Non parametric DNA copy number segmentation using kernels

M. Pierre-Jean[1], A. Célisse[4,5], G. Marot[3,5] and G. Rigaill[1,2]

[1]LaMME, Université d'Evry Val d'Essonne
[2]URGV, Université d'Evry Val d'Essonne
[3]CERIM, Université Lille 2
[4]Laboratoire Paul Painlevé, Université Lille 1
[5]INRIA, MODAL Team-Project.

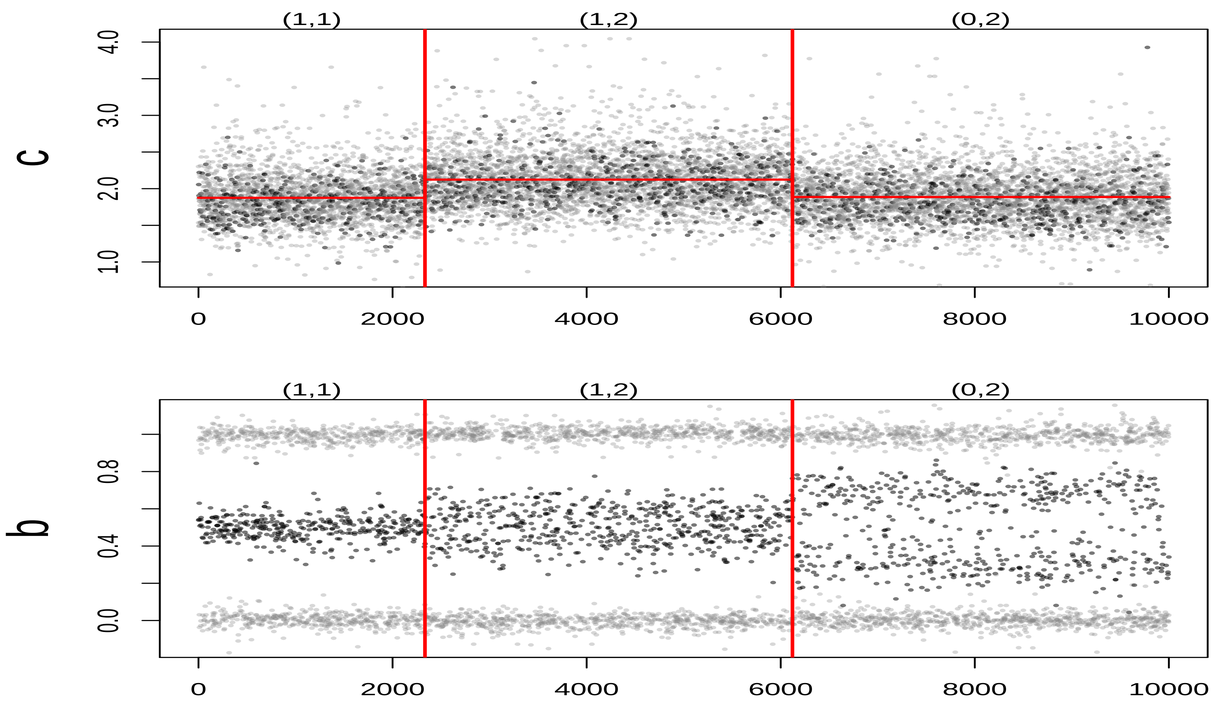
Framework and notations

Detection changes in the whole distribution

We consider the non-parametric change-point problem where one tries to recover abrupt changes arising in the distribution of random variables $X_1, \ldots, X_n$.

Application: detection of DNA alteration in cancer using SNP arrays. We jointly segment the copy number and baf signal (changes are not only in the mean or variance of the signal).

Illustration : SNP array data



Detecting changes in the mean in a RKHS

We have n independent ordered observations $(\mathrm{X}_1, \ldots, \mathrm{X}_n) \in \mathcal{X}^n$.

$\mathcal{H}$ is the reproducing kernel Hilbert space (RKHS) associated to kernel k, and $\phi = \mathcal{X} \mapsto \mathcal{H}$ is the canonical feature map defined by $\Phi(x) = k(x, \bullet)$.

with $\forall x, y \in \mathcal{X}$ we have : $\langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} = k(x, y)$.

For each locus $t \in \{1, \ldots, n\}$ and $\mathrm{X}_t \in \mathcal{X}$ we define:

$$\mathrm{Y}_t = \Phi(\mathrm{X}_t) \in \mathcal{H}.$$

Key Idea

For a characteristic kernel, the mean element ([1]) $\mu_t^\star \in \mathcal{H}$ is such that:

$$\mathbb{P}_{\mathrm{X}_i} \neq \mathbb{P}_{\mathrm{X}_j} \Rightarrow \mu_i^\star \neq \mu_j^\star$$

In other words detecting changes in the distribution is equivalent to detecting changes in the mean of the RKHS

However for a generic kernel :

- existing exact algorithms are in $\mathcal{O}(n^4)$ time and $\mathcal{O}(n^2)$ space
- existing heuristics are in $\mathcal{O}(D_{max} n^2)$ time and $\mathcal{O}(n)$ space

Algorithms and statistical performances

A faster exact algorithm and a heuristic

Our algorithm recovers for all D between 1 and D_{max} :

$$\mathbf{L}_{D,n+1} = \min_{m \in \mathcal{M}_{D,n+1}} \|Y - \hat{\mu}_m\|^2_{\mathcal{H}_n} \quad \text{(best loss)}$$

where $\mathcal{M}_{D,n+1}$ is the set of segmentation in D segments of $\{1 \ldots n\}$.

- Exact algorithm: $\mathcal{O}(D_{max} n^2)$ time and $\mathcal{O}(D_{max} n)$ space
- Heuristic based on low-rank approximation : $\mathcal{O}(p^2 n D_{max})$ time

Key Idea

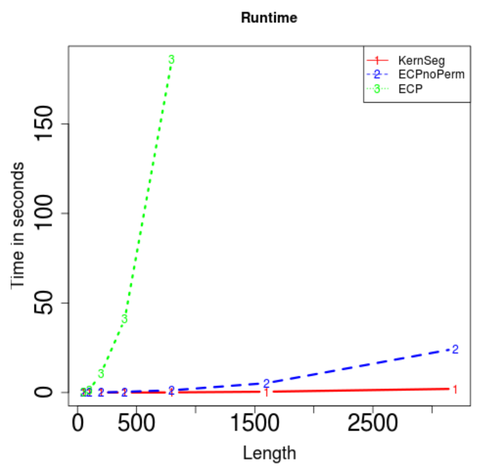
- Dynamic programming
- Iterative and on the fly computation of the cost matrix.

Simulations

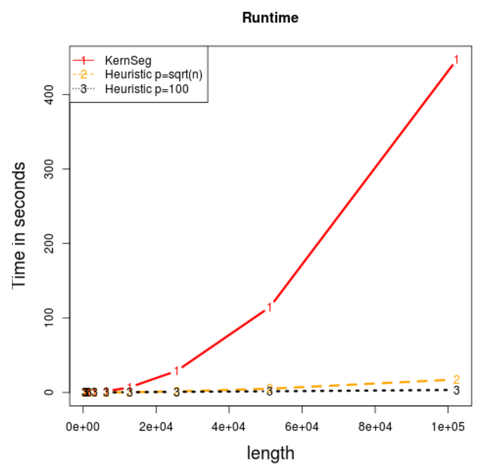
- Runtime: comparison with other non parametric approaches
- Accuracy and estimation (ℓ_2 ...): comparaison with RBS

Runtime

Faster than the ECP package



Process up to $n = 10^5$



Perfomances on realistic simulations (2)

	method	ℓ_2 risk in cn	risk in BAF	Accuracy for the true K	Accuracy with selection
100%	RBS	3.97 (0.15)	226.3 (4.8)	0.22 (0.006)	2.70 (0.09)
	Kernseg	**3.42** (0.14)	225.3 (4.9)	**0.20** (0.006)	**0.20** (0.006)
	Kernseg-approx	3.98 (0.16)	225.0 (4.9)	0.22 (0.006)	0.54 (0.06)
70%	RBS	3.35 (0.12)	219.9 (4.8)	0.43 (0.01)	6.03 (0.10)
	Kernseg	**2.88** (0.09)	216.6 (4.7)	**0.39** (0.009)	**0.39** (0.009)
	Kernseg-approx	3.28 (0.11)	219.3 (4.8)	0.42 (0.01)	0.83 (0.07)
50%	RBS	11.45 (0.53)	194.1 (4.5)	1.33 (0.04)	2.34 (0.09)
	Kernseg	**9.58** (0.40)	183.7 (3.9)	**1.20** (0.04)	**1.70** (0.06)
	Kernseg-approx	10.83 (0.36)	192.8 (3.8)	1.30 (0.03)	1.66 (0.07)

Performance for 100%, 70% and 50% purity for cn segmentation - mean (standard error)

Conclusion

Conclusions and Perspectives

1. Kernel segmentation gives better results than the state of the art Recursive Binary Segmentation
2. On SNP arrays Kernel segmentation catches true changes which are not only in the mean of the signal

Références

[1] S. Arlot, A. Celisse, Z. Harchaoui, Kernel change-point detection, Tech. Rep. hal-00671174, HAL (2012).

[2] M. Pierre-Jean, G. Rigaill, P. Neuvial, Performance evaluation of DNA copy number segmentation methods, Briefings in Bioinformatics.