

Surf 64 - XP practice

Biological Network Inference with Sparse Graphical Models

Julien Chiquet

Anglet, 27 June 2018

<http://julien.cremeriefamily.info/exposome.html>

Outline

Statistical analysis of Networks

Different questions

Understanding the network topology

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

Using the network

- Data = a given network + signal on nodes
- Questions: how the epidemic spreads along the network?

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, ...

Automatic reconstruction of biological networks

E. coli regulatory network

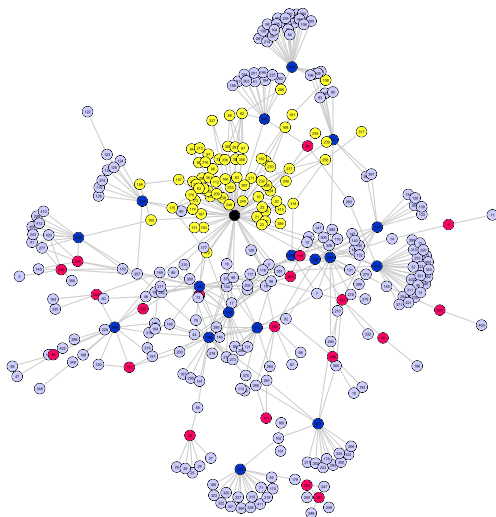
Target network

Relations between genes and their products

- highly structured
- always incomplete

Data and method

- transcriptomic data
- Gaussian graphical model with sparse methods



A challenging problem



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

⇒ 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?

Outline

- 1 Network and data modeling
 - Statistical dependence
 - Gaussian Graphical models
- 2 Network inference with sparse GGM
- 3 A tour of the huge package assessing GGM approach
- 4 Basic network analysis of transcriptomics exposome data set

Outline

- 1 Network and data modeling
 - Statistical dependence
 - Gaussian Graphical models
- 2 Network inference with sparse GGM
- 3 A tour of the huge package assessing GGM approach
- 4 Basic network analysis of transcriptomics exposome data set

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

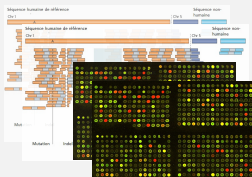
Biological microarrays in comparable conditions

Notations

• a set $\mathcal{P} = \{1, \dots, p\}$ of variables

The data

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



stacked in

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

- 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).

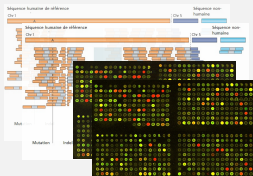
Canonical model settings

Biological microarrays in comparable conditions

Notations

The data

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



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

- 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

sparse Gaussian Graphical Models

- 1 Network and data modeling
 - Statistical dependence
 - Gaussian Graphical models
- 2 Network inference with sparse GGM
- 3 A tour of the huge package assessing GGM approach
- 4 Basic network analysis of transcriptomics exposome data set

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 $A \perp B$. Equivalently,

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

Example (class vs party)

	party			party	
	Labour	Tory		Labour	Tory
class			class		
working	0.42	0.28	working	0.60	0.40
bourgeoisie	0.06	0.24	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".
Estimating¹ $\mathbb{P}(A, B)$, $\mathbb{P}(A)$ and $\mathbb{P}(B)$ in a sample would lead to

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

¹stupidly

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

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

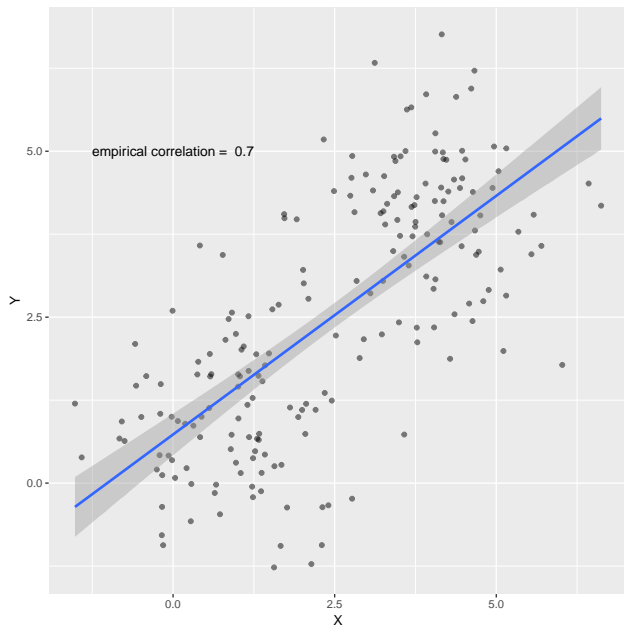
which is usually denoted by $A \perp\!\!\!\perp B|C$

Example (Does QI depends on weight?)

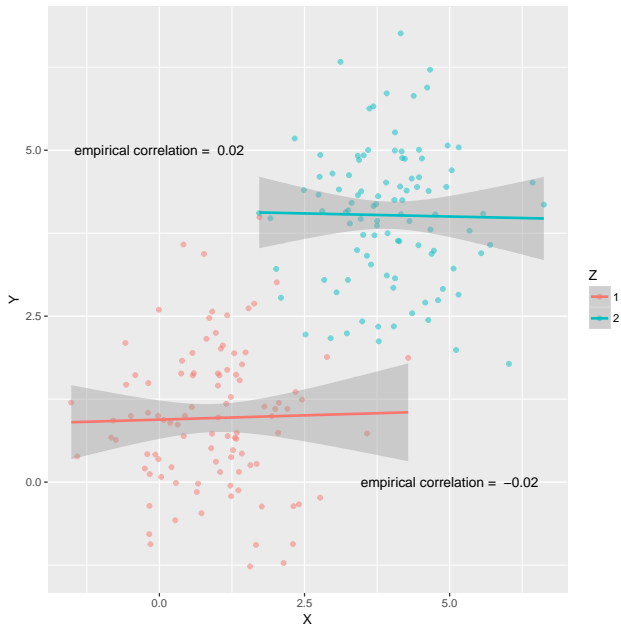
Consider the events A = "having low QI", B = "having low weight". But in fact, introducing C = "having a given age",

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

Limits of correlation for network reconstruction



Limits of correlation for network reconstruction



Outline

sparse Gaussian Graphical Models

- 1 Network and data modeling
 - Statistical dependence
 - Gaussian Graphical models
- 2 Network inference with sparse GGM
- 3 A tour of the huge package assessing GGM approach
- 4 Basic network analysis of transcriptomics exposome data set

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 $X \sim \mathbb{P}$.

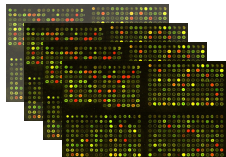
Conditional independence graph

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

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

The Gaussian case

The data



Inference

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

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

Greatly simplifies the inference:

- ↪ naturally links independence and conditional independence to the covariance and partial covariance,
- ↪ 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

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

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

Proposition

- $\text{cov}(X, X) = \mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)],$
- $\text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z),$
- $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\text{cov}(X, Y).$
- $X \perp\!\!\!\perp Y \Rightarrow \text{cov}(X, Y) = 0.$
- $X \perp\!\!\!\perp Y \Leftrightarrow \text{cov}(X, Y) = 0$ *when X, Y are Gaussian.*

The bivariate Gaussian distribution

The Covariance Matrix

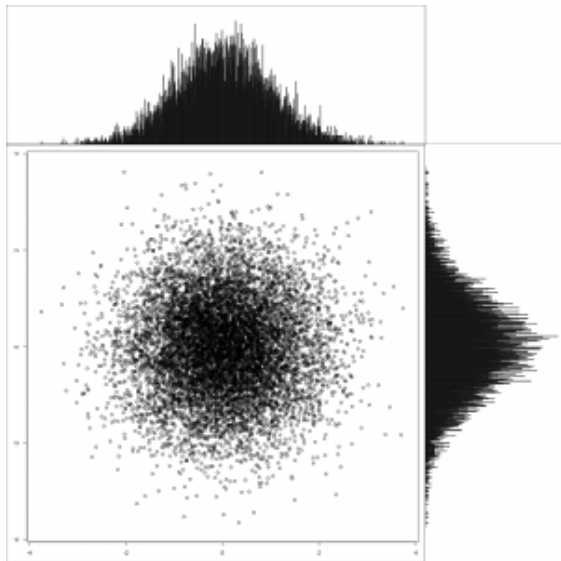
Let

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

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

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

The shape of the 2-D distribution evolves accordingly.



The bivariate Gaussian distribution

The Covariance Matrix

Let

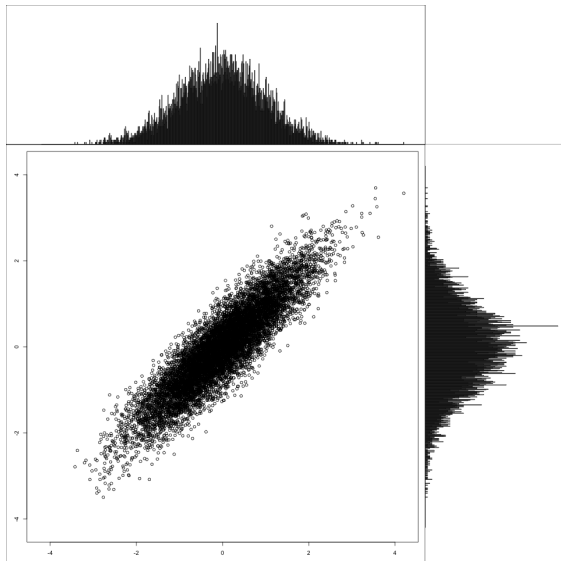
$$X \sim \mathcal{N}(\mathbf{0}, \Sigma),$$

with unit variance and

$$\rho_{XY} = 0.9$$

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

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

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

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

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

Proposition

When X, Y, Z are jointly Gaussian, then

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

Important properties of Gaussian vectors

Proposition (Gaussian vector and conditioning)

Consider a Gaussian vector with the following decomposition

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \Omega = \Sigma^{-1} = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}.$$

Then,

$$Z_2 | Z_1 = \mathbf{z} \sim \mathcal{N}(-\Omega_{22}^{-1} \Omega_{21} \mathbf{z}, \Omega_{22}^{-1})$$

and

$$\Omega_{22}^{-1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}.$$

Corollary

Partial correlations are related to the inverse of the covariance matrix:

$$\text{cor}(Z_i, Z_j | Z_k, k \neq i, j) = -\frac{\Omega_{ij}}{\sqrt{\Omega_{ii} \Omega_{jj}}}$$

Gaussian Graphical Model: canonical settings

Biological experiments in comparable Gaussian conditions

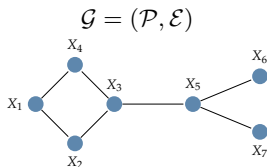
Profiles of a set $\mathcal{P} = \{1, \dots, p\}$ of genes is described by $X \in \mathbb{R}^p$ such as

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

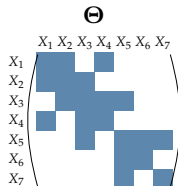
Conditional independence structure

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

Graphical interpretation



\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

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

↪ "Neighborhood" selection

Outline

- 1 Network and data modeling
 - Statistical dependence
 - Gaussian Graphical models
- 2 Network inference with sparse GGM
- 3 A tour of the huge package assessing GGM approach
- 4 Basic network analysis of transcriptomics exposome data set

Some families of methods for network reconstruction

Test-based methods

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

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 $f_X(x; \Theta)$, where Θ are the model parameters.

Maximum likelihood estimator

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

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

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

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 Θ

The multivariate Gaussian log-likelihood

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

The log-likelihood

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

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

Application to GGM: the "Graphical-Lasso"

A penalized likelihood approach

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

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

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

A penalized least-square approach

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

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

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

Model selection

Cross-validation

Optimal in terms of **prediction**, not in terms of selection

Information based criteria

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

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

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

↪ Network inference is a very difficult problem

↪ Some biological questions can be answered without network inference

Outline

- 1 Network and data modeling
 - Statistical dependence
 - Gaussian Graphical models
- 2 Network inference with sparse GGM
- 3 A tour of the huge package assessing GGM approach
- 4 Basic network analysis of transcriptomics exposome data set

Assess the standard GGMs approaches

Full analysis can be found at http://julien.cremeriefamily.info/doc/teachings/exposome/td_exposome_correction.html

```
suppressMessages(library(huge, quietly = TRUE))
```

① Simulated data

- Test that an approach is working under some simple conditions
- Especially usefull when the approach has no underlying model
- Essential sanity check

② Breast cancer data (pinpoint interesting genes/pathways)

- Several hundred breast cancers (estrogen receptor + and -)
- Several thousand genes
- Goal: How can GGMs approaches help ?

Simple simulations (network with hubs)

```
set.seed(11)
n <- 80; d <- 10;
rd.net <- huge.generator(
  n, ## number of samples
  d, ## number of genes
  graph="hub", ## type of net
  g = 2, ## number of group)
verbose=FALSE)
```

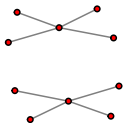
Simple simulations (network with hubs)

```
plot(rd.net)
```

Adjacency Matrix



Covariance Matrix



Empirical Covariance Matrix



Inference using GGMs and correlation

Inference

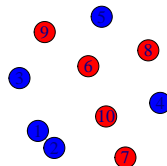
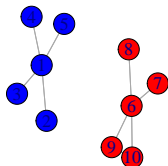
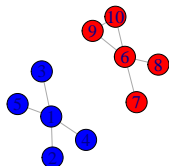
```
## glasso, mb and ct
glasso <- huge(rd.net$data, method="glasso",
              nlambda=50, verbose=F)
mb <- huge(rd.net$data, method="mb",
           nlambda=50, verbose=F)
corthr <- huge(rd.net$data, method="ct",
               nlambda = 50, verbose=F)
```

Selection

```
## glasso, mb and ct
glasso.sel <- huge.select(glasso, "stars", verbose = FALSE)
mb.sel      <- huge.select(mb, "stars", verbose = FALSE)
corthr.sel  <- huge.select(corthr, "stars", verbose = FALSE)
```

Inference using GGMs and correlation (results)

```
gr.glasso <- graph.adjacency(glasso.sel$refit)
V(gr.glasso)$label.cex <- 2
V(gr.glasso)$color <- rep(c("blue", "red"), each=5)
par(mfrow=c(1, 3))
plot(gr.glasso, vertex.size=30, edge.arrow.mode = "-")
plot(gr.mb, vertex.size=30, edge.arrow.mode = "-")
plot(gr.cor, vertex.size=30, edge.arrow.mode = "-")
```



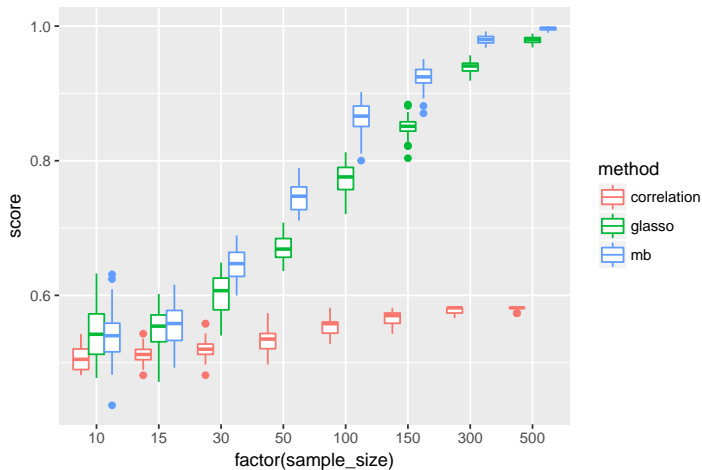
A bit of code to run a simulation

```
library(tidyverse)
one.simu <- function(i) {
  ## simulation settings
  d <- 25; seq.n <- c(10, 15, 30, 50, 100, 150, 300, 500)
  out <- Reduce("rbind", lapply(seq.n, function(n) {
    ## data generation
    exp <- huge.generator(n, d, graph="cluster", g = 3, prob = 1, verbose = F)
    ## inference
    gl <- huge(exp$data, method="glasso", nlambda = 50, verbose = F)
    mb <- huge(exp$data, method="mb", nlambda = 50, verbose = F)
    cthr <- huge(exp$data, method="ct", nlambda = 50, verbose = F)
    ## performance
    res <- setNames(
      c(perf.auc(perf.roc(gl$path, exp$theta)),
        perf.auc(perf.roc(cthr$path, exp$theta)),
        perf.auc(perf.roc(mb$path, exp$theta)), n, i),
      c("glasso", "correlation", "mb", "sample_size", "simu")
    )
    res
  }))) %>% as_data_frame() %>%
  gather(key = "method", value = "score", glasso, correlation, mb)
  out
}
```

Run

```
suppressMessages(library(parallel))  
res <- Reduce('rbind', mclapply(1:40, one.simu, mc.cores = 4))
```

Simulation results (cluster - clique)



Breast cancer: transcriptomics for ER+ and ER- tumors

We look at a large public datasets from Guedj et al. 2011 with two main subgroups

- Estrogen receptor positive
- Estrogen receptor negative

```
load ("huge/breast_cancer_guedj11.RData")
load ("huge/gen_name.RData")
gene.name <- unlist(gene.name)
data.raw <- expr
table(class.ER)
```

```
## class.ER
## ERm ERp
## 162 375
```

Filtering Unknown genes

```
toDiscard <- which(gene.name == "Not.Known")  
gene.name <- gene.name[-toDiscard]  
data.raw <- data.raw[-toDiscard, ]
```

We get

```
dim(data.raw)
```

```
## [1] 41248 537
```

Differential analysis

Do we detect some gene expression differences ?

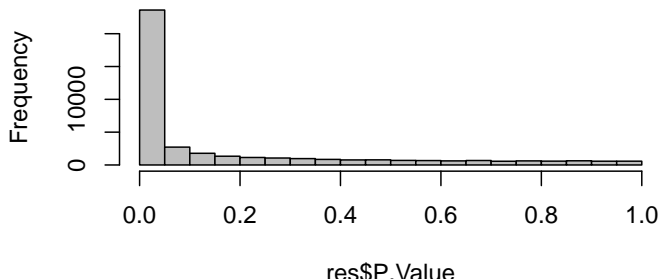
```
load ("huge/breast_cancer_guedj11.RData")
suppressMessages(library(limma))
design <- cbind(Moy=1, Erp=(class.ER == "ERp")+0)
fit <- lmFit(data.raw, design=design)
fit <- eBayes(fit)
res <- topTable(fit, coef="Erp", number=10^5,
               genelist=fit$genes, adjust.method="BH",
               sort.by="none", resort.by=NULL,
               p.value=1, lfc=0, confint=FALSE)
```


Many genes are differentially expressed

- The histogram of p-values looks good
- This is a well known fact (ER+ and ER- are very different)

```
sum(res$adj.P.Val < 10-5)  
  
## [1] 5907  
  
hist(res$P.Value, breaks=30, col="grey",  
      main="P-values ER- vs ER+")
```

P-values ER- vs ER+



What to do with this list of genes?

ESR1 has the most significant p-values

```
gene.name[order(res$adj.P.Val)[1]]
```

```
## 205225_at
```

```
## "ESR1"
```

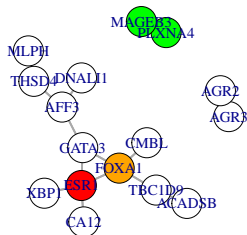
Network analysis

- Could we find partners of ESR1 that are specific to ER+?
- We cannot infer a network on 41000 genes (Verzelen 2011)
 - ~> Most differentially expressed genes
 - ~> Most varying genes
 - ~> Look at a specific pathway ...

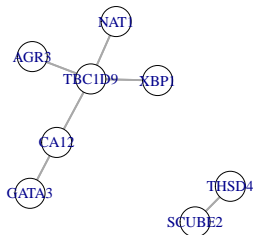
Selecting some probes

Take the 20 most differentially expressed plus some random

ER+ specific



ER- specific



FOXA1, ESR1, GATA3 a well known interaction

- 1 FOXA1 is a key determinant of estrogen receptor function and endocrine response. Antoni Hurtado et al. 2011 (Nat. Genet.):
~> "FOXA1 is a key determinant that can influence differential interactions between ER and chromatin"
- 2 GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. Theodorou et al. 2013 (Genome Res.)
- 3 Estrogen receptor regulation of carbonic anhydrase XII through a distal enhancer in breast cancer. Barnett DH et al 2008 (Cancer Res.)
~> "we show that CA12 is robustly regulated by estrogen via ER alpha in breast cancer cells"

Outline

- 1 Network and data modeling
 - Statistical dependence
 - Gaussian Graphical models
- 2 Network inference with sparse GGM
- 3 A tour of the huge package assessing GGM approach
- 4 Basic network analysis of transcriptomics exposome data set

Tutorial

Let us have a look together at

`http://julien.cremeriefamily.info/doc/teachings/exposome/
transcriptomics_networks_inference.html`