



**HAL**  
open science

# A statistical approach to detect chromosomal domains of gene expression

Tristan Mary-Huard, Stephane Robin

► **To cite this version:**

Tristan Mary-Huard, Stephane Robin. A statistical approach to detect chromosomal domains of gene expression. International Federation of Classification Societies 2009 Conference, Mar 2009, Dresde, Germany. 14 p. hal-01197555

**HAL Id: hal-01197555**

**<https://hal.science/hal-01197555>**

Submitted on 6 Jun 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# A statistical approach to detect chromosomal domains of gene expression

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

UMR AgroParisTech/INRA 518



We consider microarray data from cancer experiments.

For normal tissues :

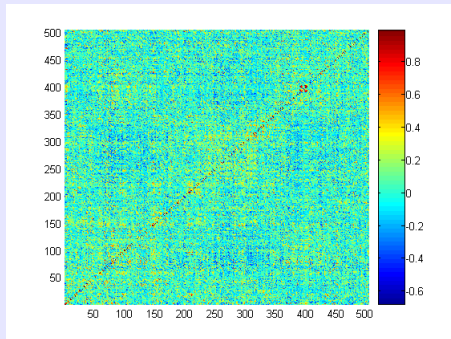
- ★ Two copies of each gene,
- ★ Correlations between expression profiles of neighbor genes is low.

For tumorous tissues :

- ★ Amplification/Deletion may occur (CGH microarray)  
⇒  $k$  copies of a gene, with  $k = 0, \dots, 13, \dots$ ,
- ★ Small regions of genes with highly correlated expression  
⇒ 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  $X^1, \dots, X^p$  be an ordered sequence of variables, that breakdowns into  $K$  clusters of variables  $\mathcal{C}_1, \dots, \mathcal{C}_K$ , such that

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

$A^k$  : cluster variable,  $V(A^k) = \sigma_k^2 I_n$

$E^j$  : residual variable,  $V(E^j) = \sigma^2 I_n$

$\text{cov}(A^k, E^j) = 0$ ,  $\text{cov}(A^k, A^{k'}) = 0$ ,  $\text{cov}(E^j, E^{j'}) = 0$

- ★ In the following, we assume that variables are standardized.
- ★ We note  $\rho_k$  the correlation between 2 variables in cluster  $\mathcal{C}_k$ .
- ★ 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  $\mathcal{C}_1, \dots, \mathcal{C}_K$ ),
- ★ Test for each cluster whether  $\rho_k = \rho_0$  or  $\rho_k > \rho_0$ .

## Loss for clusters

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

$$\star \hat{A}^k = \frac{1}{p_k} \sum_{j=\ell+1}^{\ell+p_k} X^j = \bar{X}^{(k)} \text{ (BLUP)}$$

$$\Rightarrow L(X^{(k)}) = p_k - \sum_{j=\ell+1}^{\ell+p_k} \text{cov}(X^j, \bar{X}^{(k)})$$

## Optimal clustering

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

$$\mathcal{C}^* = \underset{\mathcal{C}}{\text{Argmin}} L(X^1, \dots, X^p)$$

$$= \underset{\mathcal{C}}{\text{Argmin}} \sum_{k=1}^K L(X^{(k)}) .$$

# Clustering Algorithm

## Exhaustive search

- ★ Variables are ordered  $\Rightarrow$  clusters of adjacent variables only
  - ★ The loss to optimize is additive on clusters
- $\Rightarrow$  Optimal solution can be found using Dynamic Programming

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

## Heuristic search

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

- $\Rightarrow$  Use of a constrained version of Hierarchical Clustering  
Algorithm

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

Distance between clusters :

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

# Estimation of the number of clusters

## Objective :

Find the break in the slope of the clustering curve.

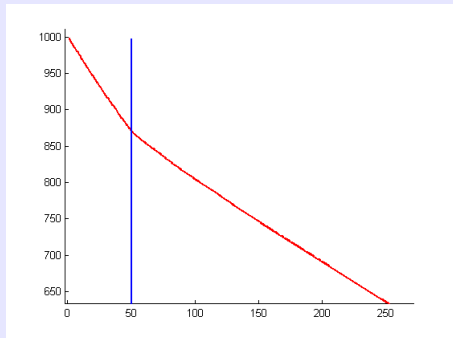
Same objective as for adaptive penalty in model selection context

Lavielle (2003)

## Choice of $K$

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

$$\hat{K} = \underset{K}{\operatorname{Argmin}} \{ \beta_K > S \}$$





# Testing the clusters

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

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

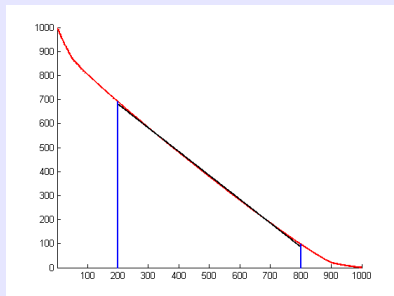
$$\begin{aligned}\bar{X}_i^{(k)} &= \frac{1}{\rho_k} \sum_{j \in \mathcal{C}_k} X^j \sim \mathcal{N} \left( 0, \frac{1}{\rho_k} (1 + (\rho_k - 1)\rho_k) \right) \\ \Rightarrow \hat{V}(\bar{X}^{(k)}) &= \frac{1}{n} \sum_i^n \left( \bar{X}_i^{(k)} - \overline{\bar{X}^{(k)}} \right)^2 \sim \frac{(1 + (\rho_k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

$$\Rightarrow n \left( \rho_k - L(X^{(k)}) \right)_{\mathcal{H}_0} \sim (1 + (\rho_k - 1)\rho_0) \chi_{(n-1)}^2$$

**Reject  $\mathcal{H}_0$**  if  $n \left( \rho_k - L(X^{(k)}) \right) > (1 + (\rho_k - 1)\hat{\rho}_0) \chi_{n-1, 1-\alpha}^2$

A 3-phase clustering curve :

- ★ Clusterings in correlated regions
- ★ Clusterings in  $\mathcal{H}_0$  regions
- ★ Clustering between regions



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

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

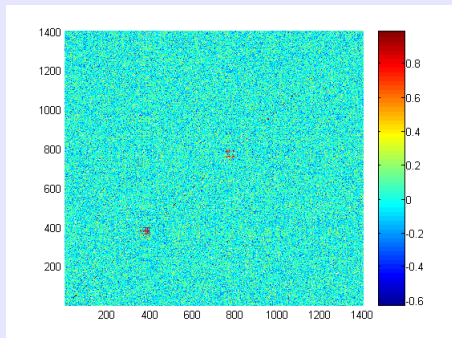
$\Rightarrow \rho_0$  may be estimated with the slope estimate of regression  
between  $b_{low}$  and  $b_{up}$  .

# First example : Simulated data

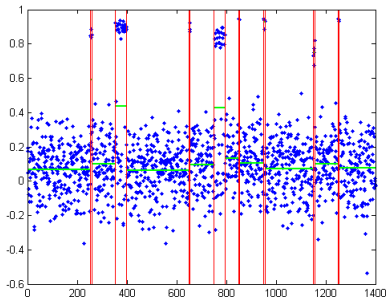
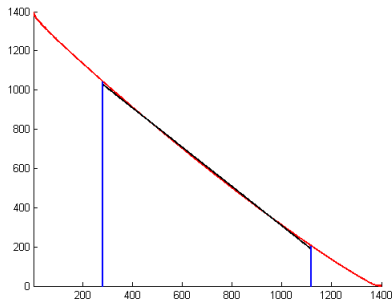
50 observations, 1400 variables  
Background correlation :  $\rho_0 = 0$

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  $\rho_0$  by regression,  $\hat{\rho}_0 = 0$ .

Right :

Green : average correlation with the aggregated variable

Blue : correlation between a gene and its aggregated variable

Number of clusters = 17, correlated : 8

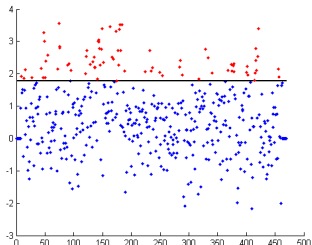
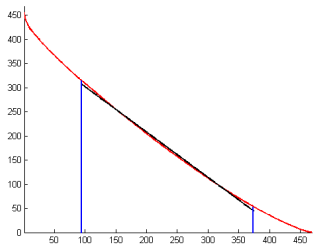
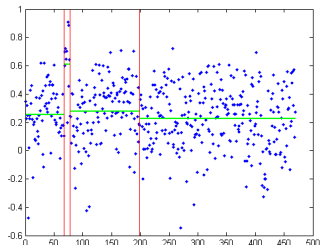
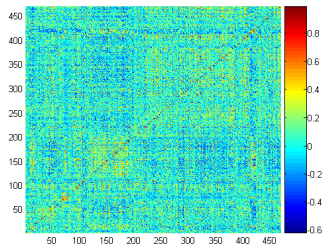
# Identified regions

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



## Improvements

- ★ Break point detection for the number of clusters
- ★ Efficient control of Type I error
- ★ Study of the bias of the  $\rho_0$  estimator

## Toward complete model-based clustering

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