PCA for Point Processes

Franck Picard*, Victor Panaretos †, Vincent Rivoirard‡, Angelina Roche‡

*Laboratoire Biologie et Modélisation de la Cellule, CNRS ENS-Lyon
† Institute of Mathematics, École Polytechnique Fédérale de Lausanne
‡ CEREMADE, Université Paris-Dauphine

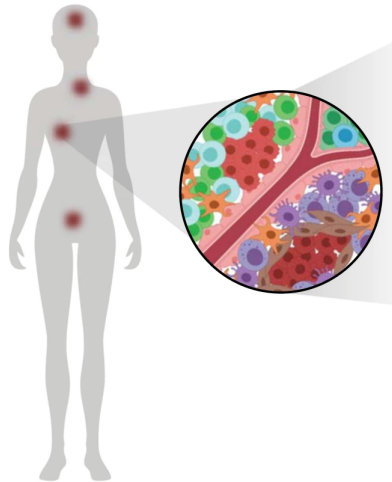
Séminaire de modélisation du vivant, Mars 2026

franck.picard@ens-lyon.fr

Outline

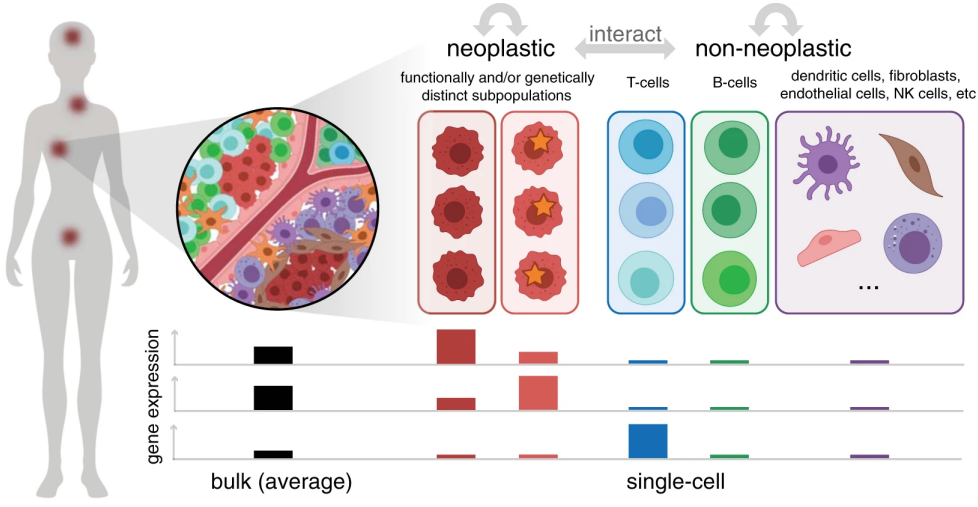
The cellular scale of biology

- Cells are the basic unit of structure and function in living organisms
- Cells are characterized by their 'types' that are diverse
- Physiology emerge thanks to complex interactions between different cell-types



[1]

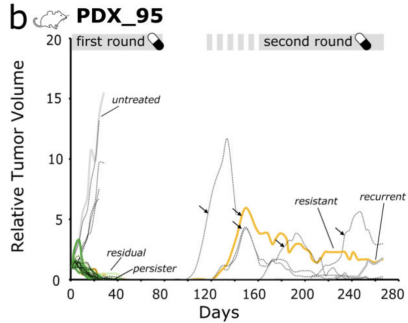
From bulk to distributions of molecular features

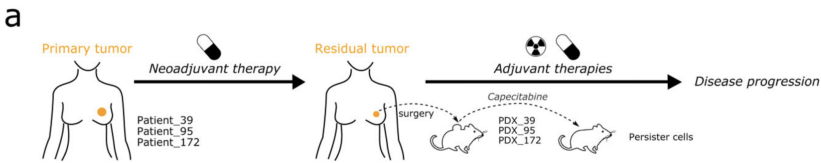


[1]

ChemoResistance in Triple Negative Breast Cancer

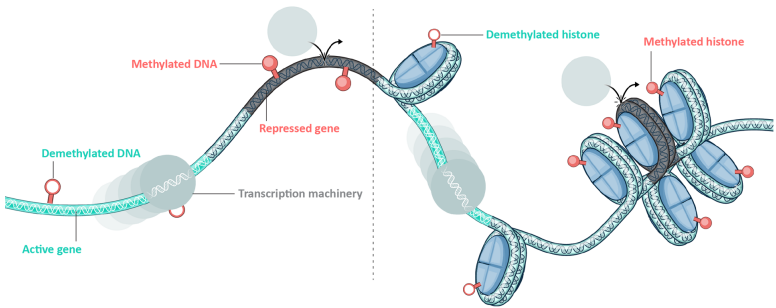
- Emergence of resistant phenotypes is a multi-step process
- After drug insult only a pool of drug-tolerant persister cells manage to tolerate the treatment and survive.
- Reservoir from which drug-resistant cells can ultimately emerge.





Gene Regulations by Chromatin Modifications

- Chromatin is the molecule that makes chromosomes
- Chromatin modifications can modulate gene activity
- ChemoResistance is associated with chromatin modifications



https://tunetx.com/

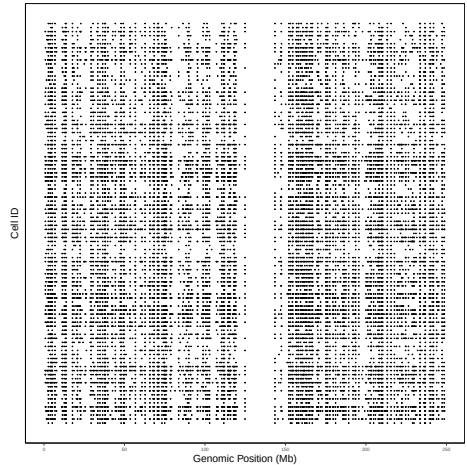
Single-Cell Chromatin Modification Data

- The data contain the positions of chromatin modifications along a chromosome (1D)
- Each cell is characterized by a collection of random positions

$$N_i = (T_1^i, T_2^i, \ldots), \quad T_1^i < T_2^i < \ldots$$

- We observe n replicated Point Processes:

$$(N_1, \ldots, N_n)$$



[3]

Description of the data

- A Point Process is described thanks to the intensity of the process

$$\mathbb{P}(dN_t = 1 \mid \mathcal{F}^{t-}) = w(t)$$

- Homogeneous Poisson Process:

$$w(t) = w_0$$

- The intensity can depend on the past occurrences (Hawkes)

$$w(t) = w_0 + \sum_{T \in N} h(t - T)$$

The Cumulative Mass Function of a Point Process

- Equivalent representation through the counting process

$$F(t) = \sum_{T \in N} 1_{\{t \leq T\}}$$

- This is the cumulative mass function of the process (CMF)
- In the Poisson case

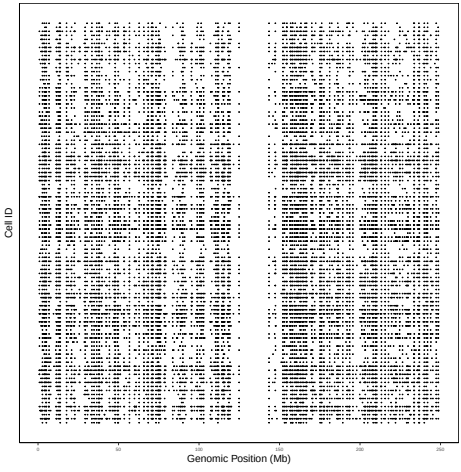
$$\mathbb{E}\big[F(t)\big] = w_0 t$$

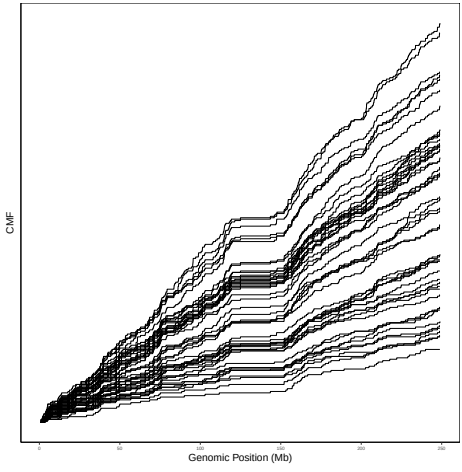
- Hawkes process with $h(t) = \alpha e^{-\beta t}$

$$\mathbb{E}\big[F(t)\big] = w_0 \frac{\beta t}{\beta - \alpha}$$

- The CMF connects Point Processes models to functional models

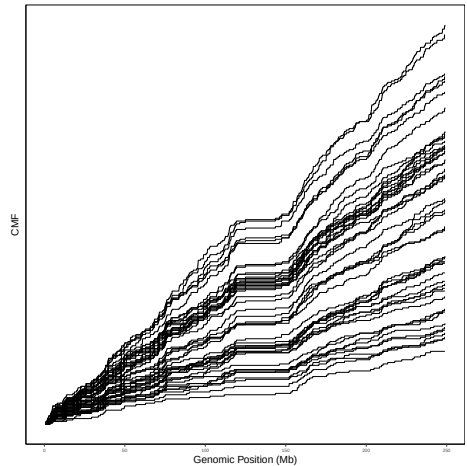
From Point Processes to Functional Data





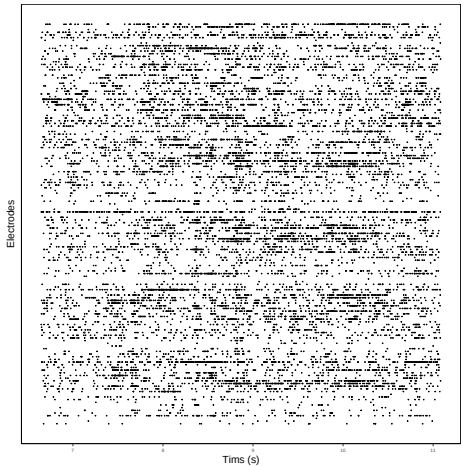
Objectives of the analysis

- How to describe the variability of chromatin modifications among cells ?
- How to represent the spatial variability ?
- Can we find atypical cells ? clusters ?
- Is there a reservoir of persister cells ?



General Setting: PCA for Point Processes

- Many application fields generate data that can be modelled as replicated Point Processes
- NeuroSciences : spike train data [2]
- GeoSciences : earthquakes data
- How to perform PCA on these data ?



Outline

Replicated Point Processes

- We observe n point processes on $[0,1]$:

$$\big(N_1, \ldots, N_n\big)$$

- For process N_i we define the random measure

$$\Pi_i(B) = \sum_{T \in N_i} \mathbf{1}_{\{T \in B\}} = \mathrm{Card}(N_i \cap B), \qquad B \in \mathcal{B}$$

- We observe a $n-$sample of random measures on $([0,1], \mathcal{B})$

$$\big(\Pi_1, \ldots, \Pi_n\big).$$

- We denote by W the first moment of Π:

$$W(B) = \mathbb{E}\big[\Pi(B)\big], \qquad B \in \mathcal{B}.$$

Examples of Point Processes

- The simplest model is the Poisson Process with intensity w

$$W(B) = \mathbb{E}\big[\Pi(B)\big] = \int_B w(u)du$$

- In this case:

$$\Pi(B) \sim \mathcal{P}\Big(W(B)\Big).$$

- If the process is homogeneous, the accumulation rate is constant over time

$$W(B) = w_0 \times |B|$$

- In the following, we do not make any assumption on the model for Π.

A Signed-Measure Model

- Objective: investigate the variations of points occurrences around W
- We consider the signed measure

$$\Delta_i(B) = \Pi_i(B) - W(B)$$

- Then we define a second-order moment for Δ using the signed measure of covariance

A Second-Order Measure of Covariance

- The tensor product for two measures μ and ν on $([0,1], \mathcal{B})$ is:

$$(\mu \otimes \nu)(B \times B') = \mu(B)\nu(B'), \qquad B, B' \in \mathcal{B}.$$

- For all $B \times B' \in \mathcal{B} \otimes \mathcal{B}$

$$\begin{aligned} C_\Delta(B \times B') &= \mathrm{Cov}\big[\Pi(B), \Pi(B')\big] \\ &= \mathbb{E}\left[\sum_{T \in N, T' \in N} \mathbf{1}_{\{(T,T') \in B \times B'\}}\right] - (W \otimes W)(B \times B') \end{aligned}$$

- In the case of the homogeneous Poisson Process

$$C_\Delta(B \times B') = w_0 \times |B \cap B'|$$

Mercer theorem for C_Δ

- Theorem (Picard F., Panaretos V., Rivoirard, V., Roche A.)
- Suppose

$$\mathbb{E}\big[\Pi([0,1])^2\big] < +\infty$$

- There exists a sequence of eigenvalues $\big(\lambda_j\big)_j$ and measures $\big(\mu_j\big)_j$ such that:

$$C_\Delta = \sum_{j=1}^{\infty} \lambda_j \times \big(\mu_j \otimes \mu_j\big).$$

- Uniform convergence on Sobolev balls of regularity 2:

$$\left\| \sum_{j=1}^{J} \lambda_j \times \big(\mu_j \otimes \mu_j\big) - C_\Delta \right\|_{\mathcal{H}^{-2}} \xrightarrow[J\to\infty]{} 0.$$

Principal Measures and the K-L decomposition

- Principal measures $(\mu_j)_j$ contain the dynamics of the process
- Karhunen-Loève decomposition for random measures:

$$\Delta_i = \sum_{j=1}^{\infty} \sqrt{\lambda_j} \times \xi_{ij} \times \mu_j$$

- The sequence $\{\xi_j\}_{j\geq 1}$ is called *the scores associated to the measure* Δ.
- Each individual measure Δ_i can be approximated from the K-L decomposition:

$$\Pi_i(B) \approx W(B) + \sum_{j=1}^{J} \sqrt{\lambda_j} \times \xi_{i,j} \times \mu_j(B), \qquad B \in \mathcal{B}$$

Outline

From Point Processes to functional data

- We need to get back to the data for statistical inference
- Consider the cumulative mass function

$$F_\Delta(t) = \Delta([0,t]) = \Pi([0,t]) - W([0,t])$$

- A sample of Point Processes gives us a sample of centered random functions on $[0,1]$

$$(\Pi_1, \ldots, \Pi_n) \longrightarrow (F_{\Delta_1}, \ldots, F_{\Delta_n})$$

- Deploy the fPCA framework on cumulative mass functions !

Covariance Kernel

- The covariance kernel of these functions is, $s, t \in [0, 1]$

$$\begin{aligned} K_\Delta(s,t) &= \mathrm{Cov}(F_\Delta(s), F_\Delta(t)) \\ &= C_\Delta([0,s] \times [0,t]). \end{aligned}$$

- The kernel K_Δ is associated with the integral operator

$$\Gamma_\Delta = \mathbb{E}\big[F_\Delta \otimes F_\Delta\big]$$

- The kernel K_Δ between CMFs is a positive semi-definite kernel for which the Mercer theorem applies

From the fPCA to the principal measures

- There exists a basis $(\eta_j)_{j\geq 1}$ of $\mathbb{L}^2([0,1])$, a sequence $(\lambda_j)_{j\geq 1}$ of positive numbers such that

$$K_\Delta(s,t) = \sum_{j=1}^{\infty} \lambda_j \times \eta_j(s) \times \eta_j(t) \qquad s,t \in [0,1].$$

- The eigenfunctions are related to the principal measures of C_Δ such that

$$\eta_j(t) = \mu_j([0,t]) = F_{\mu_j}(t) \qquad t \in [0,1].$$

From the fPCA to the principal measures

- The principal functions of K_Δ are the CMF of the principal measures

$$K_\Delta(s,t) = \sum_{j=1}^{\infty} \lambda_j \times F_{\mu_j}(s) \times F_{\mu_j}(t) \qquad s,t \in [0,1].$$

- Similarly for the integral operator:

$$\Gamma_\Delta = \sum_{j=1}^{\infty} \lambda_j \times F_{\mu_j} \otimes F_{\mu_j}$$

- Principal functions $(F_{\mu_j})_j$ contain the dynamics of the Cumulative Mass Functions

Eigen Functions and the K-L decomposition

- Karhunen-Loève decomposition for random measures writes

$$\Pi_i([0,t]) \approx W([0,t]) + \sum_{j=1}^{J} \sqrt{\lambda_j} \times \xi_{i,j} \times F_{\mu_j}(t)$$

- The scores are defined such as:

$$\xi_{ij} = \lambda_j^{-1/2} \langle F_{\mu_j}, F_{\Delta_i} \rangle$$

- The processes $\left(N_1, \ldots, N_n\right)$ can be represented with their new coordinates $\left(\xi_{ij}\right)_{ij}$

Outline

Empirical measures

- Based on the observed measures:

$$(\Pi_1, \ldots, \Pi_n)$$

- The average measure :

$$\widehat{W} = \frac{1}{n} \sum_{i=1}^{n} \Pi_i,$$

- The observed signed measure :

$$\widehat{\Delta}_i = \Pi_i - \widehat{W}$$

- The observed empirical cumulative mass functions:

$$\widehat{F}_{\Pi_i}(t) = \sum_{T \in N_i} 1_{\{t \leq T\}}$$

Empirical quantities

- The empirical covariance measure:

$$\widehat{C}_{\widehat{\Delta}}(B \times B') = \frac{1}{n} \sum_{i=1}^{n} \sum_{T, T' \in N_i} \mathbf{1}_{\{(T,T') \in B \times B'\}} - \widehat{W}(B) \times \widehat{W}(B'), \quad B, B' \in \mathcal{B},$$

- This provides the empirical kernel:

$$\widehat{K}_{\widehat{\Delta}}(s,t) = \widehat{C}_{\widehat{\Delta}}([0,s] \times [0,t]), \quad s,t \in [0,1]$$

- The empirical integral operator writes:

$$\widehat{\Gamma}_{\widehat{\Delta}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{F}_{\widehat{\Delta}_i} \otimes \widehat{F}_{\widehat{\Delta}_i}$$

Observed integral operator

- Consider all occurrences sorted in non-decreasing order:

$$\mathcal{T} = \bigcup_{i=1}^{n} N_i \bigcup \{0; 1\} = \Big\{0, T_1, T_2, \ldots, 1\Big\}$$

- We define the histogram system associated to this grid

$$e_\ell(t) = \frac{1}{\sqrt{T_\ell - T_{\ell-1}}} 1_{[T_{\ell-1}; T_\ell)}(t), \quad \ell = 1, \ldots, |\mathcal{T}|,$$

- We set

$$\widehat{G}_{\widehat{\Delta}} = \Big(\langle \widehat{\Gamma}_{\widehat{\Delta}} e_\ell, e_{\ell'} \rangle \Big)_{\ell, \ell'}.$$

Eliciting eigen elements

- Some simplifications are welcome:

$$F_{\widehat{W}}(T_{\ell-1}) = (\ell - 1)/n$$

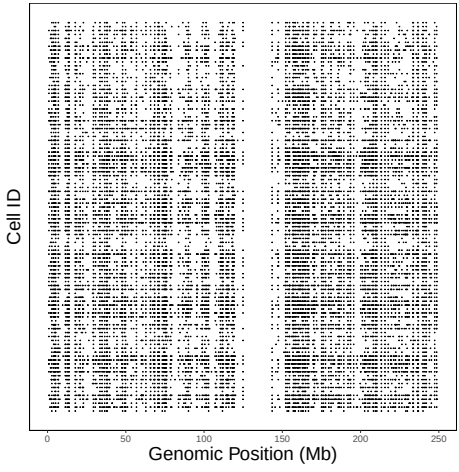
- We can compute the general term $\left(\widehat{G}_{\widehat{\Delta}}\right)_{\ell,\ell'}$ easily

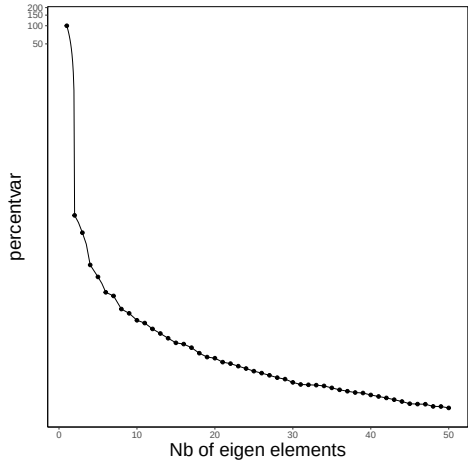
$$\frac{\sqrt{(T_\ell - T_{\ell-1})(T_{\ell'} - T_{\ell'-1})}}{n} \times \sum_{i=1}^{n} \left(F_{\Pi_i}(T_{\ell'-1}) F_{\Pi_i}(T_{\ell-1}) - \frac{\ell - 1}{n} \frac{\ell' - 1}{n} \right).$$

- Eigen elements of matrix $\widehat{G}_{\widehat{\Delta}}$ provide $\left(\widehat{\lambda}_j\right)_j$, $\left(\widehat{\eta}_j\right)_j$, $\left(\widehat{\xi}_{ij}\right)_j$

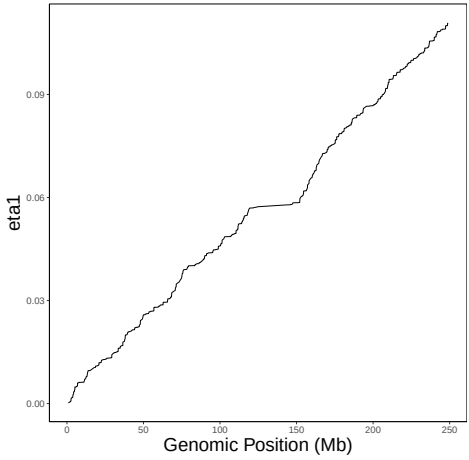
Outline

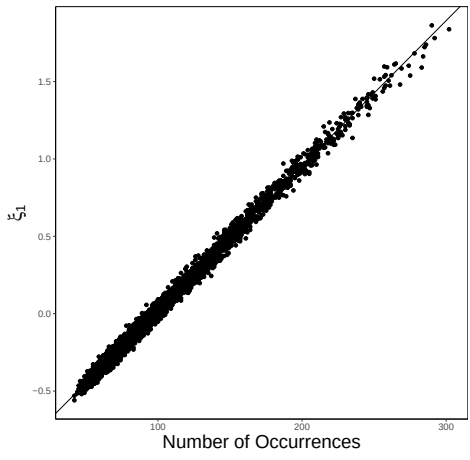
Summary of the data



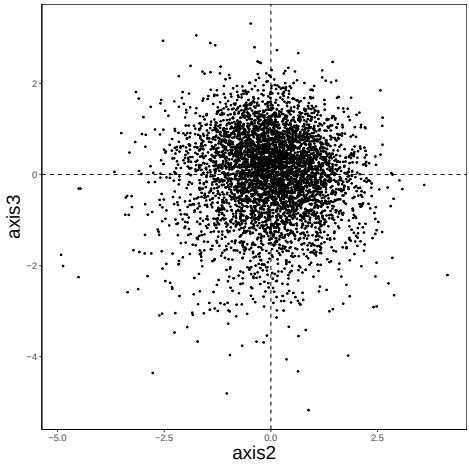


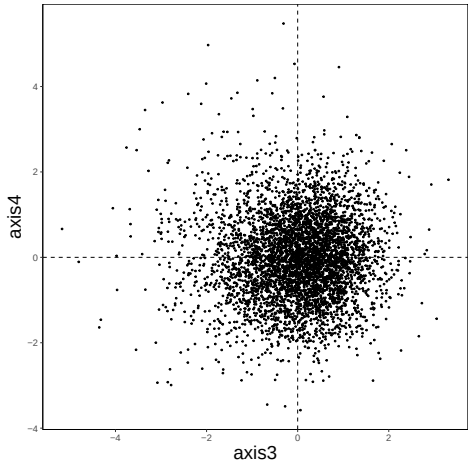
The first axis is correlated to the number of points



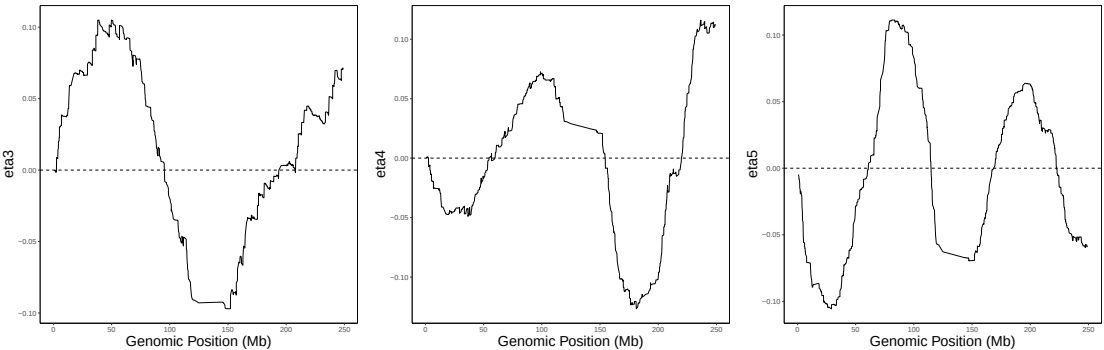


Inspecting Scores





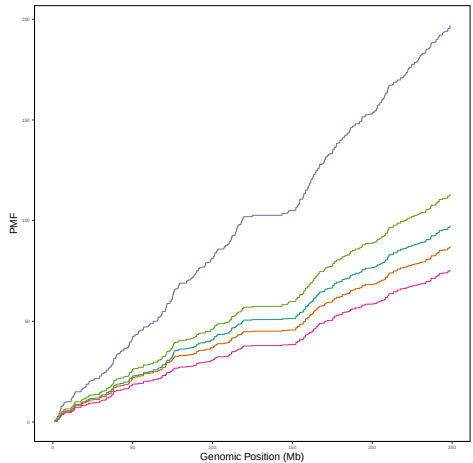
Inspecting Eigenfunctions



Clustering based on scores

- k-means clustering based on $\widehat{\xi} ij$
- Obtain a partition $\widehat{Z}_{ik}$
- Represent

$$\widehat{F}_{\Pi}^{k}(t) = \frac{1}{n_k} \sum_{i=1}^{n} \widehat{Z}_{ik} \times \Pi_i([0, t])$$

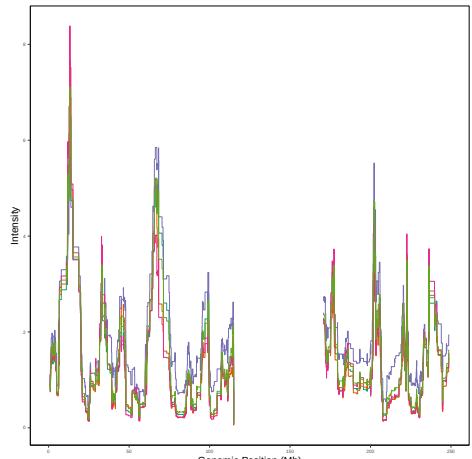


Clustering based on scores

- k-means clustering based on $\widehat{\xi}_{ij}$
- Obtain a partition $\widehat{Z}_{ik}$
- Represent the intensity by group

$$\frac{\partial}{\partial t}\widehat{F}^k_{\Pi}(t)$$

- Some cells show particular accumulations of chromatin modifications along chromosome 1 !



Outline

A new framework to analyses repeated Point Process

- We provide a new framework to handle, visualize and analyse point process data
- Existence of equivalent of the Karhunen-Loève decomposition for random measures
- Convergence results and convergence rates for the estimators of PC measures
- Fine study of principal measures for Poisson and Hawkes processes.
- Preprint and package soon available !

Perspectives and Extensions

- Consider spatial processes
- Consider multivariate processes
- Consider the supervised setting
- Any suggestion ?

References

[1] J. Fan, K. Slowikowski, and F. Zhang. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Exp Mol Med*, 52(9):1452–1465, Sep 2020.

[2] G. Gauvain, H. Akolkar, A. Chaffiol, F. Arcizet, M. A. Khoei, M. Desrosiers, C. Jaillard, R. Caplette, O. Marre, S. Bertin, C. M. Fovet, J. Demilly, V. Forster, E. Brazhnikova, P. Hantraye, P. Pouget, A. Douar, D. Pruneau, J. Chavas, J. A. Sahel, D. Dalkara, J. Duebel, R. Benosman, and S. Picaud. Optogenetic therapy: high spatiotemporal resolution and pattern discrimination compatible with vision restoration in non-human primates. *Commun Biol*, 4(1):125, Jan 2021.

[3] J. Marsolier, P. Prompsy, A. Durand, A. M. Lyne, C. Landragin, A. Trouchet, S. T. Bento, A. Eisele, S. Foulon, L. Baudre, K. Grosselin, M. Bohec, S. Baulande, A. Dahmani, L. Sourd, E. , A. V. Salomon, E. Marangoni, L. , and C. Vallot. H3K27me3 conditions chemotolerance in triple-negative breast cancer. *Nat Genet*, 54(4):459–468, Apr 2022.