A clustering method for ordered variables to detect up-correlated genomic regions

Tristan Mary-Huard

A clustering method for ordered variables to detect up-correlated genomic regions

T. Mary-Huard, E. Lebarbier, S. Robin

UMR AgroParisTech/INRA 518

Basics on Cancer mechanisms

We consider microarray data from cancer experiments.

For normal tissues :

⋆ Two copies of each gene,

⋆ Correlations between expression profiles of neighbor genes is low.

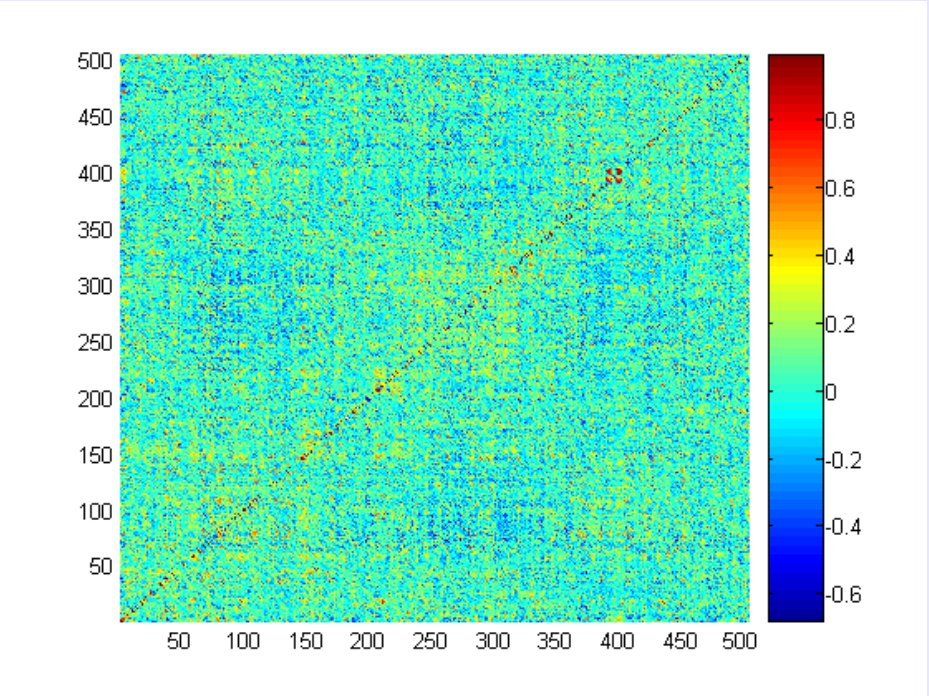
For tumorous tissues :

⋆ Amplification/Deletion may occur (CGH microarray)
$\Rightarrow$ k copies of a gene, with $k = 0, ..., 13, ...,$

⋆ Small regions of genes with highly correlated expression
$\Rightarrow$ Chromosomal domains of interest.

Example of chromosomal domains

	Gene1	Gene2	...
Sample1	x_1^1	x_1^2	...
Sample2	x_2^1	x_2^2	...
Sample3	x_3^1	x_3^2	...
...	...	...	...



Chromosomal domains may be due to amplification, deletion, epigenetic...

Objective : identify the up-correlated regions.

Variable clustering

Model :

Let $X^1, ..., X^p$ be an ordered sequence of variables, that breakdowns into K clusters of variables $\mathscr{C}_1, ..., \mathscr{C}_K$, such that

$$\forall j \in \mathscr{C}_k, \quad X^j = A^k + E^j$$

A^k : cluster variable, $V(A^k) = \sigma_k^2 I_n$
E^j : residual variable, $V(E^j) = \sigma^2 I_n$
$\text{cov}(A^k, E^j) = 0, \quad \text{cov}(A^k, A^{k'}) = 0, \quad \text{cov}(E^j, E^{j'}) = 0$

⋆ In the following, we assume that variables are standardized.

⋆ We note ρ_k the correlation between 2 variables in cluster $\mathscr{C}_k$.

⋆ We assume that for most of the clusters $\rho_k = \rho_0$, and for a small number of clusters $\rho_k > \rho_0$.

Optimal clustering

Two step strategy

⋆ Identify the clusters (K and $\mathscr{C}_1, ..., \mathscr{C}_K$),

⋆ Test for each cluster whether $\rho_k = \rho_0$ or $\rho_k > \rho_0$.

Loss for clusters

⋆ If variables $X^{\ell+1}, ..., X^{\ell+p_k}$ belong to the same cluster $\mathscr{C}_k$, they are noisy copies of variable A^k.

⋆ $\widehat{A}^k = \dfrac{1}{p_k} \displaystyle\sum_{j=\ell+1}^{\ell+p_k} X^j = \overline{X}^{(k)}$ (BLUP)

$$\Rightarrow L(X^{(k)}) = p_k - \sum_{j=\ell+1}^{\ell+p_k} \operatorname{cov}\left(X^j, \overline{X}^{(k)}\right)$$

Optimal clustering

We look for clustering $\mathscr{C}^*$ that satisfies

$$\begin{aligned} \mathscr{C}^* &= \operatorname*{Argmin}_{\mathscr{C}} L\left(X^1, ..., X^p\right) \\ &= \operatorname*{Argmin}_{\mathscr{C}} \sum_{k=1}^{K} L(X^{(k)}) \ . \end{aligned}$$

Clustering Algorithm

Exhaustive search

⋆ Variables are ordered ⇒ clusters of adjacent variables only

⋆ The loss to optimize is additive on clusters

⇒ Optimal solution can be found using Dynamic Programming

Complexity : $\mathcal{O}(p^2)$

Heuristic search

We look for applicability to large sequences of variables ($p \sim 10^4$)

⇒ Use of a constrained version of Hierarchical Clustering Algorithm

Complexity : $\mathcal{O}(p)$

Distance between clusters :

$$D(X^{(k)}, X^{(\ell)}) \quad = \quad L(X^{(k)}, X^{(\ell)}) - L(X^{(k)}) - L(X^{(\ell)})$$

Estimation of the number of clusters

Objective :

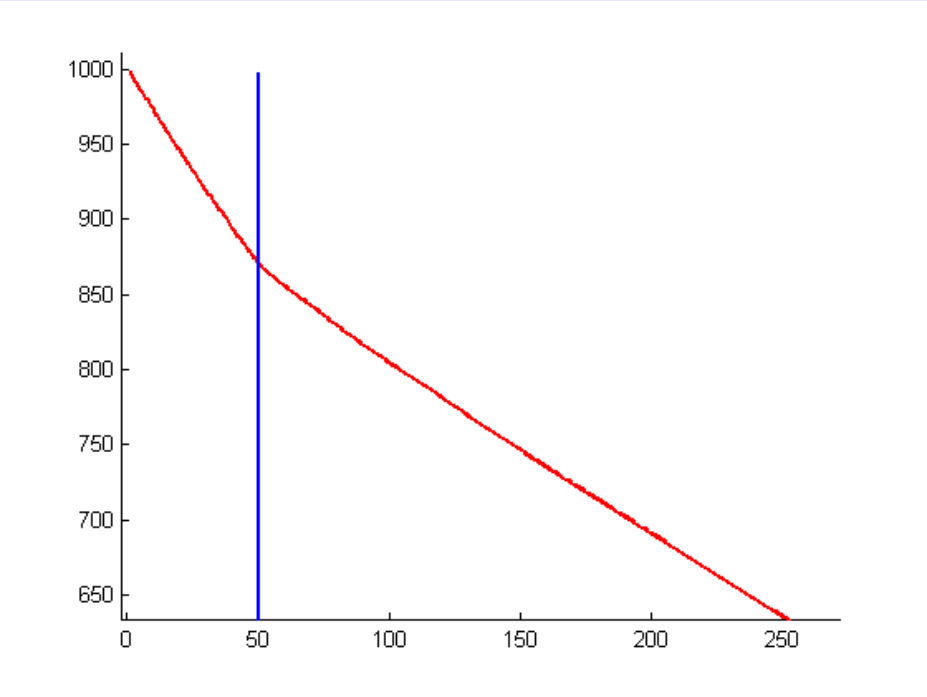
Find the break in the slope of the clustering curve.

Same objective as for adaptive penalty in model selection context Lavielle (2003)

Choice of K

$$\beta_K = (L_K - L_{K+1}) - (L_{K+1} - L_{K+2})$$

$$\widehat{K} = \underset{K}{\text{Argmin}} \{\beta_K > S\}$$



Testing the clusters

Test : $\mathscr{H}_0 : \{\rho_k = \rho_0\}$ vs $\mathscr{H}_1 : \{\rho_k > \rho_0\}$ (conditionally to the clustering)

Assuming A^k and E^j to be gaussian, for observation i we have :

$$\begin{aligned} \overline{X}_i^{(k)} &= \frac{1}{p_k} \sum_{j \in \mathscr{C}_k} X^j \sim \mathscr{N}\left(0,\ \frac{1}{p_k}\left(1 + (p_k - 1)\rho_k\right)\right) \\ \Rightarrow \widehat{V}\left(\overline{X}^{(k)}\right) &= \frac{1}{n}\sum_i^n \left(\overline{X}_i^{(k)} - \overline{\overline{X}^{(k)}}\right)^2 \sim \frac{\left(1 + (p_k - 1)\rho_k\right)}{n p_k} \chi^2_{(n-1)} \end{aligned}$$

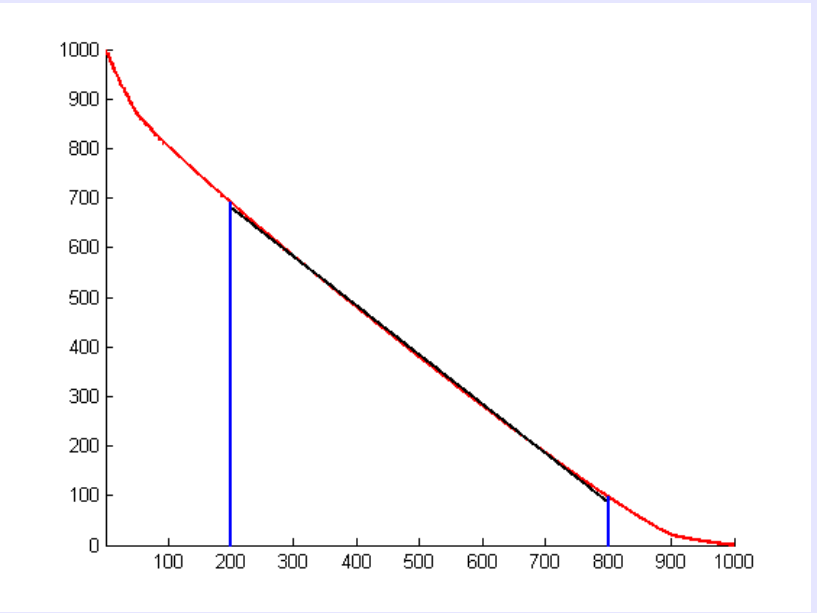
$$\Rightarrow n\left(p_k - L(X^{(k)})\right) \underset{\mathscr{H}_0}{\sim} \left(1 + (p_k - 1)\rho_0\right)\chi^2_{(n-1)}$$

Reject $\mathscr{H}_0$ if $n\left(p_k - L(X^{(k)})\right) > \left(1 + (p_k - 1)\widehat{\rho}_0\right)\chi^2_{n-1,1-\alpha}$

Estimation of ρ_0

A 3-phase clustering curve :

- ⋆ Clusterings in correlated regions
- ⋆ Clusterings in $\mathscr{H}_0$ regions
- ⋆ Clustering between regions



Theoretical distance for an $\mathscr{H}_0$ clustering step :

$$\begin{aligned} D(X^{(k)}, X^{(\ell)}) &= L(X^{(k)}, X^{(\ell)}) - L(X^{(k)}) - L(X^{(\ell)}) \\ &= 1 - \rho_0 \end{aligned}$$

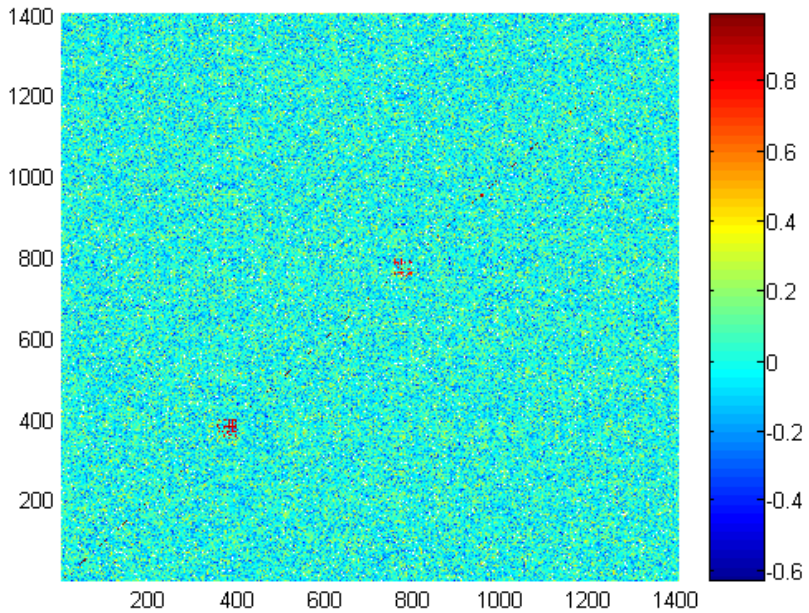
$\Rightarrow \rho_0$ may be estimated with the slope estimate of regression between b_{low} and b_{up} .

First example : Simulated data

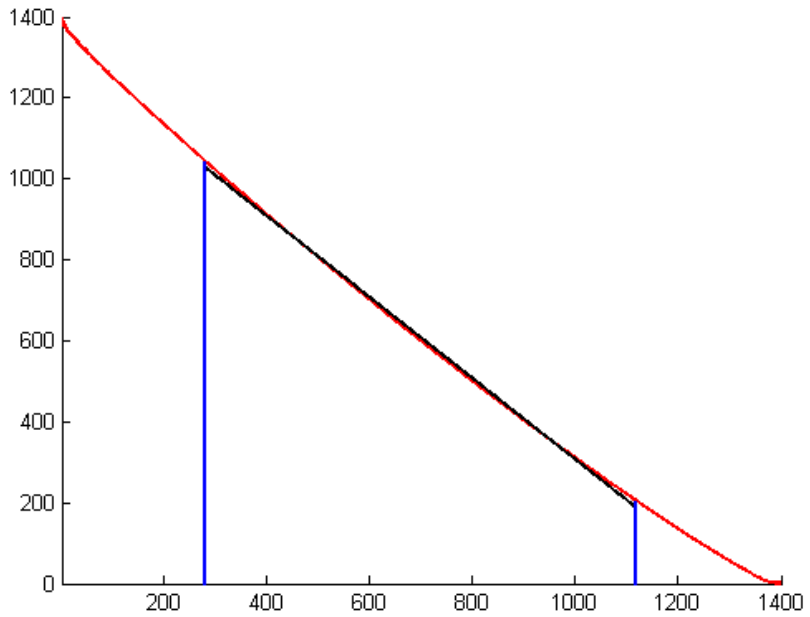
50 observations, 1400 variables
Background correlation : $\rho_0 = 0$

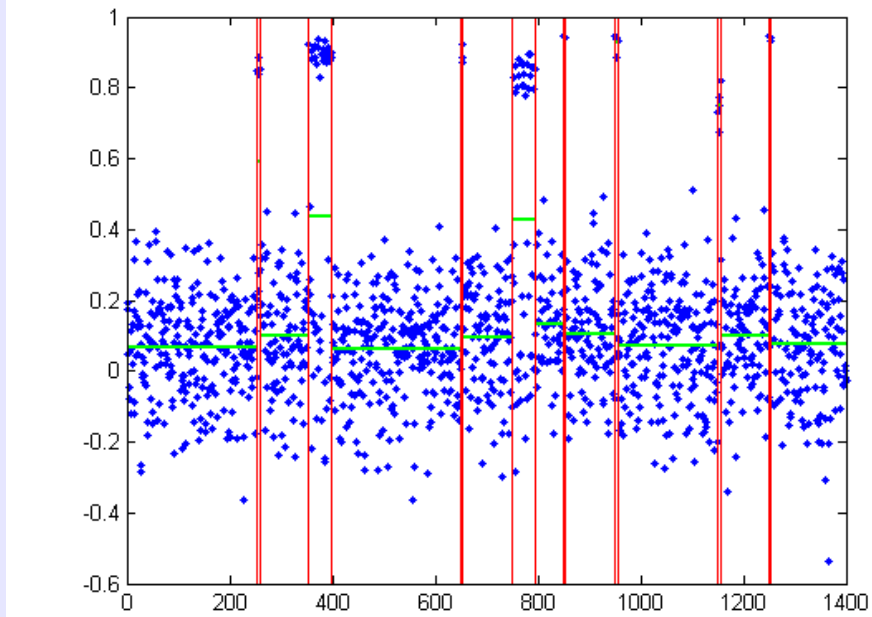
14 chromosomal domains :

- ⋆ 4 amplified regions (3/5/10/50)
- ⋆ 4 deleted regions (3/5/10/50)
- ⋆ 2 epigenetic regions (3/5)
- ⋆ 4 correlated regions (3/5/3/5)



Clustering step





Left :

Red : Clustering curve, from no clustering (right) to one cluster (left).

Black : estimation of ρ_0 by regression, $\widehat{\rho}_0 = 0$.

Right :

Green : average correlation with the aggregated variable

Blue : correlation between a gene and its aggregated variable

Number of clusters = 17, correlated : 8

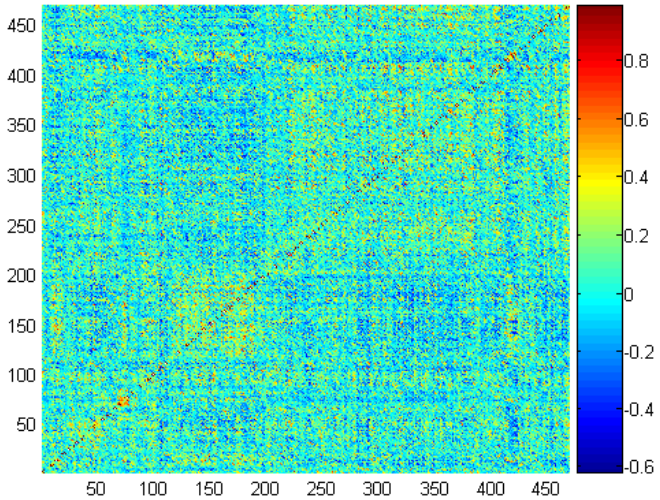
Identified regions

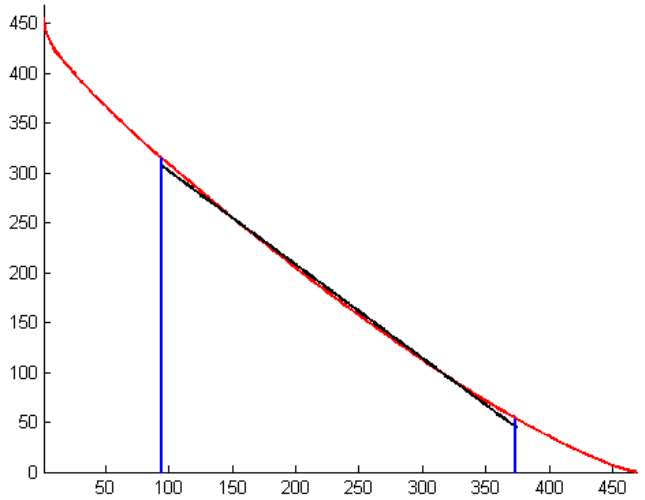
Type	Truth	Clustering	Sliding Windows
Amplifications	51-53		
	151-155		
	251-260	253-259	257,...,259
	351-400	353-398	365,...,398
Deletions	451-453		
	551-555		
	651-660	652-654	
	751-800	752-796	756,...,796
Epigenetics	851-853	851-853	
	951-955	951-955	951-955
Correlated	1051-1053		
	1151-1155	1151-1155	1155
	1251-1253	1251-1253	
	1351-1355		

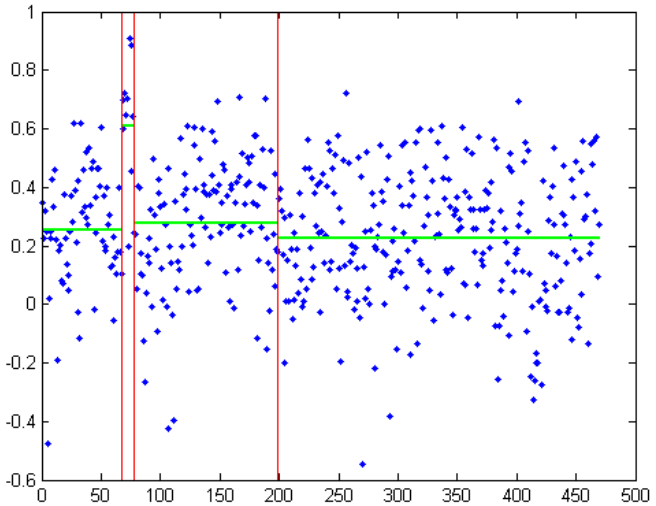
Red : poorly recovered regions

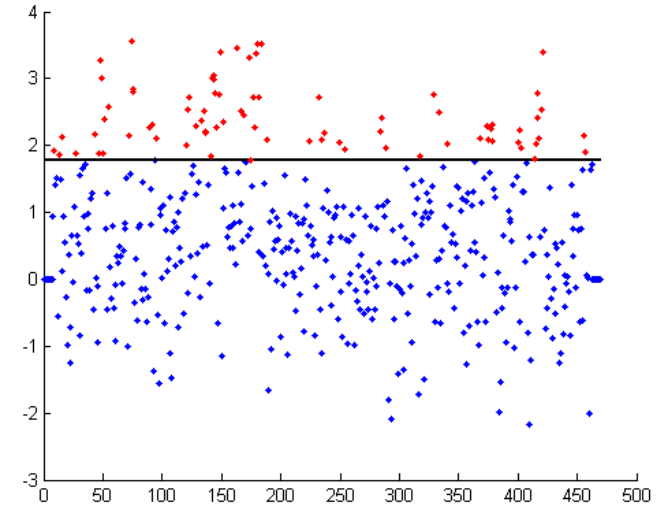
Second example : Bladder Cancer (Stransky et al.)

Chromosome 3









Some conclusions

Improvements

- ⋆ Break point detection for the number of clusters
- ⋆ Efficient control of Type I error
- ⋆ Study of the bias of the ρ_0 estimator

Toward complete model-based clustering

- ⋆ Maximum likelihood method
- ⋆ Adapted to more complex situations (long range correlations)
- ⋆ Computational cost may be prohibitive...