Detection of spatial cluster using nearest neighbour distance

Avner Bar-Hen & Mathieu Emily

Paris: October 2011

What is it?

Point process: $X_1, \ldots, X_n$ a set of n points observed in a window W of $\mathbb{R}^2$ (position and n are random).

Example of statistic: number of points within a ball of radius r, distance between points, and so on ...

Statistical question is well defined

Clusters: areas with high concentration of points

first order statistic (eg intensity too high) ? second order (distance between points) ?

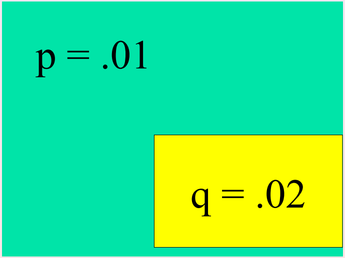
Statistical question is not defined

Few examples

- Epidemiology
- Ecology
- Imagery
- Earthquake
- Mines
- Astrophysic
- etc...

Scan statistic

- $L(Z, p, q)$ the likelihood of an area Z such that the probability of having a point **within** Z is q and the probability of having a point **outside** Z is p
- $H_0 : p = q$ versus $H1 : q > p$
- $\lambda = \frac{Sup_{Z \in W, q > p} L(Z, p, q)}{Sup_{Z \in W, p = q} L(Z, p, q)}$
- simulate λ under H_0
- questions : which Z, # clusters, multiple tests , comput. burden



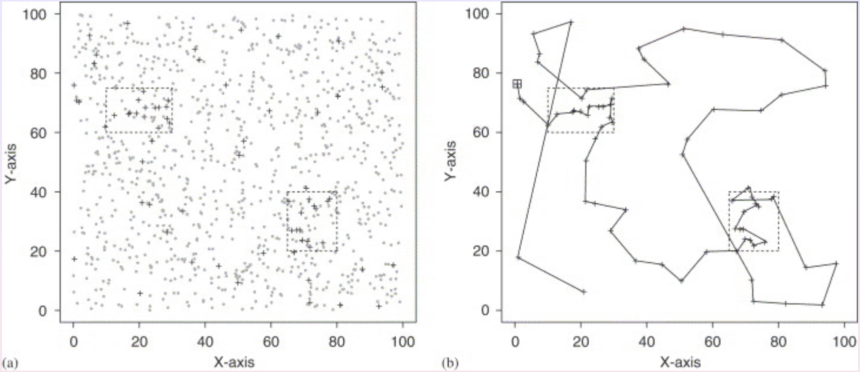
Example of scan statistic: Bernoulli model

Sane/Unsane. n_W: total number of cases within W and n_Z number of cases within $Z \in W$

- Under H_0 $N(B) \sim \mathscr{B}(\mu(B), p)$ for all B
- Under H_1 $N(B) \sim \mathscr{B}(\mu(B), p)$ for all $B \in Z$ and $N(B) \sim \mathscr{B}(\mu(B), q)$ for all $B \in Z^c$
- $L(Z, p, q) = p^{n_z}(1-p)^{\mu(Z)-n_z} q^{n_W - n_z}(1-q)^{\mu(W)-\mu(Z)-(n_W - n_Z)}$
- For Z fixed $L(Z) = \max_{p>q} L(Z, p, q)$ then: $p = \frac{n_Z}{\mu(Z)}$ and $q = \frac{n_W - n_Z}{\mu(W)-\mu(Z)}$
- and $L_0 = \left(\frac{n_W}{\mu(W)}\right)^{n_W} \left(\frac{\mu(W)-n_W}{\mu(W)}\right)^{\mu(W)-n_W}$
- $$\lambda = \frac{\sup_{Z \in W} L(Z)}{L_0}$$

L_0 obtained with simulations

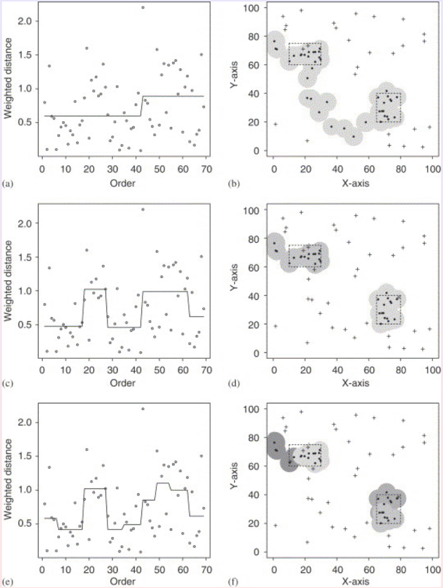
Transform $\mathbb{R}^2$ to $\mathbb{R}$ (Ch. Demattei)



- $(X_{n,j})_{j=1,\ldots,n-1} \sim \mathcal{U}[0,1]$
- spacings: $U_{n,j} = n(X_{n,(j)} - X_{n,(j-1)})$ ($\sim \beta(1, n-1) \to \exp(1)$ when $n \to +\infty$)

$$d_k^w = d_x \times E_{H_0}(D_k | X_{(1)} = x_{(1)}, \ldots, X_{(k)} = x_{(k)})$$

Transform $\mathbb{R}^2$ to $\mathbb{R}$ (Ch. Demattei)



Proposal (1/2)

$D_i = \text{distance}(X_{(i)}, X_{(i+i)}) \quad 1 \leq i \leq n-1,$

$n-1$ vector of distances: $[D_1, \ldots, D_{n-1}]$

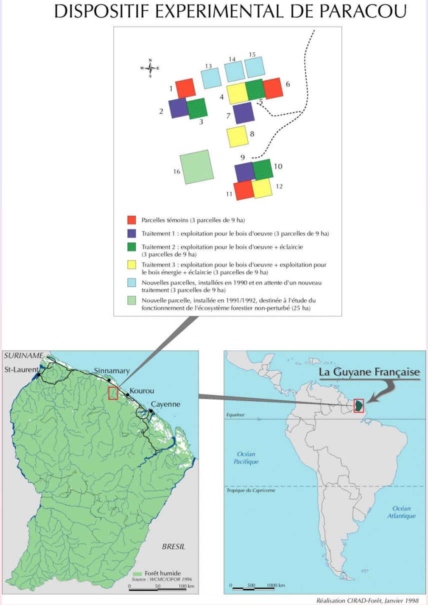
Probability that X_2 at a distance D_1 of X_1 : $\lambda\pi D_1^2$ (surface of $\mathscr{B}(X_1, D_1)$)

Probability that X_3 at a distance D_2 of X_2 : $\lambda\pi D_2^2$ (surface of $\mathscr{B}(X_2, D_2)$)

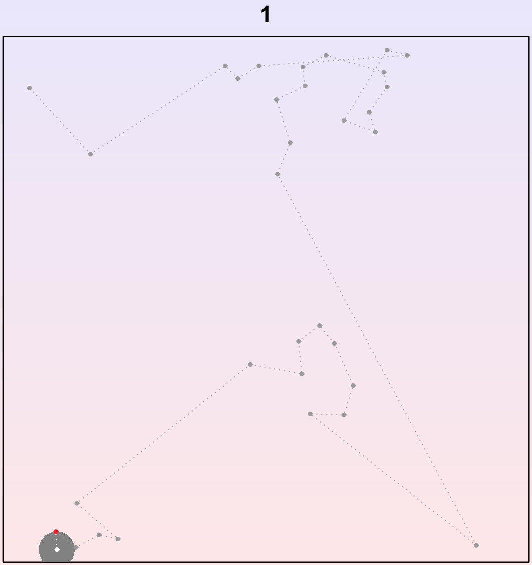
BUT Probability that X_3 at a distance D_2 of X_2 conditionally on X_1 : (surface of $\mathscr{B}(X_1, D_1)$) \ (surface of $\mathscr{B}(X_2, D_2)$)

Finally $[D_1, \ldots, D_{n-1}]$ becomes $[p_1, , \ldots, p_{n-1}]$, vector of probabilities

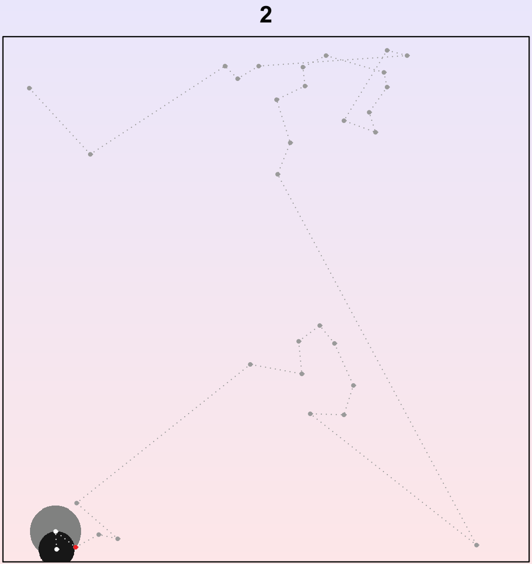
Illustration : Paracou



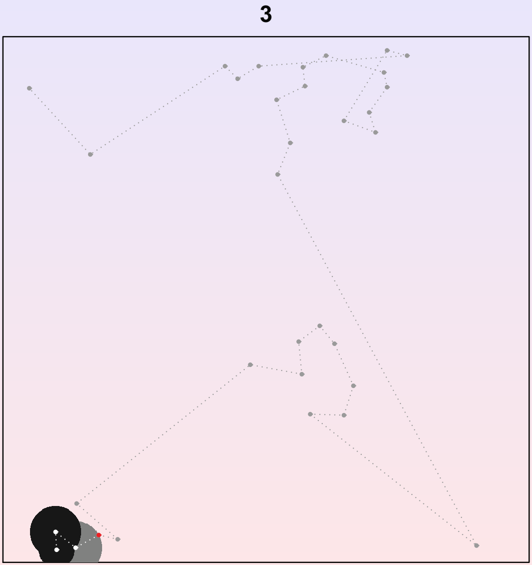
Proposal (1/2): illustration



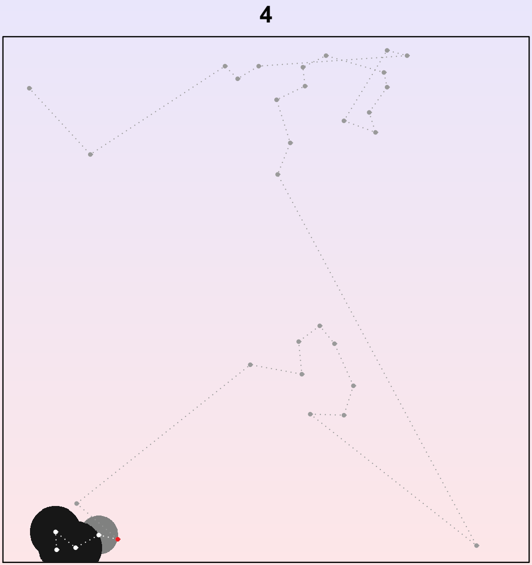
Proposal (1/2): illustration



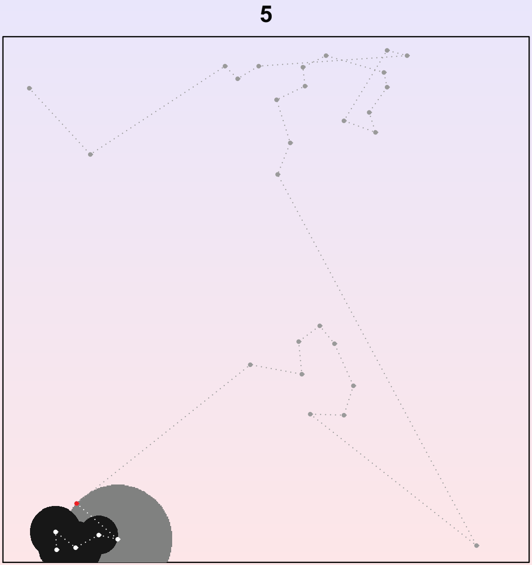
Proposal (1/2): illustration



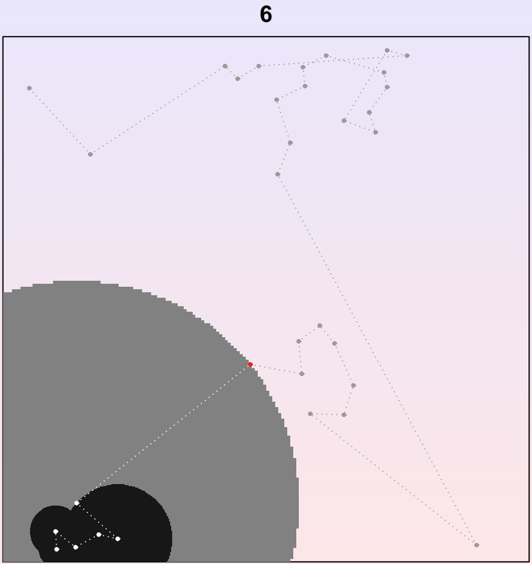
Proposal (1/2): illustration



Proposal (1/2): illustration



Proposal (1/2): illustration



Proposal (2/2) (Godehardt, 96)

For a given $d \in [0,1]$, connect X_i and X_j if $|X_i - X_j| \le d$

Let C_n be the number of components in a random interval graph $G_{n,d}$.

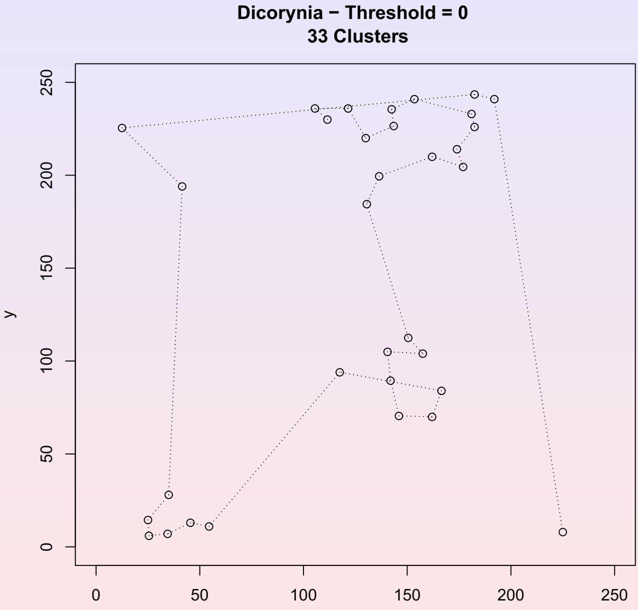
$$\mathbb{P}(C_n = r) = \sum_{j=r-1}^{\min(n-1,\lfloor 1/d \rfloor)} \binom{n-1}{j} \binom{j}{r-1} (-1)^{j+r-1} (1 - jd)^n$$

for $r = 1, 2, \ldots, \min(n-1, \lfloor 1/d \rfloor) + 1$.

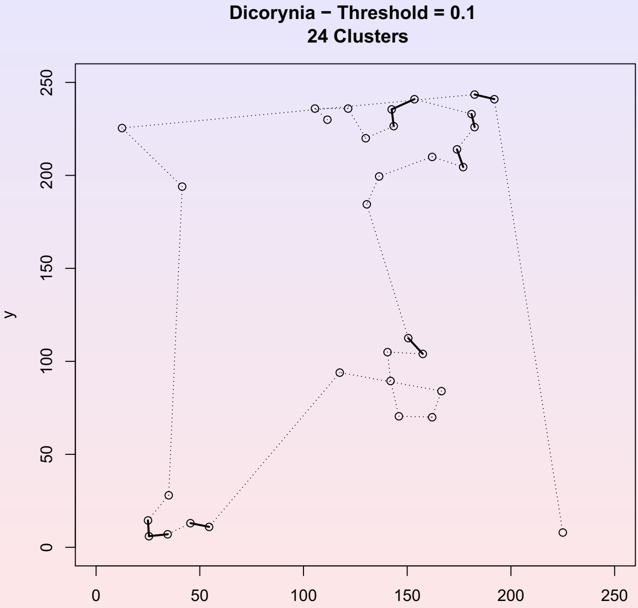
Expected number of components of size greater than m:

$$\sum_{k=m+1}^{n} \mathbb{E}(C_n^k) = \sum_{j=0}^{\min(m+1,\lfloor 1/d \rfloor)} \binom{m+1}{j} (-1)^j (1-jd)^n + (n-m) \sum_{j=0}^{\min(m,\lfloor 1/d \rfloor)-1} \binom{m}{j} (-1)^j (1-(j+1)d)^n$$

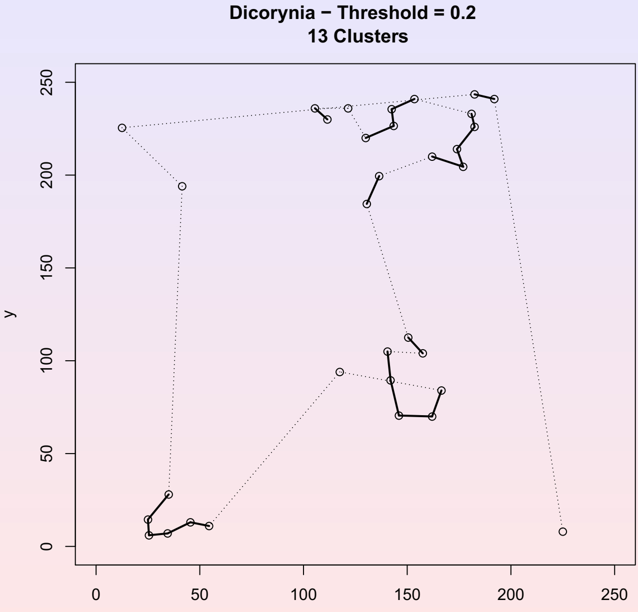
Proposal (2/2): illustration



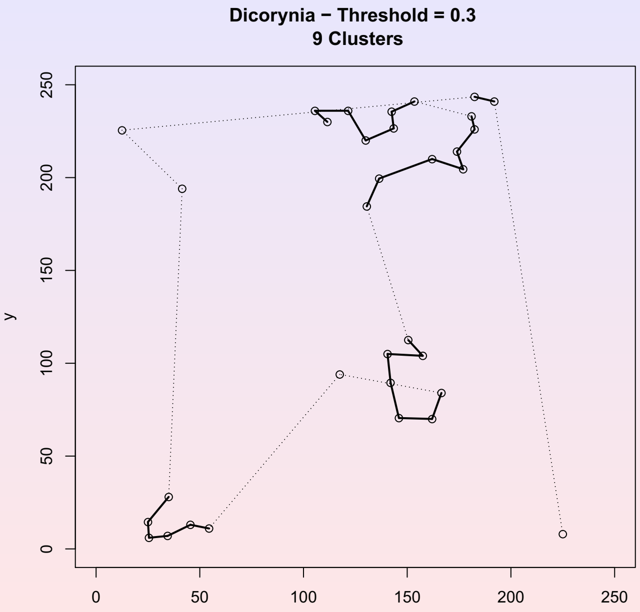
Proposal (2/2): illustration



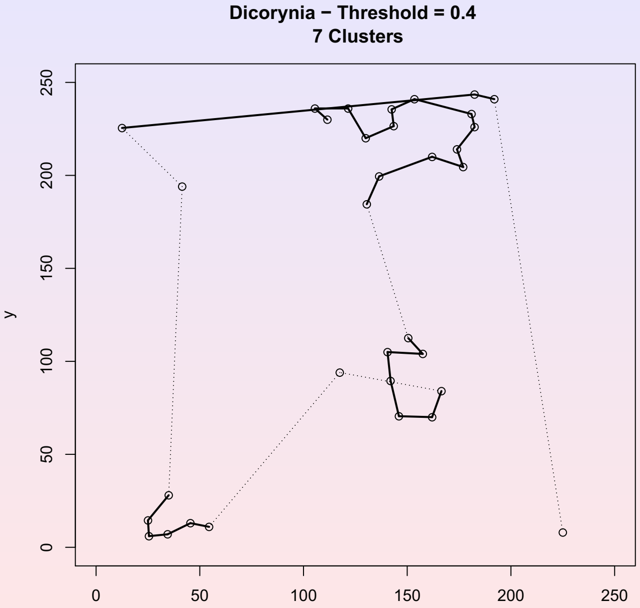
Proposal (2/2): illustration



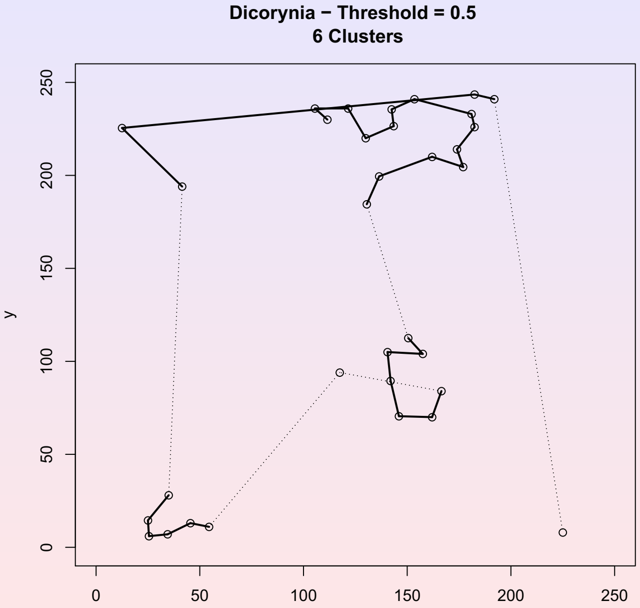
Proposal (2/2): illustration



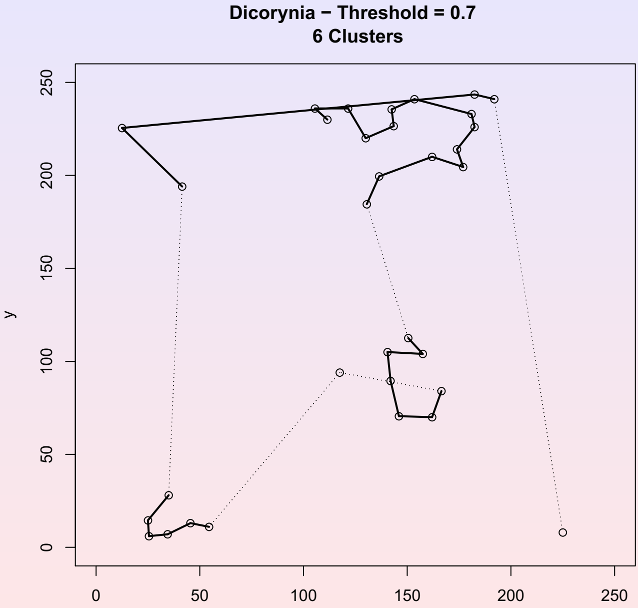
Proposal (2/2): illustration



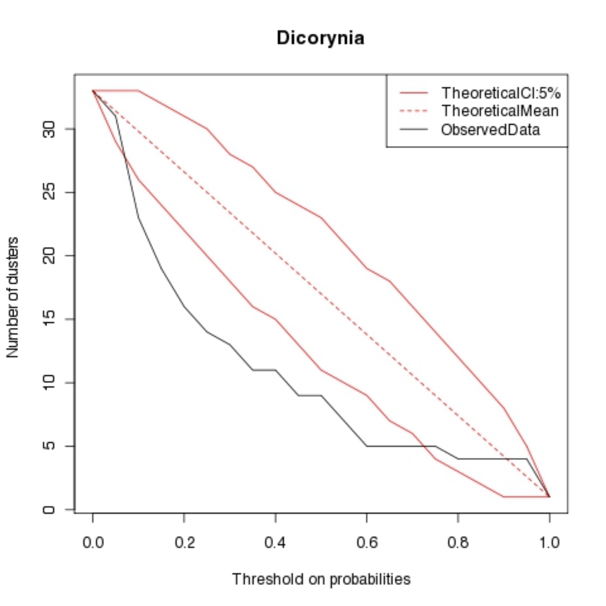
Proposal (2/2): illustration



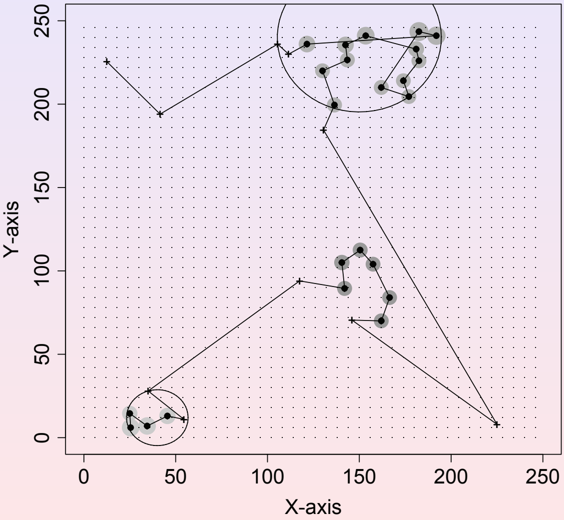
Proposal (2/2): illustration



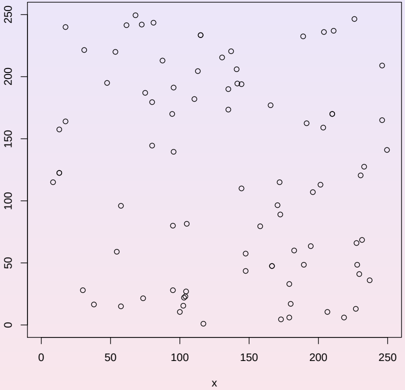
Proposal (2/2): illustration



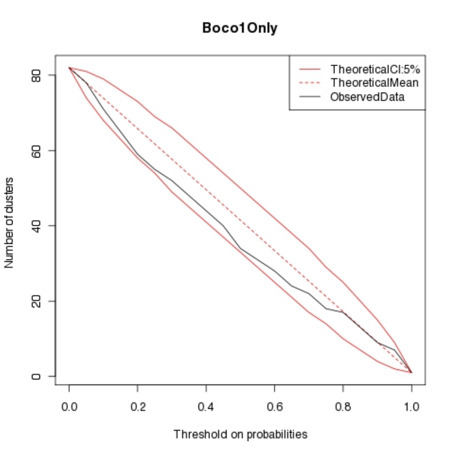
Angélique: scan statistic and Demattei's approach



Boco2

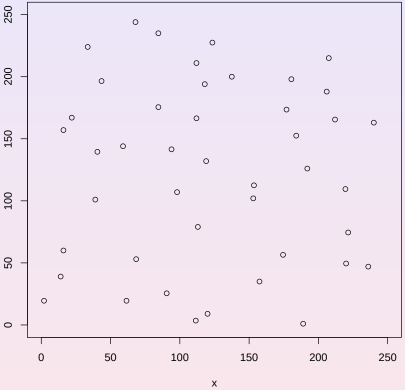


0
50
100
150
200
250
x

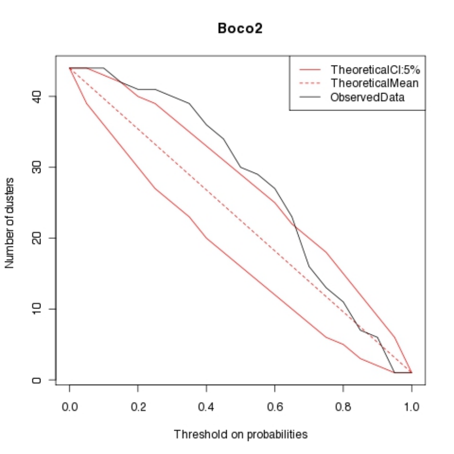


Boco1Only
TheoreticalCI:5%
TheoreticalMean
ObservedData
Number of clusters
Threshold on probabilities
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80

Boco7



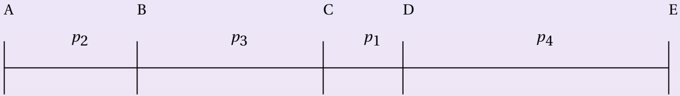
250
200
150
100
50
0
0
50
100
150
200
250
x



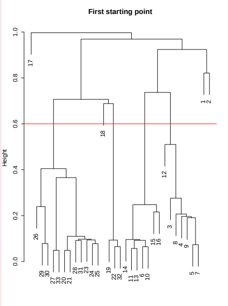
Boco2
TheoreticalCI:5%
TheoreticalMean
ObservedData
Number of clusters
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
Threshold on probabilities

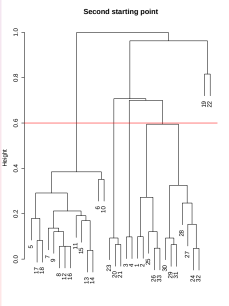
Stability of the procedure : importance of the first point ?

For a sequence $d_1 < d_2 < \ldots < d_n$, the connected components corresponds to a nested sequence of clusters (hierarchy)



Our proposal is equivalent to construct a hierarchical clustering based on minimum distance



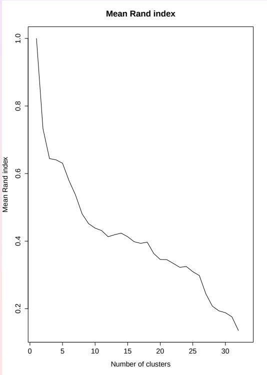


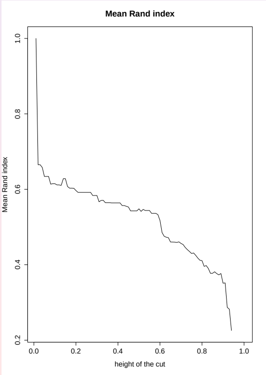
Stability of the procedure : importance of the first point ?

Comparison of the hierarchy based on the various starting points with Rand index :

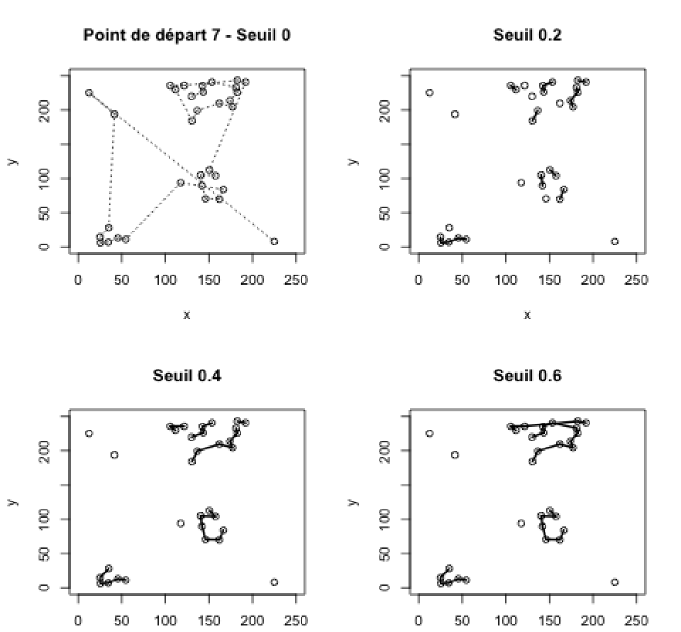
$$R = \frac{a+b}{\binom{n}{2}}$$

a: nb pairs in the same set in the two partitions ;
b: nb pairs not in the same set in the two partitions

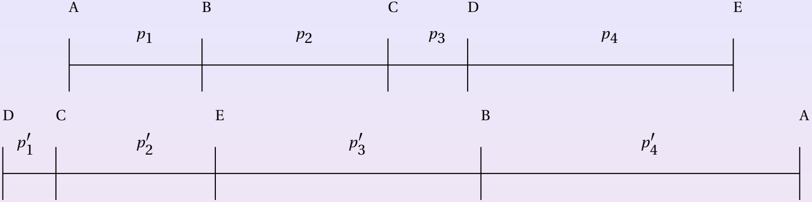




Yet Another Problem



Stability of the procedure : proposal



First idea: for each pair of points X, Y, mean over all paths of p_{XY}.

$$d(\mathrm{A},\mathrm{B}) = \frac{p_1 + p'_4}{2}\ ,\ d(\mathrm{C},\mathrm{E}) = p'_2$$

BUT unequal variance

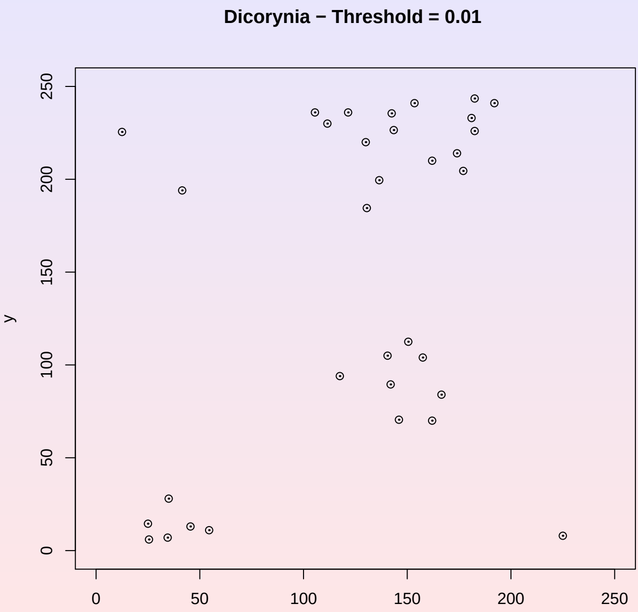
Second idea: for each pair of points X, Y, mean over all paths of connecting probability of X and Y

$$d(\mathrm{A},\mathrm{B}) = \frac{p_1 + p'_4}{2}\ ,\ d(\mathrm{C},\mathrm{E}) = \frac{p'_2 + \max(p_3, p_4)}{2}$$

Then connect X and Y if $d(\mathrm{X},\mathrm{Y}) < d$ (for a given d)

Resulting structure is no more a line but a graph

Evolution of the cluster with respect to the size



Law of the number of components for Erdös graph (with M. Koskas and N. Picard)

- Erdös' graph with n vertices and p the probability of having an edge
- connected components are sets of vertices with a path between all vertices of the component and no path with vertices outside the component
- $p_{k,n}$ probability of having k connected components among n vertices
- $p_{k,n} = \frac{1}{k} \sum_{l=1}^{n-(k-1)} \binom{n}{l} p_{1,l} p_{k-1,n-l} q^{l(n-l)}$
- $p_{1,n} = 1 - \sum_{k=2}^{n} p_{k,n}$
- $p_{k,n} = \frac{1}{k!} \sum_{\substack{\forall 1 \leq i \leq k,\, l_i \geq 1, \\ l_1+l_2+\cdots+l_k=n}} \binom{n}{l_1, l_2, \ldots, l_k} p_{1,l_1} p_{1,l_2} \ldots p_{1,l_k} q^{\sum_{1 \leq a < b \leq k} l_a l_b}.$
- $p_{1,n} = 1 - \sum_{d=2}^{n} \frac{1}{d!} \sum_{\substack{l_1+\cdots+l_d=n \\ l_i \geq 1}} \binom{n}{l_1, \ldots, l_d} p_{1,l_1} p_{1,l_2} \ldots p_{1,l_d} q^{\sum_{1 \leq a < b \leq d} l_a l_b}$

Related results

- Let K (the number of connected component) be a random variable taking integer values $1, \ldots, n$ with probability function defined by $p_{k,n}$, then:

$$\mathbb{E}(\mathrm{K}) = \sum_{l=1}^{n} \binom{n}{l} p_{1,l} q^{l(n-l)}$$

- $p'_{n,d}$ be the probability that the connected component including s is of size d: $p'_{n,d} = \binom{n-1}{d-1} p_{1,d} q^{d(n-d)}$
- Let D (the size of a component) be a random variable taking integer values $1, \ldots, n$ with probability distribution function defined by $p'_{n,d}$. Then

$$\mathbb{E}(\mathrm{D}^{-1})^{-1} = n / \mathbb{E}(\mathrm{K})$$

  Harmonic expectation of the size of a connected component taken at random is equal to the size of the graph divided by its expected number of connected components

Few practical remarks

- $p_{k,n}$ probability of having k connected components among n vertices

- $p_{k,n} = \frac{1}{k} \sum_{l=1}^{n-(k-1)} \binom{n}{l} p_{1,l} p_{k-1,n-l} q^{l(n-l)}$
- $p_{1,n} = 1 - \sum_{k=2}^{n} p_{k,n}$

- precision is an issue: difficult pour $n > 30$
- Symbolic calculus: computational time increases

What about isolates ?

$\mathrm{T}_{k,n,d}$ be the probability of having k connected components of size greater or equal than d.

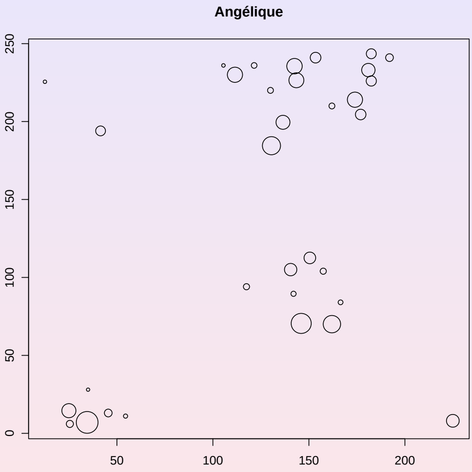
$$\mathrm{T}_{k,n,d} = \sum_{s=kd}^{n} \binom{n}{s} \mathrm{T}''_{k,s,d} \sum_{k'=\lceil \frac{n-s}{d-1} \rceil}^{n-s} \mathrm{T}'_{k',n-s,d-1} q^{s(n-s)}$$

where $\lceil x \rceil = \min\{n \in \mathbb{Z}, n \geq x\}$ and

- $\mathrm{T}''_{k,n,d}$ is the probability of having k connected components of size greater or equal to d with no component of size strictly less that d,
- $\mathrm{T}'_{k,n,d}$ is the probability of having k connected components of size smaller than d.

$\mathrm{T}'_{k,n,d} = \frac{1}{k} \sum_{l=1}^{\min(d,n-1)} \binom{n}{l} p_{1,l} \mathrm{T}'_{k-1,n-l,d} q^{l(n-l)}$ si $kd \geq n \geq k-1$

Angélique



Law of the number of components for Erdös' graph for multivariate process

- Erdös' graph with c classes, $\mathrm{V}_1, \ldots, \mathrm{V}_c$ of size $(n_1, \ldots, n_c)$
- Probability of connection $\mathrm{P} = \left(p_{i,j}\right)_{1 \le i,j \le c}$

$$p_{k,n_1,\ldots,n_c} = \frac{1}{k} \sum_{\substack{0 \le l_1 \le n_1 \\ \vdots \\ 0 \le l_c \le n_c}} \prod_{i=1}^{c} \binom{n_i}{l_i} p_{1,l_1,\ldots,l_c} p_{k-1,n_1-l_1,\ldots,n_c-l_c} \prod_{1 \le i \le j \le c} (1-p_{i,j})^{l_i(n_j-l_j)}$$

- Same computational burden..

What next?

- Computational issues
- Cut-off for the number of clusters
- Inhomogeneous Poisson Process
- Other suggestions

Thank you for your attention