Graphical Model Structure Inference Using Trees

Loïc Schwaller, Stéphane Robin, Michael Stumpf

INRA MIA/AgroParisTech
Imperial College London

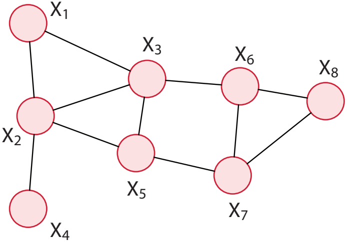
Netbio
September 19 2014

Graphical Models

- $\mathbf{X} = (X_1, ..., X_p)$ random vector

Graphical Models

- $\mathbf{X} = (X_1, ..., X_p)$ random vector
- $G = (V, E_G)$ **undirected** graph with $V = \{1, ..., p\}$



Graphical Models

- $\mathbf{X} = (X_1, ..., X_p)$ random vector
- $G = (V, E_G)$ **undirected** graph with $V = \{1, ..., p\}$

Definition

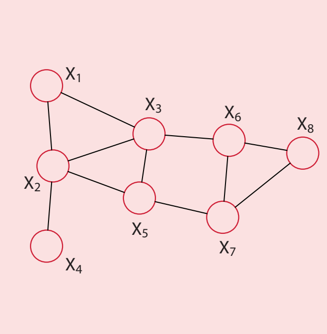
A graphical model following G is a probabilistic model for which the conditional independence structure of $\mathbf{X}$ is given by G.

$$m_G = (G, \mathcal{F}_G)$$

Inference Problem

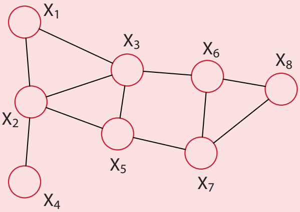
Inference Problem

Graph



Inference Problem

Graph

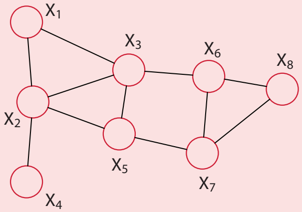


PDF

$$p(X_1, ..., X_8)$$

Inference Problem

Graph



PDF

$$p(X_1, ..., X_8)$$

Sample

$$\mathbf{x}^{(1)} = (x_1^{(1)}, ..., x_8^{(1)})$$
$$\mathbf{x}^{(2)} = (x_1^{(2)}, ..., x_8^{(2)})$$
$$...$$
$$\mathbf{x}^{(n)} = (x_1^{(n)}, ..., x_8^{(n)})$$

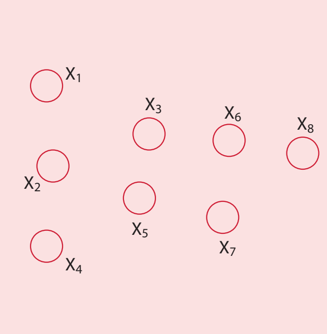
Inference Problem

Sample

$$\mathbf{x}^{(1)} = (x_1^{(1)}, ..., x_8^{(1)})$$
$$\mathbf{x}^{(2)} = (x_1^{(2)}, ..., x_8^{(2)})$$
$$...$$
$$\mathbf{x}^{(n)} = (x_1^{(n)}, ..., x_8^{(n)})$$

Inference Problem

Graph

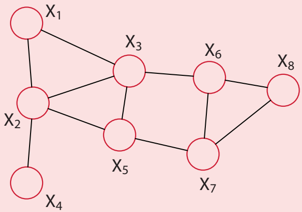


Sample

$$\mathbf{x}^{(1)} = (x_1^{(1)}, \dots, x_8^{(1)})$$
$$\mathbf{x}^{(2)} = (x_1^{(2)}, \dots, x_8^{(2)})$$
$$\dots$$
$$\mathbf{x}^{(n)} = (x_1^{(n)}, \dots, x_8^{(n)})$$

Inference Problem

Graph



Inference

$\longleftarrow$

Sample

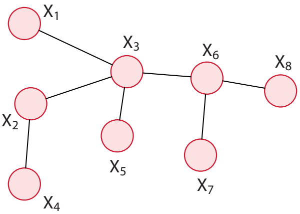
$$\mathbf{x}^{(1)} = (x_1^{(1)}, \ldots, x_8^{(1)})$$
$$\mathbf{x}^{(2)} = (x_1^{(2)}, \ldots, x_8^{(2)})$$
$$\ldots$$
$$\mathbf{x}^{(n)} = (x_1^{(n)}, \ldots, x_8^{(n)})$$

Spanning Trees

Spanning Tree

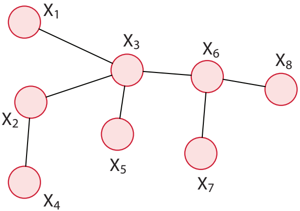
A spanning tree T on the set of vertices V is a connected graph with no cycles.

$$\mathcal{T} = \{\text{spanning trees}\}$$



Tree distribution

- T spanning tree



$$P_T(\mathbf{x}) = \prod_{i \in V} p_i(x_i) \prod_{\{i,j\} \in E_T} \frac{p_{ij}(x_i, x_j)}{p_i(x_i) p_j(x_j)}$$

Chow & Liu Algorithm[1]

- $\mathbf{X} = (X_1, ..., X_p)$ random **discrete** vector
- $D = (\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)})$ i.i.d. sample

[1] C.K. Chow and C.N. Liu. "Approximating Discrete Probability Distributions with Dependence Trees". In: *IEEE Transactions on Information Theory* IT-14.3 (1968), pp. 462–467

Chow & Liu Algorithm[1]

- $\mathbf{X} = (X_1, ..., X_p)$ random **discrete** vector
- $D = (\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)})$ i.i.d. sample

Aim: finding P_T maximizing the likelihood of D.

[1] C.K. Chow and C.N. Liu. "Approximating Discrete Probability Distributions with Dependence Trees". In: *IEEE Transactions on Information Theory* IT-14.3 (1968), pp. 462–467

Chow & Liu Algorithm

- $\mathbf{X} = (X_1, ..., X_p)$ random **discrete** vector
- $D = (\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)})$ i.i.d. sample

Log-Likelihood

$$l_{P_T}(D) = \sum_{k=1}^{n} log\left(P_T(x^{(k)})\right)$$

Chow & Liu Algorithm

- $\mathbf{X} = (X_1, ..., X_p)$ random **discrete** vector
- $D = (\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)})$ i.i.d. sample

T fixed

$$\max_{P_T} l_{P_T}(D) = \max_{P_T} \sum_{k=1}^{n} log\left(P_T(x^{(k)})\right)$$

Chow & Liu Algorithm

- $\mathbf{X} = (X_1, \dots, X_p)$ random **discrete** vector
- $D = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$ i.i.d. sample

T fixed

$$\max_{P_T} l_{P_T}(D) = \max_{P_T} \sum_{k=1}^{n} log\left(P_T(x^{(k)})\right)$$
$$= l_{\widehat{P}_T}(D)$$
$$= \sum_{\{i,j\}\in E_T} \widehat{I}(X_i, X_j) + K$$

- $\widehat{P}_T$ empirical distribution on T
- $\widehat{I}(X_i, X_j)$ empirical mutual information between X_i and X_j.

Maximal Spanning Tree

$$\max_{P_T} l_{P_T}(D) = \sum_{\{i,j\} \in E_T} \widehat{I}(X_i, X_j) + K$$

Maximal Spanning Tree

$$\max_{P_T} l_{P_T}(D) = \sum_{\{i,j\} \in E_T} \widehat{I}(X_i, X_j) + K$$

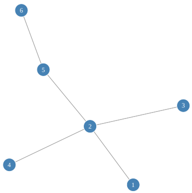
$$\text{edge } \{i,j\} \longleftarrow \widehat{I}(X_i, X_j)$$

$$\begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

Maximal Spanning Tree

$$\max_{P_T} l_{P_T}(D) = \sum_{\{i,j\}\in E_T} \widehat{I}(X_i, X_j) + K$$

$$\text{tree } T \longleftarrow \sum_{\{i,j\}\in E_T} \widehat{I}(X_i, X_j)$$

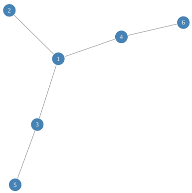


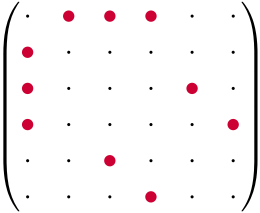
$$\begin{pmatrix} \cdot & \bullet & \cdot & \cdot & \cdot & \cdot \\ \bullet & \cdot & \bullet & \bullet & \bullet & \cdot \\ \cdot & \bullet & \cdot & \cdot & \cdot & \cdot \\ \cdot & \bullet & \cdot & \cdot & \cdot & \cdot \\ \cdot & \bullet & \cdot & \cdot & \cdot & \bullet \\ \cdot & \cdot & \cdot & \cdot & \bullet & \cdot \end{pmatrix}$$

Maximal Spanning Tree

$$\max_{P_T} l_{P_T}(D) = \sum_{\{i,j\} \in E_T} \widehat{I}(X_i, X_j) + K$$

$$\text{tree } T' \longleftarrow \sum_{\{i,j\} \in E_{T'}} \widehat{I}(X_i, X_j)$$





Maximal Spanning Tree

$$\max_{P_T} l_{P_T}(D) = \sum_{\{i,j\} \in E_T} \widehat{I}(X_i, X_j) + K$$

Maximum Spanning Tree Problem[1]

[1] Joseph B. Kruskal. "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem". In: *Proceedings of the American Mathematical Society* 7.1 (Feb. 1956), pp. 48–50

Maximal Spanning Tree

$$\max_{P_T} l_{P_T}(D) = \sum_{\{i,j\} \in E_T} \widehat{I}(X_i, X_j) + K$$

Maximum Spanning Tree Problem[1]

Best Tree T^* $\longrightarrow$ Best Distribution $\widehat{P}_{T^*}$

[1] Joseph B. Kruskal. "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem". In: *Proceedings of the American Mathematical Society* 7.1 (Feb. 1956), pp. 48–50

Algorithm

Input : $\mathbf{D} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$
MaxSpanningTree procedure
Output: best Tree structure T
best Tree distribution P_T

for *each edge* $\{i, j\}$ **do**
  Compute empirical frequencies $\widehat{\gamma_{ij}}$ and $\widehat{\gamma}_i$;
  Compute empiricial MI $\widehat{I}(X_i, X_j)$;
end
$T \longleftarrow MaxSpanningTree(\widehat{I})$;
$P_T \longleftarrow \widehat{P}_T$.

Rationale

- T^* (unknow) underlying tree
- $\mathcal{T} = \{\text{Trees on } V\}$

$$\mathbf{1}_{\{k,l\}}(T) = \begin{cases} 1 & \text{if } \{k,l\} \in E_T \\ 0 & \text{otherwise} \end{cases}$$

Posterior probability of an edge

$$p(\{k,l\} \in E_{T^*} | D) = \sum_{T \in \mathcal{T}} \mathbf{1}_{\{k,l\}}(T) p(T|D)$$

Posterior Distribution on Trees

- $D = (\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)})$ i.i.d. sample

$$P(T|D) \propto P(D|T)P(T)$$

Posterior Distribution on Trees

- $D = (\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)})$ i.i.d. sample

$$P(T|D) \propto P(D|T)P(T)$$

- Prior $P(T)$

$$\mathcal{U}(\mathcal{T})$$

Posterior Distribution on Trees

- $D = (\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)})$ i.i.d. sample

$$P(T|D) \propto P(D|T)P(T)$$

- Prior $P(T)$

$$\mathcal{U}(\mathcal{T})$$

- Marginal Likelihood $P(D|T)$

$$P(D|T) = \int P_T(D)\rho(P_T)dP_T$$

Posterior Distribution on Trees

- $D = (\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)})$ i.i.d. sample

$$P(T|D) \propto P(D|T)P(T)$$

- Prior $P(T)$

$$\mathcal{U}(\mathcal{T})$$

- Marginal Likelihood $P(D|T)$

$$P(D|T) \approx \widehat{P}_T(D)$$

Pseudo-Posterior Distribution on Trees

$$\widehat{P}_T(D) = \exp\left(l_{\widehat{P}_T}(D)\right)$$

Pseudo-Posterior Distribution on Trees

$$\widehat{P}_T(D) = \exp\left(l_{\widehat{P}_T}(D)\right)$$
$$\propto \exp\left(\sum_{\{i,j\}\in E_T} \widehat{I}(X_i, X_j)\right)$$

Pseudo-Posterior Distribution on Trees

$$\begin{aligned}\widehat{P}_T(D) &= \exp\left(l_{\widehat{P}_T}(D)\right) \\ &\propto \exp\left(\sum_{\{i,j\}\in E_T} \widehat{I}(X_i, X_j)\right) \\ &\propto \prod_{\{i,j\}\in E_T} \omega_{ij}(D)\end{aligned}$$

$$\omega_{ij}(D) = \exp\left(\widehat{I}(X_i, X_j)\right)$$

Pseudo-Posterior Distribution on Trees

- $\widehat{P}_T(D) \propto \prod_{\{i,j\} \in E_T} \omega_{ij}(D)$

- $P(T|D) \propto P(D|T)$ (since $P(T) \sim \mathcal{U}(\mathcal{T})$)

Pseudo-Posterior Distribution on Trees

- $\widehat{P}_T(D) \propto \prod_{\{i,j\} \in E_T} \omega_{ij}(D)$
- $P(T|D) \propto P(D|T)$ (since $P(T) \sim \mathcal{U}(\mathcal{T})$)

$$P(T|D) \approx \frac{1}{Z} \prod_{\{i,j\} \in E_T} \omega_{ij}(D)$$

$$Z = \sum_{T \in \mathcal{T}} \prod_{\{i,j\} \in E_T} \omega_{ij}(D)$$

Matrix-Tree Theorem

$\omega = (\omega_{ij})$ symmetric weight matrix $(\forall i, \omega_{ii} = 0)$

Laplacian Matrix

The Laplacian matrix $Q = (Q_{ij})$ relatively to the weights ω is given by

$$Q_{ij} = \begin{cases} -\omega_{ij} & \text{if } i \neq j \\ \sum_j \omega_{ij} & \text{if } i = j \end{cases}$$

Matrix-Tree Theorem

Theorem

Let Q be the Laplacian matrix associated to weights ω. Let $\overline{Q}_{ij}$ denote the $(i,j)^{th}$ minor of Q.

- *All $\overline{Q}_{ij}$ are equal.*
- *The following identity holds*

$$\sum_{T \in \mathcal{T}} \prod_{\{i,j\} \in E_T} \omega_{ij} = \overline{Q}_{ij}$$

Seth Chaiken. *A Combinatorial Proof of the All Minors Matrix Tree Theorem.* 1982

In Practice

$$p(\{k,l\} \in E_{T^*}|D) = \sum_{T \in \mathcal{T}} \mathbf{1}_{\{k,l\}}(T)p(T|D)$$

In Practice

$$p(\{k,l\} \in E_{T^*}|D) = \sum_{T \in \mathcal{T}} \mathbf{1}_{\{k,l\}}(T) p(T|D)$$

$$p(\{k,l\} \in E_{T^*}|D) \approx 1 - \frac{Z^{(kl)}}{Z}$$

$$Z = \sum_{T \in \mathcal{T}} \prod_{\{i,j\} \in E_T} \omega_{ij}(D)$$

$$Z^{(kl)} = \sum_{T \not\ni \{k,l\}} \prod_{\{i,j\} \in E_T} \omega_{ij}(D)$$

In Practice

$$p(\{k,l\} \in E_{T^*}|D) = \sum_{T \in \mathcal{T}} \mathbf{1}_{\{k,l\}}(T) p(T|D)$$

$$p(\{k,l\} \in E_{T^*}|D) \approx 1 - \frac{Z^{(kl)}}{Z}$$

$$\begin{array}{rcl} \omega(D) & \longrightarrow & Z \\ \omega^{(kl)}(D) & \longrightarrow & Z^{(kl)} \end{array}$$

$$\omega_{ij}^{(kl)}(D) = \begin{cases} 0 & \text{if } \{i,j\} = \{k,l\} \\ \omega_{ij}(D) & \text{otherwise} \end{cases}$$

Algorithm

Input : $\mathbf{D} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$
MatTree procedure
Output: edgewise score matrix α

for *each edge* $\{i, j\}$ **do**
  Compute empirical frequencies $\widehat{\gamma_{ij}}$ and $\widehat{\gamma}_i$;
  Compute empiricial MI $\widehat{I}(X_i, X_j)$;
end
Compute ω from $\widehat{I}$;
$Z \longleftarrow MatTree(\omega)$;
for *each edge* $\{i, j\}$ **do**
  $Z^{(ij)} \longleftarrow MatTree(\omega^{(ij)})$;
  $\alpha_{ij} \longleftarrow 1 - \frac{Z^{(ij)}}{Z}$.
end

Remarks

Real Posterior Distribution

$$P(T|\mathbf{x}) \propto P(\mathbf{x}|T)P(T)$$

- Prior $P(T)$

$$\mathcal{U}(\mathcal{T})$$

- Marginal Likelihood $P(\mathbf{x}|T)$

$$P(\mathbf{x}|T) = \int P_T(\mathbf{x})\rho_T(P_T)dP_T$$

Remarks

Real Posterior Distribution

- Particular choice of prior ρ_T
  - Strong Compatible Hyper Markov prior

$$P(\mathbf{x}|T) = \prod_{i \in V} f_i(x_i) \prod_{\{i,j\} \in E_T} \frac{f_{ij}(x_i, x_j)}{f_i(x_i) f_j(x_j)}$$

$$f_{ij}(x_i, x_j) = \int P_{ij}(x_i, x_j) \rho_{ij}(P_{ij}) dP_{ij}$$

$$f_i(x_i) = \int P_i(x_i) \rho_i(P_i) dP_i$$

Remarks

Real Posterior Distribution

- Particular choice of prior ρ_T
  - Strong Compatible Hyper Markov prior

$$P(\mathbf{x}|T) \propto \prod_{\{i,j\}\in E_T} \omega_{ij}(x_i, x_j)$$

$$\omega_{ij}(x_i, x_j) = \frac{f_{ij}(x_i, x_j)}{f_i(x_i) f_j(x_j)}$$

Remarks

Real Posterior Distribution

- Particular choice of prior ρ_T
  - Strong Compatible Hyper Markov prior
- Gaussian distribution with Normal-Wishart prior
- Multinomial distribution with Dirichlet prior

Simulation Scheme

- Choice of a size ($p = 25$) and topology (*Tree, Hub, Erdös-Rényi*)
- $\mathbf{A} \longleftarrow$ undirected adjacency matrix
- $\mathbf{\Lambda} \longleftarrow$ precision matrix
  - $\Lambda \longleftarrow -\mathbf{A} + 3I_p$
- $\mathbf{\Sigma} \longleftarrow \mathbf{\Lambda}^{-1}$

$$y^{(i)} \sim \mathcal{N}(0, \Sigma) \text{ i.i.d.}$$
$$x^{(i)} \longleftarrow \mathit{discretize}(y^{(i)}, d = 5)$$

$i = 1, ..., n$
$n = 25, 50, 75, 100$
10 repetitions per sample size

Chow & Liu VS Pseudo-posterior

- Chow & Liu
  - Tree structure with $(p-1)$ edges
- Pseudo-posterior on Tree
  - $(p-1)$ most probable edges

Assessment

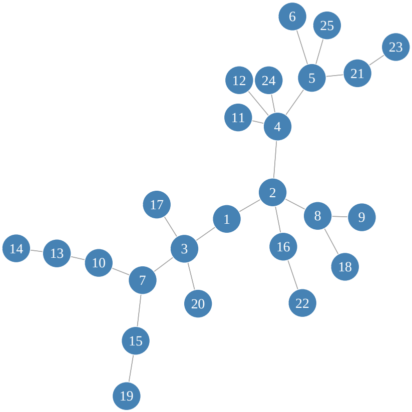
True Positive Rate (TPR)
for $(p-1)$ edges

$$TPR = \frac{TP}{P}$$

False Positive Rate (FPR)
for $(p-1)$ edges

$$FPR = \frac{FP}{N}$$

Tree



Tree

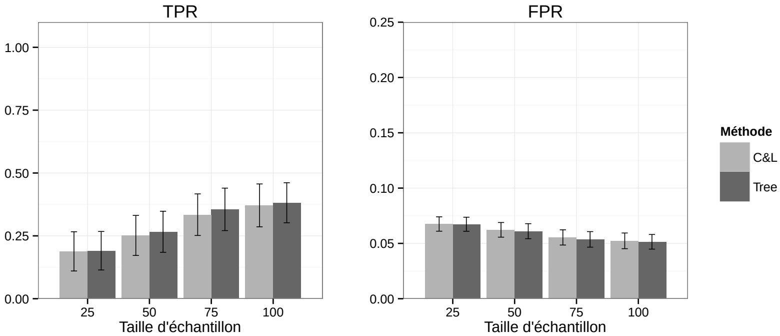
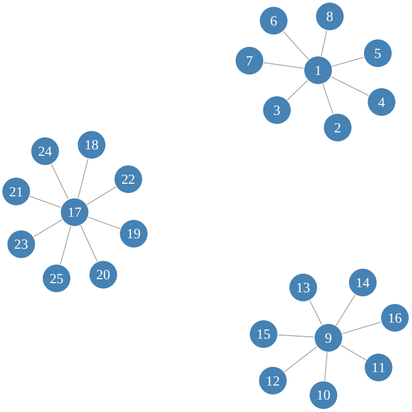


Figure: 25 Nodes.

Hub



Hub

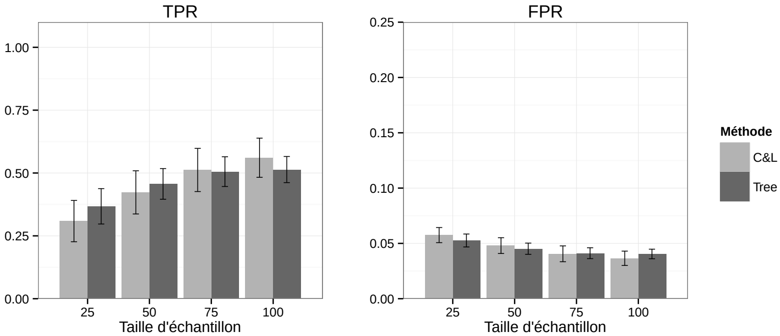
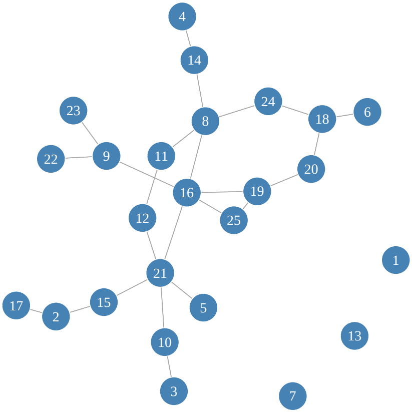


Figure: 25 Nodes.

Erdös-Rényi



Erdös-Rényi

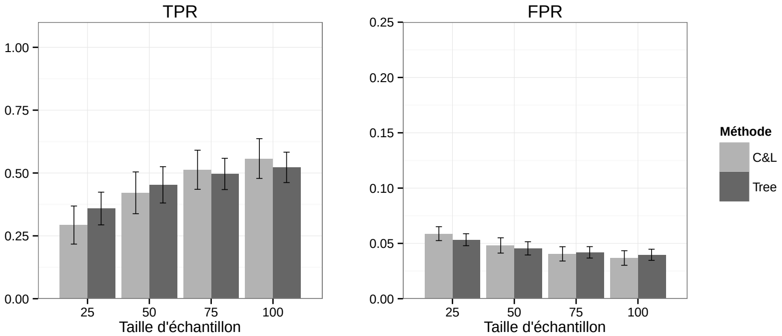


Figure: 25 Nodes, Connection Probability 2/p

Inference Methods

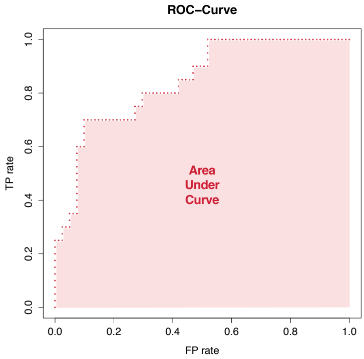
- Partial Order MCMC[3]
  - Structure sampling

- Pseudo-posterior on Trees

[3] Teppo Niinimaki, Pekka Parviainen, and Mikko Koivisto. "Partial Order MCMC for Structure Discovery in Bayesian Networks." In: *UAI*. ed. by Fabio Gagliardi Cozman and Avi Pfeffer. 2011

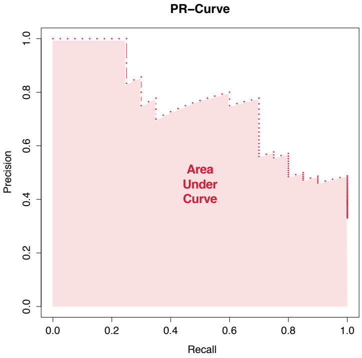
Assessment

ROC Curve



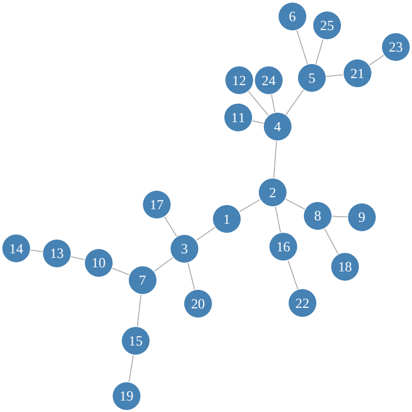
$$TPR = \frac{TP}{P} \qquad FPR = \frac{FP}{N}$$

PR Curve



$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{P}$$

Tree



Tree

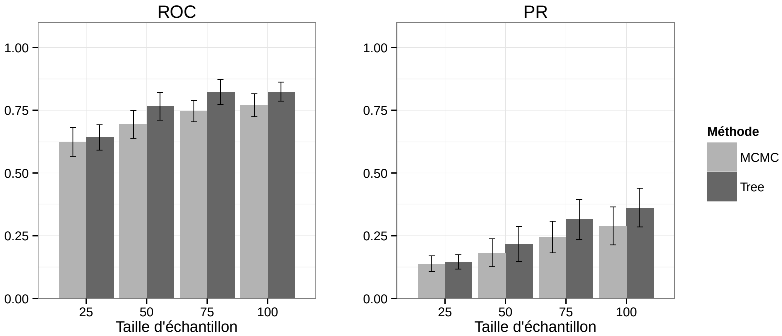
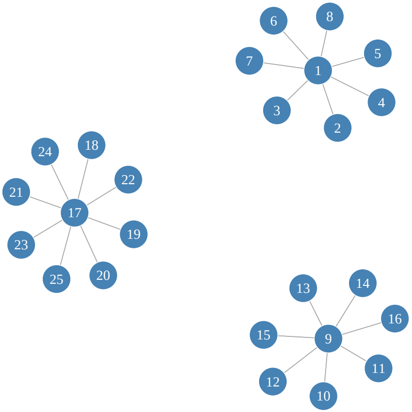


Figure: 25 Nodes.

Hub



Hub

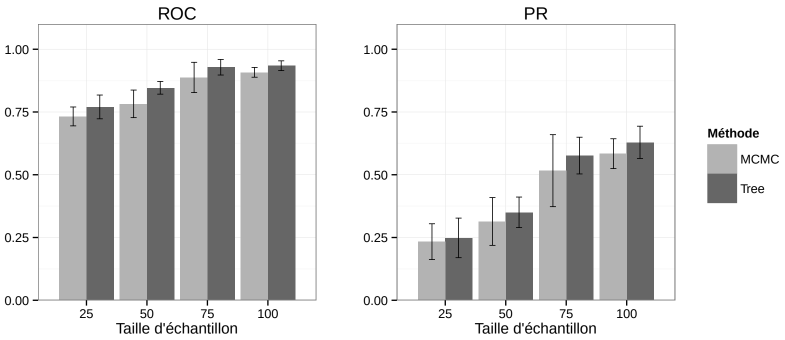
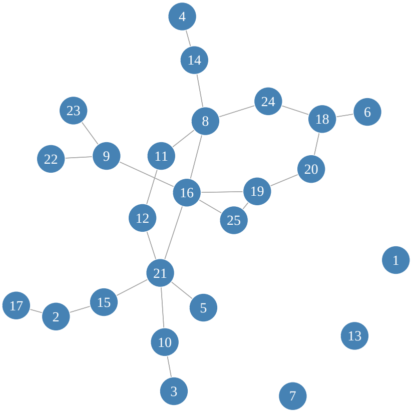


Figure: 25 Nodes.

Erdös-Rényi



Erdös-Rényi

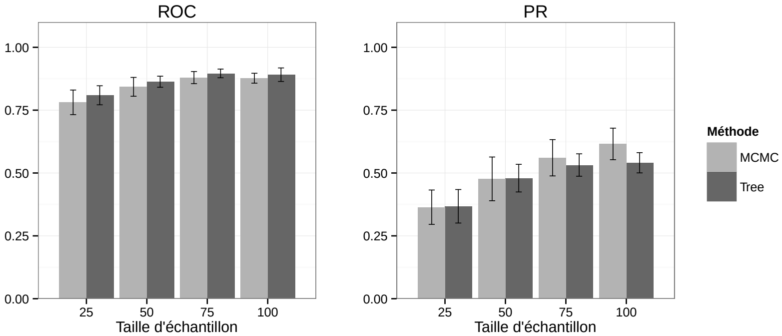
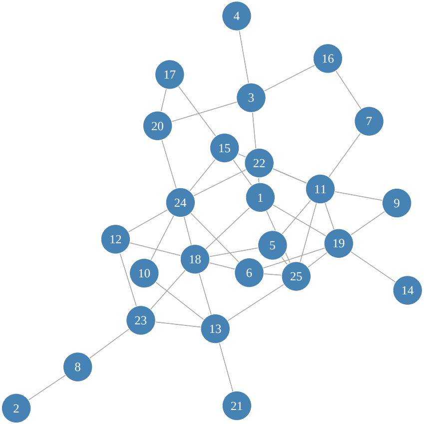


Figure: 25 Nodes, Connection Probability 2/p

Erdös-Rényi



Erdös-Rényi

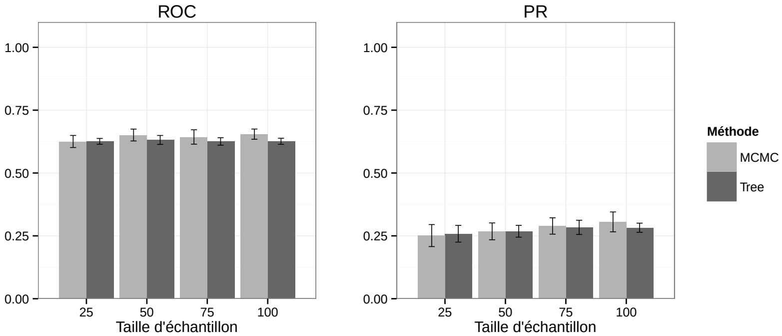


Figure: 25 Nodes, Connection Probability 4/p

Running Time

Network Size	MCMC	Tree
p=25	17 s	0.6 s
p=50	317.5 s	2.8 s
p=75	1913.7 s	7.6 s

Figure: Running Time for MCMC & Tree inference. Sample of size $n = 100$, $d = 10$ levels per variable.

RAF Network[4]

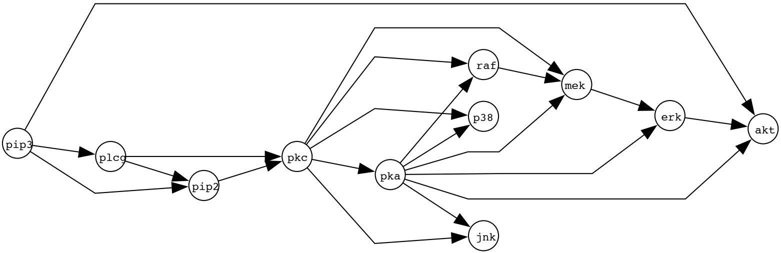


Figure: Cellular signalling network describing the interactions of 11 phosphorylated proteins and phospholipids in human immune cells.

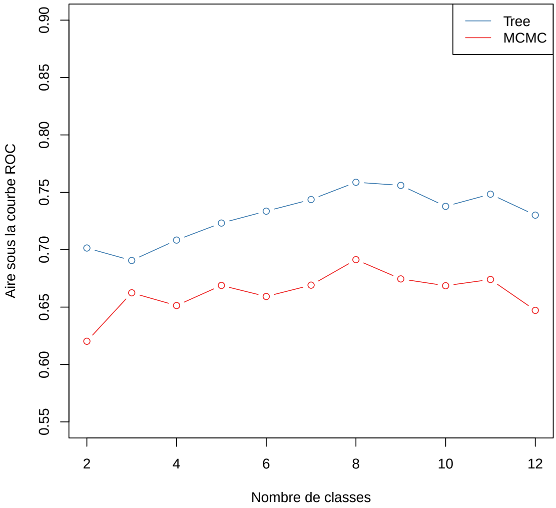
[4] Adriano V. Werhli, Marco Grzegorczyk, and Dirk Husmeier. *Comparative Evaluation of Reverse Engineering Gene Regulatory Networks with Relevance Networks, Graphical Gaussian Models and Bayesian Networks*. 2006.

RAF Network

- 5 samples of size n=100
- Continuous data, discretized at d=10

	MCMC	Tree
ROC	0.67	0.74
PR	0.59	0.63

RAF Network


Tree
MCMC
Aire sous la courbe ROC
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
2
4
6
8
10
12
Nombre de classes

Conclusion

- Build on the work of Chow & Liu
- Algebraic theorem to compute the pseudo-posterior on Trees
  - Matrix-Tree theorem
- Broad Framework
  - Pseudo-posterior
  - Posterior

Perspectives

- General method
  - Posterior probability of an edge
  - Posterior probability of more complexe motifs (fork, chain, etc)
  - Network Comparison

Seth Chaiken. *A Combinatorial Proof of the All Minors Matrix Tree Theorem*. 1982.

C.K. Chow and C.N. Liu. "Approximating Discrete Probability Distributions with Dependence Trees". In: *IEEE Transactions on Information Theory* IT-14.3 (1968), pp. 462–467.

Joseph B. Kruskal. "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem". In: *Proceedings of the American Mathematical Society* 7.1 (Feb. 1956), pp. 48–50.

Teppo Niinimaki, Pekka Parviainen, and Mikko Koivisto. "Partial Order MCMC for Structure Discovery in Bayesian Networks." In: *UAI*. Ed. by Fabio Gagliardi Cozman and Avi Pfeffer. 2011.

Adriano V. Werhli, Marco Grzegorczyk, and Dirk Husmeier. *Comparative Evaluation of Reverse Engineering Gene Regulatory Networks with Relevance Networks, Graphical Gaussian Models and Bayesian Networks*. 2006.