Canonical model settings

Biological microarrays in comparable conditions

## Notations

1. a set $\mathcal{P} = \{1, \dots, p\}$ of $p$ variables:
   these are typically the genes (could be proteins);
2. a sample $\mathcal{N} = \{1, \dots, n\}$ of individuals associated to the variables:
   these are typically the microarray (could be sequence counts).

## Basic statistical model

This can be view as

- a random vector $X$ in $\mathbb{R}^p$, whose $j$th entry is the $j$th variable,
- a $n$-size sample $(X^1, \dots, X^n)$, such as $X^i$ is the $i$th microarrays,
  - could be independent identically distributed copies (steady-state)
  - could be dependent in a certain way (time-course data)
- assume a parametric probability distribution for $X$ (Gaussian).

# Canonical model settings

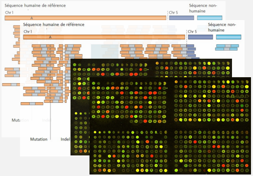
## Biological microarrays in comparable conditions

### Notations

1. a set $\mathcal{P} = \{1, \dots, p\}$ of $p$ variables:
   these are typically the genes (could be proteins);

### The data

Stacking $(X^1, \dots, X^n)$, we met the usual individual/variable table $\mathbf{X}$



stacked in

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

- a $n$-size sample $(X^1, \dots, X^n)$, such as $X^i$ is the $i$th microarrays,
  - could be independent identically distributed copies (steady-state)
  - could be dependent in a certain way (time-course data)
- assume a parametric probability distribution for $X$ (Gaussian).

# Outline

Motivations

Network and data modeling

- Statistical dependence
- Gaussian Graphical models

Network Inference

# Modeling relationship between variables (1)

## Independence

### Definition (Independence of events)

Two events $A$ and $B$ are independent if and only if

$$\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B),$$

which is usually denoted by $A \perp\!\!\!\perp B$. Equivalently,

- $A \perp\!\!\!\perp B \Leftrightarrow \mathbb{P}(A|B) = \mathbb{P}(A)$,
- $A \perp\!\!\!\perp B \Leftrightarrow \mathbb{P}(A|B) = \mathbb{P}(A|B^c)$

### Example (class vs party)

|  | party |  |
|---|---|---|
| class | Labour | Tory |
| working | 0.42 | 0.28 |
| bourgeoisie | 0.06 | 0.24 |

|  | party |  |
|---|---|---|
| class | Labour | Tory |
| working | 0.60 | 0.40 |
| bourgeoisie | 0.20 | 0.80 |

Table: Joint probability (left) vs. conditional probability (right)

# Modeling relationship between variables (1)

## Independence

### Definition (Independence of events)

Two events $A$ and $B$ are independent if and only if

$$\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B),$$

which is usually denoted by $A \perp\!\!\!\perp B$. Equivalently,

- $A \perp\!\!\!\perp B \Leftrightarrow \mathbb{P}(A|B) = \mathbb{P}(A)$,
- $A \perp\!\!\!\perp B \Leftrightarrow \mathbb{P}(A|B) = \mathbb{P}(A|B^c)$

### Example (class vs party)

|  | party |  |
|---|---|---|
| class | Labour | Tory |
| working | 0.42 | 0.28 |
| bourgeoisie | 0.06 | 0.24 |

|  | party |  |
|---|---|---|
| class | Labour | Tory |
| working | 0.60 | 0.40 |
| bourgeoisie | 0.20 | 0.80 |

Table: Joint probability (left) vs. conditional probability (right)

# Modeling relationships between variables (2)

## Conditional independence

Generalizing to more than two events requires strong assumptions (mutual independence). Better handle with

### Definition (Conditional independence of events)

Two events $A$ and $B$ are conditionally independent if and only if

$$\mathbb{P}(A, B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C),$$

which is usually denoted by $A \perp\!\!\!\perp B|C$

### Example (Does QI depends on weight?)

Consider the events $A =$ "having low QI", $B =$ "having low weight".

# Modeling relationships between variables (2)

## Conditional independence

Generalizing to more than two events requires strong assumptions (mutual independence). Better handle with

### Definition (Conditional independence of events)

Two events $A$ and $B$ are conditionally independent if and only if

$$\mathbb{P}(A, B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C),$$

which is usually denoted by $A \perp\!\!\!\perp B|C$

### Example (Does QI depends on weight?)

Consider the events $A =$ "having low QI", $B =$ "having low weight".

# Modeling relationships between variables (2)

## Conditional independence

Generalizing to more than two events requires strong assumptions (mutual independence). Better handle with

### Definition (Conditional independence of events)

Two events $A$ and $B$ are conditionally independent if and only if

$$\mathbb{P}(A, B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C),$$

which is usually denoted by $A \perp\!\!\!\perp B|C$

### Example (Does QI depends on weight?)

Consider the events $A =$ "having low QI", $B =$ "having low weight". Estimating[1] $\mathbb{P}(A, B)$, $\mathbb{P}(A)$ and $\mathbb{P}(B)$ in a sample would lead to

$$\mathbb{P}(A, B) \neq \mathbb{P}(A)\mathbb{P}(B)$$

[1] stupidly

# Modeling relationships between variables (2)

## Conditional independence

Generalizing to more than two events requires strong assumptions (mutual independence). Better handle with

### Definition (Conditional independence of events)

Two events $A$ and $B$ are conditionally independent if and only if

$$\mathbb{P}(A, B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C),$$

which is usually denoted by $A \perp\!\!\!\perp B|C$

### Example (Does QI depends on weight?)

Consider the events $A =$ "having low QI", $B =$ "having low weight". But in fact, introducing $C =$ "having a given age",

$$\mathbb{P}(A, B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C)$$

# Outline

Motivations

Network and data modeling

- Statistical dependence
- Gaussian Graphical models

Network Inference

# 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 $X \sim \mathbb{P}$.

## Definition

The conditional independence graph of a random vector $X$ is the undirected graph $\mathcal{G} = \{\mathcal{P}, \mathcal{E}\}$ with the set of node $\mathcal{P} = \{1, \ldots, p\}$ and where

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

# 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.) $X = \{X_1, \dots, X_p\}$ with distribution $\mathbb{P}$,

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

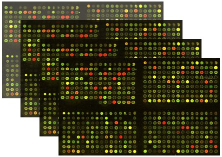
## Definition

The conditional independence graph of a random vector $X$ is the undirected graph $\mathcal{G} = \{\mathcal{P}, \mathcal{E}\}$ with the set of node $\mathcal{P} = \{1, \dots, p\}$ and 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:

- $\leadsto$ naturally links independence and conditional independence to the covariance and partial covariance,
- $\leadsto$ 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

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

$$\rho_{XY} = \operatorname{cor}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{Var}(X) \;\cdot\; \operatorname{Var}(Y)}}.$$

### Proposition

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

# Why Gaussianity helps?

## Case of 2 variables or size-2 random vector

Let $X$, $Y$ be two real random variables.

### Definitions

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

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

### Proposition

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

# The bivariate Gaussian distribution

## The Covariance Matrix

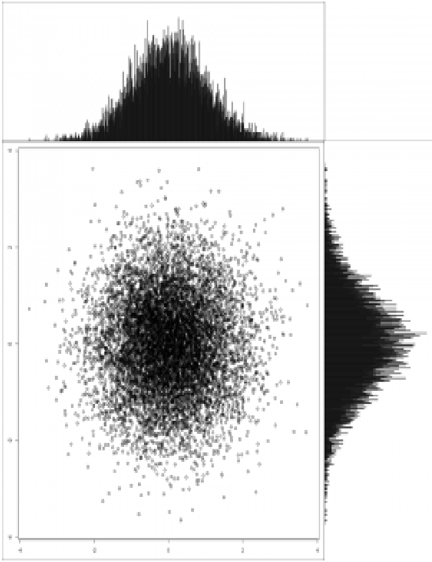
Let

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

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

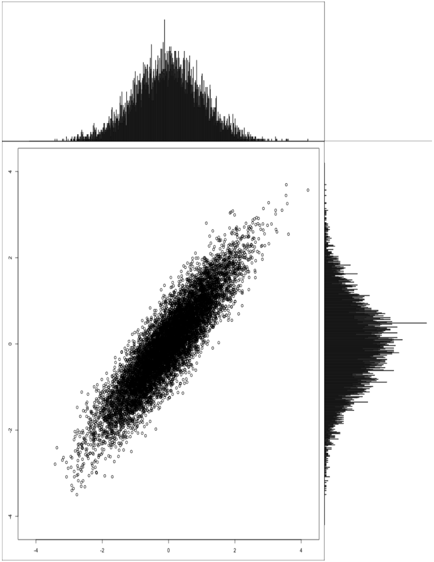
Let

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

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

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

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

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

### Proposition

*When $X, Y, Z$ are jointly Gaussian, then*

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

# Generalization: multivariate Gaussian vector

## Now need partial covariance and partial correlation

Let $X, Y, Z$ be real random variables.

### Definitions

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

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

$\rightsquigarrow$ 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 $\mathcal{P} = \{1, \dots, p\}$ of genes is described by $X \in \mathbb{R}^p$ such as

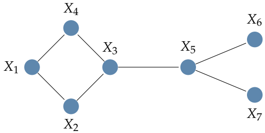
1. $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ the precision matrix.
2. a sample $(X^1, \dots, X^n)$ of exp. stacked in an $n \times p$ data matrix $\mathbf{X}$.

## Conditional independence structure

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

## Graphical interpretation

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



$$\boldsymbol{\Theta}$$

$$\begin{array}{c|ccccccc|}
 & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \\
X_1 & \blacksquare & \blacksquare & & \blacksquare & & & \\
X_2 & \blacksquare & \blacksquare & \blacksquare & & & & \\
X_3 & & \blacksquare & \blacksquare & \blacksquare & \blacksquare & & \\
X_4 & \blacksquare & & \blacksquare & \blacksquare & & & \\
X_5 & & & \blacksquare & & \blacksquare & \blacksquare & \blacksquare \\
X_6 & & & & & \blacksquare & \blacksquare & \\
X_7 & & & & & \blacksquare & & \blacksquare
\end{array}$$

⇝ “Covariance” selection

# Outline

Motivations

Network and data modeling

Network Inference

Inducing sparsity for edge selection

Limitations and extensions of sparse GGM

# Outline

Motivations

Network and data modeling

Network Inference

- Inducing sparsity for edge selection
- Limitations and extensions of sparse GGM

# 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 $\ell$ 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 $i$th 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 $\boldsymbol{\Theta}$.

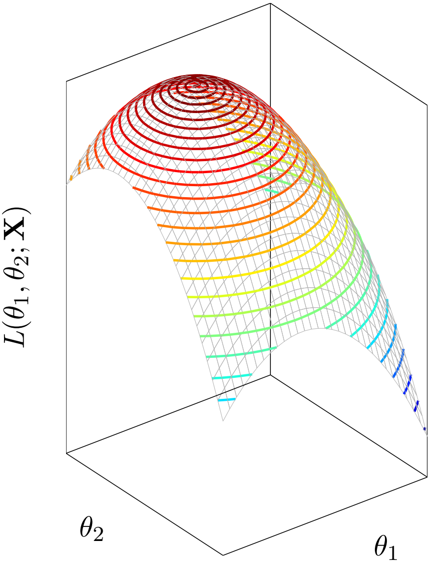
## The log-likelihood

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

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

# A Geometric View of Shrinkage

## Constrained Optimization



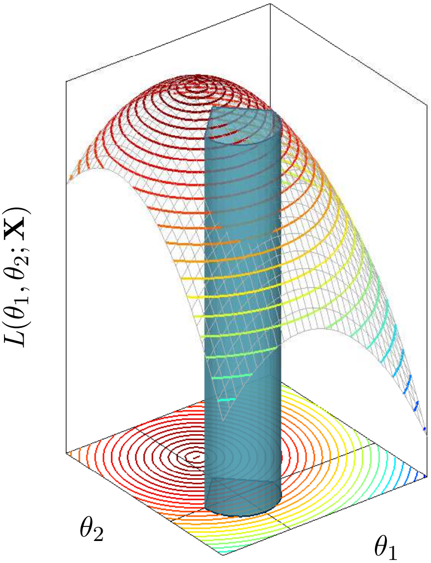
We basically want to solve a problem of the form

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

where $\ell$ is typically a concave likelihood function.

# A Geometric View of Shrinkage

## Constrained Optimization



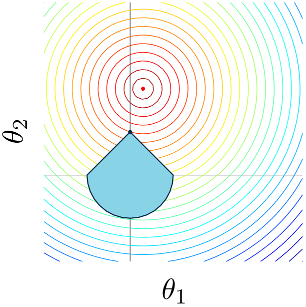
$$\begin{cases} \underset{\theta_1, \theta_2}{\text{maximize}} & \ell(\theta_1, \theta_2; \mathbf{X}) \\ \text{s.t.} & \Omega(\theta_1, \theta_2) \leq c \end{cases},$$

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

How shall we define $\Omega$ ?

# A Geometric View of Shrinkage

## Constrained Optimization



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

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

How shall we define $\Omega$ ?

# The Lasso

## Least Absolute Shrinkage and Selection Operator

### Idea

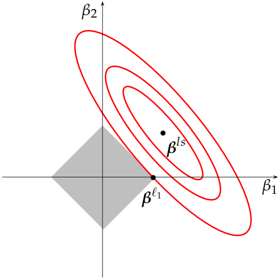
Suggest an admissible set that induces sparsity (force several entries to exactly zero in $\hat{\boldsymbol{\beta}}$).

### Lasso as a regularization problem

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

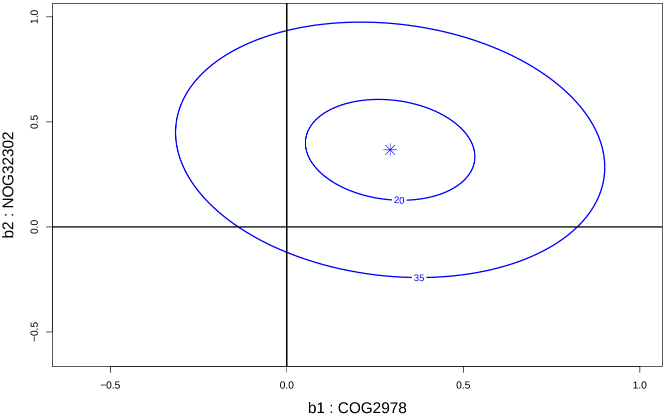
$$\hat{\boldsymbol{\theta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\theta}} -\ell(\boldsymbol{\theta}), \quad \text{s.t. } \sum_{j=1}^{p} |\theta_j| \leq s,$$

where $s$ is a shrinkage factor.



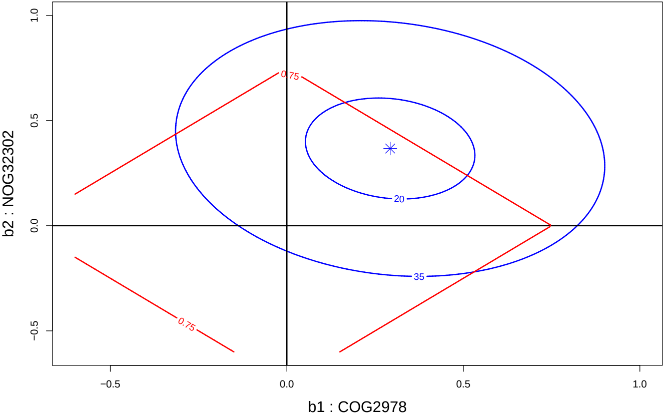
# Insights: 2-dimensional example with the square loss

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



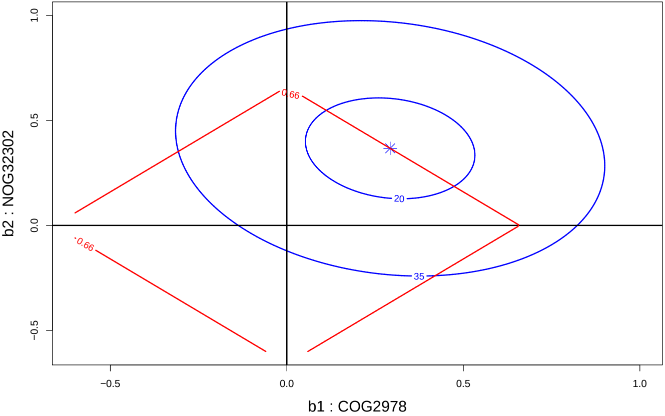
# Insights: 2-dimensional example with the square loss

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



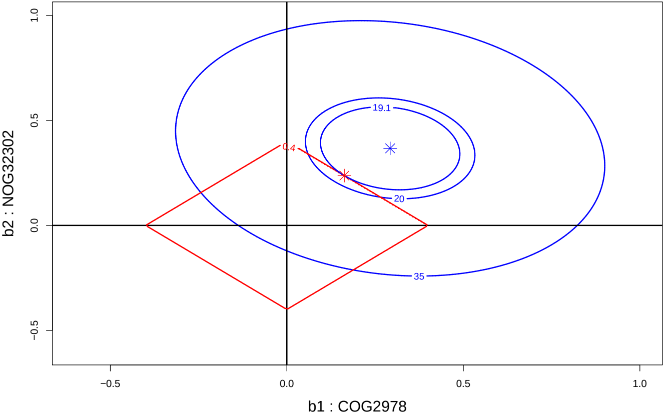
# Insights: 2-dimensional example with the square loss

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



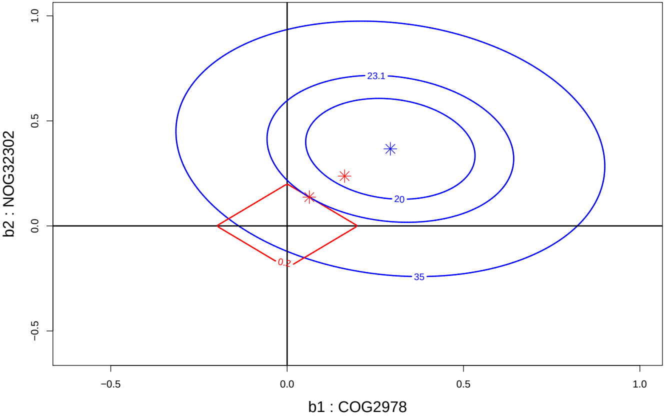
# Insights: 2-dimensional example with the square loss

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



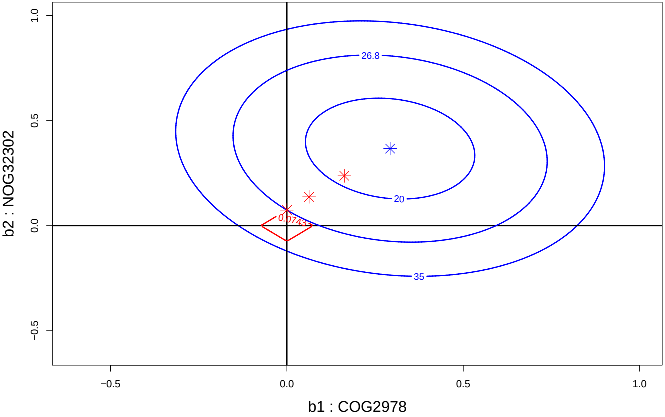
# Insights: 2-dimensional example with the square loss

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



# 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

## A penalized likelihood approach

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

where

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

# The plasmodium data I

```
library(Matrix)
load("plasmodium_expression.Rdata")
dim(Y)
```

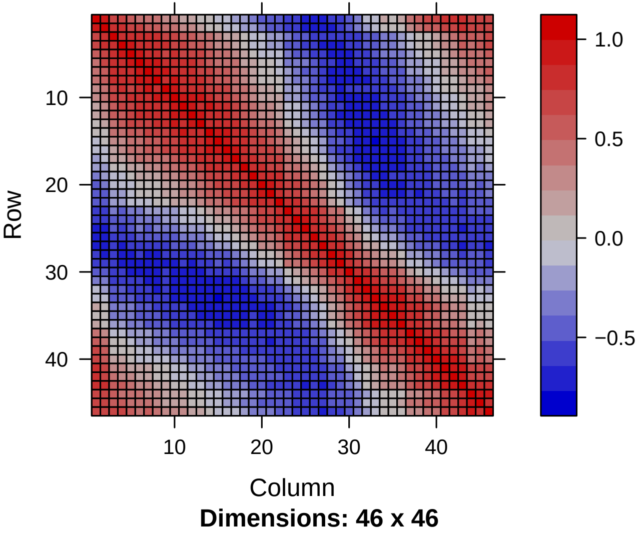
```
## [1] 3490   46
```

```
head(Y)[, 1:5]
```

```
##                 TP1    TP2    TP3    TP4    TP5
## MAL13P1.100 0.4510 0.6532 1.0760 0.5515 0.4238
## MAL13P1.102 1.5320 1.8920 0.8803 1.0300 0.9328
## MAL13P1.103 0.5218 0.5213 0.5328 0.3719 0.3258
## MAL13P1.105 0.5515 0.5527 0.8627 0.4541 0.4299
## MAL13P1.107 0.5630 0.4463 1.0760 0.4035 0.2082
## MAL13P1.112 0.5390 0.5393 0.5642 0.5326 0.4469
```

```
image(Matrix(cor(Y)))
```

# The plasmodium data II



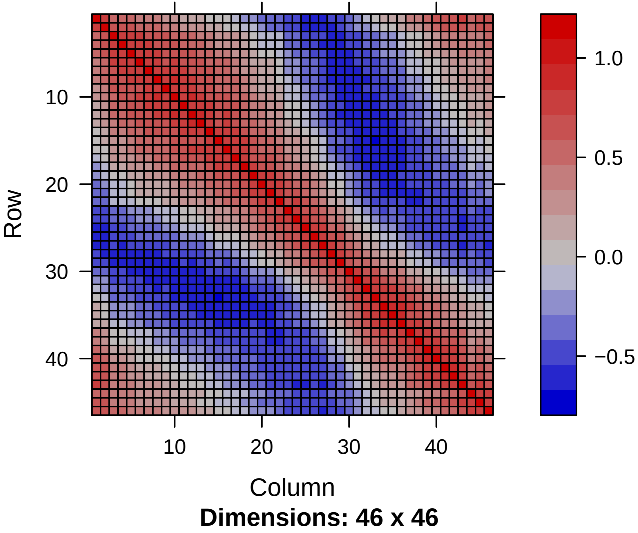
# Covariance structure between the conditions I

## Sparse Estimation

```
library(huge)
huge.out <- huge(as.matrix(Y), method="glasso", cov.output=TRUE)

## Conducting the graphical lasso (glasso) with lossless screening....in progress:0
Conducting the graphical lasso (glasso) with lossless screening....in progress:9%
Conducting the graphical lasso (glasso) with lossless screening....in progress:19%
Conducting the graphical lasso (glasso) with lossless screening....in progress:30%
Conducting the graphical lasso (glasso) with lossless screening....in progress:40%
Conducting the graphical lasso (glasso) with lossless screening....in progress:50%
Conducting the graphical lasso (glasso) with lossless screening....in progress:60%
Conducting the graphical lasso (glasso) with lossless screening....in progress:70%
Conducting the graphical lasso (glasso) with lossless screening....in progress:80%
Conducting the graphical lasso (glasso)....done.

sel.out  <- huge.select(huge.out)

## Conducting extended Bayesian information criterion (ebic) selection....done

image(sel.out$opt.cov)
```

# Covariance structure between the conditions II

## Sparse Estimation



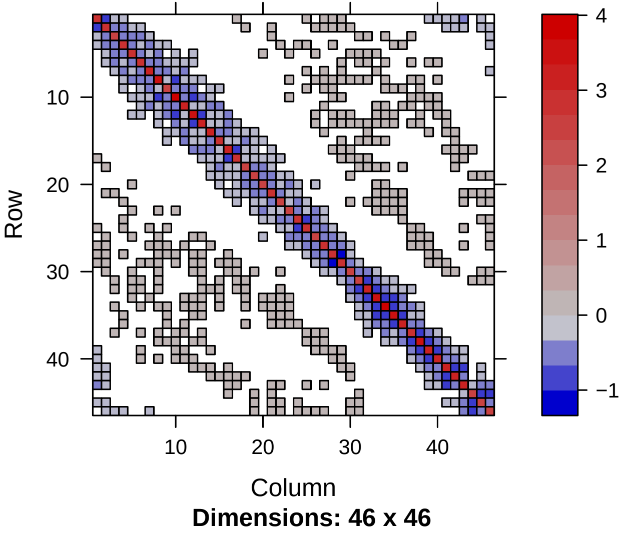
# Covariance structure between the conditions I

## Sparse Estimation of the inverse covariance

```
sum(abs(sel.out$opt.icov) != 0)

## [1] 760

ncol(sel.out$opt.icov) ** 2

## [1] 2116

image(sel.out$opt.icov)
```

# Covariance structure between the conditions II

## Sparse Estimation of the inverse covariance



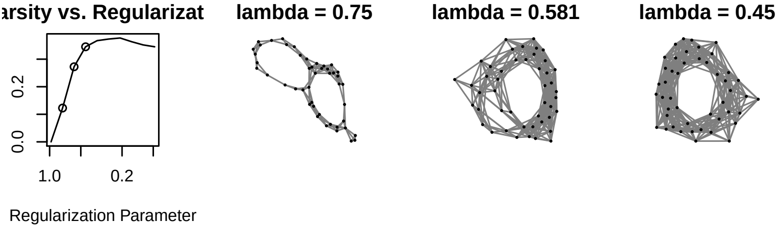
# Covariance structure between the conditions I

## Associated network

```
plot(huge.out)
```

# Covariance structure between the conditions II

## Associated network



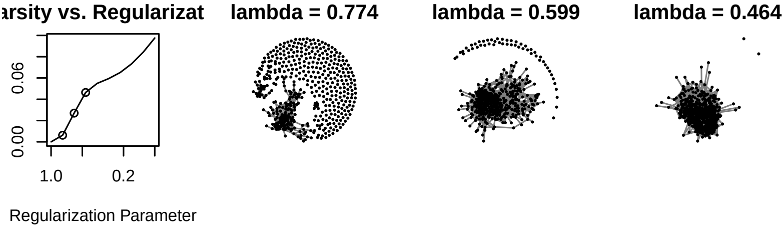
# Network between the genes I

## Sparse Estimation

```
library(huge)
genes.subset <- order(apply(Y,1,var))[1:500]
huge.out <- huge(as.matrix(t(Y[genes.subset, ])), method="glasso", cov.output=TRUE)

## Conducting the graphical lasso (glasso) with lossless screening....in progress:0
Conducting the graphical lasso (glasso) with lossless screening....in progress:9%
Conducting the graphical lasso (glasso) with lossless screening....in progress:19%
Conducting the graphical lasso (glasso) with lossless screening....in progress:30%
Conducting the graphical lasso (glasso) with lossless screening....in progress:40%
Conducting the graphical lasso (glasso) with lossless screening....in progress:50%
Conducting the graphical lasso (glasso) with lossless screening....in progress:60%
Conducting the graphical lasso (glasso) with lossless screening....in progress:70%
Conducting the graphical lasso (glasso) with lossless screening....in progress:80%
Conducting the graphical lasso (glasso)....done.

plot(huge.out)
```

# Network between the genes II

## Sparse Estimation



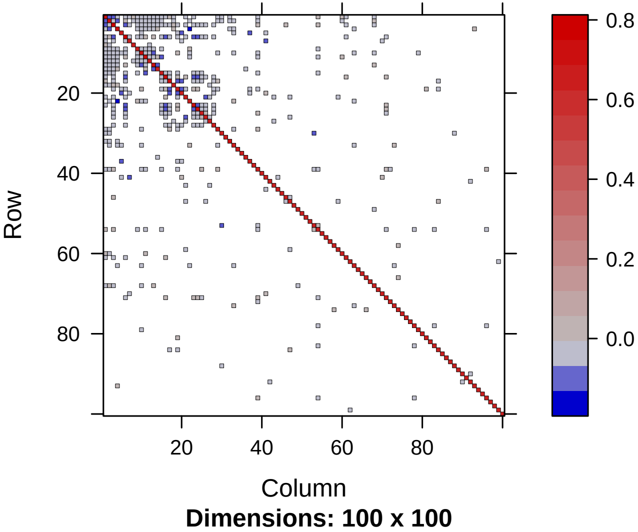
# Network between the genes I

## Inverse covariance

```
library(huge)
huge.out$df

##  [1]     0.0   776.0  3368.0  5790.0  6851.5  7416.5  8128.0  9159.0
##  [9] 10515.0 12172.5

image(Matrix(huge.out$icov[[3]][1:100, 1:100]))
```

# Network between the genes II

## Inverse covariance



# Network between the genes I

## Covariance

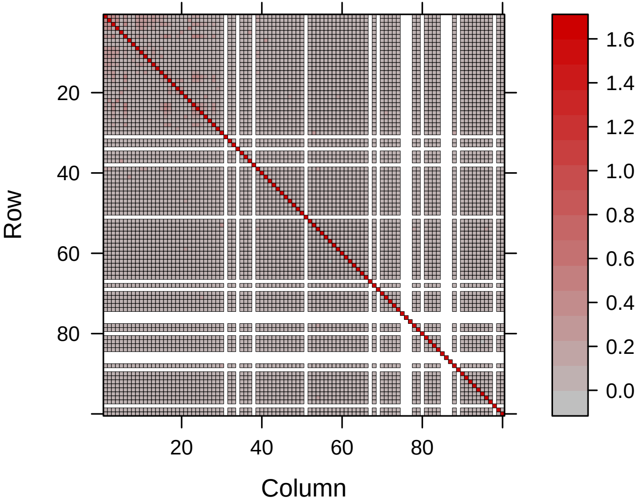
```
library(huge)
huge.out$df
```

```
##  [1]     0.0   776.0  3368.0  5790.0  6851.5  7416.5  8128.0  9159.0
##  [9] 10515.0 12172.5
```

```
image(Matrix(huge.out$cov[[3]][1:100, 1:100]))
```

# Network between the genes II

## Covariance



**Dimensions: 100 x 100**

# Outline

Motivations

Network and data modeling

Network Inference

- Inducing sparsity for edge selection
- Limitations and extensions of sparse GGM

# Practical implications of theoretical results

## Selection consistency (Ravikumar, Wainwright, 2009-2012)

Denote $d = \max_{j\in\mathcal{P}}(\mathrm{degree_j})$. Consistency for an appropriate $\lambda$ 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 $d = \max_{j \in \mathcal{P}}(\mathrm{degree_j})$. Consistency for an appropriate $\lambda$ 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

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

*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.

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

## Resampling/subsampling

Keep edges frequently selected on an range of $\lambda$ after sub-samplings

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

# Limitations towards biological network inference

- Sparse GGM
  - + very solid statistical and computational framework

- DREAM 5 benchmark, 2012 (+ personal experiences).
  - + competitive to other inference methods
  - − performances remain questionable on real data, as for other methods

## Ideas

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)
- accounting for data features (especially NGS)
  - ⇝ extend to non strictly normal distribution
  - ⇝ deal with a very large number of actors

# Network inference for count data

## Data transformation

Consider $\mathbf{X} = (X^1, \dots, X^n)$ some count data with size $n \times p$.

### Simple transformation

Often surprisingly efficients

- log transformation $\log(1 + \mathbf{X})$
- compute $\boldsymbol{S}_n$ by means of Spearman's correlation

### Non paranormal transformation (Liu et al 2009)

The random vector $X$ has non-paranormal distribution if there exist

$$f(X) = f(X_1, \dots, X_p) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Theta}^{-1}).$$

- Distribution of $X$ is a Gaussian copula if $f$ is monotone differentiable
- $X_i \perp\!\!\!\perp X_j | X_{\setminus i,j}$ iff $\boldsymbol{\Theta}_{ij} = 0$.

# Network inference for count data

## Poisson graphical models

### Poisson graphical Lasso (Allen et al, 2012)

Assuming that $X_j | X_k \sim \mathcal{P}(\exp(\beta_j + \sum_{j \neq k} \beta_k X_k))$

$$\hat{\boldsymbol{\beta}} \operatorname*{arg\,min}_{\hat{\boldsymbol{\beta}} \in \mathbb{R}^p} \left\{ - \sum_{i=1}^{n} \sum_{k \neq j} X_{ij} X_{ik} \beta_k - \exp\{X_{ik} \beta_k\} \right\} + \lambda \|\boldsymbol{\beta}\|_1 .$$

- ⇝ Log-linear version of neighborhood selection
- ⇝ Other extensions in Yang et al, 2014 (truncated Poisson).

+ Better performance than GGM...

− ...on simulated Poisson data

− Computationally less efficient

# Dealing with the growing number of feature

### Problem

The number of OTU $p$ may be huge in metagenomics studies

- Statistical limitation (depends on $d, n$)
- Computational limitation (depends on your time but max. 1e6)

### How should we limit the size of the problem?

- Screening (discarding of irrelevant variables)
- Clustering (aggregation of similar actors)

$\rightsquigarrow$ How does this affect the inferred networks?

# Conclusion

## Sparse Gaussian Graphical Model

Well established framework with a vast, growing literature

1. Nice modeling tool (conditional dependencies),
2. Good theoretical framework (which I have not much talked about),
3. Powerful algorithms
   - that scale the dimension (large $p$ large $n$)
   - that allow resampling/parallelization (for robustness)

$\rightsquigarrow$ Great tool for covariance estimation/selection in a reasonably high dimensional settings.

### Still...

- an interaction is not even well defined
- $\rightsquigarrow$ carefull with interpreatation of the networks
- metagenomics data do have some specificities
- $\rightsquigarrow$ adaptation needed