fig/LogoLille2

Comparison of segmentation methods in cancer samples

Morgane Pierre-Jean, Guillem Rigaill, Pierre Neuvial

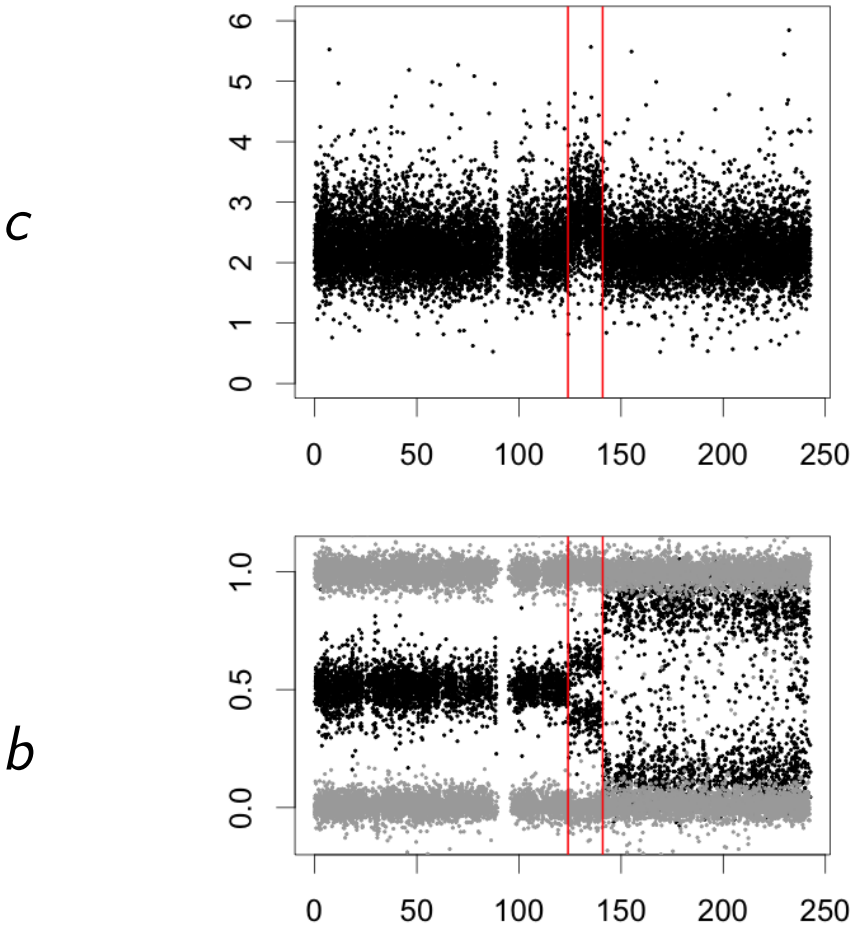
Laboratoire Statistique et Génome
Université d'Évry Val d'Éssonne
UMR CNRS 8071 USC INRA

CERIM
Université Lille 2 Droit et Santé
Faculté de médecine

2013-04-16

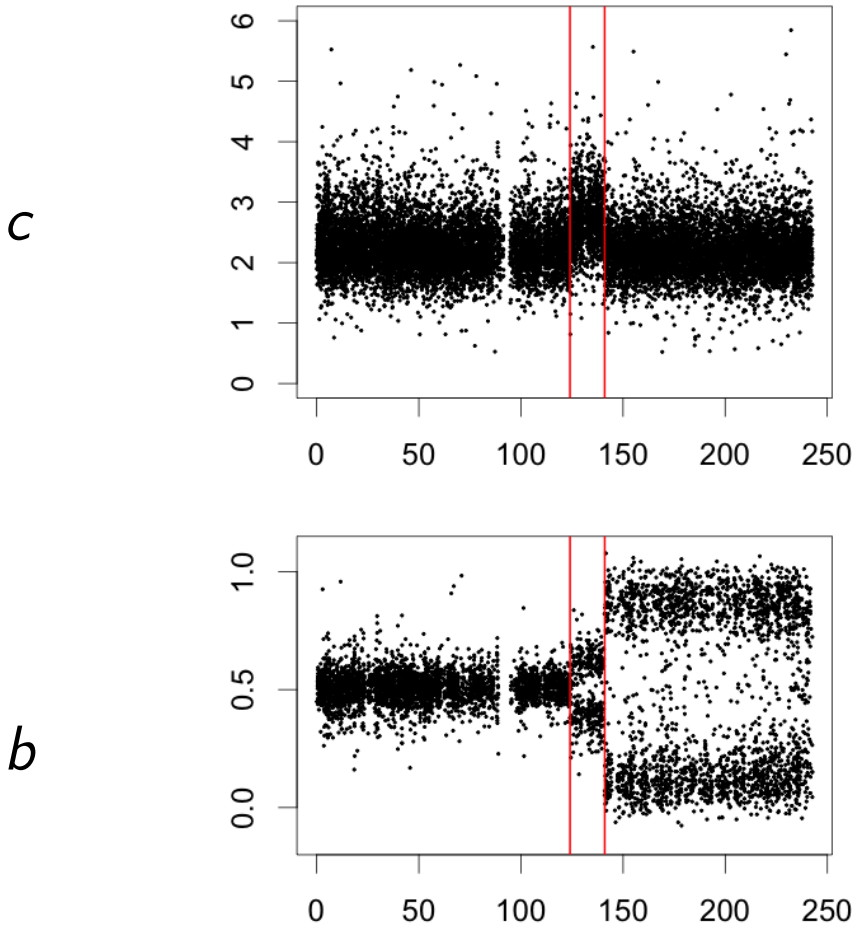
Outline

Outline



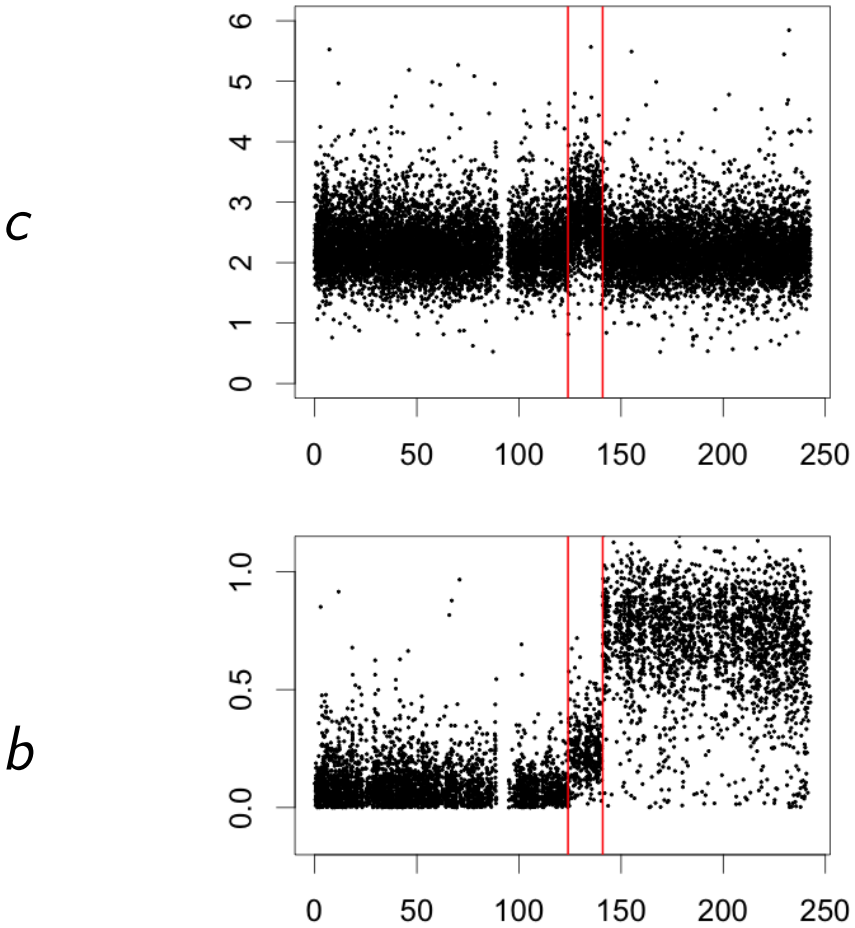
Chromosome 2 (Mb)

Breakpoints occur at exact same position in the two dimensions



Chromosome 2 (Mb)

Breakpoints occur at exact same position in the two dimensions



Chromosome 2 (Mb)

Breakpoints occur at exact same position in the two dimensions

Goal of DNA copy number studies : Identification of altered 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 or symmetrized B allele frequency 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 or symmetrized B allele frequency 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 $\forall k \in \{1, \ldots, K+1\}, \forall j \in [t_{k-1}, t_k[\qquad \gamma_j = \Gamma_k$

Complexity

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

State of the art : Exact solution

One dimension

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

Two dimensions

- Extension of [Picard et al. (2005)] : complexity in $\mathcal{O}(dKn^2)$ for smaller problems
- [Mosen-Ansorena, D et al (2013)] : complexity in $\mathcal{O}(dKnl)$ where l is the maximum length of segments

State of the art : Heuristics

Type	Name	Method	Dimension
Convex relaxation	FLASSO	total variation distance with a complexity in $\mathcal{O}(Kn)$	1 d
	GFLASSO	Group fused Lasso solved by LARS $\mathcal{O}(Knd)$	$\geq$2 d
Binary segmentation	CBS	Circular binary segmentation	1d
	CART	Classification and regression tree	1 d
	MCBS	Multivariate circular binary segmentation	$\geq$2 d
	PSCBS	CBS on copy number then on B allele frequency	2 d
	RBS	Recursive binary segmentation in 2 dimensions adaptation of CART	2 d
Other	PSCN	HMM (hidden Markov Model)	2 d

Two-step approaches for joint segmentation

[Gey,S and Lebarbier,E (2008)] and [Bleakley and Vert(2011)] proposed two-step approaches.
So, we implemented a fast joint segmentation using CART in 2d following by a pruning.
First step :

- Running a fast but approximate segmentation method (RBS)

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 $\max_{1 \leq i \leq n} |Z_i|$

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

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 : $t_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$

fig/RBS0.pdf

fig/RBS1.pdf

First step : Recursive Binary Segmentation (RBS)

Complexity : $O(dnlog(K))$

- First breakpoint
- For each i : we compute Z_i : $t_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$

fig/RBS2.pdf

First step : Recursive Binary Segmentation (RBS)

- Second breakpoint :
  - $\max_{1 \leq i \leq t_1} \|Z_i\|_2^2$
  - $\max_{t_1 < i \leq n} \|Z_i\|_2^2$
- Compute RSE for each segment.
- Keep the RSE which bring the maximum gain
- Add the breakpoint to the active set

fig/RBS3.pdf

First step : Recursive Binary Segmentation (RBS)

- Second breakpoint :
  - $\max_{1 \leq i \leq t_1} \|Z_i\|_2^2$
  - $\max_{t_1 < i \leq n} \|Z_i\|_2^2$
- Compute RSE for each segment.
- Keep the RSE which bring the maximum gain
- Add the breakpoint to the active set

fig/RBS4.pdf

First step : Recursive Binary Segmentation (RBS)

- Third breakpoint :
  - $\max_{1 \leq i \leq t_1} \|Z_i\|_2^2$
  - $\max_{t_1 < i \leq t_2} \|Z_i\|_2^2$
  - $\max_{t_2 < i \leq n} \|Z_i\|_2^2$
- Compute RSE for each segment.
- Keep the RSE which bring the maximum gain
- Add the breakpoint to the active set

fig/RBS5.pdf

First step : Recursive Binary Segmentation (RBS)

- Third breakpoint :
  - $\max_{1 \leq i \leq t_1} \|Z_i\|_2^2$
  - $\max_{t_1 < i \leq t_2} \|Z_i\|_2^2$
  - $\max_{t_2 < i \leq n} \|Z_i\|_2^2$
- Compute RSE for each segment.
- Keep the RSE which bring the maximum gain
- Add the breakpoint to the active set

fig/RBS6.pdf

First step : Recursive Binary Segmentation (RBS)

- Third breakpoint :
  - $\max_{1 \leq i \leq t_1} \|Z_i\|_2^2$
  - $\max_{t_1 < i \leq t_2} \|Z_i\|_2^2$
  - $\max_{t_2 < i \leq n} \|Z_i\|_2^2$
- Compute RSE for each segment.
- Keep the RSE which bring the maximum gain
- Add the breakpoint to the active set

fig/RBS7.pdf

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
- State of segments are also known

Affymetrix

fig/profileAffy50100.pdf

Illumina

fig/profileIllu50100.pdf

fig/TNTP.pdf

Illumina : Use 2 dimensions provides good results

100 profiles, $n = 5000$, $K = 5$, purity = 79%, precision = 1

fig/figIllumina/CRL2324,BAF,ROC,n=5000,K=5,regSize=0,minL=

fig/figIllumina/CRL2324,BAF,ROC

Illumina : Use 2 dimensions provides good results

100 profiles, $n = 5000$, $K = 5$, purity = 79%, precision = 1

fig/figIllumina/CRL2324,BAF,ROC,n=5000,K=5,regSize=0,minL=

fig/figIllumina/CRL2324,BAF,ROC

Illumina : Univariate methods are as good as bivariate

100 profiles, $n = 5000$, $K = 5$, purity = 100%, precision = 1

fig/figIllumina/CRL2324,BAF,ROC,n=5000,K=5,regSize=0,minL=

fig/figIllumina/CRL2324,BAF,ROC

Outline

Conclusion

Results

- Creation of realistic simulated data
- R package development 'jointSeg' on R-forge. https ://r-forge.r-project.org/R/ ?group_id=1562
- Bivariate methods are not uniformly better than univariate
- No superiority of one method

Perspective

- Kernel approaches
- Labelling
- Other applications (several profiles, methylation data)

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)

Gey, S. and Lebarbier, E.,
Using CART to Detect Multiple Change Points in the Mean for Large Sample,
Statistics for Systems Biology research group, (2008)

Olshen, Adam B and Bengtsson, Henrik and Neuvial, Pierre and Spellman, Paul T and Olshen, Richard A and Seshan, Venkatraman E,
Parent-specific copy number in paired tumor-normal studies using circular binary segmentation
Boinformatics, (2011)

Mosen-Ansorena, David and Aransay, Ana Maria,
Bivariate segmentation of SNP-array data for allele-specific copy number analysis in tumour samples
BMC Bioinformatics, (2013)