A segmentation-clustering problem for the analysis of array CGH data

F. Picard, S. Robin, E. Lebarbier, J-J. Daudin

UMR INA P-G / ENGREF / INRA MIA 518

APPLIED STOCHASTIC MODELS AND DATA ANALYSIS
Brest May 2004

Microarray CGH technology

- Known effects of big size chromosomal aberrations (ex: trisomy).

  $\rightarrow$experimental tool: **Karyotype** (Resolution $\sim$ chromosome).

- Change of scale: what are the effects of small size DNA sequences deletions/amplifications?

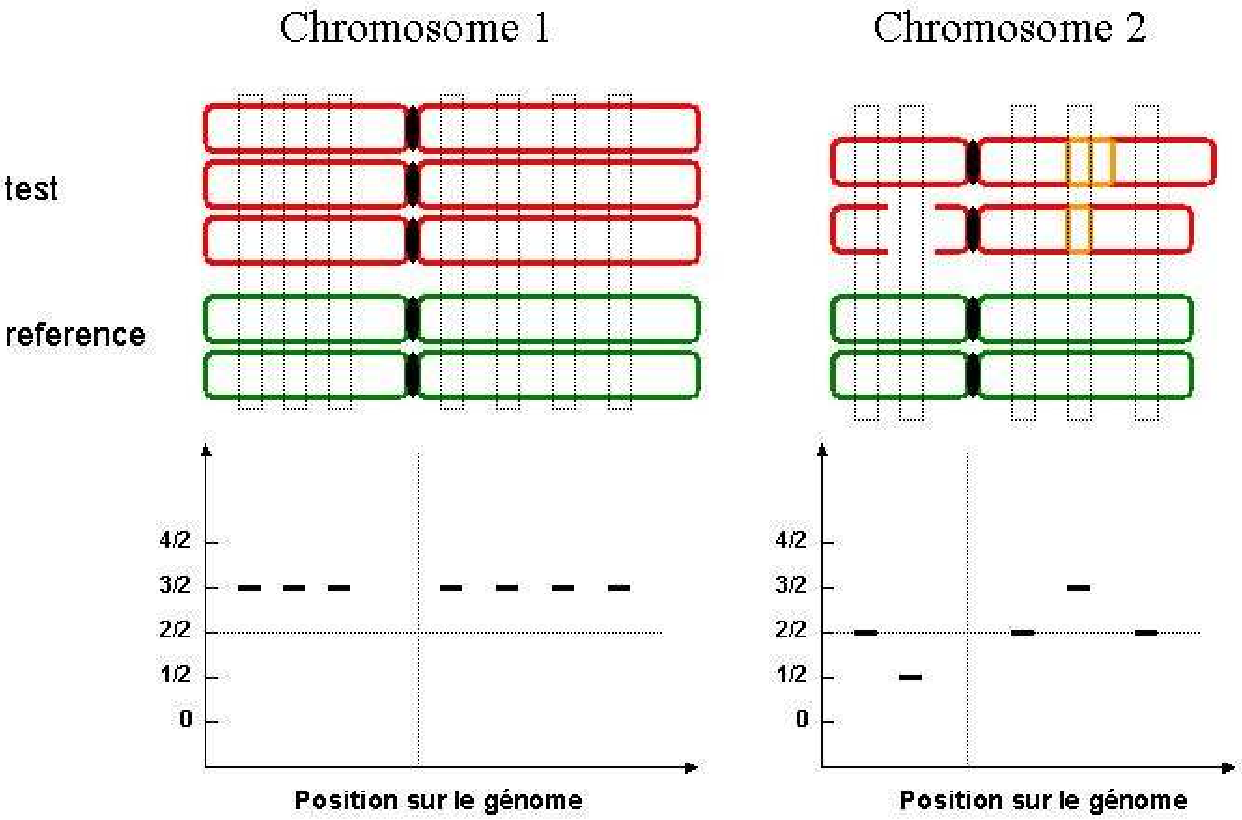
  $\rightarrow$ experimental tool: **"conventional" CGH** (resolution $\sim$ 10Mb).

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

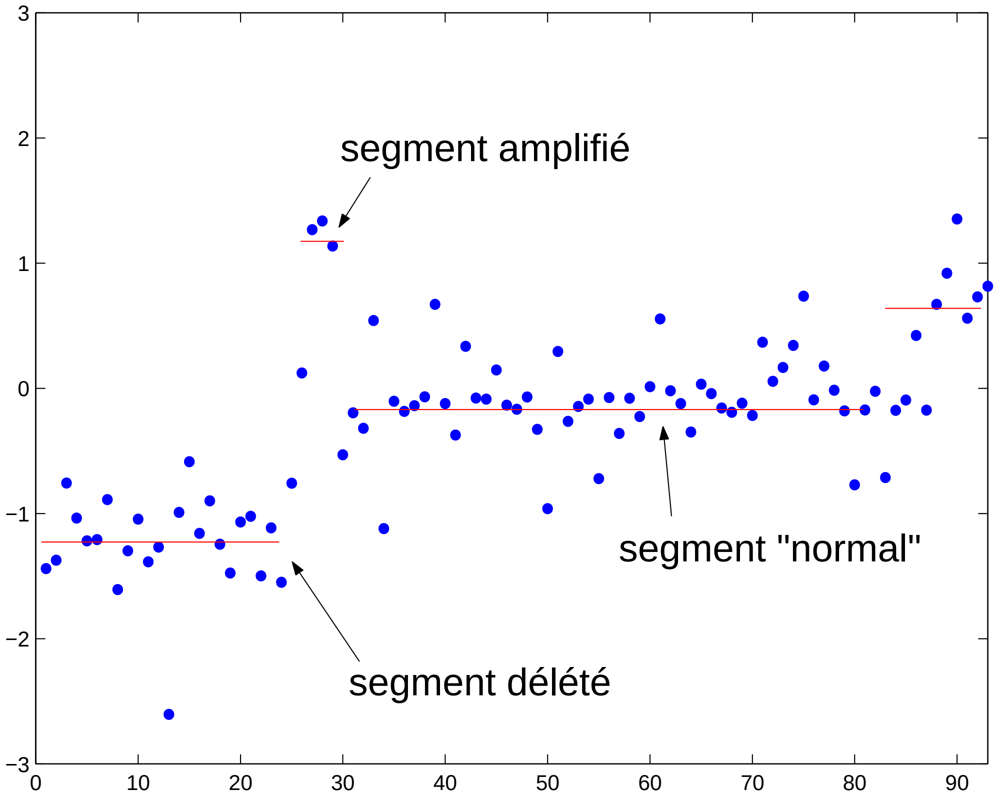
  $\rightarrow$ Application of the **microarray** technology to CGH : 1997.

  $\rightarrow$last generation of chips: resolution $\sim$ 100kb.

Microarray technology in its principle



Interpretation of a CGH profile



A dot on the graph represents

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

First step of the statistical analysis

Break-points detection in a gaussian signal

- $Y = (Y_1, ..., Y_n)$ a random process such that $Y_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$.
- Suppose that the parameters of the distribution of the $Y s$ are affected by K-1 abrupt-changes at unknown coordinates $T = (t_1, ..., t_{K-1})$.
- Those break-points define a partition of the data into K segments of size n_k:

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

- Suppose that those parameters are constant between two changes:

$$\forall t \in I_k,\ Y_t \sim \mathcal{N}(\mu_k, \sigma_k^2).$$

- The parameters of this model are :

$$\begin{aligned} T &= (t_1, ..., t_{K-1}), \\ \Theta &= (\theta_1, \ldots, \theta_K), \theta_k = (\mu_k, \sigma_k^2). \end{aligned}$$

- Break-points detection aims at studying the **spatial structure of the signal**.

Estimating the parameters in a model of abrupt-changes detection

Log-Likelihood

$$\mathcal{L}_K(T,\Theta) = \sum_{k=1}^{K} \log f(y^k;\theta_k) = \sum_{k=1}^{K}\sum_{t\in I_k} \log f(y_t;\theta_k)$$

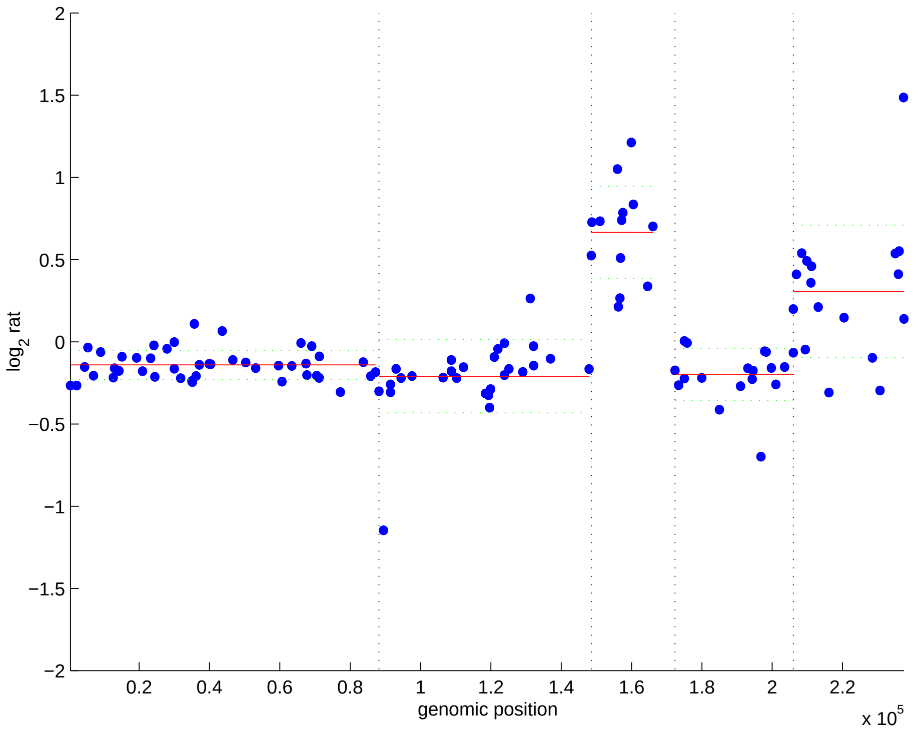
Estimating the parameters with K fixed by maximum likelihood

- Joint estimation of T and Θ with dynamic programming.
- Necessary property of the likelihood : additivity in K (sum of local likelihoods calculated on each segment).

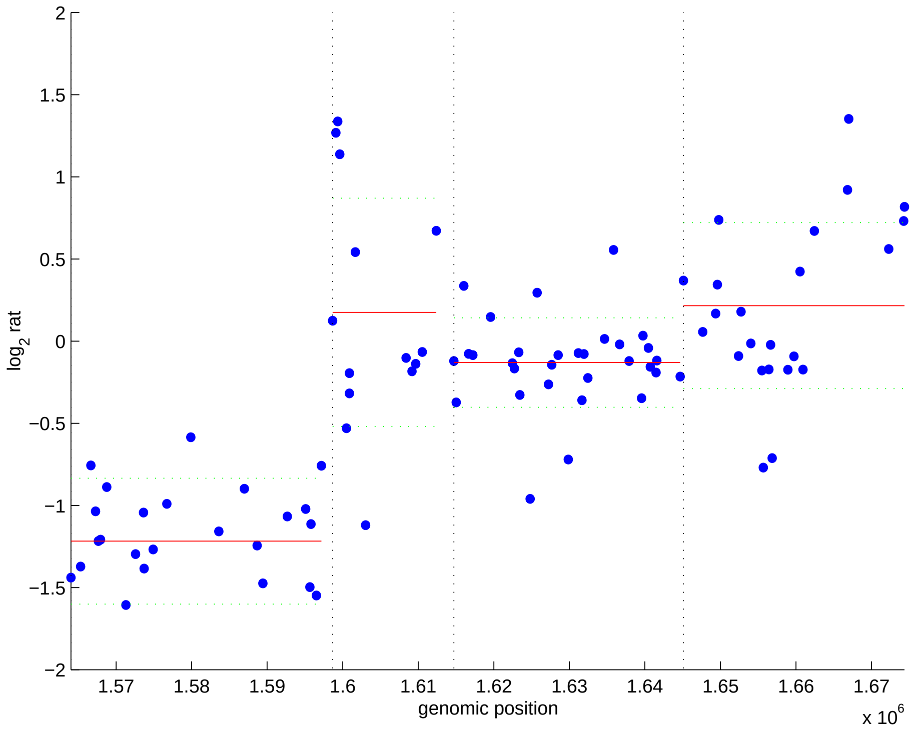
Model Selection : choice of K

- Penalized Likelihood : $\hat{K} = \underset{K}{Argmax}\left(\hat{\mathcal{L}}_K - \beta \times pen(K)\right)$.
- With $pen(K) = 2K$.
- β is adaptively estimated to the data (Lavielle(2003)).

Example of segmentation on array CGH data

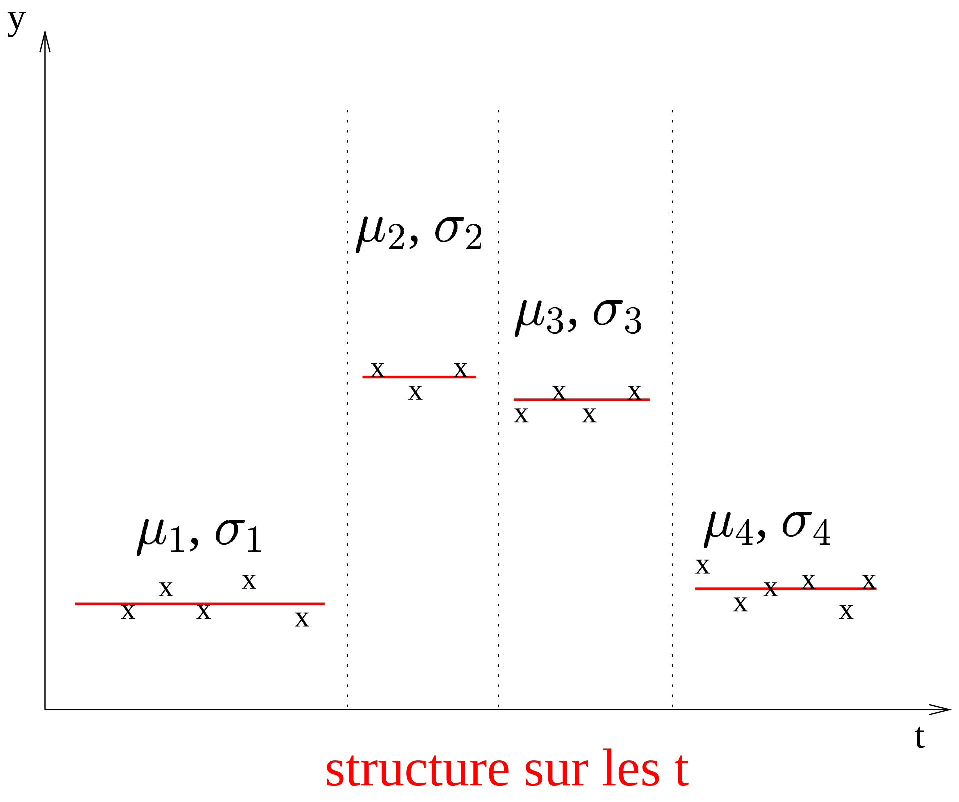


BT474 chromosome 1, $\hat{K} = 5$



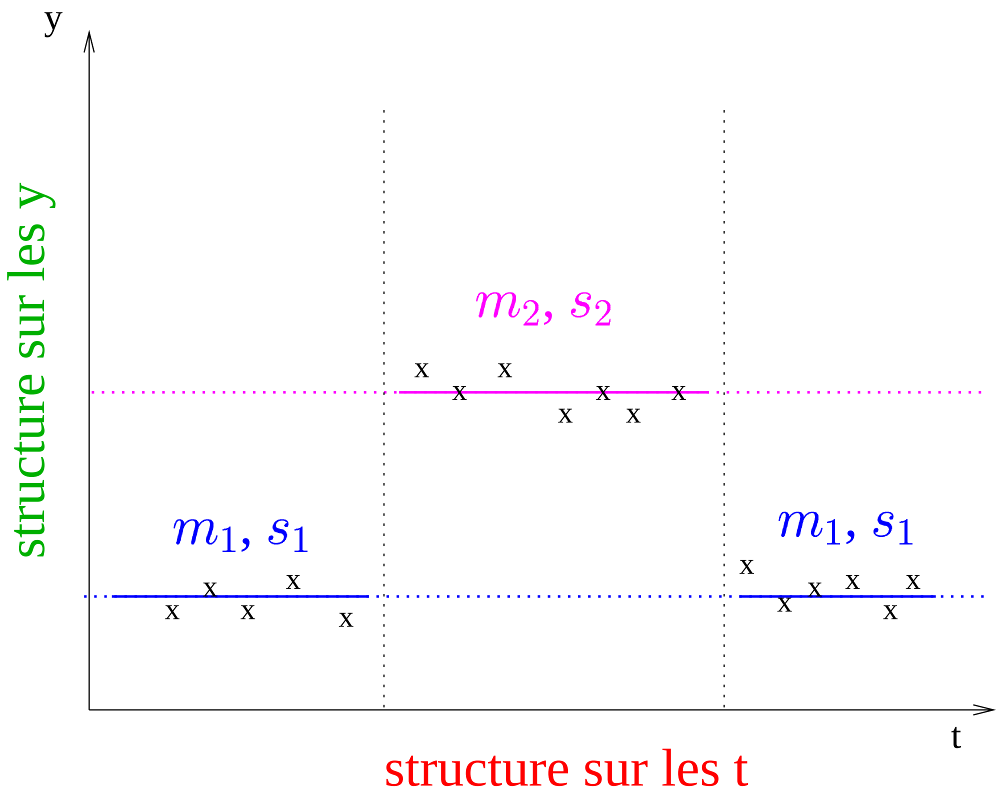
BT474 chromosome 9, $\hat{K} = 4$

Considering biologists objective and the need for a new model



Segmentation: structure spatiale du signal

$$\theta_k = (\mu_k, \sigma_k^2)$$



Segmentation/Classification

$$\theta_p = (m_p, s_p^2)$$

A new model for segmentation-clustering purposes

- We suppose there exists a **secondary underlying structure** of the segments into P populations with weights $\pi_1, ..., \pi_P (\sum_p \pi_p = 1)$.

- We introduce hidden variables, Z_{kp} indicators of the population of origin of **segment** k .

- Those variables are supposed independent, with multinomial distribution:

$$(Z_{k1}, \ldots, Z_{kP}) \sim \mathcal{M}(1; \pi_1, \ldots, \pi_P).$$

- Conditionnally to the hidden variables, we know the distribution of Y :

$$Y^k | Z_{kp} = 1 \sim \mathcal{N}(\mathbb{1}_{n_k} m_p, s_p^2 I_{n_k}).$$

- It is a model of **segmentation/clustering**.
- The parameters of this model are

$$\begin{aligned} T &= (t_1, ..., t_{K-1}), \\ \Theta &= (\pi_1, \ldots, \pi_P; \theta_1, \ldots, \theta_P), \text{ avec } \theta_p = (m_p, s_p^2). \end{aligned}$$

Likelihood and statistical units of the model

- Mixture Model of segments :

- $\star$ the statistical units are segments :Y^k,
- $\star$ the density of Y^k is a mixture density:

$$\log \mathcal{L}_{KP}(T,\Theta) = \sum_{k=1}^{K} \log f(y^k;\Theta) = \sum_{k=1}^{K} \log \left\{ \sum_{p=1}^{P} \pi_p f(y^k;\theta_p) \right\}$$

- $\star$ If the $Y_t s$ are independent, we have:

$$\log \mathcal{L}_{KP}(T,\Theta) = \sum_{k=1}^{K} \log \left\{ \sum_{p=1}^{P} \pi_p \prod_{t\in I_k} f(y_t;\theta_p) \right\}.$$

- Classical mixture model :

- $\star$ the statistical units are the Y_ts,

$$\log \mathcal{L}_{P}(\Theta) = \sum_{k=1}^{K} \log \left\{ \prod_{t\in I_k} \sum_{p=1}^{P} \pi_p f(y_t;\theta_p) \right\}$$

An hybrid algorithm for the optimization of the likelihood

Alternate parameters estimation with K and P known

1 When T is fixed, the **EM** algorithm estimates Θ:

$$\hat{\Theta}^{(\ell+1)} = \underset{\Theta}{Argmax} \left\{ \log \mathcal{L}_{KP} \left(\Theta, T^{(\ell)} \right) \right\}.$$

$$\log \mathcal{L}_{KP}(\hat{\Theta}^{(\ell+1)}; \hat{T}^{(\ell)}) \geq \log \mathcal{L}_{KP}(\hat{\Theta}^{(\ell)}; \hat{T}^{(\ell)})$$

2 When Θ is fixed, **dynamic programming** estimates T:

$$\hat{T}^{(\ell+1)} = \underset{T}{Argmax} \left\{ \log \mathcal{L}_{KP} \left(\hat{\Theta}^{(\ell+1)}, T \right) \right\}.$$

$$\log \mathcal{L}_{KP}(\hat{\Theta}^{(\ell+1)}; \hat{T}^{(\ell+1)}) \geq \log \mathcal{L}_{KP}(\hat{\Theta}^{(\ell+1)}; \hat{T}^{(\ell)})$$

An increasing sequence of likelihoods:

$$\log \mathcal{L}_{KP}(\hat{\Theta}^{(\ell+1)}; \hat{T}^{(\ell+1)}) \geq \log \mathcal{L}_{KP}(\hat{\Theta}^{(\ell)}; \hat{T}^{(\ell)})$$

Mixture Model when the segmentation is knwon

Mixture model parameters estimators

$$\hat{\tau}_{kp} = \frac{\hat{\pi}_p f(y^k; \hat{\theta}_p)}{\sum_{\ell=1}^{P} \hat{\pi}_\ell f(y^k; \hat{\theta}_\ell)}.$$

- the estimator the the mixing proportions is: $\hat{\pi}_p = \frac{\sum_k \hat{\tau}_{kp}}{K}$.
- In the gaussian case, $\theta_p = (m_p, s_p^2)$:

$$\begin{aligned} \hat{m}_p &= \frac{\sum_k \hat{\tau}_{kp} \sum_{t \in I_k} y_t}{\sum_k \hat{\tau}_{kp} n_k}, \\ \hat{s}_p^2 &= \frac{\sum_k \hat{\tau}_{kp} \sum_{t \in I_k} (y_t - \hat{m}_p)^2}{\sum_k \hat{\tau}_{kp} n_k}. \end{aligned}$$

- Big size vectors will have a bigger impact in the estimation of the parameters, via the term $\sum_k \hat{\tau}_{kp} n_k$

Influence of the vectors size on the affectation (MAP)

- The density of Y^k can be written as follows:

$$f(y^k;\theta_p) = \exp\left\{-\frac{n_k}{2}\left(\log(2\pi s_p^2) + \frac{1}{s_p^2}\left[(\overline{y_k^2} - \bar{y}_k^2) + (\bar{y}_k - m_p)^2\right]\right)\right\}$$

$\star$ $(\bar{y}_k - m_p)^2$: distance of the mean of vector k to population p

$\star$ $(\overline{y_k^2} - \bar{y}_k^2)$: intra-vector k variability

- Big size Individuals will be affected with certitude to the closest population

$$\begin{array}{rcl|rcl} \lim_{n_k \to \infty} \tau_{kp_0} & = & 1 & \lim_{n_k \to \infty} \tau_{kp} & = & 0 \\ \lim_{n_k \to 0} \tau_{kp_0} & = & \pi_{p_0} & \lim_{n_k \to 0} \tau_{kp} & = & \pi_p \end{array}$$

Segmentation with a fixed mixture

Back to dynamic programming

- the incomplete mixture log-likelihood can be written as a sum of local log-likelihoods:

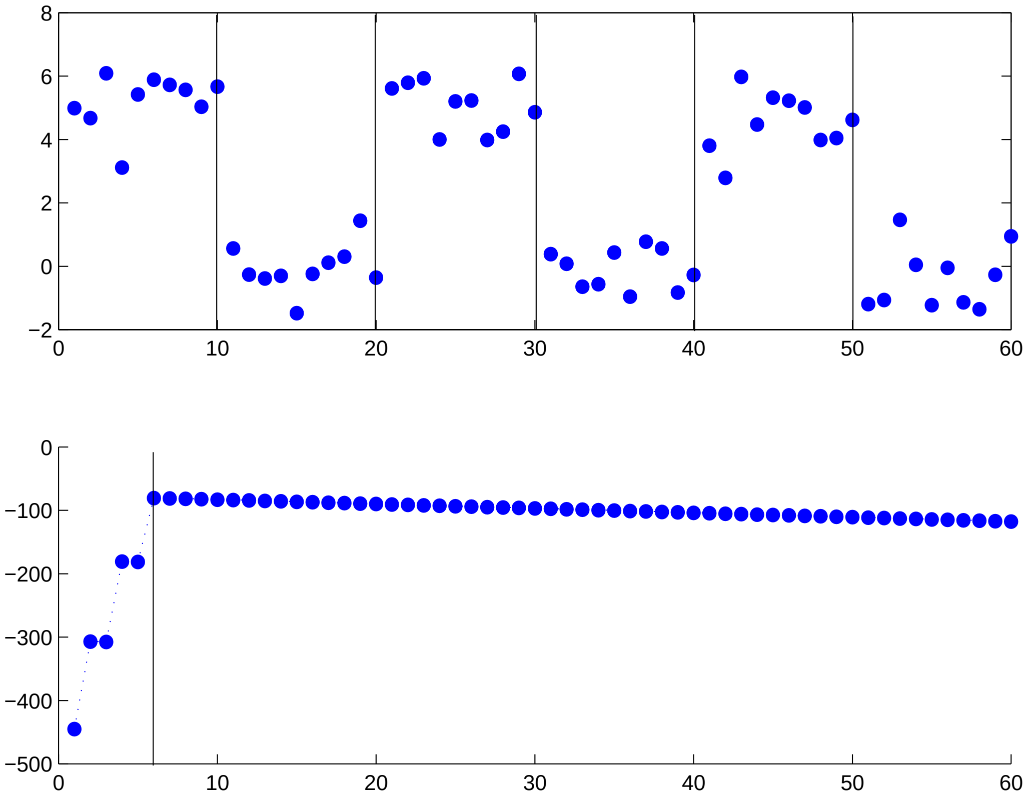
$$\mathcal{L}_{KP}(T,\Theta) \;=\; \sum_k \ell_{kP}(y^k;\Theta)$$

- the local log-likelihood of segment k corresponds to the mixture log-density of vector Y^k

$$\ell_{kP}(y^k;\Theta) = \log\left\{\sum_{p=1}^{P}\pi_p \prod_{t\in I_k} f(y_t;\theta_p)\right\}.$$

- $\log \mathcal{L}_{KP}(T,\Theta)$ can be optimized in T with Θ fixed, by dynamix programming.

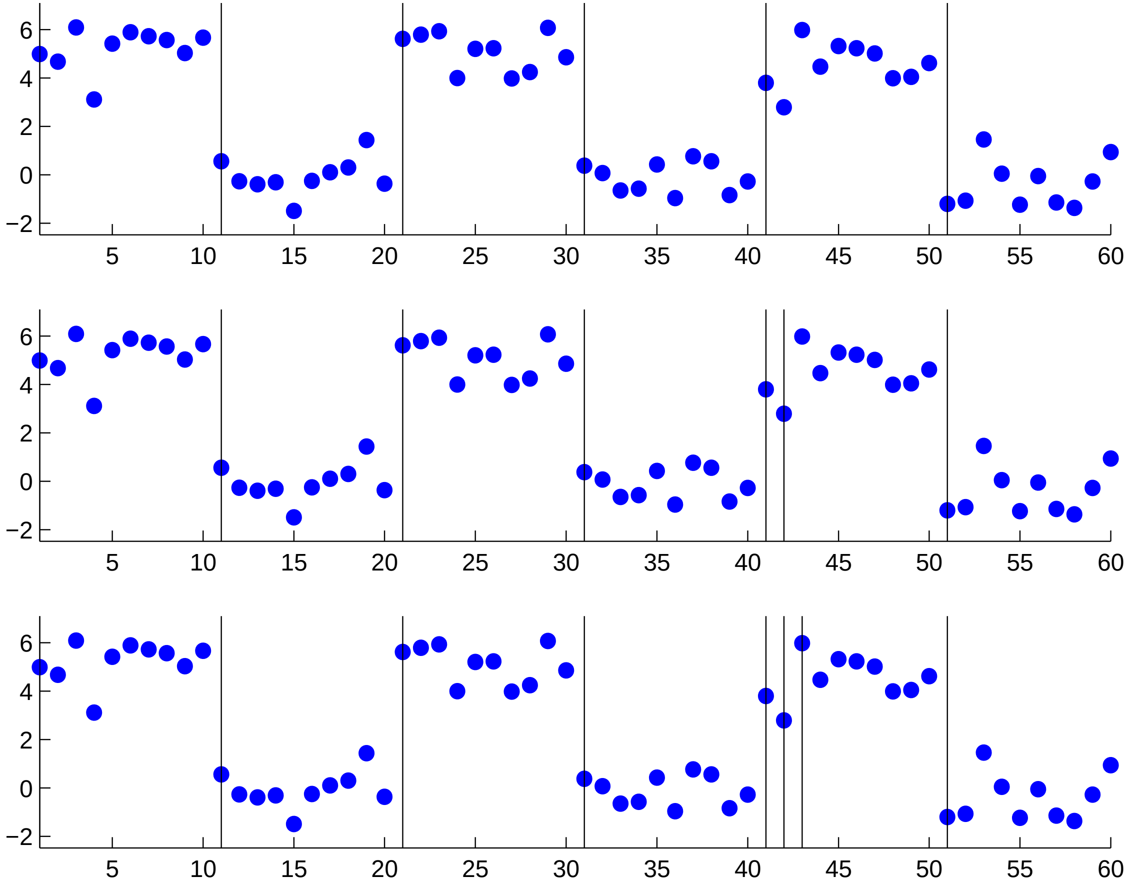
A decreasing log-Likelihood?



Evolution of the incomplete log-likelihood with respect to the number of segments.

$$f(y^k; \Theta) \;=\; 0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(5,1)$$

What is going on?



When the true number of segments is reached (6), segments are cut on the edges.

Explaining the behavior of the likelihood

Optimization of the incomplete likelihood with dynamic programming:

$$\begin{aligned}
\log \mathcal{L}_{KP}(T;\Theta) &= Q_{KP}(T;\Theta) - H_{KP}(T;\Theta) \\
Q_{KP}(T;\Theta) &= \sum_k \sum_p \tau_{kp} \log(\pi_p) + \sum_k \sum_p \tau_{kp} \log f(y^k;\theta_p) \\
H_{KP}(T;\Theta) &= \sum_k \sum_p \tau_{kp} \log \tau_{kp}
\end{aligned}$$

Hypothesis:

1 We suppose that the true number of segments is K^* and that the partitions are nested for $K \geq K^*$.

⋆ Segment Y^K is cut into (Y_1^K, Y_2^K):

$$f(Y^K;\theta_p) = f(Y_1^K;\theta_p) \times f(Y_2^K;\theta_p).$$

2 We suppose that if $Y^K \in p$ then $(Y_1^K, Y_2^K) \in p$:

$$\tau_p(Y^K) \simeq \tau_p(Y_1^K) \simeq \tau_p(Y_2^K) \simeq \tau_p.$$

An intrinsic penality

Under hypothesis 1-2:

$$\forall K \geq K^*, \log \hat{\mathcal{L}}_{(K+1),P} - \log \hat{\mathcal{L}}_{(K),P} \simeq \sum_p \hat{\pi}_p \log(\hat{\pi}_p) - \sum_p \hat{\tau}_p \log(\hat{\tau}_p) \leq 0$$

The log-likelihood is decomposed into two terms

- A term of **fit** that increases with K, and is constant from a certain K^* (nested partitions)

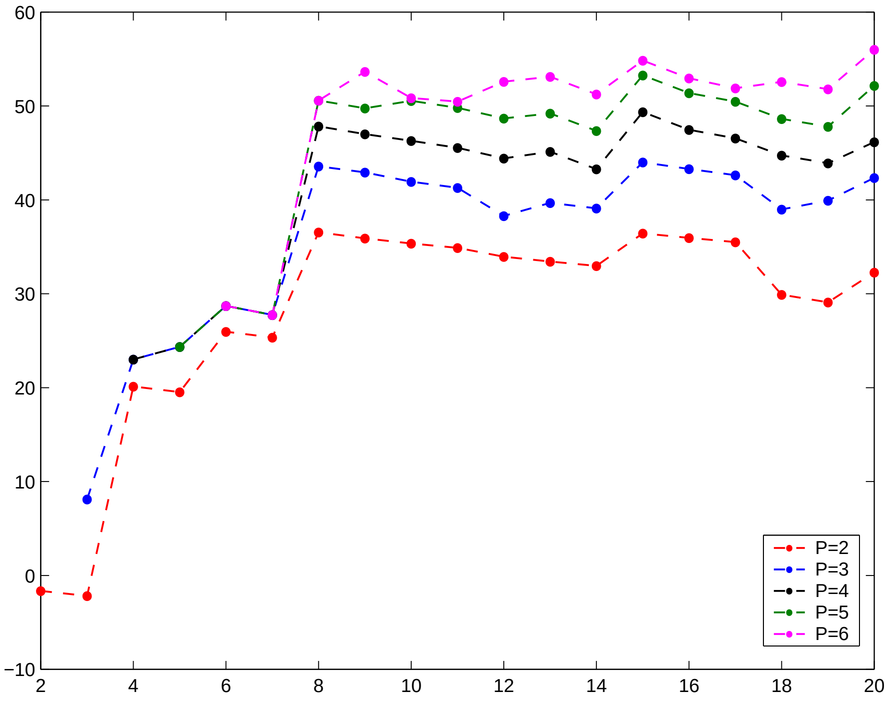
$$\sum_k \sum_p \hat{\tau}_{kp} \log f(y^k; \hat{\theta}_p).$$

- A term of **differences of entropies** that decreases with K: plays the role of penalty for the choice of K

$$K \sum_p \hat{\pi}_p \log(\hat{\pi}_p) - \sum_k \sum_p \hat{\tau}_{kp} \log \hat{\tau}_{kp}.$$

Choosing the number of segments K when P is fixed can be done with a penalized likelihood

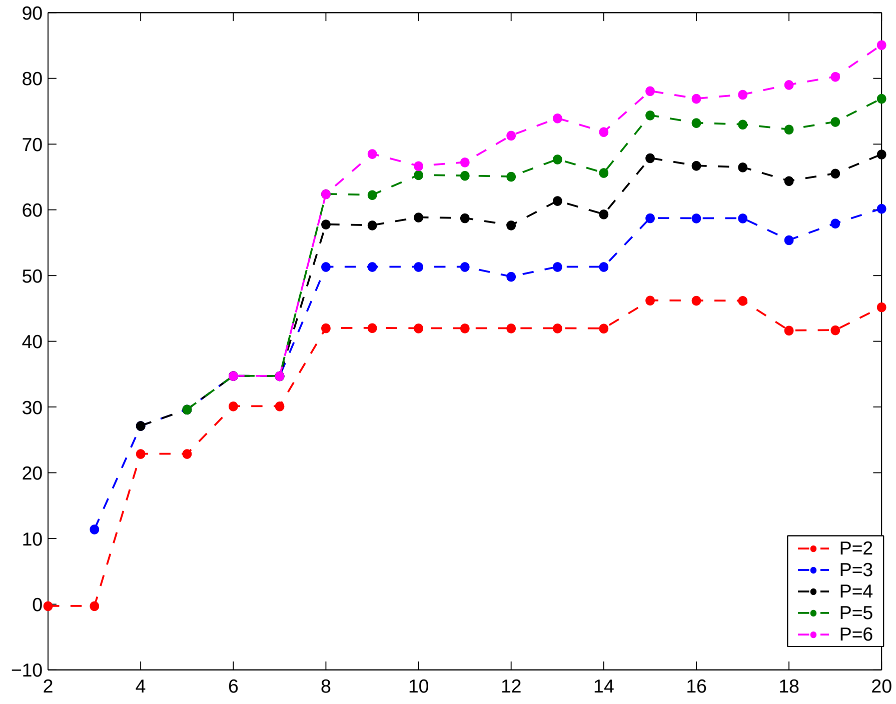
Incomplete Likelihood behavior with respect to the number of segments



The incomplete log-likelihood is decreasing from de $K = 8$

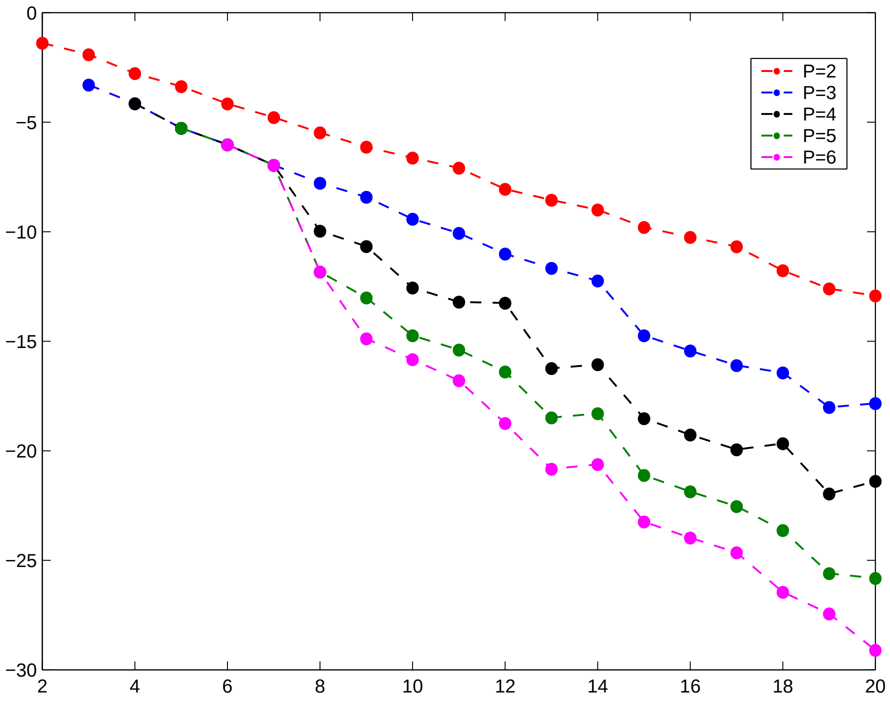
$$\hat{\mathcal{L}}_{KP}(\hat{T};\hat{\Theta}) = \sum_k \log\left\{\sum_p \hat{\pi}_p f(y^k;\hat{\theta}_p)\right\}.$$

Decomposition of the log-likelihood



term of fit

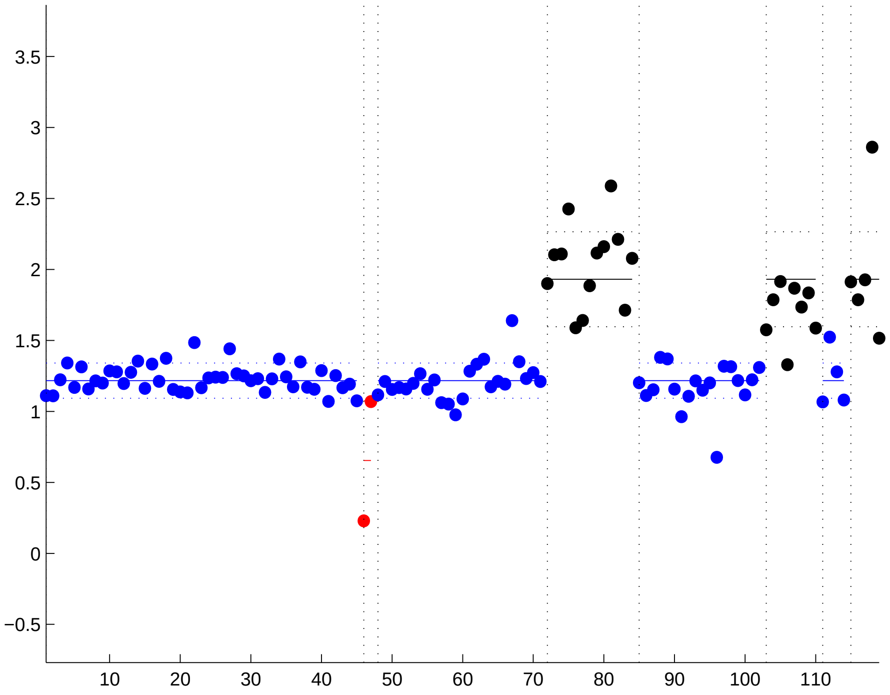
$$\sum_k \sum_p \hat{\tau}_{kp} \log f(y^k; \hat{\theta}_p)$$



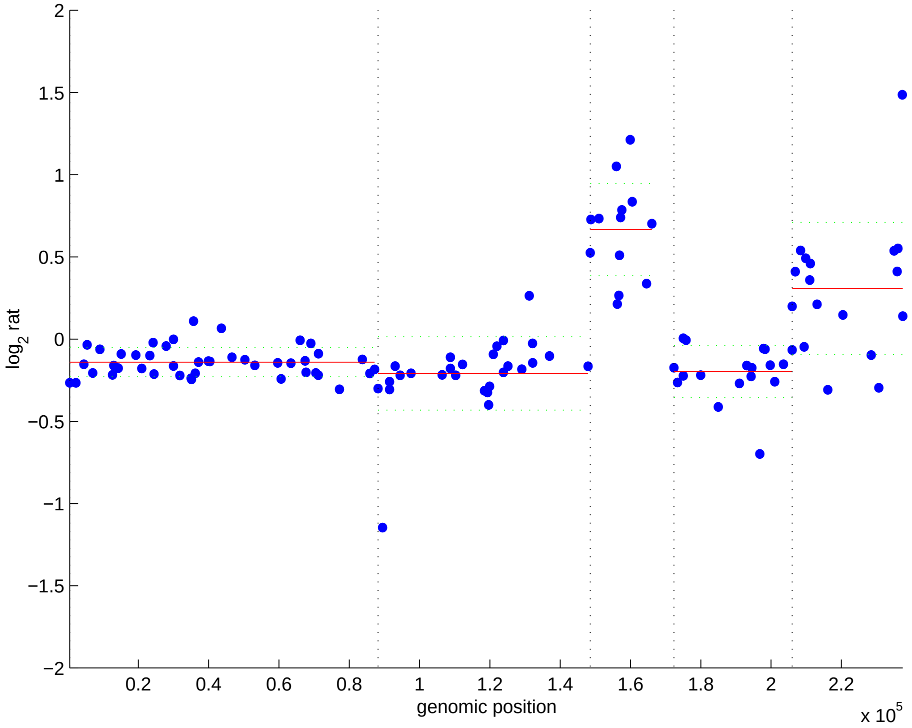
differences of entropies

$$K \sum_p \hat{\pi}_p \log(\hat{\pi}_p) - \sum_k \sum_p \hat{\tau}_{kp} \log \hat{\tau}_{kp}$$

Resulting clusters

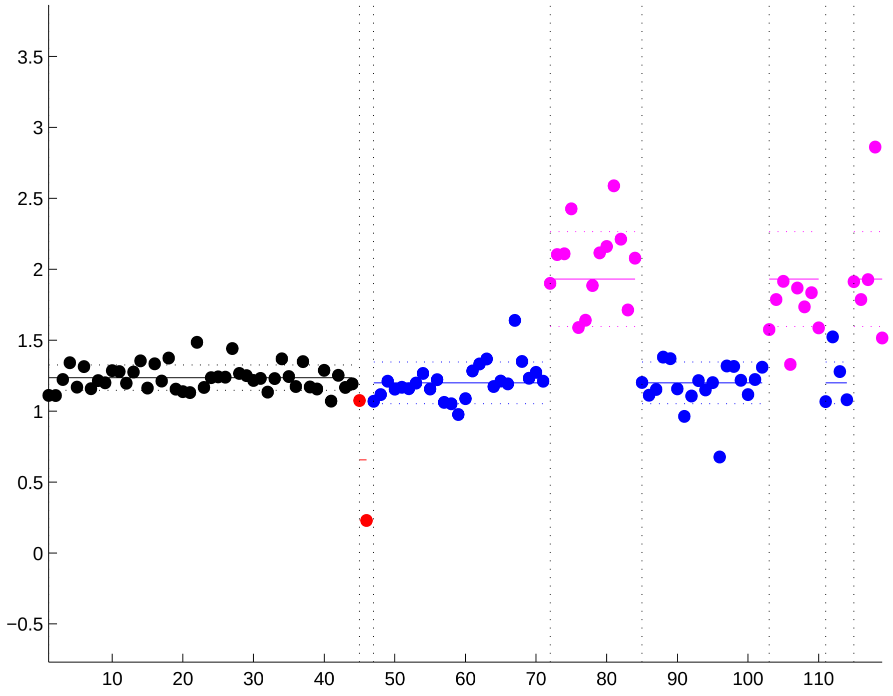


Segmentation/Clustering $P = 3$, $K = 8$

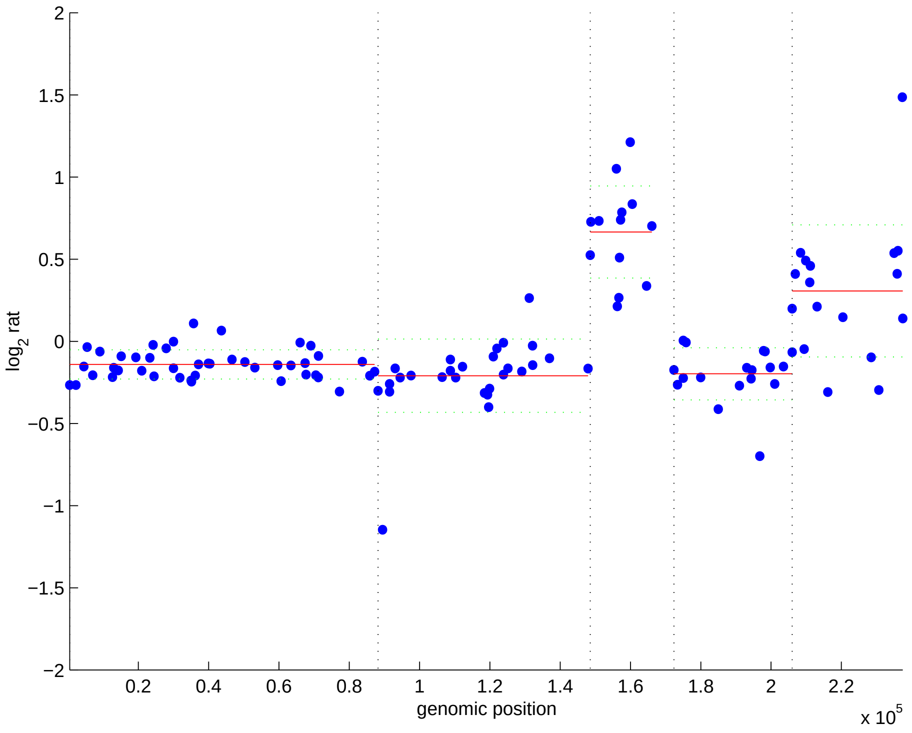


Segmentation $K = 5$

Resulting clusters

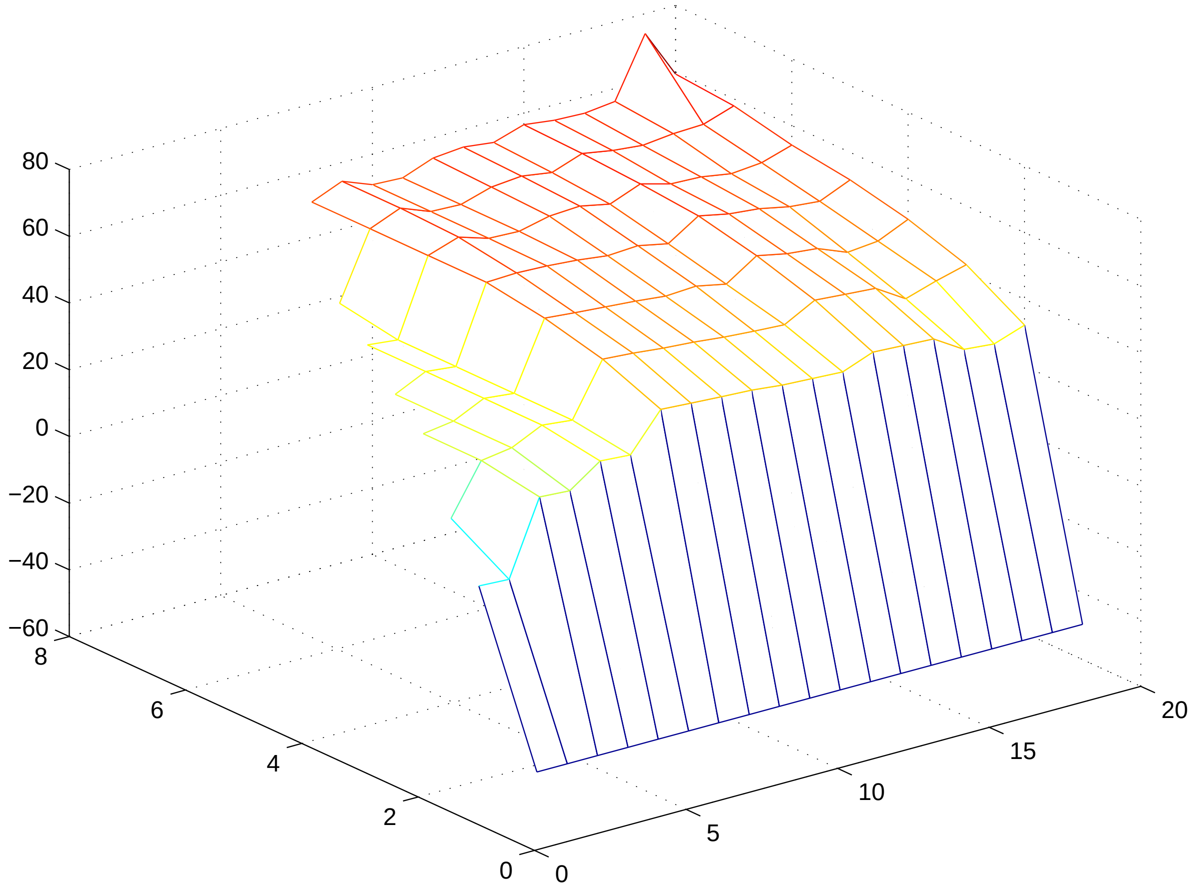


Segmentation/Clustering $P = 4$, $K = 8$



Segmentation $K = 5$

Perspective : simultaneous choice for K and P



Incomplete Log-likelihood with respect to K and P.

This is the end

Conclusions:

- Definition of a new model that considers the *a priori* knowledge we have about the biological phenomena under study.
- Development of an hybrid algorithm (EM/dynamic programming) for the parameters estimation (problems linked to EM : initializtion, local maxima, degeneracy).
- Still waiting for an other data set to assess the performance of the clustering.

Perspectives:

- Modeling :
  - ⋆ Comparison with Hidden Markov Models
- Model choice:
  - ⋆ Develop an adaptive procedure for two components.
- Other application field
  - ⋆ DNA sequences (in progress)