Exact posterior distributions over the segmentation space and model selection for multiple change-point detection problems

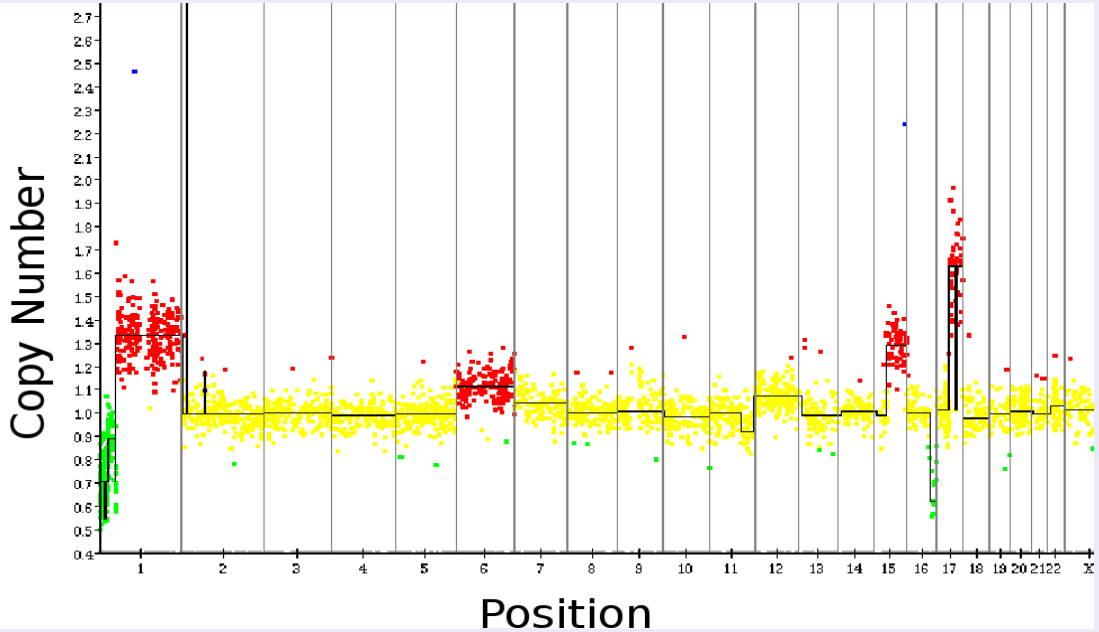
Guillem Rigaill, Emilie Lebarbier and Stéphane Robin,

August 2010

Application to DNA Copy number

DNA Copy number analysis

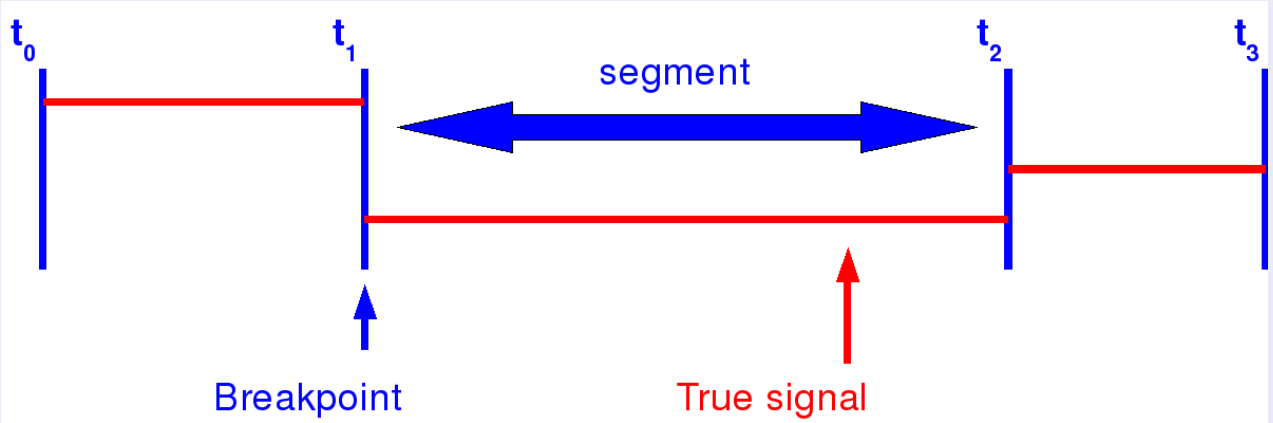
- In normal cells: copy number = 2 (pairs of chromosome)
- In tumor cells: copy number $\neq$ 2 on many points of the genome
- Gain and loss of DNA:
  - chromosomes
  - smaller regions up to 10Kb



Multiple change-point detection

The data

- The signal we observe Y_t is noisy
- The true signal is affected by abrupt changes



Segments and segmentations

$\mathcal{M}_K$ the set of all possible segmentations with K segments

$m \in \mathcal{M}_K$ a specific segmentation

$r \in m$ a segment of m with n_r observations

A model, a simple example

Normal heteroscedastic segmentation

$$\forall t \in r \qquad Y_t \sim \mathcal{N}(\mu_r, \sigma_r^2) \qquad \{Y_t\}_t \text{ are independent}$$

Parameter estimation

- Given the breakpoint positions, the estimation of other parameters is straightforward
- For example, using maximum likelihood we get:

$$\hat{\mu}_r = \frac{1}{n_r} \sum_{t \in r} Y_t$$

Estimation of breakpoint positions?

Problems

- For n points, there are 2^{n-1} possible segmentations
- Breakpoints are discrete parameters

- How to select one segmentation out of so many?
- How to explore the segmentation space?

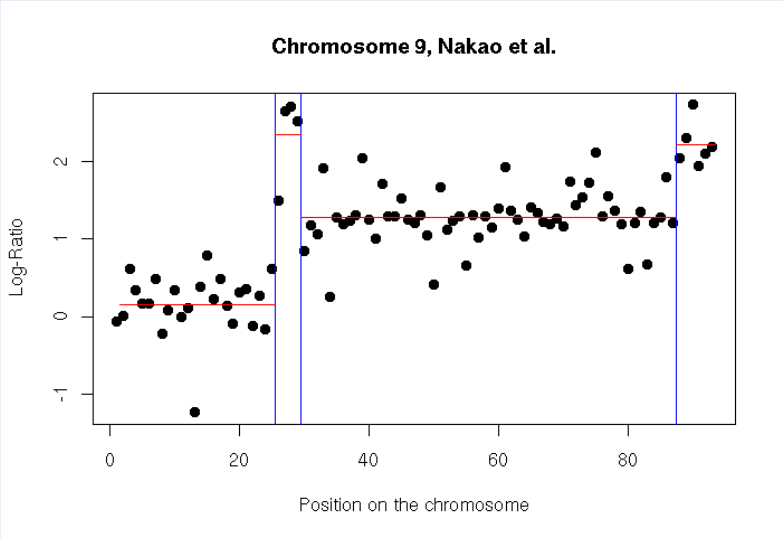
Some solutions

- Dynamic Programming (DP) to recover the optimal solution: $O(n^2)$
- Various model selection criteria:
  - The BIC criteria is not theoretically justified
  - [Zhang and Siegmund(2007)] proposed a modified BIC criteria

One example

Application to a DNA copy number profile

1. Algorithm
   - DP to recover the best segmentation in $K = 1$ up to $K = 30$ segments
2. Select K
   - with the modified BIC



Questions

- Is the optimal segmentation far better than others?
- Quality of the segment/breakpoint localizations?

Bayesian framework

Some probabilities

$P(m)$	prior distribution of segmentation m
$P(K)$	prior distribution of the number of segments
$P(Y\|\theta_m, m)$	distribution of the data given m and θ_m

Assumption: Factorisability

- If the segment are independent: $P(Y|m) = \Pi_{r \in m} P(Y^r|r)$
- $P(Y^r|r) = \int P(Y^r|\theta_r)P(\theta_r)d\theta_r$, with θ_r parameters or segment r

Computation

Quantities of interest

$P(m|Y)$ posterior probability of a segmentation m

$P(K|Y)$ posterior probability of the number of segments

$S_K(r)$ posterior probability of the segment r

$ICL(K)$ Integrated Completed Likelihood [Biernacki et al.(2000)]

$$\mathrm{ICL}(K) = -\log P(Y, K) + \mathcal{H}(K)$$

- ICL favours the K where the best segmentation is by far the best one

$\mathcal{H}(K)$ entropy: $\mathcal{H}(K) = -\sum_{m \in \mathcal{M}_K} P(m|Y, K) \log P(m|Y, K)$

- Small entropy means that the best segmentation in K is by far the best fit to the data

$P(m|Y)$ and $P(K|Y)$

$P(m|Y)$

$$P(m|Y) = P(Y|m).P(m) = \Pi_{r \in m} P(Y^r|r).P(m)$$

- $P(Y^r|r) = \int P(Y^r|\theta_r)P(\theta_r)d\theta_r$, with θ_r parameters or segment r
- BIC criteria is derived from an approximation of this $P(m|Y)$
- In fact, it can be computed exactly

$P(K|Y)$

$$P(Y, K) = \sum_{m \in \mathcal{M}_K} P(Y, m)$$

- $P(K|Y)$ can be computed as successive matrix-vector products
- Similar computations were proposed by using backward-forward like algorithms [Fearnhead(2005), Guédon(2008)]
- $P(K|Y)$ can be used to select the number of segments

Posterior probability of a segment

Posterior probability of a segment

$\mathcal{S}_{K,k}([\![t_1, t_2]\!])$ segmentations having $r = [\![t_1, t_2[\![$ as their k-th segment.

- Compute exactly their probability $S_{K,k}([\![t_1, t_2]\!])$ in $O(n^2)$:

$k-1$ seg. before t_1 $\times$ 1 between t_1 & t_2 $\times$ $K-k$ after t_2

$$\mathcal{M}_{k-1}([\![1, t_1]\!]) \quad \times \quad \{[\![t_1, t_2]\!]\} \quad \times \mathcal{M}_{K-k}([\![t_2, n+1]\!])$$

$\mathcal{S}_K([\![t_1, t_2]\!])$ segmentations including segment $[\![t_1, t_2[\![$

$$\mathcal{S}_K([\![t_1, t_2]\!]) = \bigcup_k \mathcal{S}_{K,k}([\![t_1, t_2]\!])$$

$$S_K([\![t_1, t_2]\!]) = \sum_k S_{K,k}([\![t_1, t_2]\!])$$

Entropy

Entropy

- Exact computation in $O(K.n^2)$, uses the posterior probability of segments

$$\begin{aligned}\mathcal{H}(K) = & -\sum_{m\in\mathcal{M}_K} P(m|Y,K)\log P(m|Y,K) \\ = & -\sum_{m\in\mathcal{M}_K} P(m|Y,K)\log(\Pi_{r\in m}P(Y^r|r).P(m)) \\ = & -\sum_r S_K(r)\log P(Y^r|r) + \log P(K|Y)\end{aligned}$$

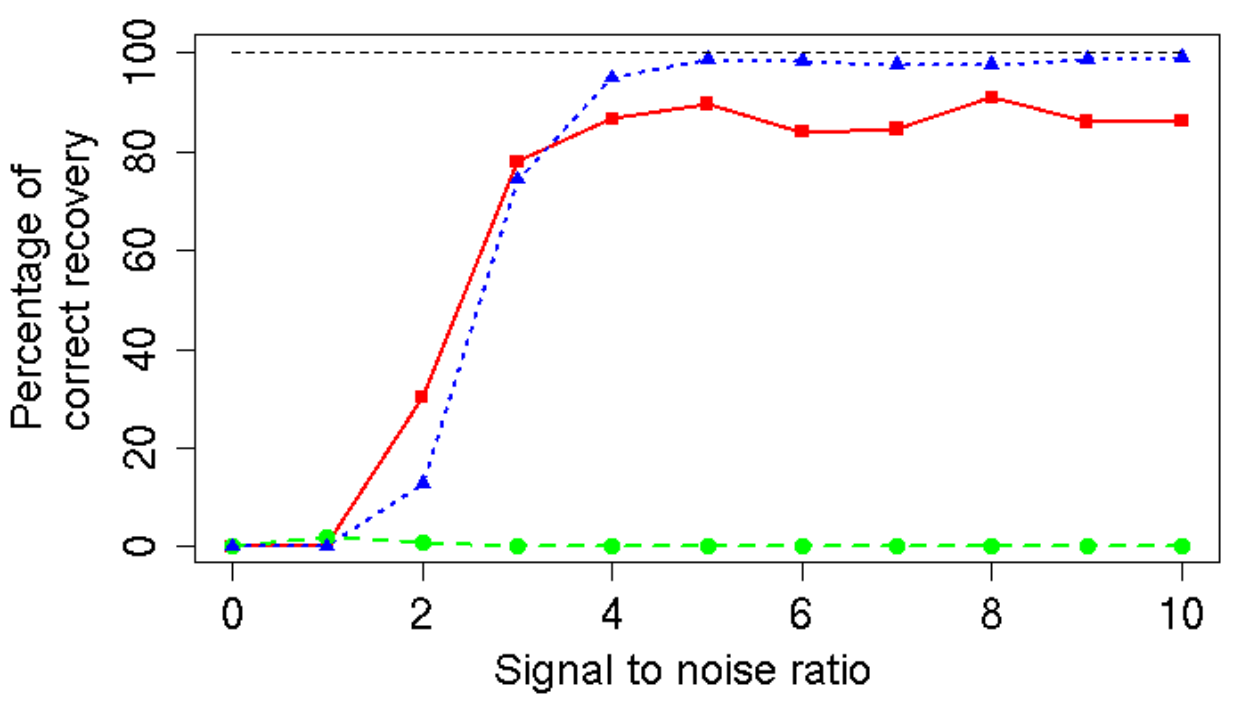
ICL

$$\mathrm{ICL}(K) = -\log P(Y,K) + \mathcal{H}(K)$$

Simulation

Design and results

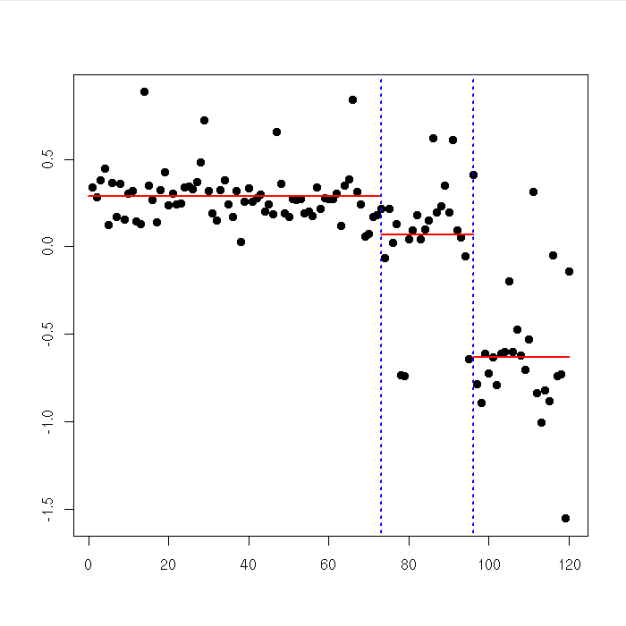
- Simulated sequence of 150 observations
- 6 change-points (positions: 21, 29, 68, 82, 115, 135).
- Do $P(m|Y)$, $P(K|Y)$ and $\mathrm{ICL}(K)$ recover the correct number of breakpoints (in relation with the level of noise)?



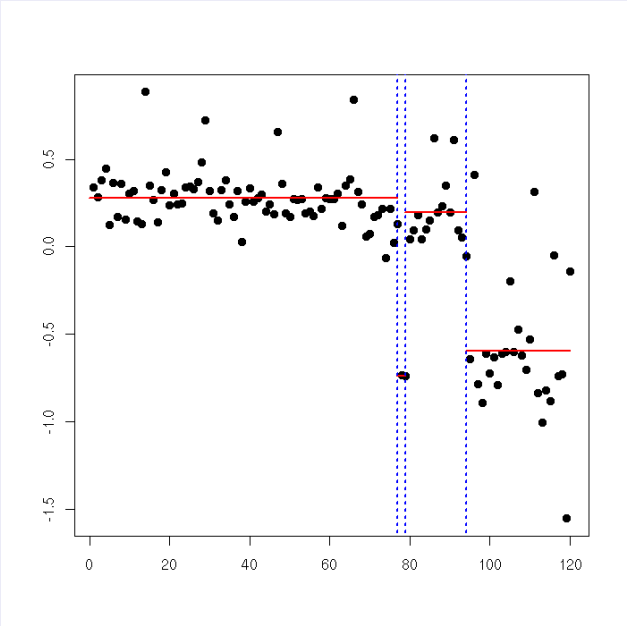
A CGH example

CGH Profiles

$P(m|Y)$: 3 segments



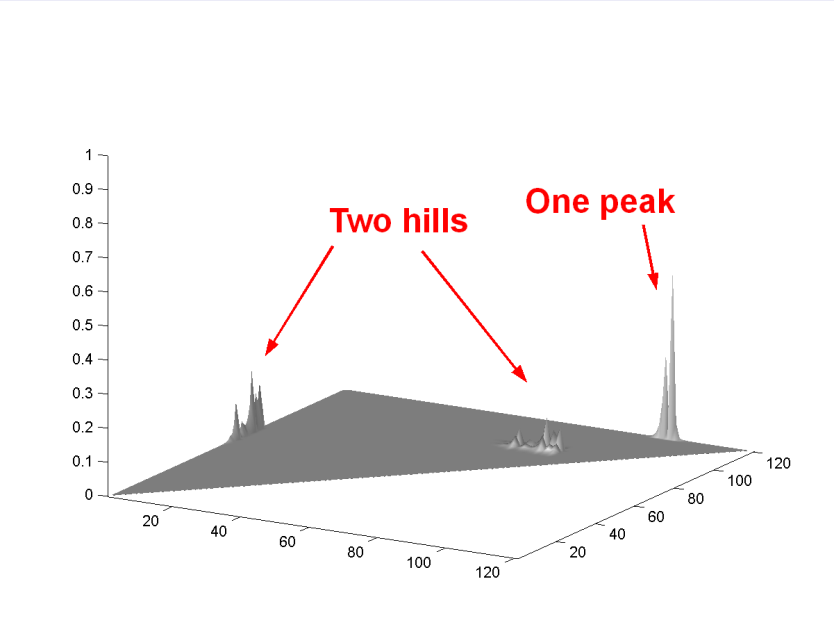
ICL(K): 4 segments



A CGH example

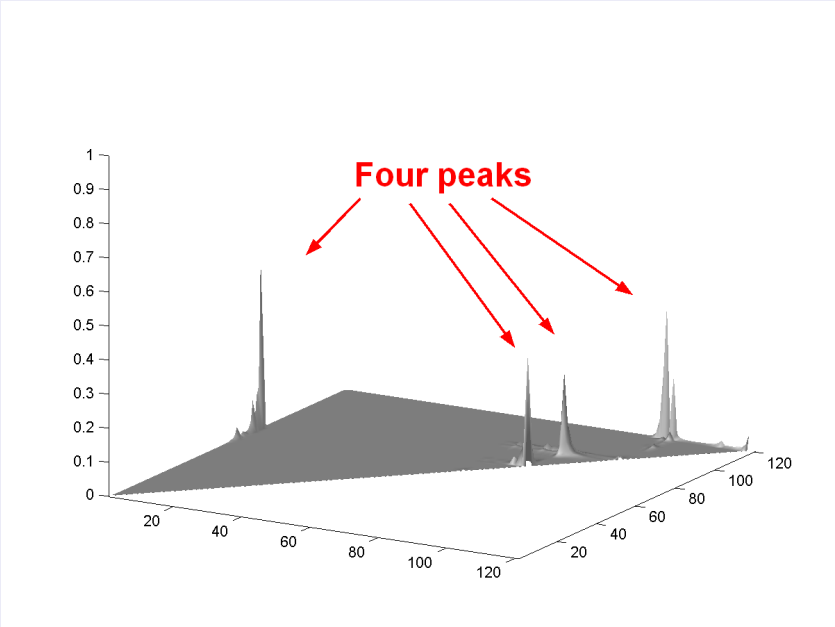
ICL favors segmentations with small entropy

$P(m|Y)$: 3 segments



Segments probability if $K = 3$

ICL(K): 4 segments



Segments probability if $K = 4$

Conclusion

- Exact computation in $O(Kn^2)$
  - Posterior Probability of a segment
  - Entropy of the segmentation space

- Model selection
  - Exact computation of $P(m|Y)$
  - Exact computation of $P(K|Y)$
  - Exact computation of ICL(K) (using the entropy)

References

Zhang, N. R. and Siegmund, D. O. (2007)
A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data
Biometrics, 63, 22–32, PMID: 17447926

Biernacki, C. and Celeux, G. and Govaert, G. (2000)
Assessing a mixture model for clustering with the integrated completed likelihood
IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 719–725.

Fearnhead, P. (2005),
Exact Bayesian curve fitting and signal segmentation,
IEEE Transactions on Signal Processing, 53, 2160–2166.

Guédon, Y. (2008),
Exploring the segmentation space for the assessment of multiple change-point models, Tech. Rep. 6619, INRIA.