Identification of causal relationships in gene networks, from observational and interventional expression data

Gilles Monneret

Université Pierre & Marie Curie
Institut National de la Recherche Agronomique

Introduction

Current Goal in Genetics

Since the end Human Genome Project (2003), the new main goal in genomic is to understand functions and links between genes, and find way of action to achieve a particular phenotype.

Applications

- **Medecine**
  - Breast Cancer
  - Laron syndrome
- **Agronomy**
  - Quantitative trait, like meat or milk
  - Animal's robustness
  - Hornless cattle

Detecting Gene influence

Correlation

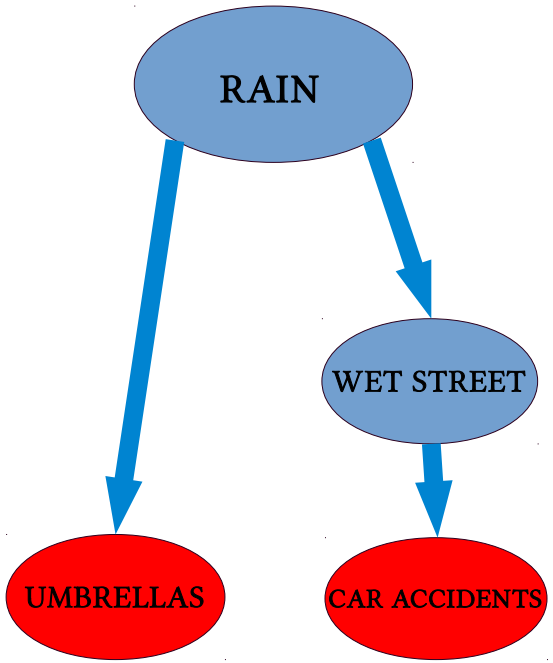
Traditionnal use of statistics give to us correlation$\neq$causality.

- Correlation can be the expression of a (indirect) causal effect
- Can be the consequence of another variable that have a causal effect on both
- Or means nothing, just bad luck or methodological bias (e.g : normalization)

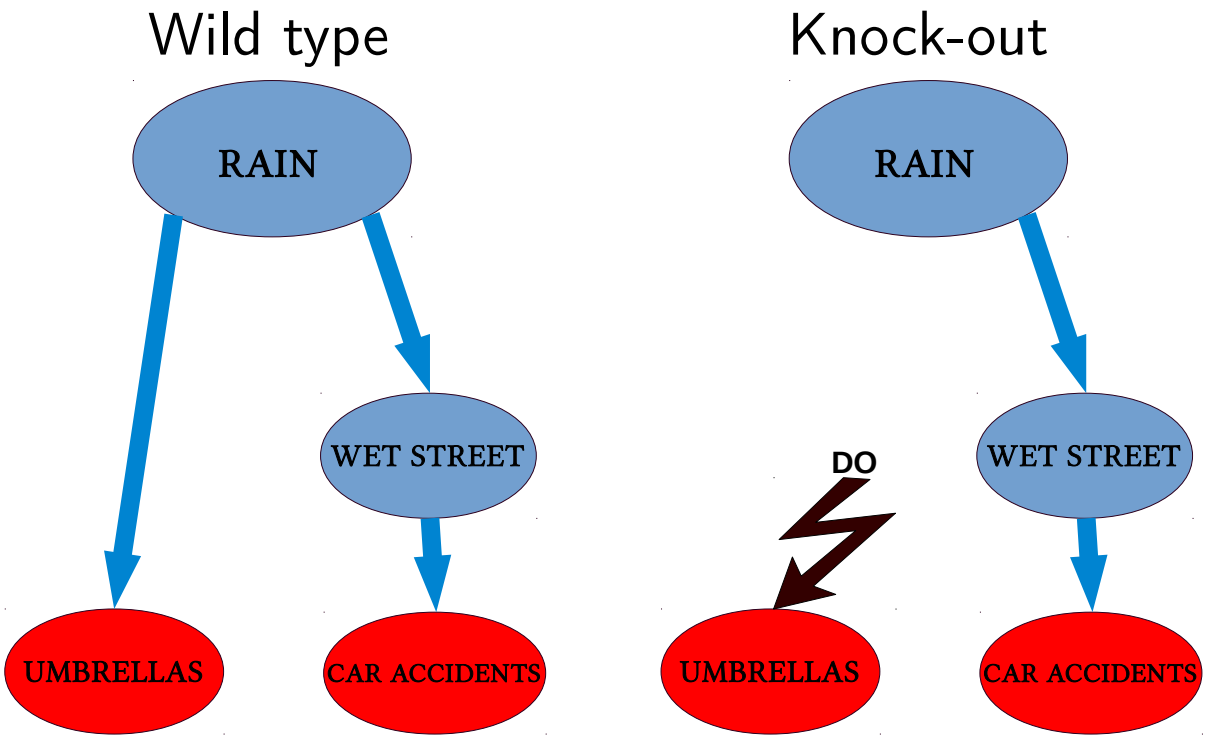
Genes can be correlated because they are located in the same geographic area.

Confounder

Does umbrellas cause car accidents ?

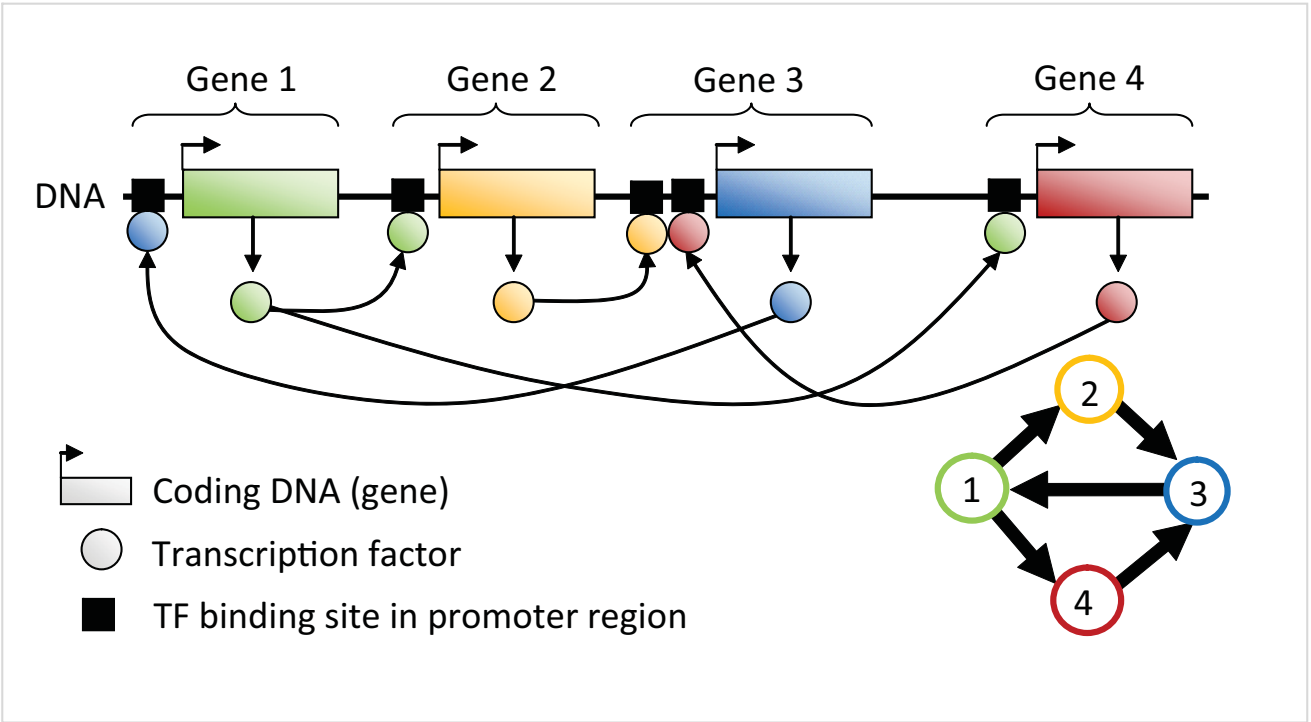


Do-calculus



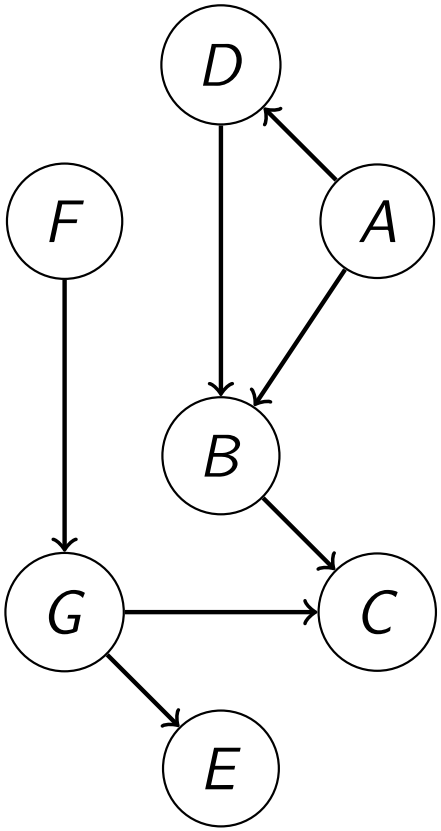
$\mathbb{P}(\text{CAR ACCIDENTS}|\text{UMBRELLAS=YES}) \neq \mathbb{P}(\text{CAR ACCIDENTS}|\,\text{do}(\text{UMBRELLAS=YES}))$

Gene Regulatory Network



MCMC-Mallows

Bayesian Network

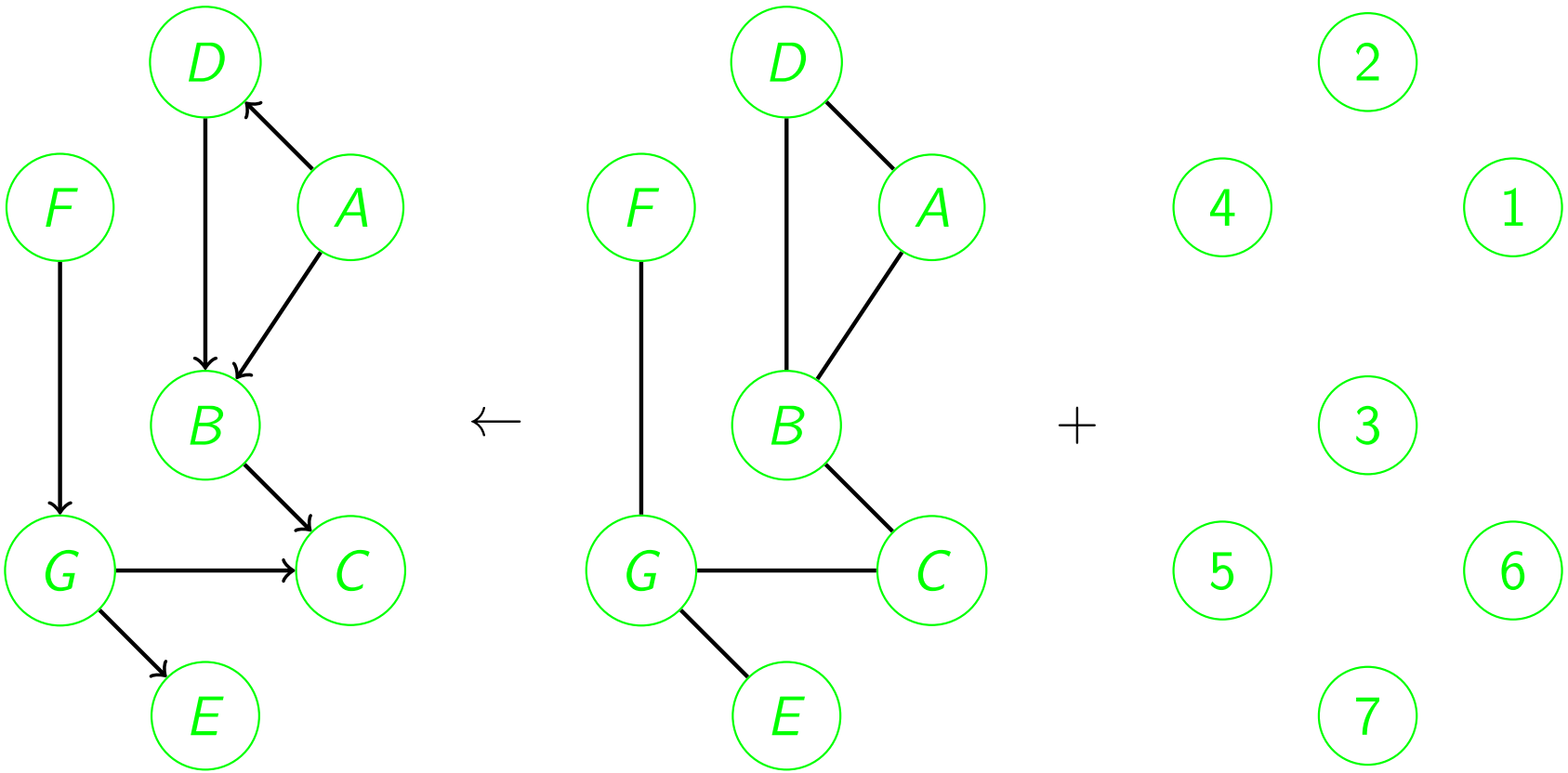


- Directed Acyclic Graph $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$, No directed cycles are allowed.
- Bayesian Network $\mathcal{B} = (\mathcal{G}, \mathbb{P}_\theta)$.
- $\boldsymbol{V}$ are linked to $\boldsymbol{X} \sim \mathbb{P}_\theta$.
- $\mathbb{P}_\theta(\boldsymbol{X}) = \prod_i \mathbb{P}_\theta(X_i | \text{pa}(X_i))$.
- Causal Network : experiments, interventions... $\mathbb{P}_\theta \rightarrow \tilde{\mathbb{P}}_\theta$

First use of directed graph : Geneticist Wright (1921).

Wright, S. (1934). The method of path coefficients. The annals of mathematical statistics, 5(3), 161-215.

Topological ordering



Causal Gaussian Bayesian Network

X_j^k is the expression of gene $j \in 1, \ldots, p$ in experiment $k \in 1, \ldots, N$

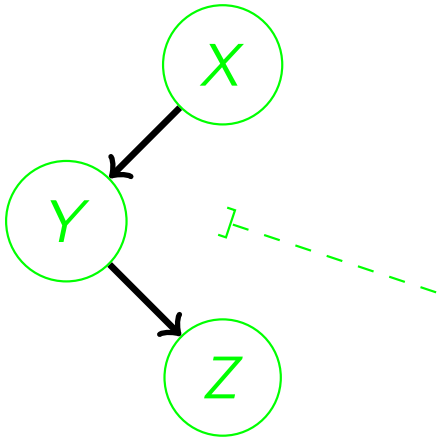
$$X_j^k = m_j + \sum_{i \in \text{pa}(j)} W_{i,j} X_i^k + \varepsilon_j \text{ with } \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$$

with $W_{i,j} \neq 0$ if and only if $i \in \text{pa}(j)$ and nodes ordered such that $i \in \text{pa}(j) \Rightarrow i < j$ (i.e., $\mathbf{W} = (W_{i,j})$ is upper triangular). Model parameters are $\theta = (\mathbf{W}, \mathbf{m}, \boldsymbol{\sigma})$.

- Direct causal effects are $\mathbf{W}$
- Total causal effects are $\mathbf{L} = (\mathbf{I} - \mathbf{W})^{-1} = \mathbf{I} + \mathbf{W} + \ldots + \mathbf{W}^{p-1}$

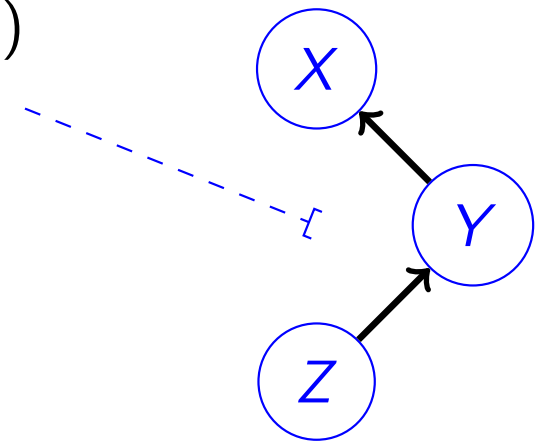
$$W_{i,j} = \frac{d}{dx}\mathbb{E}[X_j | \text{do}\left(X_{-j}\right), X_i = x] \quad L_{i,j} = \frac{d}{dx}\mathbb{E}[X_j | \text{do}(X_i = x)]$$

Markov equivalence



$$\mathbb{P}(\text{Graph}) = \mathbb{P}(X)\,\mathbb{P}(Y|X)\,\mathbb{P}(Z|Y)$$

$$\mathbb{P}(\text{Graph}) = \mathbb{P}(Z)\,\mathbb{P}(Y|Z)\,\mathbb{P}(X|Y)$$



With only observational data,
we can not choose between these models.

Posterior Causal Ordering

For any given ordering $\boldsymbol{o} = o_1, o_2, \ldots, o_p$ we assume the full model : $W_{i,j} \neq \forall i < j$ (not suitable for large p without some kind of regularization).

Posterior Causal Ordering is defined as :

$$\mathbb{P}(\boldsymbol{o}|\text{data}) \propto \mathbb{P}(\text{data}|\hat{\boldsymbol{\theta}}_{\boldsymbol{o}}) \times \mathbb{P}(\boldsymbol{o})$$

where $\hat{\boldsymbol{\theta}}_{\boldsymbol{o}}$ is the MLE of the full model with causal ordering $\boldsymbol{o}$ and $\mathbb{P}(\boldsymbol{o})$ is a prior distribution.

Causal effect estimates :

$$\hat{\boldsymbol{W}} = \sum_{\boldsymbol{o}} \mathbb{P}(\boldsymbol{o}|\text{data}) \times \hat{\boldsymbol{W}}_{\boldsymbol{o}} \quad \text{and} \quad \hat{\boldsymbol{L}} = \sum_{\boldsymbol{o}} \mathbb{P}(\boldsymbol{o}|\text{data}) \times \hat{\boldsymbol{L}}_{\boldsymbol{o}}$$

log-likelihood (1)

Consider experiment k with intervention on $\mathcal{J}_k$ ($\mathcal{J}_k = \emptyset$ means no intervention), where $\mathcal{K}_j = \{k, j \notin \mathcal{J}_k\}$ and $N_j = |\mathcal{K}_j|$.

The log-likelihood of the model can be written as :

$$\ell(\mathbf{m}, \boldsymbol{\sigma}, \mathbf{W}) = \text{Cst} - \sum_j N_j \log(\sigma_j) - \frac{1}{2} \sum_j \frac{1}{\sigma_j^2} \sum_{k \in \mathcal{K}_j} (x_j^k - \mathbf{x}^k \mathbf{W} \mathbf{e}_j^T - m_j)^2$$

Then

$$m_j = \frac{1}{N_j} \sum_{k \in \mathcal{K}_j} (x_j^k - \mathbf{x}^k \mathbf{W} \mathbf{e}_j^T)$$

log-likelihood (2)

The log-likelihood of the model can then be rewritten as :

$$\tilde{\ell}(\boldsymbol{\sigma}, \mathbf{W}) = \text{Cst} - \sum_j N_j \log(\sigma_j) - \frac{1}{2} \sum_j \frac{1}{\sigma_j^2} \sum_{k \in \mathcal{K}_j} (y_j^{k,j} - \mathbf{y}^{k,j} \mathbf{W} \mathbf{e}_j^T)^2$$

where for (k, j) such that $k \in \mathcal{K}_j$: $\mathbf{y}^{k,j} = \mathbf{x}^k - 1/N_j \sum_{k' \in \mathcal{K}_j} \mathbf{x}^{k'}$

Then $\mathbf{W}$ solution of the following linear system :

$$\sum_{i', (i',j) \in \mathcal{E}} W_{i',j} \sum_{k \in \mathcal{K}_j} y_i^{k,j} y_{i'}^{k,j} = \sum_{k \in \mathcal{K}_j} y_i^{k,j} y_j^{k,j} \quad \text{for all } (i,j) \in \mathcal{E}$$

and

$$\sigma_j^2 = \frac{1}{N_j} \sum_{k \in \mathcal{K}_j} (y_j^{k,j} - \mathbf{y}^{k,j} \mathbf{W} \mathbf{e}_j^T)^2$$

Metropolis-Hasting

Objective : draw samples from $\mathbb{P}(\boldsymbol{o}|\text{data})$ (which is only known up to a normalization factor).

Metropolis-Hasting algorithm :

1. start from arbitrary order $\boldsymbol{o}^{(0)}$
2. for $i = 1, \ldots, N$:
   - propose $\boldsymbol{o}'$ according to proposal distribution $Q(\boldsymbol{o}'|\boldsymbol{o}^{(i-1)})$
   - compute acceptance rate

$$\min\left(1, \frac{\mathbb{P}(\boldsymbol{o}'|\text{data}) \times Q(\boldsymbol{o}^{(i-1)}|\boldsymbol{o}')}{\mathbb{P}(\boldsymbol{o}^{(i-1)}|\text{data}) \times Q(\boldsymbol{o}'|\boldsymbol{o}^{(i-1)})}\right)$$

   - if move accepted $\boldsymbol{o}^{(i)} = \boldsymbol{o}'$ else $\boldsymbol{o}^{(i)} = \boldsymbol{o}^{(i-1)}$
3. $\boldsymbol{o}^{(0)}, \boldsymbol{o}^{(1)}, \boldsymbol{o}^{(N)}$ is a (dependent) sample of the target distribution.

Limitations

Problem in High Dimension :

1. If we do not have enough Data, we can not solve

$$\sum_{i',(i',j)\in\mathcal{E}} W_{i',j} \sum_{k\in\mathcal{K}_j} y_i^{k,j} y_{i'}^{k,j} = \sum_{k\in\mathcal{K}_j} y_i^{k,j} y_j^{k,j} \quad \text{for all } (i,j)\in\mathcal{E}$$

2. Large number of parameters. $\frac{p(p+1)}{2} + 2p$ parameters to estimate, with generally p observation/intervention data or less : overfitting.
3. The search space is huge : $p!$ orders. We need to explore this space, so the MCMC-Mallows takes a long time until convergence.

Penalized Maximum Likelihood

The penalized log-likelihood can be written as :

$$\ell(\boldsymbol{m}, \boldsymbol{\sigma}, \boldsymbol{W}) = Cst - \frac{1}{2}\sum_j \frac{1}{\sigma_j^2} \sum_{k \in \mathcal{K}_j} (y_j^{k,j} - \boldsymbol{y}^{k,j} \boldsymbol{W} \boldsymbol{e}_j^T)^2 - \frac{\lambda}{2} \sum_{(i,j) \in \mathcal{E}} w_{ij}^2$$

$\boldsymbol{W}$ become the solution of the following linear system, for all (i, j) s.t. $i \in \text{pa}_j$:

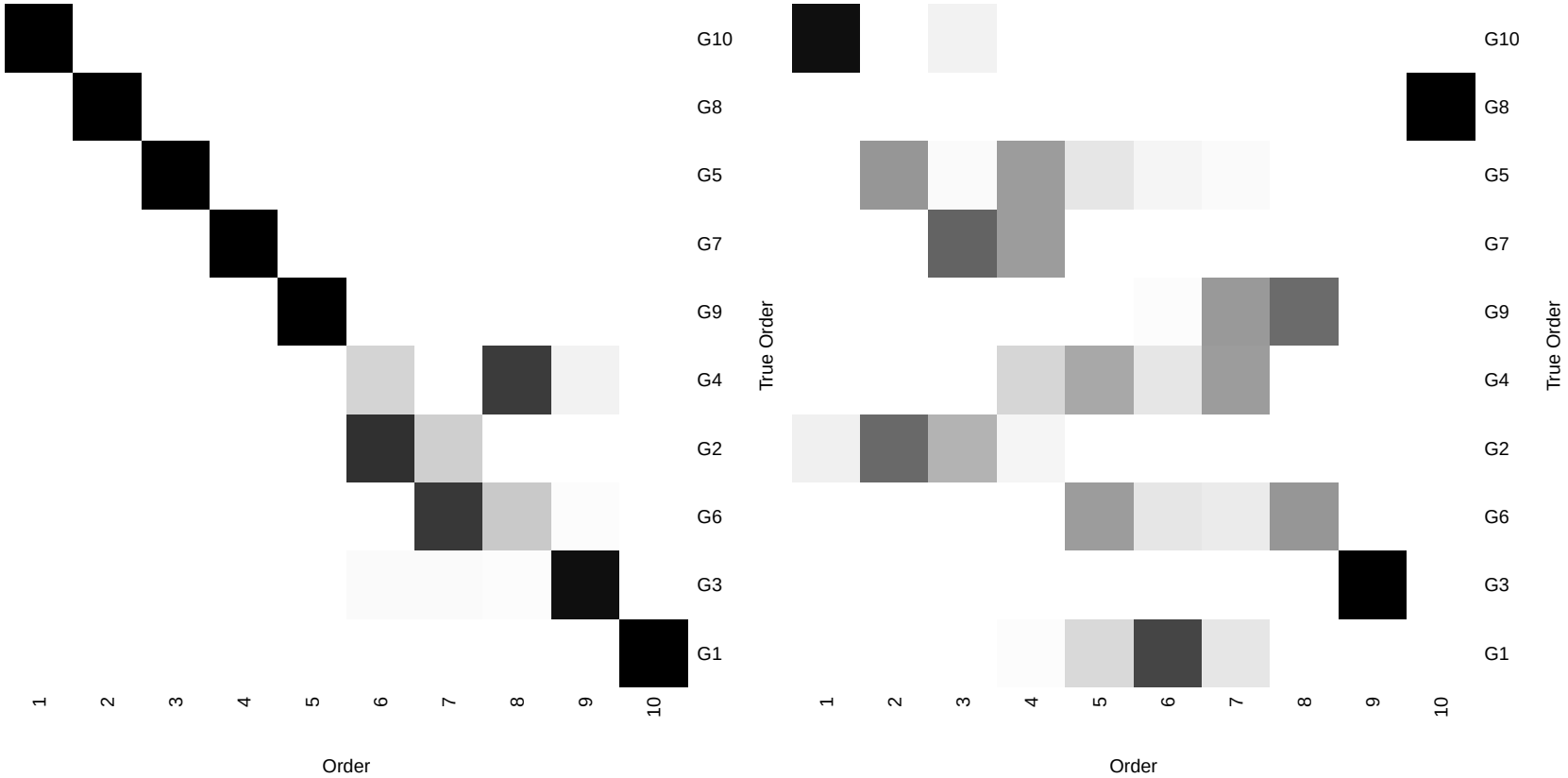
$$\sum_{k \in \mathcal{K}_j} y_i^{k,j} \sum_{i', (i',j) \in \mathcal{E}} w_{i',j} y_{i'}^{k,j} = \sum_{k \in \mathcal{K}_j} y_i^{k,j} y_j^{k,j} - \lambda \sigma_j^2 w_{i,j}$$

Skeleton - Simulation

- Fixed DAG of 10 nodes
- 100 run with 10 random observations each.
- Estimation of skeleton with **glasso**

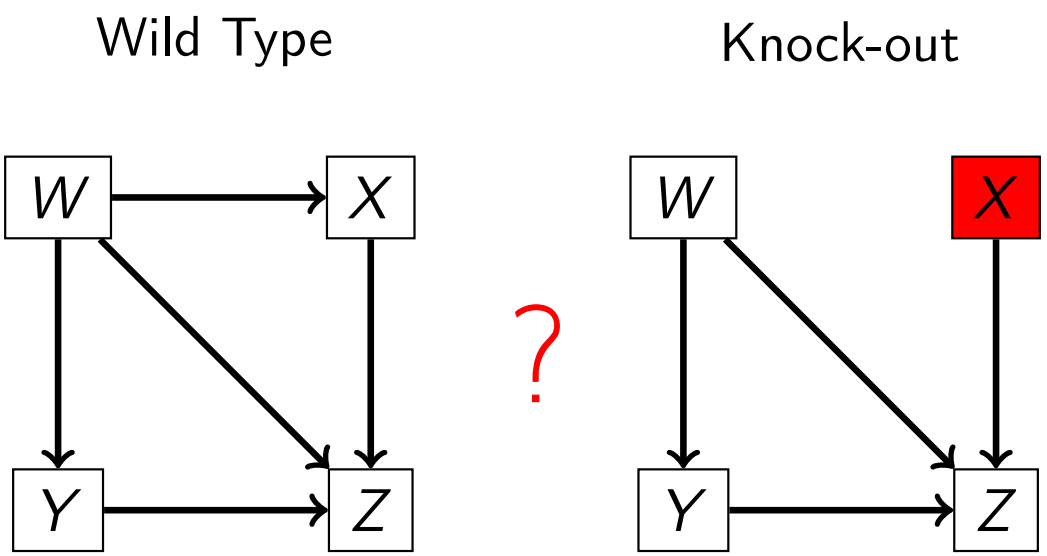
Posterior Order 10 nodes

A posteriori order with and without skeleton.



Marginal approach

Gaussian Bayesian Network



We assume that data are generated under a **directed acyclic graph**.
Our approach : marginal linear causal analysis

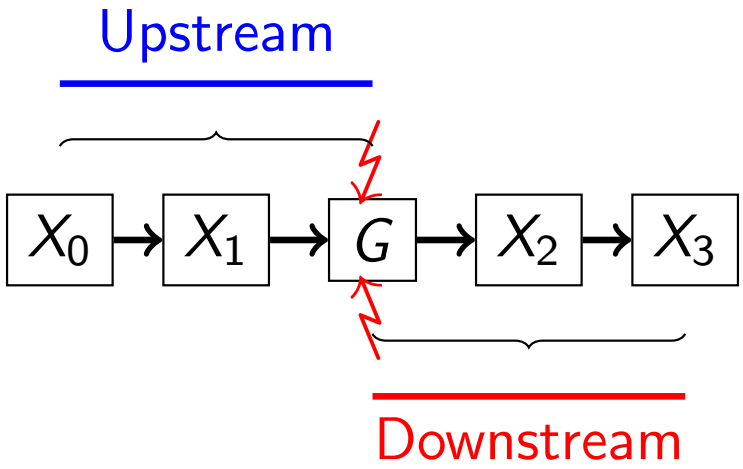
$$Z = \alpha X + \mu_Z + \epsilon_Z$$

$$\mu_Z = \beta W + \gamma Y + \tilde{\mu_Z}$$

Where $\epsilon_Z \sim \mathcal{N}(0, \sigma_Z^2)$: we want to find *causal links*.

Marginal causality : Upstream/Downstream

G "do-node"



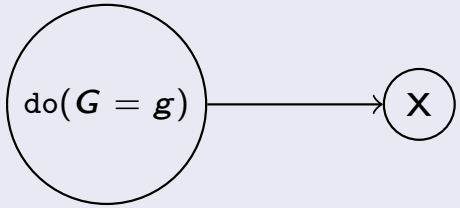
- Upstream : $X_0 \rightarrow G$, $X_1 \rightarrow G$, node G ko
- Downstream : $G \rightarrow X_2$, $G \rightarrow X_3$ node G ko

Models

M1

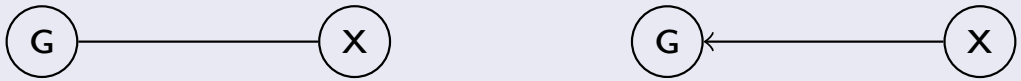
Wild type

Knock-out

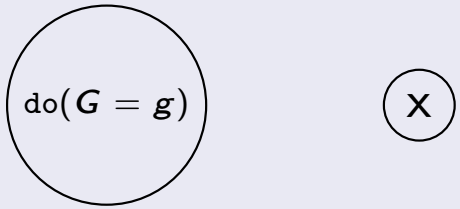


M0

Wild type



Knock-out



Correlation/Causality model and Bayes Factor

Markov equivalence

$$M1 : Z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2) \quad Z_2 \sim \mathcal{N}(\mu_2, \sigma_2^2) \quad G = Z_1 \quad X = \alpha Z_1 + Z_2$$

$$M0 : \tilde{Z_1} \sim \mathcal{N}(\tilde{\mu_1}, \tilde{\sigma_1}^2) \quad \tilde{Z_2} \sim \mathcal{N}(\tilde{\mu_2}, \tilde{\sigma_2}^2) \quad G = \beta \tilde{Z_1} + \tilde{Z_2} \quad X = \tilde{Z_1}$$

$$M0 : \quad \begin{pmatrix} G \\ X \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \begin{pmatrix} s_1^2 & \rho s_1 s_2 \\ \rho s_1 s_2 & s_2^2 \end{pmatrix} \right)$$

$$\mu_1 = m_1 \quad \mu_2 = m_2 - \alpha m_1 \quad \sigma_1 = s_1 \quad \alpha = \rho s_2 / s_1 \quad \sigma_2 = \sqrt{s_2^2 - \alpha^2 s_1^2}$$

$$\tilde{\mu_1} = m_2 \quad \tilde{\mu_2} = m_1 - \beta m_2 \quad \tilde{\sigma_1} = s_2 \quad \beta = \rho s_1 / s_2 \quad \tilde{\sigma_2} = \sqrt{s_1^2 - \beta^2 s_2^2}$$

Bayes Factor

$$\begin{aligned} B &= \exp\left(\ell_0(\hat{\theta}_0)\right) / \exp\left(\ell_1(\hat{\theta}_1)\right) \\ B &= \mathbb{P}(\text{data}|M_0)/\mathbb{P}(\text{data}|M_1) \end{aligned}$$

Likelihood

$$\ell_{\mathbf{WT}}(\theta) = \sum_{k \in \mathbf{WT}} \log \Phi\left(X_k | \mu_X + \alpha G_k, \sigma_X^2\right) + \log \Phi\left(G_k | \mu_G, \sigma_G^2\right),$$

$$\ell_{\mathbf{KO}}^{1}(\theta) = \sum_{k \in \mathbf{KO}} \log \Phi\left(X_k | \mu_X + \alpha G_k, \sigma_X^2\right),$$

$$\ell_{\mathbf{KO}}^{2}(\theta) = \sum_{k \in \mathbf{KO}} \log \Phi\left(X_k | \mu_X + \alpha \mu_G, \alpha^2 \sigma_G^2 + \sigma_X^2\right),$$

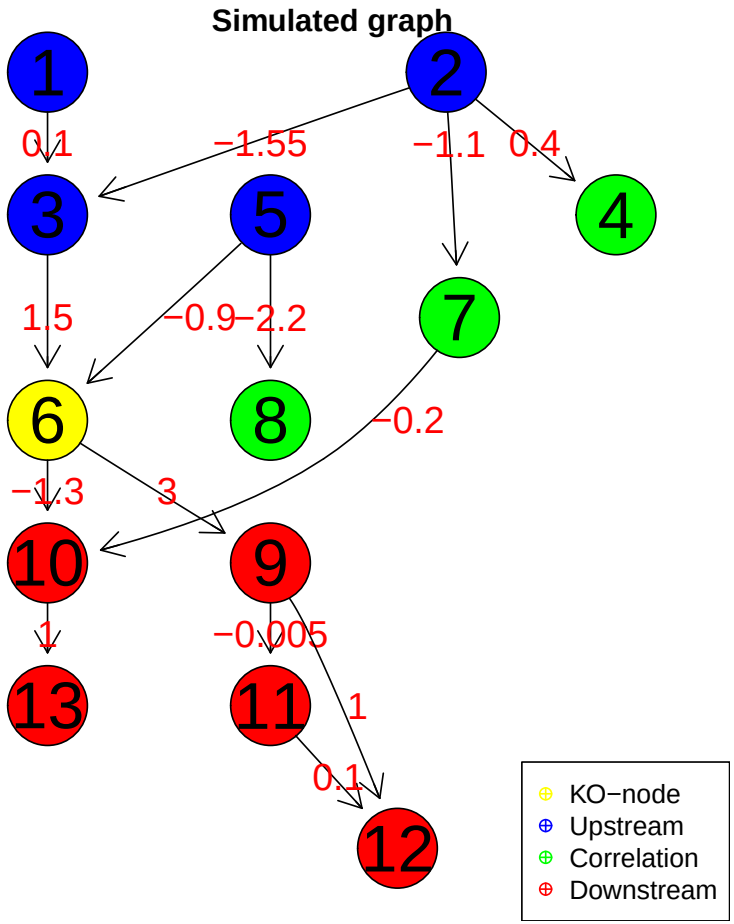
M1 : Downstream causality

$$\ell_{M_1}(\theta) = \ell_{\mathbf{WT}}(\theta) + \ell_{\mathbf{KO}}^{1}(\theta)$$

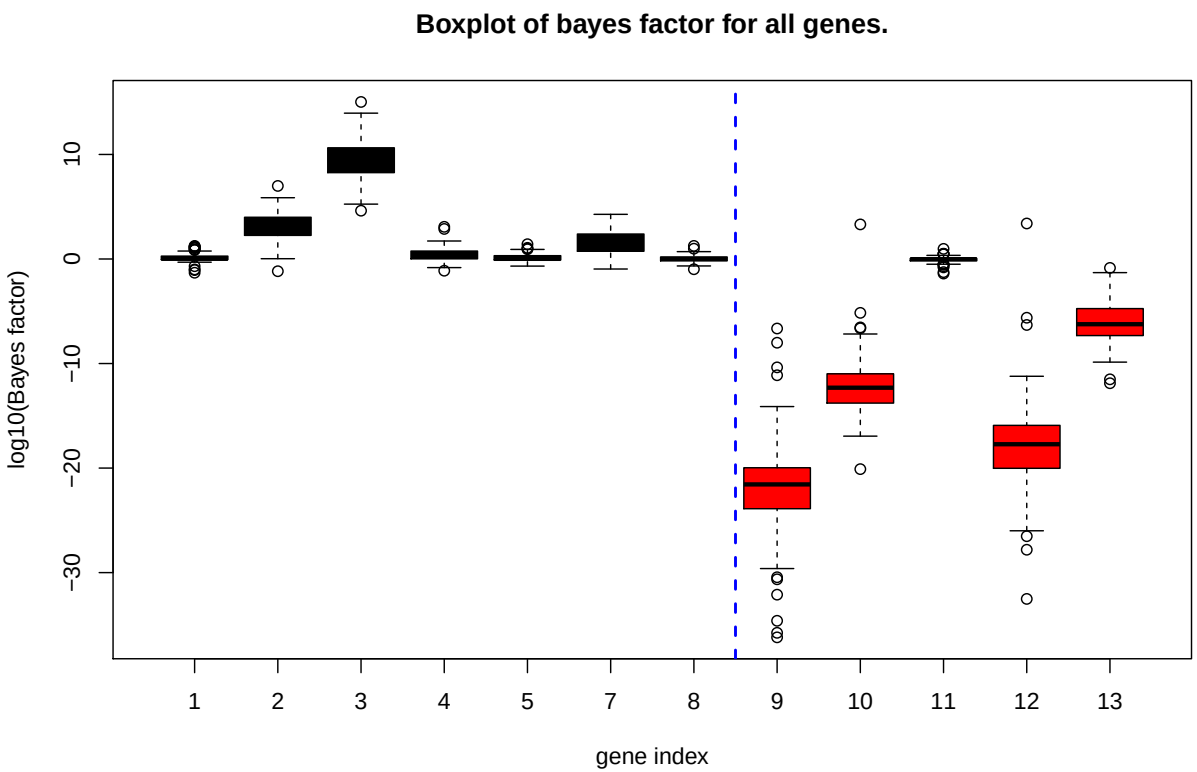
M0 : Upstream/Correlation case

$$\ell_{M_0}(\theta) = \ell_{\mathbf{WT}}(\theta) + \ell_{\mathbf{KO}}^{2}(\theta).$$

Simulation


Simulated graph
1
2
3
5
4
7
6
8
10
9
13
11
12
0.1
−1.55
−1.1
0.4
1.5
−0.9
−2.2
−0.2
−1.3
3
1
−0.005
1
0.1
KO−node
Upstream
Correlation
Downstream

Results

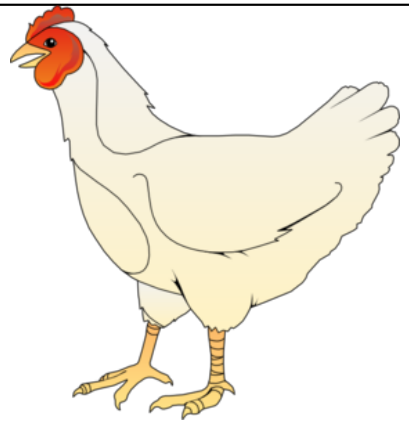


1	2	3	4	5	6	7	8	9	10	11	12	13
3.80	1.06	1.09	0.88	0.66	0.87	0.76	5.27	1.21	1.30	1.11	1.33	2.45

Table – Residual standard deviation

Application on real dataset

Data analysis : Causality

		
Number of samples	24	24
Growth Hormone Receptor state	active	inactive
Genes	43088	
Differentially expressed genes	16276	

- How to find the causal relationship between GHR and other genes ? Causality paradigm
- Classical analysis : differential analysis with limma... but does not identify real causal relationships.

Differential Analysis

Two sample t-test

X_{WT} and X_{KO} are n samples for two random variable, with respective mean $\mathbb{E}(X_{\mathrm{WT}})$ and $\mathbb{E}(X_{\mathrm{KO}})$. $\hat{s} = \sqrt{\left(\hat{s}_{WT}^2 + \hat{s}_{KO}^2\right)/2}$

$$H_0 : \beta = \mathbb{E}(X_{\mathrm{WT}}) - \mathbb{E}(X_{\mathrm{KO}}) = 0.$$

Corresponding statistics :

$$t = \frac{\hat{\beta}\sqrt{n}}{\hat{s}} \quad \sim t\text{-distribution with } 2(n-1) \text{ degrees of freedom}$$

Do not use any links *between* genes.

limma

Hierarchical Empirical Bayes

Somes prior (not all priors are listed here) :

- $\frac{1}{s^2} = \frac{1}{d_0 s_0^2}\chi_{d_0}$
- $\hat{\beta}|\ \beta, s^2 \sim \mathcal{N}\left(\beta, 2\sigma^2\right)$
- $\hat{s^2}|s^2 \sim \frac{s^2}{d_g}\chi_{d_g}$

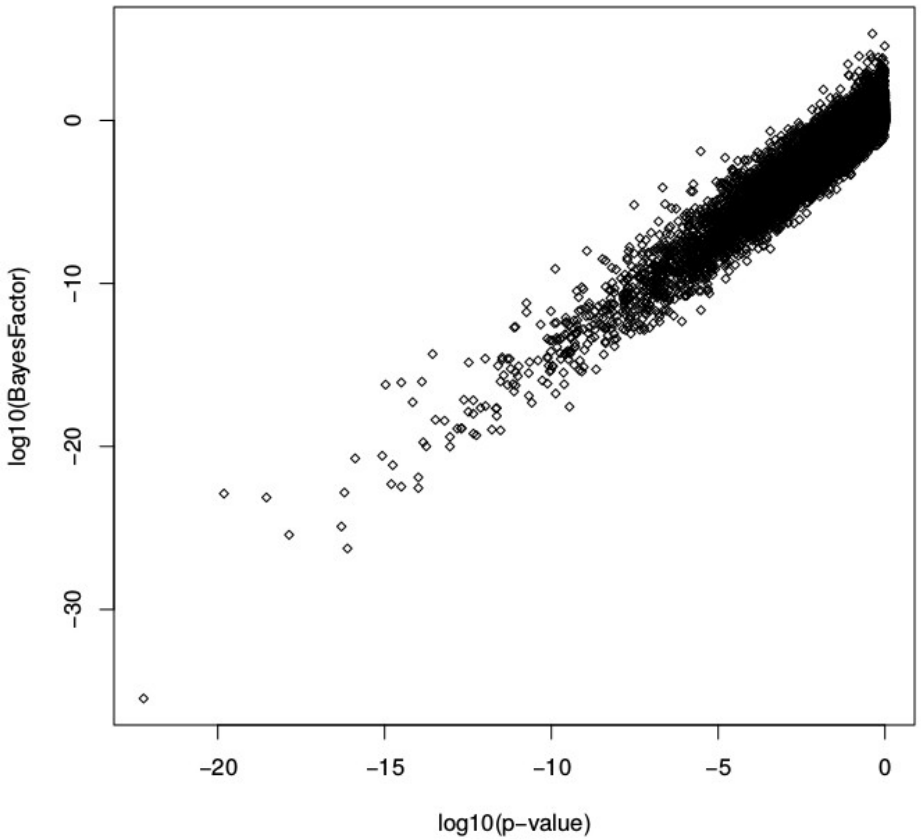
$$\text{Which result to :} \quad \tilde{s} = \frac{d_0 * s_0^2 + d_g * \hat{s}_g^2}{d_0 + d_g},$$

where d_0 and s_0 being some precalculated prior.

moderated t-statistic

$$t = \frac{\hat{\beta}\sqrt{n}}{\tilde{s}} \quad \sim t\text{-distribution with } d_g + d_0 \text{ degrees of freedom}$$

Marginal test



Discussion

Summary - 1

MCMC-Mallows

- MCMC approach can find an posterior ordering based only on data.
- Can handle many interventions in the same time.
- To have good quality for the solution, needs of many interventions.
- MCMC approach is not tractable for more than 100 genes.

Summary - 2

Marginal approach

- Explicit use of do-calculus
- Likelihood derivation
- Simulations that validate this model
- Very related to differential analysis.

Conclusion

What next :

- Parallel-tempering inside MCMC approach
- Differential analysis// Causal analysis
- Integrate several KO
- Interventions models (bypath, knock-down)

Funding : ⋆ ED386 : PhD grant
⋆ INRA : Data (Causality)
Thanks to my supervisors : **Florence Jaffrezic, Andrea Rau** and **Gregory Nuel**.

Thank you !

References I

Amer, P. (2012).

Turning science on robust cattle into improved genetic selection decisions.
Animal, 6(04) :551–556.

Monneret, G., Jaffrézic, F., Rau, A., Zerjal, T., and Nuel, G. (2017).

Identification of marginal causal relationships in gene networks from observational and interventional expression data.
PloS one, 12(3) :e0171142.

Nuel, G., Rau, A., and Jaffrézic, F. (2013).

Joint likelihood calculation for intervention and observational data from a gaussian bayesian network.
arXiv preprint arXiv :1305.0709.

Rau, A., Jaffrézic, F., and Nuel, G. (2013).

Joint estimation of causal effects from observational and intervention gene expression data.
BMC Systems Biology, 7(1) :111.

Smyth, G. K. (2004).

Linear models and empirical bayes methods for assessing differential expression in microarray experiments.
In *Statistical Applications in Genetics and Molecular Biology*, pages 1–25.

Mallows' Proposal

Mallows' Ranking Distribution : with parameter $\phi \in]0,1[$ and reference ordering $\boldsymbol{r}$ is defined by

$$\mathbb{P}(\boldsymbol{o}; \phi, \boldsymbol{r}) = \phi^{d(\boldsymbol{o},\boldsymbol{r})}$$

where $d(\boldsymbol{o}, \boldsymbol{r})$ counts the number of pairwise disagreements.

Properties :

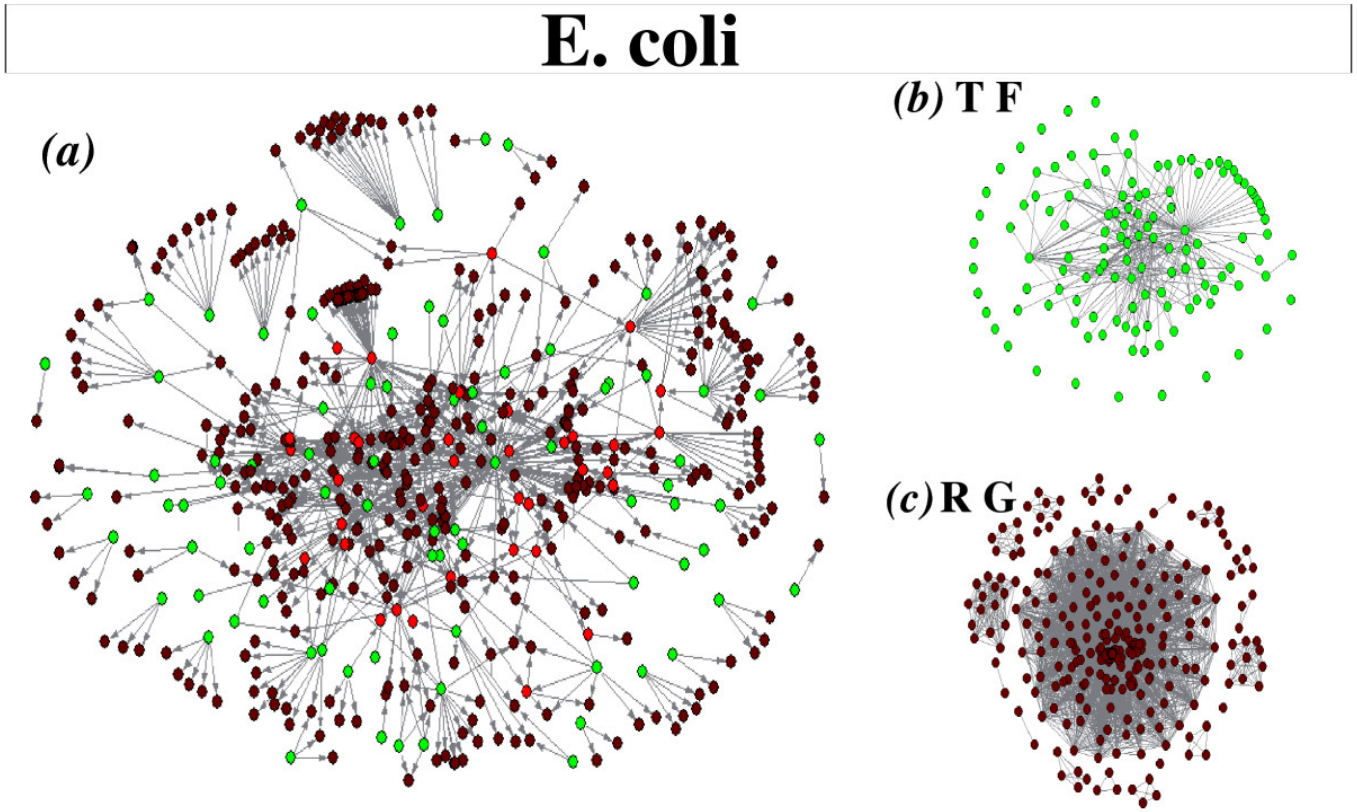
- mode is in $\boldsymbol{r}$
- $\phi \to 0$ corresponds to a dirac distribution
- $\phi \to 1$ corresponds to the uniform distribution
- normalization factor is $1 \times (1+\phi) \times \ldots \times (1+\phi+\ldots+\phi^{p-1})$
- sampling in $O(p)$ with the Repeated Insertion Method

Mallow's distribution in action

$\phi = 0.1$	$\phi = 0.3$	$\phi = 0.6$	$\phi = 0.9$
1 2 4 3 5 1	1 2 3 4 5 2	1 3 4 5 2 1	3 4 2 5 1 1
2 3 4 5 1 3	1 3 4 5 3 1	3 4 5 2 1 5	4 5 3 2 3 2
2 4 5 2 1 3	2 4 5 1 2 3	3 2 4 1 2 3	4 5 1 1 2 3
4 5 1 2 3 4	4 5 1 2 3 5	4 5 4 5 3 1	4 5 2 1 5 3
5 1 2 3 4 5	4 2 1 4 3 5	2 1 3 2 4 5	4 2 4 5 1 3
1 2 3 4 5 1	1 2 4 3 5 1	3 1 5 2 4 1	3 4 2 5 1 4
3 2 5 4 1 2	2 3 4 5 1 2	2 3 5 4 1 2	2 1 3 5 3 4
3 4 5 1 2 4	3 4 5 1 3 4	4 3 5 1 3 4	2 1 5 1 5 3
3 5	5 2	5 2	4 2

Table – Example illustrating ten draws from the Mallows model with a reference ordering of $\boldsymbol{r} = (1\ 2\ 3\ 4\ 5)$ for different temperatures ($\phi = 0.1, 0.3, 0.6, 0.9$).

Examples of real-life GRN

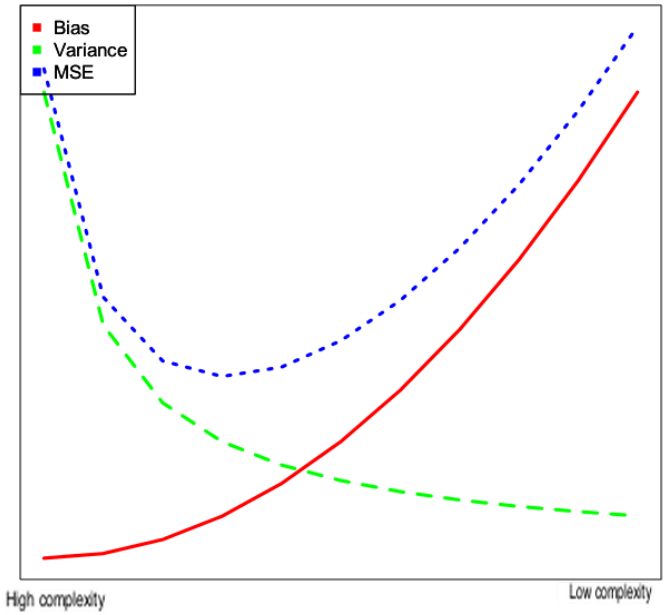


Bias-variance dilemna

We use the MSE to test our model. We can show that :

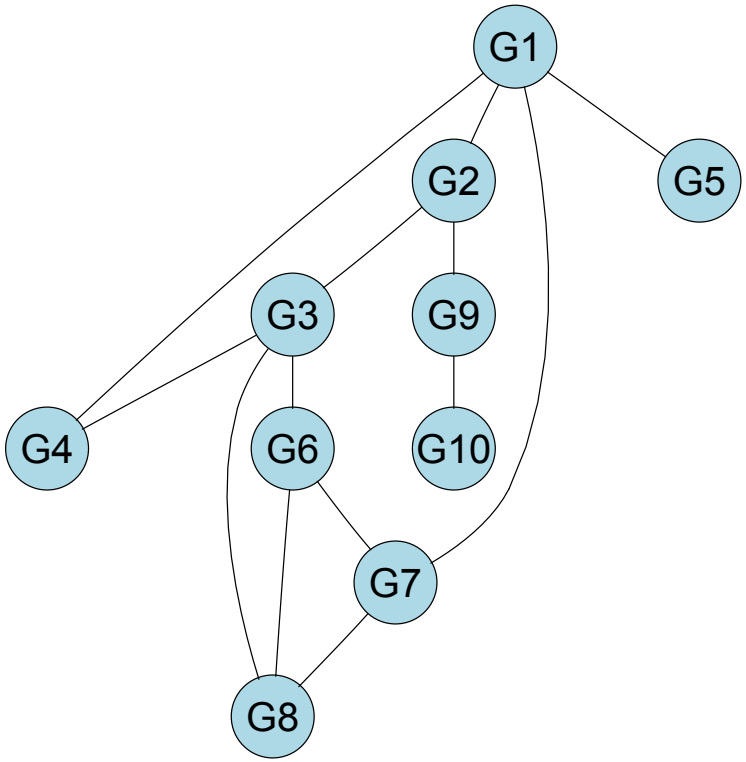
$$\text{MSE} = \text{Cst} + \text{Bias}^2 + \text{Variance}$$

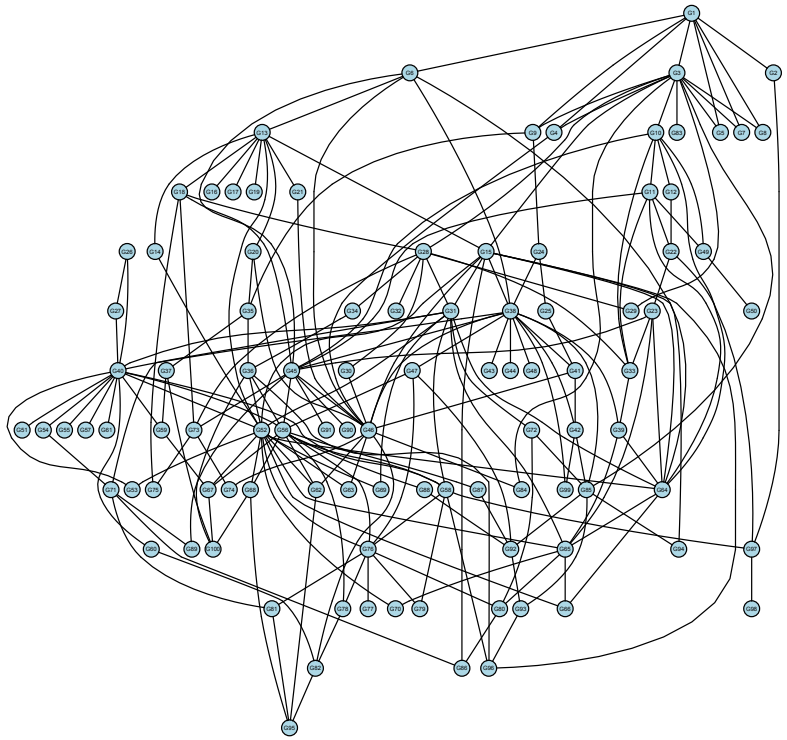
The idea is to put some constraint to decrease greatly the variance of our estimator, even if that increases slightly the bias.



Example of DREAM 10 and 100 Graphs

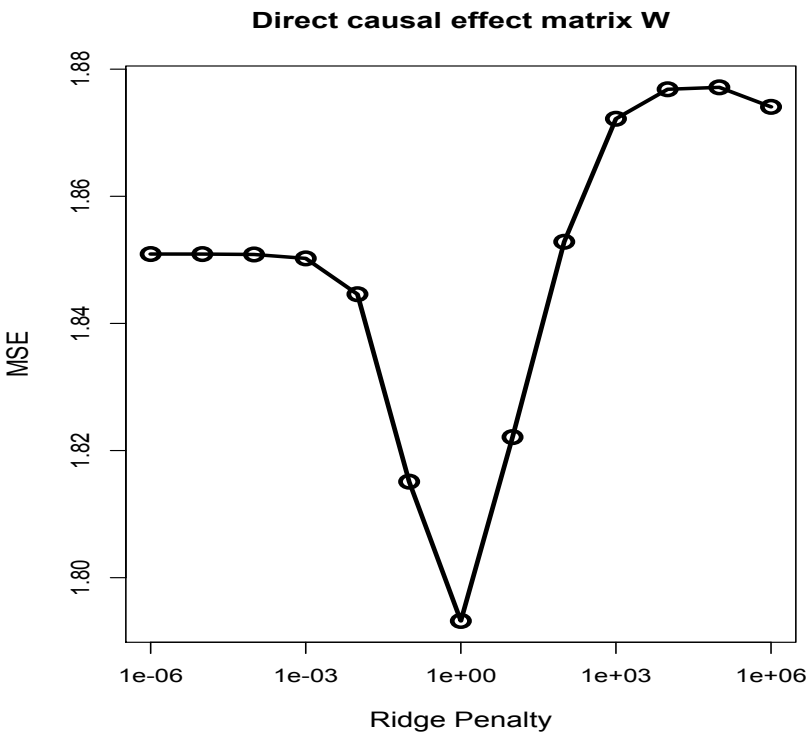
Skeleton coming from the DREAM 4 challenge.
We simulate DAG and data from this skeleton.

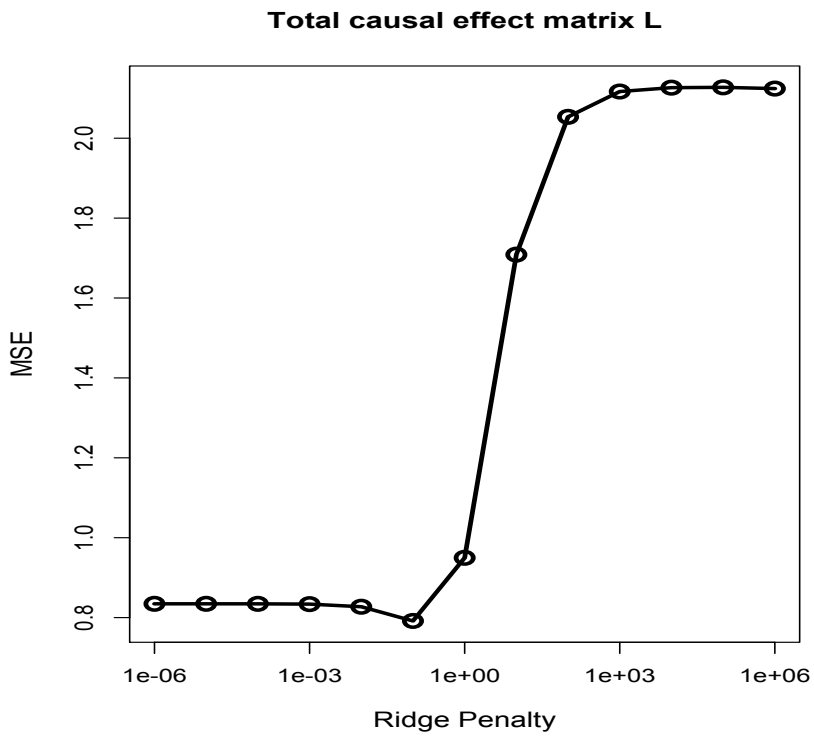




Quality of estimation : W and L

Effect of the constraint on both direct and total causal effect, 10 nodes.





Including Skeleton

If we know the underlying skeleton, we can integrate this information to improve our model. Both speed and quality of the output are improved as shown below :

Figure – Acceptance rate for several temperatures

Temperature	Without skeleton	With skeleton
0.1	0.9995	1.0000
0.3	0.8285	0.8915
0.5	0.5170	0.6430
0.7	0.2910	0.3940
0.9	0.1520	0.2325
1	0.1240	0.2000
10	0.0040	0.0050

Skeleton and speed

From DREAM 10 and 100 graphs, we have more or less 5% of edges.

	Number of equations	Computational time for MLE	Computational time for 50.000 iterations of MCMC
Without skeleton	5050	30s	17 days
With skeleton	250	<1s	2 hours

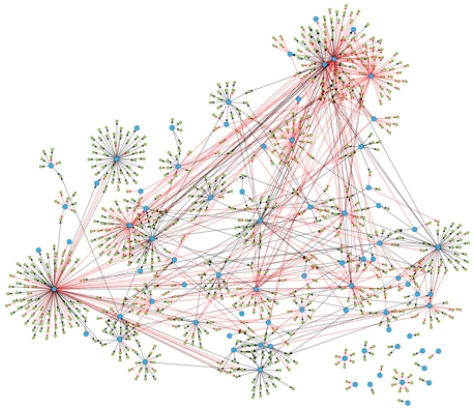
In fact, computation is in order of $\mathcal{O}(p^6)$.

Estimation enhancement with structure

Expectation of estimation quality enhancement.
Fixed DAG of 10 nodes with known order, 100 run with 10 random observations each.

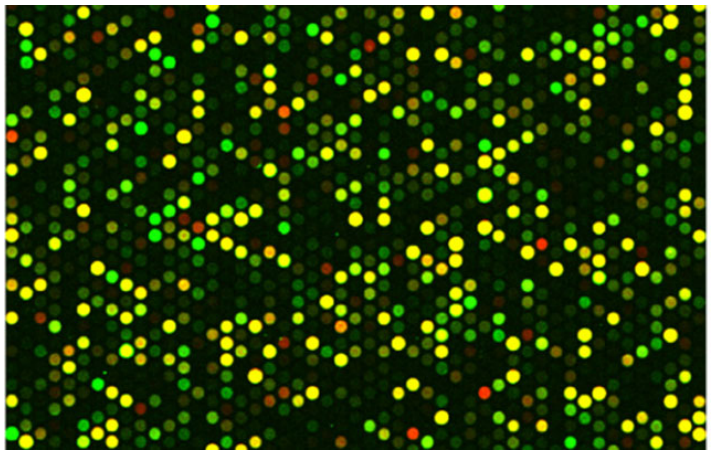
	MSE W	MSE L
Without Skeleton	271.6166	21.18093
With Skeleton	0.1726994	0.02466682

Biological context



← Gene regulation

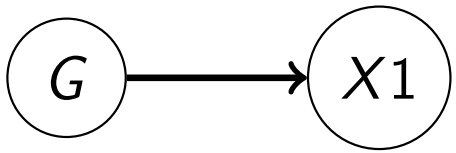
Transcriptomic data →



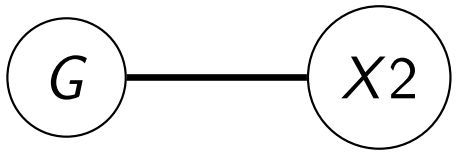
Interventional data (knock-out)

Simulation-1

We simulate two kinds of data : some are causally related (1), and others are only associated (2).

(1) M1 : 

Downstream model

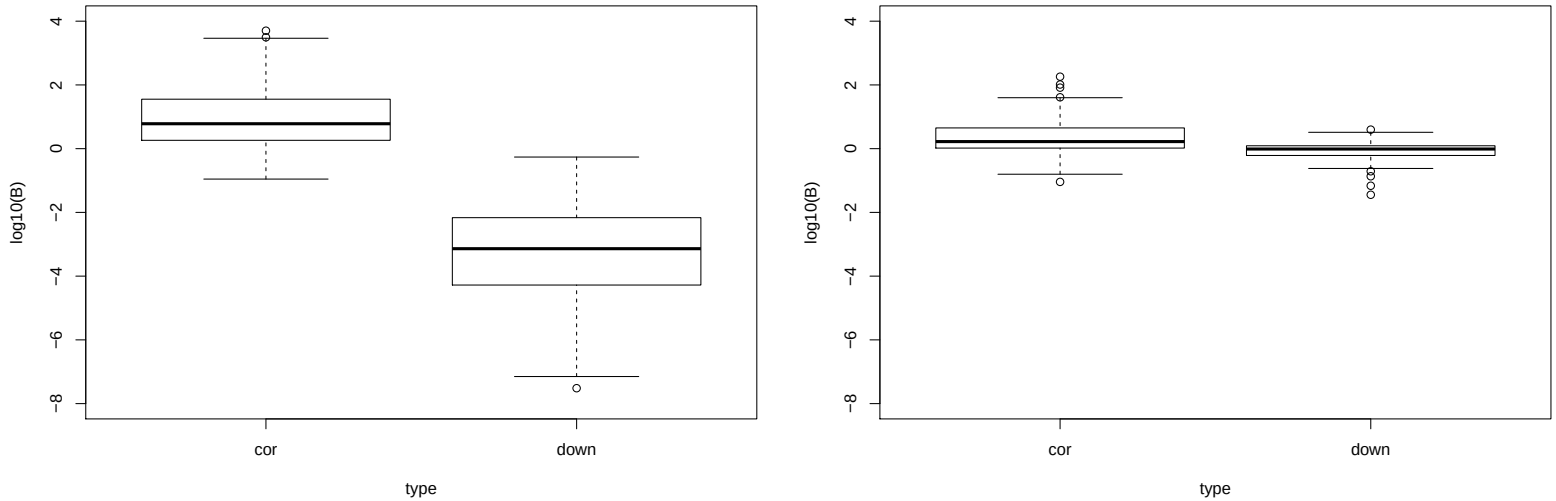
(2) M0 : 

Correlation model

G is knocked-down. 24 observations/24 interventions for each model $\rightarrow$ 96 samples. Repeated 100 times.

Results-1

Figure – Bayes factor to choose between correlation and causality



Parameters : $N = 24, \alpha = 0.3, \mu_1 = 2.0, \mu_2 = -3.1,$
Left : $\sigma_1 = 0.3, \sigma_2 = 0.5,$ Right : $\sigma_1 = 3, \sigma_2 = 5.$

Interactome