Stochastic processes analysis for Genomics: MTD model and Dynamic Bayesian Network inference.

Sophie Lèbre

-

Thesis defense - Supervisor Bernard Prum

-

14 September 2007

Université d'Évry-Val-d'Essonne - Laboratoire Statistique et Génome

Outline

1. Sequence modeling: an EM algorithm for MTD models
2. Genetic network: inferring DBN with partial order dependence
3. Inferring non-homogneous DBN with rjMCMC

DNA sequences: MTD modeling?

$\rightsquigarrow$ in collaboration with Pierre-Yves Bourguignon

Let $\mathbf{Y} = Y_1 \dots Y_n$ be a random sequence in $\mathcal{Y}$, $\quad |\mathcal{Y}| = q$,

- m^{th}-order Markov model,

$$\forall t > m, \quad \mathbb{P}(Y_t | \mathbf{Y}_1^{t-1}) = \mathbb{P}(Y_t | \mathbf{Y}_{t-m}^{t-1}).$$

- *Mixture Transition Distribution* model (Raftery, 1985)

$$\begin{aligned} \mathbb{P}(Y_t | \mathbf{Y}_1^{t-1}) &= \sum_{g=1}^{m} \varphi_g \ \mathbb{P}(Y_t | Y_{t-g}), \\ &= \sum_{g=1}^{m} \varphi_g \ \boldsymbol{\pi}_g(y_{t-g}, y_t). \end{aligned}$$

with $\varphi_g > 0$, $\quad \sum_{g=1}^{m} \varphi_g = 1$ and π_g stochastic matrices.

MTD model: very parsimonious but estimation?

- Number of independent parameters:

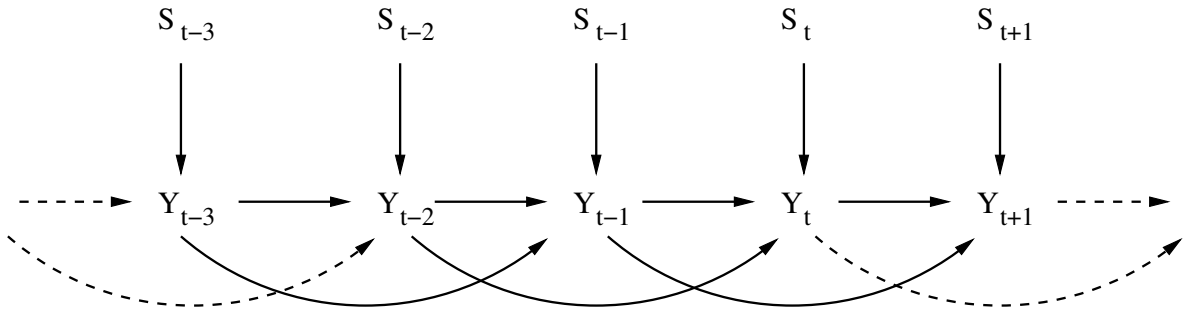
Full Markov Model vs MTD

$\prod_{[q^m \times q]}$ $(\varphi_g, \pi_{g[q \times q]})_{g=1..m}$

Order m	Full MM	MTD
1	12	12
2	48	25
3	192	38
4	**768**	**51**
5	**3 072**	**64**

- No expression of the Maximum Likelihood Estimate
  - $\rightsquigarrow$ Estimation with constraints (Berchtold, 2001)
  - $\rightsquigarrow$ Drawbacks

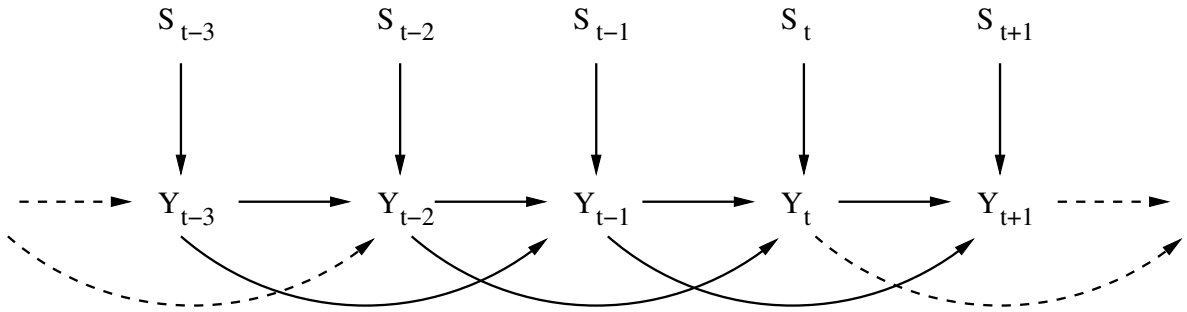
Estimation ?



- A mixture model: hidden process S_t
  $\forall\ 1 \leq g \leq m$,

$$\begin{aligned} \mathbb{P}(S_t = g) &= \varphi_g \\ \mathbb{P}(Y_t | S_t = g, \mathbf{Y}_{t-m}^{t-1}) &= \pi_g(y_{t-g}, y_t) \end{aligned}$$

Estimation ?



- A mixture model: hidden process S_t
  $\forall\ 1 \leq g \leq m$,

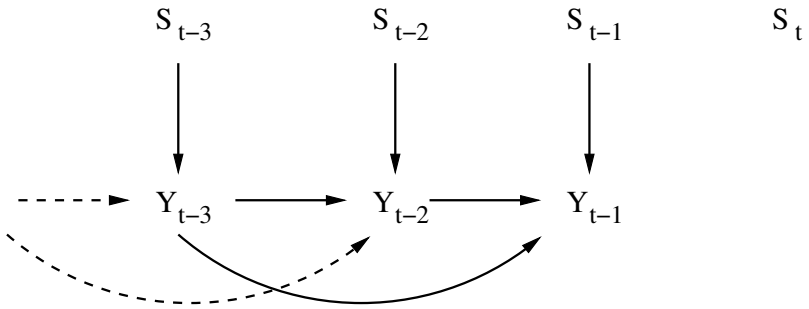
$$\begin{aligned} \mathbb{P}(S_t = g) &= \varphi_g \\ \mathbb{P}(Y_t | S_t = g, \mathbf{Y}_{t-m}^{t-1}) &= \pi_g(y_{t-g}, y_t) \end{aligned}$$

$\rightsquigarrow$ EM algorithm

- E-Step: compute $\mathbb{P}(S_t = g | Y, \theta)$

$$\begin{aligned} &= \mathbb{P}(S_t = g | Y_{t-m}^{t}, \theta) \\ &= \frac{\mathbb{P}(Y_t | S_t = g, \mathbf{Y}_{t-m}^{t-1}, \theta)\ \mathbb{P}(S_t = g | \mathbf{Y}_{t-m}^{t-1}, \theta)}{\sum_{l=1}^{m} \mathbb{P}(Y_t | S_t = l, \mathbf{Y}_{t-m}^{t-1}, \theta)\ \mathbb{P}(S_t = l | \mathbf{Y}_{t-m}^{t-1}, \theta)}. \end{aligned}$$

Estimation ?



- A mixture model: hidden process S_t
  $\forall\ 1 \leq g \leq m$,

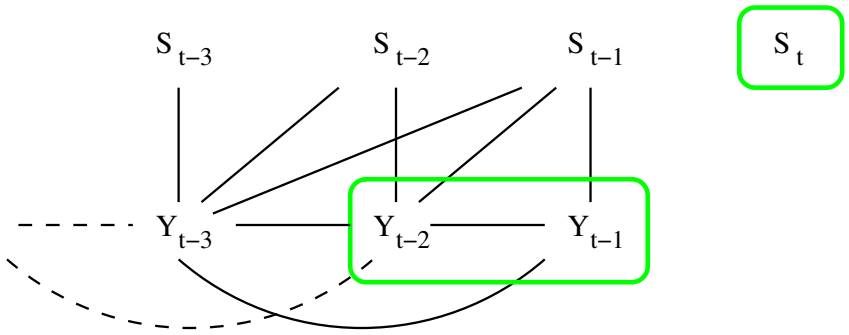
$$\begin{aligned} \mathbb{P}(S_t = g) &= \varphi_g \\ \mathbb{P}(Y_t | S_t = g, \mathbf{Y}_{t-m}^{t-1}) &= \pi_g(y_{t-g}, y_t) \end{aligned}$$

$\rightsquigarrow$ EM algorithm

- E-Step: compute $\mathbb{P}(S_t = g | Y, \theta)$

$$\begin{aligned} &= \mathbb{P}(S_t = g | Y_{t-m}^{t}, \theta) \\ &= \frac{\mathbb{P}(Y_t | S_t = g, \mathbf{Y}_{t-m}^{t-1}, \theta)\ \mathbb{P}(S_t = g | \mathbf{Y}_{t-m}^{t-1}, \theta)}{\sum_{l=1}^{m} \mathbb{P}(Y_t | S_t = l, \mathbf{Y}_{t-m}^{t-1}, \theta)\ \mathbb{P}(S_t = l | \mathbf{Y}_{t-m}^{t-1}, \theta)}. \end{aligned}$$

Estimation ?



- A mixture model: hidden process S_t
  $\forall\ 1 \leq g \leq m$,

$$\begin{aligned} \mathbb{P}(S_t = g) &= \varphi_g \\ \mathbb{P}(Y_t | S_t = g, \mathbf{Y}_{t-m}^{t-1}) &= \pi_g(y_{t-g}, y_t) \end{aligned}$$

$\rightsquigarrow$ EM algorithm

- E-Step: compute $\mathbb{P}(S_t = g | Y, \theta)$

$$\begin{aligned} &= \mathbb{P}(S_t = g | Y_{t-m}^{t}, \theta) \\ &= \frac{\mathbb{P}(Y_t | S_t = g, \mathbf{Y}_{t-m}^{t-1}, \theta)\ \mathbb{P}(S_t = g | \mathbf{Y}_{t-m}^{t-1}, \theta)}{\sum_{l=1}^{m} \mathbb{P}(Y_t | S_t = l, \mathbf{Y}_{t-m}^{t-1}, \theta)\ \mathbb{P}(S_t = l | \mathbf{Y}_{t-m}^{t-1}, \theta)}. \end{aligned}$$

EM algorithm - k^{th} iteration

$\forall g \in \{1, ..., m\}, \forall\ i_m, ..., i_1, i_0 \in \mathcal{Y}$,

- E-Step:

$$\mathbb{P}_S^{(k)}(g|\mathbf{i}_m^0) = \frac{\varphi_g^{(k)}\pi_g^{(k)}(i_g, i_0)}{\sum_{l=1}^{m}\varphi_l^{(k)}\pi_l^{(k)}(i_l, i_0)}.$$

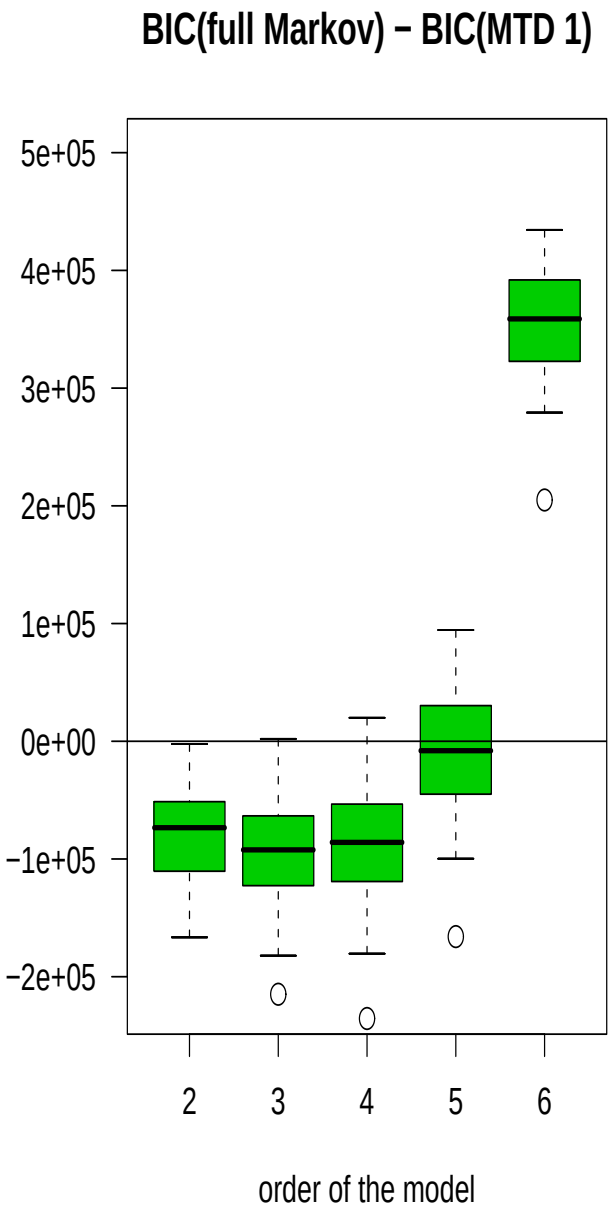
- M-Step:

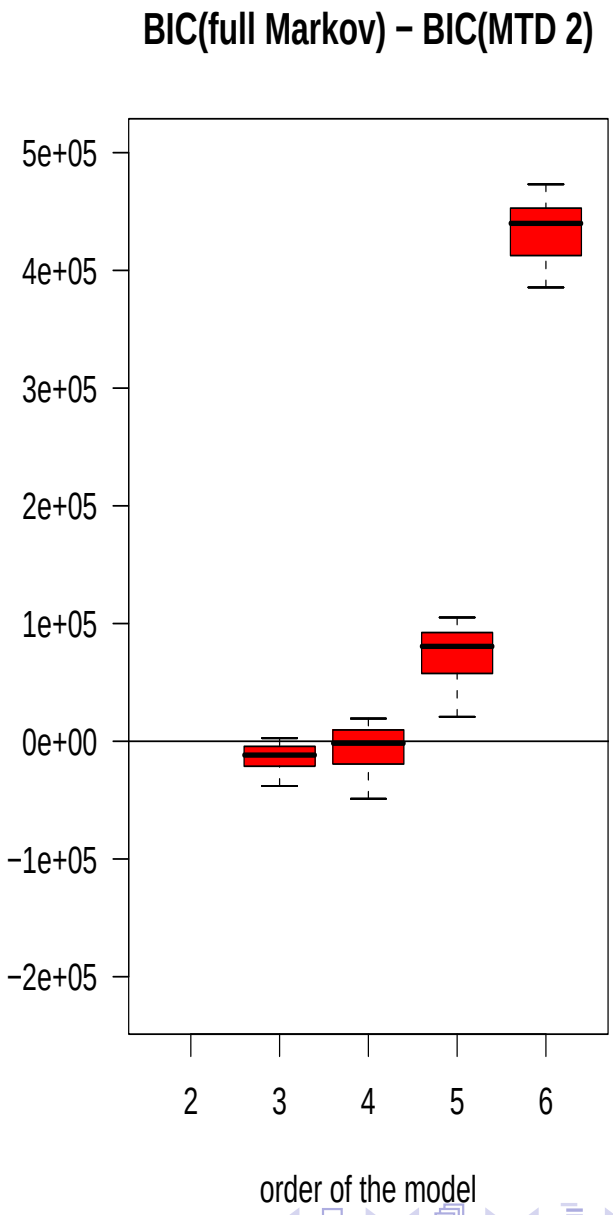
$$\varphi_g^{(k+1)} = \frac{1}{n-m}\sum_{i_m \ldots i_0}\mathbb{P}^{(k)}(g|\mathbf{i}_m^0)N(\mathbf{i}_m^0)$$

$$\pi_g^{(k+1)}(i,j) = \frac{\sum_{i_m \ldots i_{g+1} i_{g-1} \ldots i_1}\mathbb{P}^{(k)}(g|\mathbf{i}_m^{g+1} i \mathbf{i}_{g-1}^1 j)N(\mathbf{i}_m^{g+1} i \mathbf{i}_{g-1}^1 j)}{\sum_{i_m \ldots i_{g+1} i_{g-1} \ldots i_1 i_0}\mathbb{P}^{(k)}(g|\mathbf{i}_m^{g+1} i \mathbf{i}_{g-1}^0)N(\mathbf{i}_m^{g+1} i \mathbf{i}_{g-1}^0)}$$

$\rightsquigarrow$ available in seq++ (Vincent Miele)

Application to bacteria coding DNA sequences





MTD model

- New inference procedure,

  easy to use and avalaible in seq++

- Improved goodness of fit when $m >>$

  $\rightsquigarrow$ annotation, gene detection... (HMM)

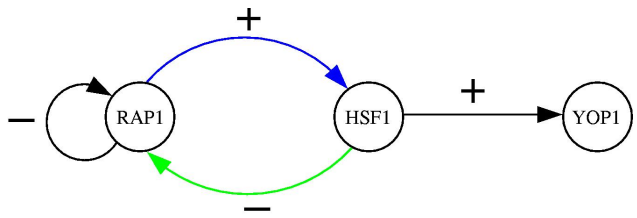
- Article: *An EM algorithm for estimation in the MTD Model.* Lèbre S. and Bourguignon, P-Y., to appear in the *Journal of Statistical Computation and Simulation*.

Outline

1. Sequence modeling: an EM algorithm for MTD models
2. Genetic network: inferring DBN with partial order dependence
3. Inferring non-homogneous DBN with rjMCMC

Genes functions?

- Recover cellular regulations:

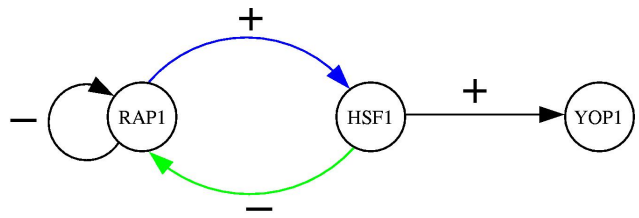


  - up/down regulation
  - retroaction, feedforwards loops...

$\Rightarrow$ Complex dynamic system

Genes functions?

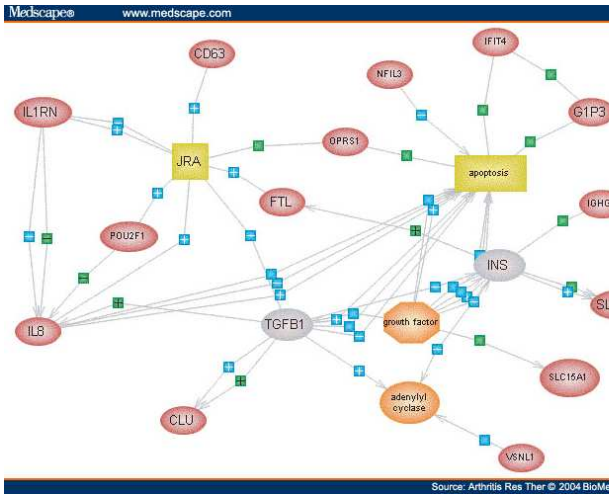
- Recover cellular regulations:



  - up/down regulation
  - retroaction, feedforwards loops...

$\Rightarrow$ Complex dynamic system

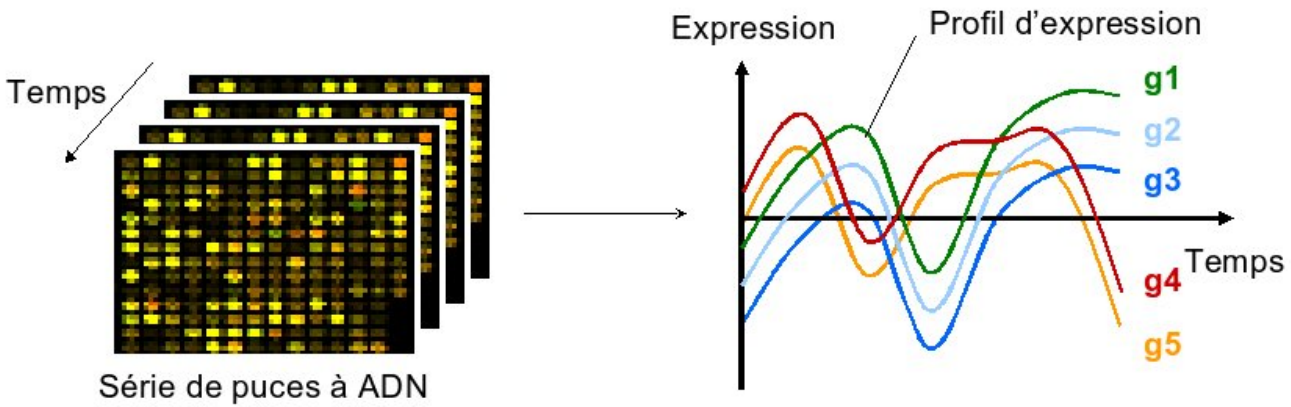
- Objective: identify this organisation in large scale.



$\Rightarrow$ Regulation networks

Temporal gene expression data

- Microarrays: **simultaneous** expression of **several thousands** of genes.



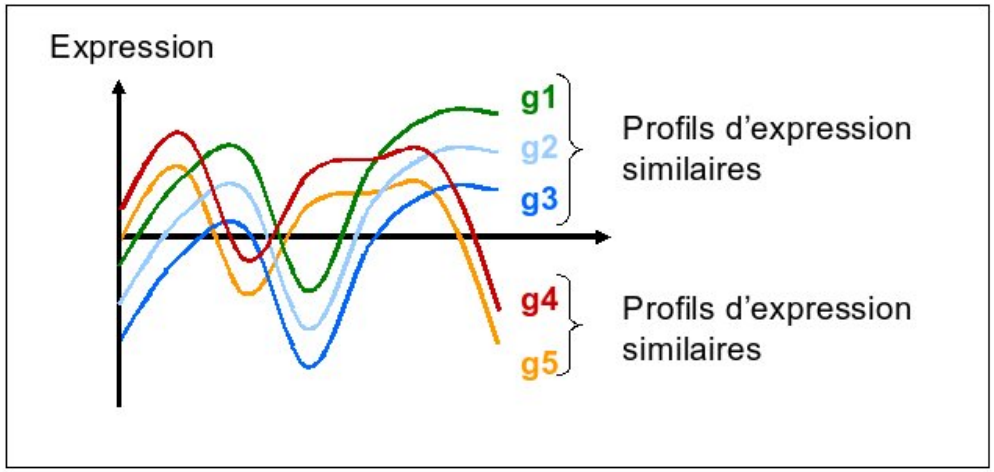
- Notations:
  Stochastic process $X = \{X_t^i;\ \forall i \in \{1, ..., p\}, \forall t \in \{1, ..., n\}\}$.
- X_t^i expression of gene i at time t,

What information extracting from expression profiles?

$\Rightarrow$ Study the interactions between genes.

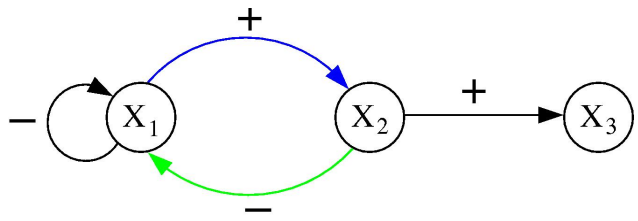
- identify coexpressed genes

  $\rightsquigarrow$ coregulated genes? $\rightsquigarrow$ same biological process?



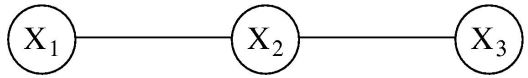
- which genes directly interact? 2 main objectives:
  - modeling temporal dependencies,
  - carrying inference when $n << p$.

How to model biological motifs ?



- A biological motif
- Gaussian Graphical Modeling

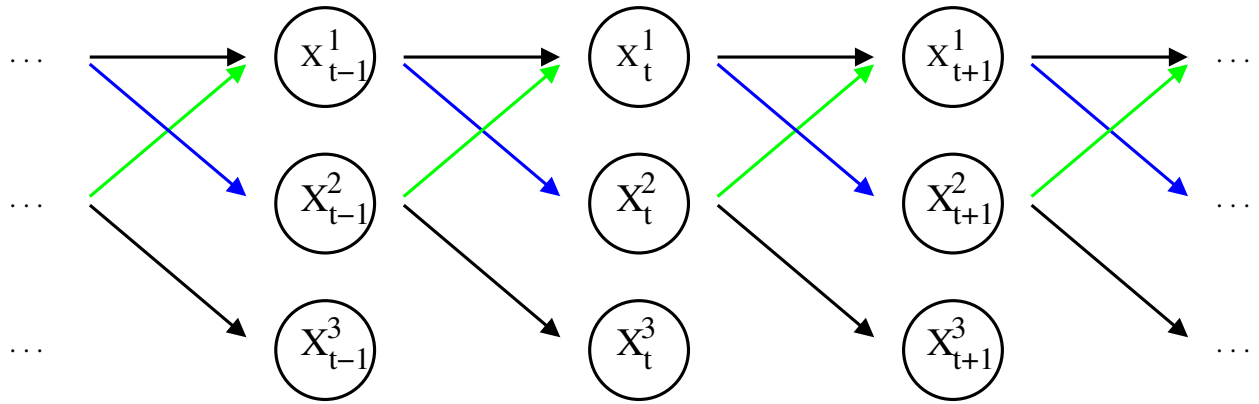
  $\longrightarrow$ Concentration graph



(Toh et al. 2002, Wang et al. 2003, Schäfer and Strimmer 2005)

- Bayesian Networks

  $\longrightarrow$ Dynamic: allows to model cycles!



Dynamic Bayesian Networks

Introduced by Murphy and Mian (1999) to model gene expression time series.

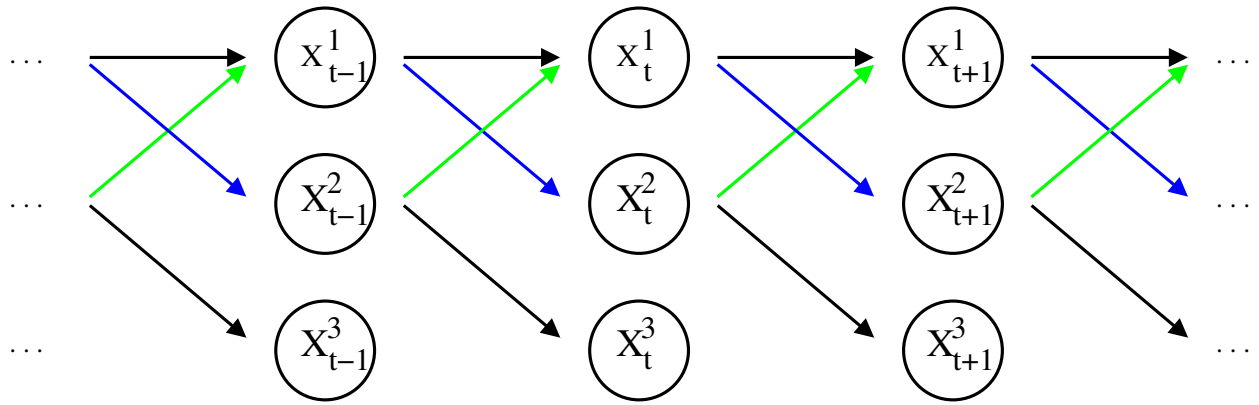
- Discrete DBN
  (Ong et al. 2002)
- HMM, State Space Models
  (Perrin et al. 2003, Beal et al. 2005)
- Non parametric additive models
  (Kim, Imoto and Miyano, 2004)

DBN modeling

Assumptions

- $(\mathcal{A}_1)$ *X* 1^{st} *order Markov process*
- $(\mathcal{A}_2)$ *'simultaneous independence' given the past,*

$$\forall t > 1, \forall i, j \in N, \quad X_t^i \perp\!\!\!\perp X_t^j \mid X_{t-1}.$$

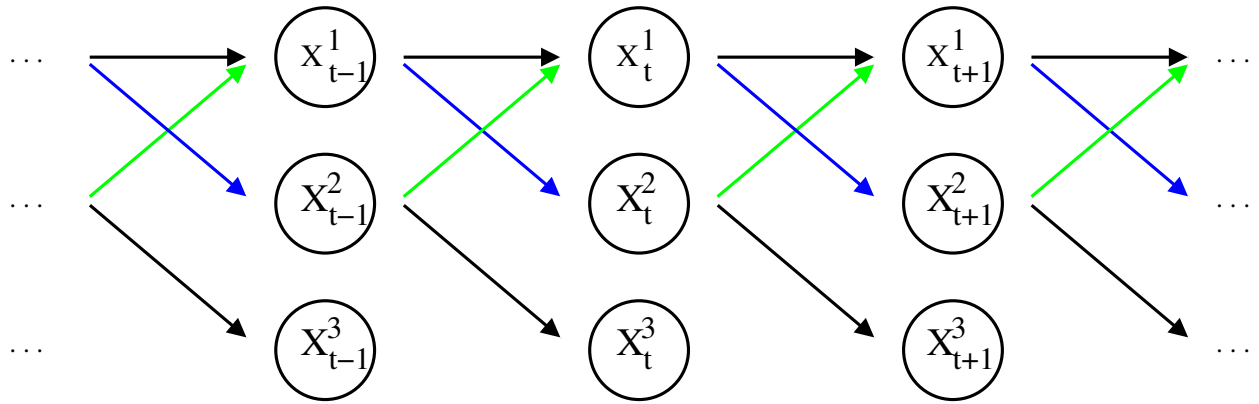


Theorem

Under $(\mathcal{A}_1)$ and $(\mathcal{A}_2)$, the probability distribution $\mathbb{P}_X$ admits a DBN representation according to DAG $\tilde{\mathcal{G}}$,

$$\tilde{\mathcal{G}} := \quad X^j_{t-1} \rightarrow X^i_t \quad \Leftrightarrow \quad X^i_t \not\perp\!\!\!\perp X^j_{t-1} \mid X_{t-1}$$

where $P = \{1, ..., p\}$. *(Proof: graphical models theory.)*



DAG $\mathcal{G}^{(1)}$ for an AR(1) process

- AR(1) process: $\forall t \geq 1,\ X_t = AX_{t-1} + B + \varepsilon_t,\quad \varepsilon_t \sim \mathcal{N}(0, \Sigma)$

$$\begin{bmatrix} X_t^1 \\ \cdot \\ \cdot \\ X_t^i \\ \cdot \\ \cdot \\ \cdot \\ X_t^p \end{bmatrix} = \begin{bmatrix} a_{11} & \cdot & \cdot & \cdot & a_{1j} & \cdot & \cdot & a_{1p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{i1} & \cdot & \cdot & \cdot & a_{ij} & \cdot & \cdot & a_{ip} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{p1} & \cdot & \cdot & \cdot & a_{pj} & \cdot & \cdot & a_{pp} \end{bmatrix} \begin{bmatrix} X_{t-1}^1 \\ \cdot \\ \cdot \\ \cdot \\ X_{t-1}^j \\ \cdot \\ \cdot \\ X_{t-1}^p \end{bmatrix} + \begin{bmatrix} b_t^1 \\ \cdot \\ \cdot \\ b_t^i \\ \cdot \\ \cdot \\ \cdot \\ b_t^p \end{bmatrix} + \begin{bmatrix} \varepsilon_t^1 \\ \cdot \\ \cdot \\ \varepsilon_t^i \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_t^p \end{bmatrix}$$

DAG $\mathcal{G}^{(1)}$ for an AR(1) process

- AR(1) process: $\forall t \geq 1,\ X_t = AX_{t-1} + B + \varepsilon_t,\quad \varepsilon_t \sim \mathcal{N}(0, \Sigma)$

$$\begin{bmatrix} X_t^1 \\ \cdot \\ \cdot \\ X_t^i \\ \cdot \\ \cdot \\ \cdot \\ X_t^p \end{bmatrix} = \begin{bmatrix} a_{11} & \cdot & \cdot & \cdot & a_{1j} & \cdot & \cdot & a_{1p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{i1} & \cdot & \cdot & \cdot & a_{ij} & \cdot & \cdot & a_{ip} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{p1} & \cdot & \cdot & \cdot & a_{pj} & \cdot & \cdot & a_{pp} \end{bmatrix} \begin{bmatrix} X_{t-1}^1 \\ \cdot \\ \cdot \\ \cdot \\ X_{t-1}^j \\ \cdot \\ \cdot \\ X_{t-1}^p \end{bmatrix} + \begin{bmatrix} b_t^1 \\ \cdot \\ \cdot \\ b_t^i \\ \cdot \\ \cdot \\ \cdot \\ b_t^p \end{bmatrix} + \begin{bmatrix} \varepsilon_t^1 \\ \cdot \\ \cdot \\ \varepsilon_t^i \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_t^p \end{bmatrix}$$

Proposition

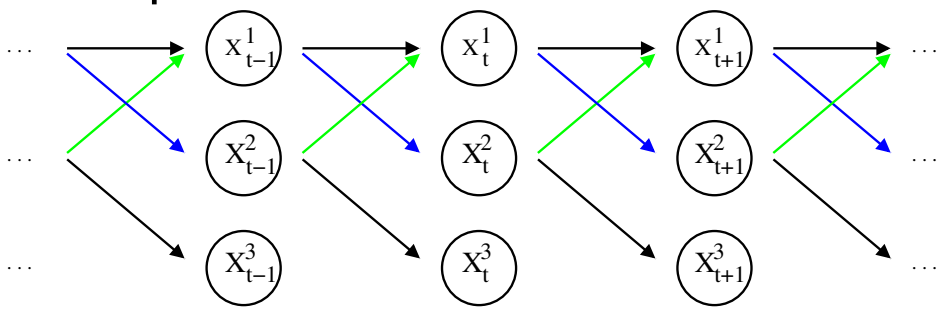
If $\Sigma = Var(\varepsilon_t)$ is diagonal then $\tilde{\mathcal{G}} := \{X_{t-1}^j \rightarrow X_t^i\} \Leftrightarrow a_{ij} \neq 0$.

DAG $\mathcal{G}^{(1)}$ for an AR(1) process

- AR(1) process: $\forall t \geq 1,\ X_t = AX_{t-1} + B + \varepsilon_t,\quad \varepsilon_t \sim \mathcal{N}(0, \Sigma)$

$$\begin{bmatrix} X_t^1 \\ \cdot \\ \cdot \\ X_t^i \\ \cdot \\ \cdot \\ X_t^p \end{bmatrix} = \begin{bmatrix} a_{11} & \cdot & \cdot & \cdot & a_{1j} & \cdot & \cdot & a_{1p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{i1} & \cdot & \cdot & \cdot & a_{ij} & \cdot & \cdot & a_{ip} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{p1} & \cdot & \cdot & \cdot & a_{pj} & \cdot & \cdot & a_{pp} \end{bmatrix} \begin{bmatrix} X_{t-1}^1 \\ \cdot \\ \cdot \\ \cdot \\ X_{t-1}^j \\ \cdot \\ \cdot \\ X_{t-1}^p \end{bmatrix} + \begin{bmatrix} b_t^1 \\ \cdot \\ \cdot \\ b_t^i \\ \cdot \\ \cdot \\ b_t^p \end{bmatrix} + \begin{bmatrix} \varepsilon_t^1 \\ \cdot \\ \cdot \\ \varepsilon_t^i \\ \cdot \\ \cdot \\ \varepsilon_t^p \end{bmatrix}$$

- Example:



$$A = \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & 0 & 0 \\ 0 & a_{32} & 0 \end{bmatrix}$$

Inference $\rightsquigarrow$ partial order dependencies approximation

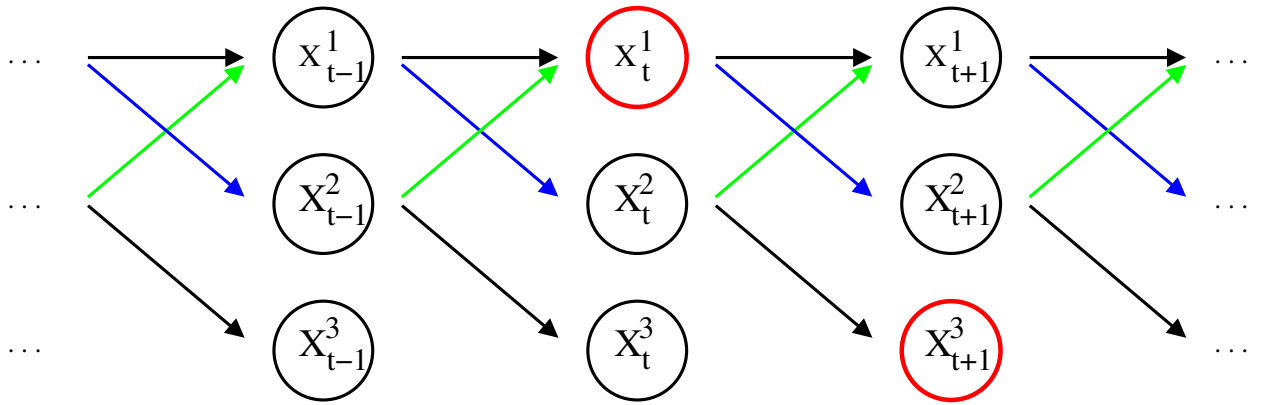
- 1^{st} order dependencies (Wille, 2004).
  $\Rightarrow$ Extention to dynamic graphs

Definition

q^{th} order conditional dependence DAG $\mathcal{G}^{(q)}$, $(q \geq 1)$

$$\exists Q \subseteq P, |Q| = q, \quad X_t^i \perp\!\!\!\perp X_{t-1}^j \mid X_{t-1}^Q \quad \Leftrightarrow \quad \{X_{t-1}^j \rightarrow X_t^i\} \notin \mathcal{G}^{(q)},$$

- Example: $\{X_t^1 \rightarrow X_{t+1}^3\} \notin \mathcal{G}^{(1)}$



Inference $\rightsquigarrow$ partial order dependencies approximation

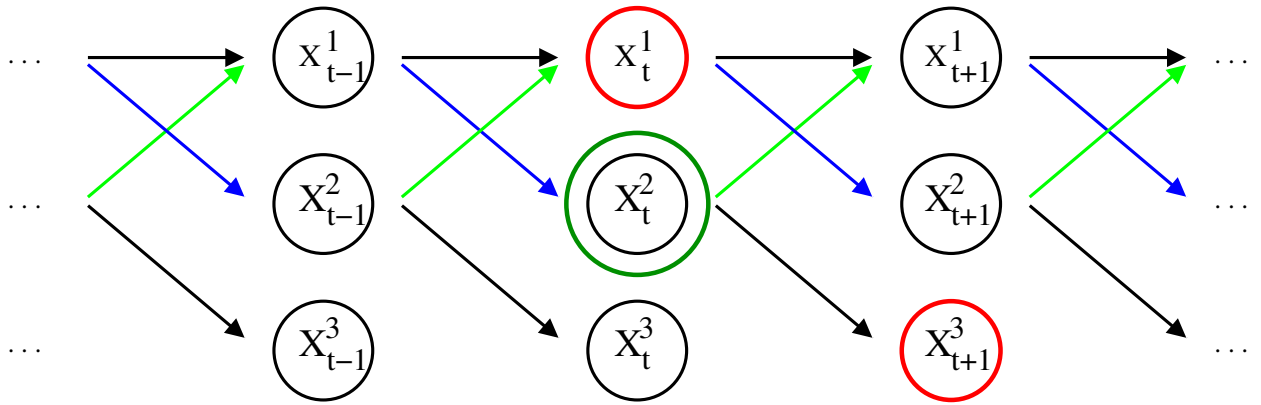
- 1^{st} order dependencies (Wille, 2004).
  $\Rightarrow$ Extention to dynamic graphs

Definition

q^{th} order conditional dependence DAG $\mathcal{G}^{(q)}$, $(q \geq 1)$

$$\exists Q \subseteq P, |Q| = q, \quad X_t^i \perp\!\!\!\perp X_{t-1}^j \mid X_{t-1}^Q \quad \Leftrightarrow \quad \{X_{t-1}^j \rightarrow X_t^i\} \notin \mathcal{G}^{(q)},$$

- Example: $\{X_t^1 \rightarrow X_{t+1}^3\} \notin \mathcal{G}^{(1)}$



Inference $\rightsquigarrow$ partial order dependencies approximation

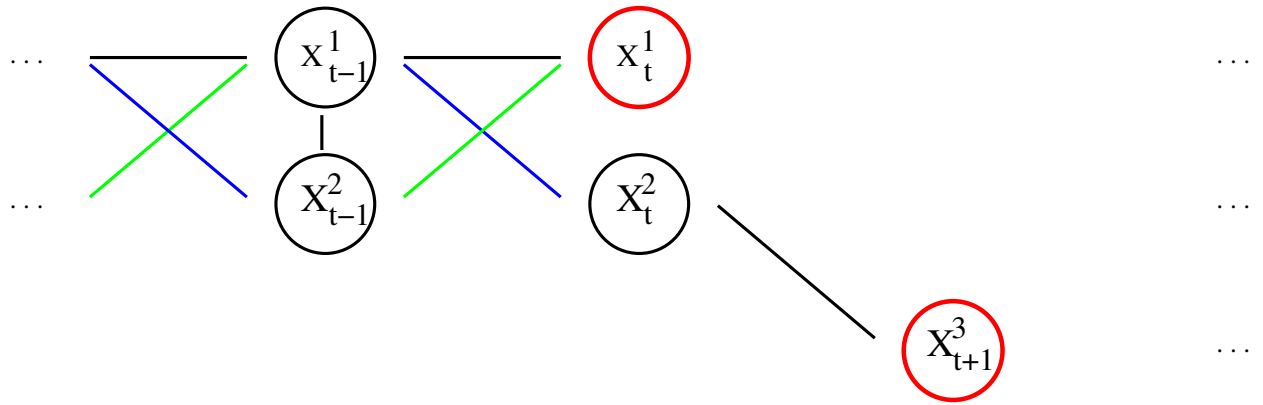
- 1^{st} order dependencies (Wille, 2004).
  $\Rightarrow$ Extention to dynamic graphs

Definition

q^{th} order conditional dependence DAG $\mathcal{G}^{(q)}$, $(q \geq 1)$

$$\exists Q \subseteq P, |Q| = q, \quad X_t^i \perp\!\!\!\perp X_{t-1}^j \mid X_{t-1}^Q \quad \Leftrightarrow \quad \{X_{t-1}^j \rightarrow X_t^i\} \notin \mathcal{G}^{(q)},$$

- Example: $\{X_t^1 \rightarrow X_{t+1}^3\} \notin \mathcal{G}^{(1)}$



Inference $\rightsquigarrow$ partial order dependencies approximation

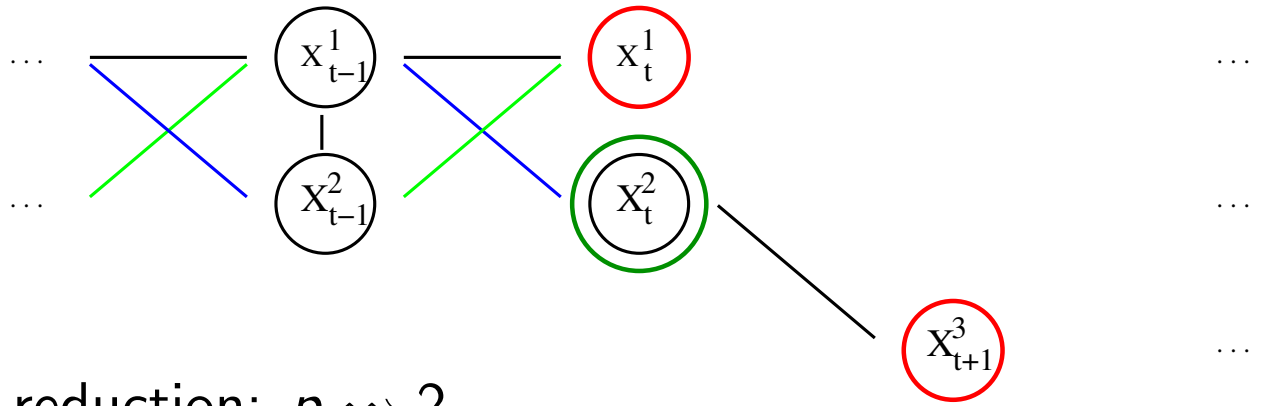
- 1^{st} order dependencies (Wille, 2004).
  $\Rightarrow$ Extention to dynamic graphs

Definition

q^{th} order conditional dependence DAG $\mathcal{G}^{(q)}$, $(q \geq 1)$

$$\exists Q \subseteq P, |Q| = q, \quad X_t^i \perp\!\!\!\perp X_{t-1}^j \mid X_{t-1}^Q \quad \Leftrightarrow \quad \{X_{t-1}^j \rightarrow X_t^i\} \notin \mathcal{G}^{(q)},$$

- Example: $\{X_t^1 \rightarrow X_{t+1}^3\} \notin \mathcal{G}^{(1)}$



Dimension reduction: $p \rightsquigarrow 2$

$\mathcal{G}^{(q)}$ properties

Proposition

If the number of parents of each vertex in $\tilde{\mathcal{G}}$ $N_{pa}^{Max}(\tilde{\mathcal{G}}) \leq q$, *then* $\tilde{\mathcal{G}} \supseteq \mathcal{G}^{(q)}$.

$\mathcal{G}^{(q)}$ properties

Proposition

If the number of parents of each vertex in $\tilde{\mathcal{G}}$ $N_{pa}^{Max}(\tilde{\mathcal{G}}) \leq q$, *then* $\tilde{\mathcal{G}} \supseteq \mathcal{G}^{(q)}$.

Definition

Faithfulness: any conditional independence can be derived from $\tilde{\mathcal{G}}$.

$\mathcal{G}^{(q)}$ properties

Proposition

If the number of parents of each vertex in $\tilde{\mathcal{G}}$ $N_{pa}^{Max}(\tilde{\mathcal{G}}) \leq q$, *then* $\tilde{\mathcal{G}} \supseteq \mathcal{G}^{(q)}$.

Definition

Faithfulness: any conditional independence can be derived from $\tilde{\mathcal{G}}$.

Proposition

If $\mathbb{P}_X$ *is faithful to* $\tilde{\mathcal{G}}$ *then,*

- $\tilde{\mathcal{G}} \subseteq \mathcal{G}^{(q)}$,

If $\mathbb{P}_X$ is 'faithful' to $\tilde{\mathcal{G}}$ then $\tilde{\mathcal{G}} \subseteq \mathcal{G}^{(q)}$.

Proof.

Contraposition

- Assume $(X^j_{t-1}, X^i_t) \notin \mathcal{G}^{(q)}$,
  then $\exists\ Q \subset P$, $|Q| = q$, such that $X^i_t \perp\!\!\!\perp X^j_{t-1} | X^Q_{t-1}$.
- From faithfulness, X^Q_{t-1} separates X^j_{t-1} and X^i_t in the moral graph of the ancestral set containing $X^i_t \cup X^j_{t-1} \cup X^Q_{t-1}$,
  then $(X^j_{t-1}, X^i_t) \notin \tilde{\mathcal{G}}$.

$\mathcal{G}^{(q)}$ properties

Proposition

If the number of parents of each vertex in $\tilde{\mathcal{G}}$ $N_{pa}^{Max}(\tilde{\mathcal{G}}) \leq q$, *then* $\tilde{\mathcal{G}} \supseteq \mathcal{G}^{(q)}$.

Definition

Faithfulness: any conditional independence can be derived from $\tilde{\mathcal{G}}$.

Proposition

If $\mathbb{P}_X$ *is faithful to* $\tilde{\mathcal{G}}$ *then,*

- $\tilde{\mathcal{G}} \subseteq \mathcal{G}^{(q)}$,
- *for all* $q \geq N_{pa}^{Max}(\tilde{\mathcal{G}})$, $\tilde{\mathcal{G}} = \mathcal{G}^{(q)}$,

$\mathcal{G}^{(q)}$ properties

Proposition

If the number of parents of each vertex in $\tilde{\mathcal{G}}$ $N_{pa}^{Max}(\tilde{\mathcal{G}}) \leq q$, then $\tilde{\mathcal{G}} \supseteq \mathcal{G}^{(q)}$.

Definition

Faithfulness: any conditional independence can be derived from $\tilde{\mathcal{G}}$.

Proposition

If $\mathbb{P}_X$ is faithful to $\tilde{\mathcal{G}}$ then,

- $\tilde{\mathcal{G}} \subseteq \mathcal{G}^{(q)}$,
- *for all* $q \geq N_{pa}^{Max}(\tilde{\mathcal{G}})$, $\tilde{\mathcal{G}} = \mathcal{G}^{(q)}$,
- *if* $q \geq N_{pa}^{Max}(\mathcal{G}^{(q)})$, *then* $\tilde{\mathcal{G}} = \mathcal{G}^{(q)}$.

$\Rightarrow$ infer $\mathcal{G}^{(1)}$ to reduce the dimension.

DBN Inference from $\mathcal{G}^{(1)}$ (package R 'G1DBN')

- **Step 1: infer $\mathcal{G}^{(1)}$ (dimension reduction)**

$$X_t^i = b_{ijk} + a_{ij|k} X_{t-1}^j + a_{ik|j} X_{t-1}^k + \eta_t^{i,j,k}$$

  - For all $i, j \in P$,
    - for all $k \neq j$, test $\mathcal{H}_0^{i,j,k}$: "$a_{ij|k} = 0$" (Student test)
    - $S_1(i,j) \leftarrow Max_{k \neq j}(p_{ij|k})$.
  - $\hat{E}_i^{(1)} = \{j \in P, S_1(i,j) < \alpha_1\}$.

- **Step 2: infer $\tilde{\mathcal{G}}$ from $\mathcal{G}^{(1)}$.**
  - For all edge in $\hat{\mathcal{G}}^{(1)}$, compute the p-value $p_{ij|\hat{E}_i^{(1)}}$.
  - $E(\tilde{\mathcal{G}}) = \{(X_{t-1}^j, X_t^i)_{t>1}, i \in P, j \in \hat{E}_i^{(1)}$ tel que $p_{ij|\hat{E}_i^{(1)}} < \alpha_2\}$.

Validation

- Package R 'G1DBN' available from the CRAN archive `http://cran.at.r-project.org`,

Validation

- Package R 'G1DBN' available from the CRAN archive `http://cran.at.r-project.org`,
- Comparative simulation study,

Validation

- Package R 'G1DBN' available from the CRAN archive `http://cran.at.r-project.org`,
- Comparative simulation study,
- Real data analysis:

Validation

- Package R 'G1DBN' available from the CRAN archive `http://cran.at.r-project.org`,
- Comparative simulation study,
- Real data analysis:
    - Yeast cell cycle S. Cerevisiae (Spellman 1998),

Validation

- Package R 'G1DBN' available from the CRAN archive `http://cran.at.r-project.org`,
- Comparative simulation study,
- Real data analysis:
  - Yeast cell cycle S. Cerevisiae (Spellman 1998),
  - Starch metabolism in Arabidopsis leaves (Smith 2004).

Simulation study

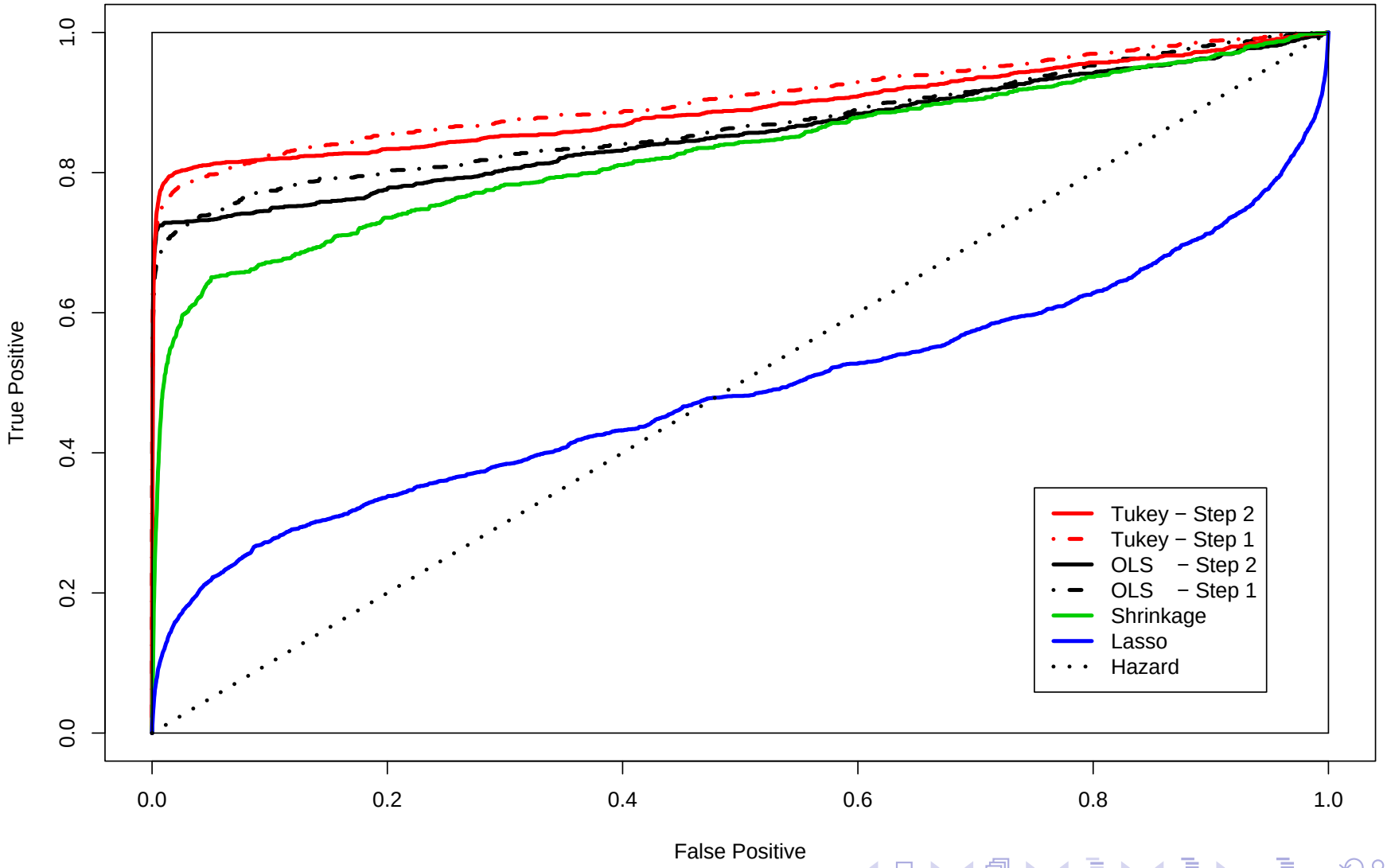
- Random generation of 100 matrices $A_{[p \times p]}$:
  - p = 50,
  - 2 % of edges: non-zero coefficients, $a_{ij} \sim \mathcal{U}(-1, 1)$.

- AR(1) process simulation

$$\forall 1 \leq t \leq n,\ X_t^i = \sum_{j=1}^{p} a_{ij} X_{t-1}^j + b_i + \varepsilon_t^i, \quad \varepsilon_t^i \sim \mathcal{N}(0, \sigma_i).$$

  - $b_i \sim \mathcal{U}(0, 1)$,
  - $\sigma_i \sim \mathcal{U}(0.03; 0.08)$,
  - $n = 20$ to 50.

ROC curves: G1DBN vs Lasso (Tibshirani, 1996) Shrinkage (Opgen-Rhein and Strimmer, 2007)

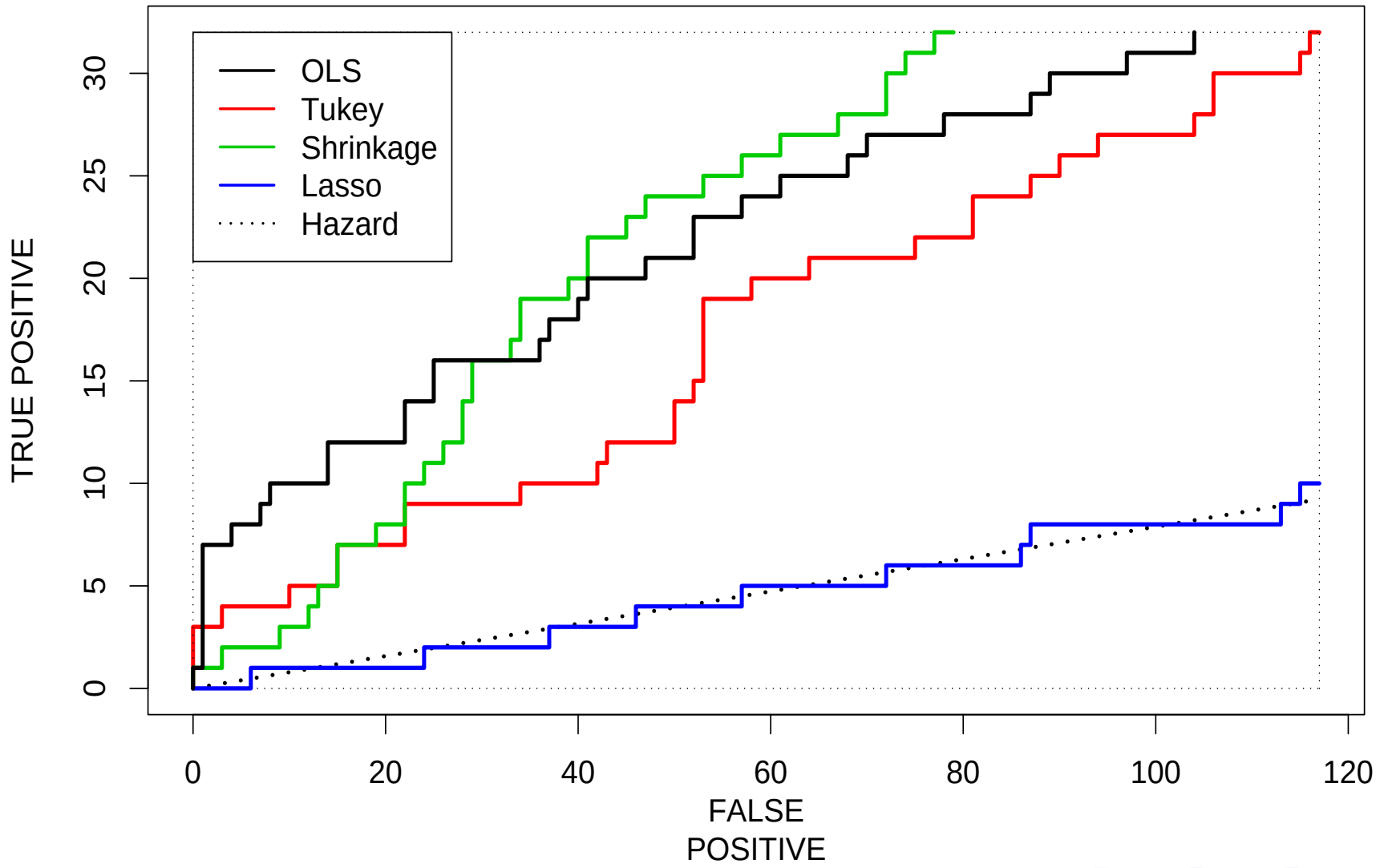


Yeast cell cycle (Spellman, 1998)

Expression data:

- 792 genes,
- 18 time points (each 7 minutes).
- 9 transcription factors,

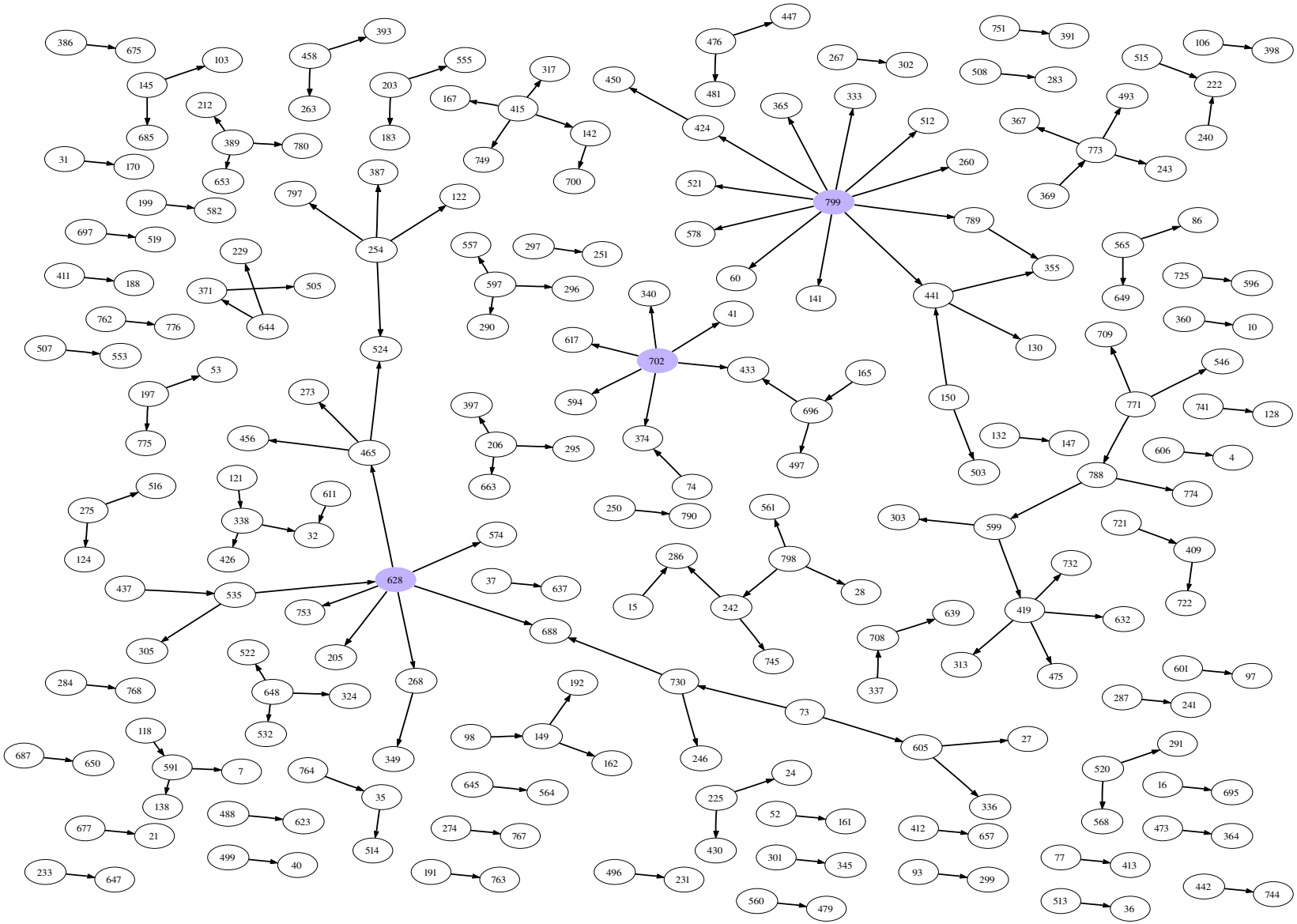
Yeast cell cycle (Spellman, 1998)



Starch metabolism in Arabidopsis leaves (Smith, 2004)

- Expression data
  - 800 genes
  - 11 time points (2 repetitions)
- Inferred network:
  - 236 genes,
  - 168 edges,
  - “hub” structure.

DAG $\hat{\mathcal{G}}^{(1)}$ for $\alpha_1 = 0.1$ (168 edges).



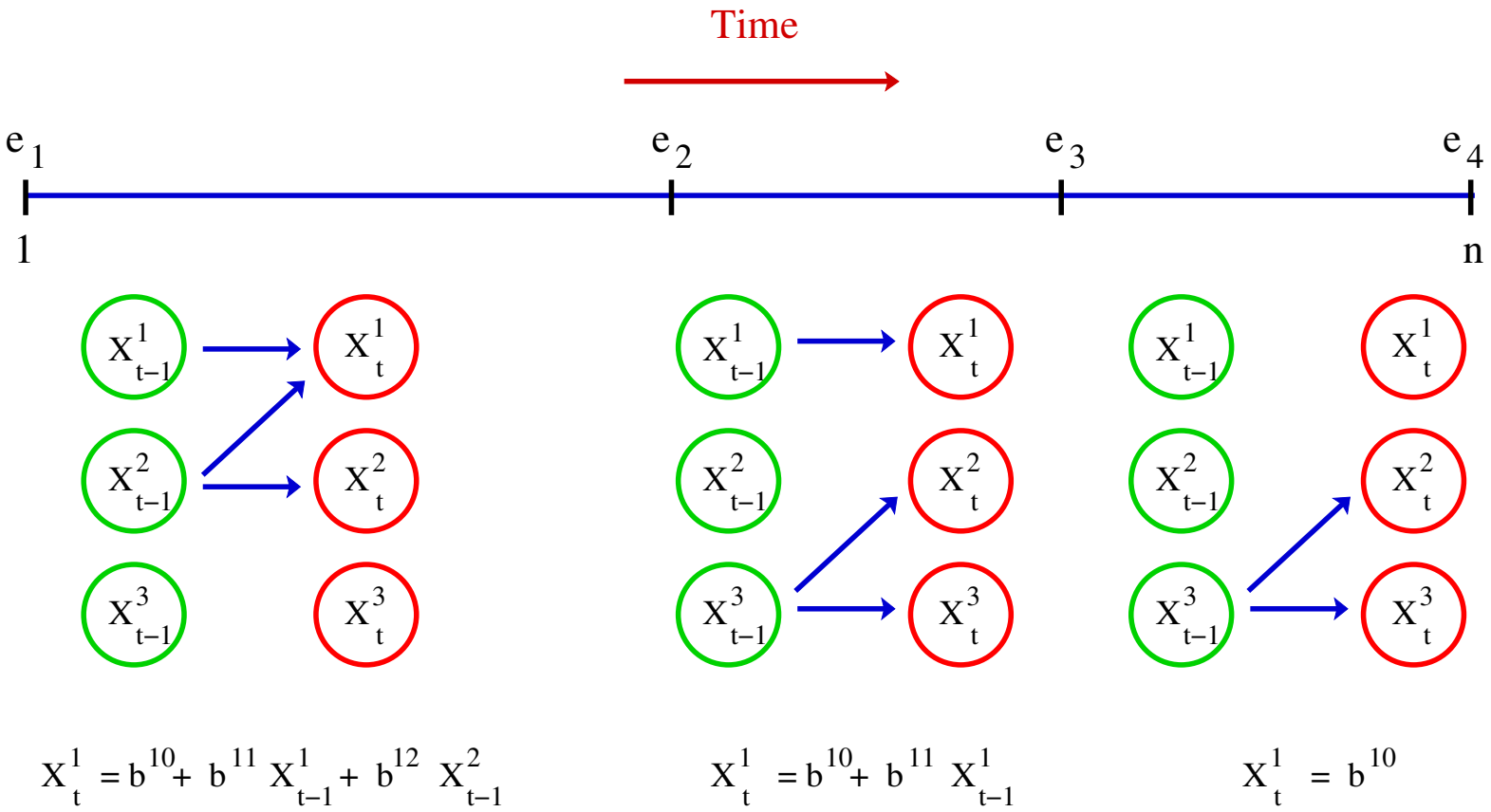
Inferring homogenous DBN with partial order dependences

- Mathematical results:
  $\tilde{\mathcal{G}}$, $\mathcal{G}^{(q)}$ definition and characterization for DBNs.
- A new DBN inference procedure
  - when $n << p$
  - performs well in comparison with existent inference procedures
  - R package 'G1DBN' available from the CRAN.
- *Inferring dynamic genetic networks with low order independencies.* Lèbre, S., under revision for Statistical Applications in Genetics and Molecular Biology.

Outline

1. Sequence modeling: an EM algorithm for MTD models
2. Genetic network: inferring DBN with partial order dependence
3. Inferring non-homogneous DBN with rjMCMC

Extension: infering a **time-dependent** network?



Multiple changepoint Model

A piecewise homogeneous network

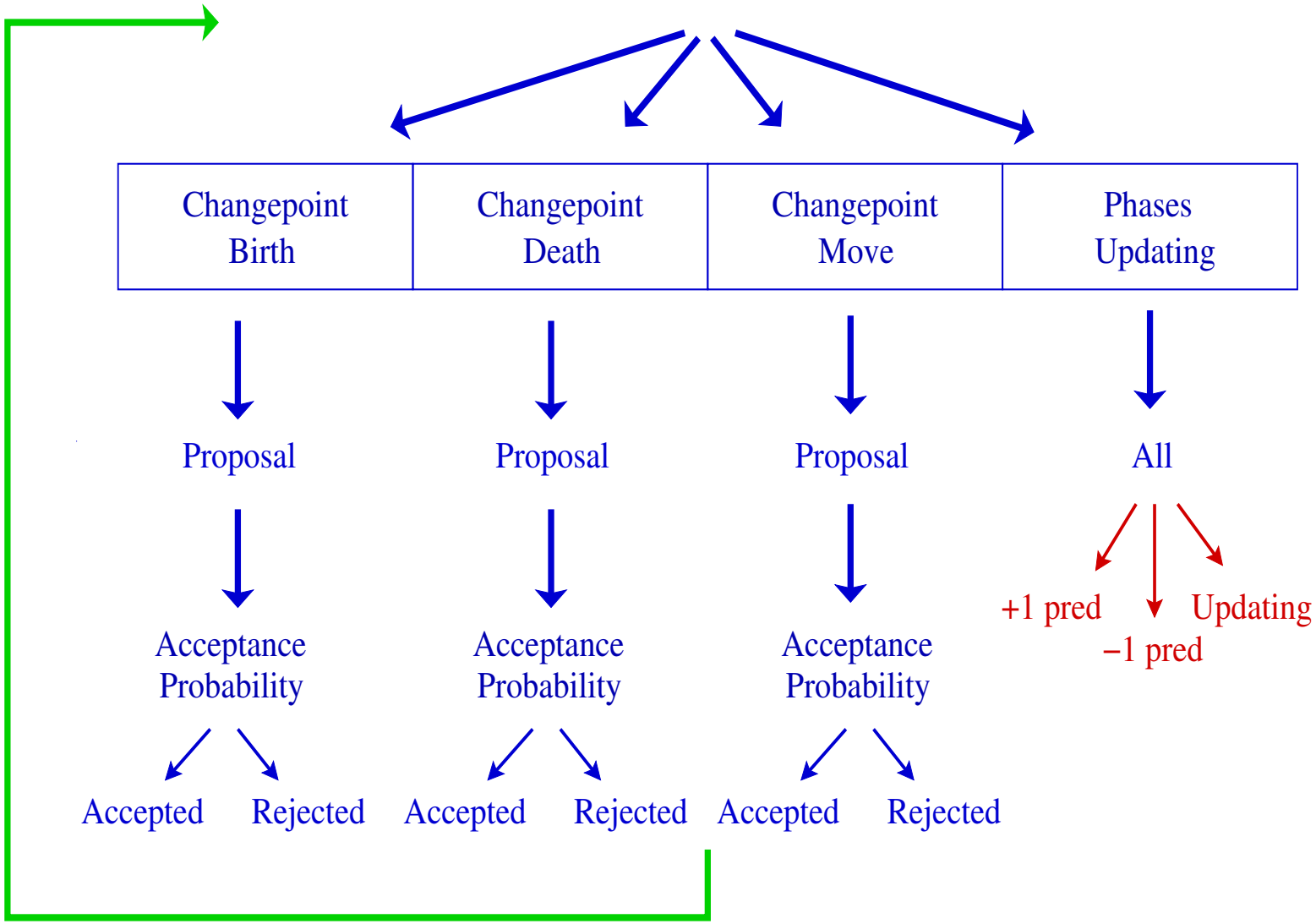
For each gene i,

- changepoint vector $e = (e_1, ..., e_{h-1}, e_h, ..., e_s)$
- in each phase h,
  - a set of k_h predictors $\tau_h = \{j_1, ..., j_{k_h}\}$
  - and a set of parameters $\theta_h = ((b_h^{ij})_{j \in \{0,...,q\}},\ \sigma_h)$,

define the regression model, for all $e_h \leq t < e_{h+1}$,

$$X_t^i \ = \ b_h^{i0} + \sum_{j \in \tau_h} b_h^{ij}\ X_{t-1}^j \ + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_h).$$

Inference: 2 steps **embedded** *reversible jump MCMC. (Green 95, Andrieu and Doucet 99)*



Phases updating: regression model change within phases. (Andrieu and Doucet, 1999)

- Priors
  - number of predictors $s_h^i \sim \mathcal{P}(\Lambda)$
  - set of predictors $\tau_h^i | s_h^i \sim \mathcal{U}niform$
- Integration of the "nuisance" parameters (a, σ)
  $\rightsquigarrow$ acceptance ratio :

$$r_{s_h^i, s_h^i+1}(\tau_h^i, \tau_h^{i+}) = \frac{1}{\sqrt{1+\delta^2}} \left(\frac{\gamma_0 + (y_h^i)^t P_{\tau_h^i} y_h^i}{\gamma_0 + (y_h^i)^t P_{\tau_h^{i+}} y_h^i} \right)^{(m^i(\xi_h^i - \xi_{h-1}^i)+\upsilon_0)/2} .$$

- Acceptance probability: $\alpha_{s_h^i, s_h^i+1} = min\{1, r_{s_h^i, s_h^i+1}(\tau_h^i, \tau_h^{i+})\}$
  $\rightsquigarrow$ Reversibility
- Convergence property

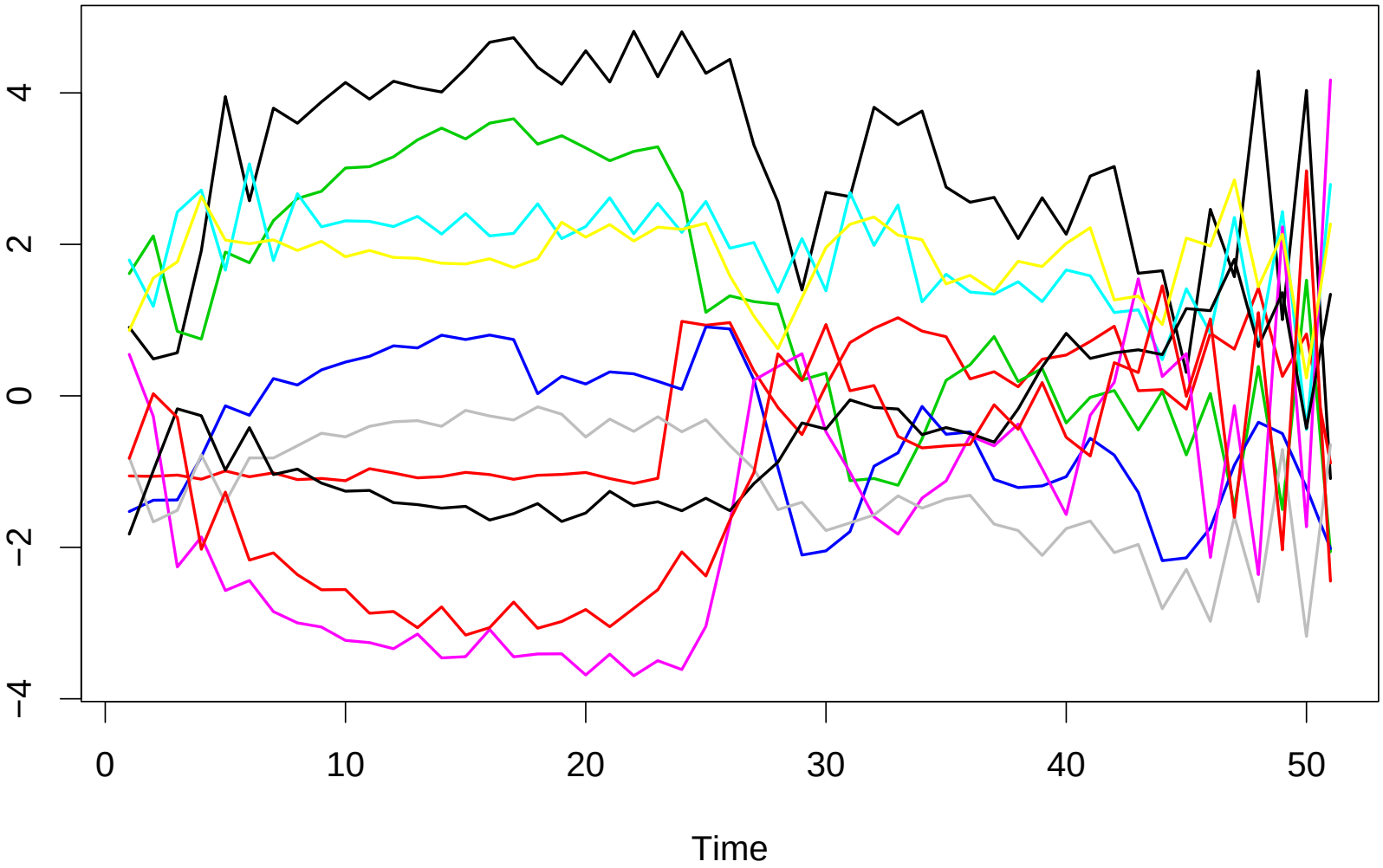
Simulation study

$$X_t^i = b_h^{i0} + \sum_{j \in \tau_h} b_h^{ij} X_{t-1}^j + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_h^i).$$

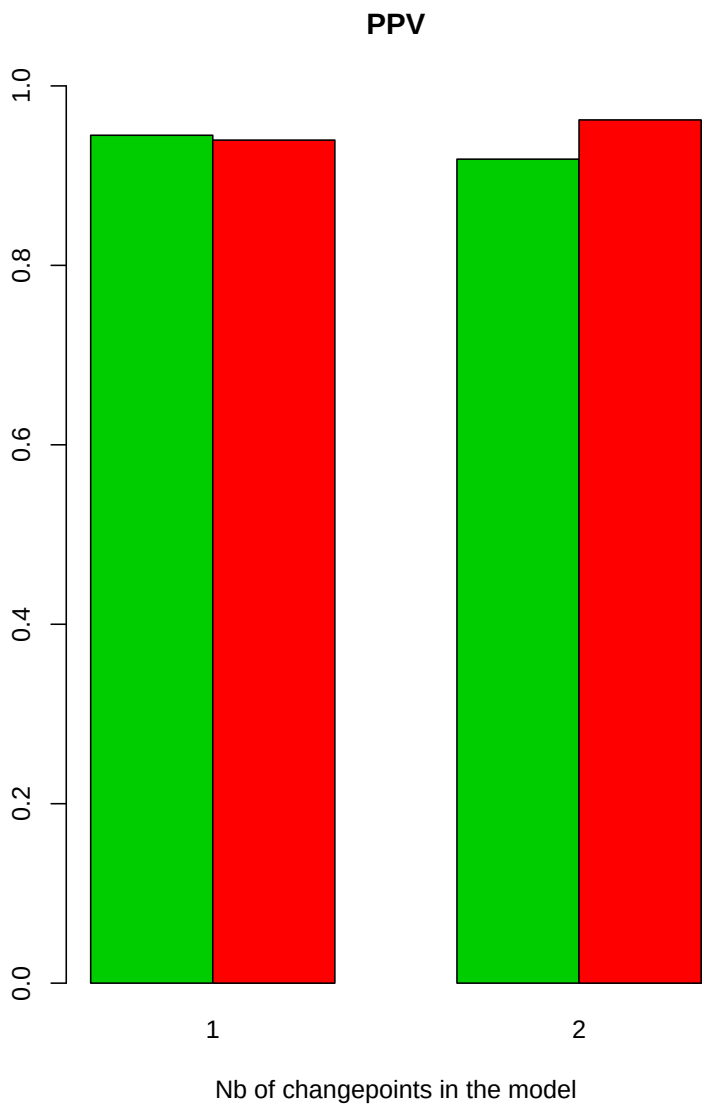
- $p = 1000$ target genes
- $s_{max} = 2$ (max number changepoints)
- $q_{max} = 5$ (max number of factor genes)
- $\sigma^i \sim \mathcal{U}(0.03; 0.08)$
- $b^{ij} \sim \mathcal{U}(0.2 + \sigma_i; 1 + \sigma_i)$
- $b^{i0} \sim \mathcal{U}([-2, -0.5] \cup [0.5, 2])$
- $n = 50$ and then 100 repeated time points.

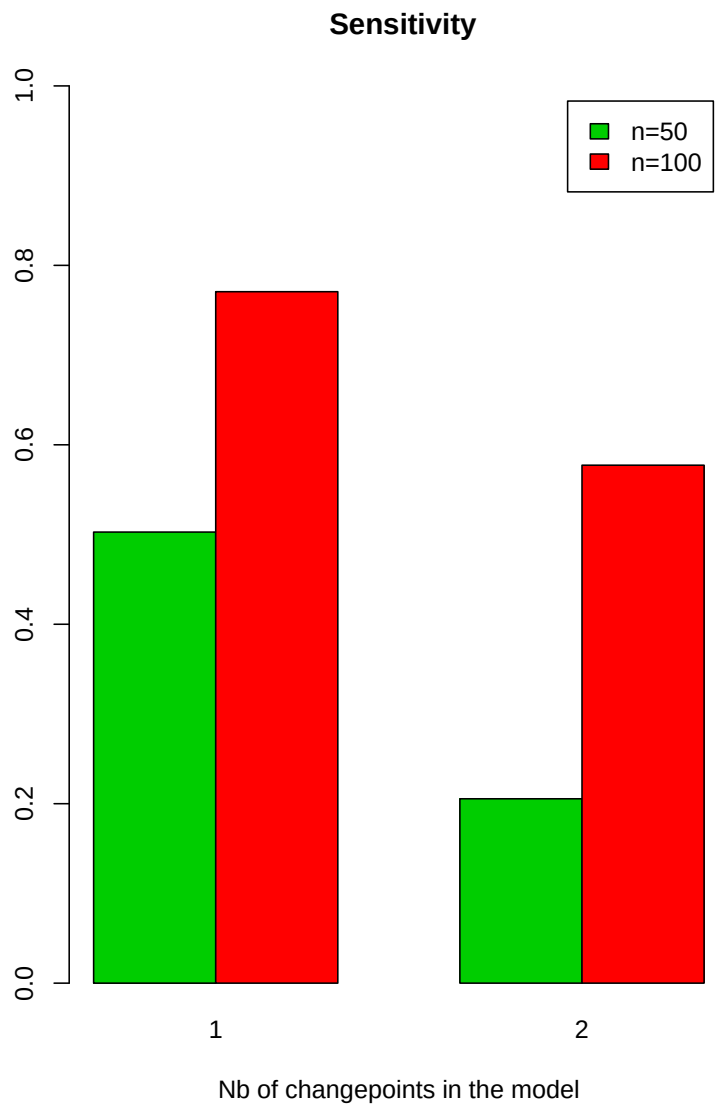
$$PPV = \frac{TP}{TP + FP} \qquad Sensitivity = \frac{TP}{TP + FN}$$

Example of simulated data (10 series)

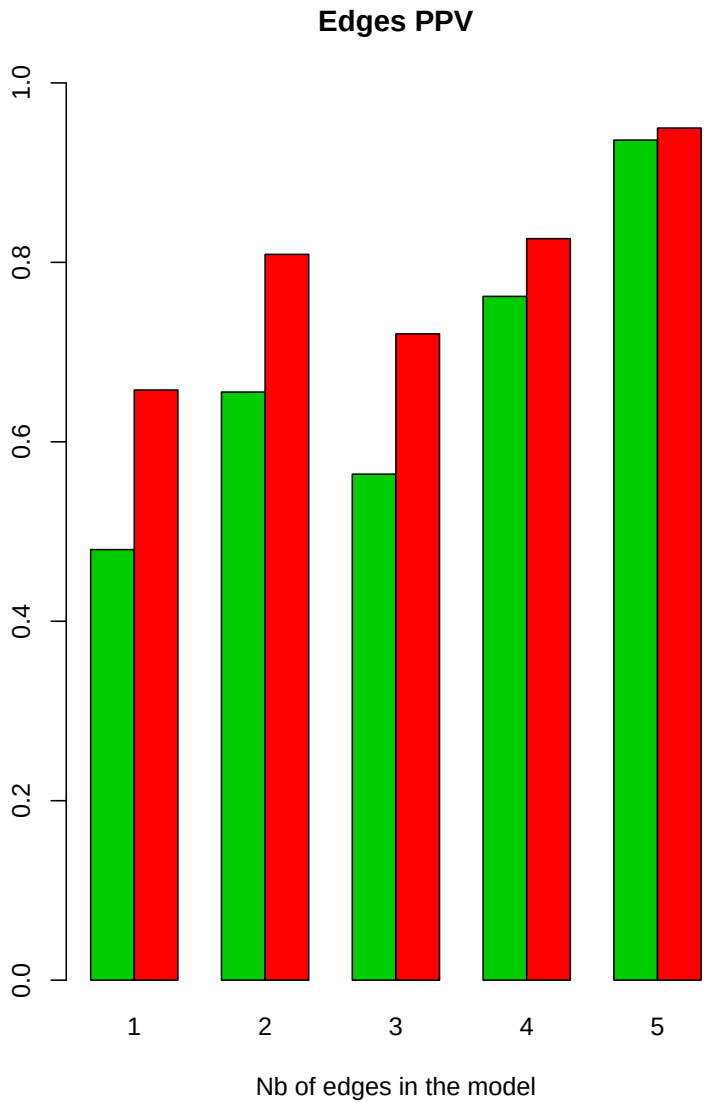


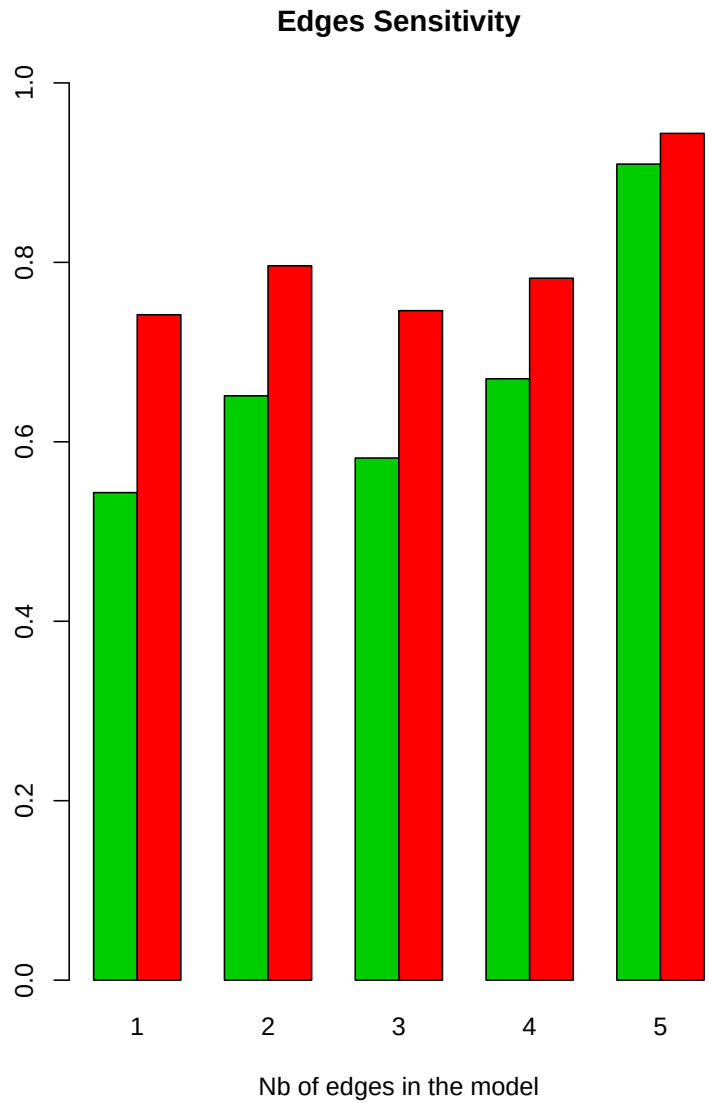
Changepoints detection





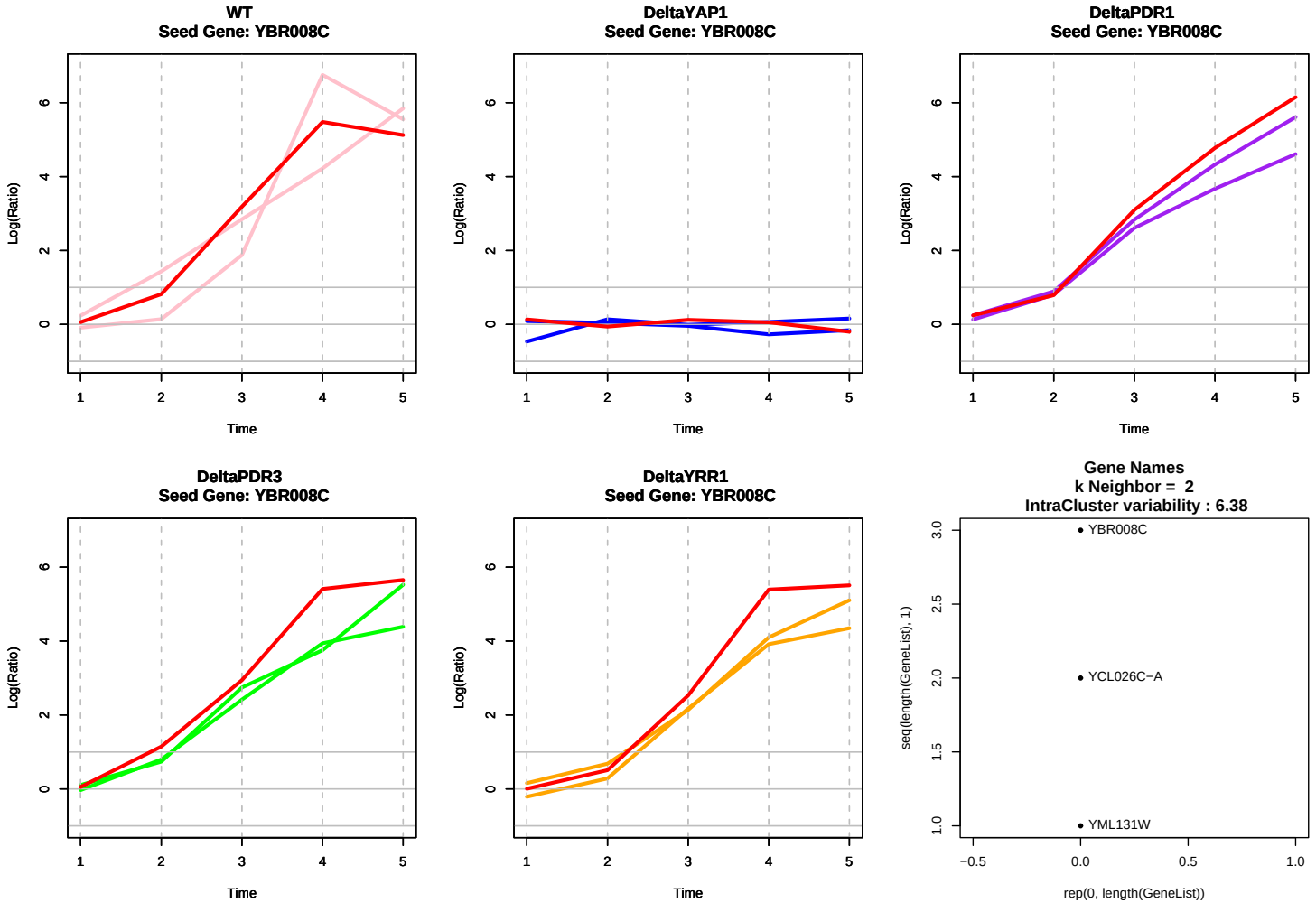
Edges detection





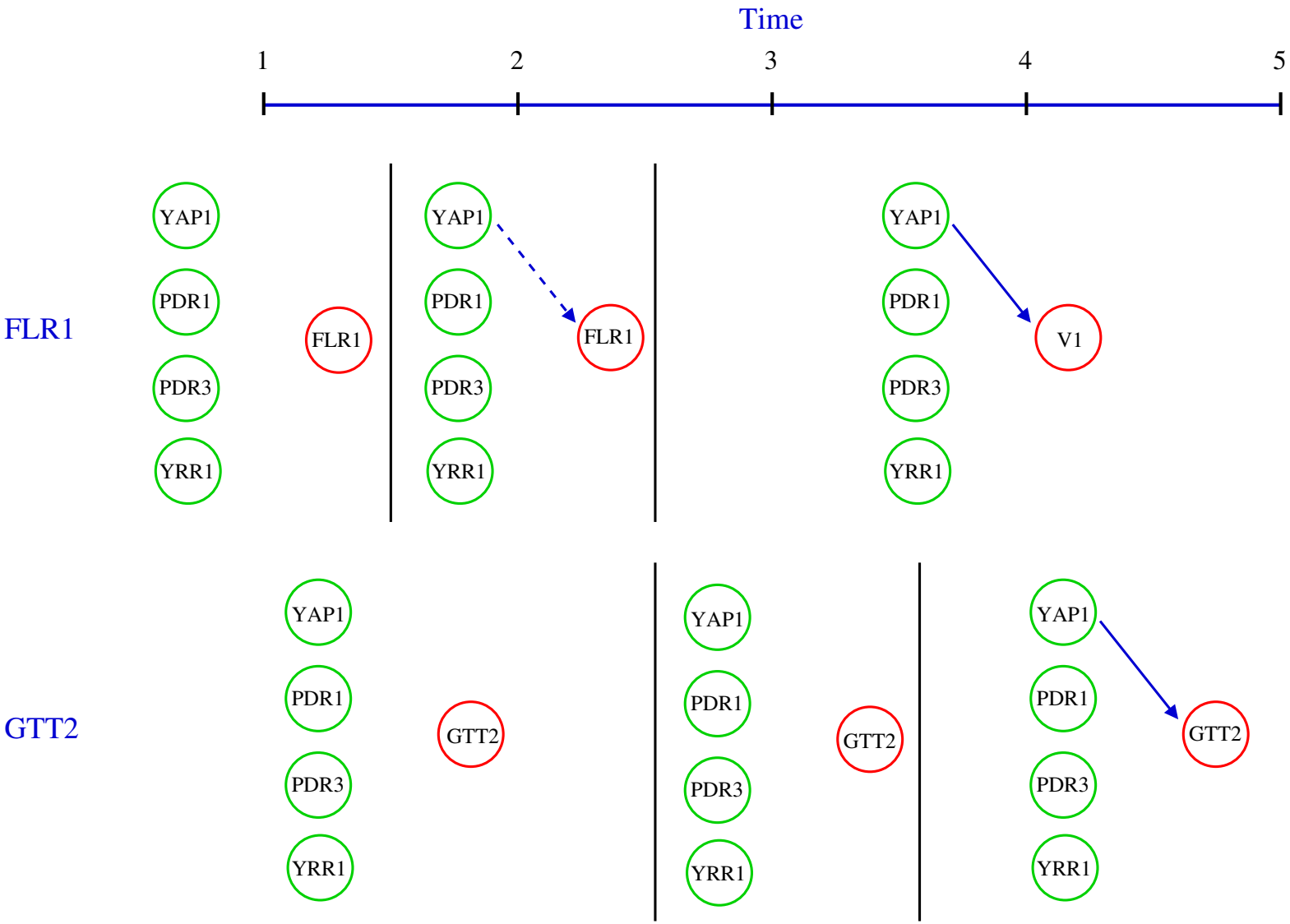
Response to benomyl addition by the yeast *S cerevisiae*

$\rightsquigarrow$ in collaboration with Gaëlle Lelandais (EBGM)



$$Y_t^i = b_h^{i0} + \sum_{j \in \tau_h} b_h^{ij} \mathbb{1}_{\{TF_t^t = 0\}} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^i).$$

FLR1 and GTT2: time-delayed YAP1 targets



Conclusions

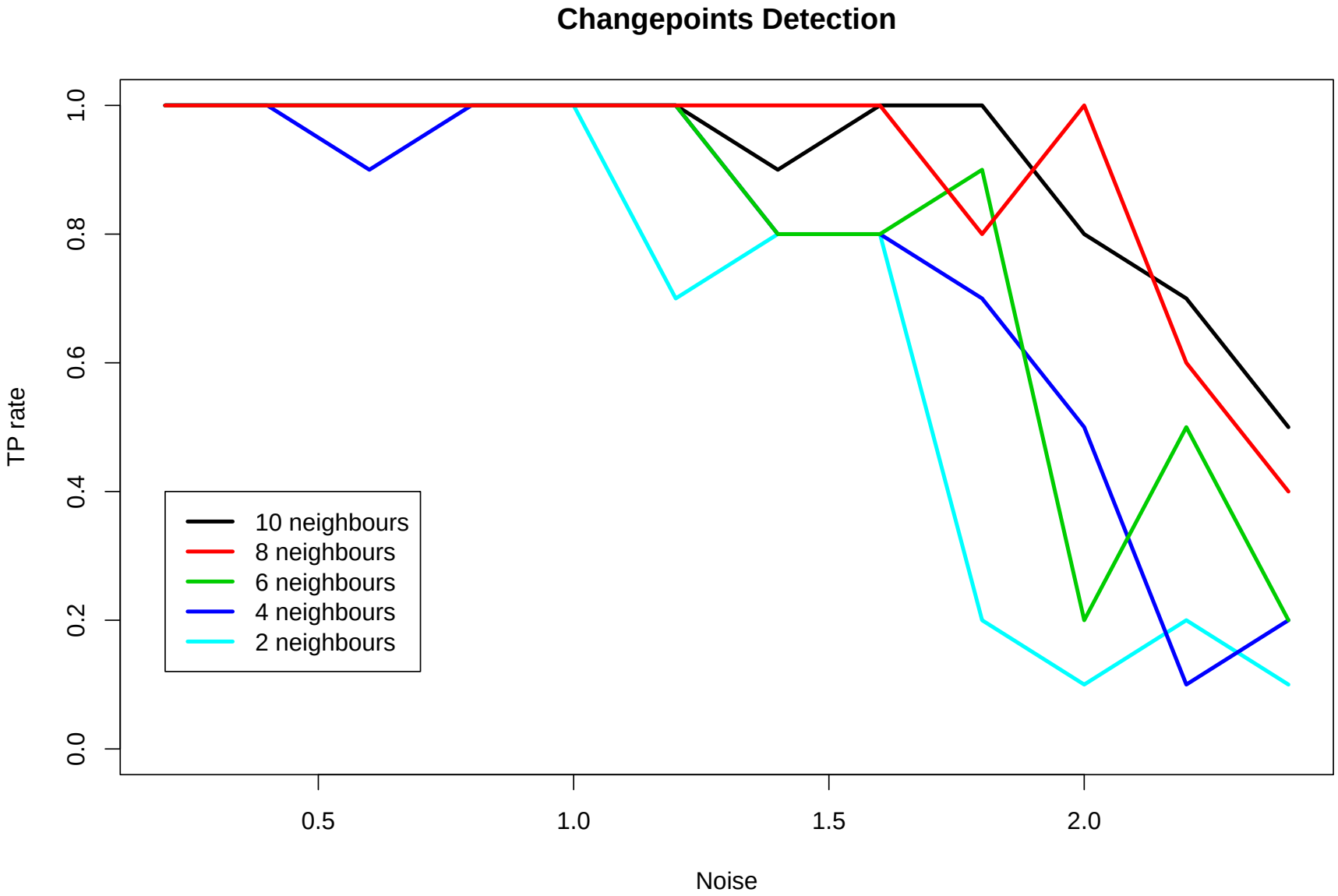
- An EM algorithm for estimation in MTD model (1 article).
  $\rightsquigarrow$ available from seq++
- A new DBN inference procedure when $n << p$
  by considering 1^{st} order dependence (1 article under revision).
  $\rightsquigarrow$ R package 'G1DBN' available from the CRAN
- Relaxation of the homogenity assumption for DBN modeling and reversible jump MCMC inference procedure.

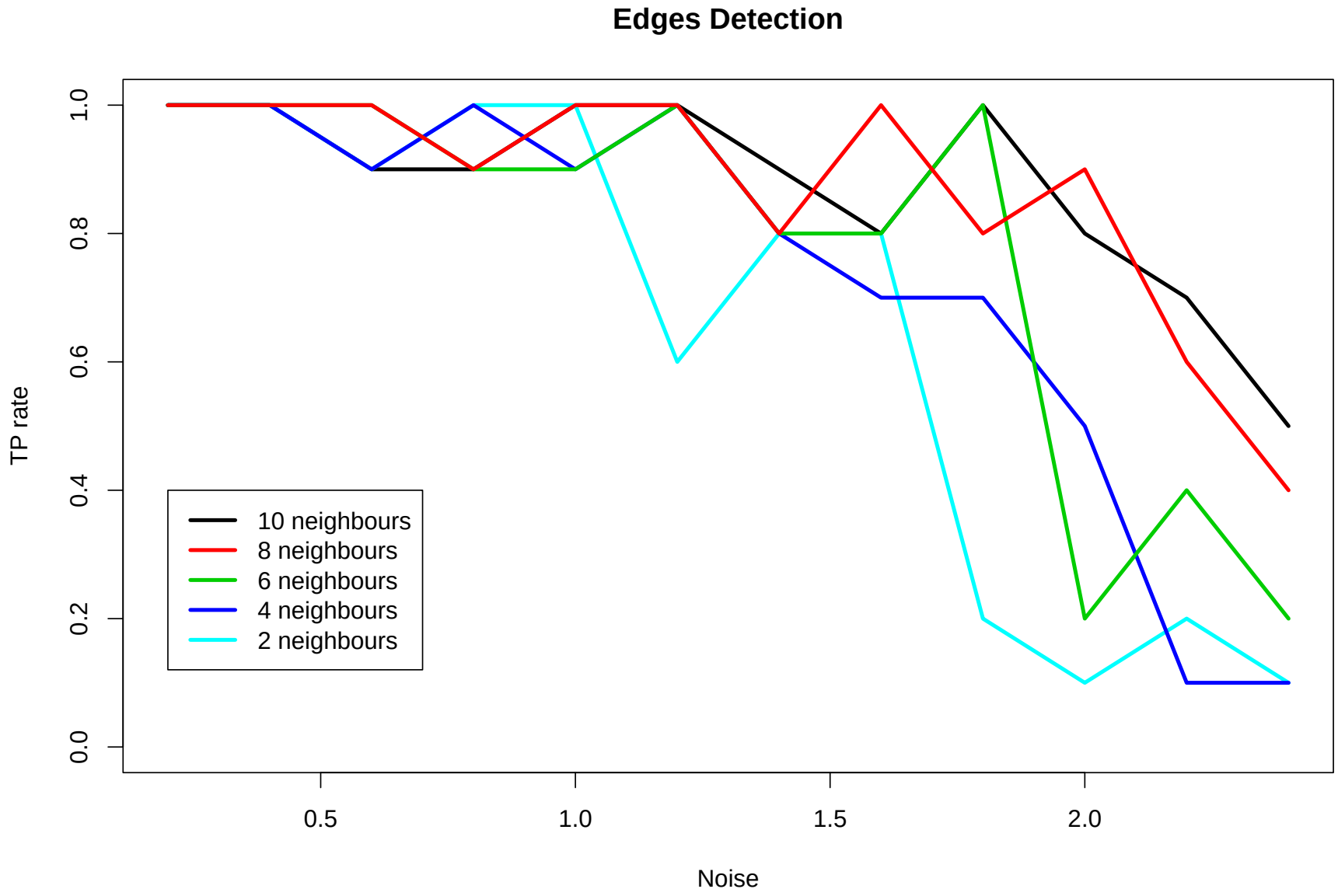
Perspectives

- RJ MCMC procedure: real data,
  - finalize reaction to benomyl analysis,
  - test on Yeast gene expression with 36 repeated time points (Tu et al., 2005).
- Use those DBN inference procedures to study stress response in E. coli and cancer data.
- Random networks and characterization from incomplete graphs.

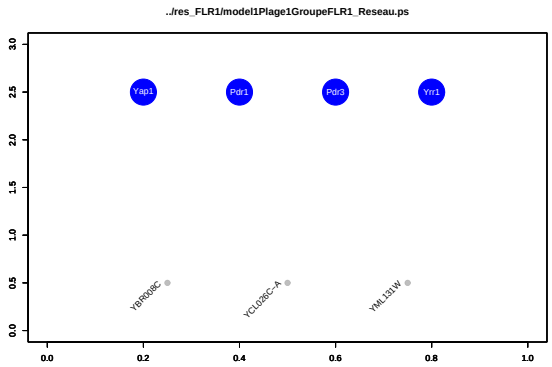
Priors

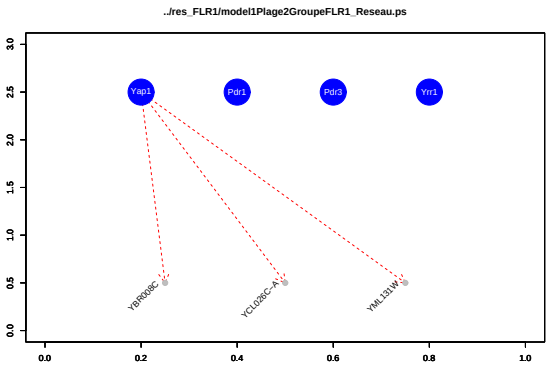
- number of changepoints $k \sim \mathcal{P}(\lambda)$
- changepoints vector $\xi|k \sim \mathcal{U}niform$
- number of predictors $s_h^i \sim \mathcal{P}(\Lambda)$
- set of predictors $\tau_h^i|s_h^i \sim \mathcal{U}niform$
- variance $(\sigma_h^i)^2 \sim \mathcal{IG}(\upsilon_0, \gamma_0), \quad \upsilon_0, \gamma_0 \ll$
- regression coefficient $a_h^i|\sigma_h^i \sim \mathcal{N}(0, (\sigma_h^i)^2\Sigma)$

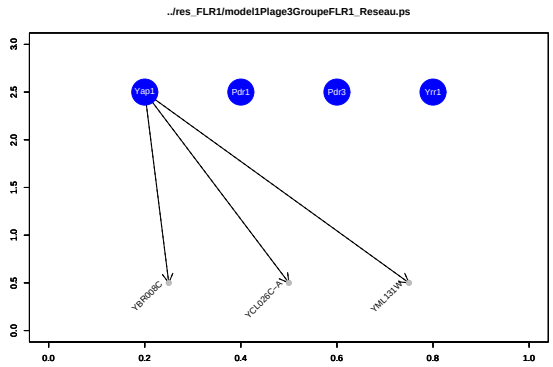

Changepoints Detection
TP rate
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.0
1.5
2.0
Noise
10 neighbours
8 neighbours
6 neighbours
4 neighbours
2 neighbours

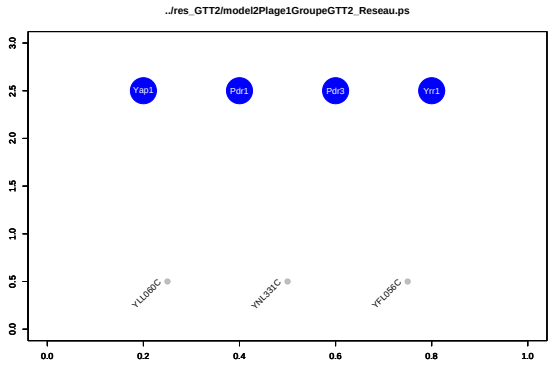

Edges Detection
TP rate
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.0
1.5
2.0
Noise
10 neighbours
8 neighbours
6 neighbours
4 neighbours
2 neighbours

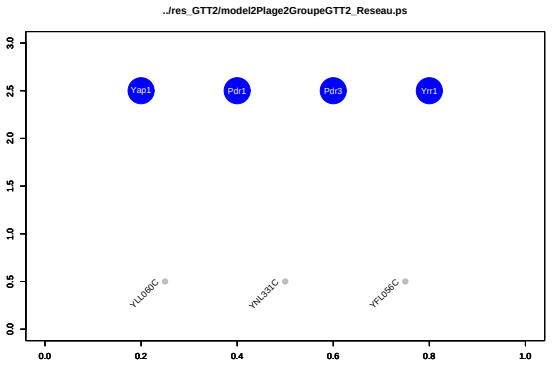
FLR1 and GTT2: time-delayed YAP1 targets

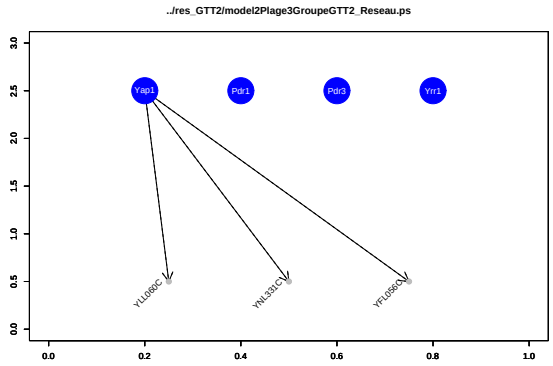




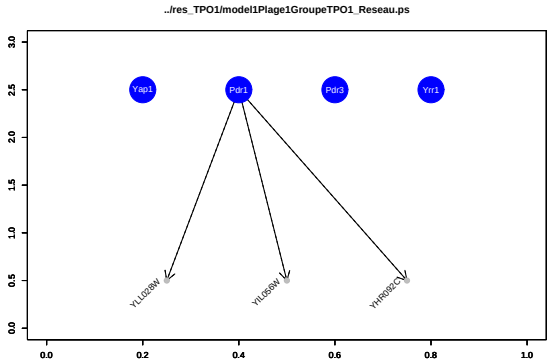


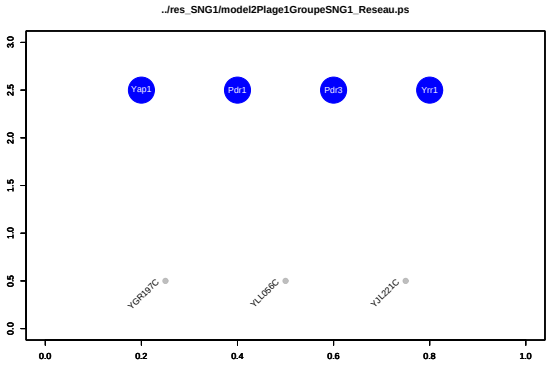


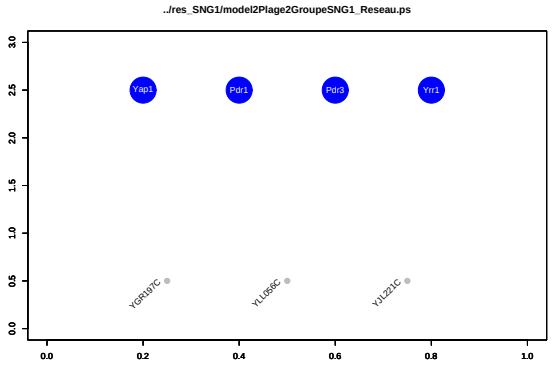


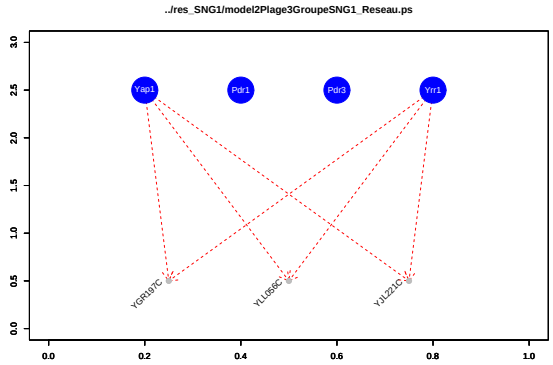


TPO1 and SNG1

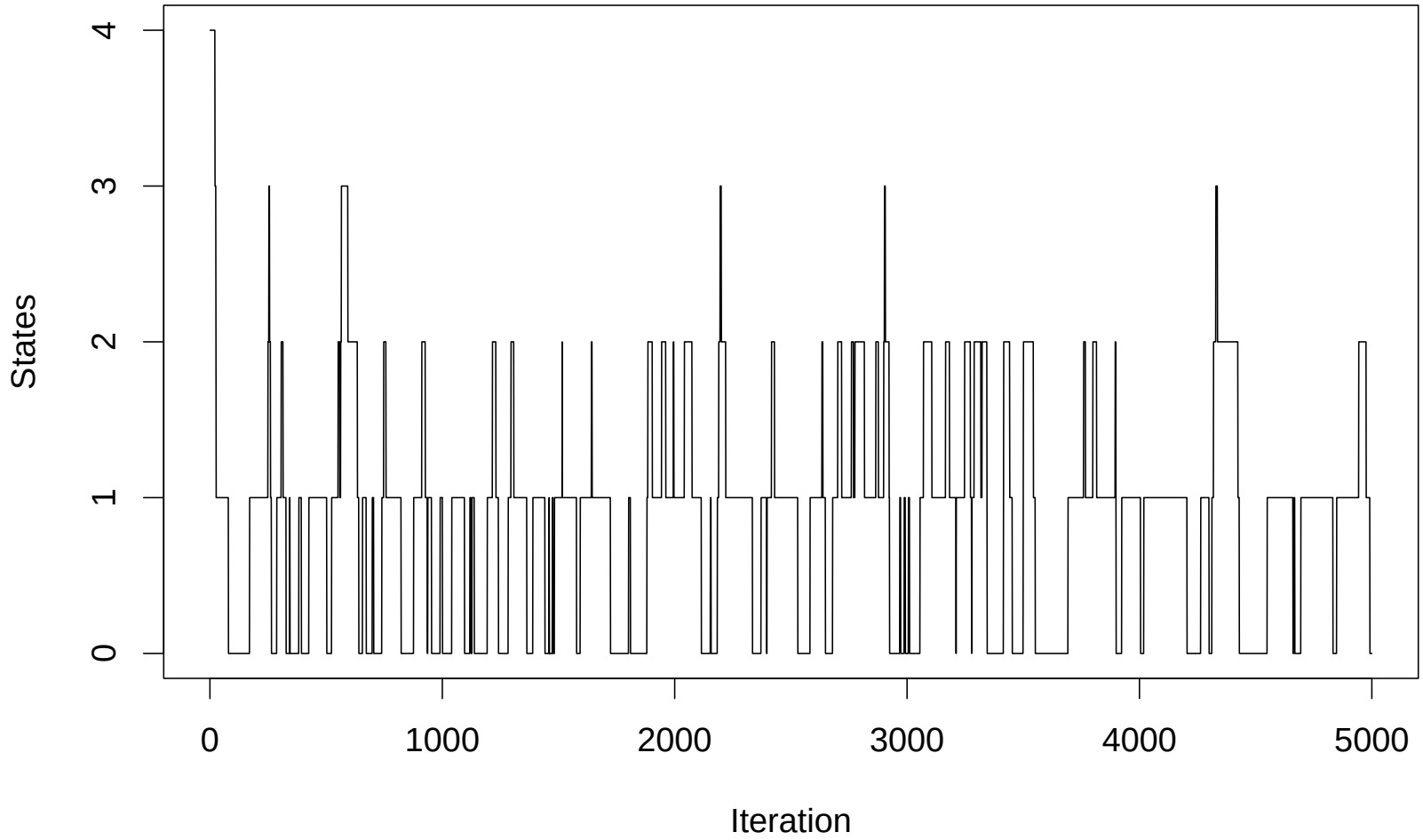




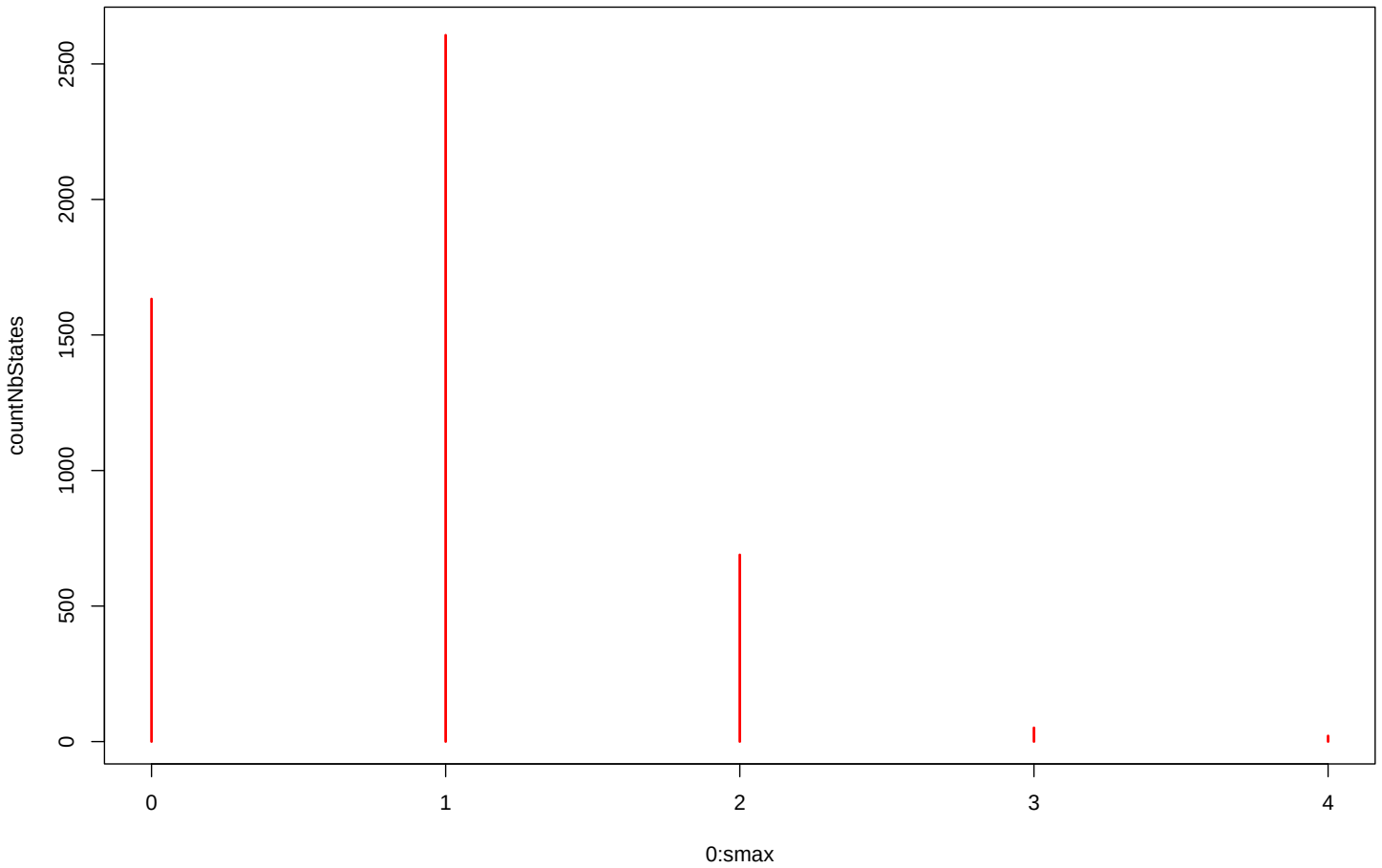




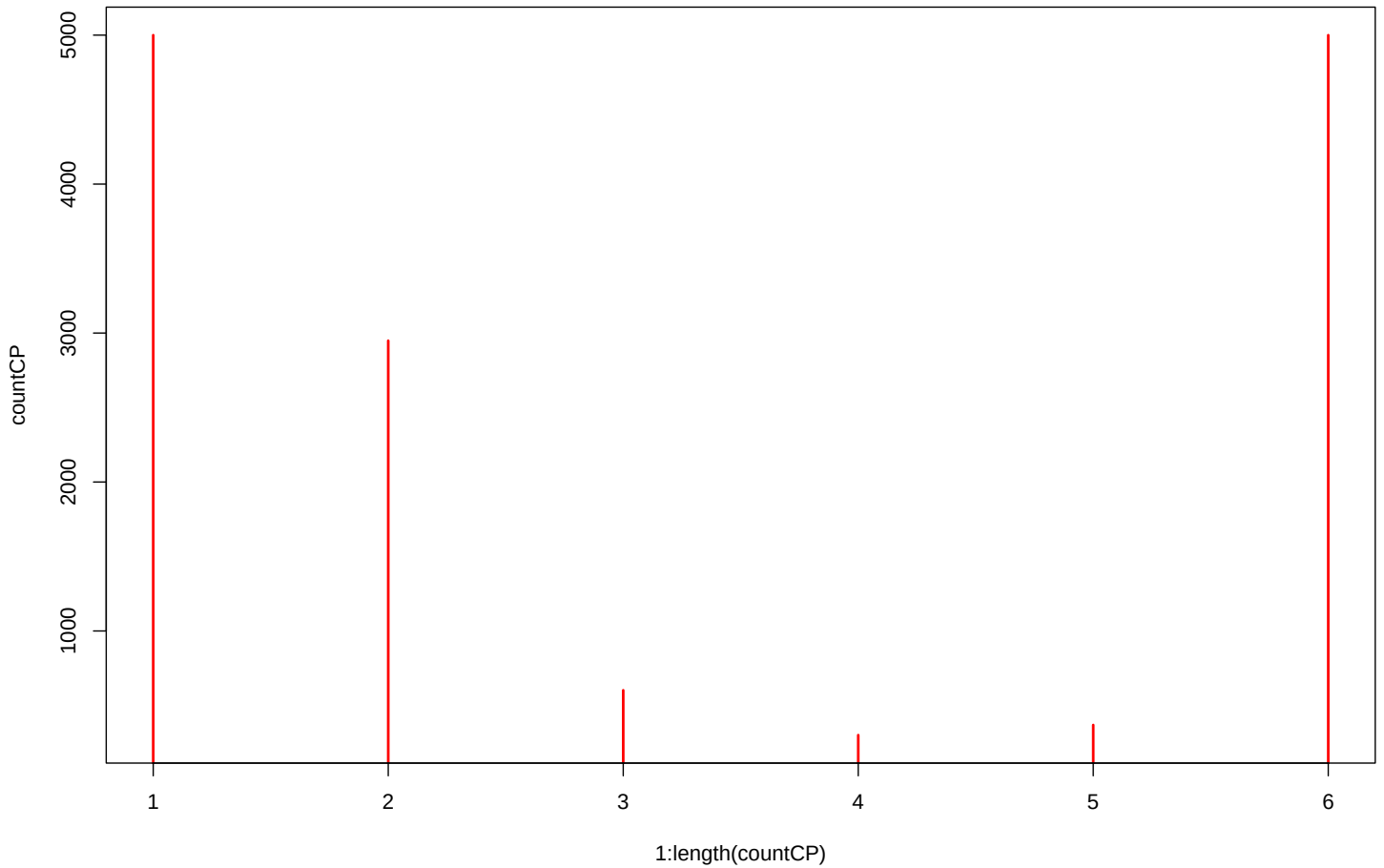
Résultats : nombre de points de ruptures.



Nombre de points de rupture.



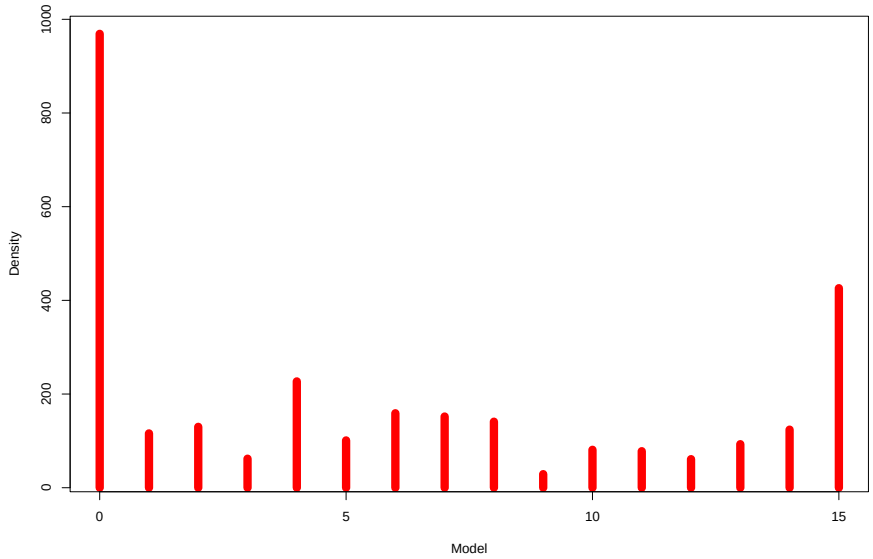
Temps de séjour pour chaque point de rupture.

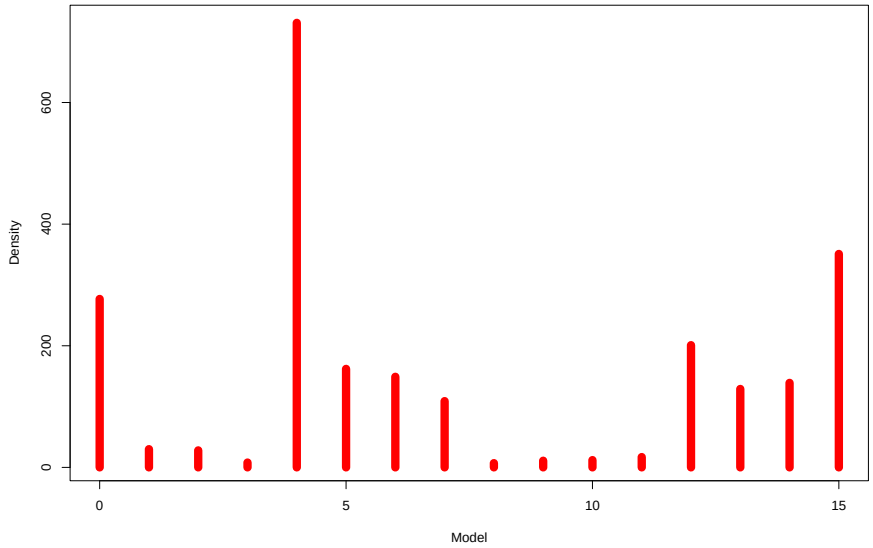


Loi a posteriori du vecteur de ruptures

Ordre	Probabilité a Posteriori	Vecteur de ruptures
1	0.4722	(1,2,6)
2	0.3266	(1,6)
3	0.0816	(1,2,3,6)
4	0.0242	(1,4,5,6)
5	0.0218	(1,2,5,6)
6	0.0206	(1,3,6)
7	0.015	(1,4,6)
8	0.0134	(1,5,6)
9	0.0046	(1,3,4,6)
10	0.0042	(1,2,3,4,5,6)

Loi a posteriori pour chaque phase.





Données Bénomyl