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

Chromosome 1

Chromosome 2

Position sur le génome

S. Robin: Breakpoint detection
CGH profile. Because of the technical variability, the observed data look like this:

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

S. Robin: Breakpoint detection
1.2 - Model = What we have in mind

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

- The value of the true log-ratio $\lambda_t$ is affected by abrupt changes:

  \[
  \begin{align*}
  \lambda_t &= \mu_1 \\
  \lambda_t &= \mu_2 \\
  \lambda_t &= \mu_3 \\
  \lambda_t &= \mu_4
  \end{align*}
  \]

  Position $t_1, t_2, ..$ are called breakpoints. $\mu_k$ 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:

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

• The parameters of this model are:

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

• The model can rewritten as a regression model:

$$Y = T\mu + E$$

where $T = unknown$ $n \times K$ segmentation matrix ($Y_{tk} = \text{I}\{i \in I_k\}$), $\mu$ = vector of the $K$ segment means.
Estimating the parameters

Log-Likelihood (with a constant variance $\sigma^2$):

$$2\mathcal{L}_K(T, \Theta) = 2 \sum_{k=1}^{K} \log \phi(\{Y_t\}_{t \in I_k}; \theta_k) = 2 \sum_{k=1}^{K} \sum_{t \in I_k} \log \phi(Y_t; \theta_k)$$

$$= -n \log \sigma^2 - \frac{1}{\sigma^2} \sum_{k=1}^{K} \sum_{t \in I_k} (Y_t - \mu_k)^2 + \text{cst.}$$

- 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: $\hat{\mu}_k = \frac{1}{n_k} \sum_{t \in I_k} Y_t$. 
How to find the breakpoints?

When $K$ is known, we have to minimise

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

- There are $\binom{n-1}{K-1}$ possible choices for the positions of the breakpoints $t_1, t_2, \ldots, t_{K-1}$:

$$\Rightarrow \text{Impossible to explore for large } n \text{ and } K$$

- $\sum_{t \in I_k} (Y_t - \hat{\mu}_k)^2$ 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 $0 \leq i < j \leq n$:

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

**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:
\[ l_{\text{min}} = 2 \]
\[ C = \text{matrix}(\text{Inf}, n, n) \]
\[
\text{for } (i \text{ in } (1:(n-l_{\text{min})))) \\
\quad \{ \text{for } (j \text{ in } ((i+l_{\text{min}}):n)) \\
\quad \quad \{ \text{reg} = \text{lm}(y[i:j] \sim x[i:j]) \\
\quad \quad \quad C[i, j] = \text{sum}(\text{reg}$\$\text{residuals}^2) \\
\quad \quad \} \\
\quad \} \\
\]

Breakpoints:
\[ t_{\text{est}} \]
\[
\begin{bmatrix}
[1,] & 40 & 0 & 0 & 0 & 0 \\
[2,] & 10 & 40 & 0 & 0 & 0 \\
[3,] & 16 & 30 & 40 & 0 & 0 \\
[4,] & 10 & 16 & 30 & 40 & 0 \\
[5,] & 10 & 16 & 24 & 30 & 40 \\
\end{bmatrix}
\]

Contrasts:
\[ J_{\text{est}} \]
\[
\begin{bmatrix}
\end{bmatrix}
\]
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: $\text{pen}(K) = K + 1$ with constant variance, $2K$ with heterogeneous variance.

- We look for the minimum of

$$J_k + \beta \text{pen}(K)$$

where $\beta$ is adaptively estimated (Lavielle(2003)).
1.3 - Example of segmentation on array CGH data

Are the variances $\sigma_k^2$ 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

$\hat{K} = 10$ segments

Heterogeneous variances

$\hat{K} = 2$ segments

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.
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,
2.2 - Regression / segmentation model

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

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

$$Y_t = bx_t + \mu_k + E_t$$

regression segmentation

Matrix form. The model can be written as

$$Y + X\theta + T\mu + E$$

where $X = \text{known matrix of regressors (vector of temperatures)}$, $\theta = \text{vector of regression coefficients } (\theta = [b])$, $T = \text{unknown segmentation matrix } (Y_{tk} = I\{i \in I_k\})$, $\mu = \text{vector of segment means}$.
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 \(F^0_t = Y_t\) and iterate until convergence of \(b^h, \{I^h_k\}, \{\mu^h_k\}\):

1. Segmentation step:

\[
\min_{\{I_k\},\{\mu_k\}} \sum_k \sum_{t \in I_k} (F^h_t - \mu_k)^2, \quad \rightarrow \quad \text{for } t \in i^h_{k+1} : \quad G_{t}^{h+1} = Y_t - \mu^h_{k+1};
\]

2. Regression step:

\[
\min_b \sum_k \sum_{t \in I_k} (G_{t}^{h+1} - bx_t)^2, \quad \rightarrow \quad F_{t}^{h+1} = Y_t - b^{h+1}x_t.
\]
**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

$K = 3 \ (1?)$
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 \(U_i\):

\[
t \in I_k \Rightarrow Y_{it} = \mu + U_i + b_k t + E_{it}
\]

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:

\[
\text{Cov}(Y_{it}, Y_{i,t'}) = \gamma^2 \quad \Rightarrow \quad \text{Corr}(Y_{it}, Y_{i,t'}) = \frac{\gamma^2}{\gamma^2 + \sigma^2}.
\]
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

$$Y = T\mu + ZU + E$$

where

- $Y$: profiles,
- $T$: segments ($unknown \rightarrow to\ estimate$),
- $\mu$: mean signal in each segment ($unknown \rightarrow to\ estimate$),
- $Z$: design matrix of the random effect,
- $U$: vector of random effect ($unobserved$): $\sim \mathcal{N}(0, G)$ ($G$ unknown $\rightarrow to\ estimate$),
- $E$: residual (unobserved): $\sim \mathcal{N}(0, R)$ ($R$ diagonal, unknown $\rightarrow to\ estimate$).
Estimation of the parameters

Direct maximisation of the likelihood. The marginal distribution of \( Y \) is

\[ Y \sim \mathcal{N}(X\theta + T\mu, V), \quad \text{where} \quad V = ZGZ' + R. \]

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

Dynamic programming can not be used to estimate \( T \) and \( \mu \).

E-M strategy. Its conditional distribution given \( U \) is

\[ (Y \mid U) \sim \mathcal{N}(X\theta + T\mu + ZU, R). \]

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

Dynamic programming can be used to estimate \( T \) and \( \mu \).
A DP-EM algorithm

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

\[ \hat{E}(U|Y), \quad \hat{V}(U|Y). \]

**M step.** Denoting \( \hat{U} = \hat{E}(U|Y) \), perform the segmentation as follows:

\[ \hat{T}\mu = \arg \min_{T\mu} \|Y - T\mu - Z\hat{U}\|^2. \]

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. \(t \in I_k\)

\[ Y_{it} = \mu + U_i + b_k t + E_{it}. \]

Estimates.

\[ \hat{b}_1 = 1.8 \times 10^{-3}, \]
\[ \hat{b}_2 = 2.5 \times 10^{-2}, \]
\[ \hat{\gamma} = 2.0, \quad \hat{\sigma} = 0.51. \]
CGH profile: Bladder cancer data

Global analysis (Inst. Curie, F. Radvanyi)

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

→ Poor probe affinity?
→ Wrong annotation?
→ Polymorphism?

The mean profile of the whole set of patients can be corrected from the probe effect:
(···) mean of raw profiles,
(○) 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 (–).