Linear models for the joint analysis of multiple array-CGH profiles

F. Picard[⋆], E. Lebarbier[†], B. Thiam [†], S. Robin[†].

[⋆] UMR 5558 CNRS Univ. Lyon 1, Lyon

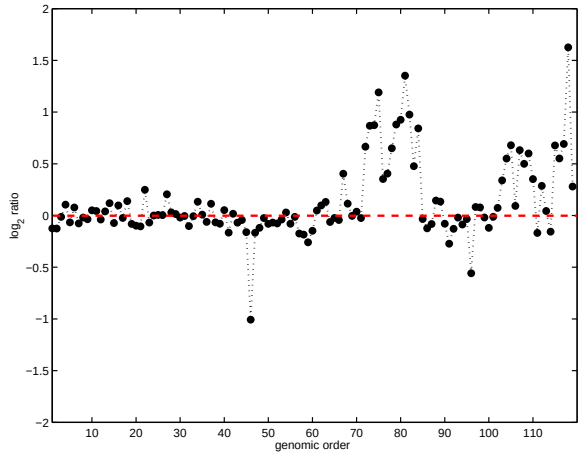
[†] UMR 518 AgroParisTech/INRA, F-75231, Paris

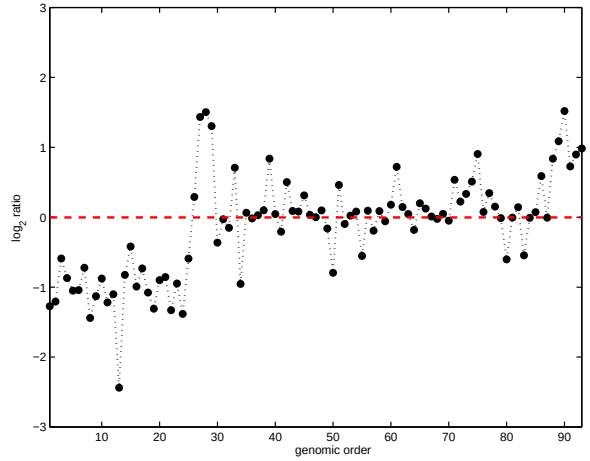
Statistics for Systems Biology (SSB) group

http://genome.jouy.inra.fr/ssb/

First years of array CGH data analysis

- **First papers**: (2002) Olshen et al., (2004) Fridlyand et al., Hupé et al., (2005) Picard et al.
- **Motivations**: find breakpoints, assign a status to segments
- **Frameworks**: segmentation, HMMs, smoothing.
- **Algorithms**: iterative split, EM, Dynamic Programming
- **Refinements**: continuous time HMMs, Bayesian segmentation, ...





Using array CGH in the clinical context

- Early motivations of array CGH experiments was to study possible associations between submicroscopic chromosomal aberrations and tumor progression or patient outcome
- Previous studies have shown that:

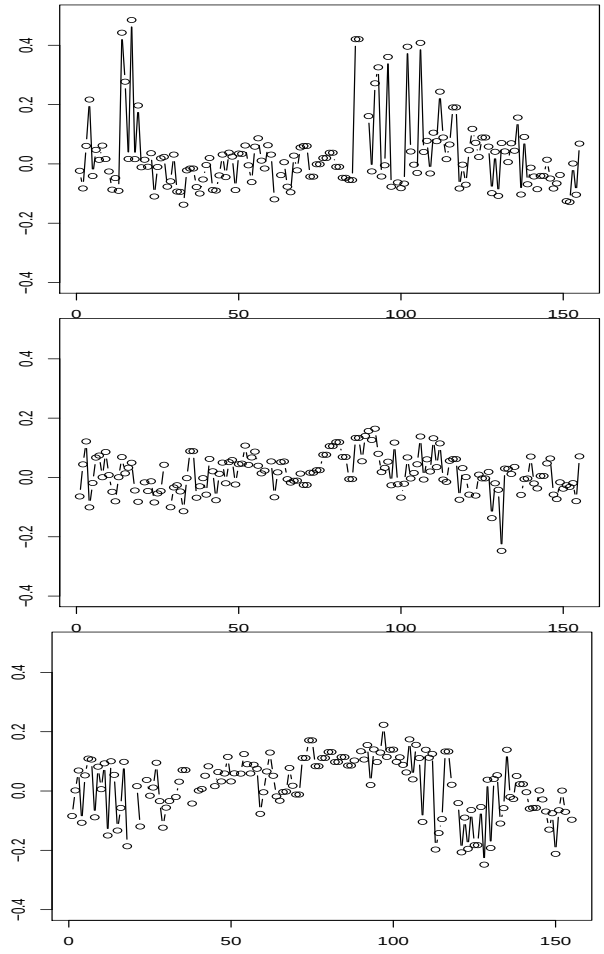
(i) Clustering analysis reveals that tumor type specific copy number patterns exist and can be used for efficient classification

(ii) chromosomal regions have been shown to be associated with overall survival of patients

(iii) genomic aberrations have been linked to differential response to various cancer therapies.

Towards multiple sample analysis

- Suppose that the cohort is made of individuals with homogeneous diagnosis
- The purpose is the joint characterization of their CGH profiles
- Broad diversity of genomic imbalances (even for patients with homogeneous diagnosis)



Lack of unified methodology to analyze multiple CGH experiments

- New challenge

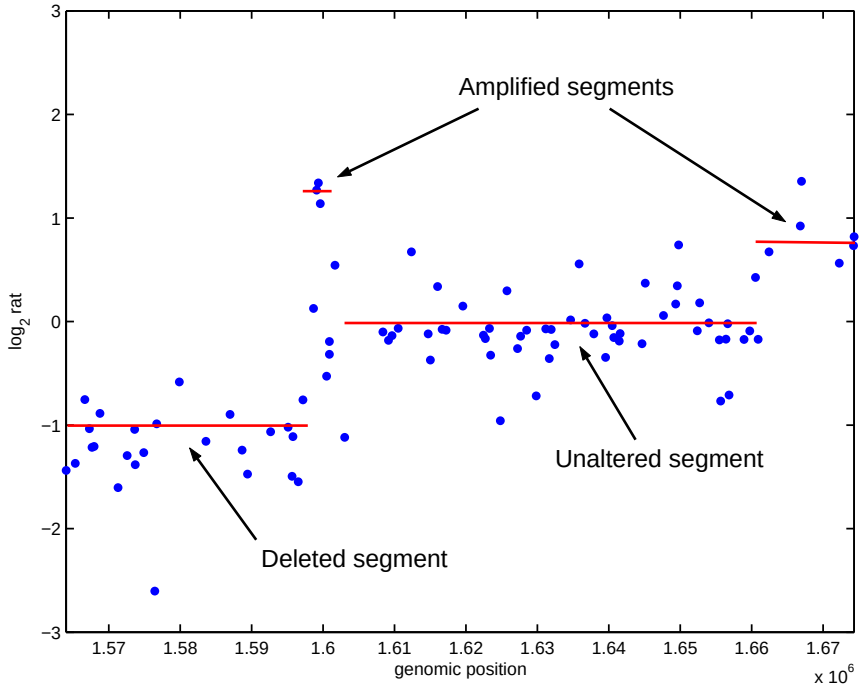
(i) Normalize the data: spatial analysis, pop-loess,

(ii) Integrate experimental design informations in the model (familial, clinical informations)

(iii) Determine chromosomal aberrations at the cohort level ?

- Linear models are proposed to do $(i) - (ii) - (iii)$ in an unified way.

Interpreting a CGH profile



One dot on the graph represents

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

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

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

Linear models for the joint segmentation of multivariate signals

- We now observe Y_t^m, the signal for patient m at position t with $m = 1, \dots M$, such that $Y_t^m \sim \mathcal{N}(\mu_t^m, \sigma^2)$
- The mean of Y_t^m is still subject to changes:

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

- We use the matricial formulation such that:

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

- $\boldsymbol{\mu}$ corresponds to the set of parameters subject to changes,
- $\mathbf{T}$ corresponds to the set of breakpoint positions,
- Pure Partial Structural Change model (Bai and Perron 2003).

Adding covariates in partial structural change models

- There are effects which concern the samples but which are not subject to changes
- How to integrate the experimental design in the global analysis ?

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

- We use the matricial formulation such that:

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

- Partial structural change model
- But some effects may not concern the expectation of the signal only

Introducing a random effect to account for dependencies

- The blind segmentation of multiple profiles is not the purpose of a joint analysis.
- We introduce some correlation of the profiles at every instants with a positional random effect:

$$\forall t \in I_k^m, \quad Y_t^m = \mu_k^m + x_t^m \theta + U_t + E_t^m,$$

- This allows us to model $\text{cov}(Y_t^m, Y_{t'}^m) = \sigma_u^2$. The positional random effect captures what is common across samples.
- With the matricial formulation: $\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}\mathbf{U} + \mathbf{E}$.
- Other random effects can be introduced, such as pedigree information
- Considering mixed model completely changes the estimation framework

Summary ofthe complete model

Notation	Interpretation	Estimation algorithm
$\mathbf{X}$	Design matrix of constant parameters	-
$\boldsymbol{\theta}$	Constant parameters	Least-Squares
$\mathbf{T}$	Breakpoint positions	Dynamic Programming
$\boldsymbol{\mu}$	parameters subject to changes	Dynamic Programming
$\mathbf{Z}$	Design matrix of random effects	-
$\mathbf{U} \sim \mathcal{N}(0, \mathbf{G})$	Random Effects	EM algorithm
$\mathbf{E} \sim \mathcal{N}(0, \mathbf{R})$	Error	Least Squares

A new computational issue for estimating breakpoint coordinates

- In the case of pure structural changes: $\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{E}$
- The purpose is to minimize the RSS:

$$\begin{aligned} RSS_K(\boldsymbol{\mu}, \mathbf{T}) = \|\mathbf{Y} - \mathbf{T}\boldsymbol{\mu}\|^2 &= \sum_{m=1}^{M} \sum_{k=1}^{K_m} RSS_k^m(\boldsymbol{\mu}_m, \mathbf{T}_m) \\ &= \sum_{m=1}^{M} \sum_{k=1}^{K_m} \sum_{t \in I_k^m} (y_{mt} - \mu_{km})^2, \end{aligned}$$

- But there is a constraint : $\sum_m K_m = K$, thus:

$$\min_{\{\mathbf{T}, \boldsymbol{\mu}\}} RSS_K(\mathbf{T}, \boldsymbol{\mu}) = \min_{K_1 + \ldots + K_M = K} \left\{ \sum_{m=1}^{M} \min_{\mathbf{T}_m, \boldsymbol{\mu}_m} RSS_{K_m}^m(\mathbf{T}_m, \boldsymbol{\mu}_m) \right\}$$

A two-stage Dynamic Programming procedure

- The blind application of DP to the multiple segmentation would lead to a procedure with complexity $\mathcal{O}(n^2 M^2)$
- Stage 1 : optimization of individual $RSS^m_{Km}(\mathbf{T}_m, \boldsymbol{\mu}_m)$ for each patient

$$\forall m \in [1, M] \;\; \{\hat{\mathbf{T}}_m, \hat{\boldsymbol{\mu}}_m\} = \min_{\mathbf{T}_m, \boldsymbol{\mu}_m} RSS^m_{Km}(\mathbf{T}_m, \boldsymbol{\mu}_m).$$

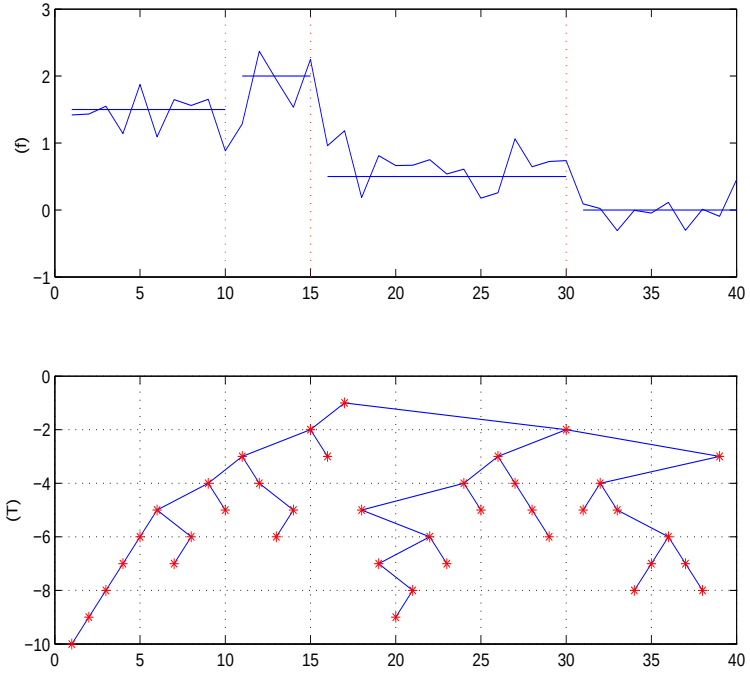
- Stage 2 : the second step consists in solving:

$$\min_{K_1 + \ldots + K_M = K} \sum_{m=1}^{M} RSS^m_{Km}(\hat{\mathbf{T}}_m, \hat{\boldsymbol{\mu}}_m).$$

- the principle of the second stage is to spread segments among M patients
- This procedure is optimal and with a complexity $\mathcal{O}(\lambda M n^2 [n + \lambda M^2])$ $(\lambda \ll 1)$

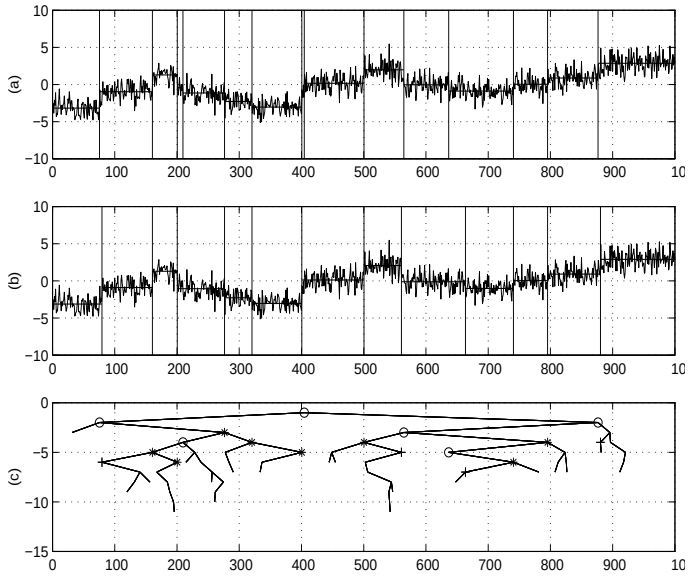
Segmenting large signals using CART - 1

- With the use of tiling arrays, the size of one signal is huge : $n \sim 10^4$
- the complexity of DP $\mathcal{O}(n^2)$
- DP performs an exhaustive search, whereas some configurations may not be relevant
- Reduce the number of configurations, and perform the exhaustive search on relevant ones only



Segmenting large signals using CART - 2

- Hybrid algorithm: Gey & Lebarbier
  1. apply CART to give some potential configurations $\mathcal{O}(n \log(n))$. (no test sample, model selection is used instead of CV).
  2. perform the exhaustive search using the obtained candidates
- simulations show that the performance of the CART-based approaches are close to the performance of the exhaustive search



top : CART solution, middle : DP solution

bottom : $\circ$ removed, $*$ kept, $+$ added breaks

Mixed models and the EM algorithm

- Mixed models can be viewed as models with incomplete data whose parameters can be estimated using the EM algorithm.
- Parameters are $\phi = (\boldsymbol{\mu}, \boldsymbol{\theta}, \mathbf{G}, \mathbf{R}, \mathbf{T})$ and the complete-data likelihood is such that:

$$\log \mathcal{L}(\mathbf{Y}, \mathbf{U}; \phi) = \log \mathcal{L}(\mathbf{Y}|\mathbf{U}; \boldsymbol{\theta}, \mathbf{T}, \boldsymbol{\mu}, \mathbf{R}) + \log \mathcal{L}(\mathbf{U}; \mathbf{G})$$

- Consequently taking the conditional expectation of this likelihood cond. to $\mathbf{Y}$ is equivalent with calculating the BLUP of $\mathbf{U}$, $\widehat{\mathbf{U}} = \mathbb{E}_{\phi}\{\mathbf{U}|\mathbf{Y}\}$
- This solves the E-step part.

CM steps : conditional maximization steps

- The maximization step is broken down into simpler conditional maximization steps
- Estimation of $\boldsymbol{\theta}$ with the classical least-squares estimator

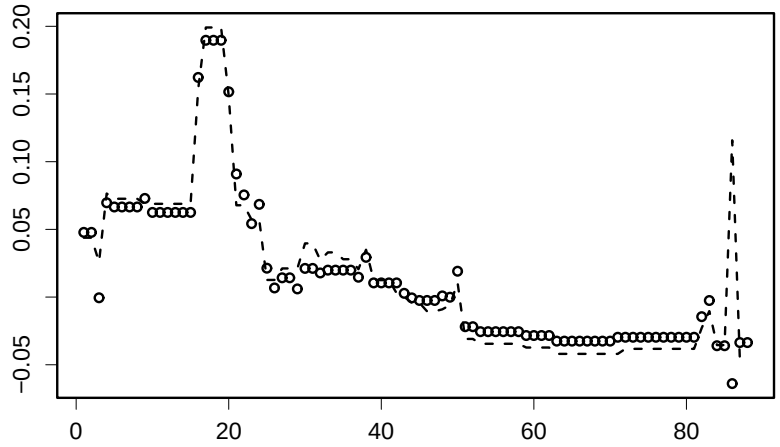
$$\mathbf{X}'\mathbf{R}^{(h)-1}\mathbf{X}\boldsymbol{\theta}^{(h+1)} = \mathbf{X}'\mathbf{R}^{(h)-1}(\mathbf{Y} - \mathbf{T}^{(h)}\boldsymbol{\mu}^{(h)} - \mathbf{Z}\widehat{\mathbf{U}}^{(h+1)}).$$

- Estimation of Variance components $\mathbf{G}^{(h+1)}$ and $\mathbf{R}^{(h+1)}$ (classical maximization)
- Estimation of $\mathbf{T}$: the computation of this particular CM-step is equivalent to the minimization of the residual sum of squares:

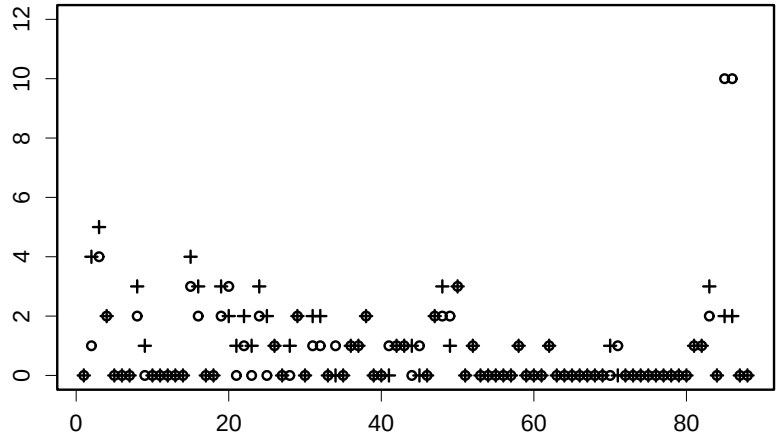
$$RSS_K(\boldsymbol{\mu}, \mathbf{T}) \quad = \quad \|\mathbf{Y} - \mathbf{X}\boldsymbol{\theta}^{(h+1)} - \mathbf{T}\boldsymbol{\mu} - \mathbf{Z}\widehat{\mathbf{U}}^{(h+1)}\|^2_{\mathbf{R}^{(h+1)-1}},$$

- This step is performed using the double-stage Dynamic Programming procedure.

Results on 57 bladder tumors

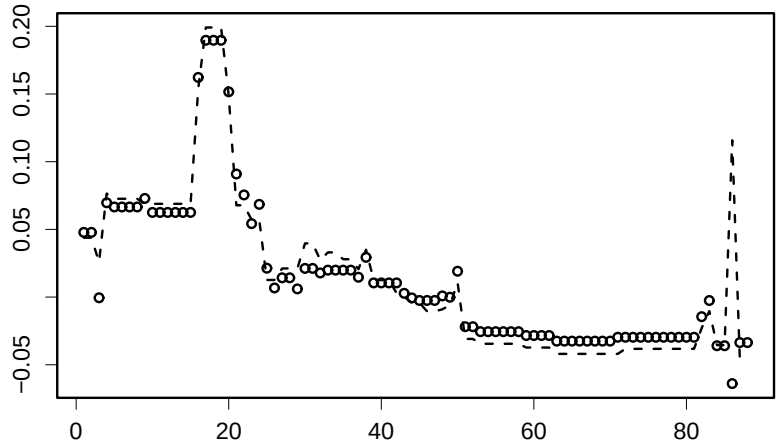


mean segmented profiles (Dotted line: without the random effect. o: with the random effect)

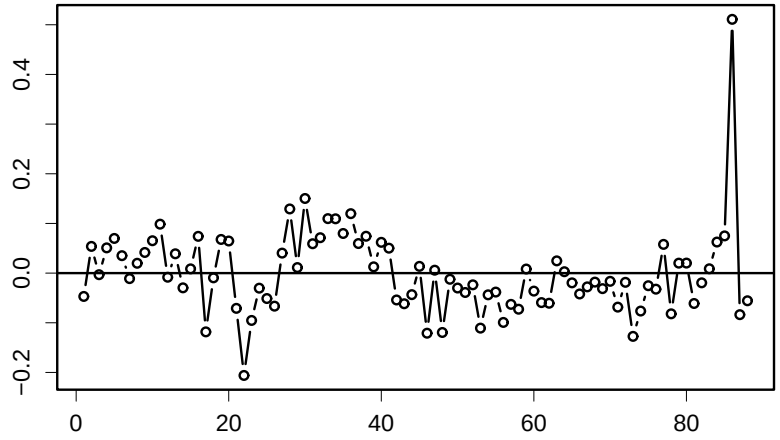


number of patients having a breakpoint at each position with (+) or without (○) random effect

Results on 57 bladder tumors



mean segmented profiles (Dotted line: without the random effect. o: with the random effect)



Predicted random effect

Interpretation

- 10 breakpoints detected without the random effect at position 85 vanish with the mixed model
- The prediction of the random effect at this particular position is very large
- Interestingly, this position is known to be subject to polymorphism, meaning that it is altered for many profiles of the cohort
- This suggests that the random effect reveals some intrinsic characteristics of the sequence at a given position or of the position on the slide on which the concerned genomic sequence is spotted (systematic technical ,artefact),
- whereas the segmentation part of the model $\mathbf{T}\boldsymbol{\mu}$ reveals the biological information specific to each profile.

Perspectives - further developments

- The next step consists in generalizing the Segmentation/Clustering framework to the multivariate case:

$$\mathbf{Y} = \mathbf{TC}\boldsymbol{\mu} + \mathbf{X}\boldsymbol{\theta} + \mathbf{ZU} + \mathbf{E}.$$

- $\mathbf{C}$ is a classification matrix which constraints the levels for every profiles. There is an underlying random label variable $\mathbf{S}$ which is multinomial.
- Consequently, the estimation procedure will be more difficult since $\mathbf{Y}|\mathbf{U}, \mathbf{S}$ is not Gaussian anymore
- A possibility would be to consider the positional effect as being fixed (which gives the same results in practice)
- We are currently developping an R package to perform multiple sample analysis.

Preprints and documents

- Linear Models for segmentation: *Joint segmentation of multivariate Gaussian processes using mixed linear models.* F. Picard, E. Lebarbier, E. Budinska and S. Robin
- CART for large samples: *Using CART to detect multiple change points in the mean for large samples.* S. Gey and E. Lebarbier
- Reccurrent aberrations: *Simultaneous occurrences of runs in independent Markov chains.* S. Robin and V. Stefanov
- all documents at `http://genome.jouy.inra.fr/ssb/preprint/`