INSTITUT de MATHEMATIQUES de TOULOUSE

Institut National de la Recherche Agronomique

Convex optimization for learning Gene Regulatory Network

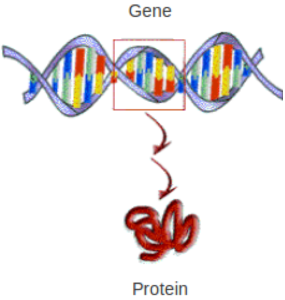
Magali Champion

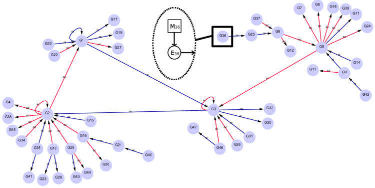
Sébastien Gadat, Christine Cierco-Ayrolles et Matthieu Vignes

14 juin 2013

Plan

Introduction (biological)



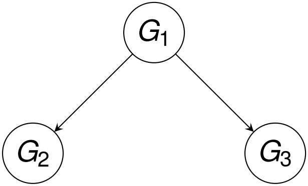


Objective : Recover the unknown gene network $\mathcal{G}$ for which :

- a node of $\mathcal{G}$ is one of the p genes,
- an edge of $\mathcal{G}$ represents an interaction between two genes.

Introduction (statistical)

- p studied genes, for which we know the expression data
- sample of size n



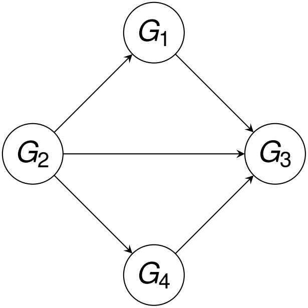
Objective : Recover the unknown gene network $\mathcal{G}$ for which :

- a node of $\mathcal{G}$ is one of the p genes,
- an edge of $\mathcal{G}$ represents an interaction between two genes.

Model I

The first idea consists in considering gene G^j as an observation and the others genes as explanatory variables.

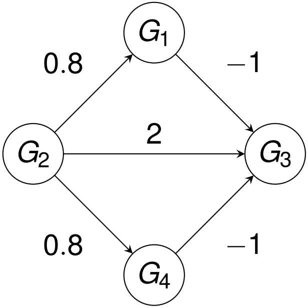
$$\forall 1 \leq j \leq p, \quad G^j = \sum_{1 \leq i \neq j \leq p} G^i + \varepsilon.$$



Model I

The first idea consists in considering gene G^j as an observation and the others genes as explanatory variables.

$$\forall 1 \leq j \leq p, \quad G^j = \sum_{1 \leq i \neq j \leq p} \theta_i^j G^i + \varepsilon.$$



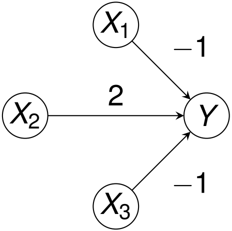
$\Theta = (\theta^1, ..., \theta^p)$ is the adjacency matrix associated to the graph $\mathcal{G}$, which support is denoted $\mathcal{S}$.

$$\Theta = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0.8 & 0 & 2 & 0.8 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{pmatrix}$$

Linear regression

We can first rewrite the model as the following way :

$$Y = X\theta + \varepsilon.$$



Use of

- Lasso,
- Boosting...

$\rightarrow$ Main disadvantage : we don't exploit the structure of the graph.

Model II

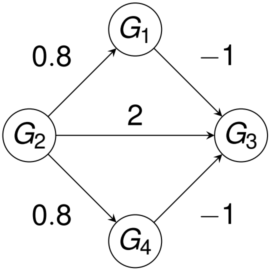
Consider the set of gaussian *Directed Acyclic Graphs*.

Proposition

Any adjacency matrix Θ *associated to a DAG* $\mathcal{G}$ *satisfies :*

$$\Theta = PT^tP,$$

where P *and* T *are permutation and lower-triangular matrices.*



$$T = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 2 & 0.8 & 0.8 & 0 \end{pmatrix}$$

Model II

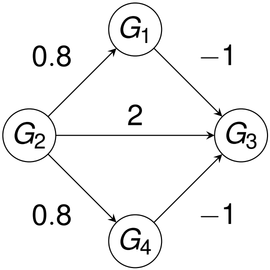
Consider the set of gaussian *Directed Acyclic Graphs*.

Proposition

Any adjacency matrix Θ *associated to a DAG* $\mathcal{G}$ *satisfies :*

$$\Theta = PT^tP,$$

where P *and* T *are permutation and lower-triangular matrices.*



$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Optimization problem

We aim at minimizing the negative penalized log-likelihood :

$$(\hat{P}, \hat{T}) = \underset{P \in \mathbb{P}_p(\mathbb{R}), T \in \mathbb{T}_p(\mathbb{R})}{\text{argmin}} \left\{ \frac{1}{n} \| G - GPT^t P \|_F^2 + \lambda \| T \|_1 \right\},$$

where

- $\mathbb{P}_p(\mathbb{R})$ is the set of permutation matrices,
- $\mathbb{T}_p(\mathbb{R})$ is the set of strict lower-triangular matrices,
- $\|M\|_F = \text{Trace}({}^t MM) = \sum_{i,j} (M_i^j)^2$,
- $\|M\|_1 = \sum_{i,j} |M_i^j|$.

Oracle inequality

Let $\hat{\Theta}$ an estimator of the parameter Θ^* and $R(.)$ a risk function. Oracle inequalities aim at comparing the risk of the proposed estimator with the risk of the "oracle", defined as the unknown parameter which minimizes the risk.

Theorem (Oracle inequality)

$$R(\hat{\Theta}) \leq \inf_{\Theta} \{R(\Theta) + \textit{residual term}\}.$$

Oracle inequality

Assumption $Re(s, c_0)$: for some integer $s \in \{1, ..., p\}$, and $c_0 \geq 0$, the following condition holds :

$$\kappa(s, c_0) := \min_{\substack{J \subset \{1,...,p\} \\ |J| \leq s}} \min_{\substack{M \neq 0 \\ \|M_{J^c}\|_1 \leq c_0 \|M_J\|_1}} \frac{\|XM\|_F}{\sqrt{n}\|M_J\|_F} > 0.$$

Theorem (Oracle inequality)

Let $\eta > 0$ and $s \leq p$. Consider the estimate $\hat{\Theta} = \hat{P}\hat{T}^t\hat{P}$ with $\lambda = A\sigma\sqrt{\frac{\log p}{n}}$, where $A > 4\sqrt{2}$. Then, with probability at least $1 - p^{2-A^2/16}$, there exists $C(\eta)$ such that :

$$\frac{1}{n}\|G\hat{\Theta} - G\Theta^*\|_F^2 \leq$$

$$(1+\eta) \inf_{\Theta, |\mathcal{S}_\Theta| \leq s} \left\{ \frac{1}{n}\|G\Theta - G\Theta^*\|_F^2 + \frac{C(\eta)A^2\sigma^2}{\kappa^2(s, 3+4/\eta)} \frac{\log p}{n} \right\}$$

For P fixed

We aim at minimizing the negative penalized log-likelihood :

$$\hat{T} = \underset{T \in \mathbb{T}_p(\mathbb{R})}{\text{argmin}} \left\{ \frac{1}{n} \|G - GPT^t P\|_F^2 + \lambda \|T\|_1 \right\}.$$

- minimization of a convex, differentiable and quadratic function + penalization

$$T_{k+1} = \underset{T}{argmin} \left\{ \frac{L}{2} \| T - \left(T_k - \frac{\nabla f(T_k)}{L} \right) \|_F^2 + \lambda \| T \|_1 \right\}.$$

- projection on the space of constraints $\mathbb{T}_p(\mathbb{R})$.

For T fixed

We aim at minimizing $\hat{P} = \underset{P \in \mathbb{P}_p(\mathbb{R})}{\text{argmin}} \left\{ \frac{1}{n} \| G - GPT^t P \|_F^2 \right\}$.

Since the space of constraints is not convex, we propose a convex relaxation of the criterion to minimize.

Definition

A bistochastic matrix $A = (a_{ij})_{1 \leq i,j \leq p}$ is a matrix such that :

- $a_{ij} \geq 0$,
- $\sum_i a_{ij} = \sum_j a_{ij} = 1$.

Proposition (Birkhoff)

The set of bistochastic matrices $\mathbb{B}_p(\mathbb{R})$ is a convex set, which permutation matrices are extreme points.

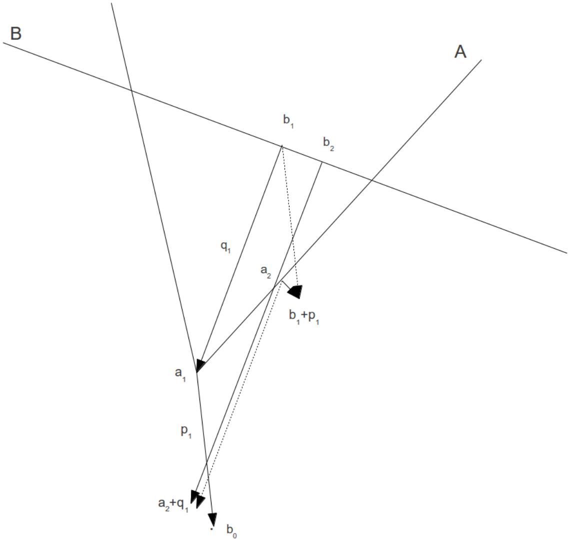
Alternate projection

We can write $\mathbb{B}_p(\mathbb{R}) = \Lambda^+ \cap \mathcal{LC}_1$ as the intersection of the two sets :

1. the convex cone
$\Lambda^+ = \left\{ M = (M_i^j)_{i,j} \in \mathcal{M}_p, \ \ \forall i, j, \ \ M_i^j \geq 0 \right\},$
2. the affine subspace
$\mathcal{LC}_1 = \left\{ M = (M_i^j)_{i,j} \in \mathcal{M}_p, \ \ \sum_{i=1}^p M_i^j = \sum_{j=1}^p M_i^j = 1 \right\}.$

We use alternate projection algorithms (algorithm of Von Neumann or Boyle-Dykstra) to find the expression of the projected bistochastic matrix.

Algorithm of Boyle-Dykstra



Alternate minimization

$$(\hat{P}, \hat{T}) = \underset{P \in \mathbb{B}_p(\mathbb{R}), T \in \mathbb{T}_p(\mathbb{R})}{\text{argmin}} \frac{1}{n} \|G - GPT^t P\|_F^2 + \lambda \|T\|_1 .$$

$$P_0 \in \mathbb{P}_p(\mathbb{R}) \xrightarrow{\text{optimization}} T_0 \xrightarrow{\text{proj}} T'_0 \in \mathbb{T}_p(\mathbb{R})$$

projected gradient descent

$$P_1 \in \mathbb{B}_p(\mathbb{R}) \xrightarrow{\text{optimization}} T_1 \xrightarrow{\text{proj}} T'_1 \in \mathbb{T}_p(\mathbb{R})$$

projected gradient descent

$$P_2 \in \mathbb{B}_p(\mathbb{R}) \longrightarrow \cdots \xrightarrow{\text{Projection over } \mathbb{P}_p(\mathbb{R})}$$

Permutation matrices

We rewrite the problem of finding the projection of any bistochastic matrix $B \in \mathbb{B}_p(\mathbb{R})$ on $\mathbb{P}_p(\mathbb{R})$ as :

$$\begin{aligned}\mathrm{Proj}_{\mathbb{P}_p(\mathbb{R})}(B) &= \underset{P \in \mathbb{P}_p(\mathbb{R})}{\mathrm{argmin}} \;\|B - P\|_F \\ &= \underset{P \in \mathbb{P}_p(\mathbb{R})}{\mathrm{argmin}} -2\langle B, P\rangle_F.\end{aligned}$$

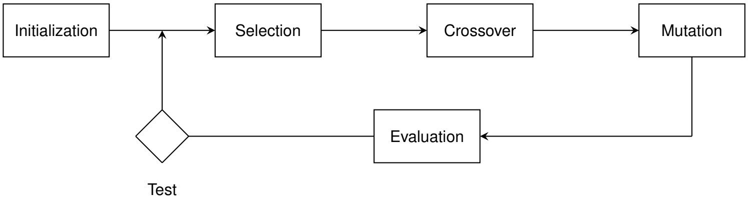
Remark that the new function $-2\langle B, P\rangle_F$ to minimize is linear, whereas the space of constraints $\mathbb{P}_p(\mathbb{R})$ is the set of all extreme points of the convex polytope.
There exists thus an extreme point solution of the relaxed problem

$$P = \underset{P \in \mathbb{B}_p(\mathbb{R})}{\mathrm{argmin}} -2\langle B, P\rangle_F.$$

Another issue

Joint work with Victor

Instead of relaxing the condition $P \in \mathbb{P}_p(\mathbb{R})$, we propose to use genetic algorithms, which are heuristic searchs that mimic the process of natural evolution.



In few words (initialization)

1. We take N permutation matrices. Each of them will be represented by a sequence of “genes”, called “chromosome”.

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

chromosome $\longrightarrow$

2	3	6	1	4	5

2. We search the strict lower-triangular matrix T associated to each chromosome.

In few words (crossover)

1. Selection for the crossover : roulette wheel selection
2. Method of crossover

1	2	3	4	5	6
6	5	1	3	2	4

$\longrightarrow$

		1	3	2	
		3	4	5	

In few words (crossover)

1. Selection for the crossover : roulette wheel selection
2. Method of crossover

1	2	3	4	5	6
6	5	1	3	2	4

→

4	5	1	3	2	6
6	1	3	4	5	2

Performances

Confusion matrix :

		Prediction	
		edge	no edge
Reality	edge	true positives	false negatives
	no edge	false positives	true negatives

We then define :

- the recall

$$R = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}.$$

- the precision

$$Pr = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}.$$

Experimental results

We also compute :

- the MSE : $\|\hat{\Theta} - \Theta^*\|_F^2$
- the MSEP : $\frac{1}{n}\|G - G\hat{\Theta}\|_F^2$.

For $n = 100$ and $p = 5$

	Optimization	G-A	Boosting	Lasso
R	0.86	0.91	0.91	0.83
Pr	0.63	0.69	0.42	0.46
MSE	2.62	0.29	2.33	
MSEP	8.02	4.88	5.26	

Experimental results

We also compute :

- the MSE : $\|\hat{\Theta} - \Theta^*\|_F^2$
- the MSEP : $\frac{1}{n}\|G - G\hat{\Theta}\|_F^2$.

For $n = 100$ and $p = 5$

	Optimization	G-A	Boosting	Lasso	Random
R	0.86	0.91	0.91	0.83	094
Pr	0.63	0.69	0.42	0.46	0.71
MSE	2.62	0.29	2.33		0.29
MSEP	8.02	4.88	5.26		4.88