Integrating biological prior knowledge for network inference with Gaussian graphical models.

Marine Jeanmougin
Teams

Statistic & Genome Laboratory
Christophe Ambroise

Michèle, Carène, Claudine, Catherine, Camille, Étienne, Pierre, Gilles, Cécile, Maurice, Marie-Luce, Anne-Sophie, Cyril, Justin, Van-Hanh, Yolande, Sarah, Marius, Bernard et Julien.

Pharnext
Mickaël Guedj

Jan, Caroline, Fabrice, Mickaël, Matthieu, Jonas, Sory.
1. Introduction to transcriptome data
   - Measure of gene expression
   - Microarray technology
   - Statistical methods for microarray analysis

2. Network inference with biological prior
   - Biological prior definition from pathways
   - Network inference

3. Application
   - Breast Cancer data
   - Study of the Estrogen Receptor status
Outline

1. Introduction to transcriptome data
   - Measure of gene expression
   - Microarray technology
   - Statistical methods for microarray analysis

2. Network inference with biological prior
   - Biological prior definition from pathways
   - Network inference

3. Application
   - Breast Cancer data
   - Study of the Estrogen Receptor status
How to measure the gene activity?

Measure of the transcription level

DNA → RNA → Protein

Use RNA measurements as a proxy for gene activity
If all cells contain the same DNA information, how can they have different functions?

The gene expression levels differ among:

- tissues
- developmental stages
- external conditions

Regulation of gene activity

- at different levels (chromatin, DNA, RNA)
- through various mechanisms: transcription factors, miRNAs, splicing factors
If all cells contain the same DNA information, how can they have different functions?

The gene expression levels differ among:

- tissues
- developmental stages
- external conditions

Regulation of gene activity

- at different levels (chromatin, DNA, RNA)
- through various mechanisms: transcription factors, miRNAs, splicing factors
Introduction

If all cells contain the same DNA information, how can they have different functions?

The gene expression levels differ among:

▶ tissues
▶ developmental stages
▶ external conditions

Regulation of gene activity

▶ at different levels (chromatin, DNA, RNA)
▶ through various mechanisms: transcription factors, miRNAs, splicing factors
Microarray technology

Expression measurement through microarray
Microarray technology

$n \approx 10s/100s$ of microarrays
$p \approx 1000s$ of genes
“large $p$, small $n$”

| sample 1 | $x_1^1$ | ... | $x_1^p$ |
| sample i | $x_i^1$ | ... | $x_i^p$ |
| sample n | $x_n^1$ | ... | $x_n^p$ |

The main statistical issue is the **high dimensional setting**.
Objectives of microarray experiments

Expression level of thousands of transcripts

Biological purpose

- Identify genes involved in a phenotype of interest
- Medical applications: diagnosis, prognosis, treatment efficacy

Methodology

- Hypothesis testing: select genes differentially expressed
Differential analysis

Data

- $X_{ig}^{(c)}$: **expression level** of the $i$th sample for gene $g$ under condition $c$;
- $Y$: **output random variable**, either Gaussian or Bernoulli.

Model

Under the assumption of homoscedasticity (i.e. homogeneity of variance) between conditions:

$$
\mathbb{E}(X_{ig}^{(c)}) = \mu_{g}^{(c)} \quad \text{and} \quad \mathbb{V}(X_{ig}^{(c)}) = (\sigma_{g})^{2},
$$

where $\sigma_{g}^{2}$ is the variance of gene $g$ expression levels, constant across conditions.
Hypothesis testing strategy

For two conditions, the null hypothesis to test comes down to

\[
\begin{align*}
H_{0,g} & : \mu_g^{(1)} = \mu_g^{(2)} \\
H_{1,g} & : \mu_g^{(1)} \neq \mu_g^{(2)}
\end{align*}
\]

Test statistic

Classical approach: \( t \)-statistic:

\[
t_g = \frac{\bar{x}_g^{(1)} - \bar{x}_g^{(2)}}{S_g^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]

with,

- \( \bar{x}_g^{(1)} \): natural estimator of \( \mu_g^{(1)} \)
- \( S_g^2 \): usual unbiased estimator of the variance \( (\sigma_g)^2 \)

Issues for gene-specific variance estimation
Differential analysis

**Limma: a shrinkage approach (Smyth, 2004)**

**Variance estimate**

- Scale-inverse-chi-square prior distribution of parameters $S^2_0$ and $d_0$
- Posterior variance

\[
S_{g \text{limma}} = \frac{d_0 S^2_0 + d_g S^2_g}{d_0 + d_g},
\]

- $d_g$: residual degrees of freedom for gene $g$
- Bayesian estimation of $S^2_0$: usually a little less than the mean of the $S_g$

\[\sim\text{Robust estimate of the gene-specific variances}\]

**Test statistic**

\[
t_{g \text{limma}} = \frac{\bar{x}^{(1)} - \bar{x}^{(2)}}{S_{g \text{limma}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.
\]
Problematic

Microarray data

Differential analysis

Signature 1 . . . . . . . . . . s

How to understand underlying biological phenomena?
Gene regulatory network inference

Definition

Graphical representation of the conditional dependencies between RNA measurements

Issue

Which gene interactions among the $p(p-1)/2$ possible are biological regulations?
Handling the scarcity of data (1)

- Few genes effectively interact (sparsity)
- Estimation of a coefficient matrix with mostly zero-valued entries
Handling the scarcity of data (1)

- Few genes effectively interact (sparsity)
- Estimation of a coefficient matrix with mostly zero-valued entries
Network inference

Handling the scarcity of data (1)

- Few genes effectively interact (sparsity)

- Estimation of a coefficient matrix with mostly zero-valued entries
Network inference

Handling the scarcity of data (2)

Inference strategy

(1) Inference of each network **independently**

**Condition 1**

\[(X_1^{(1)}, \ldots, X_{n_1}^{(1)}), X_i^{(1)} \in \mathbb{R}^{p_1}\]

**Condition 2**

\[(X_1^{(2)}, \ldots, X_{n_2}^{(2)}), X_i^{(2)} \in \mathbb{R}^{p_2}\]

**Condition 3**

\[(X_1^{(3)}, \ldots, X_{n_3}^{(3)}), X_i^{(3)} \in \mathbb{R}^{p_3}\]
Network inference

Handling the scarcity of data (2)

Inference strategy

(2) pooling all the available data

Condition 1

Condition 2

Condition 3

\[(X_1, \ldots, X_n), X_i \in \mathbb{R}^P, \text{ with } n = n_1 + n_2 + n_3.\]
Network inference

Handling the scarcity of data (2)

Inference strategy

(3) breaking the separability

Condition 1

\((X^{(1)}_1, \ldots, X^{(1)}_{n_1}), X^{(1)}_i \in \mathbb{R}^{p_1}\)

Condition 2

\((X^{(2)}_1, \ldots, X^{(2)}_{n_2}), X^{(2)}_i \in \mathbb{R}^{p_2}\)

Condition 3

\((X^{(3)}_1, \ldots, X^{(3)}_{n_3}), X^{(3)}_i \in \mathbb{R}^{p_3}\)
Network inference

Handling the scarcity of data (3)

- A vast space of possible network structures
- Technical and biological noise in microarray data

Biological prior knowledge could be used to limit the set of candidate networks
1. Introduction to transcriptome data
   - Measure of gene expression
   - Microarray technology
   - Statistical methods for microarray analysis

2. Network inference with biological prior
   - Biological prior definition from pathways
   - Network inference

3. Application
   - Breast Cancer data
   - Study of the Estrogen Receptor status
Pathways

Sets of gene products interacting in order to achieve a specific cellular function: metabolic pathways, signaling pathways...
Pathway Analysis

**Biological purpose:** Which are the pathways targeted by the set of differentially expressed genes?

**Idea:** Testing whether the signature is enriched in pathway key actors → over-representation test
Model

- Let $S$ be the subset of differentially expressed genes (i.e. the signature).
- Say $S$ is of size $s$.
- Let $K = K_1, \ldots, K_m$ be a set of pre-defined pathways.
- For all pathway $K_l$ of size $t_l$, we define:
  \[ Y_l = \{\# \text{ genes } \in S \cap K_l\}, \]
  such as $Y_l \sim H(s, p, t_l)$, with
  \[
  \mathbb{P}(Y_l = y) = \binom{s}{y} \binom{p-s}{t_l-y} \binom{p}{t_l}.
  \]
Hypothesis testing strategy

Under the null hypothesis of no over-representation:

\[
\mathbb{P}(Y_l \geq y) = 1 - \mathbb{P}(Y_l \leq y) \\
= 1 - \sum_{i=0}^{y} \frac{(s)_i}{(p-s)_{t_l-i}} \frac{(p-s)_{t_l-i}}{(p)_{t_l}}.
\]

- \text{p-value of a one-sided test}
- Significance level of the test set at 5%
Pathway Analysis in practice...

<table>
<thead>
<tr>
<th>Pathway Name</th>
<th>Genes in pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>HER-2 Signaling in Breast Cancer</td>
<td>CCNE1, CDK6, PARD6B, ERBB3, EGFR</td>
</tr>
<tr>
<td>Glioblastoma Multiforme Signalizing</td>
<td>CCNE1, RHOB, IGF1R, CDK6, EGFR</td>
</tr>
<tr>
<td>Estrogen-Dependent Breast Cancer Signalizing</td>
<td>IGF1R, ESR1, EGFR</td>
</tr>
<tr>
<td>Small Cell Lung Cancer Signalizing</td>
<td>CCNE1, CDK6, BCL2</td>
</tr>
<tr>
<td>Aryl Hydrocarbon Receptor Signalizing</td>
<td>CCNE1, TFF1, CDK6, ESR1</td>
</tr>
</tbody>
</table>

Pathways do not clearly represent distinct entities!

We need to summarize the set of pathways found significant.
Pathway Analysis in practice...

<table>
<thead>
<tr>
<th>Pathway Name</th>
<th>Genes in pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>HER-2 Signaling in Breast Cancer</td>
<td>CCNE1, CDK6, PARD6B, ERBB3, EGFR</td>
</tr>
<tr>
<td>Glioblastoma Multiforme Signaling</td>
<td>CCNE1, RHOB, IGF1R, CDK6, EGFR</td>
</tr>
<tr>
<td>Estrogen-Dependent Breast Cancer Signaling</td>
<td>IGF1R, ESR1, EGFR</td>
</tr>
<tr>
<td>Small Cell Lung Cancer Signaling</td>
<td>CCNE1, CDK6, BCL2</td>
</tr>
<tr>
<td>Aryl Hydrocarbon Receptor Signaling</td>
<td>CCNE1, TFF1, CDK6, ESR1</td>
</tr>
</tbody>
</table>

Pathways do not clearly represent distinct entities!

We need to summarize the set of pathways found significant
### Prior definition

<table>
<thead>
<tr>
<th>Pathway Name</th>
<th>Genes in pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>HER-2 Signaling in Breast Cancer</td>
<td>CCNE1, CDK6, PARD6B, ERBB3, EGFR</td>
</tr>
<tr>
<td>Glioblastoma Multiforme Signaling</td>
<td>CCNE1, RHOB, IGFR1, CDK6, EGFR</td>
</tr>
<tr>
<td>Estrogen-Dependent Breast Cancer Signaling</td>
<td>IGFR1, ESR1, EGFR</td>
</tr>
<tr>
<td>Small Cell Lung Cancer Signaling</td>
<td>CCNE1, CDK6, BCL2</td>
</tr>
<tr>
<td>Aryl Hydrocarbon Receptor Signaling</td>
<td>CCNE1, TFF1, CDK6, ESR1</td>
</tr>
</tbody>
</table>

**Pathway analysis results**

- **Binary matrix**
- **Jaccard distance**
- **Distance matrix**
- **Ward's criterion**
- **Core pathways**
Prior definition - Summary

Microarray data

Differential analysis

Signature

1 . . . . . . . . . .s

Pathway analysis

Core pathways
Gaussian Graphical Model (GGM)

Definition

Probabilistic framework representing the conditional dependency structure between random variables on a graph.

Gaussian model for an i.i.d. sample

- Let $\mathcal{P} = \{1, \ldots, p\}$ be a set of nodes (i.e. genes).
- $X = (X_1, \ldots, X_p)$, such as $X \sim \mathcal{N}(0_p, \Sigma)$ is the signal over this set (i.e. the gene expression levels).
- Let $\Theta$ be the parameter to be inferred.
  - $\Theta = (\theta_{ij})_{i,j \in \mathcal{P}} \triangleq \Sigma^{-1}$ is the concentration matrix.
  - $\text{cor}_{ij|\mathcal{P}\backslash\{i,j\}} = -\theta_{ij}/\sqrt{\theta_{ii}\theta_{jj}}$ for $i \neq j$ (Dempster, 1972)
Definition

Probabilistic framework representing the conditional dependency structure between random variables on a graph.

Gaussian model for an i.i.d. sample

- Let $\mathcal{P} = \{1, ..., p\}$ be a set of nodes (i.e. genes).
- $X = (X_1, ..., X_p)$, such as $X \sim \mathcal{N}(\mathbf{0}_p, \Sigma)$ is the signal over this set (i.e. the gene expression levels).
- Let $\Theta$ be the parameter to be inferred.
  - $\Theta = (\theta_{ij})_{i,j \in \mathcal{P}} \triangleq \Sigma^{-1}$ is the concentration matrix.
  - $\text{cor}_{ij|\mathcal{P}\setminus\{i,j\}} = -\theta_{ij}/\sqrt{\theta_{ii}\theta_{jj}}$ for $i \neq j$ (Dempster, 1972)
Interpretation

If 2 nodes $i$ and $j$ are partially uncorrelated, no edge is inferred:

$$X_i \perp \perp X_j \mid X(P\setminus\{i, j\}) \iff \theta_{i,j} = 0$$

After a simple rescaling, $\Theta$ can be interpreted as the adjacency matrix of the graph of conditional dependencies.

conditional dependency

if and only if

or

non null partial correlation
The log-likelihood

\[ \mathcal{L}(\Theta; S) = \frac{n}{2} \log \det(\Theta) - \frac{n}{2} \text{Trace}(S\Theta) + \frac{n}{2} \log(2\pi), \]

with \( S = n^{-1}X^TX \) be the empirical variance-covariance matrix.

- For \( n \geq p \) the MLE is given by \( \hat{\Theta}^{\text{MLE}} = S^{-1} \)
- The MLE is not defined for \( n < p \).

Need for regularization!
Gaussian Graphical Model (GGM)

Regularized approach

\[ \hat{\Theta}_\lambda = \arg \max_{\Theta} \mathcal{L}(\Theta; S) - \lambda \|\Theta\|_{\ell_1}, \]

with \( \lambda > 0 \), tunes the overall amount of penalty.

Banerjee et al. 2008, JMLR

- Regularization is enforced by an \( \ell_1 \) constraint on the coefficient vector.
- \( \ell_1 \) norm drives some coefficients to zero (sparsity)
Figure: Contour lines of residual sum of squares, with $\beta^{ls}$ being the least squares estimator, $\ell_2$-ball corresponding to the Ridge regression (right) and $\ell_1$-ball corresponding to the Lasso problem (left).

M. Szafranski, *Pénalités hiérarchiques pour l’intégration de connaissances dans les modèles statistiques*
Gaussian Graphical Model (GGM)

Multi-task approach

Condition 1

\[(X_1^{(1)}, \ldots, X_{n_1}^{(1)}), X_i^{(1)} \in \mathbb{R}^{p_1}\]

Condition 2

\[(X_1^{(2)}, \ldots, X_{n_2}^{(2)}), X_i^{(2)} \in \mathbb{R}^{p_2}\]

Condition 3

\[(X_1^{(3)}, \ldots, X_{n_3}^{(3)}), X_i^{(3)} \in \mathbb{R}^{p_3}\]

▶ Assumption: strong relationship between both networks
▶ Joint estimation of the graphs by coupling the estimation problems
Gaussian Graphical Model (GGM)

Breaking the separability by modifying the regularization term

\[ \max_{\Theta^{(c)}, c, \ldots, C} \sum_{c=1}^{C} \mathcal{L} \left( \Theta^{(c)}; S^{(c)} \right) - \lambda \sum_{i, j \in P, i \neq j} \left( \sum_{c=1}^{C} \left( \theta_{ij}^{(c)} \right)^2 \right)^{1/2}. \]

- Mixed \( \ell_1/\ell_2 \) norm

- \( \ell_2 \) norm: not a single zero can belong to a group with non-zeros
  \( \Rightarrow \) Enforces all graphs to share exactly the same set of edges

Grandvalet and Canu 1998, NIPS
Coop-LASSO

\[
\max_{\Theta^{(c)}, \ldots, \Theta^{(C)}} \sum_{c=1}^{C} \mathcal{L} \left( S^{(c)} ; \Theta^{(c)} \right) \\
- \lambda \sum_{i,j \in \mathcal{P}, i \neq j} \left\{ \left( \sum_{c=1}^{C} \left[ \theta_{ij}^{(c)} \right]_{+} \right)^{1/2} + \left( \sum_{c=1}^{C} \left[ \theta_{ij}^{(c)} \right]_{-} \right)^{1/2} \right\},
\]

where \([u]_+ = \max(0, u)\) and \([u]_- = \min(0, u)\).

- Group-lasso like penalty
- Disconnect the activation of up and down regulation

Chiquet et al. 2010, Statistics and Computing
Gaussian Graphical Model (GGM)

Integration of biological prior knowledge

- Let $\mathcal{Q} = \{1, \ldots, Q\}$ be the set of overlapping core-pathways.
- Let $\mathbf{Z}$ be the matrix of graph structure such as

$$
\mathbf{Z} = (Z_{iq})_{i \in \mathcal{P}, q \in \mathcal{Q}},
$$

with $Z_{iq} = 1$ if $i \in q$ and 0 otherwise.

$$
\max_{\Theta^{(c)}} \sum_{c=1}^{C} \mathcal{L} \left( \Theta^{(c)} ; \text{data} \right) - \lambda \sum_{i,j \in \mathcal{P} \atop i \neq j} \rho_{Z_i Z_j} \left\{ \left( \sum_{c=1}^{C} \left[ \theta_{ij}^{(c)} \right]^2 \right)^{1/2} + \left( \sum_{c=1}^{C} \left[ \theta_{ij}^{(c)} \right]^2 \right)^{1/2} \right\},
$$

$$
\rho_{Z_i Z_j} = \begin{cases} 
\sum_{q, \ell \in \mathcal{Q}} Z_{iq} Z_{j\ell} \frac{1}{\lambda_{\text{in}}}, & \text{if } i \neq j, \text{ and } q = \ell, \\
\sum_{q, \ell \in \mathcal{Q}} Z_{iq} Z_{j\ell} \frac{1}{\lambda_{\text{out}}}, & \text{if } i \neq j, \text{ and } q \neq \ell, \\
1, & \text{otherwise.}
\end{cases}
$$
Summary

Microarray data → Signature

Differential analysis → Pathway analysis → Network inference

Core pathways → Regulation network
Introduction to transcriptome data
- Measure of gene expression
- Microarray technology
- Statistical methods for microarray analysis

Network inference with biological prior
- Biological prior definition from pathways
- Network inference

Application
- Breast Cancer data
- Study of the Estrogen Receptor status
Breast cancer in a few words

- An heterogeneous disease (5 subtypes)

Sorlie et al. 2003, PNAS

- Presence (ER+)/absence (ER-) of estrogen receptors: an essential parameter of tumor characterization.

Understanding the molecular mechanism of ER status: a key issue for treatment and prognosis
**Method**

Inference of regulation networks under ER+ and ER- conditions

**Goal**

Comparison of *regulation patterns*
Results
ER status in breast cancer: pathway analysis

Cellular growth & proliferation

Apoptosis

Aryl Hydrocarbon Receptor Signaling
P53 Pathway
Small Cell Lung Cancer
Molecular Mechanisms of Cancer
Sphingolipid Metabolism
Valine Leucine and Isoleucine Degradation
Progesterone Mediated Oocyte Maturation
Oocyte Meiosis
CDK5 Signaling
HER-2 Signaling in Breast Cancer
Endocytosis
Calcium Signaling Pathway
ERBB Signaling Pathway
Tel Pathway
Bad Pathway
Prostate Cancer
Pathways in Cancer
Glioma
Glioblastoma Multiforme Signaling
Estrogen-Dependent Breast Cancer Signaling
RacCycD Pathway
Erk Pathway
Cell Cycle
G1 Pathway

Cell death

Protein trafficking

Small molecules biochemistry

Figure: Core pathways
Figure: **Sub-network inferred from the ER status signature** - Common edges of both ER+ and ER- networks are symbolized in gray. Dashed black edges: inferred only under the ER- condition. Solid black edges: only predicted under the ER+ condition.
Results
ER status in breast cancer: network inference

Anti-apoptotic mechanisms

Common regulations
- Estrogen receptor (ESR1) - BCL2 (Peterson et al. 2007)
- ESR1 - EGFR/IGF1R (Salvatori et al. 2000, Oesterreich et al. 2001)

Specific regulations
- EGF receptor family: ERBB3 - ERBB4 (Lee et al. 2001)
- CDK6 - IGF1R
Results
ER status in breast cancer: summary

Extracellular space
Plasma membrane
Cytoplasm
Nucleus

Growth Hormone
IGF-1

ERBB4
ERBB3
IGF1R
EGFR
ESR1
BCL2
Apoptosis

MAPT
CDK6
B

- activation
- repression
- binding
- ER+ specific regulation
- Kinase
- Other
- Transmembrane receptor
- Ligand-dependent nuclear receptor

Figure: Anti-apoptotic mechanisms
Summary

▶ Network inference: a very challenging issue
▶ Introducing **structured regularizers** and **biological priors** can help bring robustness to the inference
▶ Promising application on Breast cancer dataset

Perspectives

▶ Importance of missing covariates: need for integration of various omics data.
▶ Statistical validation of differences between graphs
Acknowledgments

J. Chiquet  
C. Charbonnier  
M. Guedj  
C. Ambroise
Chromatine

**Definition**
structure complexe constituée d’ADN et de protéines, localisé le noyau des cellules
Epissage

**Definition**
mécanismes de délétion des introns ou des exons qui ne seront pas nécessaires au codage de la protéine

![Diagram of Epissage](image)

- **Elimination simple des introns**:

- **Epissage alternatif**:
Core-pathways

- Binary matrix:

\[ M = (m_{u,v})_{1 \leq u \leq s, 1 \leq v \leq k}, \]

where \( s \) is the length of the signature and \( k \) the number of significant pathways, such that:

\[ m_{u,v} = \begin{cases} 
1 & \text{if the gene } u \text{ belongs to the pathway } v, \\
0 & \text{otherwise.}
\end{cases} \]

- Dissimilarity between 2 pathways (denoted \( v_1 \) and \( v_2 \) in the following) is assessed by using a binary metric (Jaccard distance):

\[ J_\delta = 1 - \frac{\sum_{u=1}^{y} \mathbb{I}\{m_{u,v_1}=1, m_{u,v_2}=1\}}{y - \sum_{u=1}^{y} \mathbb{I}\{m_{u,v_1}=0, m_{u,v_2}=0\}} \]

Percentage of genes that belong exclusively to either one of the two pathways of interest.

- Finally, from this dissimilarity matrix we perform a hierarchical agglomerative clustering using Ward’s criterion.
Ward’s criterion

- Agglomerative criterion
- Minimizes the increasing of within inertia
- Merge the two clusters $A$ and $B$ which are the closest with respect to the Ward’s criterion:

$$\Delta_{\text{ward}}(A, B) = \frac{I_AI_B}{I_A + I_B} d^2(\mu_A, \mu_B)$$

with $d$ the Euclidean distance, $\mu_A$ the barycentre and $I_A$ the cardinality of the set $A$. 
### Table: Outcome when testing \( n \) hypotheses.

<table>
<thead>
<tr>
<th>Reality \ Decision</th>
<th>( H_0 ) not rejected</th>
<th>( H_0 ) rejected</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 ) true</td>
<td>TN</td>
<td>FP</td>
<td>( n_0 )</td>
</tr>
<tr>
<td>( H_0 ) false</td>
<td>FN</td>
<td>TP</td>
<td>( n_1 )</td>
</tr>
<tr>
<td>total</td>
<td>( n_U = TN + FN )</td>
<td>( n_R = FP + TP )</td>
<td>( n )</td>
</tr>
</tbody>
</table>

Assuming that \( n_0 \) of the null hypotheses are true such as \( n = n_0 \):

\[
FP \sim \mathcal{B}(n, \alpha)
\]

with, \( \mathbb{E}(FP) = n\alpha \) and:

\[
\mathbb{P}_{H_0}(FP \geq 1) = 1 - \mathbb{P}_{H_0}(FP = 0) = 1 - \binom{n}{0} \alpha^0 (1 - \alpha)^n = 1 - (1 - \alpha)^n,
\]

For \( n = 10 \), \( \mathbb{P}(FP \geq 1) = 40\% \).
\[ \text{FDR}(\alpha) = \mathbb{E} \left( \frac{FP}{n_R} \mathbb{I}\{n_R > 0\} \right), \]
such as \( \mathbb{I}\{n_R > 0\} = 1 \) if if \( n_R = FP + TP \) is non-null and gives 0 otherwise.

**step-up procedure**

- Let \( p_1 \leq \ldots \leq p_n \): ordered p-values of \( n \) independent tests.
- Goal: identify the maximal threshold \( i^* \) such as the null hypothesis is rejected
- If the procedure stops at the threshold \( p_i \), then \( n_R = rg(p_i) = i \), such as \( rg(p_i) \) is the ranks of the \( i \)th p-value and:

\[
\mathbb{E}(FP(p_i)) = n_0 \times p_i = n\pi_0 \times p_i.
\]

The strategy proposed by Benjamini and Hochberg is to use the fact that \( \pi_0 \leq 1 \) to provide an upper bound of the FDR:

\[
\text{FDR}(p_i) \leq \frac{n \times p_i}{i}.
\]
Partial correlation

Two genes $g$ and $h$ are linked by an undirected edge $(g, h)$ if, conditional on all other gene expressions indexed by $G \setminus \{g, h\}$, random variables $X_g$ and $X_h$ are dependent:

$$r_{gh|G \setminus \{g, h\}} = \frac{\text{cov}(X_g, X_h|X_{G \setminus \{g, h\}})}{\sqrt{\text{var}(X_g|X_{G \setminus \{g, h\}}) \text{var}(X_h|X_{G \setminus \{g, h\}})}},$$

with,

$$\text{cov}(X_g, X_h|X_z) = \mathbb{E}(X_gX_h|X_z) - \mathbb{E}(X_g|X_z)\mathbb{E}(X_h|X_z).$$
Norms

For every size-$p$ vector $u$, we denote respectively by $\|u\|_{\ell_0}$, $\|u\|_{\ell_1}$ and $\|u\|_{\ell_2}$ its $\ell_0$, $\ell_1$ and $\ell_2$-norms, defined as follows:

$$
\|u\|_{\ell_0} = \sum_{k=1}^{p} \mathbb{I}\{u_k \neq 0\}, \quad \|u\|_{\ell_1} = \sum_{k=1}^{p} |u_k|, \quad \|u\|_{\ell_2} = \sqrt{\sum_{k=1}^{p} u_k^2}.
$$

**Figure 3.2** – Comparaisons des solutions de problèmes régularisés par une norme $\ell_1$ et $\ell_2$. 
\[ \|u\|_{p,q} = \left( \sum_i \left( \sum_j |u_{i,j}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} \]