Statistical Analysis of array CGH data

F. Picard

CNRS - Laboratoire Biométrie et Biologie Evolutive

picard@biomserv.univ-lyon1.fr

FISH and molecular cytogenetics

- Aims at studying the structure function and evolution of chromosomes.

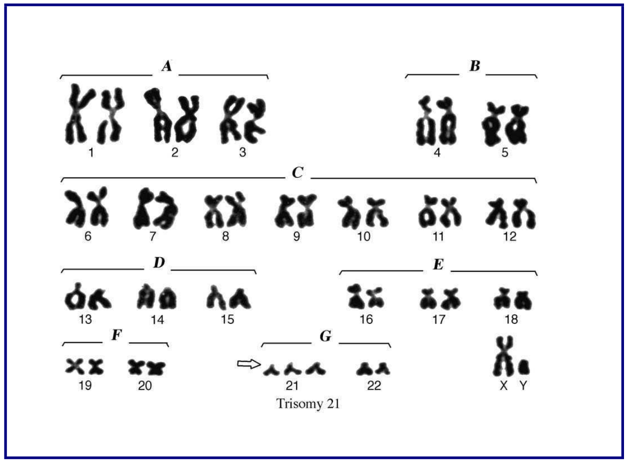
⟶ 1956: determination of the number of chromosomes in humans.

- **Objectif**: link between chromosomal defects and human pathologies.

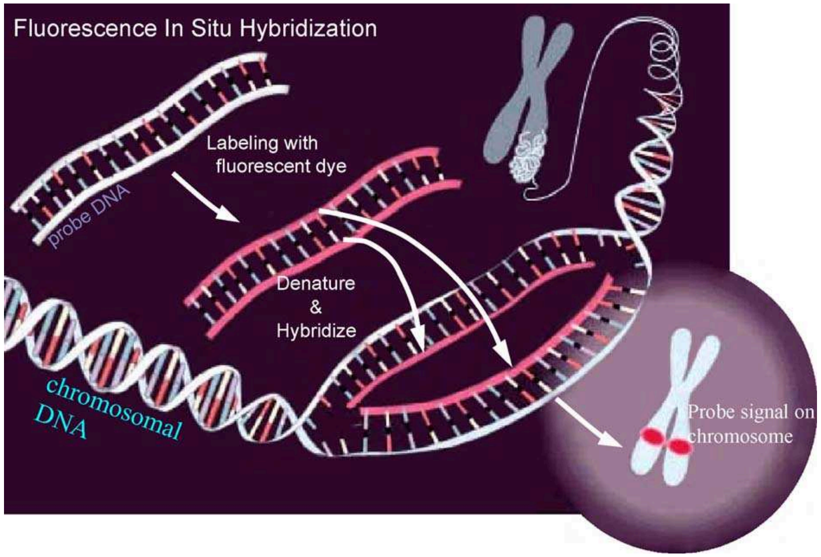
⟶ karyotype, spectral karyotyping

⟶ Fluorescence In Situ Hybridization (FISH), multiplex FISH

⟶ Comparative Genomic Hybridization (CGH), **array CGH**

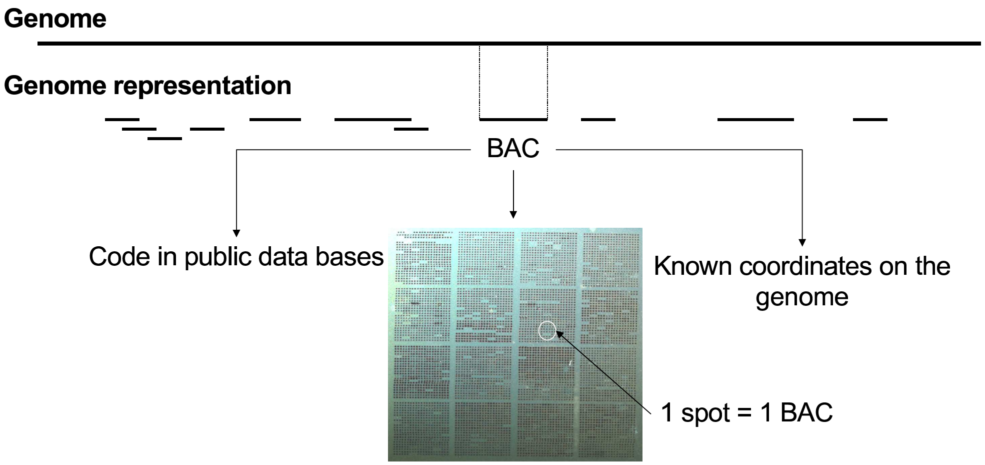


FISH and molecular cytogenetics

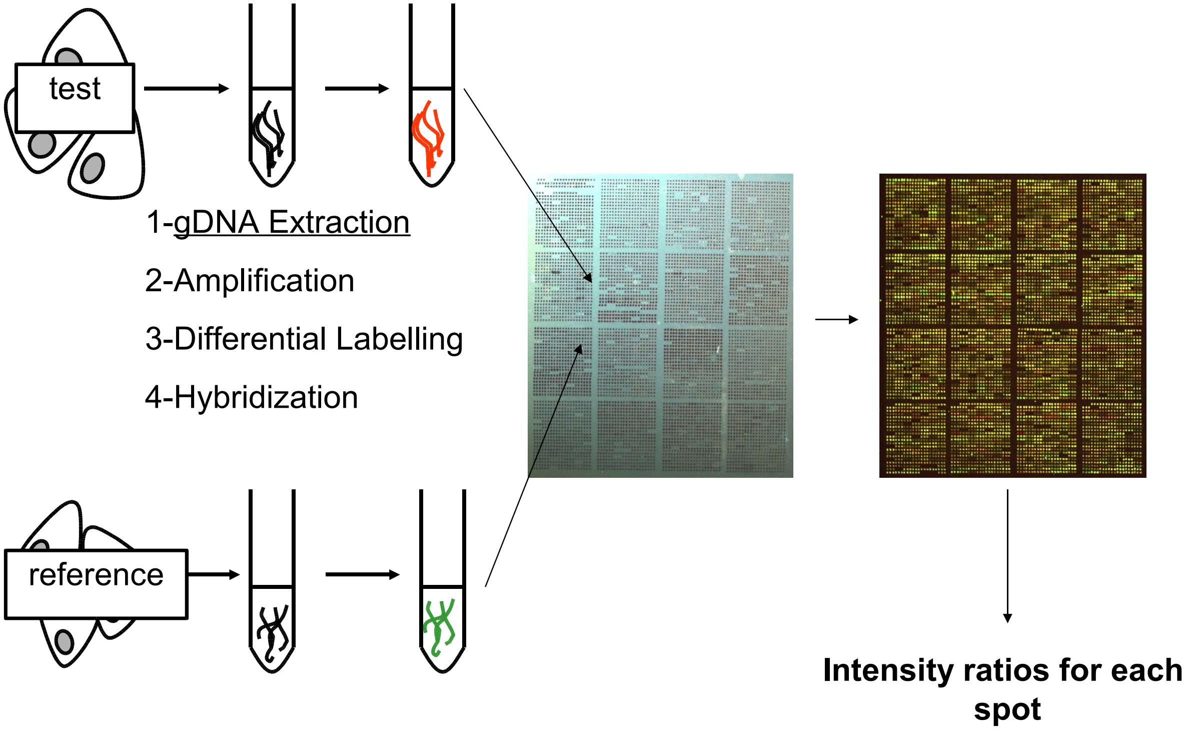

Fluorescence In Situ Hybridization
Labeling with fluorescent dye
probe DNA
Denature & Hybridize
chromosomal DNA
Probe signal on chromosome

Smeets *et al.* (2004) [15]

CGH microarrays



experimental protocole



1-gDNA Extraction

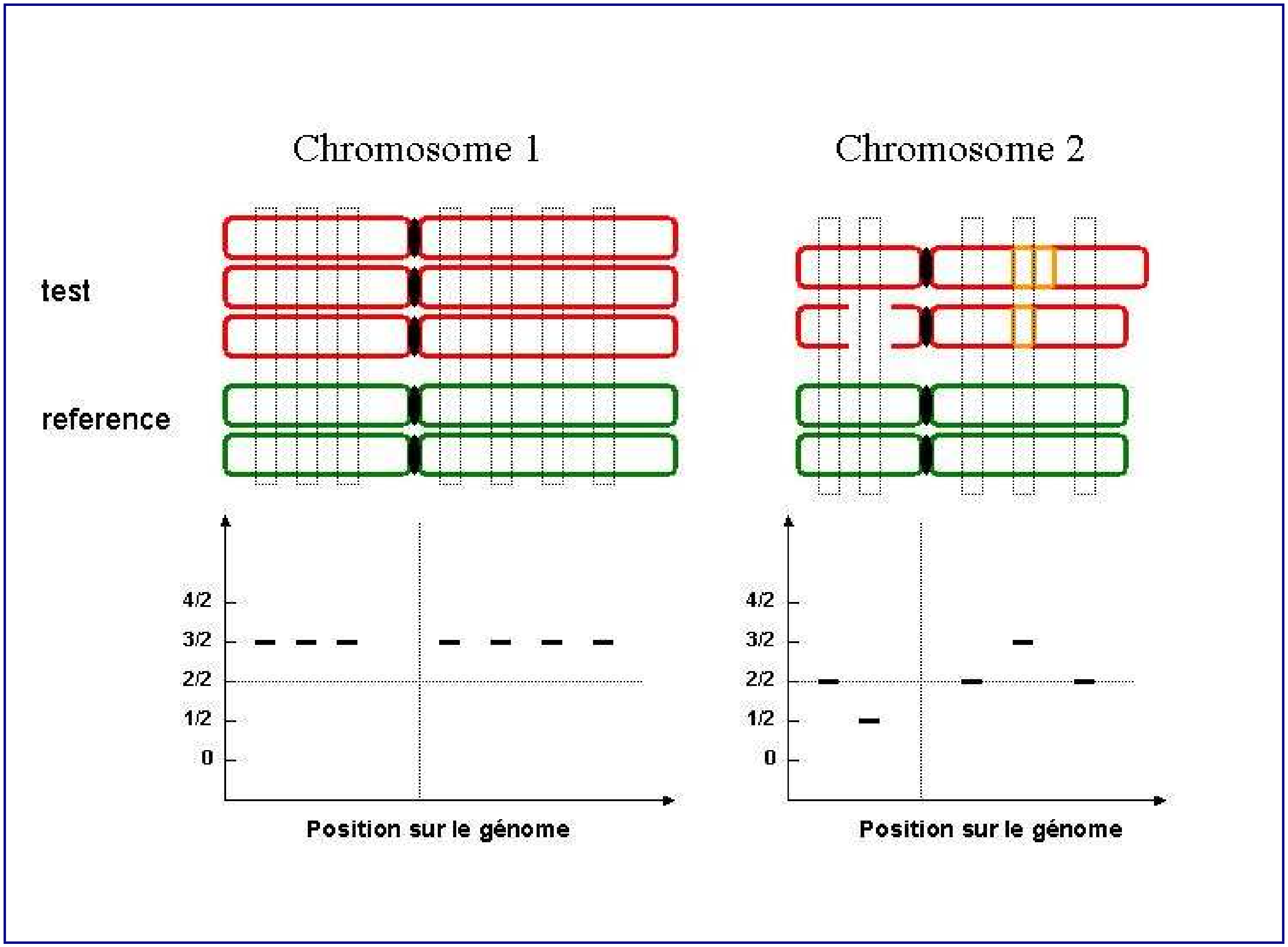
2-Amplification

3-Differential Labelling

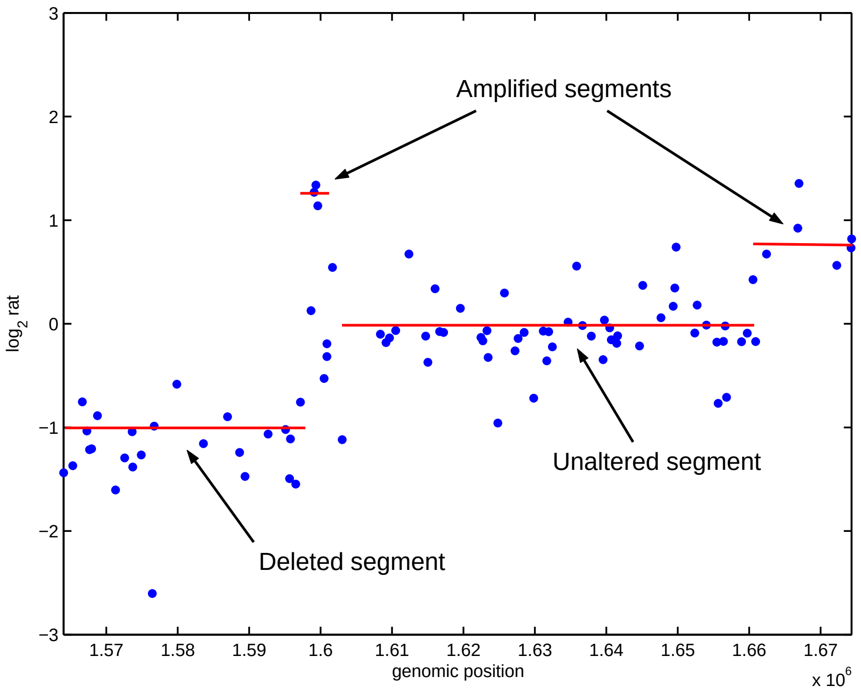
4-Hybridization

Intensity ratios for each spot

A simplified view of CGH microarray data



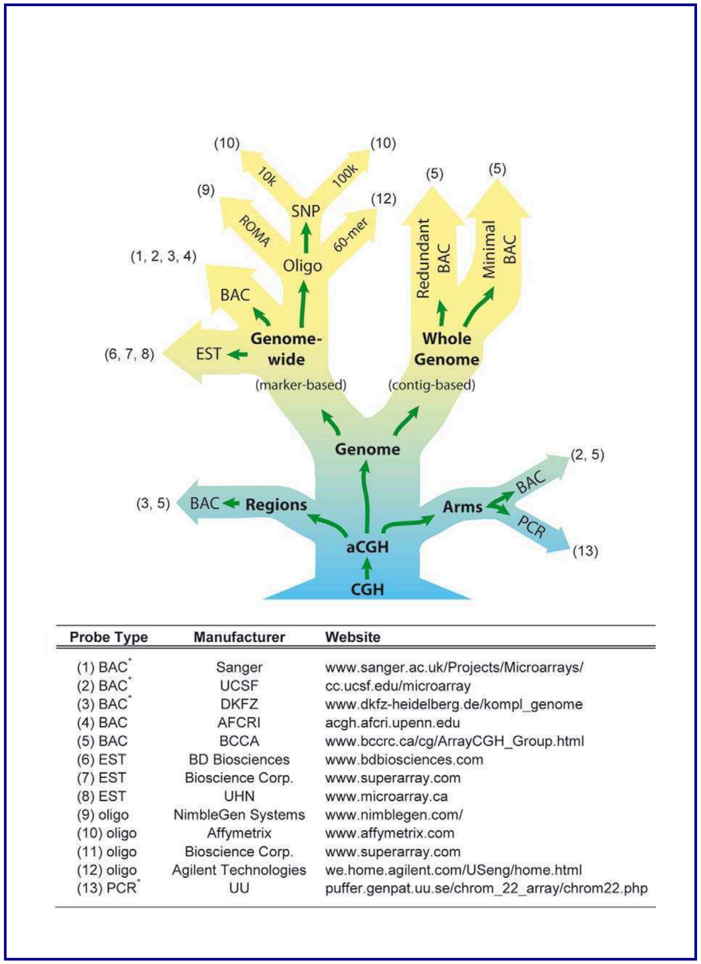
Interpreting a CGH profile



One dot on the graph represents

$$\log_2 \left\{ \frac{\sharp \text{ copies of BAC(t) in the test genome}}{\sharp \text{ copies of BAC(t) in the reference genome}} \right\}$$

Diversity of array CGH



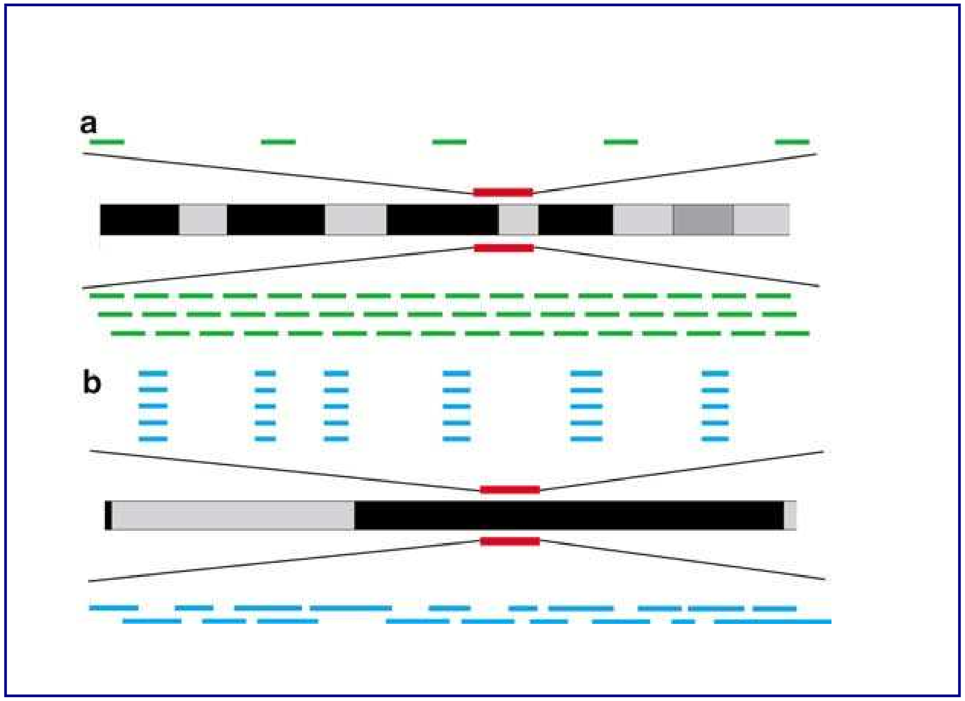
Probe Type	Manufacturer	Website
(1) BAC*	Sanger	www.sanger.ac.uk/Projects/Microarrays/
(2) BAC*	UCSF	cc.ucsf.edu/microarray
(3) BAC*	DKFZ	www.dkfz-heidelberg.de/kompl_genome
(4) BAC	AFCRI	acgh.afcri.upenn.edu
(5) BAC	BCCA	www.bccrc.ca/cg/ArrayCGH_Group.html
(6) EST	BD Biosciences	www.bdbiosciences.com
(7) EST	Bioscience Corp.	www.superarray.com
(8) EST	UHN	www.microarray.ca
(9) oligo	NimbleGen Systems	www.nimblegen.com/
(10) oligo	Affymetrix	www.affymetrix.com
(11) oligo	Bioscience Corp.	www.superarray.com
(12) oligo	Agilent Technologies	we.home.agilent.com/USeng/home.html
(13) PCR*	UU	puffer.genpat.uu.se/chrom_22_array/chrom22.php

Davies *et al.* (2005)[2]

Tiling Microarrays

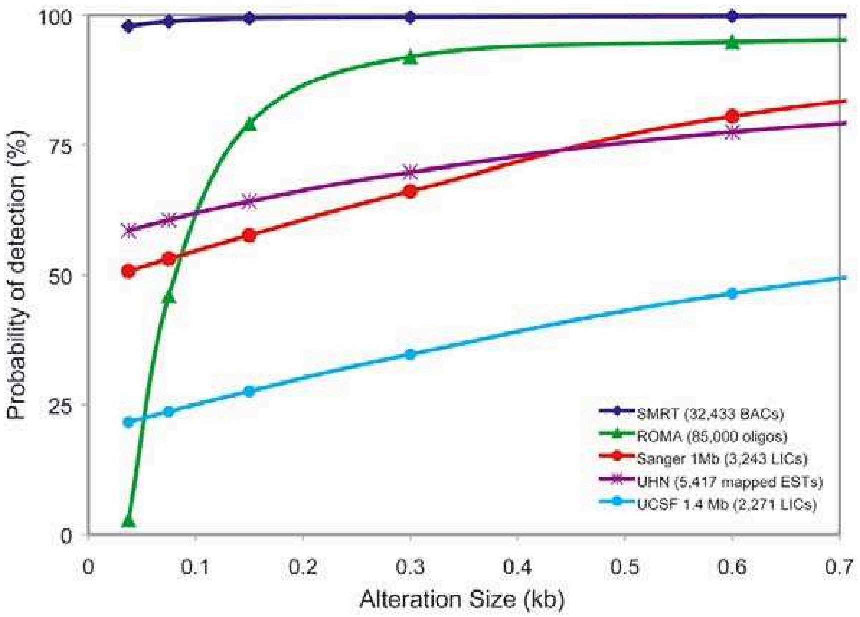
a) marker based arrays

b) tiling arrays : 32433 overlapping clones, whole genome coverage



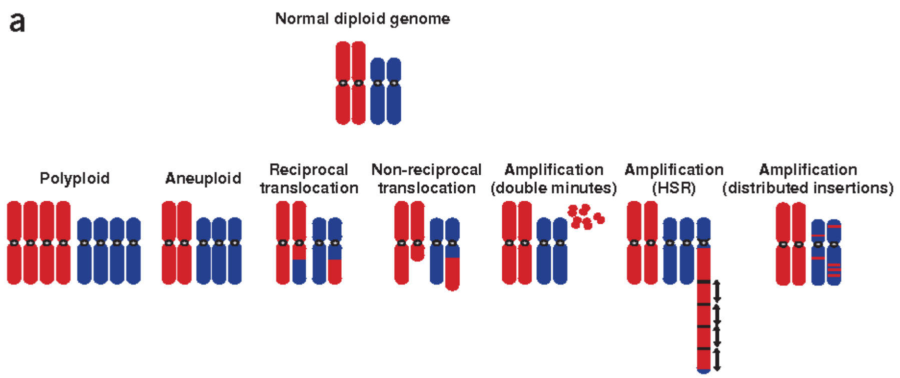
Lockwood *et al.* (2006) [9]

Performance



Davies *et al.* [2]

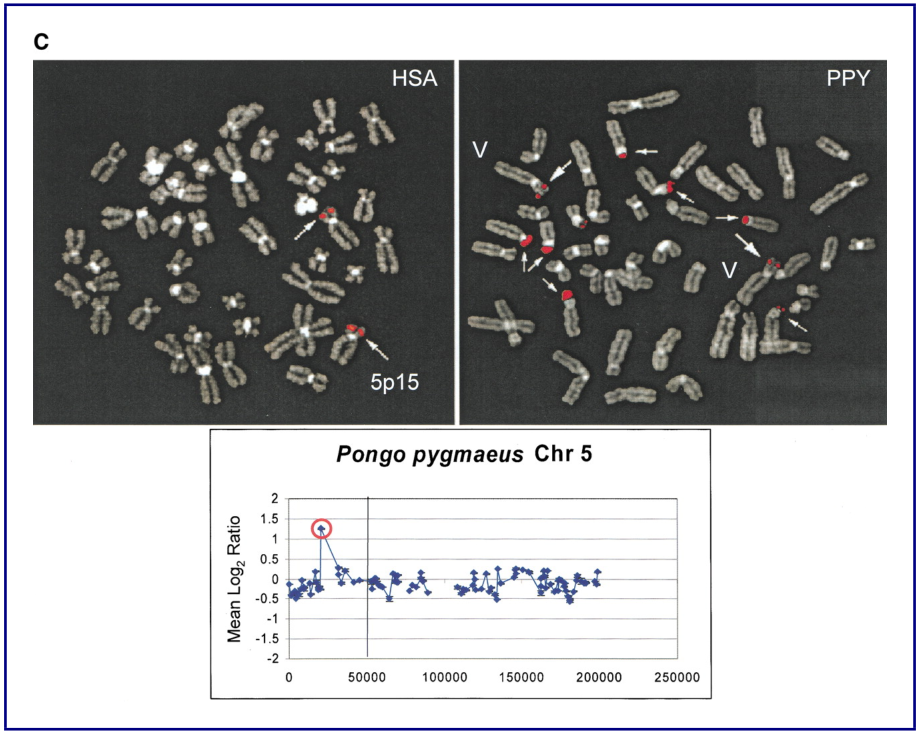
What can('t) be detected by array CGH ?


a
Normal diploid genome
Polyploid
Aneuploid
Reciprocal translocation
Non-reciprocal translocation
Amplification (double minutes)
Amplification (HSR)
Amplification (distributed insertions)

Technique	Detection						
Banding	+	+	+	+	+	+	−
SKY	+	+	+	+	+	+	−
CGH	−	+	−	+	+	+	+
LOH	−	+	−	+	+	+	+

Albertson *et al.* [1]

Application of aCGH to genome comparison



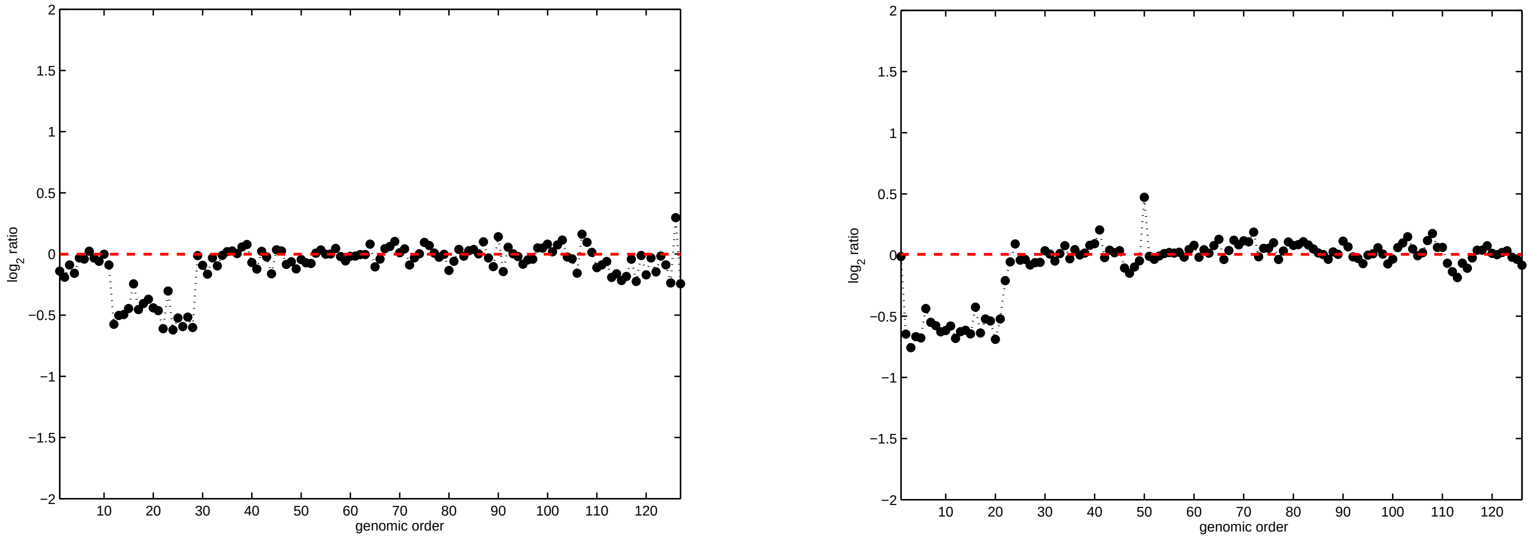
FISH validation of a duplication identified by array CGH (interphase and metaphase images). Interchromosomal duplication for an orangutan compared with 5p15 human locus. Locke *et al.* [8]

New applications in human genetics

Variation type	Definition	Frequency in the human genome
SNP	Single base pair variation found in $> 1\%$ of chromosomes in a given population	$\sim 10.10^6$SNPs in the human population
Ins./Del. Variants	Deletion or insertion of a segment of DNA. Incudes small polymorphic changes and large chromosomal aberrations. Often called CNV vhen > 1kb	$\sim 1.10^6$ Ins/Del polymorphism $>$ 1bp
Microsatellite	Sequences containing variable numbers of 1-6bp repeats totaling < 200bp in length	$> 1.10^6$ microsat. in the human genome, accounting for $\sim 3\%$ of the sequence

Feuk *et al.* [4]

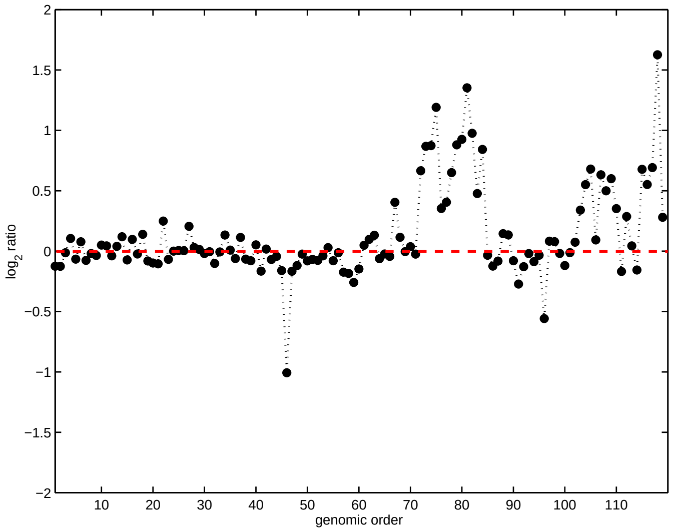
Purposes and Challenges of the statistical analysis - 1

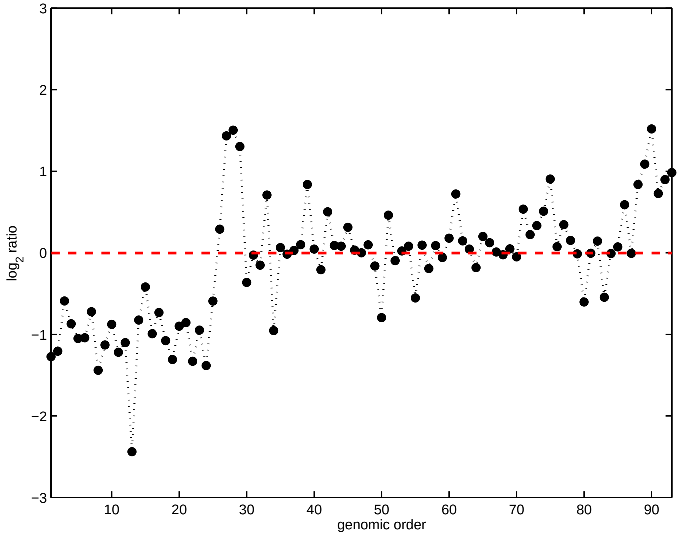


low level of noise → easy detection

Data described in Nakao et al. (2003) [11]

Purposes and Challenges of the statistical analysis - 2





high level of noise ⟶ need for automatic detection tools

Data described in Snijders et al. (2001) [16]

Hidden Markov models

Notations and construction of a statistical model

- Clones along the genome: $t = 1, \dots, n$ for the t^{th} clone at position x_t on the genome. $x_t \in \mathbb{N}$? or $x_t \in \mathbb{R}$? We consider discrete-time models. (Continuous time models are also possible [17]).
- The signal recorded for each clone $Y(x_t)$. It corresponds to the fluoresence $\log_2$ ratio. Basis 2 has been chosen for diploid organisms ! but the choice is questionable (discrete gene copy numbers do not lead to fixed mean for the corresponding signals)

First step when constructing a model : **what are the characteristics of the signal ?**

Why are copy numbers "underestimated" ?

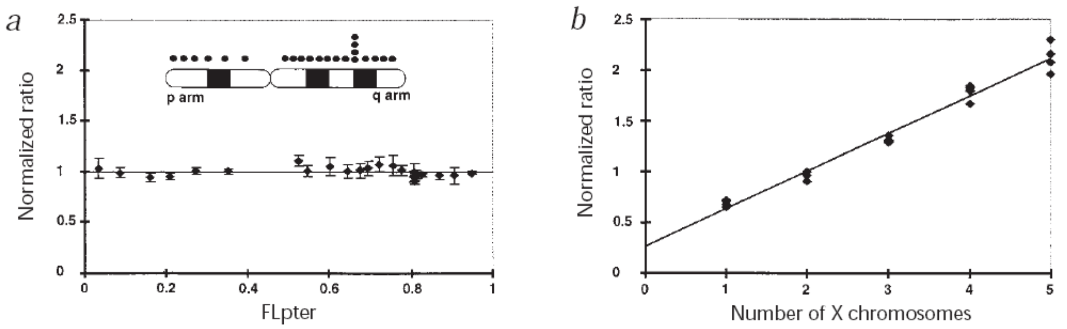


Fig. 2 Precision and accuracy of human copy number measurements. ***a***, Fluorescence ratios on the chromosome 20 array targets for six comparisons of a normal genome to itself. The inset schematically shows the location of the chromosome 20 array targets. The vertical set of targets on the q arm indicates the location of four nearly contiguous clones at 20q13.2. The ratios were normalized so that the average for all targets in each hybridization was 1.0. The data points show the mean of the six normalized ratios obtained for each target, the error bars indicate the standard deviations. They are plotted at the physical location of the clone as determined by FISH, measured as a fraction of the chromosome length relative to the p terminus (Flpter). The solid horizontal line is drawn at ratio 1.0. ***b***, Normalized fluorescence ratios on X chromosome targets as a function of X chromosome copy number. Arrays containing four clones from the X chromosome, and four clones on chromosome 20p (RMC20P107, RMC20P160, RMC20P178 and RMC20P099; Table 1), were hybridized with test genomic DNA from a normal male, normal female and three cell lines containing three, four and five copies of the X chromosome. Normal female reference DNA was used for all hybridizations. The fluorescence ratios on each of the X chromosome targets was normalized by the mean ratio of the chromosome 20 targets in that hybridization. All measurements are plotted, in some cases overlapping points obscure each other. The line is a linear regression through all of the data, slope 0.37, intercept 0.27.

Even if the relationship between the number of X chromosomes and the ratio of the intensity is **linear**, the slope differs from the theoretical expected value of 0.5. Pinkel *et al.* [13]

Why are copy numbers ”underestimated” ?

- If deletions concern only **part of the clones** the resulting signal will show less dramatic differences than expected in the case of a complete clone deletion
- The presence of **repetitive sequences** depends on the clones and the efficiency of the ”blocking” procedure with Cot-1 DNA may not be 100%
- Presence of **admixed normal DNA**. In the case of tumor extraction, tissues are composed of heterogeneous cell types resulting in a mix of different types of DNA. This leads to a dilution of the aberrations.

Even if CGH aims at studying a **discrete** process (gene copy numbers), providing a quantitative answer in terms of presence/absence is not straightforward.

Construction of a Hidden Markov Model

- First category of models which a have been widely used for the analysis of array CGH (Fridlyand *et al.*[5])
- The modeling framework is natural: aims at describing the distribution of observations when part of the information is missing
- In the case of array CGH data analysis: observations: fluorescence signal, missing information: copy number values
- widely used model in Bioinformatics, mainly in the discrete framework (sequence analysis)

Notations

- Clones along the genome: $t = 1, \ldots, n$ for the t^{th} clone at position x_t on the genome. $x_t \in \mathbb{N}$.
- The signal recorded for the clones $Y(x_t) \in \mathbb{R}$, $\mathbf{Y} = \{Y(x_1), \ldots, Y(x_n)\}$ the sequence of observations, and $\mathbf{y}$ its realization.
- $S(x_t)$: the state of clone t which corresponds to the copy number of clone t, $\mathbf{S} = \{S(x_1), \ldots, S(x_n)\}$, the sequence of hidden states, and $\mathbf{s}$ its realization.
- $S_t \in \{1, \ldots, P\}$, with P the total number of possible states. This number is unknown, and is considered fixed in the following.
- **First hypothesis**: we neglect the effect of the distance between clones:

$$Y(x_t) = Y_t, S(x_t) = S_t.$$

Conditional distributions

- Main idea: when the copy number is known, the conditional distribution of the observations is known: we model $\mathbb{P}\{\mathbf{Y} = \mathbf{y} | \mathbf{S} = \mathbf{s}\}$
- Central assumption: **conditional independence**
- For one observation: $Y_t | S_t = p \sim \mathcal{N}\left(m_p, \sigma^2\right)$
- if σ is constant, the model is $homoscedastic$, and $heteroscedastic$ if σ depends on the hidden state (σ_p)
- the joint conditional distribution $\mathbb{P}\{\mathbf{Y} | \mathbf{S}\}$ is

$$\begin{aligned}
\mathbb{P}\{\mathbf{Y} | \mathbf{S}; \mathbf{m}, \sigma\} &= \prod_{t=1}^{n} \mathbb{P}\{Y_t = y_t | S_t = s_t\} \\
&= \prod_{t=1}^{n} \prod_{p=1}^{P} \left\{ \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_t - m_p)^2}{2\sigma^2} \right) \right\}^{\mathbb{I}\{S_t = p\}}
\end{aligned}$$

Hidden part of the model

- $\mathbb{P}\{S_t = p\}$ is the marginal distribution of the hidden states, such that: $\mathbb{P}\{S_t = p\} = \pi_p$
- **Simple Hypothesis** the sequence $\mathbf{S}$ is made of iid random variables with multinomial distribution:

$$S_t = p \sim \mathcal{M}(1, \pi_1, \ldots, \pi_P)$$

- In this case, the model is often called Mixture Model (Mixture of distributions)
- **Markovian Hypothesis** the sequence $\mathbf{S}$ is a Markov chain of order 1, with transition probability ϕ_{ql} such that:

$$\phi_{q\ell} = \mathbb{P}\{S_{t+1} = \ell | S_t = q\}$$

Interest of the Markovian distribution

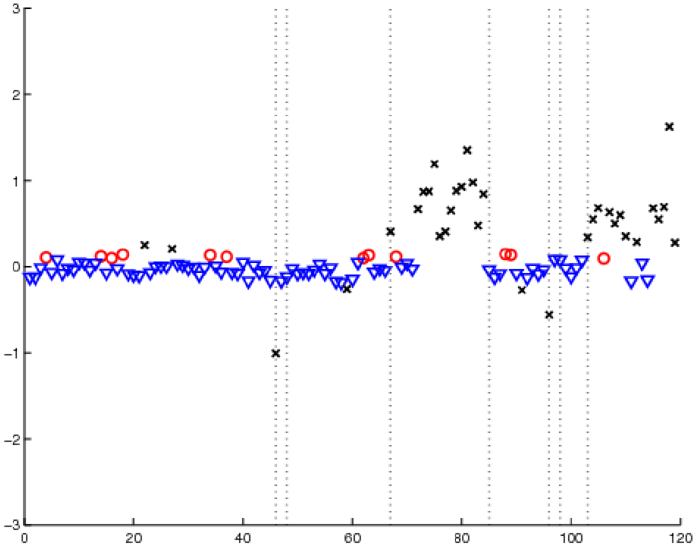
- It is used to model the spatial dependency which exists between copy number.

- In a first approximation, the chain is homogeneous (independent on t)

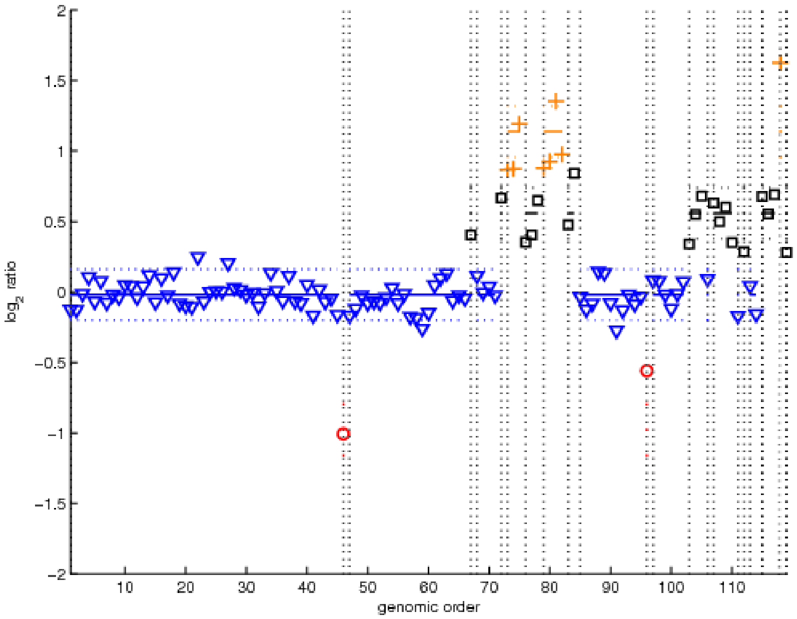
- the distribution of the hidden states is:

$$\mathbb{P}\{\mathbf{S}; \boldsymbol{\pi}, \boldsymbol{\phi}\} = \mathbb{P}\{S_1 = s_1\} \prod_{t=2}^{n-1} \mathbb{P}\{S_{t+1} = s_{t+1} | S_t = s_t\}$$

What is the interest of considering a Markov Chain for the hidden part ?



clustering results with mixture model



clustering results with HMM

Considering a Markovian sequence models the spatial structure of the signal.

Parameters and Likelihoods

- The parameters of the model are $\boldsymbol{\theta} = \{\mathbf{m}, \sigma, \boldsymbol{\pi}, \boldsymbol{\phi}\}$

- Hidden data likelihood: $\mathbb{P}\{\mathbf{S}; \boldsymbol{\pi}, \boldsymbol{\phi}\}$

- Complete data likelihood:

$$\mathbb{P}\{\mathbf{Y}, \mathbf{S}; \boldsymbol{\theta}\} = \mathbb{P}\{\mathbf{S}; \boldsymbol{\pi}, \boldsymbol{\phi}\} \times \mathbb{P}\{\mathbf{Y}|\mathbf{S}; \mathbf{m}, \sigma\}$$

- Observed incomplete data likelihood:

$$\mathbb{P}\{\mathbf{Y}; \boldsymbol{\theta}\} = \sum_{\mathbf{S}} \mathbb{P}\{\mathbf{Y}, \mathbf{S}; \boldsymbol{\theta}\}$$

- The idea is to calculate the observed data likelihood indirectly using:

$$\mathbb{P}\{\mathbf{Y}, \mathbf{S}; \boldsymbol{\theta}\} = \mathbb{P}\{\mathbf{Y}; \boldsymbol{\theta}\} \times \mathbb{P}\{\mathbf{S}|\mathbf{Y}; \boldsymbol{\theta}\}$$

How to calculate the observed-data likelihood ? - 1

- The EM algorithm was developed to optimize the observed data likelihood in the context of models with hidden structure. Dempster *et al.* [3]
- The hidden structure can take many forms, and clustering models (HMMs, mixture models) are widely used in applied Statistics
- Note that the complete-data likelihood is a random variable since the missing variables $\mathbf{S}$ are random and unknown: $\mathbb{P}\{\mathbf{Y}, \mathbf{S}; \boldsymbol{\theta}\} = h_{\mathbf{Y};\theta}(\mathbf{S})$
- One way to calculate this quantity is to evaluate its conditional expectation given $\mathbf{Y}$:

$$\mathbb{E}_\theta \left\{ \log \mathbb{P}\{\mathbf{Y}, \mathbf{S}; \boldsymbol{\theta}\} \middle| \mathbf{Y} \right\} = \mathbb{E}_\theta \left\{ \log h_{\mathbf{Y};\theta}(\mathbf{S}) \middle| \mathbf{Y} \right\}$$

- Recall that

$$\mathbb{E}_\theta \left\{ \log h(\mathbf{S}) \middle| \mathbf{Y} \right\} = \sum_{\mathbf{S}} \log h(\mathbf{S}) \times \mathbb{P}\{\mathbf{S} | \mathbf{Y}; \boldsymbol{\theta}\}$$

How to calculate the observed-data likelihood ? - 2

- Recalling that: $\mathbb{P}\{\mathbf{Y},\mathbf{S};\boldsymbol{\theta}\} = \mathbb{P}\{\mathbf{Y};\boldsymbol{\theta}\} \times \mathbb{P}\{\mathbf{S}|\mathbf{Y};\boldsymbol{\theta}\}$
- When considering the conditional expectation we get:

$$\begin{aligned} \mathbb{E}_{\theta^{(h)}}\left\{\log \mathbb{P}\{\mathbf{Y};\boldsymbol{\theta}\}|\ \mathbf{Y}\right\} &= \mathbb{E}_{\theta^{(h)}}\left\{\log \mathbb{P}\{\mathbf{Y},\mathbf{S};\boldsymbol{\theta}\}|\ \mathbf{Y}\right\} - \mathbb{E}_{\theta^{(h)}}\left\{\log \mathbb{P}\{\mathbf{S}|\mathbf{Y};\boldsymbol{\theta}\}|\ \mathbf{Y}\right\} \\ \log \mathcal{L}(\mathbf{Y};\boldsymbol{\theta}) &= \mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(h)}) - \mathcal{H}(\boldsymbol{\theta};\boldsymbol{\theta}^{(h)}) \end{aligned}$$

- $\boldsymbol{\theta}^{(h)}$ is considered **to calculate** the expectation, $\boldsymbol{\theta}$ parameter that will be **optimized**
- $\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(h)})$ term which is easier to calculate and to optimize,
- $\mathcal{H}(\boldsymbol{\theta};\boldsymbol{\theta}^{(h)})$ conditional entropy of the hidden structure which does not need to be calculated (trick! because of the following)

How to maximize the observed-data likelihood ?

- When estimating a parameter by maximum likelihood, we aim at solving:

$$\frac{\partial \log \mathcal{L}(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

- The basis of the EM algorithm lies in this relationship:

$$\frac{\partial \log \mathcal{L}(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbb{E}_{\theta}\left\{ \frac{\partial \log \mathbb{P}\{\mathbf{Y},\mathbf{S};\boldsymbol{\theta}\}}{\partial \boldsymbol{\theta}} \middle| \mathbf{Y} \right\}$$

- Consequently, maximizing the incomplete data log-likelihood can be done thanks to the maximization of $\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(h)})$.

Proof - Fisher identity (1925)

$$
\begin{aligned}
\frac{\partial \log \mathbb{P}(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{1}{\mathbb{P}(\mathbf{Y};\boldsymbol{\theta})} \times \frac{\partial \mathbb{P}(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
&= \frac{1}{\mathbb{P}(\mathbf{Y};\boldsymbol{\theta})} \int_{\mathbf{S}} \frac{\partial \mathbb{P}(\mathbf{Y},\mathbf{S};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} d\mathbf{S} \\
&= \frac{1}{\mathbb{P}(\mathbf{Y};\boldsymbol{\theta})} \int_{\mathbf{S}} \frac{\partial \log \mathbb{P}(\mathbf{Y},\mathbf{S};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbb{P}(\mathbf{Y},\mathbf{S};\boldsymbol{\theta}) d\mathbf{S} \\
&= \frac{1}{\mathbb{P}(\mathbf{Y};\boldsymbol{\theta})} \int_{\mathbf{S}} \frac{\partial \log \mathbb{P}(\mathbf{Y},\mathbf{S};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbb{P}(\mathbf{Y};\boldsymbol{\theta})\mathbb{P}(\mathbf{S}|\mathbf{Y};\boldsymbol{\theta}) d\mathbf{S} \\
&= \int_{\mathbf{S}} \frac{\partial \log \mathbb{P}(\mathbf{Y},\mathbf{S};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbb{P}(\mathbf{S}|\mathbf{Y};\boldsymbol{\theta}) d\mathbf{S} \\
&= \mathbb{E}_{\theta} \left\{ \frac{\partial \log \mathbb{P}\{\mathbf{Y},\mathbf{S};\boldsymbol{\theta}\}}{\partial \boldsymbol{\theta}} \middle| \mathbf{Y} \right\}
\end{aligned}
$$

Two steps for the EM algorithm

- **E-Step**: Expectation step. This step consists in the calculation of $\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(h)})$ which is a conditional expectation
- **M-Step**: Maximization step. This step consists in the maximization of $\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(h)})$ to get an up-dated value of the estimator $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(h+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(h)})$$

- Those steps are repeated alternatively until $|\boldsymbol{\theta}^{(h+1)} - \boldsymbol{\theta}^{(h)}| < \varepsilon$
- Warning : the stopping criterion should not be $|\log\mathcal{L}(\mathbf{Y};\boldsymbol{\theta}^{(h+1)}) - \log\mathcal{L}(\mathbf{Y};\boldsymbol{\theta}^{(h)})| < \varepsilon$

Properties of the EM algorithm

- With the maximization step,

$$\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(h+1)}) \geq \mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(h)})$$

- It can be shown that

$$\mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(h+1)}) \leq \mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(h)})$$

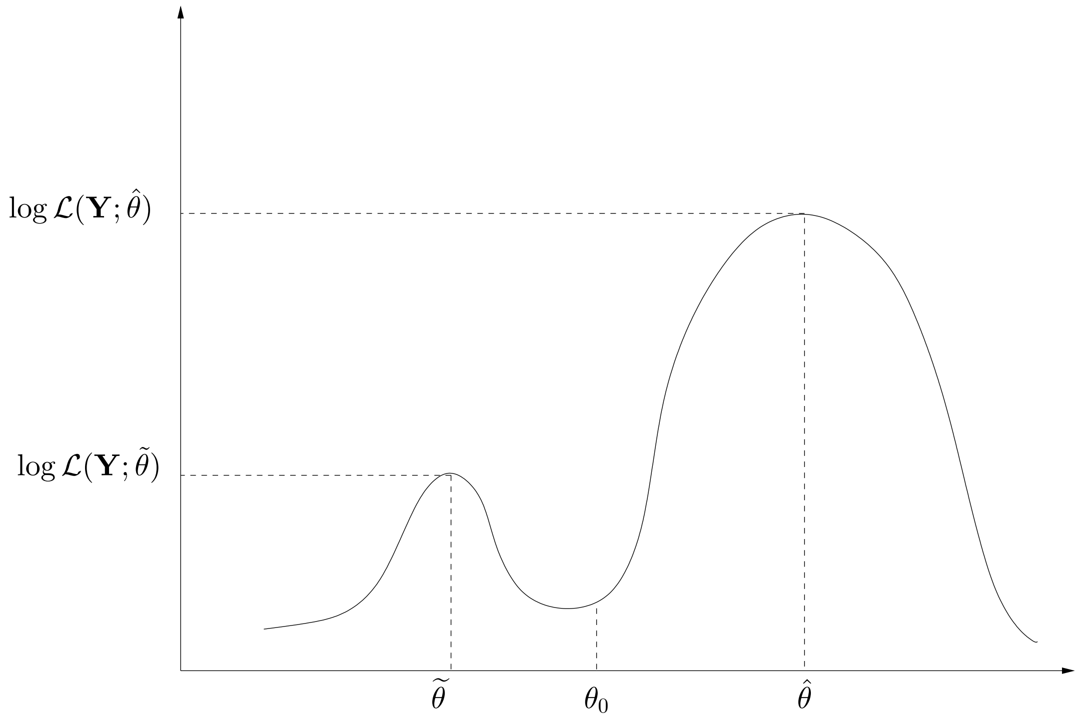
- This implies a **monotonicity** property such that:

$$\mathcal{L}(\mathbf{Y}; \boldsymbol{\theta}^{(h+1)}) \geq \mathcal{L}(\mathbf{Y}; \boldsymbol{\theta}^{(h)}),$$

- This property means that the EM algorithm improves the incomplete data likelihood at each step.
- Does the likelihood reach its maximum ?

Convergence properties of the EM algorithm

What is the convergence of an iterative algorithm ? Very difficult to show from a theoretical point of view! The convergence generally concerns the reach of local optima and strongly depends on the starting point !!!


$\log \mathcal{L}(\mathbf{Y};\hat{\theta})$
$\log \mathcal{L}(\mathbf{Y};\tilde{\theta})$
$\widetilde{\theta}$
θ_0
$\hat{\theta}$

Calculating $\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(h+1)})$ for Mixture Models

$$
\begin{aligned}
\mathbb{P}\{\mathbf{Y},\mathbf{S};\boldsymbol{\theta}\} &= \prod_{t=1}^{n}\prod_{p=1}^{P}\left[\pi_p f(y_t;m_p)\right]^{\mathbb{I}\{S_t=p\}} \\
\log \mathbb{P}\{\mathbf{Y},\mathbf{S};\boldsymbol{\theta}\} &= \sum_{t=1}^{n}\sum_{p=1}^{P}\mathbb{I}\{S_t=p\}\log\pi_p + \sum_{t=1}^{n}\sum_{p=1}^{P}\mathbb{I}\{S_t=p\}\times\log f(y_t;m_p) \\
\mathbb{E}\left\{\log \mathbb{P}\{\mathbf{Y},\mathbf{S};\boldsymbol{\theta}\}|\mathbf{Y}\right\} &= \sum_{t=1}^{n}\sum_{p=1}^{P}\mathbb{E}\left\{\mathbb{I}\{S_t=p\}|\mathbf{Y}\right\}\log\pi_p \\
&+ \sum_{t=1}^{n}\sum_{p=1}^{P}\mathbb{E}\left\{\mathbb{I}\{S_t=p\}|\mathbf{Y}\right\}\times\log f(y_t;m_p) \\
&= \sum_{t=1}^{n}\sum_{p=1}^{P}\mathbb{E}\left\{\mathbb{I}\{S_t=p\}|Y_t\right\}\log\pi_p \\
&+ \sum_{t=1}^{n}\sum_{p=1}^{P}\mathbb{E}\left\{\mathbb{I}\{S_t=p\}|Y_t\right\}\times\log f(y_t;m_p)
\end{aligned}
$$

EM algorithm for mixture models

- Let us denote by τ_{tp} the posterior probability of membership to state p:

$$\tau_{tp} = \mathbb{E}\left\{\mathbb{I}\{S_t = p\} | Y_t\right\} = \Pr\{S_t = p | Y_t = y_t\}$$

- Using this notation we get :

$$\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(h+1)}) = \sum_{t=1}^{n} \sum_{p=1}^{P} \tau_{tp}^{(h+1)} \log \pi_p^{(h)} + \sum_{t=1}^{n} \sum_{p=1}^{P} \tau_{tp}^{(h+1)} \times \log f(y_t; m_p^{(h)})$$

- The E-step consists in the update of $\{\tau_{tp}\}$ using the Bayes formula

$$\tau_{tp}^{(h+1)} = \frac{\pi_p^{(h)} f(y_t; m_p^{(h)})}{\sum_\ell \pi_\ell^{(h)} f(y_t; m_\ell^{(h)})}$$

- The M step consists in the maximization of $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(h+1)})$

Calculating $\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(h+1)})$ for HMMs

$$\begin{aligned}
\mathbb{P}\{\mathbf{Y},\mathbf{S};\boldsymbol{\theta}\} &= \prod_{\ell=1}^{P}\left[\pi_\ell\right]^{\mathbb{I}\{S_1=\ell\}} \times \prod_{t=1}^{n-1}\prod_{q=1}^{P}\prod_{\ell=1}^{P}\left[\phi_{q\ell}\right]^{\mathbb{I}\{S_{t+1}=\ell,S_t=q\}} \\
&\times \prod_{t=1}^{n}\prod_{\ell=1}^{P}\left[f(y_t;m_\ell)\right]^{\mathbb{I}\{S_t=\ell\}} \\
\log\mathbb{P}\{\mathbf{Y},\mathbf{S};\boldsymbol{\theta}\} &= \sum_{\ell}\mathbb{I}\{S_1=\ell\}\log\pi_\ell + \sum_{t,q,\ell}\mathbb{I}\{S_{t+1}=\ell,S_t=q\}\log\phi_{q\ell} \\
&+ \sum_{t,\ell}\mathbb{I}\{S_t=\ell\}\log f(y_t;m_\ell) \\
\mathbb{E}\left\{\log\mathbb{P}\{\mathbf{Y},\mathbf{S};\boldsymbol{\theta}\}|\mathbf{Y}\right\} &= \sum_{\ell}\mathbb{E}\left\{\mathbb{I}\{S_1=\ell\}|\mathbf{Y}\right\}\log\pi_\ell \\
&+ \sum_{t,q,\ell}\mathbb{E}\left\{\mathbb{I}\{S_{t+1}=\ell,S_t=q\}|\mathbf{Y}\right\}\log\phi_{q\ell} \\
&+ \sum_{t,\ell}\mathbb{E}\left\{\mathbb{I}\{S_t=\ell\}|\mathbf{Y}\right\}\log f(y_t;m_\ell)
\end{aligned}$$

Towards the forward backward equations

Recalling that if $X \in \{0, 1\}$ then:

$$\mathbb{E}(X) = \sum_{x=0,1} x \times \mathbb{P}(X = x) = \mathbb{P}(X = 1)$$

The calculation of $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(h+1)})$ requires the calculation of the following conditional probabilities:

$$\begin{aligned}
\mathbb{E}\{\mathbb{I}\{S_t = q\}|\mathbf{Y}; \boldsymbol{\theta}\} &= \mathbb{P}\{S_t = q|\mathbf{Y}; \boldsymbol{\theta}\}, \\
\mathbb{E}\{\mathbb{I}\{S_{t+1} = \ell, S_t = q\}|\mathbf{Y}; \boldsymbol{\theta}\} &= \mathbb{P}\{S_{t+1} = \ell, S_t = q|\mathbf{Y}; \boldsymbol{\theta}\}
\end{aligned}$$

which is done using the Forward-Backward algorithm.

In the following we will use a new notation : $\mathbf{Y}_1^t = \{Y_1, \ldots, Y_t\}$

Forward equations

The recursion starts as follows:

$$\begin{aligned}
\mathbb{P}\{S_1 = q; \boldsymbol{\theta}\} &= \pi_q \\
\forall t > 1, \mathbb{P}\{S_t = \ell | \mathbf{Y}_1^{t-1}; \boldsymbol{\theta}\} &= \sum_{q=1}^{Q} \phi_{q\ell} \times \mathbb{P}\{S_{t-1} = q | \mathbf{Y}_1^{t-1}; \boldsymbol{\theta}\}
\end{aligned}$$

Then calculate:

$$\mathbb{P}\{S_t = \ell | \mathbf{Y}_1^{t-1}; \boldsymbol{\theta}\} = \frac{f(y_t; m_\ell) \times \mathbb{P}\{S_t = \ell | \mathbf{Y}_1^{t-1}; \boldsymbol{\theta}\}}{\sum_{q=1}^{Q} f(y_t; m_q) \times \mathbb{P}\{S_t = q | \mathbf{Y}_1^{t-1}; \boldsymbol{\theta}\}}$$

Backward equations

It consists in calculating:

$$\mathbb{P}\{S_t = \ell, S_{t+1} = q|\mathbf{Y};\boldsymbol{\theta}\} = \frac{\mathbb{P}\{S_t = \ell|\mathbf{Y}_1^t;\boldsymbol{\theta}\} \times \phi_{q\ell} \times \mathbb{P}\{S_{t+1} = q|\mathbf{Y};\boldsymbol{\theta}\}}{\mathbb{P}\{S_{t+1} = q|\mathbf{Y}_1^t;\boldsymbol{\theta}\}}$$

From which is deduced the desired quantity:

$$\begin{aligned}\mathbb{P}\{S_t = \ell|\mathbf{Y};\boldsymbol{\theta}\} &= \sum_{q=1}^{Q} \mathbb{P}\{S_t = \ell, S_{t+1} = q|\mathbf{Y};\boldsymbol{\theta}\} \\ &= \sum_{q=1}^{Q} \xi_t(q,\ell) \\ &= \xi_t(+,\ell)\end{aligned}$$

This quantity is the *posterior* probability of being in state ℓ at position t.

Maximization step

The quantity to maximize is :

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(h+1)}) &= \sum_{\ell} \xi_1^{[h]}(+,\ell) \log \pi_\ell \\
&+ \sum_{t,q,\ell} \xi_t^{[h]}(q,\ell) \log \phi_{q\ell} \\
&+ \sum_{t,\ell} \xi_t^{[h]}(+,\ell) \log f(y_t; m_\ell, \sigma^2)
\end{aligned}$$

under the following constraints:

$$\begin{aligned}
\sum_{p=1}^{Q} \pi_p &= 1 \\
\forall q \in [1,P], \sum_{\ell=1}^{Q} \phi_{q\ell} &= 1
\end{aligned}$$

Estimators

The maximization of $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(h+1)})$ under constraint leads to the following estimators:

$$
\begin{aligned}
\pi_\ell^{[h+1]} &= \xi_1^{[h]}(+, \ell) \\
\phi_{q\ell}^{[h+1]} &= \frac{\sum_{t=1}^{n-1} \xi_t^{[h]}(q, \ell)}{\sum_{t=1}^{n-1} \xi_t^{[h]}(q, +)} \\
m_\ell^{[h+1]} &= \frac{\sum_{t=1}^{n-1} \xi_t^{[h]}(+, \ell) y_t}{\sum_{t=1}^{n-1} \xi_t^{[h]}(+, \ell)} \\
\sigma^{2[h+1]} &= \frac{\sum_{t=1}^{n-1} \sum_{\ell=1}^{P} \xi_t^{[h]}(+, \ell)(y_t - m_\ell^{[h+1]})^2}{\sum_{t=1}^{n-1} \xi_t^{[h]}(+, \ell)}
\end{aligned}
$$

with $\sum_{t=1}^{n-1} \xi_t^{[h]}(q, \ell)$ being the estimated number of transitions from state q to state ℓ.

Different strategies to recover the hidden states

- This is the final goal of the analysis: recover the *optimal* sequence of hidden states $\{\hat{s}_1, \ldots, \hat{s}_n\}$.
- This can be done once the parameters of the model have been estimated.
- Difficulty to define what is the optimal sequence !
- One possibility is to choose the states which are individually most likely:

$$\hat{s}_t = \arg\max_q \mathbb{P}\{S_t = q | \mathbf{Y}; \boldsymbol{\theta}\}$$

- Another strategy is to consider the complete sequence of hidden states: this is the *Viterbi* algorithm.

The Viterbi algorithm - 1

The objective is to calculate: $\{\hat{s}_1, \ldots, \hat{s}_n\} = \arg\max_{s_1^n} \mathbb{P}\{S_1^n = s_1^n | Y_1^n = y_1^n; \boldsymbol{\theta}\}$.

It starts with a forward recurrence initialized with $\mathbb{P}\{S_1 = q, y_1\} = \pi_q f(y_1; m_q, \sigma^2)$,

and then $\forall t > 1$

$$\begin{aligned}
\max_{s_1^t} \mathbb{P}\{S_1^{t-1} = s_1^{t-1}, S_t = q, y_1^t\} &= \max_{s_1^{t-1}} \Big(f(y_t; m_q, \sigma^2)\phi(s_{t-1}, q) \\
&\times \max_{s_1^{t-2}} \mathbb{P}\{S_1^{t-2} = s_1^{t-2}, S_{t-1} = s_{t-1}, y_1^{t-1}\} \Big) \\
\psi_t(q) &= \arg\max_{s_1^t} \mathbb{P}\{S_1^{t-1} = s_1^{t-1}, S_t = q, y_1^t\}
\end{aligned}$$

The Viterbi algorithm - 2

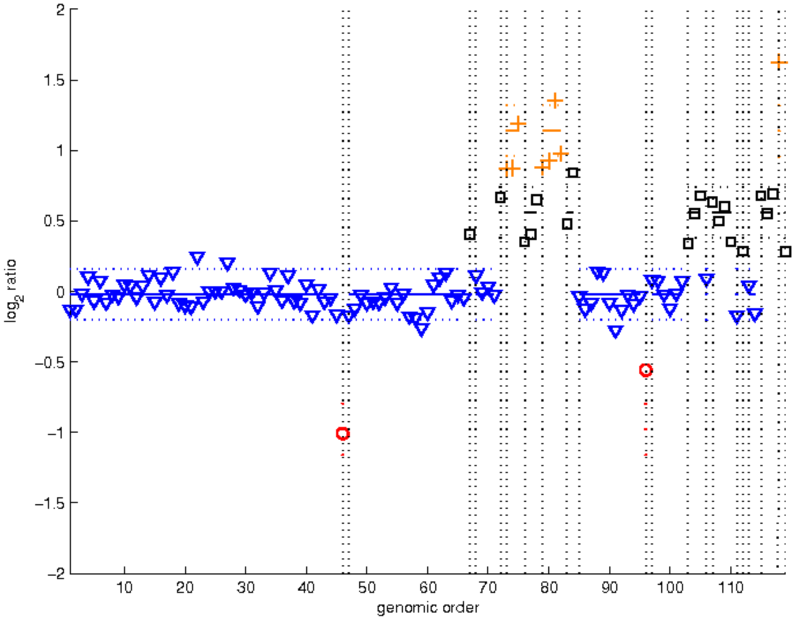
The backward recurrence finds $\hat{s}_1^n = (\hat{s}_1, \ldots, \hat{s}_n)$ such that:

$$\hat{s}_1^n = \arg\max_q \left(\max_{s_1^{n-1}} \mathbb{P}\{S_1^{n-1} = s_1^{n-1}, S_n = q, y_1^n\} \right)$$

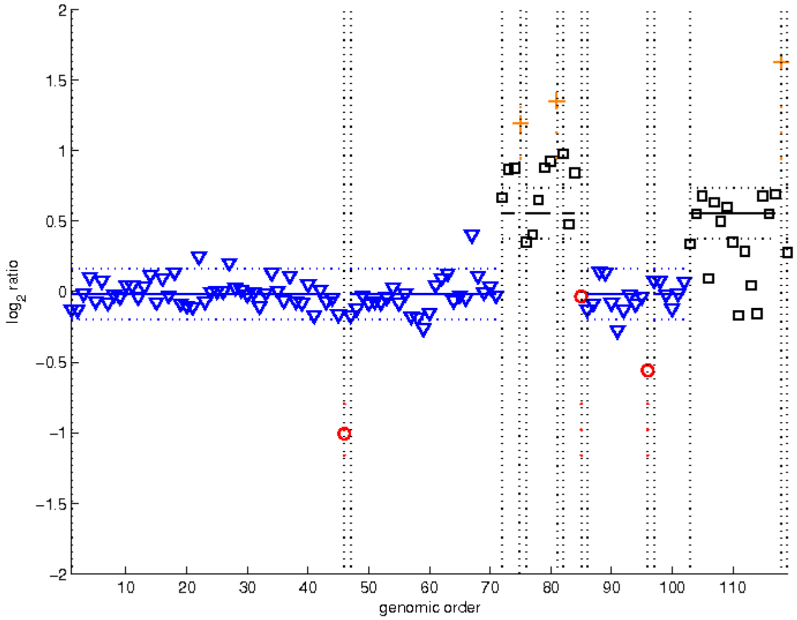
Termination: $\forall t = n-1, \ldots, 1$

$$\hat{s}_t = \psi_t(\hat{s}_{t+1})$$

Examples and comparison

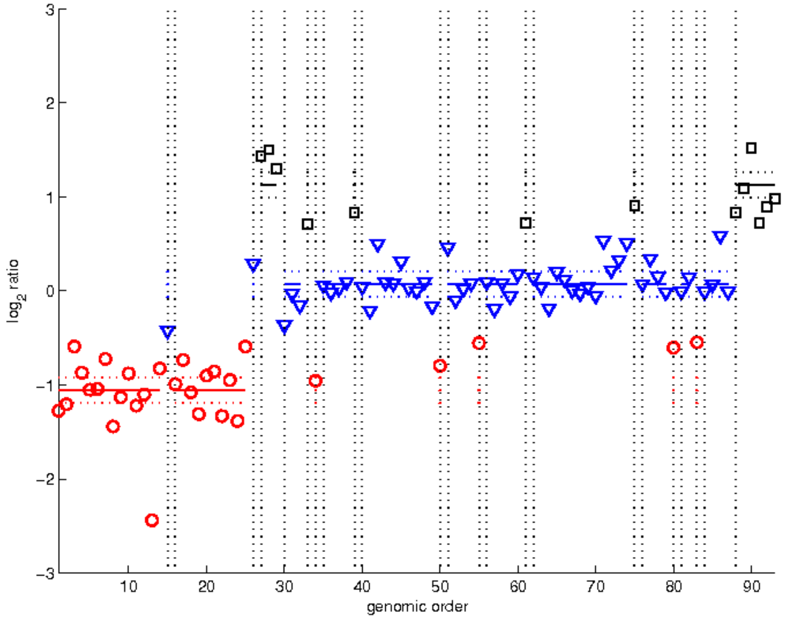


Forward-Backward

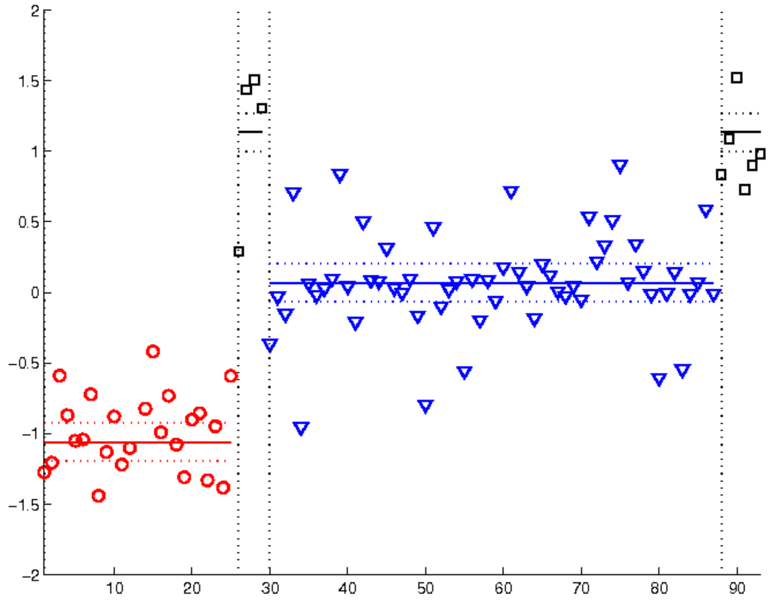


Viterbi

Examples and comparison



Forward-Backward



Viterbi

Model selection - in brief

- In practice, the total number of hidden states is unknown, and should be estimated
- This is done using a penalized criterion:

$$\hat{P} = \arg\max_{P} \log \mathcal{L}(\mathbf{Y}; \hat{\boldsymbol{\theta}}) - \beta \mathsf{pen}(P)$$

- The motivation of such criterion is to establish a trade-off between a good quality of fit and a reasonnable number of parameters to estimate
- Parcimony is the rule !
- Different penalty fonction according to different objectives which makes the choice controversial.
- This is an important field of research in Statistics

Other developments for HMMs and aCGH

- HMMs are very appropriate for aCGH modelling !
- Other refinements can be found in the literature to account for aCGH specificity
- Account for distances between clones using heterogeneous HMMs [10]:

$$\phi_{ql}(t) = f(x_t, x_{t+1})$$

- Account for possible overlap between clones with continuous-time HMMs [17]
- Other estimation strategies : the Bayesian framework [14].

Segmentation models

Presentation and motivations for segmentation models

- A first motivation is that the signal shows discontinuities when copy numbers change.
- Then the mean of the signal is constant between two changes.
- This is what is called a **segmentation model**
- We aim at recovering the change points (also called break-points): their number and their position.
- Segmentation models belong to a wide variety of models with many applications in the field of signal processing. Also widely used for segmenting biological sequences (discrete framework).

Definitions and notations for segmentation models

- Suppose we observe the process $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ such that the Y_ts are i.i.d. with distribution $\mathcal{N}(\mu_t, \sigma^2)$
- Then we suppose that there exists a sequence of change-points $t_1, \ldots, t_K$ such that the mean of the signal is constant between two changes and different from a change to another
- we denote by $I_k =]t_{k-1}, t_k]$ this interval of stationarity and μ_k the mean of the signal between two changes. Then the model is

$$\forall t \in I_k, \;\; Y_t = \mu_k + \varepsilon_t, \;\; \varepsilon_t \sim \mathcal{N}(0, \sigma^2)$$

Parameters and estimation

- The parameters of the model are $\mathbf{T} = \{t_1, \dots, t_K\}$, $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_K\}$ and σ^2
- The estimation is done for a given number of segments K, and K is estimated afterwards.
- The likelihood of the model is:

$$\begin{aligned}\mathcal{L}_K(\mathbf{Y}; \mathbf{T}, \boldsymbol{\mu}, \sigma^2) &= \prod_{k=1}^{K} \prod_{t=t_{k-1}+1}^{t_k} f(y_t; \mu_k, \sigma^2) \\ &= \prod_{k=1}^{K} \prod_{t=t_{k-1}+1}^{t_k} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y_t - \mu_k)^2\right\}\end{aligned}$$

A 3-step methodology for parameter estimation

- When K and $\mathbf{T}$ are known, how to estimate $\boldsymbol{\mu}$? Is $\boldsymbol{\mu}$ the only parameter to be concerned by abrupt changes ?
- When K is known, how to estimate $\mathbf{T}$? Optimization of the likelihood. This is a partitioning problem which is solved by Dynamic Programming which can be applied thanks to the additivity property of the log-likelihood.
- How to choose K ? Model selection

Step 1, estimation of the mean

- When K and $\mathbf{T}$ are known the estimation of $\boldsymbol{\mu}$ is straightforward:

$$\begin{aligned}\hat{\mu}_k &= \frac{1}{\hat{t}_k - \hat{t}_{k-1}} \sum_{t=\hat{t}_{k-1}+1}^{\hat{t}_k} y_t \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{k=1}^{K} \sum_{t=\hat{t}_{k-1}+1}^{\hat{t}_k} (y_t - \hat{\mu}_k)^2\end{aligned}$$

- Is the model homoskedastic ?

- Otherwise the estimation of heterogeneous variances is also straightforward:

$$\hat{\sigma}_k^2 = \frac{1}{\hat{t}_k - \hat{t}_{k-1}} \sum_{t=\hat{t}_{k-1}+1}^{\hat{t}_k} (y_t - \hat{\mu}_k)^2$$

Step 2, estimation of the change points

Objectif : find $T\hat{b}f$ such that:

$$\hat{\mathbf{T}} = \arg\max_{\mathbf{T}} \left\{ \log \mathcal{L}_K(\mathbf{Y}; \mathbf{T}, \boldsymbol{\mu}, \sigma^2) \right\}.$$

- Many strategies have been considered, in particular splitting strategies.
- Dynamic Programming is an efficient algorithm to solve this partitioning problem
- The problem is to partition n data points into K segments: the theoretical complexity of an exhaustive search would be $\mathcal{O}(n^K)$.
- Dynamic programming reduces the complexity to $\mathcal{O}(n^2)$ when K is fixed.
- It is based on the Bellman principal : "subpaths of optimal paths are themselves optimal". Analogy with the shortest path problem

Dynamic programming for segmentation

- Denoting by $J_k(i,j)$ the cost of the path connecting i to j in k steps (k segments)

$$\forall 0 \leq i < j \leq n, \ \ J_1(i,j) = \sum_{t=i+1}^{j} (y_t - \bar{y}_{ij})^2$$

- Then to calculate the next steps:

$$\begin{aligned} J_2(1,n) &= \min_h \{J_1(1,h) + J_1(h+1,n)\} \\ J_3(1,n) &= \min_h \{J_2(1,h) + J_1(h+1,n)\} \end{aligned}$$

- The general recursion is given by

$$\forall 1 \leq k \leq K-1, \ \ J_{k+1}(1,j) = \min_{1 \leq h \leq j} \{J_k(1,h) + J_1(h+1,j)\}$$

Model selection for segmentation models

- the number of segments K should be estimated.
- This is done using a penalized criterion:

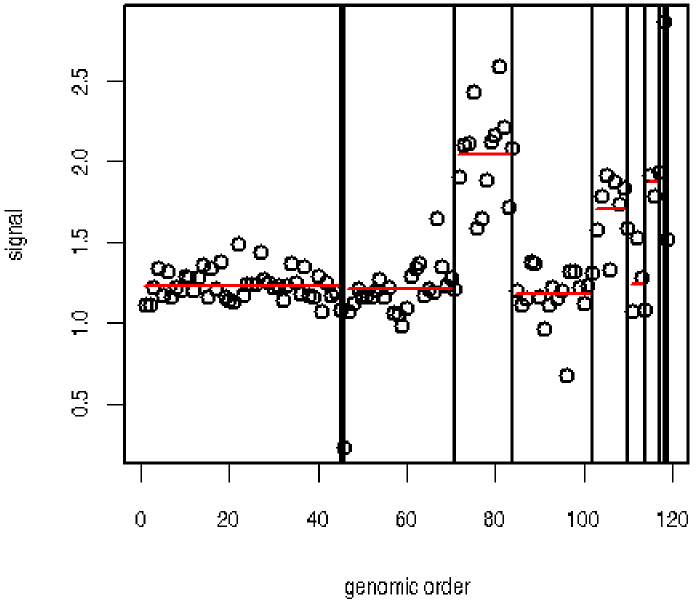
$$\hat{K} = \arg\max_K \log \mathcal{L}_K(\mathbf{Y}; \hat{\mathbf{T}}, \hat{\boldsymbol{\mu}}, \hat{\sigma}^2) - \beta \text{pen}(K)$$

- Once again, the penalty pen(K) is difficult to derive, since there exists C_{n-1}^{K-1} possible partitions for a model with K segments
- Theoretical results exists [7] and gives a general form to the penalty function:

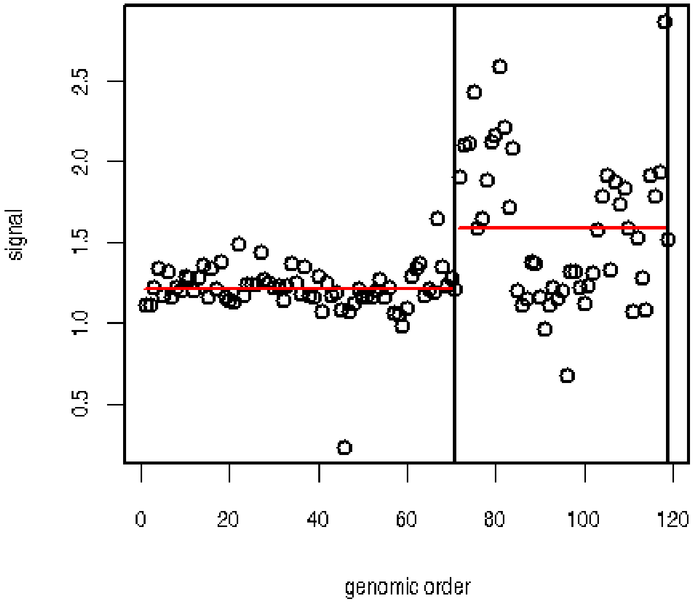
$$\beta pen(K) = \frac{K}{n}\sigma^2 \times \left(c_1 + c_2 \log \frac{n}{K}\right)$$

- Other methods have been developed, based on an adaptive estimation of K [6, 12].

Examples and comparison

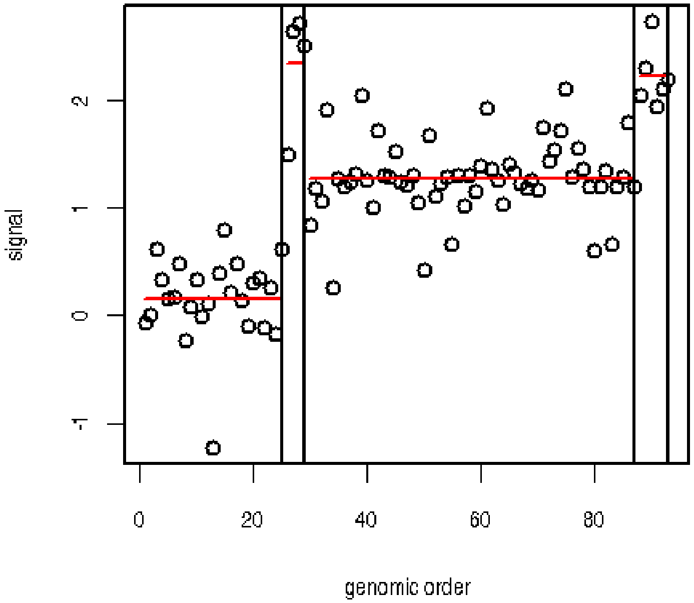


Homogeneous Variance

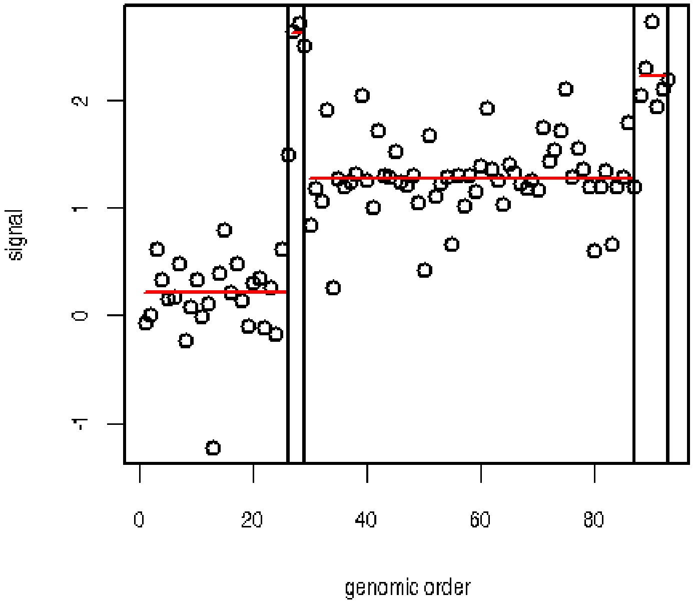


Heterogeneous Variances

Examples and comparison



Homogeneous Variance



Heterogeneous Variances

References

[1] D.G. Albertson, C. Collins, F. McCormick, and JW. Gray. Chromosome aberrations in solid tumors. *Nature Genetics*, 34(4):369–376, 2003.

[2] J.J. Davies, I.M. Wilson, and W.L. Lam. Array cgh technologies and their applications to cancer genomes. *Journal Chromosome Research*, 13(3):237–248, 2005.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38, 1977.

[4] L. Feuk, C.R. Marshall, R.F. Wintle, and S.W. Scherer. Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum. Mol. Genet.*, 15(1):57–66, 2006.

[5] J. Fridlyand, A. Snijders, D. Pinkel, D.G. Albertson, and A.N. Jain. Hidden markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*, 90(1):132–1533, 2004.

[6] M. Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501–1510, 2005.

[7] E. Lebarbier. Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85:717–736, 2005.

[8] D. P. Locke, R. Segraves, L. Carbone, N. Archidiacono, D.G. Albertson, D. Pinkel, and E.E. Eichler. Large-Scale Variation

Among Human and Great Ape Genomes Determined by Array Comparative Genomic Hybridization. *Genome Res.*, 13(3):347–357, 2003.

[9] W.W. Lockwood, C.B. Chi, and W.L. Lam. Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *European Journal of Human Genetics*, 14:139–148, 2006.

[10] J.C. Marioni, N.P. Thorne, and S. Tavare. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, 22(9):1144–1146, 2006.

[11] K. Nakao, K.R. Mehta, J. Fridlyand, D. H. Moore, A.J.Jain, A. Lafuente, J.W. Wiencke, J.P. Terdiman, and F.M. Waldman. High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*, 25(8):1345–1357, 2004.

[12] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J-J. Daudin. A statistical approach for CGH microarray data analysis. *BMC Bioinformatics*, 6:27, 2005.

[13] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. Kuo, C. Chen, Y. Zhai, S.H. Dairkee, B. Ljung, and J.W. Gray. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20:207–211, 1998.

[14] O.M. Rueda and R. Díaz-Uriarte. Flexible and accurate detection of genomic copy-number changes from acgh. *PLOS Computational Biology*, 3(6):1115–1122, 2007.

[15] D.F. Smeets. Historical prospective of human cytogenetics: from microscope to microarray. *Clin Biochem*, 37(6):439–446, 2004.

[16] A. M. Snijders, N. Nowak, R. Segraves, S. Blakwood, N. Brown, J. Conroy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J. P. Yue, J. W. Gray, A.N. Jain, D. Pinkel, and D. G. Albertson. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, 29:263–264, 2001.

[17] S. Stjernqvist, T. Ryden, M. Skold, and J. Staaf. Continuous-index hidden Markov modelling of array CGH copy number data. *Bioinformatics*, 23(8):1006–1014, 2007.