MUSÉUM
NATIONAL
D'HISTOIRE
NATURELLE

SEMIN-R

Breakpoint detection in biological and environmental sequences

Stéphane ROBIN

UMR 518 AgroParisTech / INRA
robin@agroparistech.fr

Joint work with

- E. Lebarbier, B. Thiam (UMR 518 AgroParisTech /INRA)
- F. Picard (CNRS, Lyon)

Outline

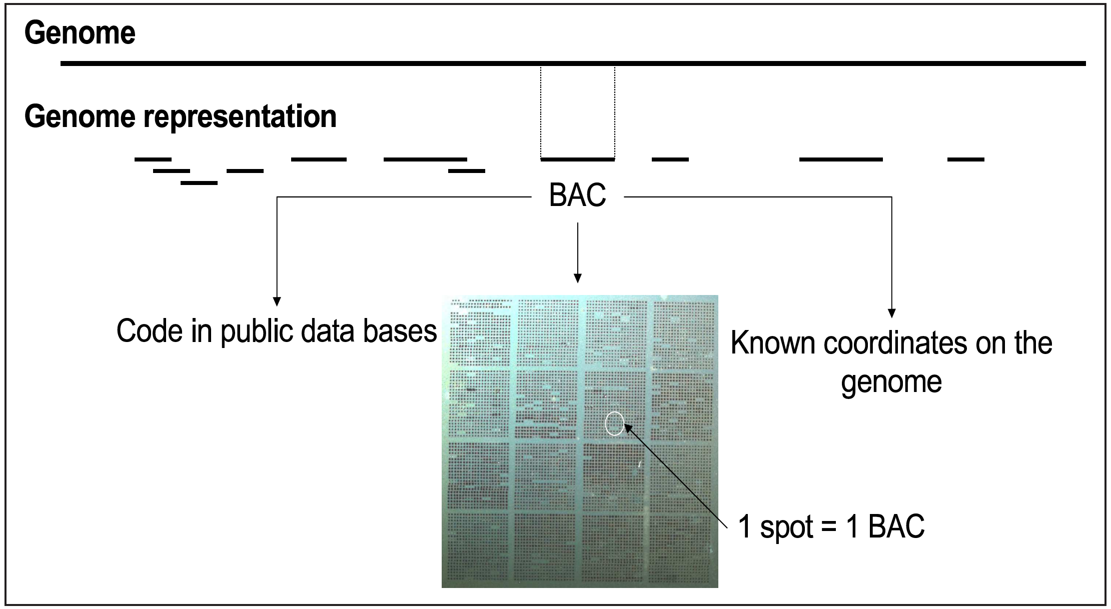
1. Simple segmentation problem: CGH array data
2. Breakpoint detection with covariates
3. Multiple segmentation

1 - Simple segmentation problem: CGH array data

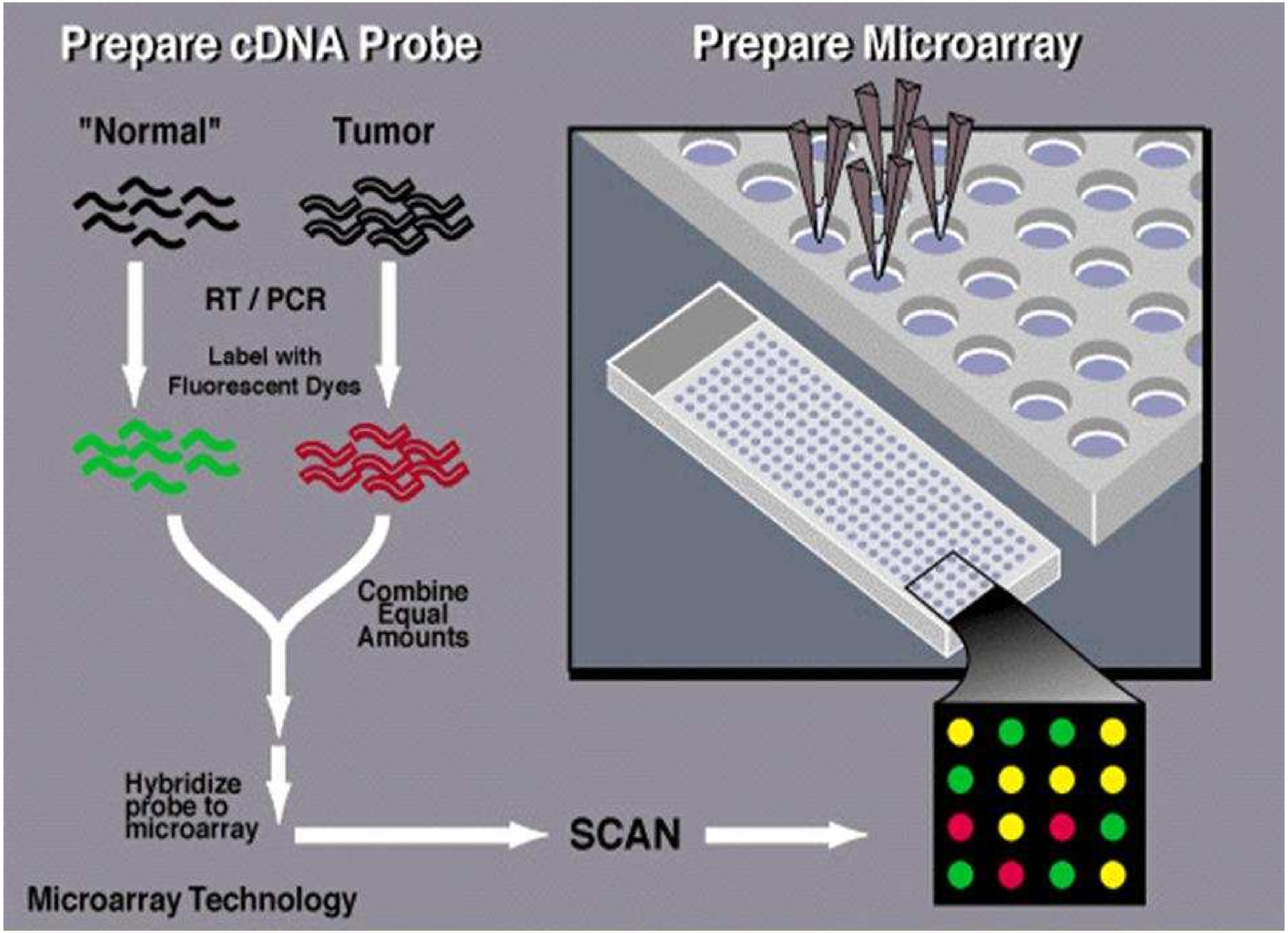
1.1 - Chromosomal aberrations and CGH arrays

CGH = Comparative Genomic Hybridization: method for the comparative measurement of relative DNA copy numbers between two samples (normal/disease, test/reference).

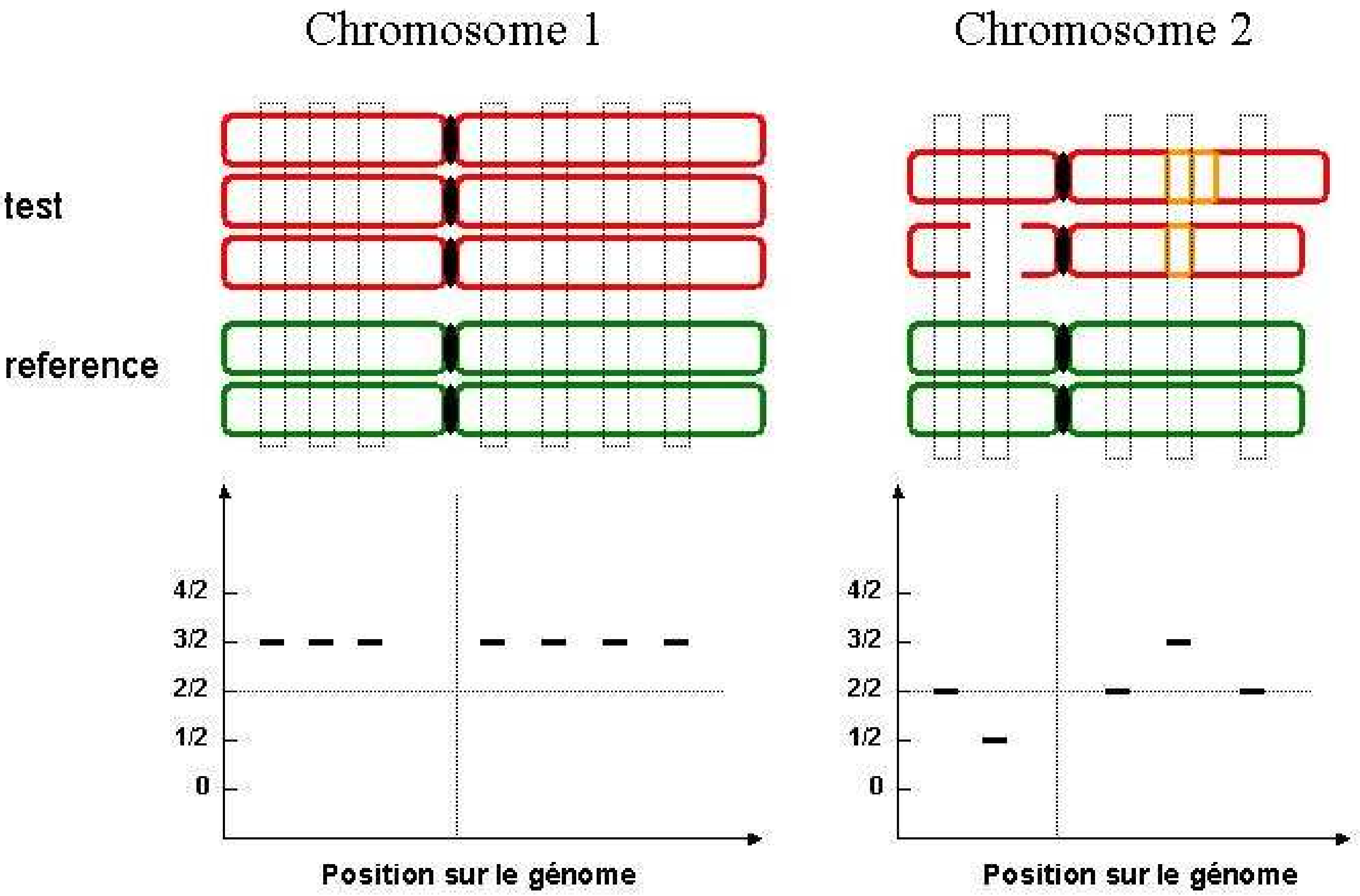
→ Application of the microarray technology to CGH (resolution ∼ 100kb).



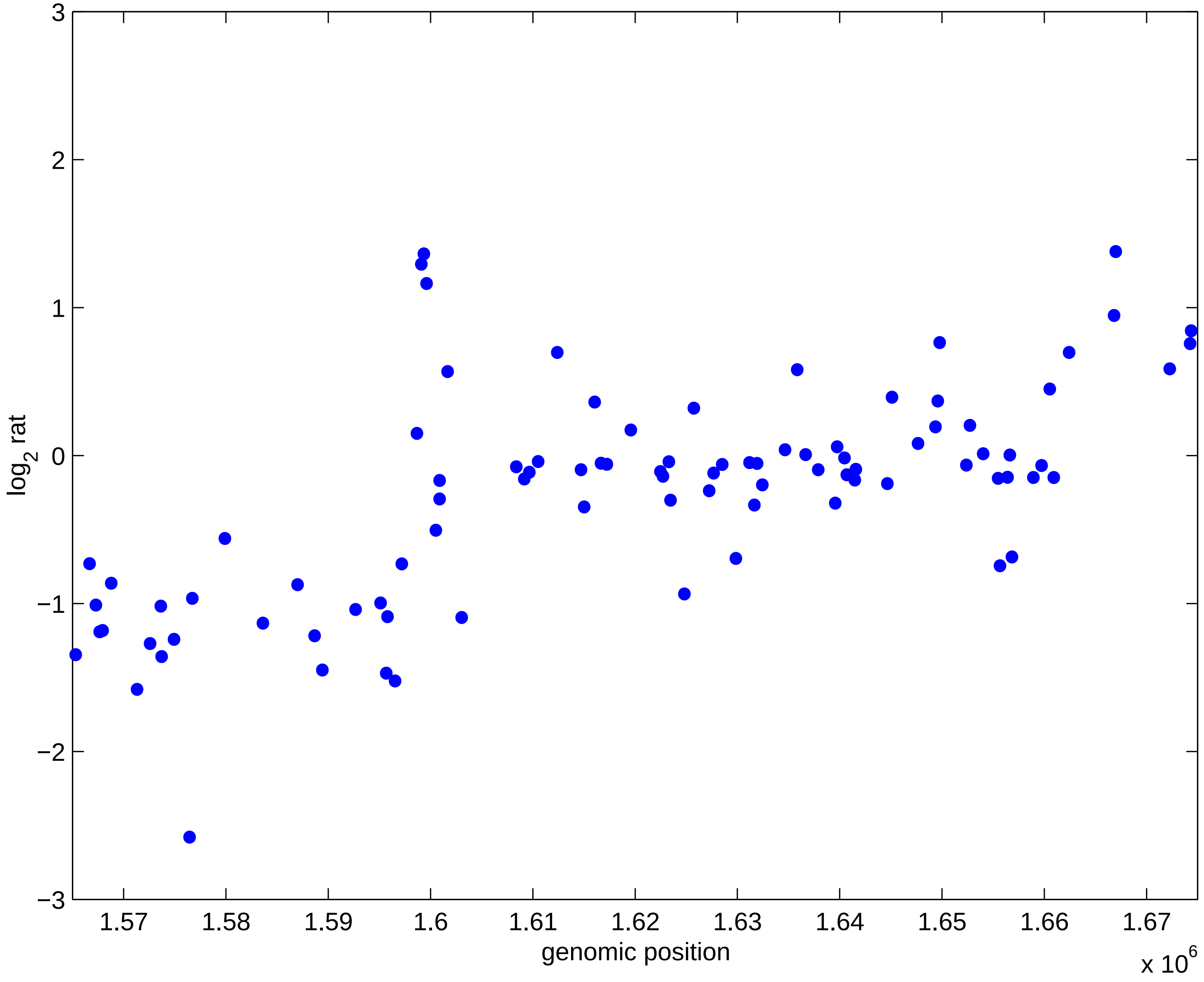
Microarray technology in its principle



Plotting the ratio along the chromosome

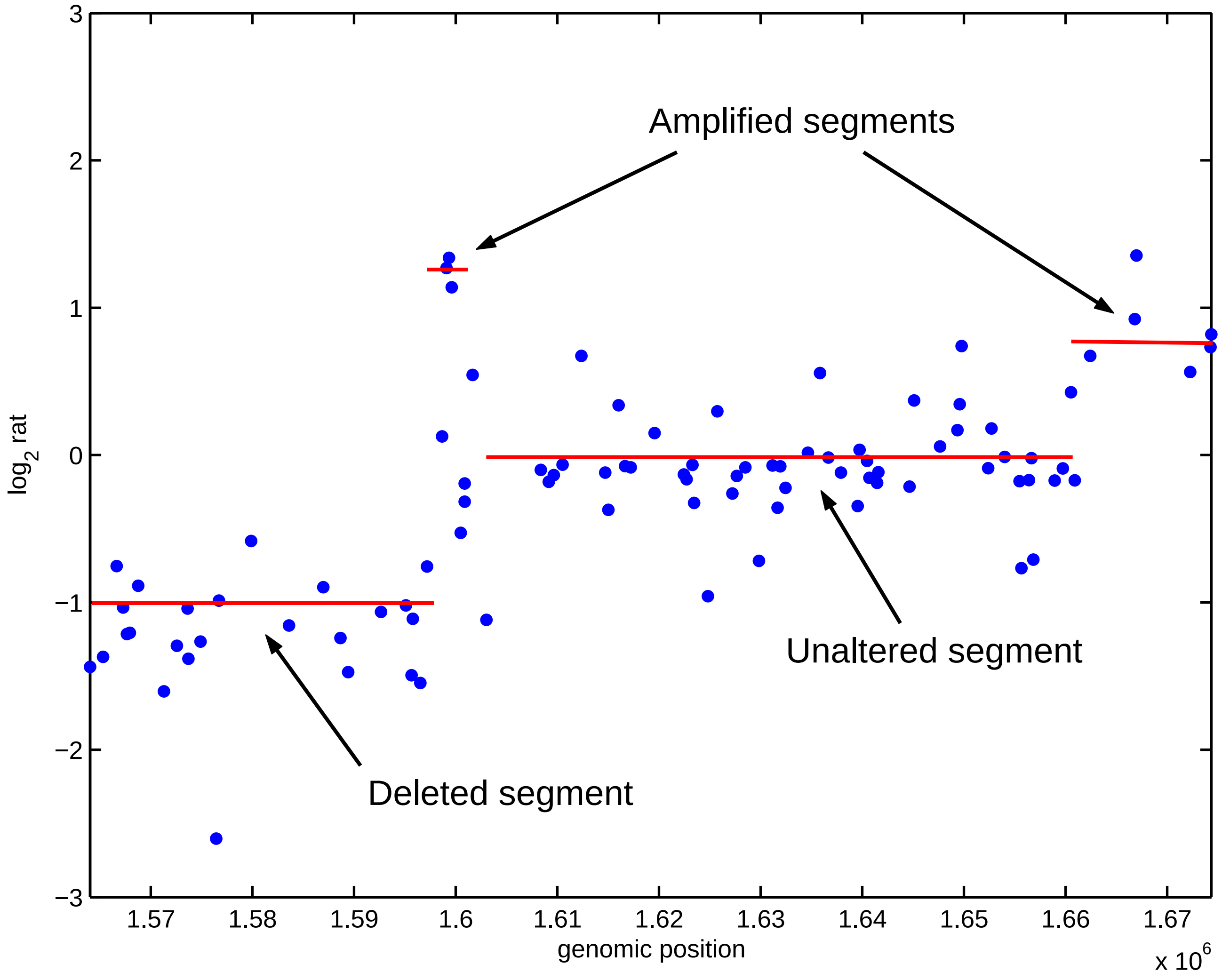


CGH profile. Because of the technical variability, the observed data look like this:



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

Interpretation of a CGH profile



Amplified segments

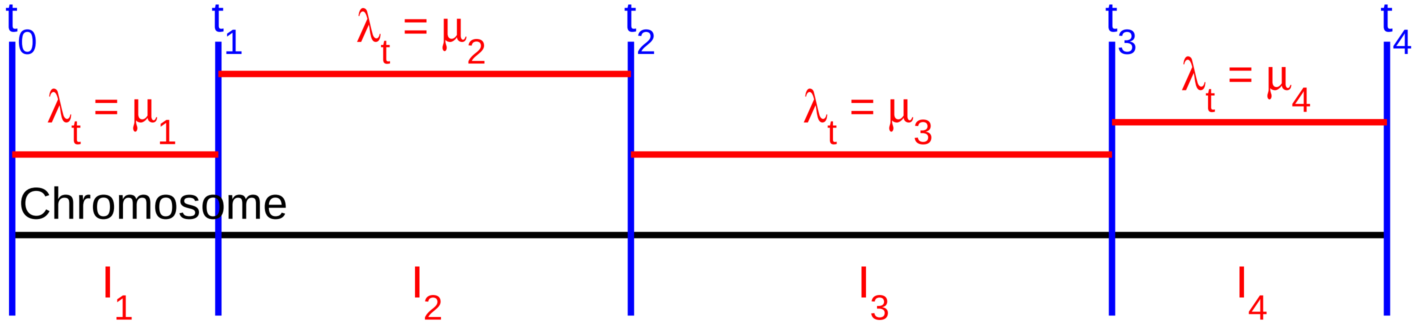
Unaltered segment

Deleted segment

1.2 - Model = What we have in mind

- At position t, there exists a 'true' log-ratio λ_t, which depends on the relative copy number.

- The value of the true log-ratio λ_t is affected by abrupt changes:



Position t_1, t_2, .. are called *breakpoints*. μ_k is the true log-ratio in segment I_k.

- The observed signal Y_t is noisy:

$$Y_t = \lambda_t + E_t.$$

Breakpoints detection aims at studying the spatial structure of the signal.

Statistical model

- The breakpoints define a partition of the data into K segments of size n_k:

$$I_k = \{t, t \in]t_{k-1}, t_k]\}.$$

- Suppose that those parameters are constant between two changes:

$$\text{if position } t \text{ is in segment } I_k, \qquad Y_t = \mu_k + E_t \sim \mathcal{N}(\mu_k, \sigma^2_{(k)}).$$

- The parameters of this model are:

$$T = (t_1, ..., t_{K-1}), \qquad \Theta = (\theta_1, \dots, \theta_K), \quad \theta_k = (\mu_k, \sigma^2_{(k)}).$$

- The model can rewritten as a *regression model*:

$$\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{E}$$

$$\begin{array}{lccl} \text{where} & \mathbf{T} & = & \textit{unknown } n \times K \text{ segmentation matrix } (Y_{tk} = \mathbb{I}\{i \in I_k\}), \\ & \boldsymbol{\mu} & = & \text{vector of the } K \text{ segment means.} \end{array}$$

Estimating the parameters

Log-Likelihood (with a constant variance σ^2):

$$\begin{aligned} 2\mathcal{L}_K(T,\Theta) &= 2\sum_{k=1}^K \log\phi(\{Y_t\}_{t\in I_k};\theta_k) = 2\sum_{k=1}^K \sum_{t\in I_k} \log\phi(Y_t;\theta_k) \\ &= -n\log\sigma^2 - \frac{1}{\sigma^2}\sum_{k=1}^K \sum_{t\in I_k}(Y_t-\mu_k)^2 + \text{cst}. \end{aligned}$$

- Because the data are supposed to be independent, the log-likelihood is a sum over all the segments (*additive contrast*).
- Because the data are supposed to be Gaussian, maximum likelihood estimation is equivalent to *least squares* fitting.
- When the segments are known, estimation is straightforward: $\widehat{\mu}_k = \frac{1}{n_k}\sum_{t\in I_k} Y_t$.

How to find the breakpoints?

When K is known , we have to minimise

$$J_k(1, n) = \sum_{k=1}^{K} \sum_{t \in I_k} (Y_t - \widehat{\mu}_k)^2.$$

- There are $\left(\begin{array}{c} n-1 \\ K-1 \end{array}\right)$ possible choices for the positions of the breakpoints $t_1, t_2, \dots, t_{K-1}$:

$\Rightarrow$ Impossible to explore for large n and K

- $\sum_{t \in I_k} (Y_t - \widehat{\mu}_k)^2$ can be viewed as the 'cost' of segment I_k, i.e. the cost of putting data $Y_{t_{k-1}+1}$ to $Y_{t_{k+1}}$ in a single segment.

- The optimisation problem is actually a shortest path problem that can be solved thanks to dynamic programming.

Dynamic programming. Based on Bellmann's optimality principle:

Sub-paths of the optimal path are themselves optimal.

Initialisation: For $0 \leq i < j \leq n$:

$$J_1(i,j) = \sum_{t=i+1}^{j} (Y_t - \widehat{\mu})^2.$$

Step k: For $2 \leq k \leq K$:

$$J_k(i,j) = \min_{i \leq h \leq j} \left[J_{k-1}(1,h) + J_1(h+1,j) \right].$$

J_k is called the cost matrix.

The global optimum is given by $J_k(1,n)$.

Example with R

Cost matrix:

```
lmin = 2
C = matrix(Inf, n, n)
for (i in (1:(n-lmin)))
  {
  for (j in ((i+lmin):n))
  {
    reg = lm(y[i:j] ~ x[i:j])
    C[i, j] = sum(reg$residuals^2)
    }
  }
```

Breakpoints:

```
$t.est
     [,1] [,2] [,3] [,4] [,5]
[1,]   40    0    0    0    0
[2,]   10   40    0    0    0
[3,]   16   30   40    0    0
[4,]   10   16   30   40    0
[5,]   10   16   24   30   40
```

Contrasts:

```
$J.est
 [1] 23.8693554  9.8660559  2.6290695  1.5546431  1.2213389
```

One last problem: the selection of K

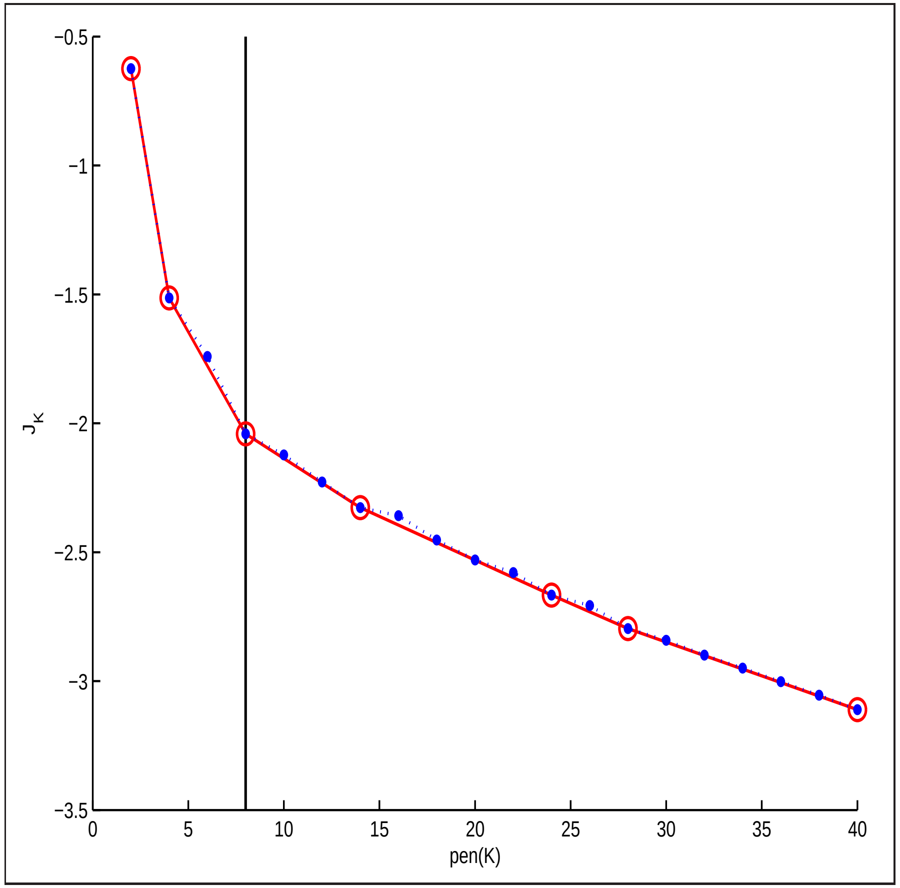
- The contrast J_K necessarily decreases when the model becomes more complex.

- The penalty function measures this complexity: $pen(K) = K+1$ with constant variance, $2K$ with heterogeneous variance.

- We look for the minimum of

$$J_k + \beta pen(K)$$

where β is adaptively estimated (*Lavielle(2003)*).



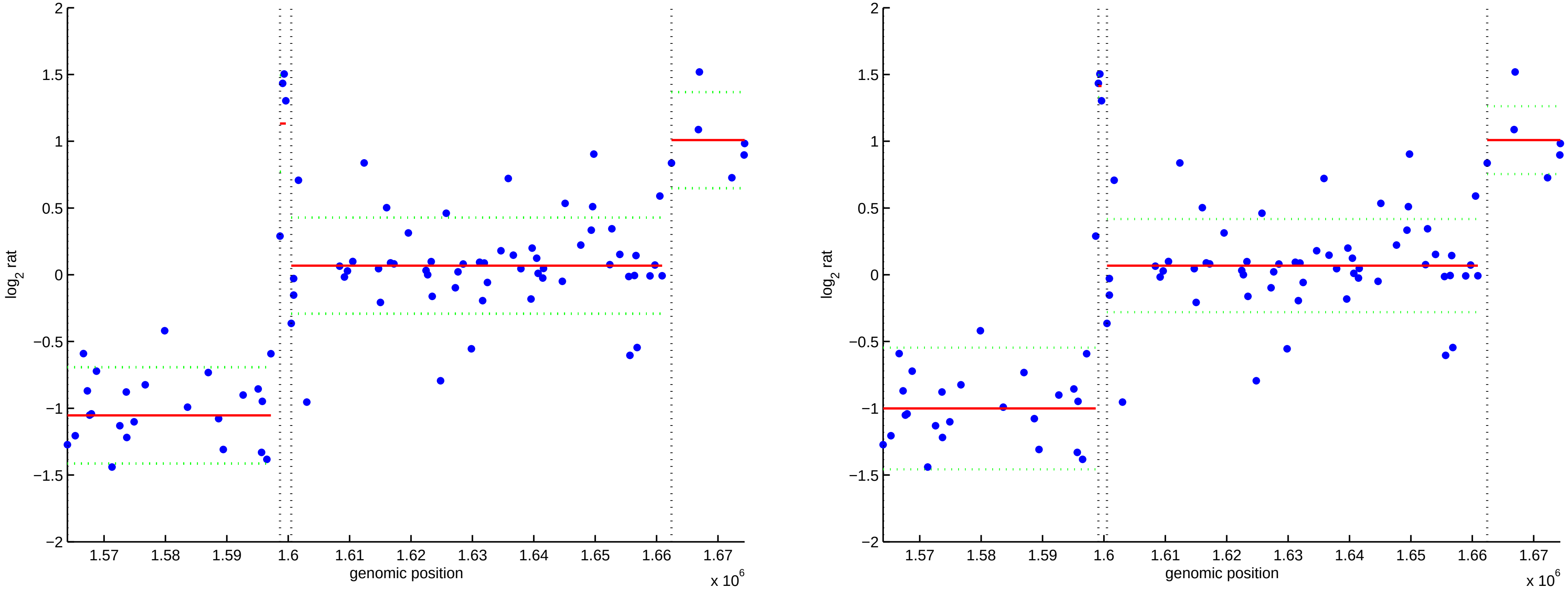
1.3 - Example of segmentation on array CGH data

Are the variances σ_k^2 homogeneous? BT474 cell line, chromosome 9:

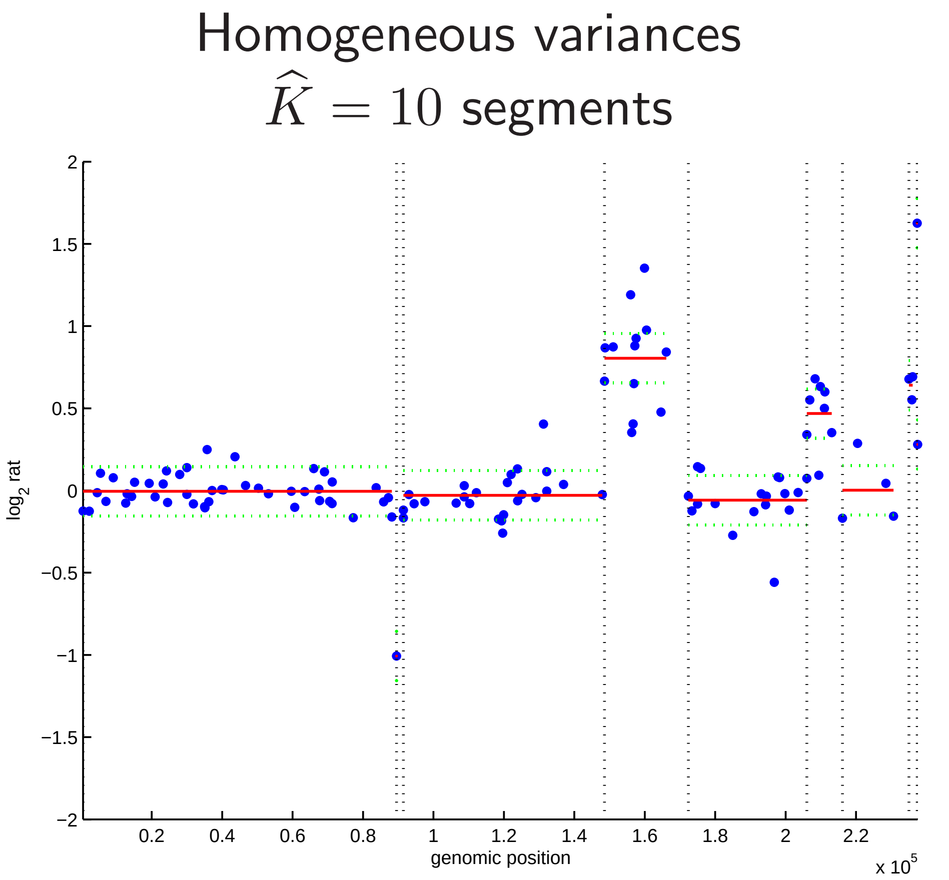
Homogeneous variances

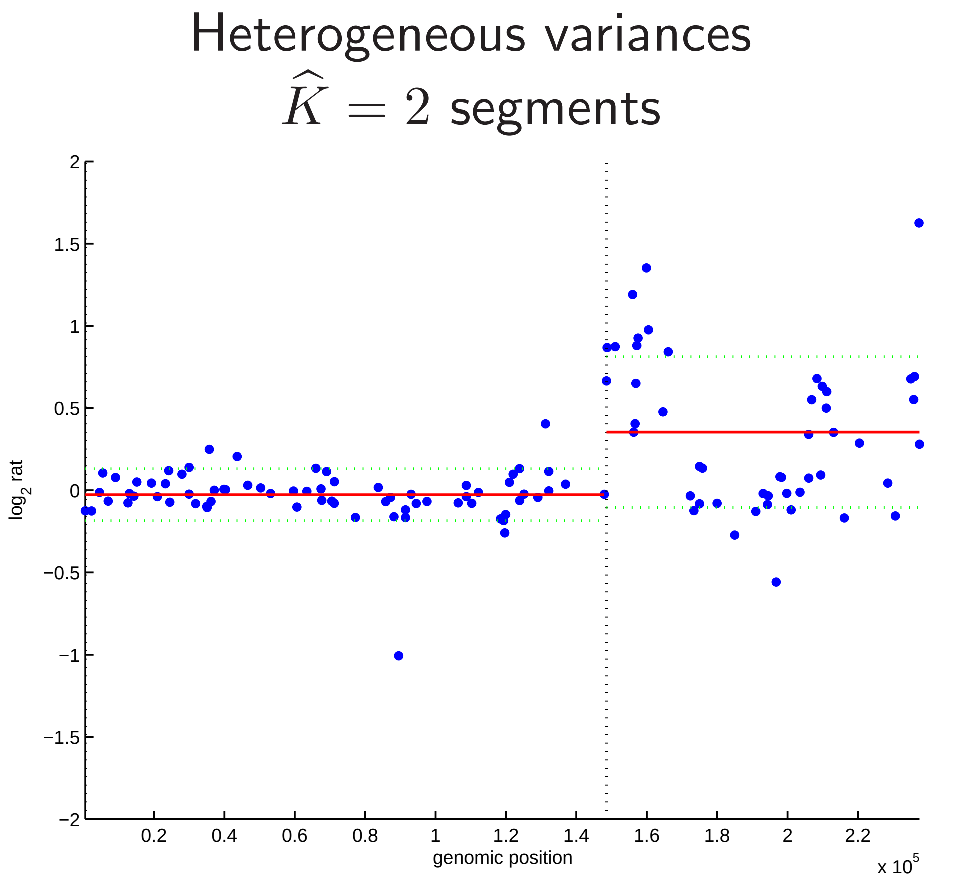
Heterogeneous variances

$K = 4$ segments



Adaptive choice of the number of segments. BT474 cell line, chromosome 1:

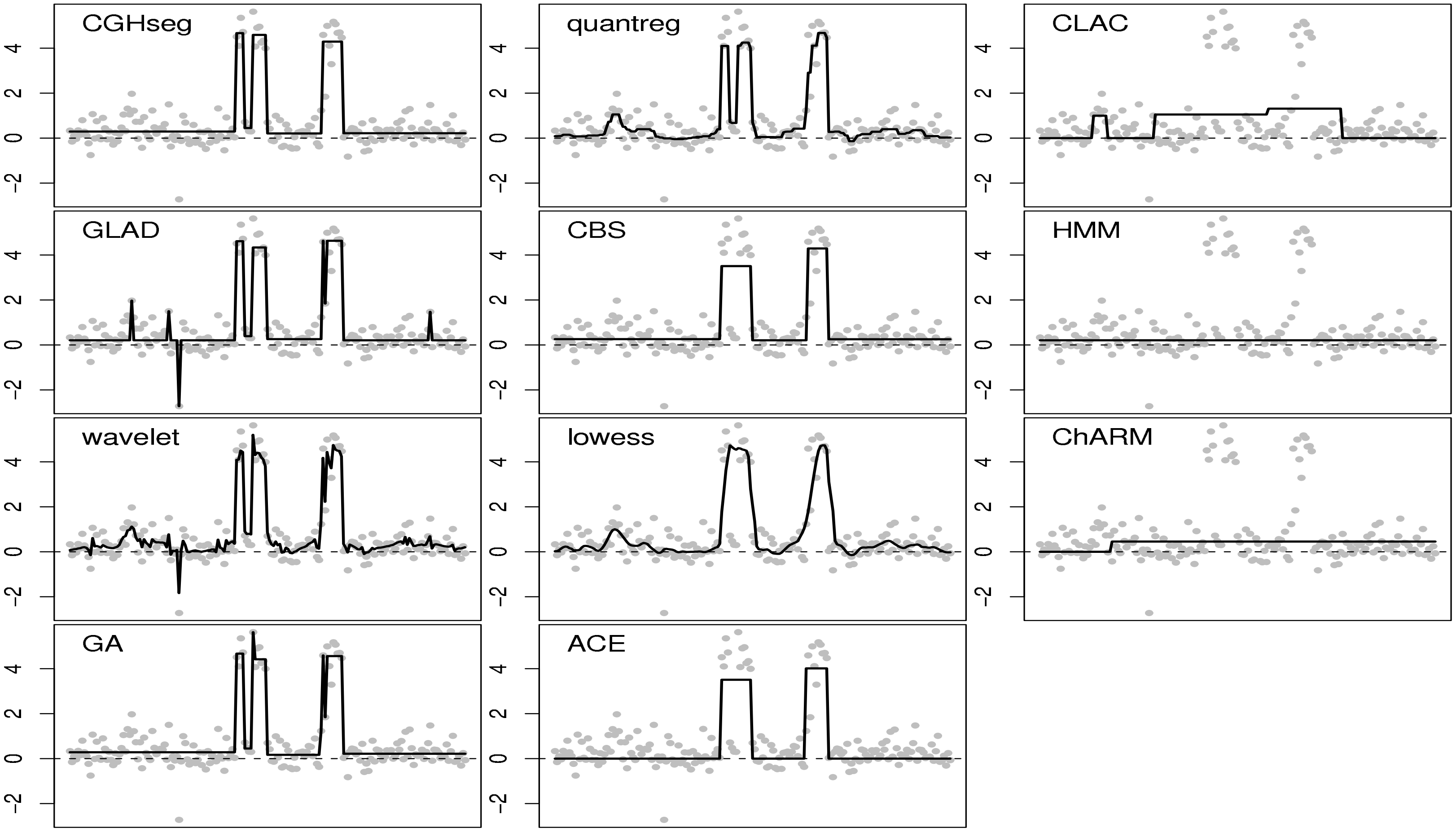




Homogeneous variances result in smaller segments. *Picard & al, 05*

Comparative study

Lai & al. (Bioinformatics, 05). On both synthetic and real data (GBM brain tumor data), the methods performs well.



2 - Breakpoint detection with covariates

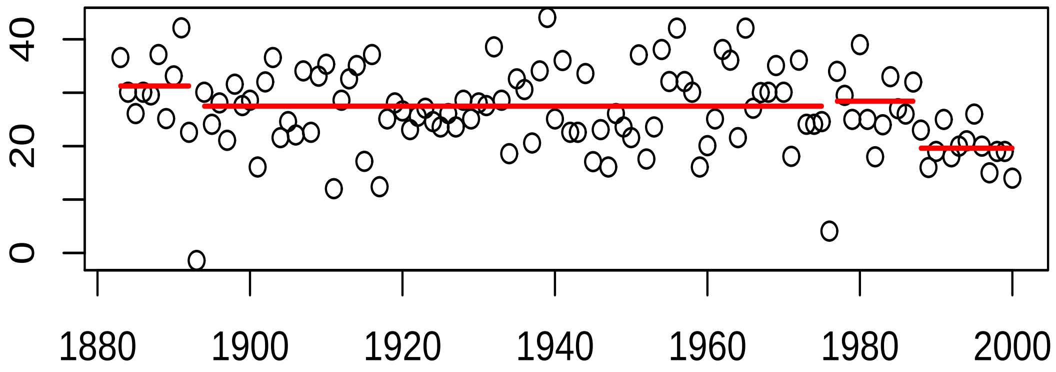
2.1 - Harvest data

Data = Harvest dates and temperatures in Ouges (Burgundy) since 1882. *Chuine, 04*

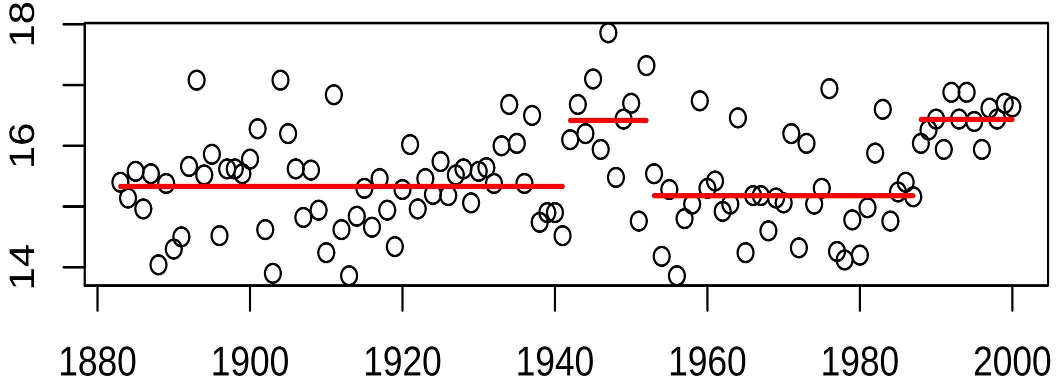
A breakpoint is detected in *both series in 1986*.

Is the 1986 breakpoint observed in harvest data caused by the corresponding rupture in the temperatures,

Harvest time



Mean summer temperature



2.2 - Regression / segmentation model

Denote Y_t = harvest date at year t,
x_t = temperature at year t,
I_k = k-th segment ($I_k = \{t, t \in]t_{k-1}, t_k]\}$).

The model is, for $t \in I_k$,

$$Y_t = \underbrace{bx_t}_{\text{regression}} + \underbrace{\mu_k}_{\text{segmentation}} + E_t$$

Matrix form. The model can be written as

$$\mathbf{Y} + \mathbf{X}\boldsymbol{\theta} + \mathbf{T}\boldsymbol{\mu} + \mathbf{E}$$

where $\mathbf{X}$ = known matrix of regressors (vector of temperatures),
$\boldsymbol{\theta}$ = vector of regression coefficients ($\boldsymbol{\theta} = [b]$),
$\mathbf{T}$ = *unknown* segmentation matrix ($Y_{tk} = \mathbb{I}\{i \in I_k\}$),
$\boldsymbol{\mu}$ = vector of segment means.

Heuristic estimation procedure

Least squares criterion. We look for

$$\min_{b,\{I_k\},\{\mu_k\}} \sum_k \sum_{t\in I_k} (Y_t - bx_t - \mu_k)^2,$$

which is *not additive*, since b is common to all segments.

Iterative heuristic. Set $F_t^0 = Y_t$ and iterate until convergence of $b^h, \{I_k^h\}, \{\mu_k^h\}$:

1. Segmentation step:

$$\min_{\{I_k\},\{\mu_k\}} \sum_k \sum_{t\in I_k} (F_t^h - \mu_k)^2, \qquad \longrightarrow \qquad \text{for } t \in i_k^{h+1}: \quad G_t^{h+1} = Y_t - \mu_k^{h+1};$$

2. Regression step:

$$\min_b \sum_k \sum_{t\in I_k} (G_t^{h+1} - bx_t)^2, \qquad \longrightarrow \qquad F_t^{h+1} = Y_t - b^{h+1}x_t.$$

Results When accounting for temperature, the breakpoint at $t = 1986$ vanishes.

Segmentation for harvest dates

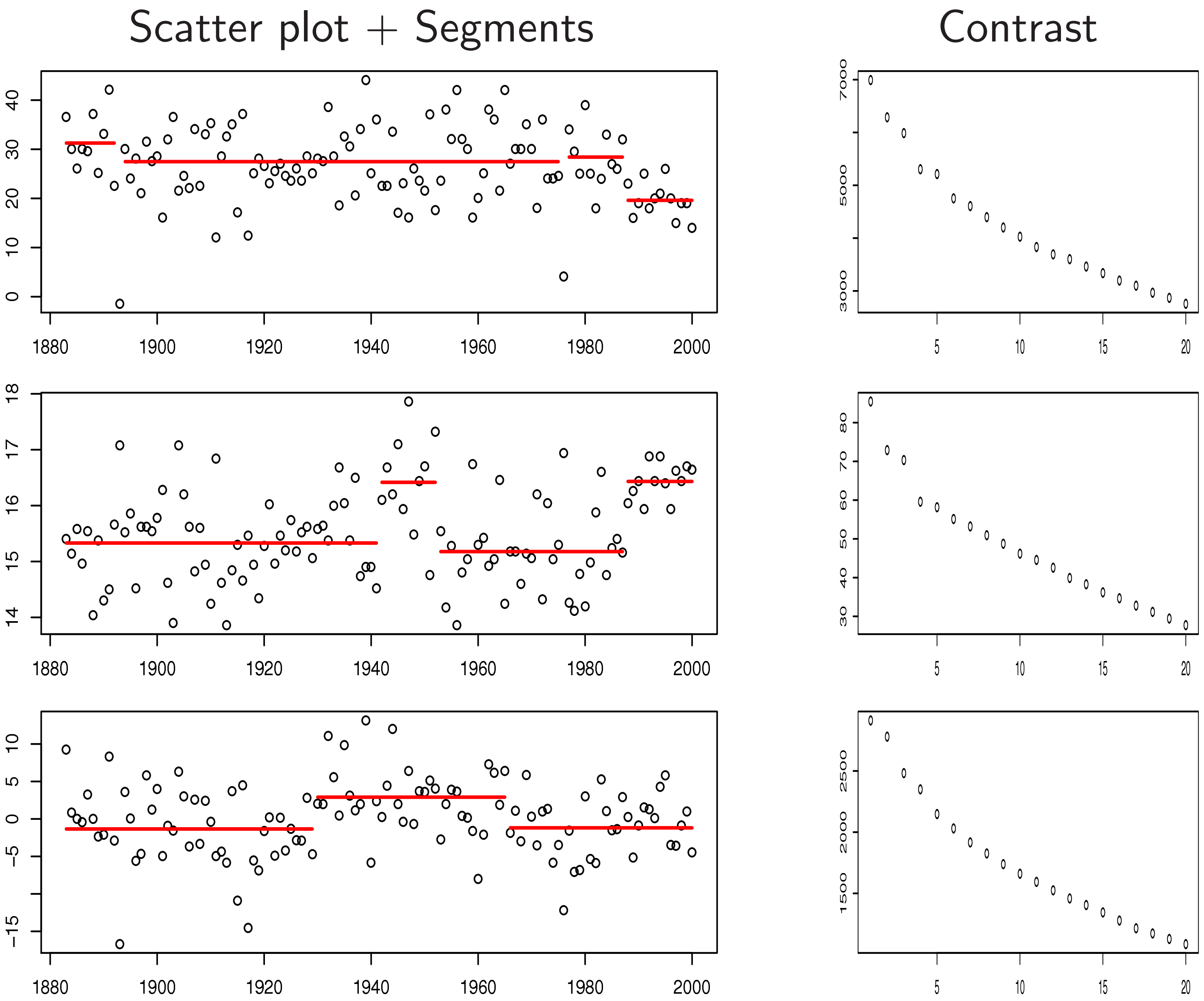
$K = 4$ (2, 6?)

Segmentation for temperatures

$K = 4$ (2?)

Segmentation / regression for harvest dates

$K = 3$ (1?)



Scatter plot + Segments

Contrast

3 - Multiple segmentation

3.1 - Examples

Breakpoints in temperature series

Consider the temperatures series $\{Y_{it}\}$ in several French cities ($i = 1..m$), we look for *common breakpoints* in the climate slope b accounting for a (random) *city effect U_i*:

$$t \in I_k \quad \Rightarrow \quad Y_{it} = \mu + U_i + b_k t + E_{it}$$

where $\{U_i\}$ are i.i.d. $\mathcal{N}(0, \gamma^2)$ and $\{E_{it}\}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$.

This model induces a correlation between all temperatures collected in the same city:

$$\mathbb{C}\text{ov}(Y_{it}, Y_{i,t'}) = \gamma^2 \qquad \Rightarrow \qquad \mathbb{C}\text{orr}(Y_{it}, Y_{i,t'}) = \frac{\gamma^2}{\gamma^2 + \sigma^2}.$$

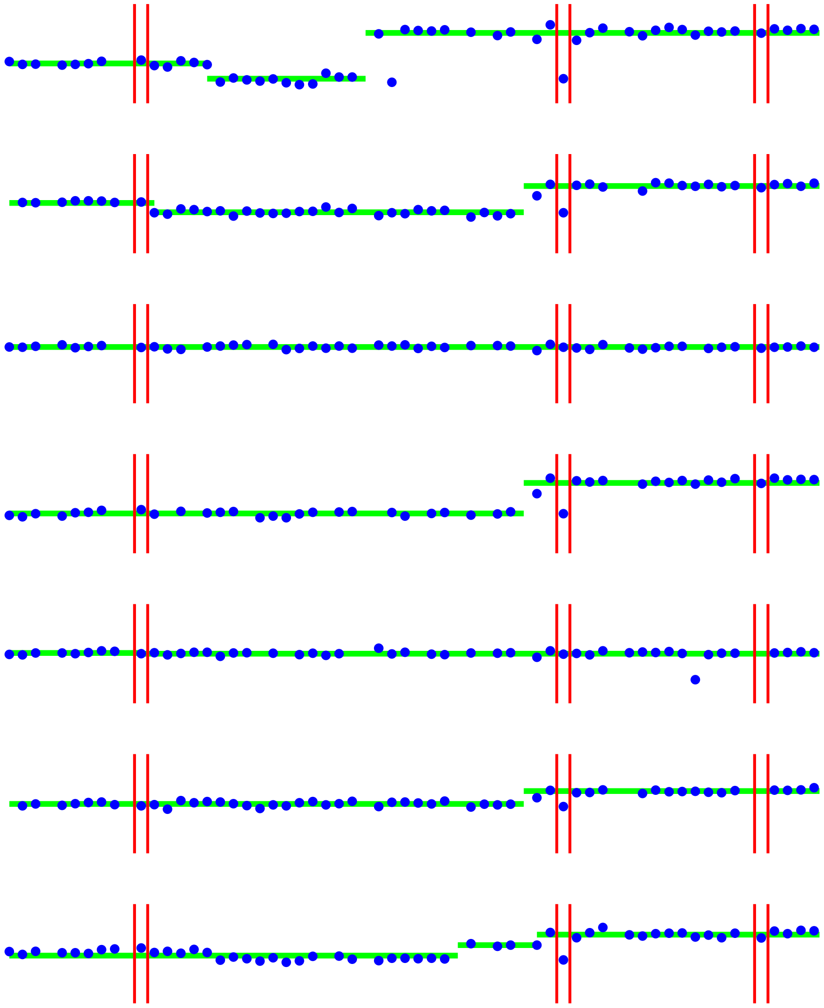
Chromosomal aberrations in a set of patients

Consider the CGH profiles $\{Y_{it}\}$ of a set of patients ($i = 1..m$), we look for *individual breakpoints* accounting for a (random) *probe effect U_t*:

$$t \in I_{ik} \quad \Rightarrow \quad Y_{it} = \mu_{ik} + U_t + E_{it}.$$

U_t accounts for different probe affinities that *may alter all the profiles* at the same position.

The random term induces a correlation between all these measurements.



3.2 - Mixed linear model with breakpoints

The general formulation of the model is

$$\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{Z}\mathbf{U} + \mathbf{E}$$

where

$\mathbf{Y}$: profiles,

$\mathbf{T}$ segments (*unknown → to estimate*),

$\boldsymbol{\mu}$ mean signal in each segment (*unknown → to estimate*),

$\mathbf{Z}$ design matrix of the random effect,

$\mathbf{U}$ vector of random effect (*unobserved*): $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$ ($\mathbf{G}$ *unknown → to estimate*),

$\mathbf{E}$ residual (unobserved): $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ ($\mathbf{R}$ *diagonal, unknown → to estimate*).

Estimation of the parameters

Direct maximisation of the likelihood. The marginal distribution of $\mathbf{Y}$ is

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta} + \mathbf{T}\boldsymbol{\mu}, \mathbf{V}), \qquad \text{where } \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}.$$

Because, $\mathbf{V}$ is not diagonal, the direct maximisation of the observed log-likelihood $\mathcal{L}(\mathbf{Y})$ leads to the minimisation of a non additive contrast.

Dynamic programming *can not be used* to estimate $\mathbf{T}$ and $\boldsymbol{\mu}$

E-M strategy. Its conditional distribution given $\mathbf{U}$ is

$$(\mathbf{Y} \mid \mathbf{U}) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta} + \mathbf{T}\boldsymbol{\mu} + \mathbf{Z}\mathbf{U}, \mathbf{R}).$$

In the E-M algorithm (*Foulley, lecture notes*), the unobserved effect $\mathbf{U}$ is predicted, so we have to maximise $\mathcal{L}(\mathbf{Y} \mid \mathbf{U})$, which involves an additive contrast since $\mathbf{R}$ is diagonal.

Dynamic programming can be used to estimate $\mathbf{T}$ *and* $\boldsymbol{\mu}$

A DP-EM algorithm

E step. Calculate the conditional moments of the random effect given the data:

$$\widehat{\mathbb{E}}(\mathbf{U}|\mathbf{Y}), \qquad \widehat{\mathbb{V}}(\mathbf{U}|\mathbf{Y}).$$

M step. Denoting $\widehat{\mathbf{U}} = \widehat{\mathbb{E}}(\mathbf{U}|\mathbf{Y})$, perform the segmentation as follows:

$$\widehat{\mathbf{T}\boldsymbol{\mu}} = \arg\min_{\mathbf{T}\boldsymbol{\mu}} \|\mathbf{Y} - \mathbf{T}\boldsymbol{\mu} - \mathbf{Z}\widehat{\mathbf{U}}\|^2.$$

A *two-stage dynamic programming* is required to achieve this step for numerous patients. *Picard et al.*

Segclust package.

`http://cran.r-project.org/web/packages/segclust/index.html`

3.3 - Applications

Breakpoints in temperature series

Data. For several locations ($m = 25$), we measure the minimal daily temperature, averaged for each year from 1957 to 2004. (Source: Meteo France).

Model. $t \in I_k$

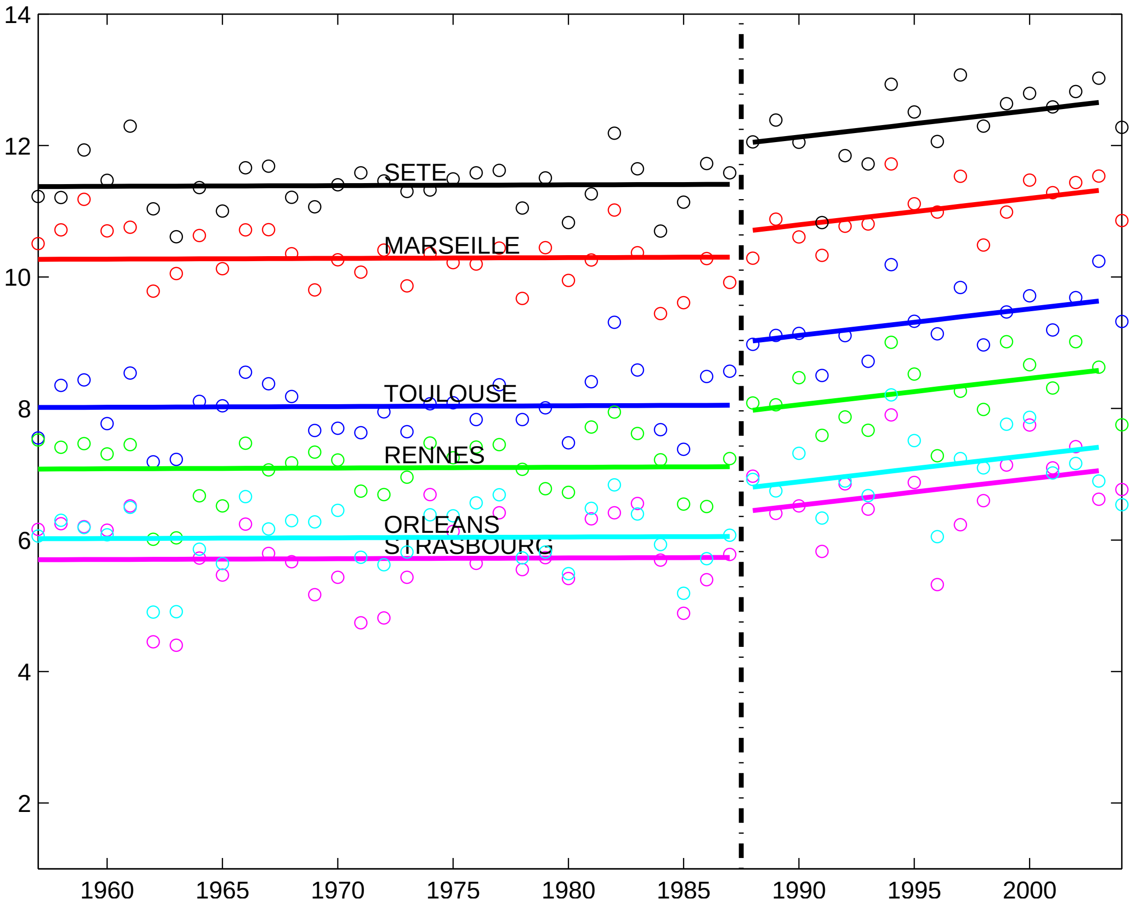
$\Rightarrow Y_{it} = \mu + U_i + b_k t + E_{it}$.

Estimates.

$\widehat{b}_1 = 1.8\,10^{-3}$,

$\widehat{b}_2 = 2.5\,10^{-2}$,

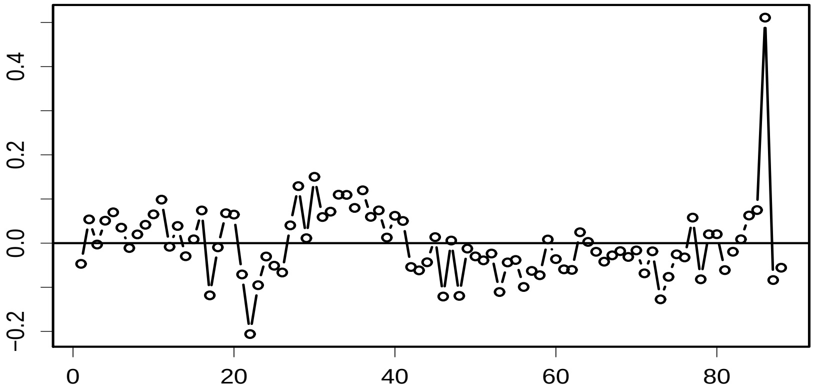
$\widehat{\gamma} = 2.0, \quad \widehat{\sigma} = 0.51$.



CGH profile: Bladder cancer data

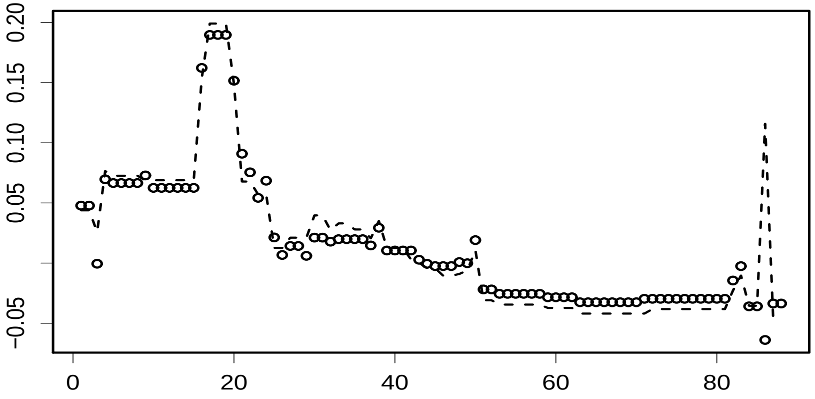
Global analysis (Inst. Curie, F. Radvanyi)

We find a *large positive random effect* U_t has at position 87.



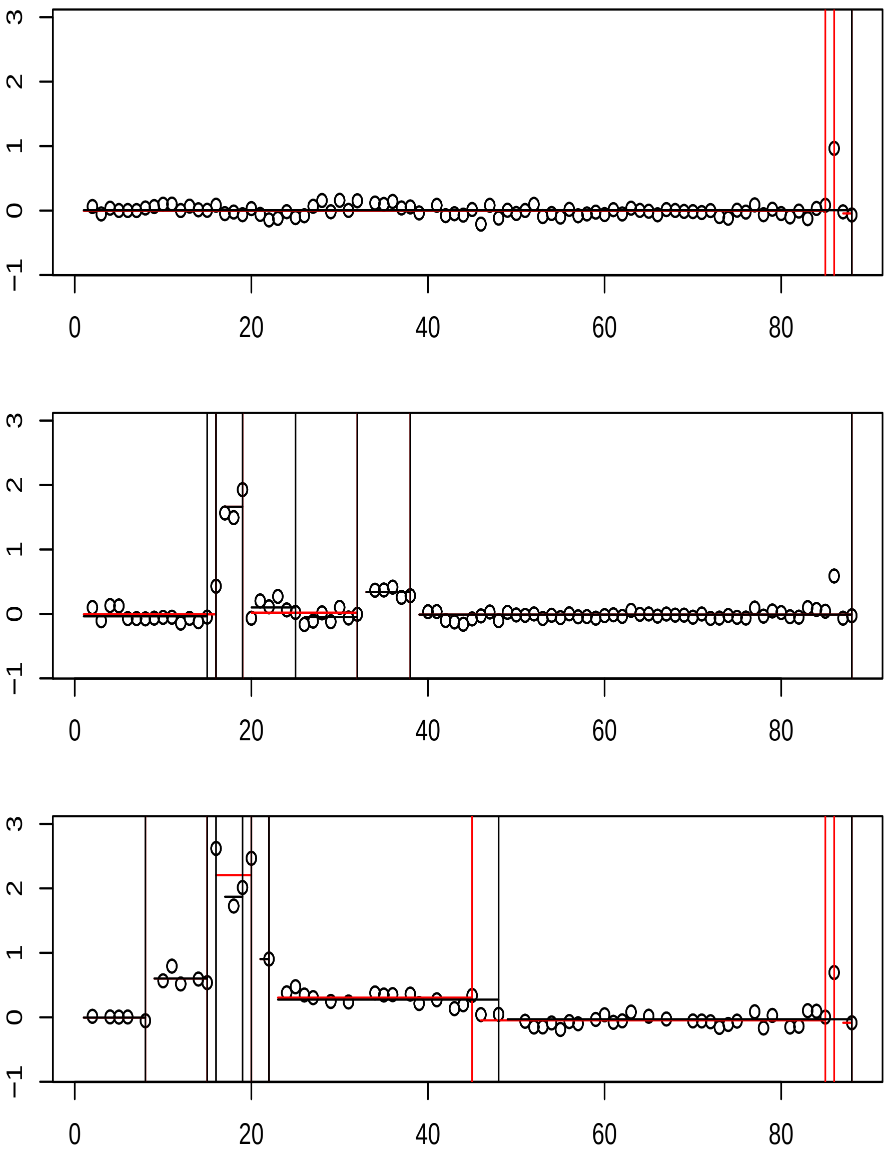
$\longrightarrow$ *Poor probe affinity*?
$\longrightarrow$ *Wrong annotation*?
$\longrightarrow$ *Polymorphism*?

The mean profile of the whole set of patients can be corrected from the probe effect:
$(\cdots)$ mean of raw profiles,
$(\circ)$ mean of corrected profiles



Individual profiles. The random effect has an influence on the segmentation.

- Breakpoints around position 86 are detected in individual profiles when analysed independently (–).
- They vanish after correction of the probe effect vanish (–).


−1
0
1
2
3
0
20
40
60
80