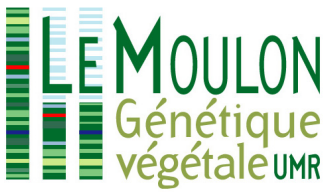
Approximation variationnelle pour des HMM multivariés en vue de la détection de CNV chez le mais

Xiaoqiang WANG, Emilie Lebarbier, Julie Aubert & Stéphane Robin

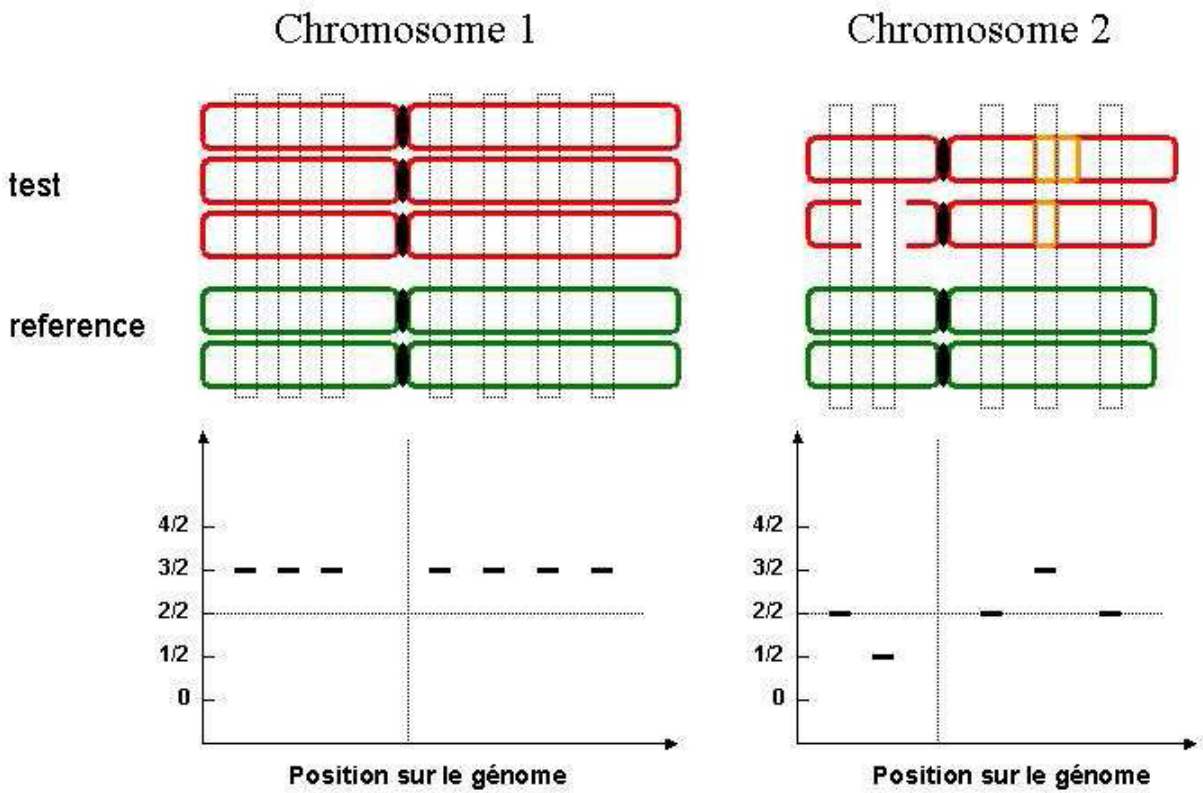
UMR de Génétique Végétale INRA/Paris Sud/CNRS & UMR Mathématiques et Informatiques Appliquées INRA/AgroParisTech

MSTGA, May 17 2013, Toulouse



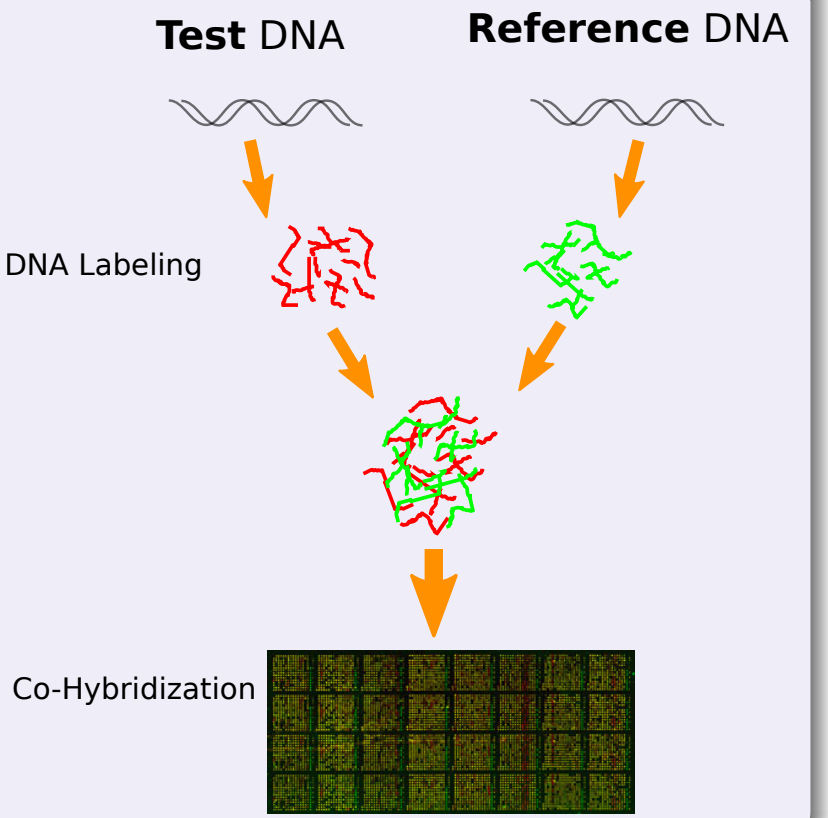
Context

- **Copy Number Variant (CNV)**
  - Alterations of the number of copies of one or more sections of the DNA in genome
- **Technology :** Comparative Genomic Hybridization (**CGH**)
- **CGH in its principle**

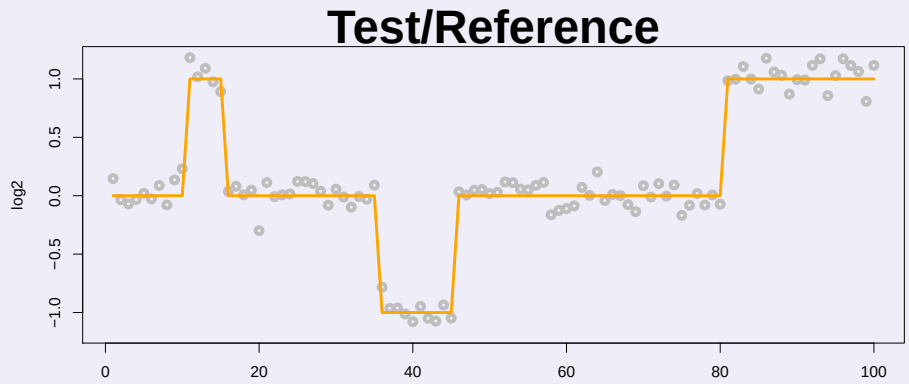


Context

CGH: CNV (Test/Reference)



Signal at position t: $Y(t)$



- X: Genomic position
- Y: log2 (♯ Test / ♯ Reference)

- **Objectif**: Classify the genomic regions as, -1 ("deleted"), 0 ("normal"), 1 ("amplified").

Methodology

- **Notations**
  - Hidden status : $\{S_t\} \sim MC(\pi)$, $\quad \pi_{k\ell} = \mathbb{P}(S_t = \ell | S_{t-1} = k)$
  - Observed data :$\{Y_t\}$ are independent conditionally to $\boldsymbol{S}$.

- **Model**

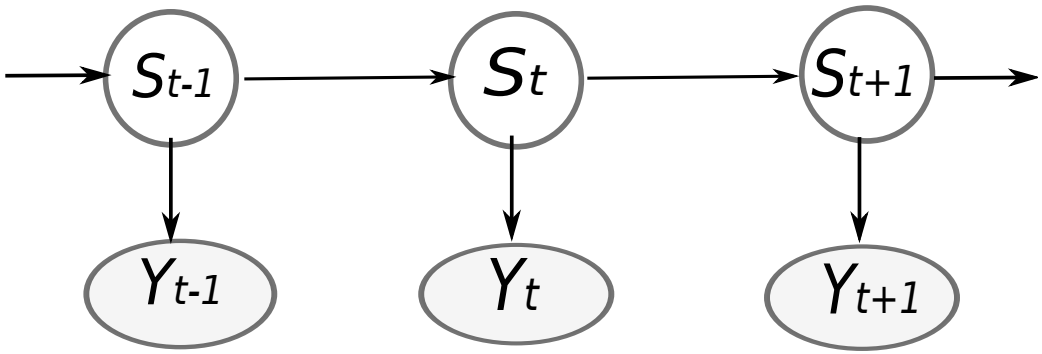
$$Y_t = \sum_k S_t^k \mu_k + \varepsilon_t,$$

where $S_t^k = \mathbb{1}_{\{S_t = k\}}$, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.

- **HMM** denoted by $\theta = (m, \pi, \gamma)$
  - m : initial state distribution
  - π : transition probability for Markov chain (S_t)
  - γ : parameter in emission probability, *i.e.*, $\gamma = (\mu, \sigma^2)$

Methodology

- **Graphical representation**



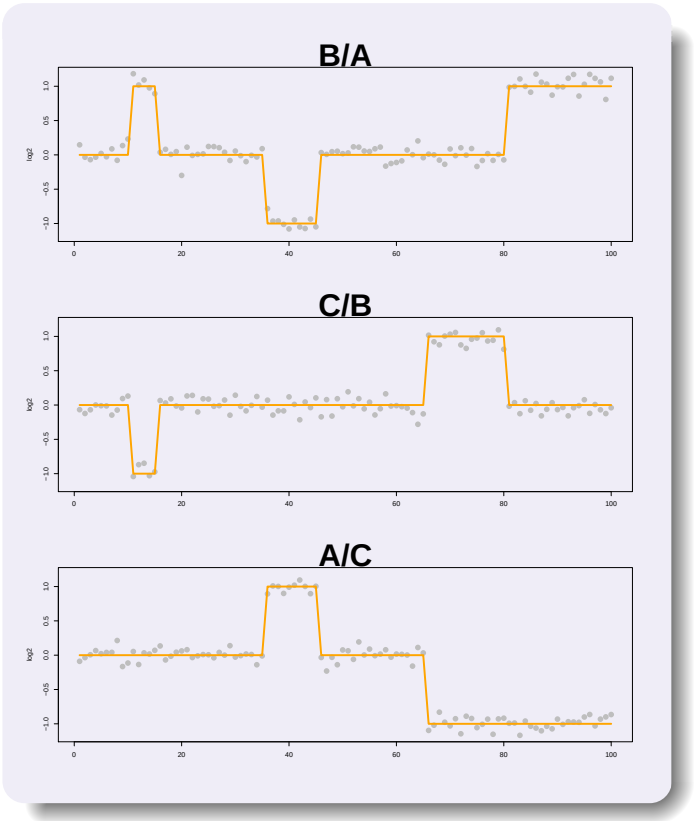
- **Complete likelihood**

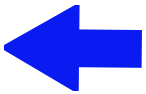
$$\mathbb{P}(\boldsymbol{Y}, \boldsymbol{S}) = \mathbb{P}(\boldsymbol{S})\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})$$

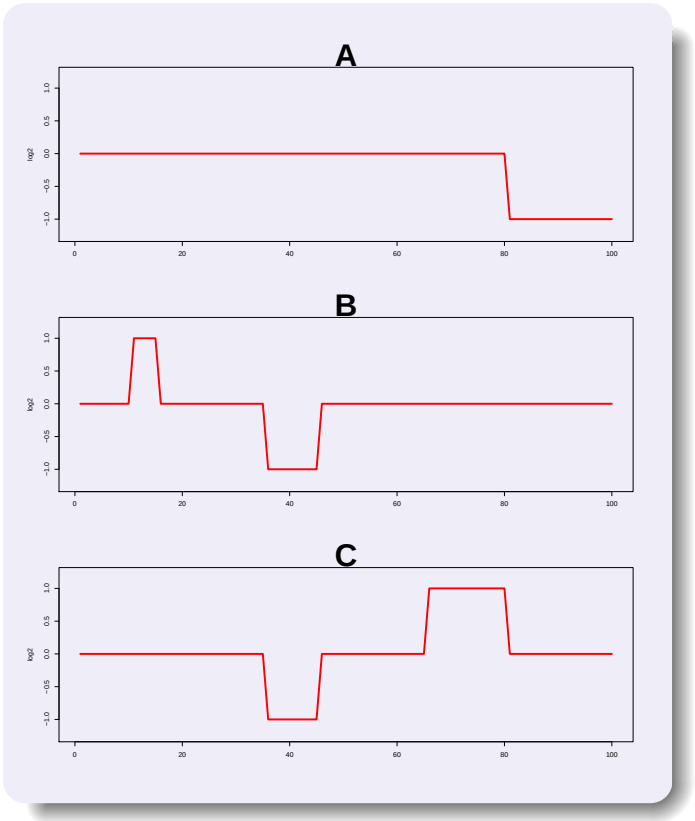
$$= \left\{ \prod_k m_k^{S_1^k} \prod_{t>1} \prod_{k,\ell} \pi_{k\ell}^{S_{t-1}^k S_t^\ell} \right\} \left\{ \prod_t \prod_\ell \phi(Y_t; \gamma_k)^{S_t^\ell} \right\}$$

- **Objectif :** Infer $\boldsymbol{S}|\boldsymbol{Y}$
- **Inference** : E-M algorithm
  - E-step: calculate $\mathbb{P}(S|Y)$ by Forward-Backward
  - M-step: estimate $\theta = \arg\max_\theta \mathbb{E}\left[\log \mathbb{P}(S, Y)|Y\right]$

Question: Dependency structure I



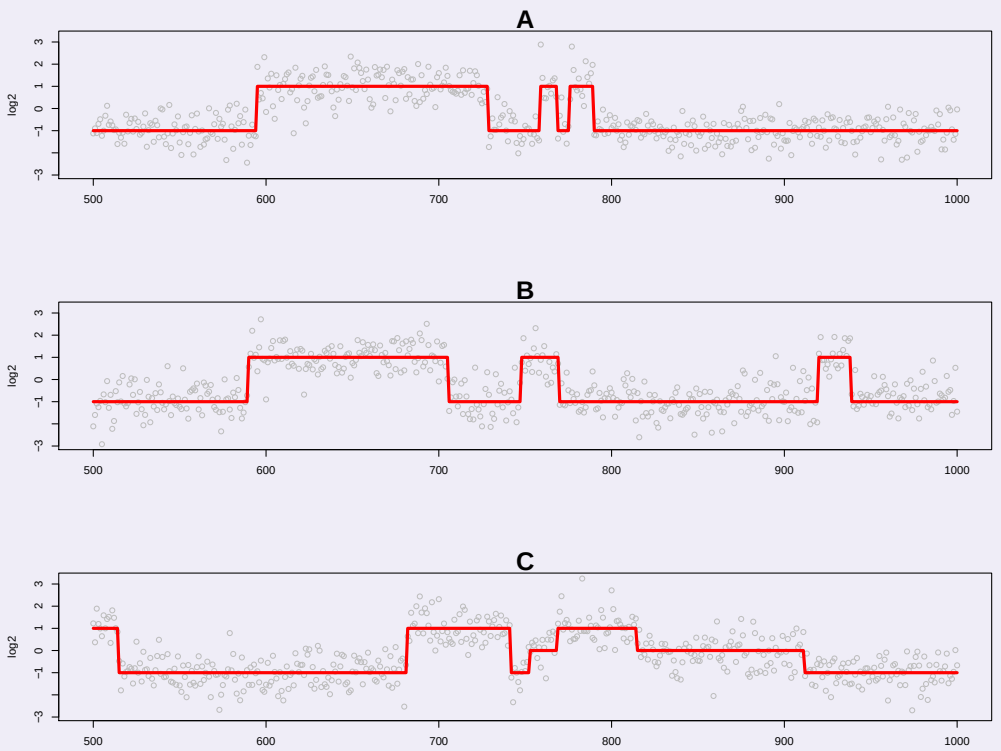




Question?

- Perform a joint analysis taking into account the dependency structure due to the experiments
- Detect CNV for lines through their comparisons

Question: Genetic dependency structure II



- The lines are genetically correlated, e.g.,
  cor(A, B) = 0.8
  cor(A, C) = 0.2
  cor(B, C) = 0.2

Question?

- Perform a joint analysis taking into account the genetic dependency structure among lines A, B and C

Multivariate HMM Model

- **Notations**
  - I: number of lines; M: number of comparisons;
  - Y_t: observation ;
  - $S_{i,t}$: hidden status and $S_{i,t}^q = \mathbb{1}_{\{S_{i,t}=q\}}$
- **Model**

$$\underset{M\times 1}{Y_t} = \underset{M\times I}{D}\ \underset{I\times 1}{\eta_t} + \varepsilon_t,$$

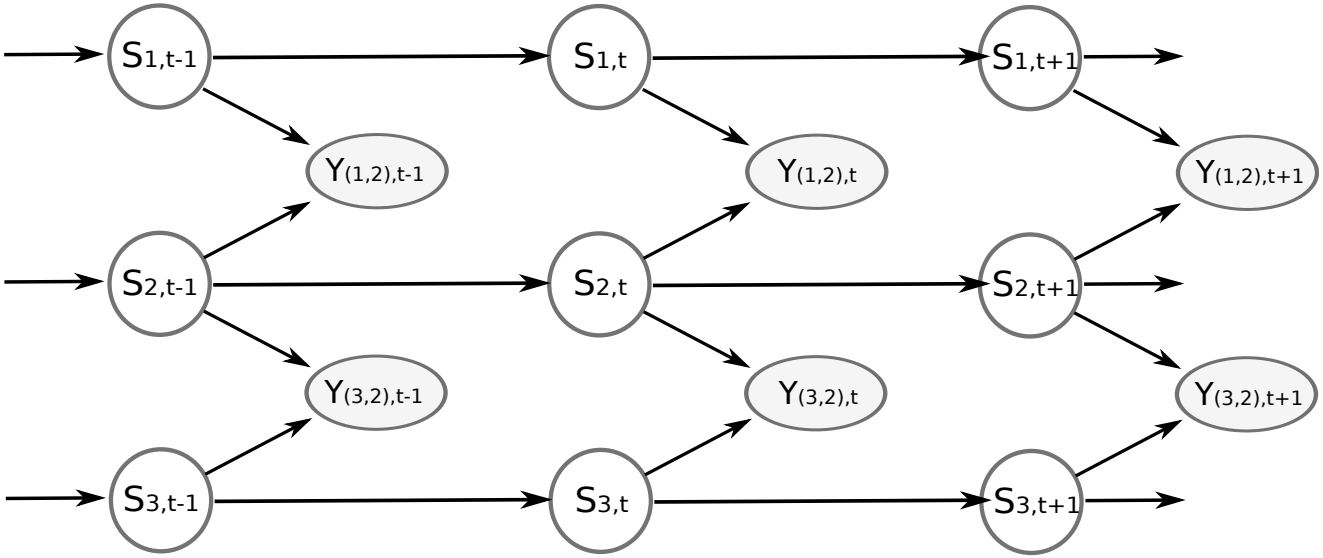
  - $\eta_t = (\eta_{1,t}, \ldots, \eta_{I,t})^T$ with $\eta_{i,t} = \sum_q \mu_q S_{i,t}^q$,
  - D : design matrix, e.g.,

$$\begin{pmatrix} Y_{\text{B/A}} \\ Y_{\text{C/B}} \\ Y_{\text{A/C}} \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} \eta_{\text{A}} \\ \eta_{\text{B}} \\ \eta_{\text{C}} \end{pmatrix}$$

  - $\varepsilon_t \sim \mathcal{N}(0, \sigma^2 I_M)$

Multivariate HMM Model

- **Graphical representation**

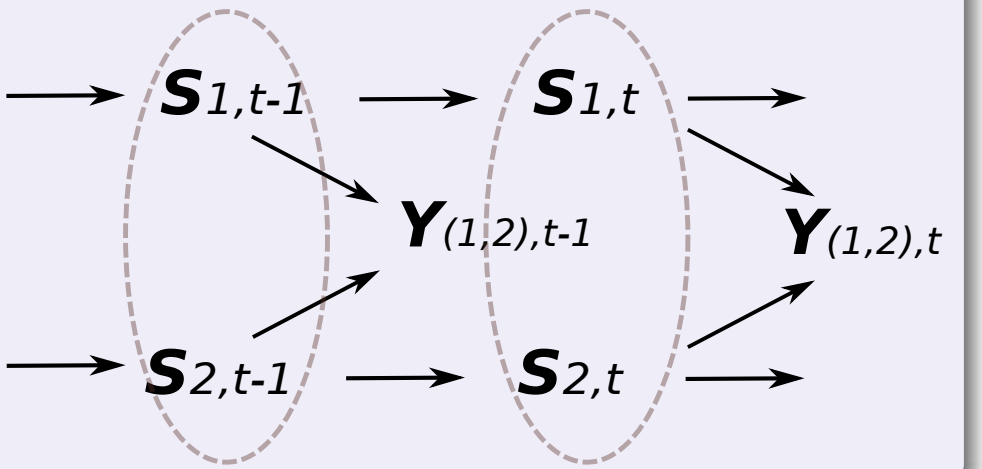


  - $S_{1,t}$: hidden status at position t for line 1, e.g. -1, 0, 1.
  - $Y_{(1,2),t}$: observation of log2(♯ line 1/ ♯ line 2) at position t.

- **Inference**
  - Small I : Exact inference by E-M algorithm
  - Big I : Approximate inference by variational E-M algorithm

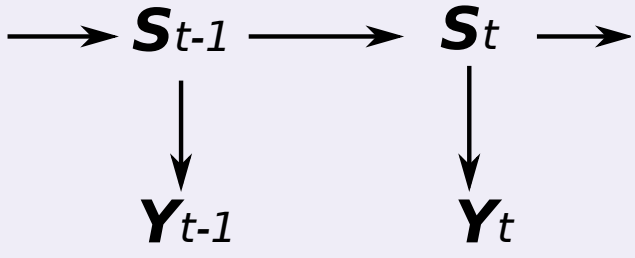
Inference when I is small

Inference



$\Rightarrow$

Global inference



$$\prod_{i,t>1}\prod_{q,r}\pi_{qr}^{S^q_{i,t-1}S^r_{i,t}}\prod_{(i,j),t}\prod_{q,r}\phi(Y_{(i,j),t};\mu_{qr},\sigma^2)^{S^q_{i,t}S^r_{j,t}} \Rightarrow \prod_{t>1}\prod_{k,\ell}\Pi_{k\ell}^{S^k_{t-1}S^\ell_t}\prod_{t}\prod_{\ell}\phi(Y_t;\gamma_k)^{S^\ell_t}$$

- For example

$$(S_{1,t},S_{2,t})=(1,1)\Leftrightarrow S_t=1$$
$$(S_{1,t},S_{2,t})=(1,2)\Leftrightarrow S_t=2$$

Inference when *I* is large

- **Question** : $\mathbb{P}(S_t|Y)$ can not be computed when ***I*** is large, e.g, if $I = 10$ and $Q = 3$, S_t has $3^{10} = 59049$ status to be considered.
- **Solution** : $\tilde{\mathbb{P}}(S) \cong \mathbb{P}(S|Y)$
- **Reminder** : Kullback-Leibler divergence

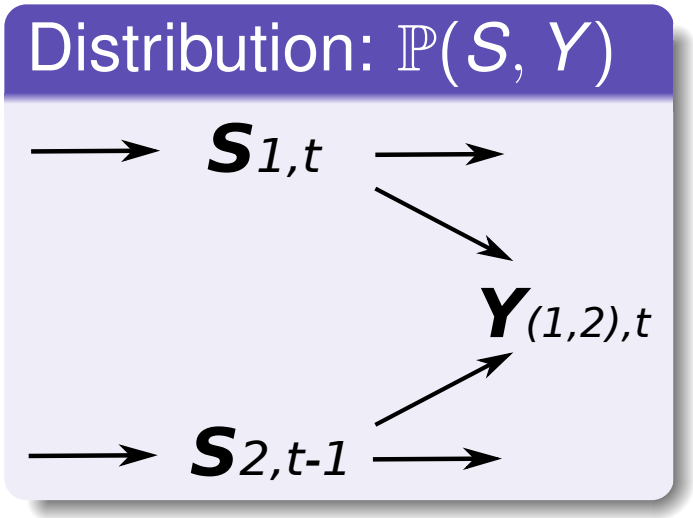
$$KL\left[\tilde{\mathbb{P}}(S)||\mathbb{P}(S|Y)\right] = \int \tilde{\mathbb{P}}(S) \log \frac{\tilde{\mathbb{P}}(S)}{\mathbb{P}(S|Y)}\, dS$$

  - KL is always non-negative
  - Null iff $\tilde{\mathbb{P}} = \mathbb{P}$

- $$\tilde{\mathbb{P}}(S) = \arg\min_{\tilde{\mathbb{P}} \in \mathcal{D}} KL\left[\tilde{\mathbb{P}}(S)||\mathbb{P}(S|Y)\right]$$

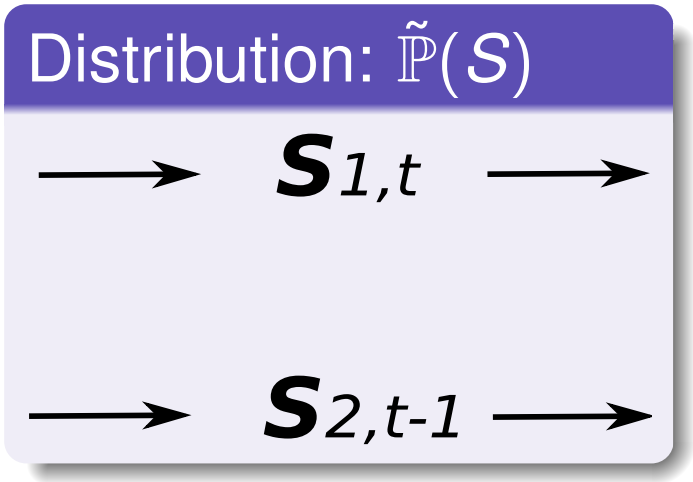
Inference when I is large

- **Graphical representation**



$$KL[\tilde{\mathbb{P}}||\mathbb{P}]$$

$$\Longleftrightarrow$$



$$\prod_t \mathbb{P}(S_t|S_{t-1})\mathbb{P}(Y_t|S_t)$$

$$\prod_i \prod_t \tilde{\mathbb{P}}(S_{i,t}|S_{i,t-1})$$

- **Key** : find $\tilde{\mathbb{P}}(S_{i,t}|S_{i,t-1})$ to break the dependency

Inference when I is large

Proposition

Let denote $p_{itqr} = \tilde{\mathbb{P}}(S_{i,t} = r | S_{i,t-1} = q)$, then we obtain a set of fixed point equations for p_{itqr}:

$$p_{itqr} \propto \pi_{qr} \prod_{j \in m^-(i), v} \phi(Y_{(i,j),t}; \mu_{rv}, \sigma^2)^{\mathbb{E}_{\tilde{\mathbb{P}}}(S^v_{j,t})} \times \prod_{j \in m^+(i), u} \phi(Y_{(j,i),t}; \mu_{ur}, \sigma^2)^{\mathbb{E}_{\tilde{\mathbb{P}}}(S^u_{j,t})}$$

where π is the transition probability and ϕ emission probability.

- **Inference by variational E-M algorithm**

Inference when I is large

- **Variational E-step**: find

$$\tilde{\mathbb{P}}^{(h)}(S) = \arg\min_{\tilde{\mathbb{P}} \in \mathcal{P}} KL\left[\tilde{\mathbb{P}}(S) \| \mathbb{P}(S|Y, \theta^{(h)})\right],$$

where $\mathcal{P} = \left\{\tilde{\mathbb{P}}(S) \,|\, \tilde{\mathbb{P}}(S) \propto \prod_i \prod_t \tilde{\mathbb{P}}(S_{i,t}|S_{i,t-1})\right\}$

- **Variational M-step**: estimate $\theta = (m, \pi, \mu, \sigma^2)$ as

$$\hat{\theta}^{(h+1)} = \arg\max_{\theta} \mathbb{E}_{\tilde{\mathbb{P}}^{(h)}} \left[\log \mathbb{P}(Y, S|\theta)\right].$$

  - $\hat{m}_q, \hat{\pi}_{qr}$

$$\hat{m}_q \propto \sum_i \mathbb{E}_{\tilde{\mathbb{P}}}(S^q_{i,1}) \qquad \hat{\pi}_{qr} \propto \sum_{i,t} \mathbb{E}_{\tilde{\mathbb{P}}}(S^q_{i,t-1} S^r_{i,t})$$

  - $\hat{\mu}, \hat{\sigma^2}$

$$(\hat{\mu}, \hat{\sigma^2}) = \arg\max_{\mu, \sigma^2} \sum_{t,(i,j)} \sum_{q,r} \mathbb{E}_{\tilde{\mathbb{P}}}(S^q_{i,t}) \mathbb{E}_{\tilde{\mathbb{P}}}(S^r_{j,t}) \log \phi(Y_{(i,j),t}; \mu_{qr}, \sigma^2)$$

Simulation and Discussion

1. **Input**

	Simulation I	Simulation II
I	3	5
M	6	20
μ	(-1,0,1)	(-1,0,1)
σ^2	0.36	0.36
T	5000	5000
Nb. Sim.	100	100

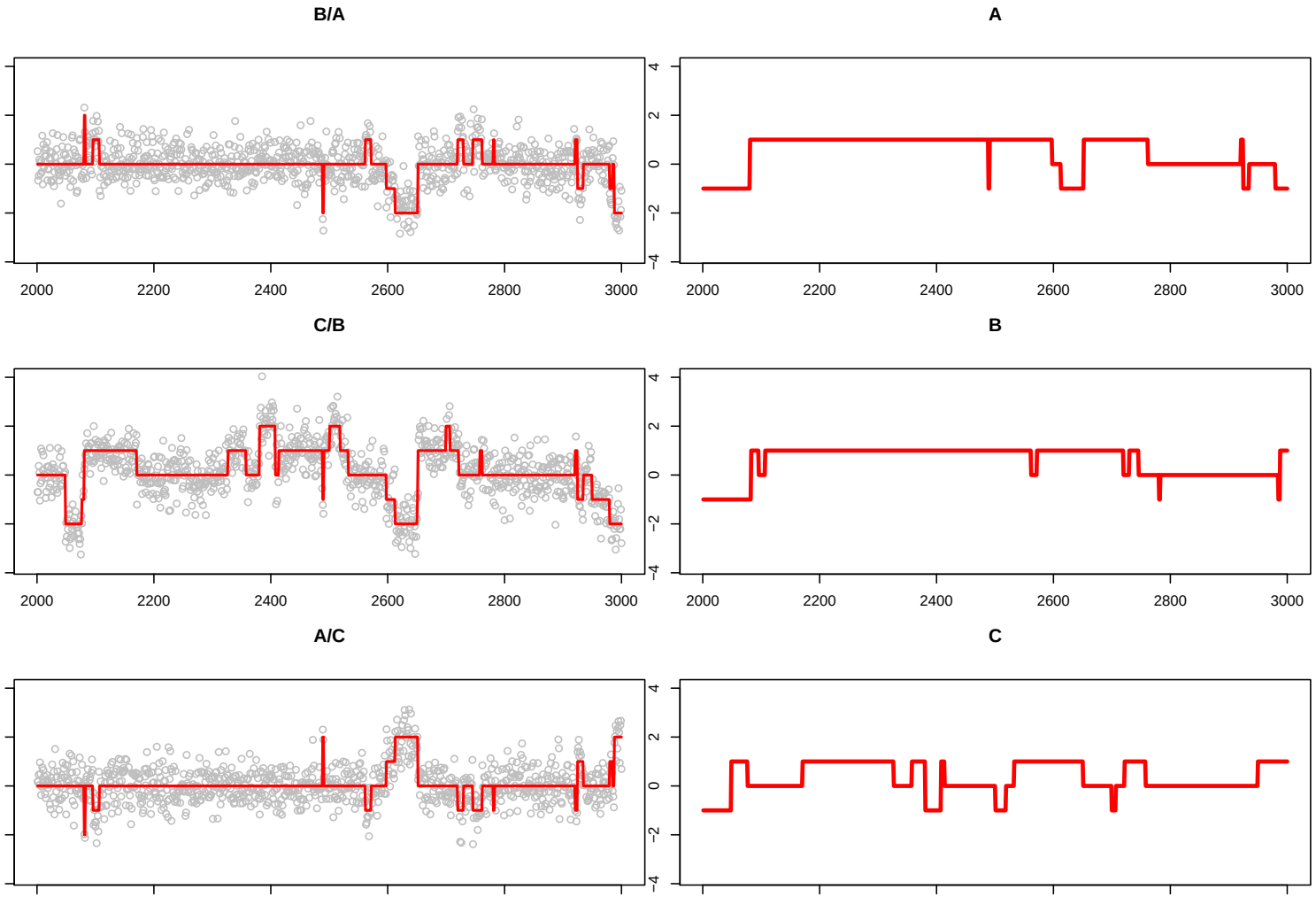
2. **Multivariate HMM**: perform a joint analysis with exact inference and variational inference, respectively.

3. **Output**

$$\hat{m}, \quad \hat{\pi}, \quad \hat{\mu}, \quad \hat{\sigma}, \quad \mathbb{P}(S_{i,t} = k|Y)$$

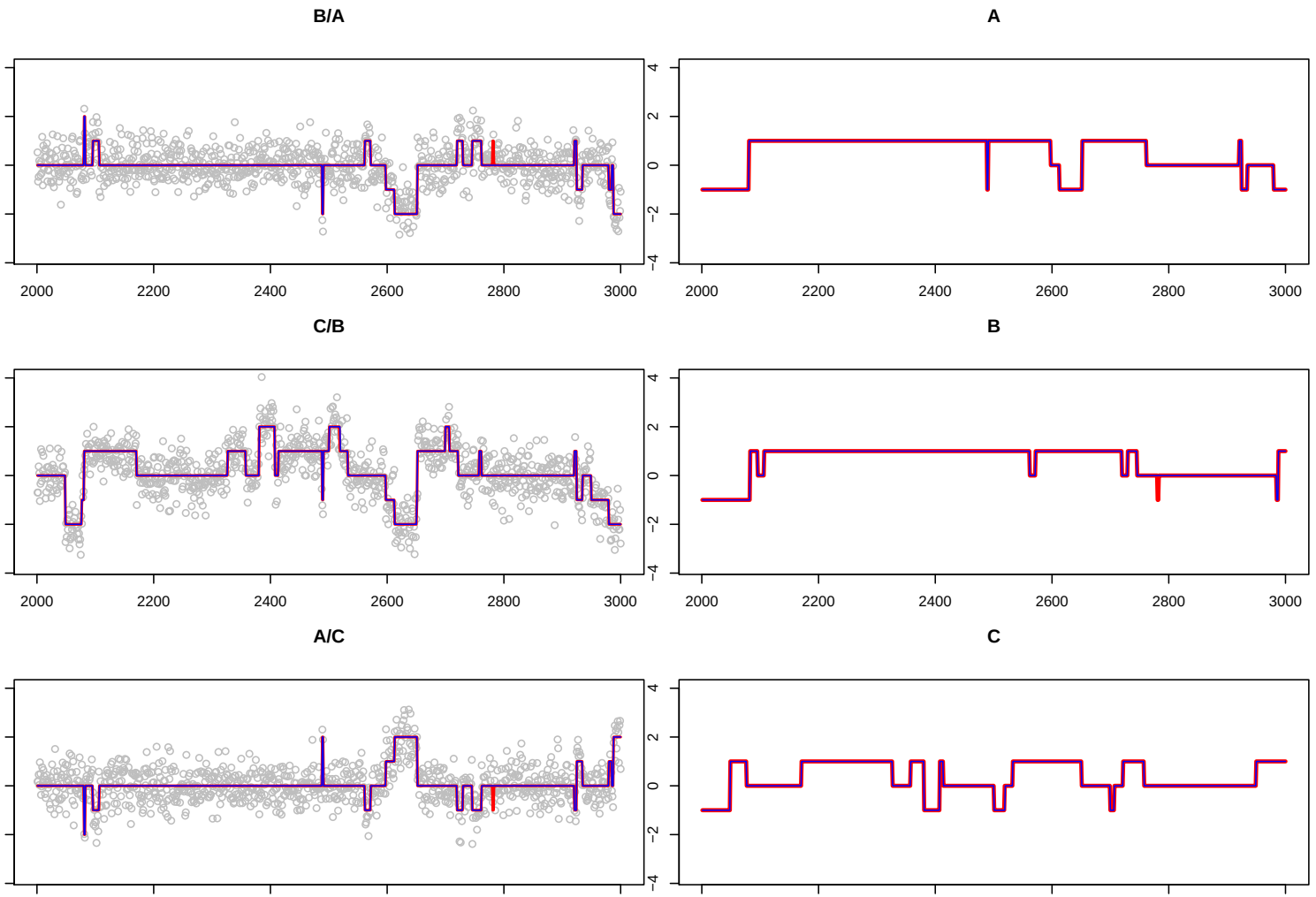
Simulation and Discussion

- One simulation :



Simulation and Discussion

- Result simulation : red (true), blue (estimation)



Simulation and Discussion

- **Time**
  - Mean time for one multivariate analysis

	Exact	Variational
$I = 3$	58 s	2 s
$I = 5$	396 s	13 s

- **Accuracy**
  - Error rate for comparison between lines, e.g. $B/A, A/C, \ldots$

	Exact	Variational
$I = 3$	0.15%	1.1%

  - Error rate for lines, e.g. $A, B, \ldots$

	Exact	Variational
$I = 3$	5.4%	29%

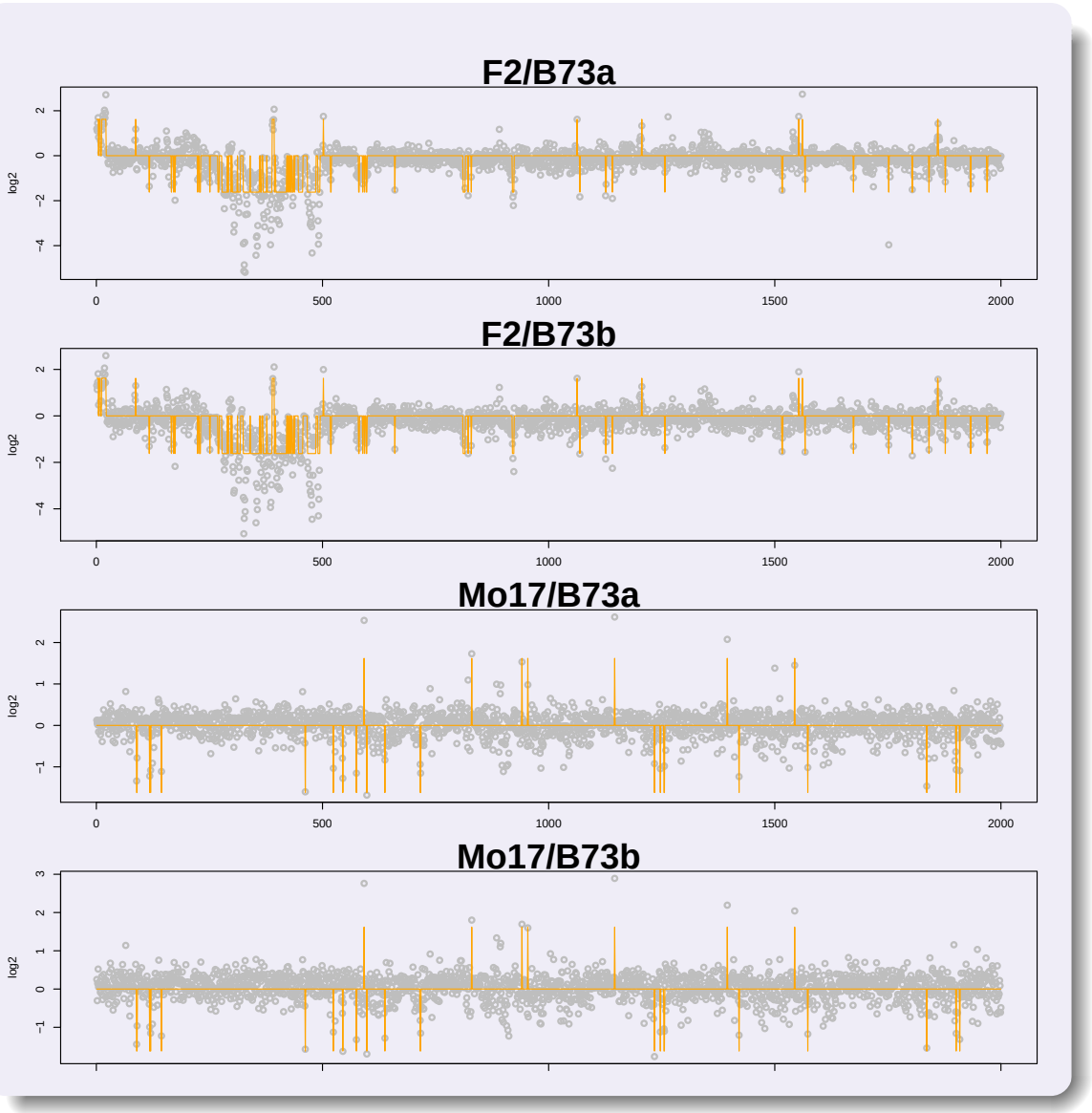
Biogemma data analysis

- Experimental design
  - **Lines:** F2, Mo17, B73 (**Reference**)
  - **Comparisons:** F2/B73a, F2/B73b, Mo17/B73a, Mo17/B73b
- Information on Chromosome
  - 2 139 527 probes (data point) on 10 chromosomes

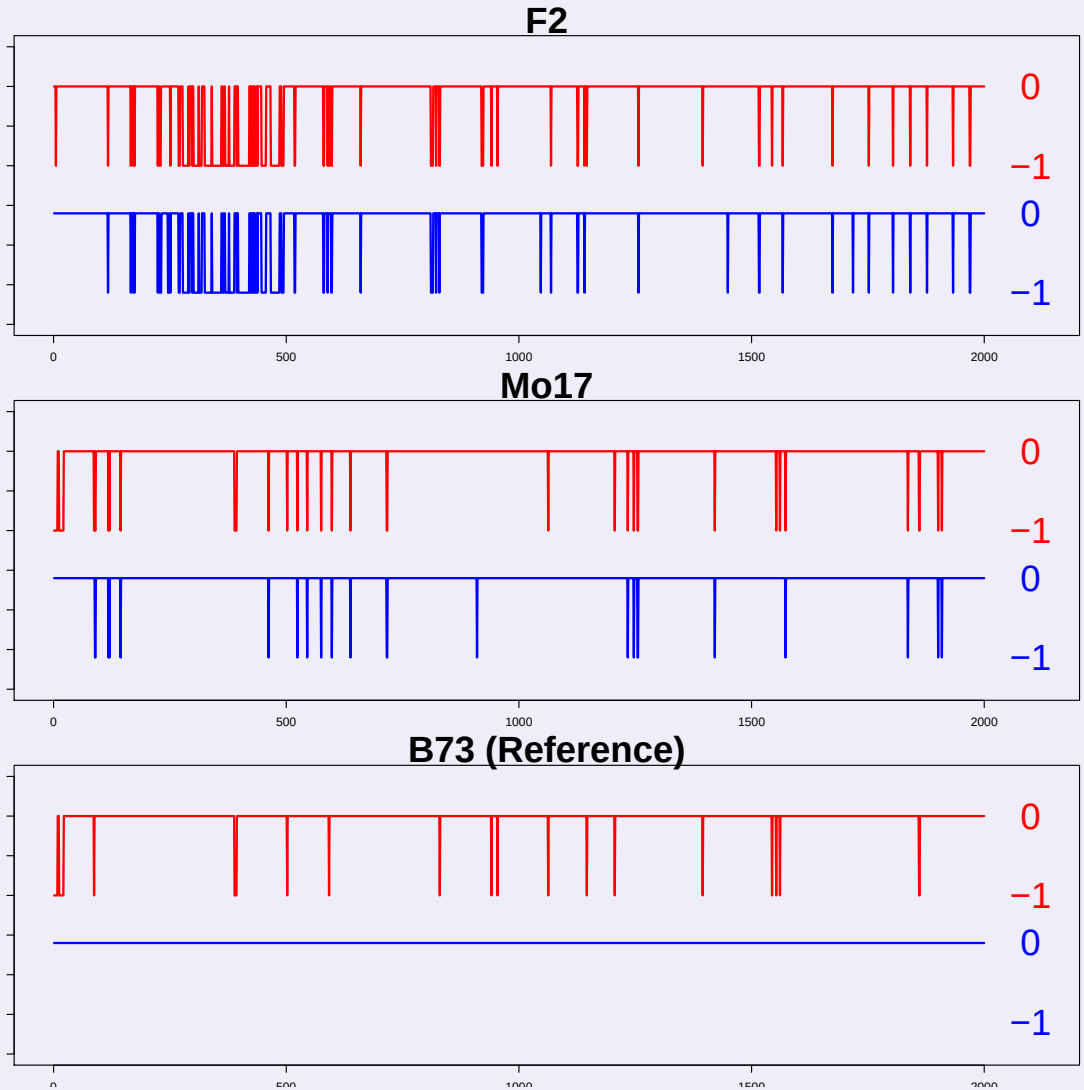
Chromosome	1	2	3	4	5
Probe size	323410	252569	243428	260526	220582

Chromosome	6	7	8	9	10
Probe size	173328	182823	175177	156925	150759

- Joint analyses one by one chromosome



- Chromosome 1
- First 2000 probes
- Orange : mean

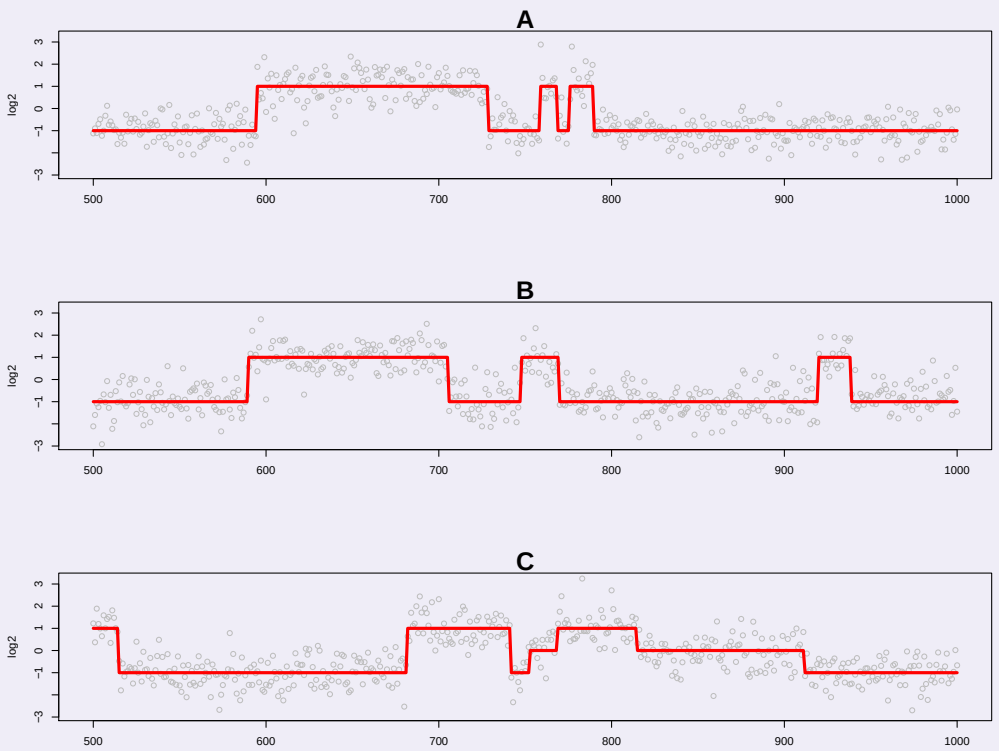


- Two joint analyses
  - Red : Multivariate HMM
  - Blue : Multivariate HMM but forcing B73 to be 0

- Different status

	Red $\neq$ Blue
F2	0.9%
Mo17	0.45%
B73	0.65%
Total	0.67%

Question: Genetic dependency structure II



- The lines are genetically correlated, e.g.,
  cor(A, B) = 0.8
  cor(A, C) = 0.2
  cor(B, C) = 0.2

Question?

- Perform a joint analysis taking into account the genetic dependency structure among lines A, B and C

Multivariate HMM Model

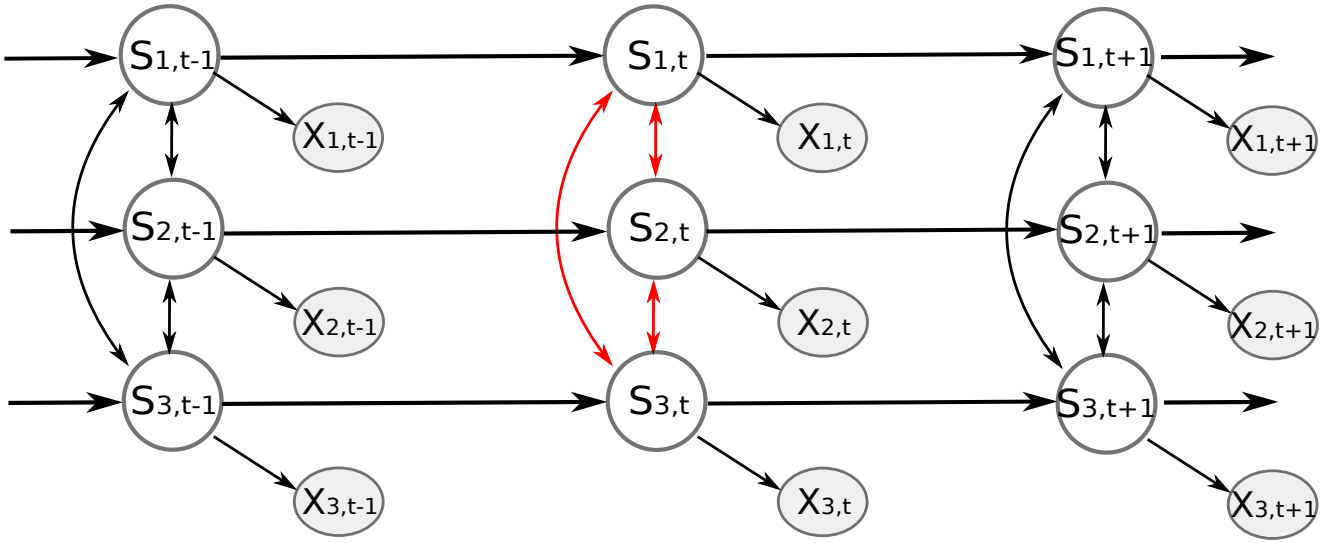
- **Notations**
  - I: number of lines
  - $X_{i,t}$: observation
  - $S_{i,t}$: hidden status but $\forall i \neq j$, $(S_{i,t}, S_{j,t})$ are not independent
- **Model**

$$X_{i,t} = \sum_q S_{i,t}^q \, \mu_q + \varepsilon_t$$

  - $S_{i,t}^q = \mathbb{1}_{\{s_{i,t}=q\}}$
  - $\varepsilon_t \sim \mathcal{N}(0, \sigma^2 I_I)$
- **Inference**
  - Small data: Exact inference by E-M algorithm
  - Big data : Approximate inference by variational E-M algorithm

Inference

- **Graphical representation**



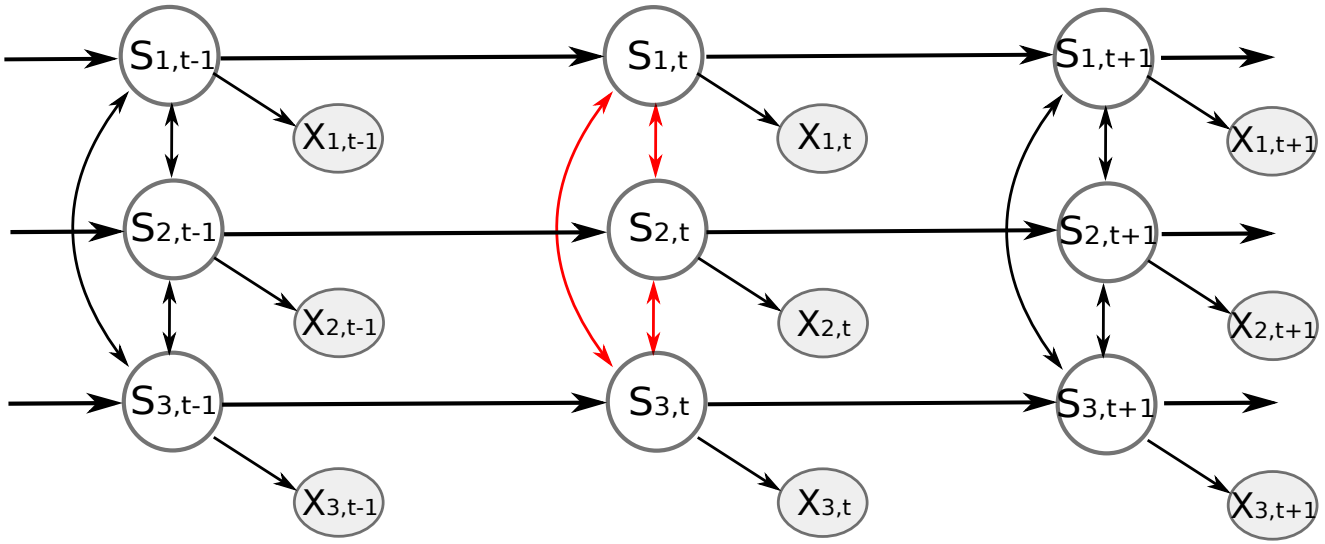
- **Complete likelihood**

$$\mathbb{P}(\boldsymbol{X}, \boldsymbol{S}) \propto \prod_k m_k^{S_1^k} \prod_{t>1} \prod_{k,\ell} \Pi_{k\ell}^{S_{t-1}^k S_t^\ell} W_\ell^{S_t^\ell} \prod_t \prod_\ell \phi(X_t; \gamma_\ell)^{S_t^\ell}$$

where $W_\ell = \mathbb{P}(S_\cdot = \ell)$.

Inference

- **Graphical representation**



- For the part of cyclic

Model

Let $W_\ell = \mathbb{P}(S_\cdot = \ell)$,

$$W_\ell \propto \prod_{i,j \neq i} \omega^{s_{ij} \mathbb{1}_{\{q_j^\ell \neq q_i^\ell\}}}$$

with $\omega < 1$, $(s_{ij})_{ij}$ is the similarity matrix.

Inference

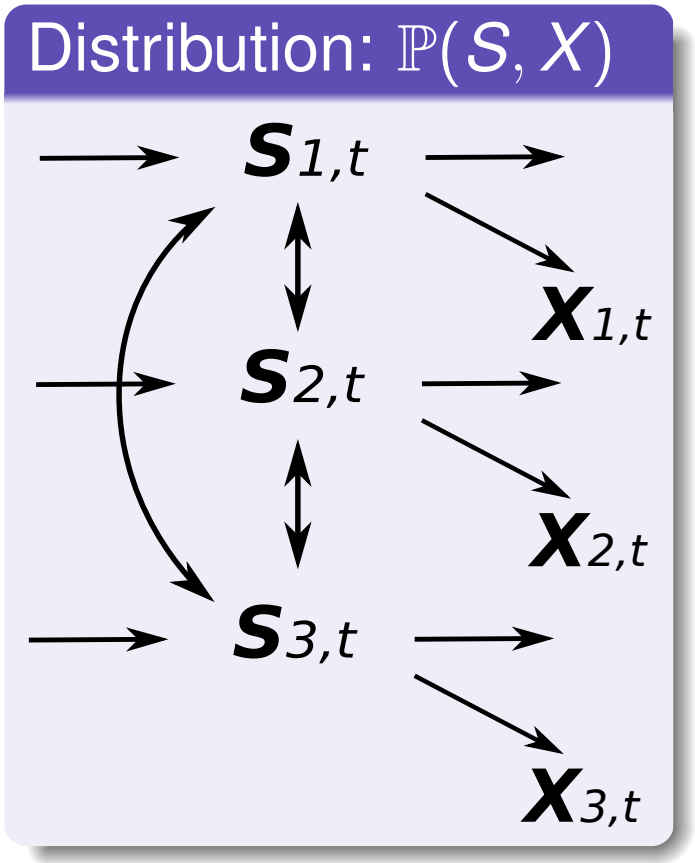
- **Small** I : exact inference with E-M algorithm is similar to that of dependency I.
- **Big** I : $\mathbb{P}(S|X)$ is not computable, find

$$\tilde{\mathbb{P}}(S) \cong \mathbb{P}(S|X)$$

in terms of Kullback-Leibler divergence.

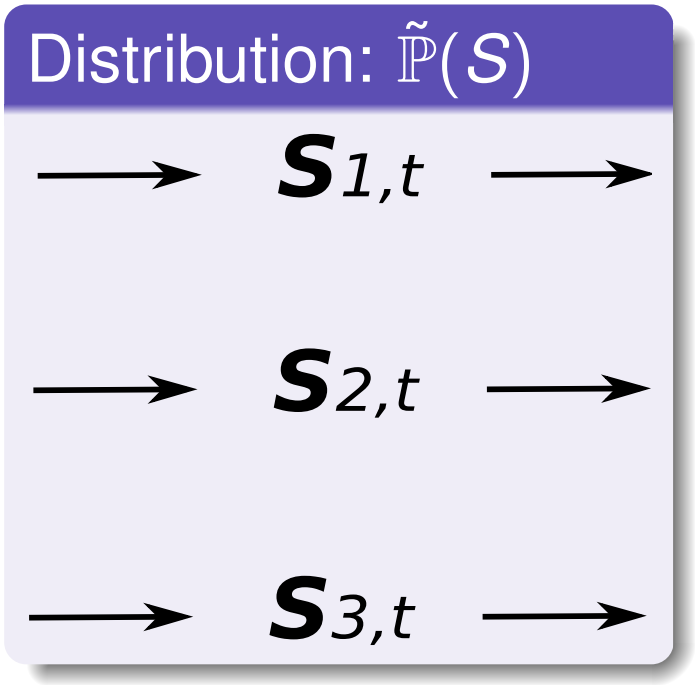
Inference when I is large

- **Graphical representation**



$$KL[\tilde{\mathbb{P}}||\mathbb{P}]$$

$$\Longleftrightarrow$$



$$\prod_i \prod_t \tilde{\mathbb{P}}(S_{i,t}|S_{i,t-1})$$

Inference when I is large

Proposition

Let denote $p_{itqr} = \tilde{\mathbb{P}}(S_{i,t} = r | S_{i,t-1} = q)$, then we obtain a set of fixed point equations for p_{itqr}:

$$p_{itqr} \propto \pi_{qr}\, \omega^{\sum_{j \neq i} \left(1 - \mathbb{E}_{\tilde{\mathbb{P}}} S^r_{j,t}\right)}\, \phi(X_{i,t}, \mu_r, \sigma^2),$$

where π is the transition probability and ϕ emission probability.

- **Variational E-step**: find

$$\tilde{\mathbb{P}}(S) = \arg\min_{\tilde{\mathbb{P}} \in \mathcal{P}} KL\left[\tilde{\mathbb{P}}(S) \| \mathbb{P}(S|X)\right],$$

where $\mathcal{P} = \left\{\tilde{\mathbb{P}}(S) \,|\, \tilde{\mathbb{P}}(S) \propto \prod_i \prod_t \tilde{\mathbb{P}}(S_{i,t}|S_{i,t-1})\right\}$;

Inference when I is large

- **Variational M-step**: estimate $\theta = (m, \pi, \mu, \sigma^2)$ as

$$\hat{\theta}^{(h+1)} = \arg\max_{\theta} \mathbb{E}_{\tilde{\mathbb{P}}(h)} \left[\log \mathbb{P}(Y, S|\theta)\right].$$

  - $\hat{m}$

$$\hat{m}_q \propto \sum_i \mathbb{E}_{\tilde{\mathbb{P}}}(S^q_{i,1})$$

  - $\hat{\pi}_{qr}$

$$\hat{\pi}_{qr} \propto \sum_{i,t} \mathbb{E}_{\tilde{\mathbb{P}}}(S^q_{i,t-1} S^r_{i,t})$$

  - $\hat{\mu}$

$$\hat{\mu}_r = \sum_{i,t} \mathbb{E}_{\tilde{\mathbb{P}}}(S^r_{i,t}) X_{i,t} / \sum_{i,t} \mathbb{E}_{\tilde{\mathbb{P}}}(S^r_{i,t})$$

  - $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \sum_{i,t,r} \mathbb{E}_{\tilde{\mathbb{P}}}(S^r_{i,t})(X_{i,t} - \hat{\mu}_r)^2 / IT$$

Simulation and Discussion

1. **Input**

	Simulation I	Simulation II
I	3	5
s_{ij}	$\begin{pmatrix} 1 & 0.8 & 0.2 \\ 0.8 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0.24 & 0.32 & 0.4 \\ 0 & 0.24 & 1 & 0.48 & 0.6 \\ 0 & 0.32 & 0.48 & 1 & 0.8 \\ 0 & 0.4 & 0.6 & 0.8 & 1 \end{pmatrix}$

2. **Multivariate HMM**: perform a joint analysis with exact inference and variational inference, respectively.

3. **Output**

$$\hat{m}, \quad \hat{\pi}, \quad \hat{\mu}, \quad \hat{\sigma}, \quad \mathbb{P}(S_{i,t} = q|X)$$

Simulation and Discussion

- **Time**
  - Mean time for one joint analysis

	Exact	Variational
$I = 3$	48 s	0.2 s
$I = 5$	450 s	0.6 s

- **Accuracy**
  - Error rate for lines, e.g. $A, B, \ldots$

	Exact	Variational
$I = 3$	0.5%	4.5%

Conclusion

- **Multivariate HMM**
  - Perform the joint analysis taking into account the dependency due to the CGH experiments
  - Perform the joint analysis taking into account the dependency among plant lines
  - Perform the joint analysis taking into account two above dependencies
- **Inference**
  - When line size is large, the inference becomes impossible.
  - Variational technique provides a fast algorithm.
- **R package** "MHMM" is built, and handles 16 different cases
  - Paired / Unpaired for observation
  - Dependent / Independent for hidden status
  - Exact / Variational for inference
  - Yes / No for reference

Thank you !

- UMR AgroParisTech/INRA

  Emilie Lebarbier
  Julie Aubert
  Stéphane Robin

- UMR Moulon

  Stéphane Nicolas
  . . . . . .

- Biogemma

  Jean-Philippe Pichon
  . . . . . .