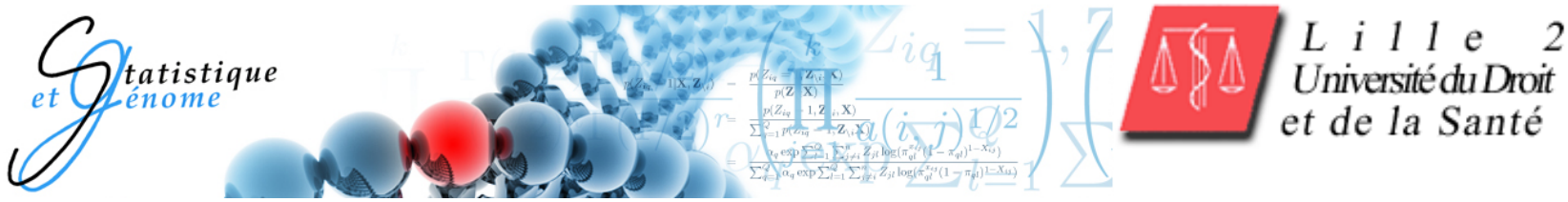


Joint copy number segmentation in cancer samples

Morgane Pierre-Jean

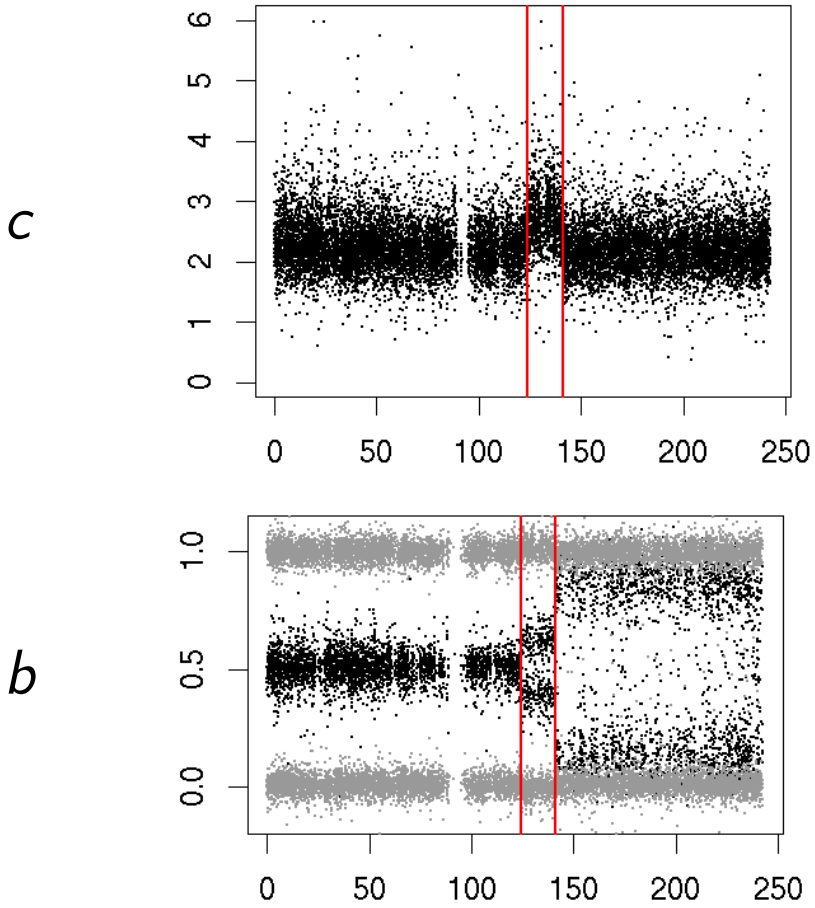
Laboratoire Statistique et Génome
Université d'Évry Val d'Éssonne
UMR CNRS 8071 USC INRA

CERIM
Université Lille 2 Droit et Santé
Faculté de médecine

2013-01-25

Outline

Outline



Breakpoints occur at exact same position in the two dimensions

Goal of DNA copy number studies : Identification of alterated genome regions.

- Understand tumor progression
- Lead to personalized therapies

We focused on identification of breakpoints

- Genomic signals from SNP arrays are bivariate
- Breakpoints occur exactly at the same position in the two-dimensions

Outline

A change-point model

- Biological assumption : DNA copy numbers are piecewise constant
- Statistical model for K change points at $(t_1, ...t_K)$:

$$\forall j = 1, \ldots, n \qquad c_j = \gamma_j + \epsilon_j$$

where $\forall k \in \{1, \ldots, K+1\}, \forall j \in [t_{k-1}, t_k[\qquad \gamma_j = \Gamma_k$

A change-point model

- Biological assumption : DNA copy numbers are piecewise constant
- Statistical model for K change points at $(t_1, \ldots t_K)$:

$$\forall j = 1, \ldots, n \qquad c_j = \gamma_j + \epsilon_j$$

where $\forall k \in \{1, \ldots, K+1\}, \forall j \in [t_{k-1}, t_k[\qquad \gamma_j = \Gamma_k$

Complexity

- Challenges : K and $(t_1, \ldots t_K)$ are unknown
- For a fixed K , the number of possible partitions : $C_{n-1}^K = \mathcal{O}(n^{K-1})$

State of the art

One dimension

More than one dimension

Exact solution by dynamic programming

[Picard et al. (2005)] : complexity in $\mathcal{O}(Kn^2)$
[Rigaill et al.(2010)] : mean complexity in $\mathcal{O}(nlog(n))$

State of the art

One dimension

More than one dimension

Exact solution by dynamic programming

[Picard et al. (2005)] : complexity in $\mathcal{O}(Kn^2)$
[Rigaill et al.(2010)] : mean complexity in $\mathcal{O}(nlog(n))$

Heuristics

[Harchaoui and Lévy-Leduc(2008)] : total variation distance with a complexity in $\mathcal{O}(Kn)$
[Olshen AB et al. (2004)] : Circular binary segmentation

State of the art

One dimension

Exact solution by dynamic programming

[Picard et al. (2005)] : complexity in $\mathcal{O}(Kn^2)$
[Rigaill et al.(2010)] : mean complexity in $\mathcal{O}(nlog(n))$

Heuristics

[Harchaoui and Lévy-Leduc(2008)] : total variation distance with a complexity in $\mathcal{O}(Kn)$
[Olshen AB et al. (2004)] : Circular binary segmentation

More than one dimension

Exact solution by dynamic programming

[Picard et al. (2005)] : complexity in $\mathcal{O}(Kn^2)$ for smaller problems

State of the art

One dimension

Exact solution by dynamic programming

[Picard et al. (2005)] : complexity in $\mathcal{O}(Kn^2)$
[Rigaill et al.(2010)] : mean complexity in $\mathcal{O}(nlog(n))$

Heuristics

[Harchaoui and Lévy-Leduc(2008)] : total variation distance with a complexity in $\mathcal{O}(Kn)$
[Olshen AB et al. (2004)] : Circular binary segmentation

More than one dimension

Exact solution by dynamic programming

[Picard et al. (2005)] : complexity in $\mathcal{O}(Kn^2)$ for smaller problems

Heuristics

[Bleakley and Vert(2011)] : group fused Lasso with complexity in $\mathcal{O}(Kn)$
[Zhang et al.(2010)] : Multivariate circular binary segmentation

State of the art

One dimension

Exact solution by dynamic programming

[Picard et al. (2005)] : complexity in $\mathcal{O}(Kn^2)$
[Rigaill et al.(2010)] : mean complexity in $\mathcal{O}(nlog(n))$

Heuristics

[Harchaoui and Lévy-Leduc(2008)] : total variation distance with a complexity in $\mathcal{O}(Kn)$
[Olshen AB et al. (2004)] : Circular binary segmentation

More than one dimension

Exact solution by dynamic programming

[Picard et al. (2005)] : complexity in $\mathcal{O}(Kn^2)$ for smaller problems

Heuristics

[Bleakley and Vert(2011)] : group fused Lasso with complexity in $\mathcal{O}(Kn)$
[Zhang et al.(2010)] : Multivariate circular binary segmentation

HMM

[Chen et al. (2011)] : HMM method using two dimensions

A two step approach for joint segmentation

The proposed joint segmention is a two-step approach.
Also used by [Bleakley and Vert(2011)]
First step :

- Running a fast but approximate segmentation method

Second step

- Pruning the final set of breakpoints using dynamic programming that is slower but exact

Binary Segmentation

- Take the simple case : dimension is equal to 1 ($d = 1$) :
- Hypothesis : $\mathcal{H}_0$: No breakpoint vs $\mathcal{H}_1$: Exactly one breakpoint.
- The likelihood ratio statistic is given by $\max_{1 \leq i \leq n} |Z_i|$

$$Z_i = \frac{\left(\frac{S_i}{i} - \frac{S_n - S_i}{n-i} \right)}{\sqrt{\frac{1}{i} + \frac{1}{n-i}}}, \qquad (1)$$

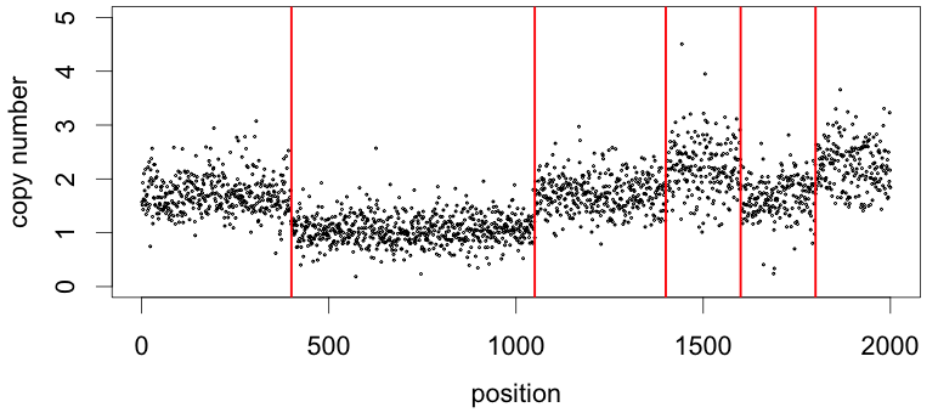
And $S_i = \sum_{1 \leq l \leq i} y_l$

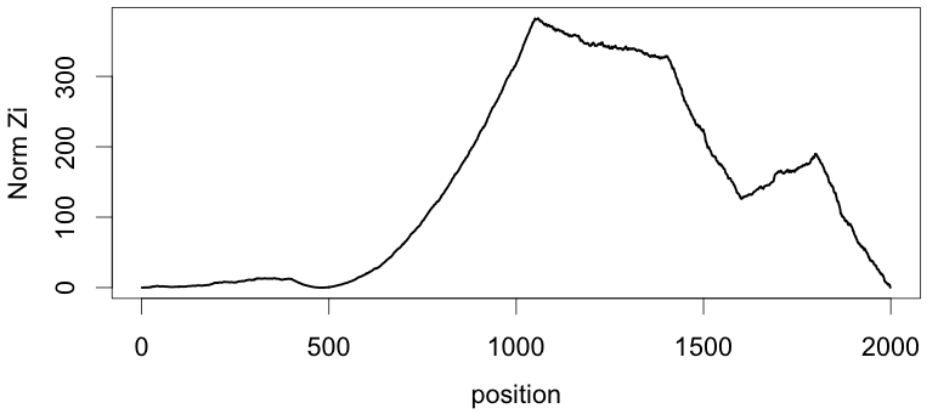
If $(d > 1)$: the likelihood ratio statistic becomes $\max_{1 \leq i \leq n} \|Z_i\|_2^2$

First step : Recursive Binary Segmentation (RBS)

Complexity : $O(dnlog(K))$

- First breakpoint
- For each i : we compute Z_i : $b_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$

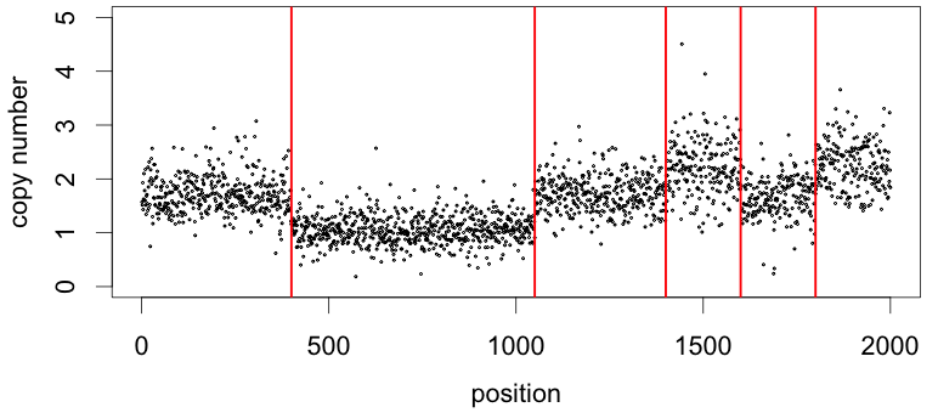


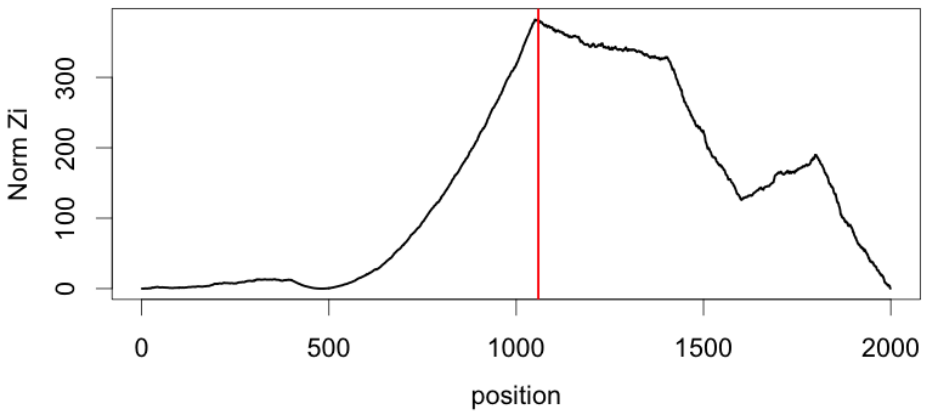


First step : Recursive Binary Segmentation (RBS)

Complexity : $O(dnlog(K))$

- First breakpoint
- For each i : we compute Z_i : $b_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$

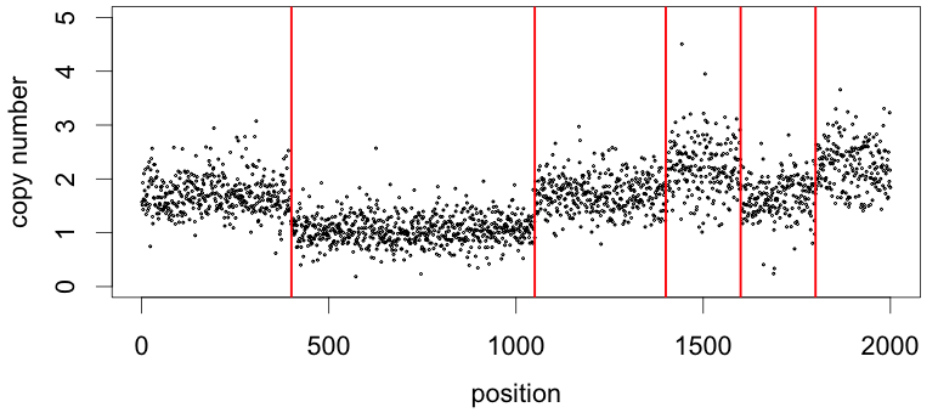


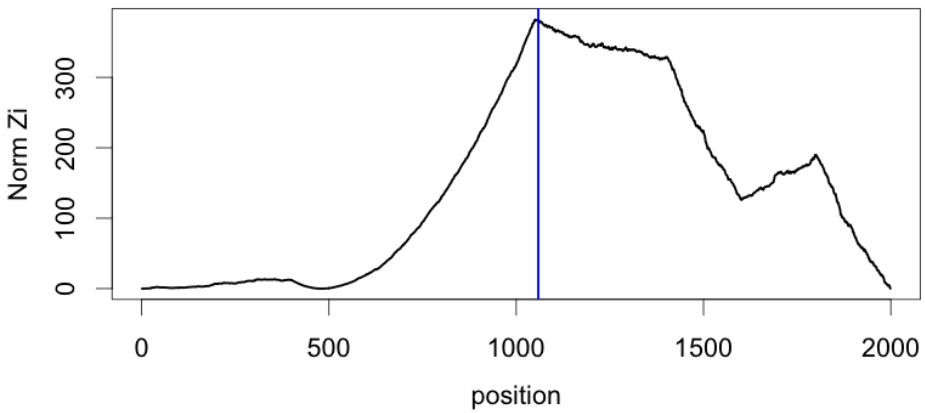


First step : Recursive Binary Segmentation (RBS)

Complexity : $O(dnlog(K))$

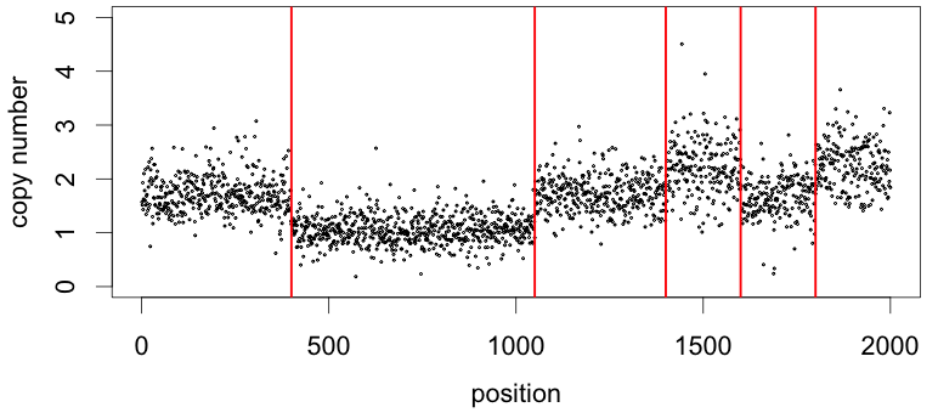
- First breakpoint
- For each i : we compute Z_i : $b_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$

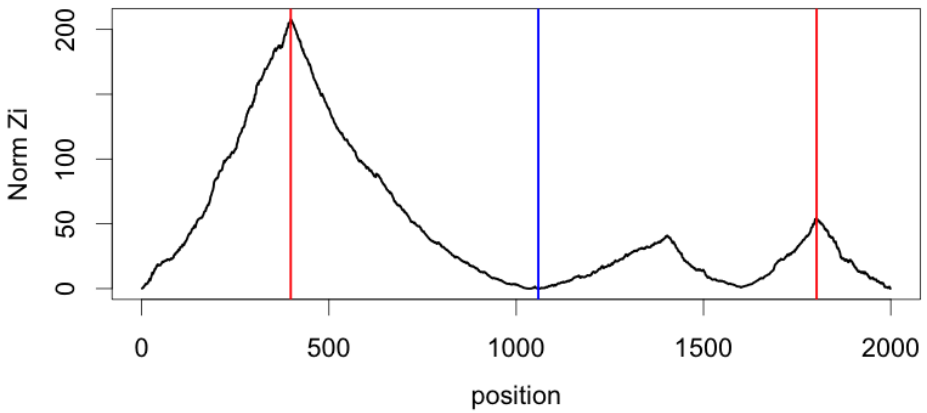




First step : Recursive Binary Segmentation (RBS)

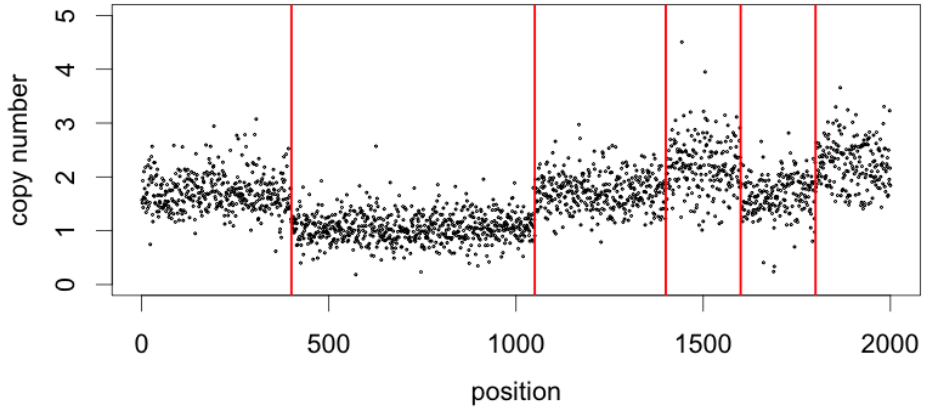
- Second breakpoint :
  - $\max_{1 \leq i \leq b_1} \|Z_i\|_2^2$
  - $\max_{b_1 \leq i \leq n} \|Z_i\|_2^2$
- Compute RSE for each segment.
- Keep the RSE which bring the maximum gain
- Add the breakpoint to the active set

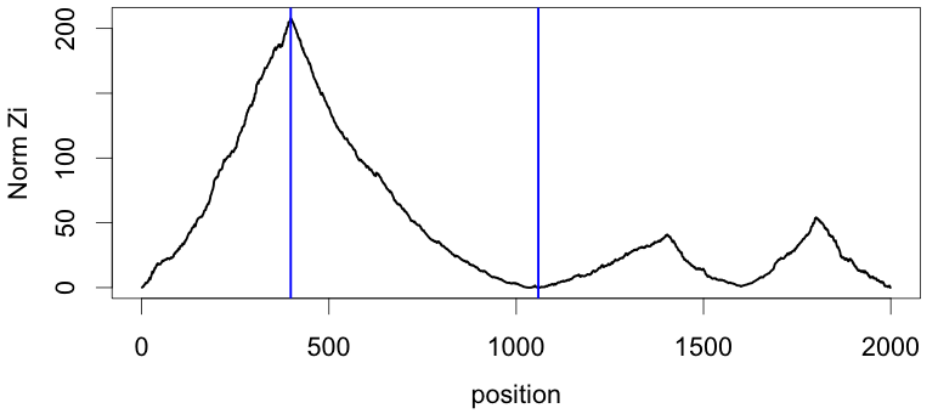




First step : Recursive Binary Segmentation (RBS)

- Second breakpoint :
  - $\max_{1 \leq i \leq b_1} \|Z_i\|_2^2$
  - $\max_{b_1 \leq i \leq n} \|Z_i\|_2^2$
- Compute RSE for each segment.
- Keep the RSE which bring the maximum gain
- Add the breakpoint to the active set



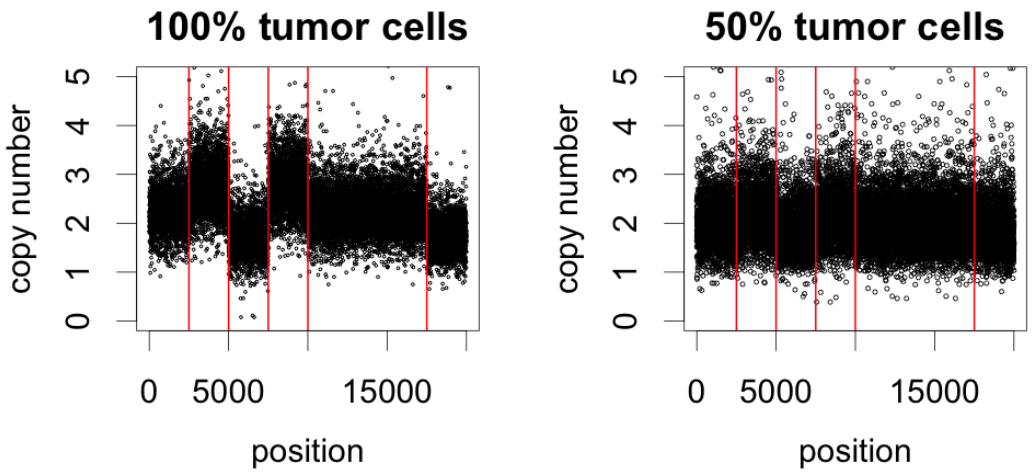


Outline

Simulated data creation

How did we create the simulated data ?

- From a real data set
  - For each technology (Illumina or Affymetrix) we have
    - Several data sets with various level of contamination by normal cells
    - Illumina : 34, 50, 79 and 100% of tumor cells
    - Affymetrix : 30, 50, 70 and 100% of tumor cells.
- Breakpoints are known

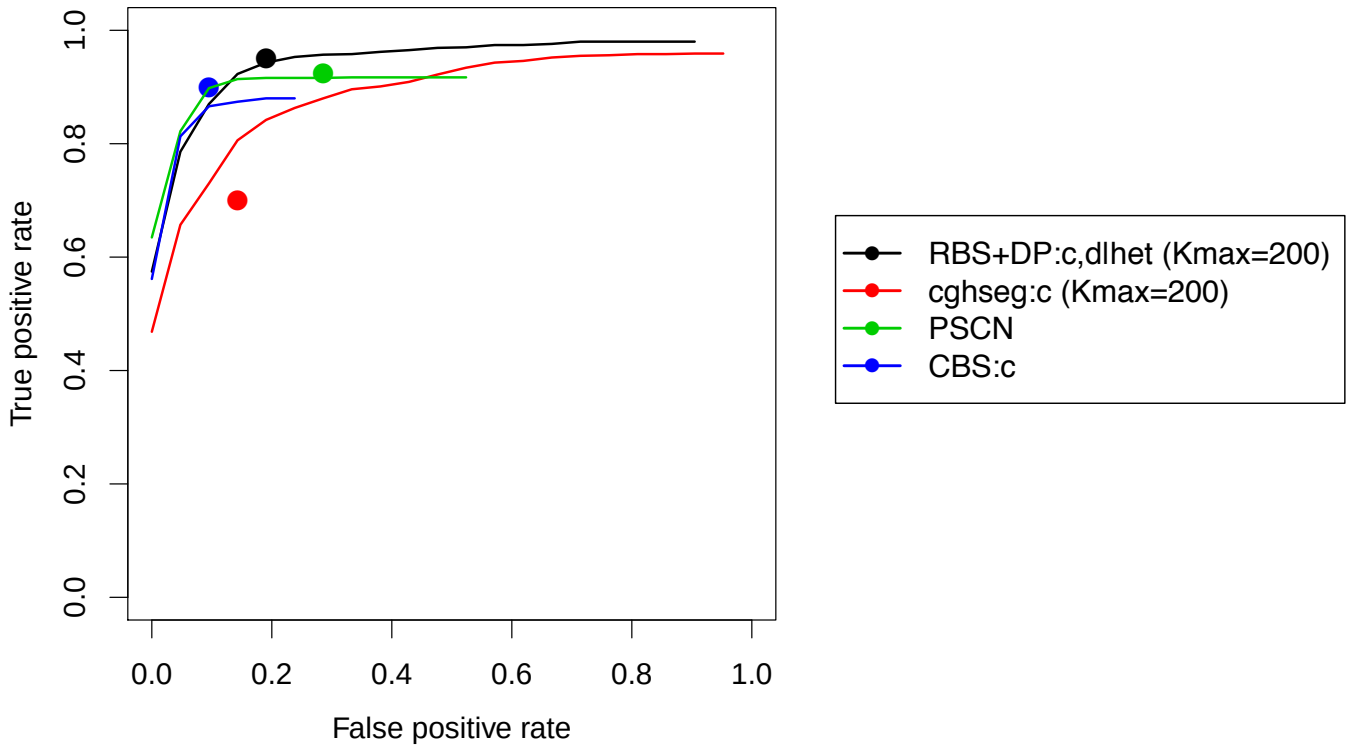


Joint segmentation results on simulated data

We created 50 profiles of length 50000 with 20 breakpoints and 70% tumor cells in sample

We assessed the precision of the methods

Precision = 20 (easy !)

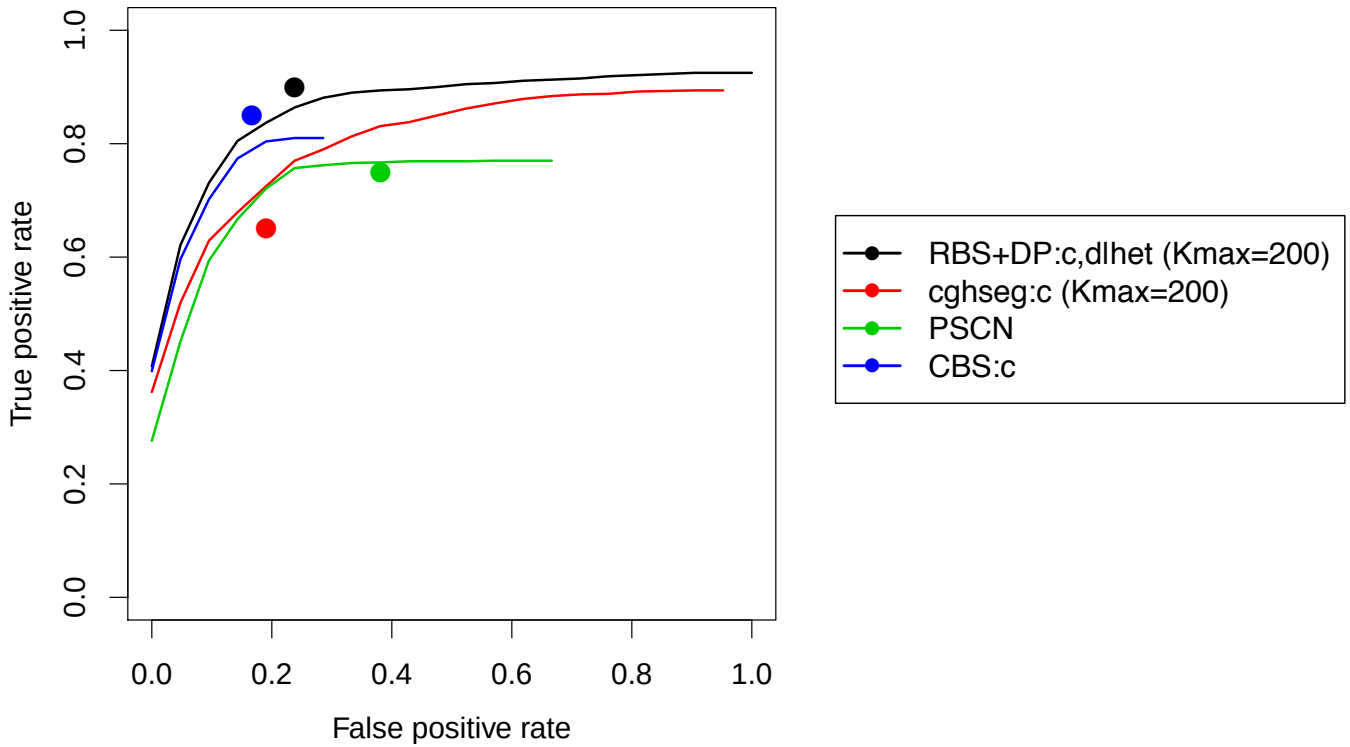


Joint segmentation results on simulated data

We created 50 profiles of length 50000 with 20 breakpoints and 70% tumor cells in sample

We assessed the precision of the methods

Precision = 10 (less easy !)

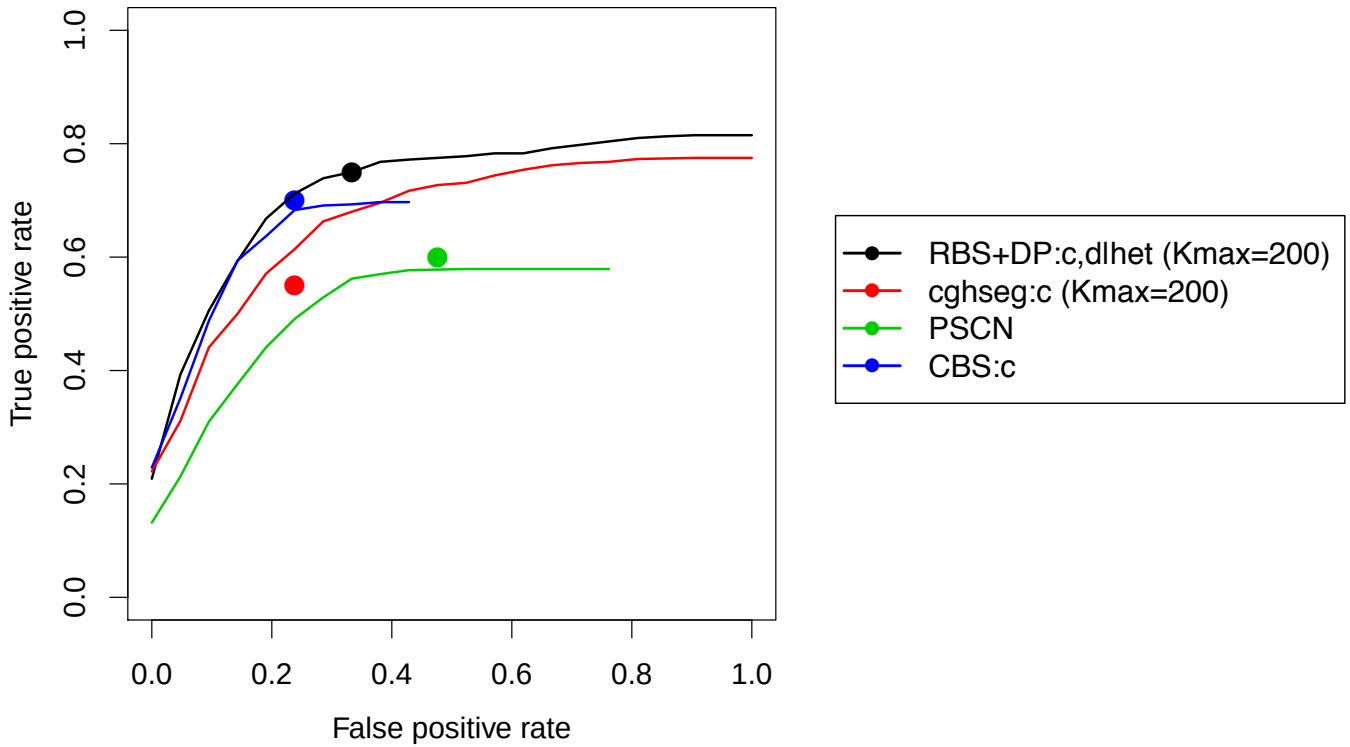


Joint segmentation results on simulated data

We created 50 profiles of length 50000 with 20 breakpoints and 70% tumor cells in sample

We assessed the precision of the methods

Precision = 5 (even less easy !)

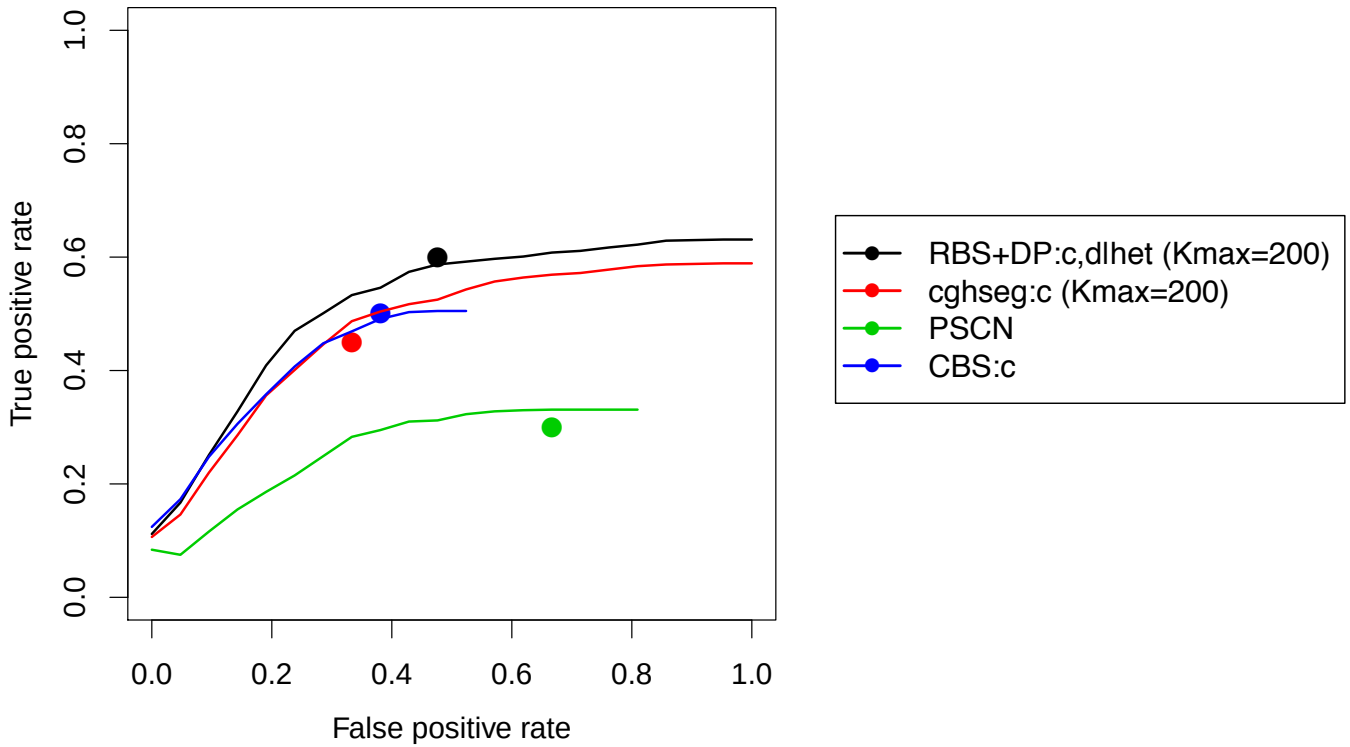


Joint segmentation results on simulated data

We created 50 profiles of length 50000 with 20 breakpoints and 70% tumor cells in sample

We assessed the precision of the methods

Precision = 2 (even less easy !)

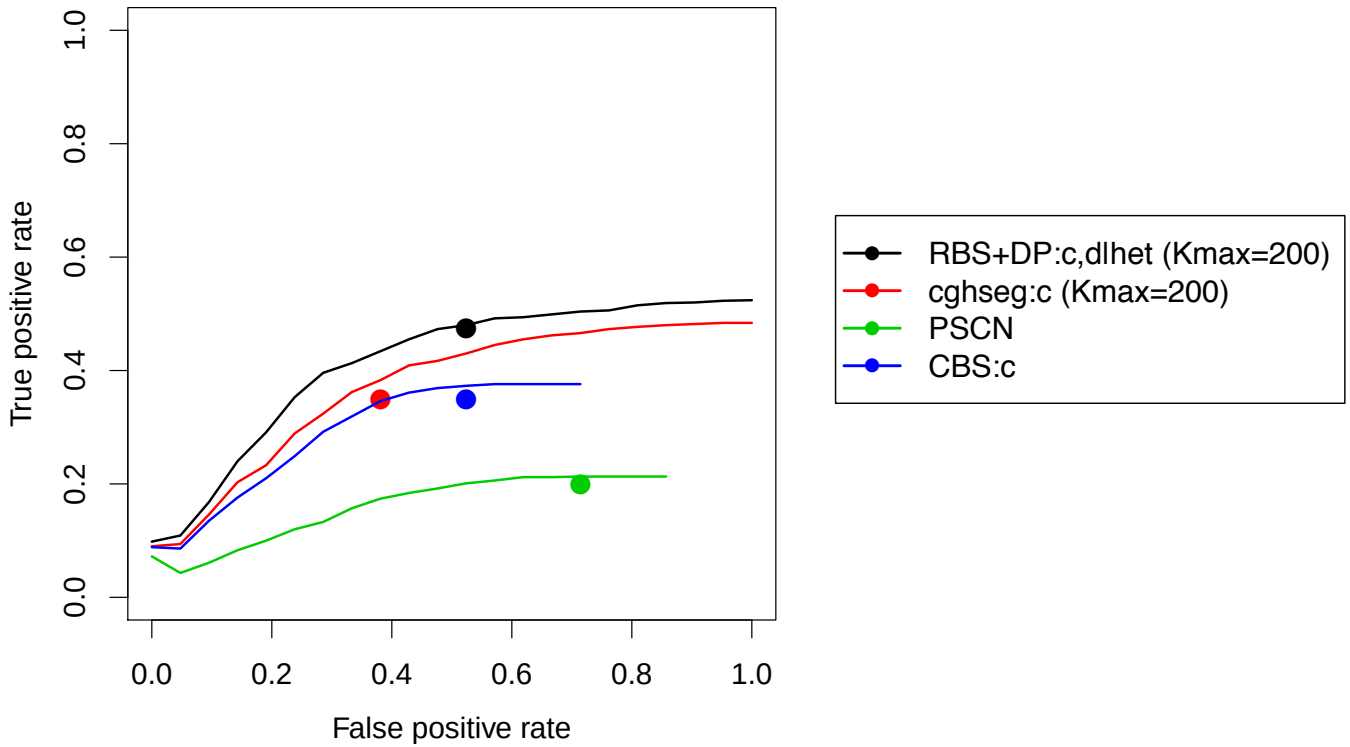


Joint segmentation results on simulated data

We created 50 profiles of length 50000 with 20 breakpoints and 70% tumor cells in sample

We assessed the precision of the methods

Precision = 1 (hard !)



Outline

Conclusion

Results

- Creation of realistic simulated data
- Fast method (RBS) using both dimensions
- More precise than existing methods, and faster.
- R package development.

Perspective

- Labeling of regions
- Extension to non gaussian settings

Thanks to Pierre Neuvial, Guillem Rigaill and Cyril Dalmasso

K. Bleakley and J.-P. Vert.
The group fused lasso for multiple change-point detection.
Technical report, Mines ParisTech, 2011.

Z. Harchaoui and C. Lévy-Leduc.
Catching change-points with lasso.
Advances in Neural Information Processing Systems, 2008.

G. Rigaill.
Pruned dynamic programming for optimal multiple change-point detection.
Technical report, http ://arXiv.org/abs/1004.0887, 2010.

G. Rigaill, E. Lebarbier, and S. Robin.
Exact posterior distributions and model selection criteria for multiple change point-criteria.
Statistics and Computing, 2012.

J.-P. Vert and K. Bleakley.
Fast detection of multiple change-points shared by many signals using group LARS.
Advances in Neutral Information Processing Systems, 2010.

F. Picard and E. Lebarbier and M. Hoebeke and G. Rigaill and B. Thiam and S. Robin.
Joint segmenation, calling and normalization of multiple CGH profiles.
Biostatistics,2011.


Chen, H., Xing, H. and Zhang, N.R.
Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays.
PLoS Comput Biol,2011.


Olshen AB, Venkatraman ES, Lucito R, Wigler M.
Circular binary segmentation for the analysis of array-based DNA copy number data.
Biostatistics, (2004).


Zhang, Nancy R. and Siegmund, David O. and Ji, Hanlee and Li, Jun Z.
Detecting simultaneous changepoints in multiple sequences.
Biometrika, (2010)


Lai, Tze Leung and Xing, Haipeng and Zhang, Nancy
Stochastic segmentation models for array-based comparative genomic hybridization data analysis
Biostat, (2008)


Zhang, Nancy R and Senbabaoglu, Yasin and Li, Jun Z,
Joint estimation of DNA copy number from multiple platforms

Boinformatics, (2010)