Graphical Modeling

Christophe Giraud

Université Paris-Sud and Ecole Polytechnique

Maths Department

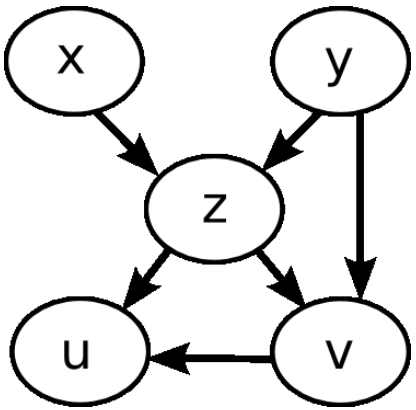
Lecture Notes on High-Dimensional Statistics :

http://www.cmap.polytechnique.fr/~giraud/MSV/LectureNotes.pdf

Please ask questions!

Topic of the lecture

Graphical models :



- Graphical modeling is a convenient representation of conditional dependences among random variables
- It is a powerful tool for
  - exploring "direct effect" between variables
  - fast computations in complex models
- It is popular in many different fields, including bioinformatics, computer vision, speech recognition, environmental statistics, economics, social sciences, etc

Warning

Two important topics :

- learning graphical models
- learning with graphical models

Warning

Two important topics :

- learning graphical models
- ~~learning with graphical models~~

Example 1

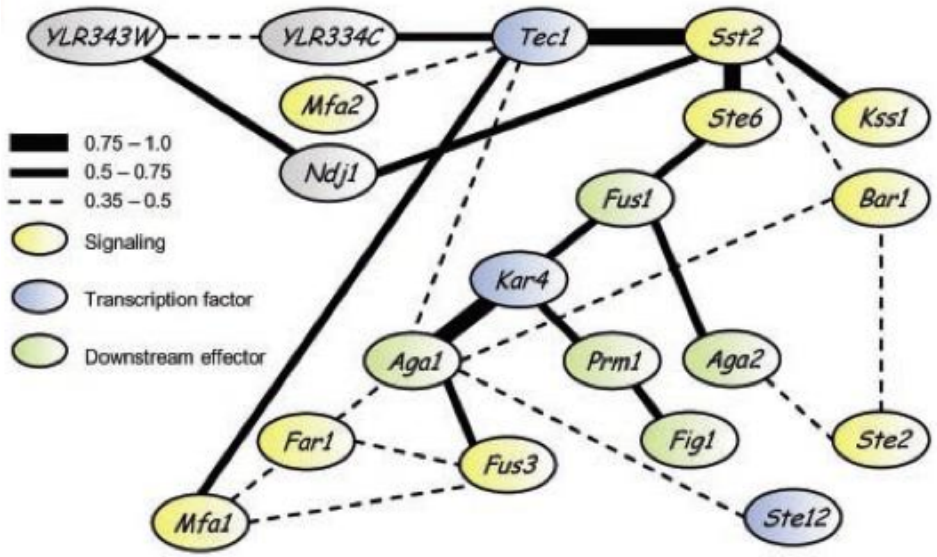


Figure: Learning Biological Regulatory Networks

Seminal reference : Friedman [6]

Example 1

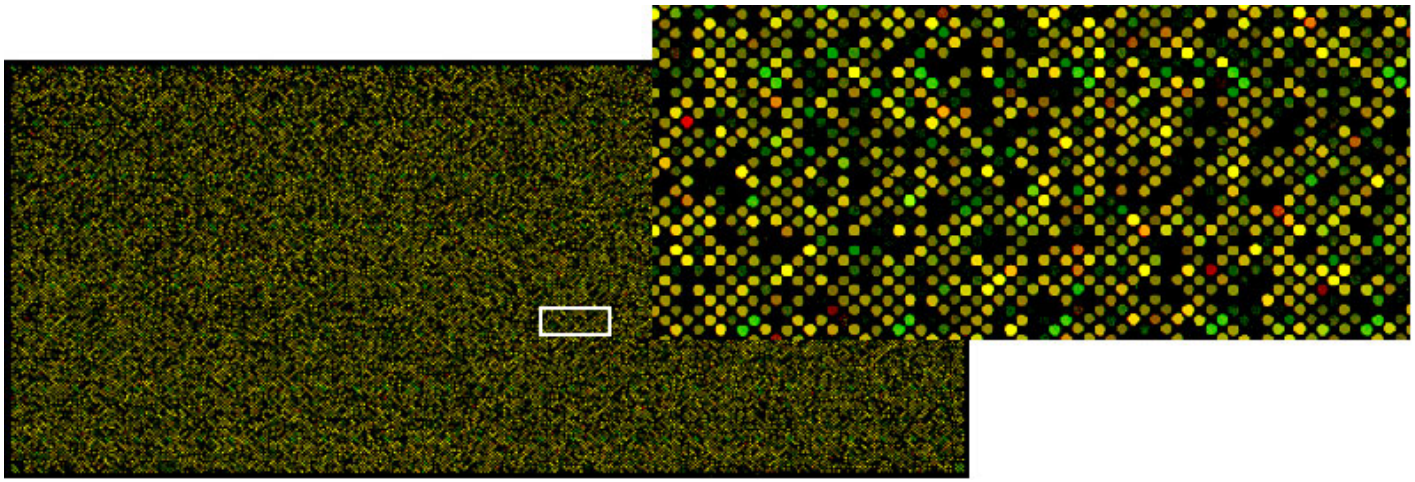
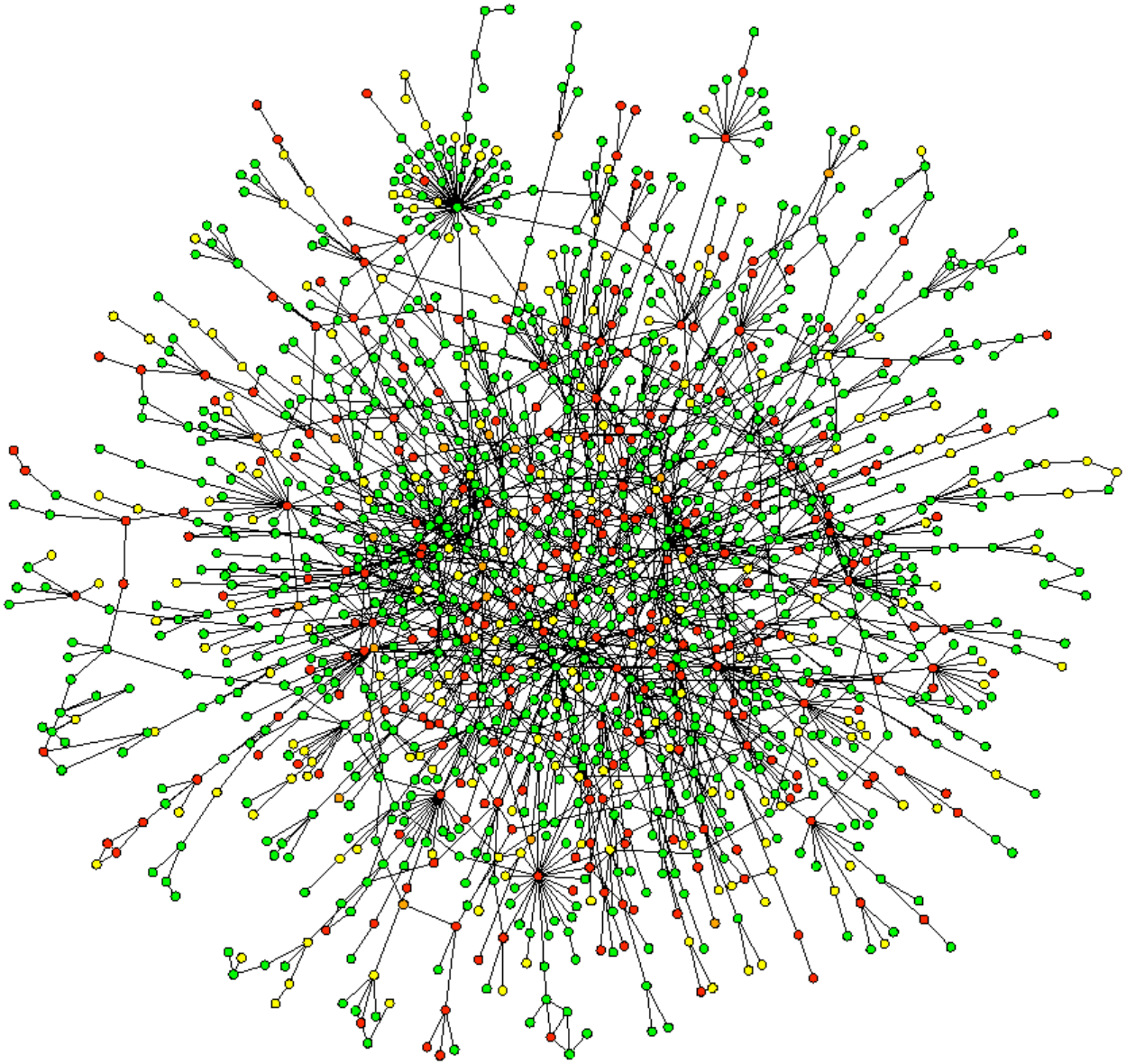


Figure: Learning Gene-Gene Regulatory Networks from microarrays

Seminal reference : Friedman [6]

Example 1



Example 2

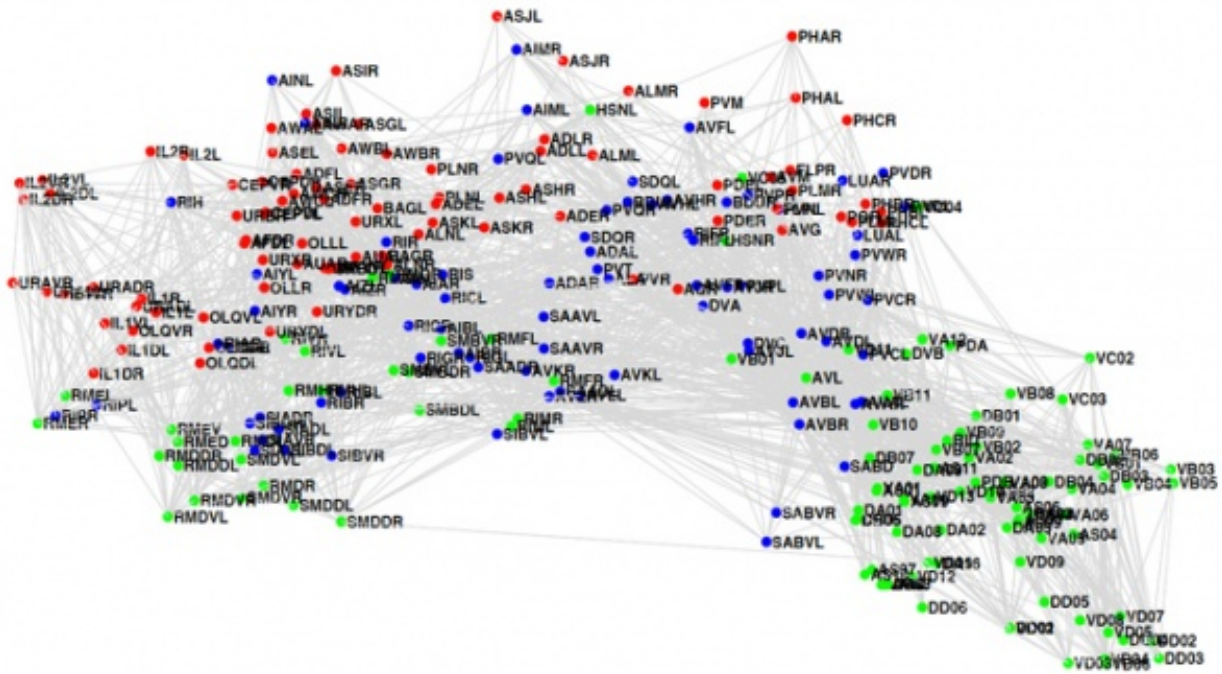


Figure: Learning Brain Connectivity Networks

Example 3

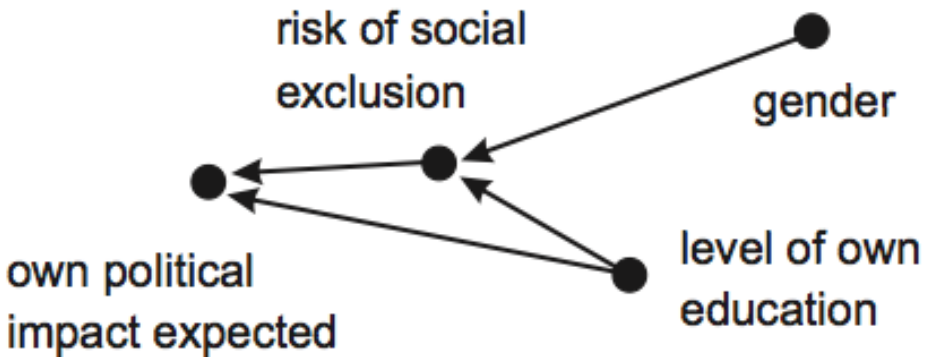


Figure: Learning "direct effects" in Social Sciences

Conditional independence

The concept of conditional dependence is more suited than the concept of dependence in order to catch "direct" dependences between variables

> Traffic jams and snowmen are correlated.
>
> But conditionally on snow falls, the size of the traffic jams and the number of snowmen are independent.

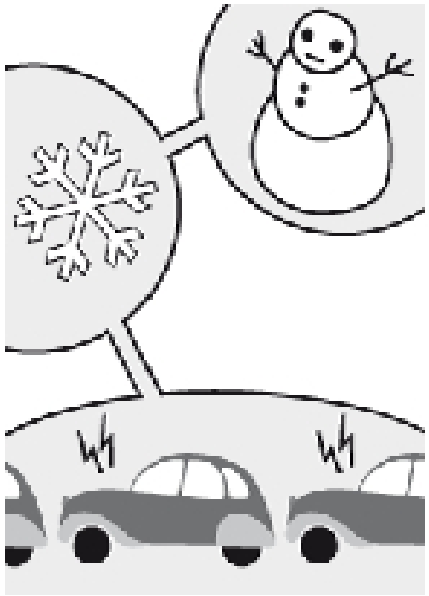


Figure: Difference between dependence and conditional dependence

Conditional independence

random variables X and Y are independent conditionally on a variable Z (we write $X \perp\!\!\!\perp Y \mid Z$) if

$$\text{law}((X, Y)|Z) = \text{law}(X|Z) \otimes \text{law}(Y|Z).$$

Caracterisation

When the distribution of (X, Y, Z) has a positive density f, then

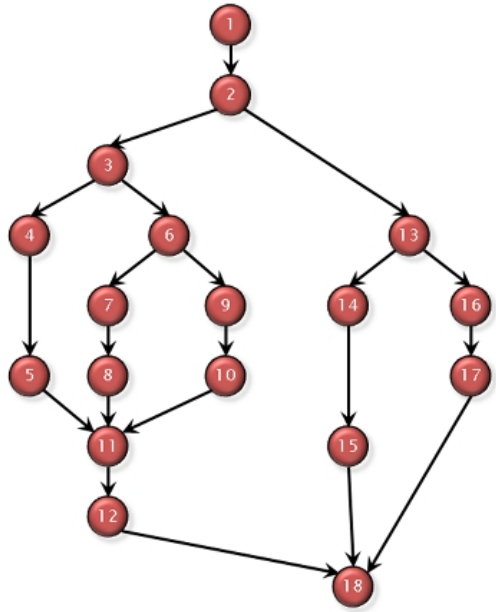
$$\begin{aligned} X \perp\!\!\!\perp Y \mid Z &\iff f(x, y|z) = f(x|z)f(y|z) \\ &\iff f(x, y, z) = f(x, z)f(y, z)/f(z) \\ &\iff f(x|y, z) = f(x|z)\,, \end{aligned}$$

Directed acyclic model

Terminology

Directed graph $\overrightarrow{g}$

set of nodes and arrows



Acyclic

no sequence of arrows forms a loop in the graph

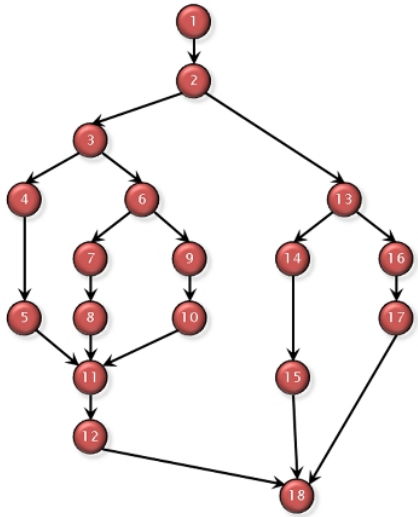
Parents

the parents of a is the set $\mathrm{pa}(a)$ of nodes b such that $b \longrightarrow a$

Descendent

the descendent of a is the set $\mathrm{de}(a)$ of nodes that can be reached from a by following some sequence of arrows.

$X_a \perp\!\!\!\perp \{X_b : a \nrightarrow \ldots \nrightarrow b\}$
conditionally on $\{X_c : c \to a\}$



Directed acyclic graphical model

The law of the random variable $X = (X_1 \ldots, X_p)$ is a graphical model according to the directed acyclic graph $\overrightarrow{g}$ if

$$\text{for all } a, \quad X_a \perp\!\!\!\perp \{X_b,\ b \notin \mathrm{de}(a)\} \mid \{X_c,\ c \in \mathrm{pa}(a)\}$$

We write $\mathcal{L}(X) \sim \overrightarrow{g}$.

Remark: if $\overrightarrow{g} \subset \overrightarrow{g}'$ and $\mathcal{L}(X) \sim \overrightarrow{g}$ then $\mathcal{L}(X) \sim \overrightarrow{g}'$.

Warning

There is no unique minimal graph in general!

Be careful with the interpretation of directed graphical models !

Example:

$$X_{i+1} = \alpha X_i + \varepsilon_i \quad \text{with} \quad \varepsilon_i \text{ independent of } X_1, \ldots, X_{i-1}.$$

Then, the two graphs

$$1 \to 2 \to \ldots \to p \quad \text{and} \quad 1 \leftarrow 2 \leftarrow \ldots \leftarrow p$$

are minimal graphs for this model.

The issue of estimating the minimal $\overrightarrow{g}$ is ill-posed in this context.

Yet,

1. it is very useful for defining / computing laws (next slides)
2. it can be used for exploring "causal effect" (last part of the talk)

Here, we assume that $\overrightarrow{g}$ is known (from expert knowledge).

Factorization formula

If $\mathcal{L}(X) \sim \overrightarrow{g}$, we have

$$f(x_1, \ldots, x_p) = \prod_{b=1}^{p} f(x_b | x_{\mathrm{pa}(b)})$$

Proof: for a leaf p

$$\begin{aligned} f(x_1, \ldots, x_p) &= f(x_p | x_1, \ldots, x_{p-1}) f(x_1, \ldots, x_{p-1}) \\ &= f(x_p | x_{\mathrm{pa}(p)}) f(x_1, \ldots, x_{p-1}) \end{aligned}$$

□

- Very useful for defining / computing f !
- Sampling with Gibbs sampler

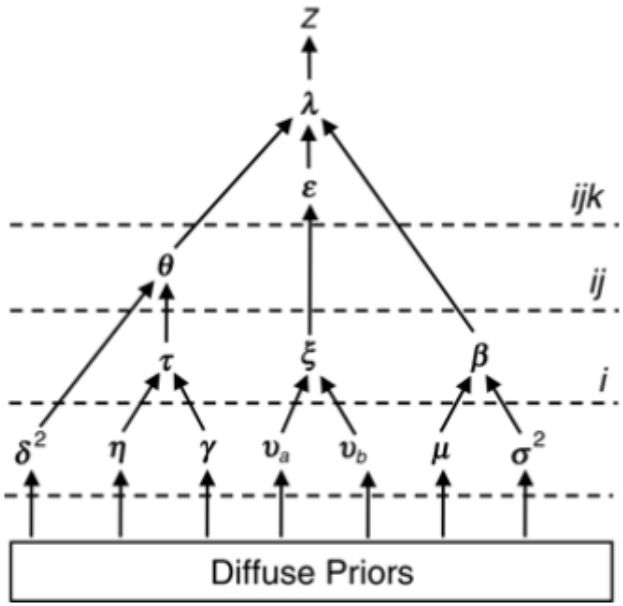
Learning with graphical models

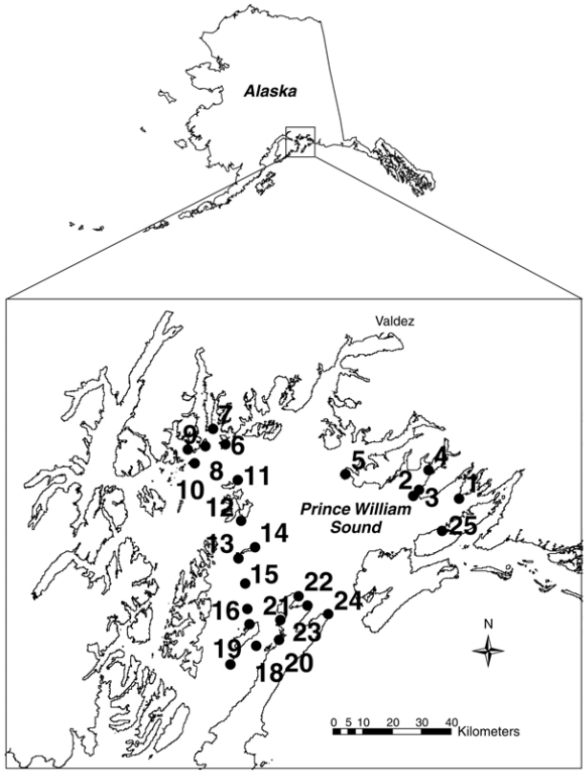
Examples of applications

- speech recognition
- computer vision
- ecological monitoring
- decision making
- diagnosis
- environmental statistics
- etc

Examples

Seals monitoring (Ver Hoef and Frost [17])





Ice streams (Berliner *et al.*)

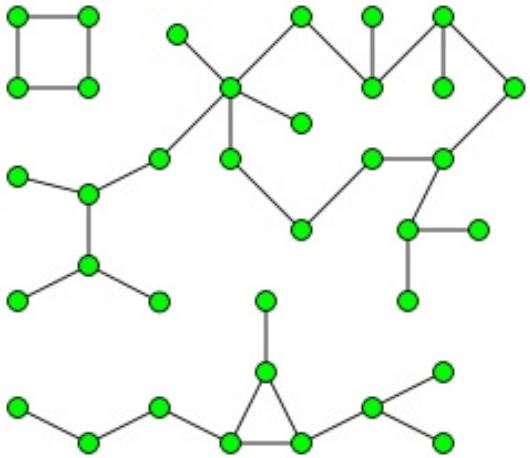
`http://www.stat.osu.edu/~sses/collab_ice.html`

Non-directed model

Terminology

Non-directed graph

set of nodes and edges



The nodes are labelled by $1, \ldots, p$.

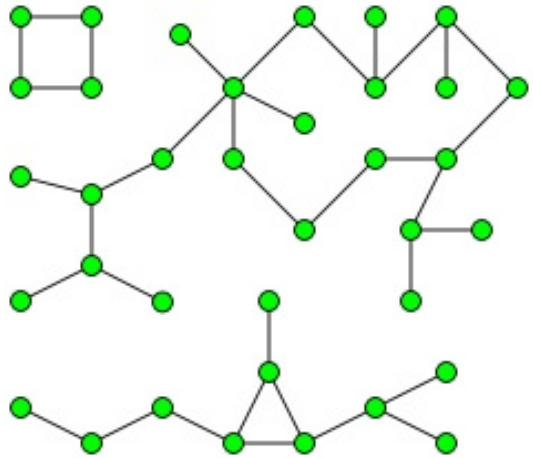
Neighbors

The neighbors of a are the nodes in $\mathrm{ne}(a) = \left\{b : b \overset{g}{\sim} a\right\}$

Class of a

We set $\mathrm{cl}(a) = \mathrm{ne}(a) \cup \{a\}$

X_a independent from $\{X_b : b \nsim a\}$
conditionally on $\{X_c : c \sim a\}$



Non-directed graphical model

The law of the random variable $X = (X_1 \ldots, X_p)$ is a graphical model according to the non-directed graph g if

$$\text{for all } a: \quad X_a \perp\!\!\!\perp \{X_b,\ b \notin \mathrm{cl}(a)\} \mid \{X_c,\ c \in \mathrm{ne}(a)\}.$$

We write $\mathcal{L}(X) \sim g$.

Remark: if $g \subset g'$ and $\mathcal{L}(X) \sim g$ then $\mathcal{L}(X) \sim g'$.

Minimal graph

Minimal graph

When X has a positive density there exists a unique minimal graph g_* such that $\mathcal{L}(X) \sim g_*$.

In the following, we call simply "graph of X" the minimal graph g_* such that $\mathcal{L}(X) \sim g_*$.

Our main goal will be to learn g_* from data.

Connection with directed acyclic graphs

Moral graph

The moral graph g^m associated to a directed acyclic graph $\overrightarrow{g}$ is obtained by

- setting an edge between each parents of each nodes
- replacing arrows by edges

Proposition

$$\mathcal{L}(X) \sim \overrightarrow{g} \quad \Longrightarrow \quad \mathcal{L}(X) \sim g^m$$

Proof :

1. From the factorization formula, $\exists\ g_1, g_2$ such that
$$f(x) = g_1(x_a, x_{\mathrm{ne}^m(a)})\, g_2(x_{\mathrm{nn}^m(a)}, x_{\mathrm{ne}^m(a)})$$
where $\mathrm{nn}^m(a) = \{1, \ldots, p\} \setminus \mathrm{cl}^m(a)$.
2. This ensures $X_a \perp\!\!\!\perp X_{\mathrm{nn}^m(a)}$ given $X_{\mathrm{ne}^m(a)}$. □

Factorization in undirected graphical models

Hammersley-Clifford formula

For a random variable X with positive density f

$$\mathcal{L}(X) \sim g \iff f(x) \propto \exp\left(\sum_{c : c \in \text{cliques}(g)} g_c(x_c) \right).$$

Proof : based on Möbius inversion formula.

Questions so far?

Gaussian graphical models

In the remaining $X \sim \mathcal{N}(0, \Sigma)$, with Σ non singular.

Reminder on Gaussian distribution (1/2)

Proposition 1 : Gaussian conditioning

We consider two sets $A = \{1, \ldots, k\}$ and $B = \{1, \ldots, p\} \setminus A$, and a Gaussian random vector $X = \begin{bmatrix} X_A \\ X_B \end{bmatrix} \in \mathbb{R}^p$ with $\mathcal{N}(0, \Sigma)$ distribution.

We write $K = \begin{bmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{bmatrix}$ for Σ^{-1} and K_{AA}^{-1} for the inverse $(K_{AA})^{-1}$ of K_{AA}.

Then, we have

$$\text{Law}(X_A \,|\, X_B) = \mathcal{N}\left(-K_{AA}^{-1} K_{AB} X_B, K_{AA}^{-1}\right)$$

which means

$$\begin{aligned} X_A &= -K_{AA}^{-1} K_{AB} X_B + \varepsilon_A, \\ &\quad \text{with } \varepsilon_A \sim \mathcal{N}\left(0, K_{AA}^{-1}\right) \text{ independent of } X_B. \end{aligned}$$

Proof: We have for some function f that we do not need to make explicit

$$g(x_A|x_B) = g(x_A, x_B)/g(x_B) = \exp\left(-\frac{1}{2}x_A^T K_{AA} x_A - x_A^T K_{AB} x_B\right) f(x_B).$$

As a consequence,

$$g(x_A|x_B) \propto \exp\left(-\frac{1}{2}(x_A + K_{AA}^{-1} K_{AB} x_B)^T K_{AA}(x_A + K_{AA}^{-1} K_{AB} x_B)\right)$$

where the factor of proportionality does not depend on x_A.

We recognize the density of the Gaussian $\mathcal{N}\left(-K_{AA}^{-1} K_{AB}\, x_B, K_{AA}^{-1}\right)$ law. □

Ref: Lauritzen [11]

Reminder on Gaussian distribution (2/2)

Partial correlation

For any $a, b \in \{1, \dots, p\}$, we have

$$\mathrm{cor}(X_a, X_b | X_c : c \neq a, b) = \frac{-K_{a,b}}{\sqrt{K_{aa} K_{bb}}} .$$

Proof : The previous proposition with $A = \{a, b\}$ and $B = A^c$ gives

$$\mathrm{cov}(X_A | X_B) = \begin{pmatrix} K_{aa} & K_{ab} \\ K_{ab} & K_{bb} \end{pmatrix}^{-1} = \frac{1}{K_{aa}K_{bb} - K_{ab}^2} \begin{pmatrix} K_{bb} & -K_{ab} \\ -K_{ab} & K_{aa} \end{pmatrix}.$$

Plugging this formula in the definition of the partial correlation gives the result. □

Reading the graph g on K

From K to g

We set $K = \Sigma^{-1}$ and define the graph g by

$$a \overset{g}{\sim} b \iff K_{a,b} \neq 0. \tag{1}$$

GGM and precision matrix

For the graph g defined by (1), we have

1. $\mathcal{L}(X) \sim g$ and g is minimal.
2. There exists $\varepsilon_a \sim \mathcal{N}(0, K_{aa}^{-1})$ independent of $\{X_b : b \neq a\}$ such that

$$X_a = - \sum_{b \in \mathrm{ne}(a)} \frac{K_{ab}}{K_{aa}} X_b + \varepsilon_a \,.$$

Proof. We apply Proposition 1 :

1. We set $A = \{a\} \cup \mathrm{nn}(a)$ and $B = \mathrm{ne}(a)$, where $\mathrm{nn}(a) = \{1, \ldots, p\} \setminus \mathrm{cl}(a)$. The precision matrix restricted to A is $K_{AA} = \begin{pmatrix} K_{aa} & 0 \\ 0 & K_{\mathrm{nn}(a)\,\mathrm{nn}(a)} \end{pmatrix}$ so its inverse is $(K_{AA})^{-1} = \begin{pmatrix} K_{aa}^{-1} & 0 \\ 0 & (K_{\mathrm{nn}(a)\,\mathrm{nn}(a)})^{-1} \end{pmatrix}$.

   The above Lemma ensures that the law of $X_{\{a\}\cup\mathrm{nn}(a)}$ given $X_{\mathrm{ne}(a)}$ is Gaussian with covariance matrix $(K_{AA})^{-1}$ so X_a and $X_{\mathrm{nn}(a)}$ are independent conditionally on $X_{\mathrm{ne}(a)}$.

2. The second point is obtained with $A = \{a\}$ and $B = A^c$.

Estimation strategies

Goal

From a n-sample $X_1, \ldots, X_n$ i.i.d. with distribution $\mathcal{N}(0, \Sigma)$, we want to estimate the (minimal) graph g such that $\mathcal{L}(X) \sim g$.

The above results suggest 3 estimations strategies:

1. by estimating the partial correlations + multiple testing
2. by a sparse estimation of K
3. by a regression approach

Estimation with partial correlation (1/3)

Reminder 1

$$a \overset{g}{\sim} b \iff \rho_{a,b} := \mathrm{cor}(X_a, X_b | X_c : \ c \neq a, b) \neq 0$$

Reminder 2

$$\rho_{a,b} = \frac{-K_{a,b}}{\sqrt{K_{aa} K_{bb}}}$$

Partial covariance estimation

For $n > p$, we estimate $\rho_{a,b}$ by

$$\widehat{\rho}_{ab} = \frac{-\left[\widehat{\Sigma}^{-1}\right]_{ab}}{\sqrt{\left[\widehat{\Sigma}^{-1}\right]_{aa}\left[\widehat{\Sigma}^{-1}\right]_{bb}}},$$

where $\widehat{\Sigma}$ is the empirical covariance.

Estimation with partial correlation (2/3)

Under the null hypothesis

when $\rho_{a,b} = 0$ and $n > p - 2$, we have

$$\widehat{t}_{a,b} := \sqrt{n-2-p} \times \frac{\widehat{\rho}_{ab}}{\sqrt{1-\widehat{\rho}_{ab}^2}} \sim \text{Student}(n-p-2).$$

Estimation procedure

1. Compute the $\widehat{t}_{a,b}$
2. Apply a multiple testing thresholding

Weakness

- when $p > n - 2$ the procedure cannot be applied
- when $n > p$ but $n - p$ small, $\widehat{t}_{a,b}$ has a large variance and the procedure is powerless

Estimation with partial correlation (3/3)

Solution 1: Shrinking the conditioning

work with $\widehat{\mathrm{cor}}(X_a, X_b | X_c : c \in S)$ with S small
Ref: Wille and Bühlmann [19], Castelo and Roverato [2], Spirtes et al. [16] or Kalisch and Bühlmann [8].

☺ $\widehat{\mathrm{cor}}(X_a, X_b | X_c : c \in S)$ is stable when S is small

☹ it is unclear what we estimate at the end (in general)

Solution 2 : Sparse estimation of K

The instability for large p comes from the instability of $\widehat{\Sigma}^{-1}$ for estimating K.

Build a more stable estimator of K capitalizing on its sparsity.

Sparse estimation of K (1/2)

The likelihood of a $p \times p$ positive symmetric matrix $K \in \mathcal{S}_p^+$ is

$$\text{Likelihood}(K) = \prod_{i=1}^{n} \sqrt{\frac{\det(K)}{(2\pi)^p}} \exp\left(-\frac{1}{2} X_i^T K X_i\right).$$

Negative log-likelihood

The negative-log-likelihood

$$K \to -\frac{n}{2} \log(\det(K)) + \frac{n}{2} \langle K, \widehat{\Sigma} \rangle_F$$

is convex.

Graphical Lasso : sparse estimation of K

$$\widehat{K}_\lambda = \underset{K \in \mathcal{S}_p^+}{\text{argmin}} \left\{ -\frac{n}{2} \log(\det(K)) + \frac{n}{2} \langle K, \widehat{\Sigma} \rangle_F + \lambda \sum_{a \neq b} |K_{ab}| \right\}$$

Sparse estimation of K (2/2)

- Efficient optimization algorithms.
  Ref: Friedman *et al.* [5], Banerjee *et al.* [1]
- Poor empirical results reported by Villers *et al.* [18]
- Theoretical guaranties under some "compatibility conditions" hard to check/interpret (by Ravikumar *et al.* [14])

keep the (good) idea of exploiting the sparsity, but move to the more classical regression framework.

Regression approach (1/4)

Definitions

- Θ = the set of $p \times p$ matrices with zero on the diagonal
- θ : matrix in Θ defined by $\theta_{ab} = -K_{ab}/K_{bb}$ for $a \neq b$

Characterization

$$\theta = \underset{\theta \in \Theta}{\operatorname{argmin}} \|\Sigma^{1/2}(I - \theta)\|_F^2$$

Proof:

$\mathbb{E}[X_a | X_b : b \neq a] = \sum_b \theta_{ba} X_b$ since $X_a = \sum_b \theta_{ba} X_b + \varepsilon_a$. So:

$$\begin{aligned} \theta &= \underset{\theta \in \Theta}{\operatorname{argmin}} \mathbb{E}\Big[\sum_{a=1}^{p} \Big(X_a - \sum_{b:b \neq a} \theta_{ba} X_b \Big)^2 \Big] \\ &= \underset{\theta \in \Theta}{\operatorname{argmin}} \mathbb{E}\left[\|X - \theta^T X\|^2 \right] = \underset{\theta \in \Theta}{\operatorname{argmin}} \|\Sigma^{1/2}(I - \theta)\|_F^2 \end{aligned}$$

Regression approach (2/4)

Replacing Σ by $\widehat{\Sigma}$, we obtain

$$\langle (I-\theta), \widehat{\Sigma}(I-\theta)\rangle_F = \frac{1}{n}\|\mathbf{X}(I-\theta)\|_F^2.$$

Estimation procedure

$$\widehat{\theta}_\lambda = \underset{\theta\in\Theta}{\operatorname{argmin}} \left\{ \frac{1}{n}\|\mathbf{X}-\mathbf{X}\theta\|_F^2 + \lambda\Omega(\theta) \right\}$$

with $\Omega(\theta)$ enforcing coordinate sparsity.

Examples :

1. ℓ^1 **penalty :** $\Omega(\theta) = \sum_{a\neq b} |\theta_{ab}|$
2. ℓ^1/ℓ^2 **penalty :** $\Omega(\theta) = \sum_{a<b} \sqrt{\theta_{ab}^2 + \theta_{ba}^2}$

Regression approach (3/4)

With the ℓ^1 penalty : (Meinshausen and Bühlmann [13])

☺ We can split the minimization into p problems in $\mathbb{R}^{p-1}$

$$\left[\widehat{\theta}^{\ell^1}_{ba}\right]_{b:b\neq a} = \underset{\beta\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ \frac{1}{n}\|\mathbf{X}_a - \sum_b \beta_b \mathbf{X}_b\|^2 + \lambda|\beta|_{\ell^1} \right\}$$

Very efficient algorithms by coordinate descent.

☹ No constraint enforces that $\widehat{\theta}^{\ell^1}_{ab} \neq 0$ when $\widehat{\theta}^{\ell^1}_{ba} \neq 0$.

$\Longrightarrow$ choose an arbitrary decision rule to build $\widehat{g}$ from $\widehat{\theta}^{\ell^1}$.

Examples:

1. set an edge between $a \sim b$ in $\widehat{g}$ when either $\widehat{\theta}^{\ell^1}_{ab} \neq 0$ or $\widehat{\theta}^{\ell^1}_{ba} \neq 0$.
2. set an edge $a \sim b$ in $\widehat{g}$ when both $\widehat{\theta}^{\ell^1}_{ab} \neq 0$ and $\widehat{\theta}^{\ell^1}_{ba} \neq 0$.

Regression approach (4/4)

With the ℓ^1/ℓ^2 penalty :

☺ Symmetric zeros

$\Longrightarrow$ no ambiguity to define $\widehat{g}$ from $\widehat{\theta}_\lambda^{\ell^1/\ell^2}$

☹ Computational cost

The minimization problem cannot be split into p subproblems and it is less easy to minimize it in large dimensions.

Algorithm : iterate on couple (a, b) until convergence

1. set $\Delta = \begin{pmatrix} \Delta_{ab} \\ \Delta_{ba} \end{pmatrix}$ with $\Delta_{ab} = \frac{1}{n}\mathbf{X}_a^T(\mathbf{X}_b - \sum_{k \neq a,b} \widehat{\theta}_{kb}\mathbf{X}_k)$.
2. set

$$\begin{pmatrix} \widehat{\theta}_{ab} \\ \widehat{\theta}_{ba} \end{pmatrix} \leftarrow \left(1 - \frac{\lambda}{2\|\Delta\|}\right)_+ \begin{pmatrix} \Delta_{ab} \\ \Delta_{ba} \end{pmatrix}.$$

Bayesian approach

A series of papers [20, 4, 15] investigate the Bayesian approach.

Issues

1. design of sensible priors
2. efficient posterior sampling

To the best of my knowledge, cannot handle large dimensional problems

Conclusion(?)

Conclusion

- we have the choice between multiple procedures
- for each procedure, there is at least one (non-scale free) tuning parameter to choose

$$\Longrightarrow \text{we need a selection criterion}$$

Classical selection criterion

We have a collection $\mathcal{G}$ of graphs.

Unbiased risk estimation

$$AIC = -2\log(L(g)) + 2|g|$$

Bayesian criterion

$$-2\log(\mathbb{P}(g|\mathbf{X})) \overset{n\to\infty}{\approx} BIC = -2\log(L(g)) + |g|\log(n) - 2\log(\mathbb{P}(g))$$

Only mathematically grounded in asymptotic setting :
p fixed and $n \to \infty$.

Resampling criterions

Cross-Validation schemes

train	train	train	train	test
train	train	train	test	train
train	train	test	train	train
train	test	train	train	train
test	train	train	train	train

Figure recursive data splitting for 5-fold Cross-Validation

No guaranty in high-dimensional settings : $p \gg n$ or $p \approx n$.

GGMselect

R package (available on `http://cran.r-project.org/`) which

1. generates a collection $\widehat{\mathcal{G}}$ of candidates graphs according to the above procedures ($+$ some variants)
2. selects "the best" graph among $\widehat{\mathcal{G}}$

Quality criterion

For a graph g

$$\begin{aligned} \text{MSEP}(g) &= \text{Mean Square Error of Prediction related to } g \\ &= \text{bias}(g) + \text{variance}(g) \end{aligned}$$

where

- bias(g) quantifies how important are the missing edges
- variance(g) is roughly proportional to the number of edges in g divided by n.

Why MSEP?

It is a way to quantify the importance of each edge.

Ideal

Select $g^* = \text{argmin}\{\text{MSEP}(g) : g \in \widehat{\mathcal{G}}\}$
$\longrightarrow$ g^* unknown!

Selection criterion

"select $\widehat{g}$ which minimizes some penalized empirical MSEP"

where the penalty term:

- roughly penalizes each node of $\widehat{g}$ according to its degree (number of edges),
- is based on quantiles of Fisher random variables.

Theorem : oracle-like inequality for GGMselect

$$\text{If} \quad \max_{g \in \widehat{\mathcal{G}}} \{\mathsf{deg}(g)\} \leq \rho \, \frac{n}{2\left(1.1 + \sqrt{\log p}\right)^2}, \quad \text{for some } \rho < 1,$$

then the estimated graph $\widehat{g}$ fulfills

$$\mathrm{MSEP}(\widehat{g}) \leq c_\rho \, \mathbb{E}\left[\inf_{g \in \widehat{\mathcal{G}}} \{\mathrm{bias}(g) + \log(p)\,\mathrm{var}(g)\}\right] + R_n,$$

where $R_n = O(\mathrm{Tr}(\Sigma) e^{-c'_\rho n} + \mathrm{CVar}(\Sigma)\log(p)/n)$
with $\mathrm{CVar}(\Sigma) = \sum_a \left(\Sigma^{-1}_{aa}\right)^{-1}$.

Ref: Giraud *et al.* [7]

Theorem : oracle-like inequality for GGMselect

$$\text{If} \quad \max_{g \in \widehat{\mathcal{G}}} \{\deg(g)\} \leq \rho \, \frac{n}{2\left(1.1 + \sqrt{\log p}\right)^2}, \qquad \text{for some } \rho < 1,$$

then the estimated graph $\widehat{g}$ fulfills

$$\mathrm{MSEP}(\widehat{g}) \leq c_\rho \, \mathbb{E}\left[\inf_{g \in \widehat{\mathcal{G}}} \{\mathrm{bias}(g) + \log(p)\,\mathrm{var}(g)\}\right] + R_n,$$

where $R_n = O(\mathrm{Tr}(\Sigma) e^{-c'_\rho n} + \mathrm{CVar}(\Sigma)\log(p)/n)$
with $\mathrm{CVar}(\Sigma) = \sum_a \left(\Sigma^{-1}_{aa}\right)^{-1}$.

Ref: Giraud *et al.* [7]

Optimality?

- Optimal selection criterion?
  - "minimal" size of the penalty to avoid overfitting
  - minimax estimation rates when $\widehat{\mathcal{G}}$ contains good graphs

- What about the condition on the degree? $(n/2\log p)$ unavoidable, otherwise estimation rate gets worse.

Gaussianity?

Hammersley-Clifford formula

For a random variable X with positive density f

$$\mathcal{L}(X) \sim g \iff f(x) \propto \exp\left(\sum_{c : c \in \text{cliques}(g)} g_c(x_c) \right).$$

Gaussianity?

Data transformation: $f_j(X_j) := \Phi^{-1}(F_j(X_j)) \sim \mathcal{N}(0,1)$

Assumption: $(f_1(X_1), \ldots, f_p(X_p)) \sim \mathcal{N}(0, \Sigma)$

Key point: $\text{graph}(X_1, \ldots, X_p) = \text{graph}(f_1(X_1), \ldots, f_p(X_p))$

Estimation: work with $\widehat{f}_j(X_j) = \Phi^{-1}(\widehat{F}_j(X_j))$ for some estimator $\widehat{F}_j$.

Ref: Data transformations proposed by Lafferty *et al.* [10]

Hidden variables?

we may only observe part of the relevant variables:

$X = \begin{pmatrix} X_O \\ X_H \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \Sigma_{OO} & \Sigma_{HO} \\ \Sigma_{OH} & \Sigma_{HH} \end{pmatrix}\right)$ with X_O observed and X_H unobserved.

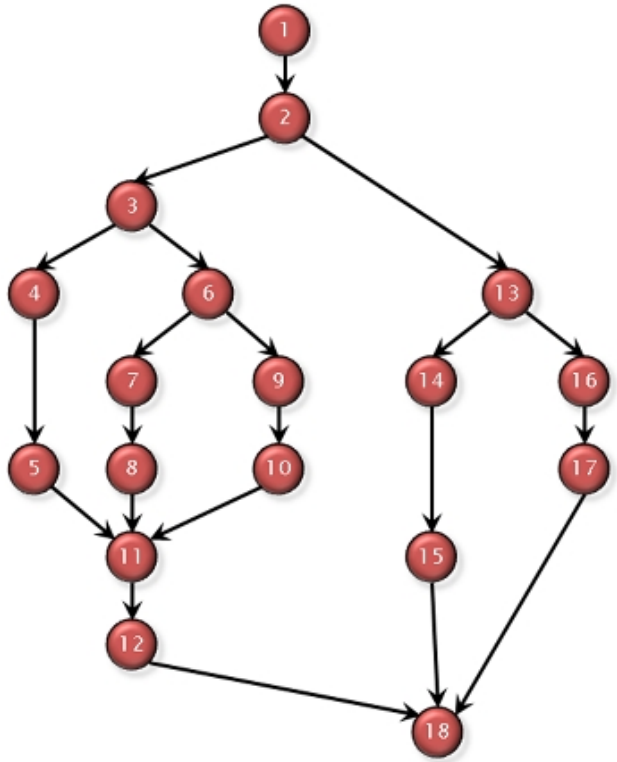
We only have access to $(\Sigma_{OO})^{-1} = K_{OO} - K_{OH}(K_{HH})^{-1}K_{HO}$

Ref: Chandrasekaran *et al.* [3] proposes a sparse + low rank estimation to recover K_{OO} when H is small

Back to directed models

Reminder

$X_a \perp\!\!\!\perp \{X_b : a \nrightarrow \ldots \nrightarrow b\}$
conditionally on $\{X_c : c \rightarrow a\}$

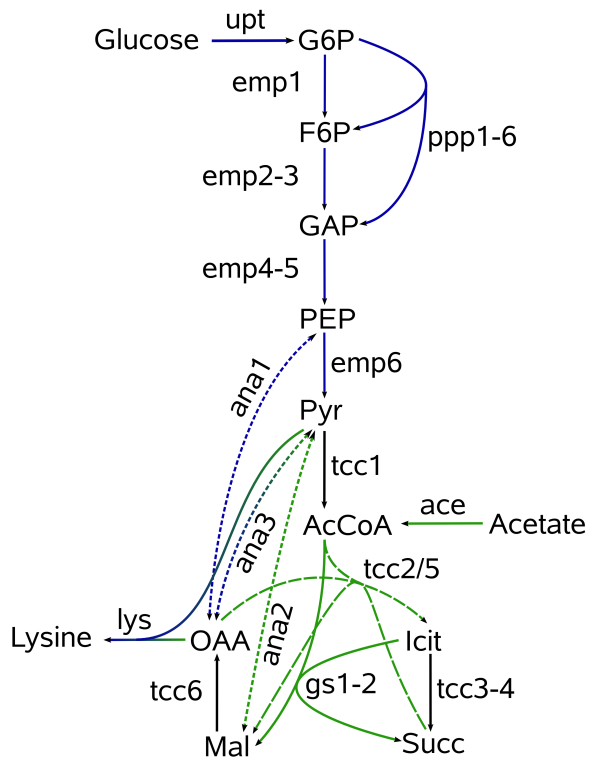


Setting

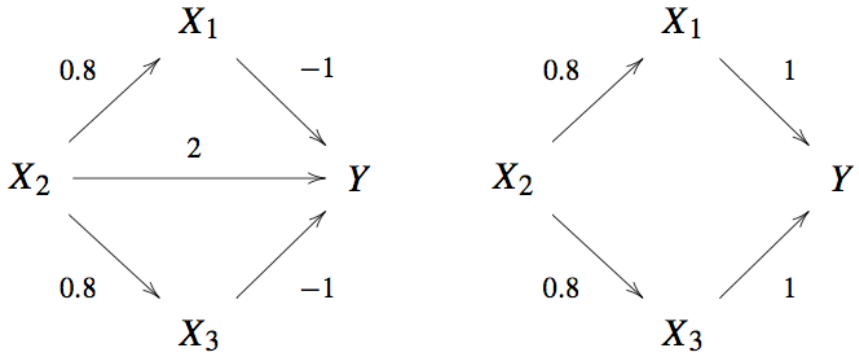
- We have p covariables $X_1, \ldots, X_p$ and Y a variable of interest.
- $[X, Y]$ is a Gaussian graphical model according to $\overrightarrow{g}$.
  No arrows from Y to $X_1, \ldots, X_p$

Example

- Y is the end product of a metabolic network
- $X_1, \ldots, X_p$ are protein abundances



"Causal" inference



Direct versus "Causal" effect

- Direct effect given by θ_a in

$$\mathbb{E}[Y|X_1,\ldots,X_p] = \sum_b \theta_b X_b$$

- Causal effect (relative to $\vec{g}$) given by β_a in

$$\mathbb{E}[Y|X_a, X_b : b \in \mathrm{pa}(a)] = \beta_a X_a + \sum_{b\in\mathrm{pa}(a)} \beta_b X_b$$

Main issue

Causal effects are defined relative to $\overrightarrow{g}$ and there is no unique minimal directed graph...

- find all the DAG $\overrightarrow{g}_{(1)}, \ldots, \overrightarrow{g}_{(m)}$ such that $\mathcal{L}(X) \sim \overrightarrow{g}_{(k)}$
- compute a lower bound of the causal effect:

$$\beta_* = \min\left\{\beta_{(1)}, \ldots, \beta_{(m)}\right\}$$

PCalg

R package (available on `http://cran.r-project.org/`) which

- estimates the DAG $\overrightarrow{g}_{(1)}, \ldots, \overrightarrow{g}_{(m)}$ from the data
- estimates $\beta_{(1)}, \ldots, \beta_{(m)}$ and β_*

Ref:

- Kalish *et al* [9]
- Maathuis *et al.* [12]

PC algorithm

Principle of PC algorithm

Init : $g =$ complete graph

Iterate :

- for $a = 1, \ldots, p$, for $b \in \mathrm{ne}(a)$: remove $a - b$ if $\widehat{\mathrm{cor}}(X_a, X_b) < t_0$
- for $a = 1, \ldots, p$, for $b \in \mathrm{ne}(a)$: remove $a - b$ if $\exists c_1 \in \mathrm{ne}(a)$ such that $\widehat{\mathrm{cor}}(X_a, X_b | X_{c_1}) < t_1$
- for $a = 1, \ldots, p$, for $b \in \mathrm{ne}(a)$: remove $a - b$ if $\exists c_1, c_2 \in \mathrm{ne}(a)$ such that $\widehat{\mathrm{cor}}(X_a, X_b | X_{c_1}, X_{c_2}) < t_2$
- $\ldots$

Output : skeleton of the DAGs

Last step: compute $\overrightarrow{g}_{(1)}, \ldots, \overrightarrow{g}_{(m)}$ from the skeleton

Riboflavin prediction

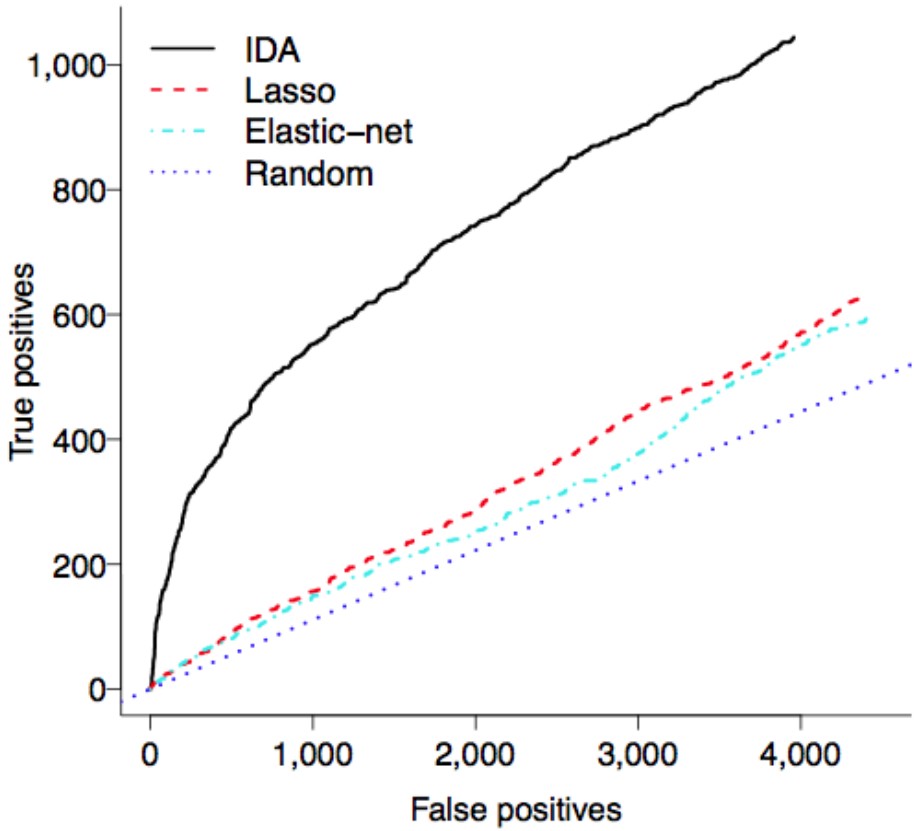


Figure: Important genes for riboflavin production

Ref : Kalish *et al* [9]

Warning

- Be aware of over-interpretation : we cannot reliably infer causal networks on i.i.d. data
- Relies on uncheckable assumptions
- But, seems promising for ranking covariables

Statistics in high-dimensional setting

- Despite theorem, do not trust to much statistical inferences in high-dimensional setting $n \ll p$
  ex: gene pre-selection, metagenes, etc
- It is not a validation tool, but rather a good tool for providing good hints
- Requires experimental validations.

References on high-dimensional statistics:

- Lecture Notes on High-Dimensional Statistics
  `http://www.cmap.polytechnique.fr/~giraud/MSV/LectureNotes.pdf`
- The Element of Statistical Learning
  by Hastie, Tibshirani, Friedman
  `www-stat.stanford.edu/~tibs/ElemStatLearn/`

Thank you!

References I

[1] O. Banerjee, L. El Ghaoui, and A. d'Aspremont.

Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data.

J. Mach. Learn. Res., 9:485–516, 2008.

[2] R. Castelo and A. Roverato.

A robust procedure for Gaussian graphical model search from microarray data with p larger than n.

J. Mach. Learn. Res., 7:2621–2650, 2006.

[3] V. Chandrasekaran, P. Parrilo, and A. Willsky.

Latent variable graphical model selection via convex optimization.

Annals of Statistics, 2012.

References II

[4] P. Dellaportas, P. Giudici, and G. Roberts.
Bayesian inference for nondecomposable graphical Gaussian models.
Sankhyā, 65(1):43–55, 2003.

[5] J. Friedman, T. Hastie, and R. Tibshirani.
Sparse inverse covariance estimation with the graphical lasso.
Biostatistics, 9(3):432–441, 2008.

[6] Nir Friedman.
Inferring Cellular Networks Using Probabilistic Graphical Models.
Science, 303(5659):799–805, 2004.

[7] C. Giraud, S. Huet, and N. Verzelen.
Graph selection with GGMselect.
Stat. Appl. Genet. Mol. Biol., 11(3):1–50, 2012.

References III

[8] M. Kalisch and P. Bühlmann.

Robustification of the pc-algorithm for directed acyclic graphs.

J. Comput. Graph. Statist., 17(4):773–789, 2008.

[9] Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann.

Causal inference using graphical models with the r package pcalg.

Journal of Statistical Software, 47(11):1–26, 5 2012.

[10] John Lafferty, Han Liu, and Larry Wasserman.

Sparse nonparametric graphical models.

Statistical Science, 27:519–537, 2012.

[11] Steffen L. Lauritzen.

Graphical Models.

Oxford University Press, 1996.

References IV

[12] Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann.

Estimating high-dimensional intervention effects from observational data.

Ann. Statist, 37(6A):3133–3164, 2009.

[13] N. Meinshausen and P. Bühlmann.

High-dimensional graphs and variable selection with the lasso.

Ann. Statist., 34(3):1436–1462, 2006.

[14] Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu.

High-dimensional covariance estimation by minimizing, ℓ^1-penalized log-determinant divergence.

Electronic Journal of Statistics, 5:935–980, 2011.

References V

[15] J.G. Scott and C. M. Carvalho.

Feature-inclusion stochastic search for gaussian graphical models.

J. Comp. Graph. Statist., 17:790–808, 2009.

[16] P. Spirtes, C. Glymour, and R. Scheines.

Causation, prediction, and search.

Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2000.

[17] J. Verhoef and K. Frost.

A bayesian hierarchical model for monitoring harbor seal changes in prince william sound, alaska.

Environmental and Ecological Statistics, 10:201–219, 2003.

References VI

[18] Fanny Villers, Brigitte Schaeffer, Caroline Bertin, and Sylvie Huet.

Assessing the Validity Domains of Graphical Gaussian Models in Order to Infer Relationships among Components of Complex Biological Systems.

Statistical Applications in Genetics and Molecular Biology, 7, 2008.

[19] A. Wille and P. Bühlmann.

Low-order conditional independence graphs for inferring genetic networks.

Stat. Appl. Genet. Mol. Biol., 5:Art. 1, 34 pp. (electronic), 2006.

[20] F. Wong, C. K. Carter, and R. Kohn.

Efficient estimation of covariance selection models.

Biometrika, 90(4):809–830, 2003.