Tests for spatial randomness based on spacings

Lionel Cucala & Christine Thomas-Agnan

Séminaire Statistique Mathématique et Applications

November 2004

Overview

- Spacings theory on $[0,1]$.
- New statistics for testing spatial randomness:

- Asymptotic normality of these statistics.
- Power of the tests.

Spacings theory on $[0, 1]$

Spacings theory on $[0,1]$

- $(U_1, \cdots, U_{n-1}) \in [0,1]^{n-1}$

Spacings theory on $[0,1]$

- $(U_1, \cdots, U_{n-1}) \in [0,1]^{n-1}$
- $\rightarrow U_{(0)} = 0 \leq U_{(1)} \leq \cdots \leq \cdots U_{(n-1)} \leq U_{(n)} = 1$

Spacings theory on $[0,1]$

- $(U_1, \cdots, U_{n-1}) \in [0,1]^{n-1}$
- $\rightarrow U_{(0)} = 0 \leq U_{(1)} \leq \cdots \leq \cdots U_{(n-1)} \leq U_{(n)} = 1$
- $\rightarrow (D_1, \cdots, D_n)$, where $D_i = U_{(i)} - U_{(i-1)}$.

Spacings theory on $[0,1]$

- $(U_1, \cdots, U_{n-1}) \in [0,1]^{n-1}$
- $\longrightarrow U_{(0)} = 0 \leq U_{(1)} \leq \cdots \leq \cdots U_{(n-1)} \leq U_{(n)} = 1$
- $\longrightarrow (D_1, \cdots, D_n)$, where $D_i = U_{(i)} - U_{(i-1)}$.
- H_0: Uniformity and independence $\Rightarrow$

  $(D_i, \;\; i = 1, \cdots, n)$ i.d. but $\sum_{i=1}^{n} D_i = 1$.

Spacings theory on $[0,1]$

- $(U_1, \cdots, U_{n-1}) \in [0,1]^{n-1}$
- $\rightarrow U_{(0)} = 0 \leq U_{(1)} \leq \cdots \leq \cdots U_{(n-1)} \leq U_{(n)} = 1$
- $\rightarrow (D_1, \cdots, D_n)$, where $D_i = U_{(i)} - U_{(i-1)}$.
- H_0: Uniformity and independence $\Rightarrow$

  $(D_i, \ \ i = 1, \cdots, n)$ i.d. but $\displaystyle\sum_{i=1}^{n} D_i = 1$.
- Idea: testing uniformity and independence by observing spacings' dispersion.

Spacings theory on $[0, 1]$

Which dispersion measures?

Spacings theory on $[0, 1]$

Which dispersion measures?

- Variance $\rightarrow V_n = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (nD_i - 1)^2$.

Spacings theory on $[0, 1]$

Which dispersion measures?

- Variance $\rightarrow V_n = \frac{1}{n}\sum_{i=1}^{n}(nD_i - 1)^2$.
- Absolute mean deviation $\rightarrow R_n = \frac{1}{n}\sum_{i=1}^{n}|nD_i - 1|$.

Spacings theory on $[0,1]$

Which dispersion measures?

- Variance $\rightarrow V_n = \dfrac{1}{n}\sum_{i=1}^{n}(nD_i - 1)^2$.

- Absolute mean deviation $\rightarrow R_n = \dfrac{1}{n}\sum_{i=1}^{n}|nD_i - 1|$.

- General formula:

$$H_0 : S_n = \frac{1}{n}\sum_{i=1}^{n} g(nD_i) \xrightarrow[n\to\infty]{d} \mathcal{N} \qquad \text{(Le Cam, 1958).}$$

Spacings theory on $[0, 1]$

Which dispersion measures?

- Variance $\rightarrow V_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(nD_i - 1)^2$.
- Absolute mean deviation $\rightarrow R_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}|nD_i - 1|$.
- General formula:

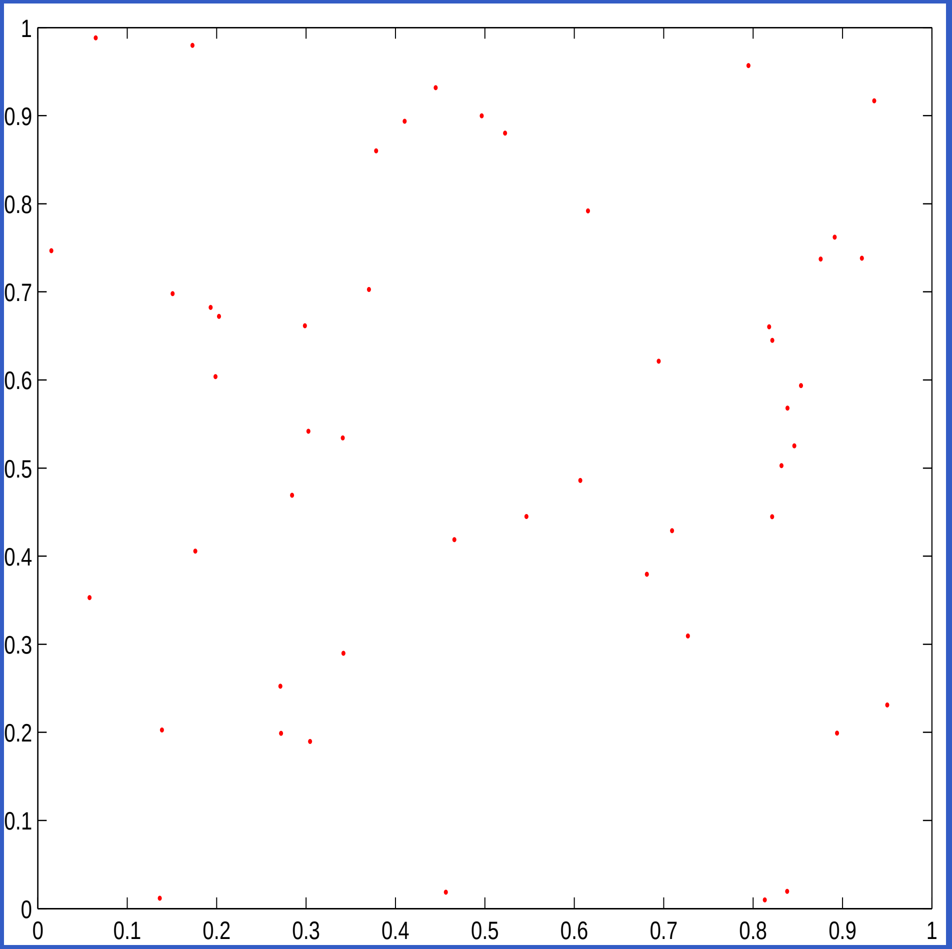
$$H_0 : S_n = \frac{1}{n}\sum_{i=1}^{n} g(nD_i) \xrightarrow[n\to\infty]{d} \mathcal{N} \qquad \text{(Le Cam, 1958).}$$

- Tests based on S_n: powerful against dependence.

Point processes

4 main types:

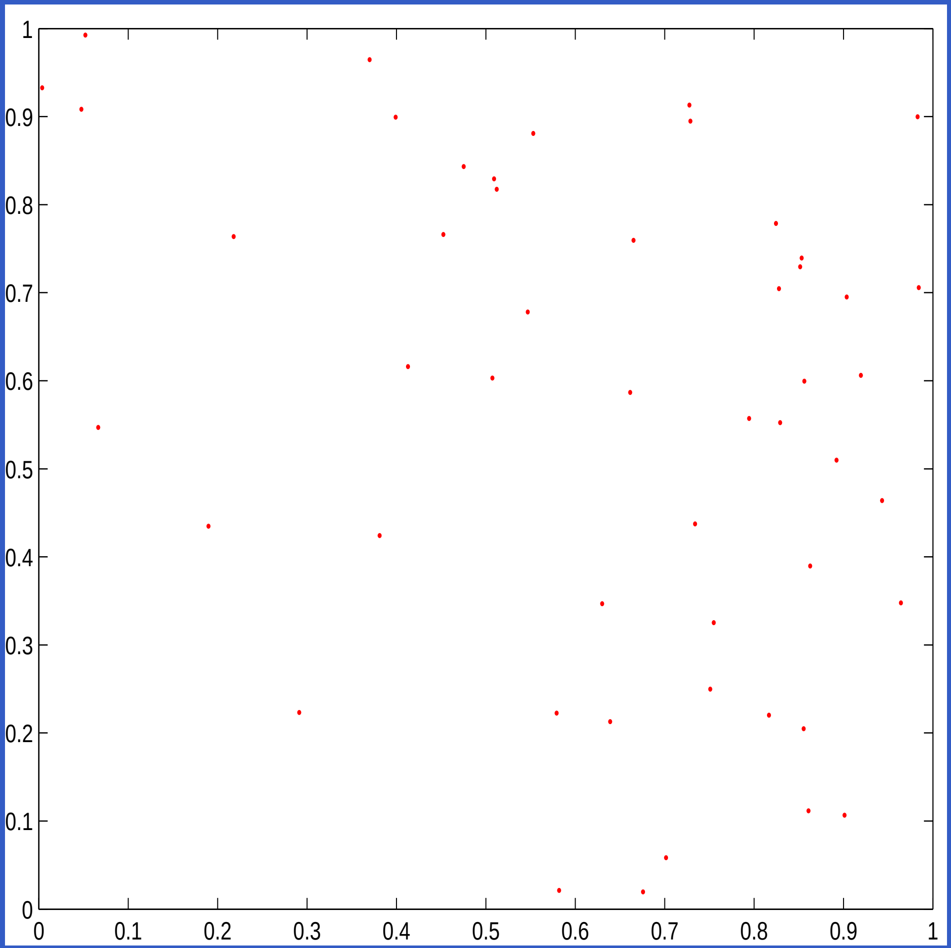
- Homogeneous Poisson process



Point processes

4 main types:

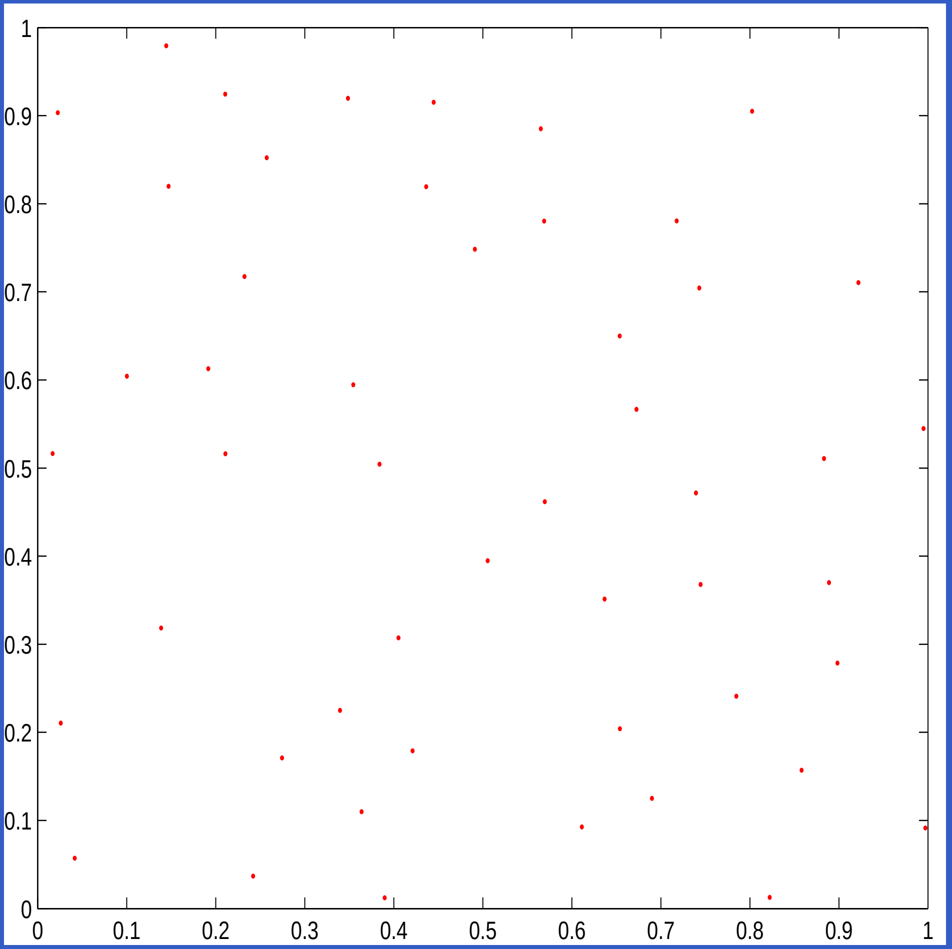
- Heterogeneous Poisson processes



Point processes

4 main types:

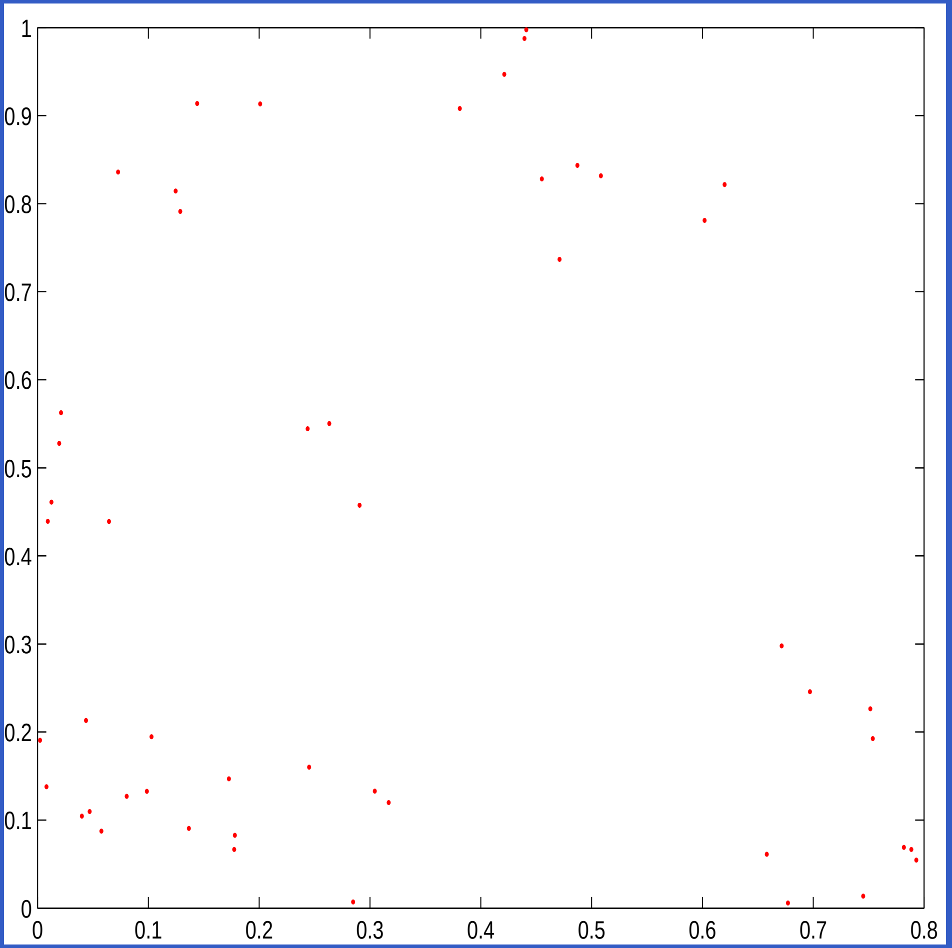
- Regular processes



Point processes

4 main types:

- Cluster processes



Complete spatial randomness

Complete spatial randomness

- $\Big((X_1, Y_1), \cdots, (X_n, Y_n)\Big)$ i.i.d. $\sim U([0,1]^2)$.

Complete spatial randomness

- $\Big((X_1, Y_1), \cdots, (X_n, Y_n)\Big)$ i.i.d. $\sim U([0,1]^2)$.

Null hypothesis as:

Complete spatial randomness

- $\Big((X_1, Y_1), \cdots, (X_n, Y_n)\Big)$ i.i.d. $\sim U([0,1]^2)$.

Null hypothesis as:

- no parameter to estimate,

Complete spatial randomness

- $\Big((X_1, Y_1), \cdots, (X_n, Y_n)\Big)$ i.i.d. $\sim U([0,1]^2)$.

Null hypothesis as:

- no parameter to estimate,
- no possible prediction.

Complete spatial randomness

- $\Big((X_1, Y_1), \cdots, (X_n, Y_n)\Big)$ i.i.d. $\sim U([0,1]^2)$.

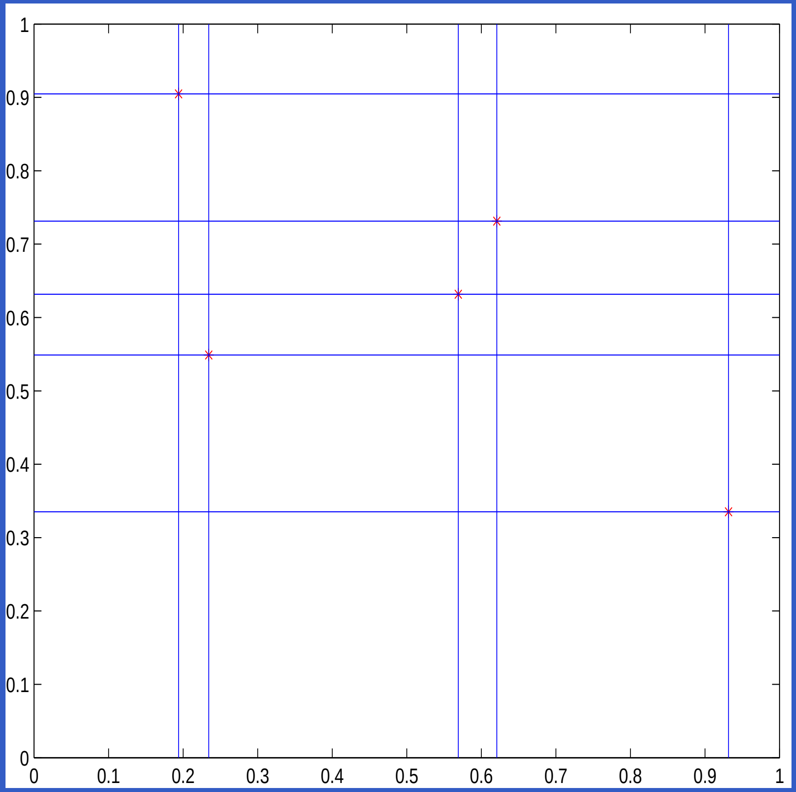
Null hypothesis as:

- no parameter to estimate,
- no possible prediction.

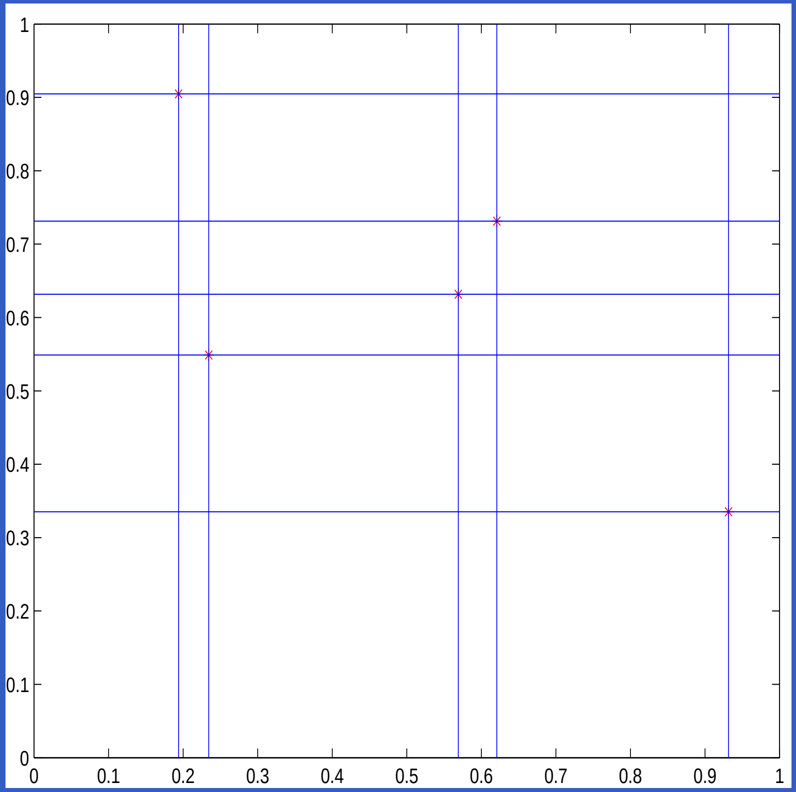
$\Rightarrow$ First reflex: CSR test.

2D spacings

2D spacings

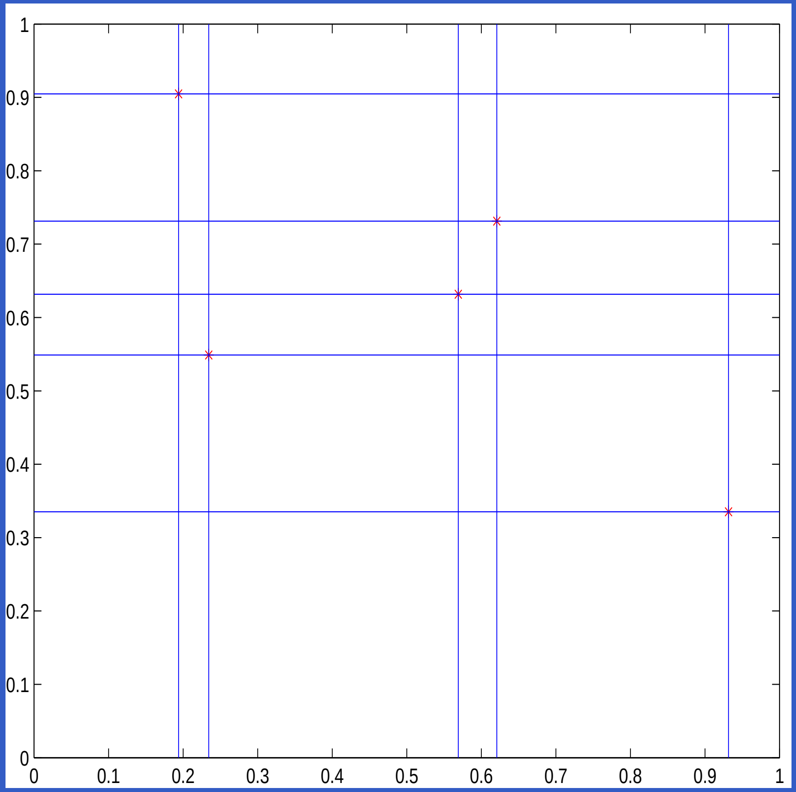


2D spacings



- $A_{ij} = D_i^x D_j^y, \qquad i = 1, \cdots, n,\ j = 1, \cdots, n.$

2D spacings



- $A_{ij} = D_i^x D_j^y, \qquad i = 1, \cdots, n,\ j = 1, \cdots, n.$
- $S_n = \dfrac{1}{n^{3/2}} \displaystyle\sum_{i=1}^{n} \sum_{j=1}^{n} g(n^2 A_{ij}).$

Generalization of the Le Cam theorem

Generalization of the Le Cam theorem

- Idea 1: use $(nD_1, \cdots, nD_n) \sim \left(\frac{E_1}{\bar{E}}, \cdots, \frac{E_n}{\bar{E}}\right)$, where $(E_1, \cdots, E_n)\, i.i.d. \sim \mathcal{E}(1)$.

Generalization of the Le Cam theorem

- Idea 1: use $(nD_1, \cdots, nD_n) \sim \left(\frac{E_1}{\bar{E}}, \cdots, \frac{E_n}{\bar{E}}\right)$, where $(E_1, \cdots, E_n)\, i.i.d. \sim \mathcal{E}(1)$.

- $\Rightarrow S_n \sim G_n = \dfrac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} g\left(\dfrac{X_i}{\bar{X}} \dfrac{Y_j}{\bar{Y}}\right)$, where $(X_1, \cdots, X_n, Y_1, \cdots, Y_n)\, i.i.d. \sim \mathcal{E}(1)$.

Generalization of the Le Cam theorem

- Idea 1: use $(nD_1, \cdots, nD_n) \sim \left(\frac{E_1}{\bar{E}}, \cdots, \frac{E_n}{\bar{E}}\right)$, where $(E_1, \cdots, E_n)\; i.i.d. \sim \mathcal{E}(1)$.
- $\Rightarrow S_n \sim G_n = \dfrac{1}{n^{3/2}} \displaystyle\sum_{i=1}^{n} \sum_{j=1}^{n} g\left(\frac{X_i}{\bar{X}} \frac{Y_j}{\bar{Y}}\right)$, where $(X_1, \cdots, X_n, Y_1, \cdots, Y_n)\; i.i.d. \sim \mathcal{E}(1)$.
- Idée 2: Taylor expansion $\rightarrow g\left(\frac{X_i}{\bar{X}} \frac{Y_j}{\bar{Y}}\right) \sim_{n\to\infty} g(X_i Y_j) - c(\bar{X} - 1) - c(\bar{Y} - 1)$, where $c = Cov(g(X_1 Y_1), X_1)$.

Generalization of the Le Cam theorem

Generalization of the Le Cam theorem

Decomposition: $G_n = U_n + R_n$,

Generalization of the Le Cam theorem

Decomposition: $G_n = U_n + R_n$,
where

$$U_n = \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} [g(X_i Y_j) - c(X_i - 1) - c(Y_j - 1)],$$ and

Generalization of the Le Cam theorem

Decomposition: $G_n = U_n + R_n$,
where

$$U_n = \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} [g(X_i Y_j) - c(X_i - 1) - c(Y_j - 1)],$$ and

$$R_n =$$

$$\frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[g(\frac{X_i}{\bar{X}} \frac{Y_j}{\bar{Y}}) - g(X_i Y_j) + c(X_i + Y_j - 2) \right].$$

Generalization of the Le Cam theorem

Decomposition: $G_n = U_n + R_n$,
where

$$U_n = \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} [g(X_i Y_j) - c(X_i - 1) - c(Y_j - 1)],$$ and

$$R_n =$$

$$\frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[g\left(\frac{X_i}{\bar{X}} \frac{Y_j}{\bar{Y}}\right) - g(X_i Y_j) + c(X_i + Y_j - 2) \right].$$

- U_n is a 2-sample U-statistic $\Rightarrow U_n \xrightarrow[n \to \infty]{d} \mathcal{N}$.

Generalization of the Le Cam theorem

Decomposition: $G_n = U_n + R_n$,
where

$$U_n = \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} [g(X_i Y_j) - c(X_i - 1) - c(Y_j - 1)], \text{ and}$$

$$R_n =$$

$$\frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[g(\frac{X_i}{\bar{X}} \frac{Y_j}{\bar{Y}}) - g(X_i Y_j) + c(X_i + Y_j - 2) \right].$$

- U_n is a 2-sample U-statistic $\Rightarrow U_n \xrightarrow[n\to\infty]{d} \mathcal{N}$.
- $\mathbb{E}(R_n^2) \xrightarrow[n\to\infty]{} 0$.

Generalization of the Le Cam theorem

Decomposition: $G_n = U_n + R_n$,
where

$$U_n = \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} [g(X_i Y_j) - c(X_i - 1) - c(Y_j - 1)], \text{ and}$$

$$R_n =$$

$$\frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[g(\frac{X_i}{\bar{X}} \frac{Y_j}{\bar{Y}}) - g(X_i Y_j) + c(X_i + Y_j - 2) \right].$$

- U_n is a 2-sample U-statistic $\Rightarrow U_n \xrightarrow[n\to\infty]{d} \mathcal{N}$.
- $\mathbb{E}(R_n^2) \xrightarrow[n\to\infty]{} 0$.

$$\Rightarrow S_n \xrightarrow[n\to\infty]{d} \mathcal{N}.$$

Used statistics and their laws

Used statistics and their laws

- Variance

$$\rightarrow V_n = \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \{(n^2 A_{ij} - 1)^2 - 3\} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 32).$$

Used statistics and their laws

- Variance

$$\rightarrow V_n = \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \{(n^2 A_{ij} - 1)^2 - 3\} \xrightarrow[n\rightarrow\infty]{d} \mathcal{N}(0, 32).$$

- Absolute mean deviation

$$\rightarrow R_n = \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \{|n^2 A_{ij} - 1| - \mu\} \xrightarrow[n\rightarrow\infty]{d} \mathcal{N}(0, \sigma^2).$$

Used statistics and their laws

- Variance

$$\to V_n = \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \{(n^2 A_{ij} - 1)^2 - 3\} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 32).$$

- Absolute mean deviation

$$\to R_n = \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \{|n^2 A_{ij} - 1| - \mu\} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2).$$

Limitation: $n \leq 100 \Rightarrow$ empirical fractiles far from the fractiles of the limit law.

Application to real data sets

Application to real data sets

4 data sets respectively considered as homogeneous, cluster, regular and heterogeneous.

Application to real data sets

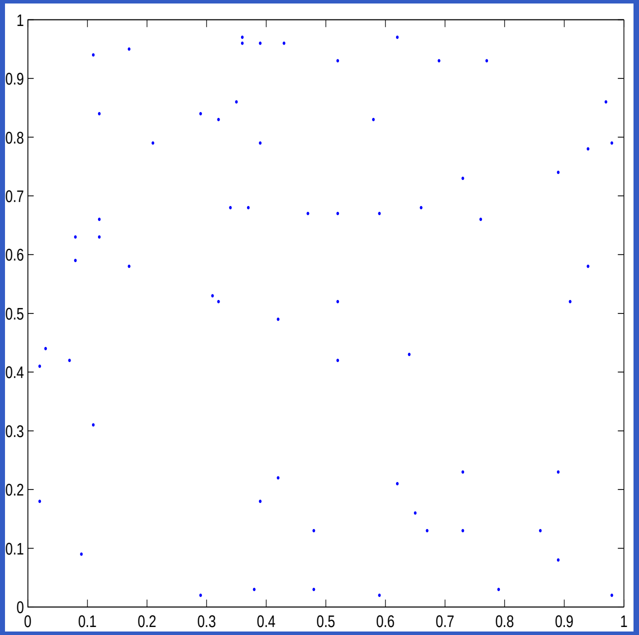
4 data sets respectively considered as homogeneous, cluster, regular and heterogeneous.

Test statistic	Results for the following data sets			
	Japanese pines	Redwoods	Biological cells	Scouring rushes
V_n	< 0.002	0.042	0.064	0.024
R_n	< 0.002	< 0.002	0.832	0.044
$\bar{\omega}^2$	0.712	0.692	0.006	0.004
D_n	0.26	0.908	0.014	0.044
T	0.915	< 0.001	< 0.001	0.936
U	0.68	< 0.01	< 0.01	0.98
V	0.50	< 0.01	< 0.01	0.92
Li	0.918	< 0.002	< 0.002	0.102
L_m	0.90	< 0.01	< 0.01	0.32

Application to real data sets

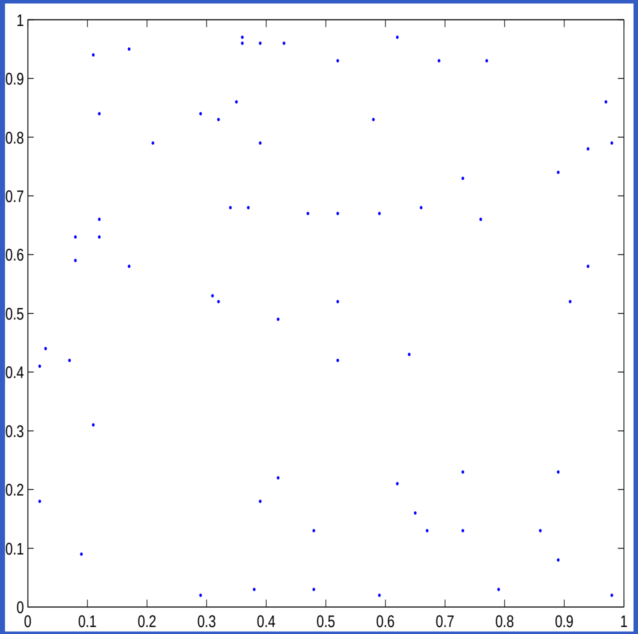
Application to real data sets

Japanese Pines:



Application to real data sets

Japanese Pines:



Heterogeneous Poisson process with intensity $\lambda(x, y) = \lambda_1(x)\lambda_2(y)$

where $\lambda_1(x) = \max_{x_1,\cdots,x_m} \exp(-c_x|x - x_i|)$

and $\lambda_2(y) = \max_{y_1,\cdots,y_l} \exp(-c_y|y - y_j|)$.

Application to simulated data sets

Application to simulated data sets

Regular process:

ϵ	Estimated power of the following:				
	L_m	$\bar{\omega}^2$	D_n	V_n	R_n
0.03	0.463	0.052	0.057	0.057	0.057
0.05	1	0.129	0.114	0.085	0.081
0.07	1	0.305	0.263	0.132	0.100

Application to simulated data sets

Regular process:

ϵ	Estimated power of the following:				
	L_m	$\bar{\omega}^2$	D_n	V_n	R_n
0.03	0.463	0.052	0.057	0.057	0.057
0.05	1	0.129	0.114	0.085	0.081
0.07	1	0.305	0.263	0.132	0.100

Cluster process:

μ	ρ	t	Estimated power of the following:				
			Li	$\bar{\omega}^2$	D_n	V_n	R_n
10	10	0.15	0.999	0.856	0.772	0.837	0.714
10	10	0.25	0.778	0.704	0.603	0.481	0.368
20	5	0.3	0.875	0.836	0.750	0.684	0.567

Application to simulated data sets

Application to simulated data sets

Heterogenous processus (planar trend):

θ_1	θ_2	Estimated power of the following:				
		Li	$\bar{\omega}^2$	D_n	V_n	R_n
4	4	0.241	0.792	0.736	0.195	0.148
6	6	0.363	0.930	0.910	0.308	0.225
8	4	0.366	0.923	0.896	0.308	0.226

Application to simulated data sets

Heterogenous processus (planar trend):

θ_1	θ_2	Estimated power of the following:				
		Li	$\bar{\omega}^2$	D_n	V_n	R_n
4	4	0.241	0.792	0.736	0.195	0.148
6	6	0.363	0.930	0.910	0.308	0.225
8	4	0.366	0.923	0.896	0.308	0.226

Grid-heterogeneous process:

m	c	Estimated power of the following:						
		Li	L_m	T	$\bar{\omega}^2$	D_n	V_n	R_n
5	25	0.253	0.065	0.402	0.026	0.042	0.663	0.722
5	30	0.375	0.194	0.589	0.033	0.045	0.886	0.930
7	30	0.044	0.016	0.076	0.036	0.028	0.293	0.458
7	40	0.106	0.047	0.162	0.029	0.036	0.695	0.862

Conclusion

Conclusion

- Spacings-based methods are useful to test whether a forest comes from a human plantation.

Conclusion

- Spacings-based methods are useful to test whether a forest comes from a human plantation.
- Estimation of the point process parameters.

Conclusion

- Spacings-based methods are useful to test whether a forest comes from a human plantation.
- Estimation of the point process parameters.
- Different alternatives to CSR $\Rightarrow$ different tests to detect them.

Conclusion

- Spacings-based methods are useful to test whether a forest comes from a human plantation.
- Estimation of the point process parameters.
- Different alternatives to CSR $\Rightarrow$ different tests to detect them.
- Generalization to 3D.