Statistical Inference for Stable Point Processes

Brunella Marta Spinelli

L'unica gioia al mondo è cominciare.
È bello vivere perché vivere è cominciare,
sempre, ad ogni istante.

C.Pavese

Acknowledgments

I would like to thank

Prof. Vincenzo Capasso for his helpfulness in the realization of this thesis project;

Prof. Sergei Zuyev for accepting to work with me, welcoming me in Chalmers and being constantly available for any question.

Contents

Introduction

Spatial models with highly irregular and bursty behavior are interesting for many applications, especially when communication systems are concerned and, for example, calling activity in a certain area has to be modeled. Figure 1 shows calling activity in Paris during the music festival *Fête de la Musique* in 2008. The height of the different torches represent the number of people calling in a certain area of the city. The data show bursty behavior in places like stadiums or squares where the concentration of people is higher. If we imagine to project the torches on the "plane" of the city we see that, if we want to fit these data, a spatial model with highly irregular spatial behavior is needed.

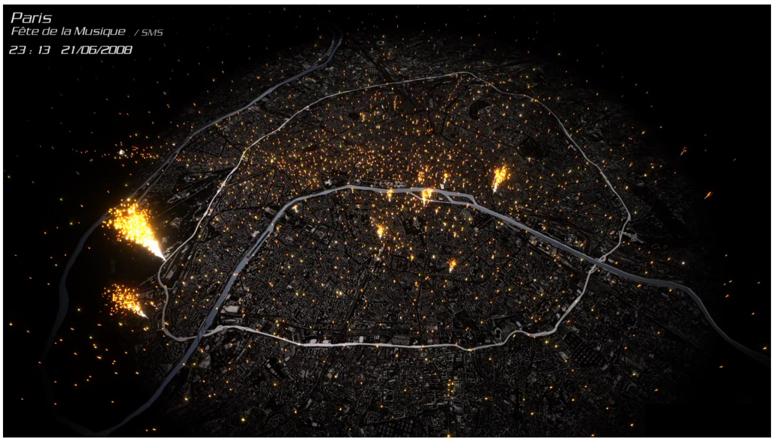


Figure 1: Paris, *Fête de la Musique*, 2008

In the present work we study parameter estimations for a particular class of point processes that, because of their bursty behavior, can be well employed in these contexts. Our model lies in the more general class of discrete α-stable point processes.

In 2011 Davydov, Molchanov and Zuyev ([4]) fully characterized discrete α-stable (DαS) point processes. These processes extend the notion of stability for random variables and random measures that arises naturally in limiting schemes to the case of point processes. Namely a random variable is called strictly α-stable (StαS) if the following equation holds in distribution for any $t \in [0, 1]$:

$$t^{1/\alpha} X' + (1-t)^{1/\alpha} X'' \stackrel{\mathcal{D}}{=} X,$$

where X' and X'' are independent copies of the random variable X. In an analogous way a random measure ξ is StαS if for any $t \in [0, 1]$:

$$t^{1/\alpha} \xi' + (1-t)^{1/\alpha} \xi'' \stackrel{\mathcal{D}}{=} \xi,$$

where ξ' and ξ'' are independent copies of the random variable ξ.
This definition of stability is extended in a natural way to point process using the operation of p-thinning as the scalar multiplication. The p-thinned point process is obtained by the original one retaining each point with probability p and deleting it with probability $1-p$. Hence, if we denote this operation by $\circ$, we say that a point process N is DαS if for any $p \in [0, 1]$

$$t^{1/\alpha} \circ N' + (1-t)^{1/\alpha} \circ N'' \stackrel{\mathcal{D}}{=} N,$$

where N' and N'' are independent copies of the point process N. In [4] it is proved that a DαS point process is a Cox (doubly stochastic) process directed by a StαS random measure. Under adequate hypotheses related to the finiteness of the StαS random measure, the DαS point process associated is a cluster process with Poisson centre process on the space of probability measures and Sibuya component processes with Sib(α) number of points each.
The model we study is exactly this cluster model, with some additional hypotheses to guarantee stationarity and ergodicity: we ask the Poisson centre process to be homogeneous with parameter λ and we also require all the clusters to be distributed according to translations of a fixed probability measure μ.
The goal of the present thesis is that of studying and implementing in **R** methods for estimating the three parameters which characterize the model: the stability exponent α, the Poisson parameter λ and the measure μ according to which the points are distributed within every single cluster.

In Chapter 1 we recall some basic definitions and results regarding the theory of point processes.
In Chapter 2 we define some operations on point processes, we discuss the notion of stability for random variables, random measures and point processes and we go through some results presented in [4] that are of interest for us.
In Chapter 3 we define our model and we discuss how to obtain an estimation of the measure μ when no information about the other two parameters is available. The idea is basically that of extracting one big cluster from our data and to use it to estimate μ or the parameters which characterize it (in case μ is known to be in a parametric class). We show that we can obtain good results by considering our data as coming from a mixture model where different components represent different clusters. Then the biggest clusters are separated and identified combining an Expectation-Maximization (EM) algorithm and the Bayesian Information Criterion.
In Chapter 4 we concentrate on the task of estimating the Poisson parameter λ and the stability exponent α under the hypothesis that the Sibuya parameter measure μ is already known. Since the processes we consider in our model are simple (i.e. they do not have multiple points) we use the characterization of the process through void probabilities. We present two strategies to make inference on our parameters. The first one is a direct estimation of the void probabilities $\mathbf{P}\{N(S_r) = 0\}$, where S_r is an arbitrary ball of radius r in the region where we can assume stationarity for the process. The second is an estimation of void probabilities for ball of fixed radius considering different filtrations of the data (i.e. we estimate void probabilities for the p-thinned process for a range of retention probabilities $p_1, .., p_n$). We show the results obtained implementing both strategies on synthetic samples and we conclude that the second method manages to exploit better all the information contained in the data.
In the Appendix we include some of the **R** codes that we used to generate the synthetic samples and to estimate the parameters and we give some ideas about how they were programmed and how they work.

Chapter 1

Basic notions regarding point processes

In this first chapter we give some basic definitions and results within the theory of point processes. Our discussion heavily relies on that of Daley and Vere-Jones in [2] and [3].
In Section 1 and 2 we introduce point processes in the more general context of random measures. In fact, basic knowledge of random measures will be also needed to define stable point processes and to study the particular class of these point processes we are focusing on.
In Section 3 we define probability generating functionals and Laplace functionals as very useful tools for classification and study of point processes.
Finally in Section 4 we present some examples of point processes that are interesting on their own but will also play a significant role in Chapter 2 when different representations for stable point processes will be given.

1.1 Definition of point process

We introduce point processes in the framework of the theory of random measures.
In the following let $\mathcal{X}$ be a complete, separable, metric space (c.s.m.s.) and $\mathcal{B}_\mathcal{X}$ the σ-algebra of its Borel sets, that is the σ-algebra generated by open sets of $\mathcal{X}$.

Definition 1 (Boundedly finite measure)**.** *A Borel measure μ on a set $\mathcal{X}$ (i.e. a measure defined on $\mathcal{B}_\mathcal{X}$) is boundedly finite if $\mu(A)$ is finite for every* *bounded Borel set A.*

Definition 2 (Counting measure)**.** *A counting measure on $\mathcal{X}$ is a boundedly finite measure which values on Borel sets are non-negative integers.*

We note that a boundedly finite measure N on $\mathcal{B}_{\mathcal{X}}$ is a counting measure if and only if $N = \sum_i k_i \varepsilon_{x_i}$ where k_i are positive integers, $\{x_i\}_i$ is a countable set of points in $\mathcal{X}$ with at most finitely many x_i in any bounded Borel set and ε denotes the Dirac measure.
We denote with $\mathcal{N}^{\#}_{\mathcal{X}}$ the set of all counting measures. On this set we define a distance as an extension of the Prohorov metric.

Proposition 1 (Prohorov distance)**.** *If we define, for every μ, ν finite measures,*

$$d(\mu, \nu) = \inf\{\varepsilon : \varepsilon \geq 0,\ \forall\, F \subseteq \mathcal{X},\ \mu(F) \leq \nu(F^{\varepsilon}) + \varepsilon \text{ and } \nu(F) \leq \mu(F^{\varepsilon}) + \varepsilon\}$$

where $F^{\varepsilon} = \{x : \rho(x, F) < \varepsilon\}$ and ρ is the metric in $\mathcal{X}$, we have that d is a distance in the space of finite measures.

Proposition 2 (Prohorov metric for boundedly finite measure)**.** *Fix a point $x_0 \in \mathcal{X}$, let $S_r = S_r(x_0)$ for $0 < r < +\infty$ and for μ, ν boundedly finite measures define*

$$d^{\#}(\mu, \nu) = \int_0^{+\infty} e^{-r} \frac{d(\mu_r, \nu_r)}{1 + d(\mu_r, \nu_r)} dr$$

where μ_r, ν_r are the restrictions of μ, ν to S_r and d is the Prohorov distance between the restrictions. Then $d^{\#}$ is a distance on the space of locally bounded measures.

The presence of a metric allows us to consider the family of Borel sets $\mathcal{B}(\mathcal{N}^{\#}_{\mathcal{X}})$. Furthermore the latter comes out to be the smallest σ-algebra with respect to which the mappings $N \to N(A)$ are measurable.
Now we are ready to define a random measure and then a point process.

Definition 3 (Random measure)**.** *A random measure ξ with phase space $\mathcal{X}$, is a measurable mapping from a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ into the set of boundedly finite measure endowed with its own σ-algebra, which we denote as $(\mathcal{M}^{\#}_{\mathcal{X}}, \mathcal{B}(\mathcal{M}^{\#}_{\mathcal{X}}))$.*

Definition 4 (Point process)**.** *A point process on $\mathcal{X}$ is a measurable mapping from a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ into the space $(\mathcal{N}^{\#}_{\mathcal{X}}, \mathcal{B}(\mathcal{N}^{\#}_{\mathcal{X}}))$*

As a consequence we can refer to point processes as *random counting measures*. The measurability of the mappings $N \to N(A)$ quoted above implies that $N : \Omega \to \mathcal{N}^{\#}_{\mathcal{X}}$ is a random measure if and only if $N(A)$ is a random variable for each bounded $A \in \mathcal{B}_{\mathcal{X}}$, providing a very useful criterion to check if a mapping $N : \Omega \to \mathcal{N}^{\#}_{\mathcal{X}}$ is indeed a point process.

Definition 5 (Simple point process)**.** *Let $\mathcal{N}^{\#*}_{\mathcal{X}}$ be the family of all simple counting measures, consisting of all those elements $N \in \mathcal{N}^{\#}_{\mathcal{X}}$ for which*

$$N(\{x\}) = 0 \text{ or } 1 \text{ for all } x \in \mathcal{X}.$$

We say that a point process N is simple when $\mathbf{P}\{N \in \mathcal{N}^{\#*}_{\mathcal{X}}\} = 1$.

We just recall an important result of characterization for random measures.

Definition 6 (Finite-dimensional distributions)**.** *The finite-dimensional distributions ("fidi") of a random measure ξ are the joint distributions, for all finite families of bounded Borel sets $A_1, ..., A_k$, of the random variables $\xi(A_1), ..., \xi(A_k)$, that is, the family of distribution functions*

$$F_k(A_1, ..., A_k; x_1, ..., x_k) = \mathbf{P}\{\xi(A_i) \leq x_i, i = 1, ..., k\}.$$

Proposition 3 (Characterization in terms of finite-dimensional distributions)**.** *The distribution of a random measure is completely determined by the fidi distributions for all finite families $(A_1, ..., A_k)$ of disjoint sets from a semiring $\mathcal{A}$ of bounded sets generating $\mathcal{B}_{\mathcal{X}}$.*

When point processes are concerned, we can consider, instead of $F_k(A_1, ..., A_k; x_1, ..., x_k)$, the fidi probabilities

$$\mathbf{P}_k(A_1, ..., A_k; n_1, ..., n_k) = \mathbf{P}\{N(A_1) = n_1, ..., N(A_k) = n_k\}.$$

In the case of simple point processes we also have the following result:

Proposition 4 (Characterization in terms of void probabilities)**.** *The distribution of a simple point process N on a c.s.m.s. $\mathcal{X}$ is determined by the values of the avoidance function $P_0 = \mathbf{P}\{N(A) = 0\}$ on the Borel sets of $\mathcal{X}$.*

We also say that $P_0(A)$ is the *void probability* of A. It will be also useful to have the following:

Definition 7 (Stationary point process)**.** *Let $\mathcal{X}$ be a vectorial c.s.m.s.. A point process N is said to be stationary if the fidi distributions are invariant under simultaneous shifts of their arguments, i.e. for each $x \in \mathcal{X}$*

$$\mathbf{P}_k(A_1, ..., A_k; n_1, ..., n_k) = \mathbf{P}_k(A_1 + x, ..., A_k + x; n_1, ..., n_k)$$

for every $A_1, ..., A_k \in \mathcal{B}_{\mathcal{X}}$ and $n_1, ..., n_k$ positive integers.

Definition 8 (Ergodic point process)**.** *A stationary point process on the state space $\mathcal{X} = \mathbb{R}^d$ with probability measure $\mathbf{P}$ on $\mathcal{B}(\mathcal{N}_{\mathcal{X}})$ is said to be ergodic if for all V, W in $\mathcal{B}(\mathcal{N}_{\mathcal{X}})$*

$$\frac{1}{\nu_d(\mathbb{U}_a^d)} \int_{\mathbb{U}_a^d} \mathbf{P}(S_x V \cap W) - \mathbf{P}(V)\mathbf{P}(W) dx \to 0 \textit{ for } a \to +\infty.$$

where $\mathcal{S}_x N(\cdot) = N(\cdot + x)$ and we write $\mathbb{U}_a^d$ a for the hypercube in $\mathbb{R}^d$ with sides of length $2a$ and vertices $(\pm a, ..., \pm a)$.

If we take $W \in \mathcal{B}(\mathcal{N}_{\mathcal{X}})$ such that $\mathbf{P}(W) \neq 0$ we can rewrite the above condition as

$$\frac{1}{\nu_d(\mathbb{U}_a^d)} \int_{\mathbb{U}_a^d} \mathbf{P}(S_x V | W) dx \to \mathbf{P}(V) \text{ for } a \to +\infty.$$

In this last formulation we can recognize that this definition of ergodicity is an extension of the well known definition for point processes on $\mathbb{R}$. In fact a stationary point process N on $\mathbb{R}$ with finite mean density is said to be ergodic when

$$\mathbf{P}\Big\{\frac{N(0, x]}{x} \to m \text{ for } x \to +\infty)\Big\} = 1$$

.

1.2 Intensity measure

We recall that if $\mathcal{X} = \mathbb{R}$ and N is a stationary point process

$$\lambda = \lim_{h \searrow 0} \frac{\mathbf{P}\{N(0, h] > 0\}}{h}$$

exists, finite or infinite, and we say that λ is the *intensity* of the process. In analogy with this, we now define the intensity for a general point process. First we need some definitions.

Definition 9 (Support counting measure). *Given $N = \sum_i k_i \varepsilon_{x_i}$ a counting measure on $\mathcal{B}_{\mathcal{X}}$, we define the support counting measure as*

$$N^* = \sum_i \varepsilon_{x_i}$$

where ε denotes the Dirac measure.

Definition 10 (Dissecting system). *The sequence $\mathcal{T} = \{\mathcal{T}_n\}_{n \in \mathbb{N}}$ of finite partitions $\mathcal{T}_n = \{A_{n,i} : i = 1, ..., k_n\}$ consisting of Borel sets of $\mathcal{X}$ is a dissecting system for $\mathcal{X}$ when*

i. *(partition property) $A_{n,i} \cap A_{n,j} = \emptyset$ for $i \neq j$ and $A_{n,1} \cup ... \cup A_{n,k_n} = \mathcal{X}$;*

ii. *(nesting property) $A_{n-1,i} \cap A_{n,j} = A_{n,j}$ or $\emptyset$;*

iii. *(point-separating property) $\forall x, y \in \mathcal{X}$ s.t. $x \neq y$ $\exists$ $n = n(x, y) \in \mathbb{N}$ s.t. $x \in A_{n,i}$ implies $y \notin A_{n,i}$.*

Definition 11 (Intensity measure). *The intensity measure of a point process N is the set function, finite or not finite, defined on each $A \in \mathcal{B}_{\mathcal{X}}$ as*

$$\Lambda(A) = \mathbb{E}[N^*(A)]$$

Proposition 5 (Khinchin's Existence Theorem). *Let N a point process on $\mathcal{X}$. For any choice of a dissecting system $\mathcal{T}$ for $\mathcal{X}$ the intensity measure of N is given by*

$$\Lambda(A) = \sup_{\mathcal{T}_n \in \mathcal{T}(A)} \sum_{i=1}^{k_n} \mathbf{P}\{N(A_{n,i}) \geq 1\},$$

meaning that it is independent of the choice of the dissecting system $\mathcal{T}$.

In case $\mathcal{X} = \mathbb{R}^d$ and the point process is stationary the intensity measure is invariant under translation and so it is proportional to the Lebesgue measure:

$$\Lambda(A) = \bar{\lambda}\nu_d(A), \quad \bar{\lambda} \in \mathbb{R}_+$$

This last formula is the exact analogous of what we have recalled for 1-dimensional point processes.

1.3 Probability generating functional and Laplace functional

We start recalling the following definition:

Definition 12 (Probability generating function)**.** *Given a random variable X on the space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ the probability generating function (p.g.f.) of X is the function defined, for each $t \in \mathbb{R}$, as $\mathbb{E}[t^X]$.*
If $\underline{X} = (X_1, ..., X_n)$ is a random vector on the space $(\mathcal{X}^n, \mathcal{B}_{\mathcal{X}^n})$ the p.g.f. of $\underline{X}$ is the function defined as $\mathbb{E}[t_1^{X_1} \cdot ... \cdot t_n^{X_n}]$ for each $(t_1, ..., t_n) \in \mathbb{R}^n$.

Let's denote with $\mathcal{V}(\mathcal{X})$ the set of $\mathcal{B}_{\mathcal{X}}$-measurable functions $h : \mathcal{X} \to \mathbb{R}$ s.t. $1 - h(x)$ vanishes out of some bounded set and $0 \leq h(x) \leq 1$ for each $x \in \mathcal{X}$.

Definition 13 (Probability generating functional)**.** *The probability generating functional (p.g.fl.) of a general point process N on the c.s.m.s. $\mathcal{X}$ is defined, for each $h \in \mathcal{V}(\mathcal{X})$, as*

$$G[h] \equiv G_N[h] = \mathbb{E}\Big[\exp\Big(\int_{\mathcal{X}} \log h(x) N(dx)\Big)\Big]$$

Since the process is a.s. finite on the set where $1 - h(x) \neq 1$ we can write

$$G_N[h] = \mathbb{E}\Big[\prod_i h(x_i)\Big]$$

where x_i are the points such that $N = \sum_i \varepsilon_{x_i}$, possibly having repetitions in the sequence $\{x_i\}_i$. We define the product in the previous formula as zero when $h(x_i) = 0$ for some x_i and we take the empty product to be unity. The main result we need is the following:

Theorem 1 (Characterization of point processes in terms of p.g.fl.)**.** *Let $G[h]$ be a real-valued functional on the space $\mathcal{V}(\mathcal{X})$. Then it is the p.g.fl. of a point process if and only if*

i. for every h of the form

$$h(x) = 1 - \sum_{k=1}^{n} (1 - z_k) \mathbb{I}_{A_k}(x)$$

where $A_1, ..., A_n$ are disjoint bounded Borel sets of $\mathcal{X}$ and $|z_i| \leq 1$, $G[h]$ reduces to the joint probability generating function $P_n(z_1, ..., z_n; A_1, ...A_n)$ of an n-dimensional integer valued random vector $\underline{X} = \underline{X}(A_1, ...A_n)$;

ii. for every sequence of functions $h_n \subset \mathcal{V}(\mathcal{X})$ with $h_n \downarrow h$ pointwise, $G[h_n] \to G[h]$ whenever $1 - h$ has bounded support;

iii. $G[\mathbf{1}] = 1$ where $\mathbf{1}$ is the function identically equal to 1 in $\mathcal{X}$.

Moreover, when these conditions are satisfied, the functional G uniquely determines the distribution of N.

Another useful instrument to handle point processes is the Laplace functional. Let's denote with $BM(\mathcal{X})$ the set of all bounded measurable functions $h : \mathcal{X} \to \mathbb{R}_+$ with compact support.

Definition 14 (Laplace functional)**.** *Let ξ be a random measure. The Laplace functional of ξ is defined by*

$$L_\xi[h] = \mathbb{E}\Big[\exp\Big(-\int_{\mathcal{X}} hd\xi\Big)\Big]$$

for each $h \in BM(\mathcal{X})$.

Remark 1 (Laplace functional and p.g.fl.)**.** *It is easy to see that if N is a point process we have the equation*

$$G_N[h] = L_N[-\log h]$$

which makes sense for each h s.t. $-\log h$ is a bounded measurable non-negative function.

1.4 Examples of point processes

1.4.1 Poisson process

Definition 15 (Poisson point process - PPP)**.** *A point process N on the space $\mathcal{X}$ is a Poisson process if there exists a boundedly finite Borel measure $\Lambda(\cdot)$ such that for every finite family of disjoint bounded Borel sets $A_1, ..., A_k$*

$$\mathbf{P}\{N(A_1) = n_1, ..., N(A_k) = n_k\} = \prod_{i=1}^{k} \frac{[\Lambda(A_i)]^{n_i}}{n_i!} e^{-\Lambda(A_i)}.$$

The measure $\Lambda(\cdot)$ is called the parameter measure of the process.

We just quote a theorem that give a characterization of the Poisson process. If we call *fixed points* the points x_0 such that $\mathbf{P}\{N(\{x_0\}) > 0\} > 0$, we have that x_0 is a fixed point for the process N if and only if Λ has an atom in x_0.

Definition 16 (Orderly process)**.** *A point process is said to be orderly if for each $x \in \mathcal{X}$, as $\varepsilon \to 0$, it satisfies*

$$\mathbf{P}\{N(S_\varepsilon(x)) > 1\} = o(\mathbf{P}\{N(S_\varepsilon(x)) > 0\}).$$

Theorem 2. *Let N be a a.s. boundedly finite point process without fixed atoms. Then N is a point process if and only if*

i. N is orderly;

ii. for each finite family of bounded Borel sets $A_1, ..., A_k$, the random variables $N(A_1), ..., N(A_k)$ are mutually independent.

The p.g.fl. of the Poisson process of parameter measure Λ is

$$G[h] = \exp\Big(- \int_{\mathcal{X}} (1 - h(x))\Lambda(dx)\Big)$$

for every $h \in \mathcal{V}(\mathcal{X})$.

1.4.2 Processes constructed via conditioning

Here we introduce models for point processes that are defined in two steps: first an initial process is laid down and then a second one, with distribution conditional on the realization of the other. As a general framework let's consider a family $\{N(\cdot|y) : y \in \mathcal{Y}\}$ of point processes defined on the c.s.m.s. $\mathcal{X}$ and indexed by the elements y of the c.s.m.s. $\mathcal{Y}$. We call this family a *measurable family* if, for each set $A \in \mathcal{B}(\mathcal{N}^{\#}_{\mathcal{X}})$, the function $\mathbf{P}(A|y)$ is $\mathcal{B}_{\mathcal{Y}}$-measurable, where

$$\mathbf{P}(A|y) \equiv \mathbf{P}\{N(\cdot|y) \in A\}.$$

The following proposition gives conditions for a measurable family to define a point process.

Proposition 6. *Suppose there is given a measurable family $\mathbf{P}(A|y)$ defined as above and a $\mathcal{Y}$-valued random variable Y with distribution Π on $\mathcal{B}_{\mathcal{Y}}$. Then*

$$\mathbf{P}(A) \equiv \mathbb{E}[\mathbf{P}(A|Y)] = \int_Y \mathbf{P}(A|y)\Pi(dy)$$

defines a probability measure $\mathbf{P}$ on $\mathcal{B}(\mathcal{N}^{\#}_{\mathcal{X}})$ and hence a point process on $\mathcal{X}$.

Actually it can be shown that if Y is a r.v. as above and a family of $\mathcal{B}_{\mathcal{Y}}$-measurable fidi probabilities $\mathbf{P}_k(A_1, ..., A_k; n_1, ..., n_k|y)$ is given, then there exists a well defined point process for which the fidi probabilities are

$$\begin{aligned} \mathbf{P}_k(A_1, ..., A_k; n_1, ..., n_k) &= \mathbb{E}[\mathbf{P}_k(A_1, ..., A_k; n_1, ..., n_k|Y)] = \\ &= \int_Y \mathbf{P}_k(A_1, ..., A_k; n_1, ..., n_k|y)\Pi(dy). \end{aligned} \tag{1.1}$$

In an analogous way if $G[h|y]$ is a family of $\mathcal{B}(\mathcal{Y})$-measurable p.g.fl.s there exists a well defined point process on $\mathcal{X}$ for which the p.g.f.l. is

$$\mathbb{E}[G[h|y]] = \int_{\mathcal{Y}} G[h|y]\Pi(dy).$$

Example 1 (Mixed Poisson Process). *Let $\mathcal{Y} = \mathbb{R}+$ and let λ play the role of y. It is easy to see that the fidi distributions for the Poisson process are measurable functions of λ. Hence if we take λ to be a random variable we have that these distributions define a point process. We will refer to such a process as mixed Poisson process.*

Let's turn to the first model we are introducing, the *Cox process.* Cox processes are obtained by randomizing the parameter measure in a Poisson Process.

Definition 17 (Cox process). *Let ξ be a random measure on the space $\mathcal{X}$. A point process N on $\mathcal{X}$ is a Cox process directed by ξ, when, conditional on ξ, realizations of N are those of a Poisson process with parameter measure ξ.*

The process is well defined thanks to the measurability with respect to ξ of the fidi probabilities of the Poisson process $N(\cdot|\xi)$.
The p.g.fl. of the Cox process is

$$G[h] = \mathbb{E}\Big[\exp\Big(- \int_{\mathcal{X}} [(1-h(x))\xi(dx)\Big)\Big].$$

Now we define our second example, the *cluster process.* This one is a process constituted by a first process N_c (the *centre process*) and by other processes (the *clusters*) which evolve from the points that support the realization of the first process. We note that the centre process is often unobserved. Furthermore it is quite usual to assume that the *component processes* (i.e. the different clusters) are independent and that the space on which the centre process is defined is included in the space $\mathcal{X}$ on which the component processes are defined.

Definition 18 (Independent cluster process). *N is an independent cluster process on the c.s.m.s. $\mathcal{X}$, with centre process N_c on the c.s.m.s. $\mathcal{Y}$ and component processes the measurable family of independent point processes $\{N(\cdot|y) : y \in \mathcal{Y}\}$, if for each $A \in \mathcal{B}_{\mathcal{X}}$*

$$N(A) = \int_{\mathcal{Y}} N(A|y)N_c(y) = \sum_{y_i \in N_c} N(A|y_i) < +\infty \quad a.s.$$

Here we write "$y_i \in N_c$" instead of "$y_i \in supp(N_c)$" with a slight abuse of notation and we mean that whenever $N_c(y) > 1$ multiple independent copies of $N(\cdot|y)$ should be taken in the sum.

We note that the definition requires the cluster process to be a.s. boundedly finite. If we denote as $G_c[\cdot]$ the p.g.fl. of the centre process and as $G_m[\cdot|y]$ the p.g.fl. of $N(\cdot|y)$, thanks to the independence assumption, the p.g.fl. of the cluster process takes the form

$$G[h] = G_c[G_m[h|\cdot]], \quad \forall h \in \mathcal{V}(\mathcal{X}).$$

In fact we have

$$G[h] = \mathbb{E}[G[h|N_c]] = \mathbb{E}\Big[\prod_{y_i \in N_c} G_m[h|y_i]\Big] = \mathbb{E}\Big[\exp\Big(\int_{\mathcal{Y}} (\log G_m[h|y]) N_c(dy)\Big)\Big].$$

Definition 19 (Poisson cluster process)**.** *A cluster process is said to be a Poisson cluster process if the centre process N_c is a Poisson process.*

Chapter 2

Operations on point processes and stability

In this chapter we are going to define some of the basic operations on point processes and to introduce a notion of stability for random measures and point processes.
The operations we define in the first sections, particularly the superposition and the p-thinning operation, are key concepts for us since they are involved in the definition of stable point process.
In Section 5 we start to talk about stability giving the definition of a stable random variable as it naturally arises in the context of limit theorems. We then generalize this notion to random measures (Section 6) and point processes (Section 7). In this last section we go through the main results given in [4] in order to characterize stable point processes in a way that will be the starting point for our work in the following chapters.

2.1 Superposition

Definition 20 (Superposition of point processes)**.** *Given two point processes N_1 and N_2 we define the superposition process N as the process $N_1 + N_2$ meaning that for each realization $N(\cdot, \omega) \in \mathcal{N}_{\mathcal{X}}$ and for each Borel set $A \in \mathcal{B}_{\mathcal{X}}$, $N(\cdot, \omega) = N_1(\cdot, \omega) + N_2(\cdot, \omega)$.*

If N_1 and N_2 are independent point processes the p.g.fl. of the superposition process is given by

$$G_N[h] = G_{N_1}[h]G_{N_2}[h].$$

Example 2. *The superposition of two independent PPP is again a PPP, since the sum of two Poisson random variables is a Poisson random variable, and the intensity measure is the sum of the two intensity measures.*

2.2 Infinitely divisible point processes

Definition 21 (Infinitely divisible point proce)**.** *A point process is infinitely divisible if, for every $k \in \mathbb{N}$, it can be represented as the sum of k independent and identically distributed point processes. In symbols*

$$N = N_1^{(k)} + ... + N_k^{(k)},$$

where $\{N_i^{(k)}\}_{i=1}^k$ are i.i.d. point processes.

Using p.g.fl.s the condition of infinite divisibility becomes

$$G_N[h] = \prod_{i=1}^{k} G_{1/k}[h].$$

In other words, thanks to the p.g.fl. being positive, we could say that a point process is infinitely divisible if and only if the only k-th root of his p.g.fl. is again a p.g.fl.

Example 3. *A Poisson process is infinitely divisible. This could be seen replacing the original parameter measure μ with the measure $\frac{\mu}{k}$ for each component. More generally, also the Poisson cluster processes are infinitely divisible and the component processes are still obtained dividing by k the intensity of the centre process.*

We have the following result of characterization of infinite divisible point processes in terms of their p.g.fl.:

Theorem 3. *Suppose that the point process N with p.g.fl. $G[\cdot]$ is a.s. finite and infinitely divisible. Then there exists a uniquely defined and a.s. finite point process $\tilde{N}$ with p.g.fl. $\tilde{G}$, such that $\mathbf{P}\{\tilde{N} = 0\} = 0$, and a finite positive number α such that*

$$G[h] = \exp(\alpha(\tilde{G}[h] - 1)) \tag{2.1}$$

for each $h \in \mathcal{V}(\mathcal{X})$. Conversely, any functional having the form above is the p.g.fl. of an a.s. finite point process that is infinitely divisible.

In view of formula (2.1) we can interpret an a.s. finite infinitely divisible process as a two-stage construction: first a Poisson distributed number n is generated and then n independent processes distributed ad $\tilde{N}$ are superposed.

2.3 Dilation of the phase space of a Poisson process

Proposition 7. *Let F be a bijection, $F : \mathcal{X} \to \mathcal{X}$ and N a Poisson process on $\mathcal{X}$ with intensity measure Λ. Then N induces on $\mathcal{X}$ a Poisson process with intensity measure $\Lambda \circ F^{-1}$.*

Proof. We use the characterization of point processes in term of p.g.fl.s. Let $G'[h]$ be the p.g.fl. of the new point process:

$$G'[h] = G[h \circ F] = \int_{\mathcal{X}} (1 - h \circ F(x))\Lambda(dx).$$

Then, with a change of variables,

$$G'[h] = \int_{\mathcal{X}} (1 - h(F(x)))\Lambda \circ F^{-1}(dx),$$

which is the p.g.fl. of a Poisson process with intensity measure
$\Lambda' = \Lambda \circ F^{-1}$. □

2.4 Thinning

Given a point process N, as in the beginning of the first section, we can write his realizations as $N(\cdot) = \sum_i k_i \varepsilon_{x_i}(\cdot)$.

Definition 22 (Independent $p(x)$-thinning)**.** *Let N be a point process on the c.s.m.s. $\mathcal{X}$ and $p(\cdot)$ a measurable function on $\mathcal{X}$ such that for each $x \in \mathcal{X}$ $0 \leq p(x) \leq 1$. $N_{p(\cdot)}$ is obtained from N by independent thinning according to $p(\cdot)$ when each realization $N_{p(\cdot)}(\cdot, \omega)$ is obtained from a realization $N(\cdot, \omega)$ of N deleting each of the $\{x_i\}$ with probability $1 - p(x_i)$ and retaining it with probability $p(x_i)$.*

Exploiting the independence of the thinnings one can obtain that the p.g.fl. of the thinned process is

$$G_{p \circ N}[h] = G_N[p \cdot h + 1 - p].$$

From this formula we also derive that the $p(x)$-thinned process of a PPP is again a PPP. Moreover, for each simple point process we have that the intensity measure of the thinned point process is nothing but

$$\Lambda_{p(\cdot)}(A) = \int_A p(x)\Lambda(dx) \quad \forall\, A \in \mathcal{B}_{\mathcal{X}}$$

2.5 Stability of random measures and point processes

In this section we give a notion of stability for random measures and point processes and we quote some results proved in [4] that will be of interest for us.

We recall that a random vector X (or its probability law) is called *strictly α-stable* (notation: StαS) if for any $t \in [0, 1]$

$$t^{1/\alpha}X' + (1-t)^{1/\alpha}X'' \stackrel{\mathcal{D}}{=} X$$

where X' and X'' are independent copies of X.

In the classical case of d-dimensional vectors, for each strictly α-stable law it is possible to give a representation, known as *LePage representation*, as a sum of points of a Poisson process Π_α in $\mathbb{R}^d$ with the intensity function having a product form $\theta_\alpha(d\rho)\sigma(ds)$ in the polar coordinates (ρ, s). Here σ is any finite measure on the unit sphere $\mathbb{S}^{d-1}$ and it is called the *spectral measure*, and $\theta_\alpha((r, +\infty)) = r^{-\alpha}$. This result is based on the fact that, thanks to the superposition property of Poisson processes and the dilation property shown in Prop. 7, the process Π_α is itself stable with

$$t^{1/\alpha}\Pi'_\alpha + (1-t)^{1/\alpha}\Pi''_\alpha \stackrel{\mathcal{D}}{=} \Pi_\alpha.$$

2.6 Strictly α-stable random measures

Definition 23 (Strictly α-stable random measure- StαS)**.** *A random measure ξ is called strictly α-stable if*

$$t^{1/\alpha}\xi' + (1-t)^{1/\alpha}\xi'' \stackrel{\mathcal{D}}{=} \xi$$

for every $0 \le t \le 1$ and ξ', ξ'' are independent copies of the random measure ξ.

We note that non-trivial StαS may exist only if $\alpha \in (0, 1]$.

Theorem 4 (Characterization in terms of Levy measures)**.** *A locally finite random measure ξ is StαS if and only if ξ is deterministic in the case $\alpha = 1$ and in the case $\alpha \in (0, 1)$ if and only if its Laplace functional is given, for each $h \in BM(\mathcal{X})$, by*

$$L_\xi[h] = \exp\Big\{ - \int_{\mathcal{M}^{\#}_{\mathcal{X}} \setminus \{0\}} (1 - \exp^{-\langle h, \mu \rangle}) \Lambda(d\mu) \Big\}$$

where Λ is a Levy measure, i.e. a Radon measure on $\mathcal{M}^{\#}_{\mathcal{X}} \setminus \{0\}$ such that

$$\int_{\mathcal{M}^{\#}_{\mathcal{X}} \setminus \{0\}} (1 - \exp^{-\langle h, \mu \rangle}) \Lambda(d\mu) < +\infty \tag{2.2}$$

for all $h \in BM(\mathcal{X})$, and Λ is homogeneous of order $-\alpha$, i.e. $\Lambda(tA) = t^{-\alpha}\Lambda(A)$ for all measurable $A \subset \mathcal{M}^{\#}_{\mathcal{X}} \setminus \{0\}$ and $t > 0$.

Now we want to outline briefly a decomposition of the Levy measure Λ in a radial and a directional part, all the details can be found in [4].
Let $B_1, B_2, ...$ a countable base of the topology on $\mathcal{X}$ made of relatively compact sets and define $B_0 = \mathcal{X}$. Let $i(\mu)$ be the smallest non-negative integer i for which $0 < \mu(B_i) < 1$. In [4] it is shown that it is possible to have a polar decomposition of $\mathcal{M}^{\#}_{\mathcal{X}}$ as $\mathcal{M}^{\#}_{\mathcal{X}} = \mathbb{S} \times \mathbb{R}_+$ where $\mathbb{S}$ is the measurable set defined as

$$\mathbb{S} = \{\mu \in \mathcal{M}^{\#}_{\mathcal{X}} : \mu(B_{i(\mu)}) = 1\}.$$

Definition 24 (Spectral measure)**.** *The spectral measure σ of a random measure ξ is given by $\sigma = \Gamma(1-\alpha)\hat{\sigma}$ where, for all measurable $A \subseteq \mathbb{S}$*

$$\hat{\sigma}(A) = \Lambda(\{t\mu : \mu \in A, t \geq 1\}).$$

The presence of the constant $\Gamma(1-\alpha)$ will be convenient in the formulation of some of the next results.
Thanks to the definition of $\hat{\sigma}$ we have

$$\Lambda(A \times [a, b]) = \hat{\sigma}(A)(a^{-\alpha} - b^{-\alpha})$$

and so we can decompose Λ as the product of $\hat{\sigma}$ and the measure θ_α defined by $\theta_\alpha([r, +\infty)) = r^{-\alpha}$.

Theorem 5 (Laplace functional for Stαs random measures)**.** *Let σ be the spectral measure of a Stαs random measure ξ with Levy measure Λ. Then the Laplace functional of ξ is*

$$L_\xi[h] = \exp\Big\{ -\int_{\mathbb{S}} \langle h, \mu\rangle^\alpha \sigma(d\mu)\Big\}, \quad h \in BM(\mathcal{X}).$$

We recall that $\{\mu_n\} \subset \mathcal{M}_{\mathcal{X}}^{\#}$ converges vaguely to $\mu \in \mathcal{M}_{\mathcal{X}}^{\#}$ if and only if $\int f d\mu_n \to \int f d\mu$ for every bounded continuous $f : \mathcal{X} \to \mathbb{R}$ vanishing outside a compact set. Then we have the following decomposition of StαS random measures:

Theorem 6 (LePage representation for Stαs random measures)**.** *A random measure ξ is StαS if and only if, in the sense of the vague convergence of measures,*

$$\xi \overset{\mathcal{D}}{=} \sum_{\mu_i \in \Psi} \mu_i,$$

where Ψ is a Poisson process on $\mathcal{M}_{\mathcal{X}}^{\#} \backslash \{0\}$ driven by an intensity measure Λ satisfying (2.2) and such that $\Lambda(tA) = t^{-\alpha}\Lambda(A)$ for all $t > 0$ and any measurable A. In this case Λ is exactly the Levy measure of ξ.

2.7 Discrete α-stable processes

Now we turn to stability for point processes. Given a constant $p \in (0, 1)$ we denote p-thinning of a point process N with $p \circ N$. This operation on point processes is associative, commutative and distributive with respect to superposition, i.e.

$$\begin{aligned} &t_1 \circ (t_2 \circ N_1) \overset{\mathcal{D}}{=} (t_1 \circ t_2) \circ N_1 \overset{\mathcal{D}}{=} t_2 \circ (t_1 \circ N_1) \\ &t \circ (N_1 + N_2) \overset{\mathcal{D}}{=} t \circ N_1 + t \circ N_2 \end{aligned} \tag{2.3}$$

for any $t, t_1, t_2 \in [0, 1]$ and independent N_1 and N_2.
Furthermore, for any disjoint Borel sets B_1 and B_2, the random variables $(t \circ N)(B_1)$ and $(t \circ N)(B_2)$ are conditionally independent given N.

Definition 25 (Discrete α-stable process - DαS)**.** *A point process N is called discrete α-stable, or α-stable with respect to thinning, if, for every $0 \le t \le 1$,*

$$t^{1/\alpha} \circ N' + (1-t)^{1/\alpha} \circ N'' \overset{\mathcal{D}}{=} N$$

where N', N'' are independent copies of the point process N.

Remark 2 (Infinite divisibility)**.** *All DαS point processes are infinite divisible. In fact from the definition of a DαS process, for each $n \in \mathbb{N}$, we have*

$$\Big(\frac{1}{n}\Big)^{1/\alpha} \circ N' + \Big(\frac{n-1}{n}\Big)^{1/\alpha} \circ N'' \overset{\mathcal{D}}{=} N$$

and now we can apply the definition of DαS process to N'' and use the associative property to write

$$\Big(\frac{1}{n}\Big)^{1/\alpha} \circ N' + \Big(\frac{1}{n}\Big)^{1/\alpha} \circ N^{(3)} + \Big(\frac{n-2}{n}\Big)^{1/\alpha} \circ N^{(4)} \overset{\mathcal{D}}{=} N$$

where N', $N^{(3)}$ and $N^{(4)}$ are independent copies of the point process N. If we repeat this operation n times we obtain that N is distributed as the sum of n independent copies of the thinned process $\frac{1}{n} \circ N$. This means that N is infinitely divisible.

Theorem 7 (Representation of Dαs point processes as Cox processes)**.** *A point process N is DαS if and only if it is a Cox process with a StαS intensity measure ξ.*

Remark 3 (Expectation of $N(B)$ and avoidance probabilities)**.** *From the previous theorem combined with Theo. 5 we can derive that a point process N is DαS if and only if its p.g.fl. has the form*

$$G_N[h] = \exp\Big\{ - \int_{\mathbb{S}} \langle 1-h, \mu\rangle^{\alpha} \sigma(d\mu)\Big\}, \quad 1-h \in BM(\mathcal{X}). \tag{2.4}$$

For every Borel set B and $s > 0$, taking $h(x) = 1 - (1-s)\chi_B(x)$, we find that the p.g.f. of the random variable $N(B)$ is

$$\mathbb{E}[s^{N(B)}] = \exp\Big\{ - (1-s)^{\alpha} \int_{\mathbb{S}} \mu(B)^{\alpha} \sigma(d\mu)\Big\}$$

and the avoidance probabilities are given by

$$\mathbf{P}\{N(B) = 0\} = \exp\Big(- \int_{\mathbb{S}} \mu(B)^{\alpha} \sigma(d\mu)\Big)$$

As a consequence of these formulas, the avoidance probabilities are always positive and the expectation of $N(B)$ is either 0 *or infinite, depending on the vanishing of the integral*

$$\int_{\mathbb{S}} \mu(B)^{\alpha} \sigma(d\mu).$$

Theorem 8 (LePage representation for Dαs point processes)**.** *A DαS process N with Levy measure Λ can be represented as a Cox process:*

$$N \stackrel{\mathcal{D}}{=} \sum_{\mu_i \in \Psi} N_{\mu_i},$$

where Ψ is a Poisson process on $\mathcal{M}^{\#}_{\mathcal{X}} \backslash \{0\}$ with intensity measure Λ.

Definition 26 (Sibuya point process)**.** *Let μ be a probability measure on $\mathcal{X}$. The point process Υ on $\mathcal{X}$ defined by the p.g.fl.*

$$G_{\Upsilon}[h] = G_{\Upsilon(\mu)}[h] = 1 - \langle 1-u, \mu \rangle^{\alpha} \tag{2.5}$$

is called the Sibuya point process with exponent α and parameter measure μ. Its distribution is denoted by Sib(α,μ).

In particular, (2.5) implies that $\Upsilon(\mathcal{X})$ has the p.g.f.

$$\mathbb{E}[u^{\Upsilon(\mathcal{X})}] = 1 - (1-u)^{\alpha}.$$

A random variable whose p.g.f. has this form is said to be *Sibuya distributed* and its distribution is denoted as Sib(α). It could be shown that the Sibuya distribution corresponds to the number of trials to get the first success in a series of Bernoulli trials with probability of success in the k-th trial being α/k. In an analogous way, for each Borel set $B \in \mathcal{B}_{\mathcal{X}}$,

$$\mathbb{E}[u^{\Upsilon(\mathcal{B})}] = 1 - \mu(B)^{\alpha}(1-u)^{\alpha}$$

implying that $\Upsilon(B)$ has infinite expectation if $\mu(B)$ does not vanish. Moreover, if for $Z \sim$ Sib(α) we denote $q_n(\alpha) = \mathbf{P}\{Z = n\}$, we have

$$\begin{aligned}
&\mathbf{P}\{\Upsilon(B) = 0\} = 1 - \mu^{\alpha}(B),\\
&\mathbf{P}\{\Upsilon(B) = 1\} = \alpha\mu^{\alpha}(B) = q_1(\alpha)\mu^{\alpha}(B),\\
&\mathbf{P}\{\Upsilon(B) = n\} = \Big(1-\alpha\Big)\Big(1-\frac{\alpha}{2}\Big)...\Big(1-\frac{\alpha}{n-1}\Big)\frac{\alpha}{n}\mu^{\alpha}(B) = q_n(\alpha)\mu^{\alpha}(B)
\end{aligned} \tag{2.6}$$

where the latter holds for each $n \geq 2$. In particular, $\Upsilon(\mathcal{X})$ is non-zero a.s. Looking again to the form of the p.g.fl. (2.5) we can understand something more about the structure of a Sibuya process:

$$\begin{aligned}
G_{\Upsilon}[h] &= \mathbb{E}\Big[\prod_{x_i \in \Upsilon} h(x_i)\Big] = 1 - (1 - \langle h, \mu \rangle)^{\alpha} =\\
&= \sum_{n=1}^{\infty} q_n(\alpha)\langle h, \mu \rangle^{\alpha} = \sum_{n=1}^{\infty} q_n(\alpha) \int_{\mathcal{X}^n} h(x_i)...h(x_n)\mu(dx_1)...\mu(dx_n),
\end{aligned} \tag{2.7}$$

meaning that the total number of points of Υ follows the Sibuya distribution and given this total number, the points are independently and identically distributed according to the probability measure μ.

Now consider a point process N with a StαS Levy measure Λ supported only by finite measures. Then the spectral measure is supported only by the set of probability measures $\mathbb{M}_1$ and from Theo. 5 and the form of the p.g.fl. of a Sibuya point process we have

$$G_N[h] = \exp \int_{\mathbb{M}_1} (G_{\Upsilon(\mu)}[h] - 1)\sigma(d\mu). \tag{2.8}$$

So we have shown the following result:

Theorem 9 (Representation of Dαs point processes as cluster processes)**.** *A DαS point process N with Levy measure Λ supported only by finite measures can be represented as a cluster process with Poisson centre process on $\mathbb{M}_1$ driven by intensity measure σ and component processes being Sibuya processes Sib(α, μ), $\mu \in \mathbb{M}_1$.*

In this way we have also established the following fact: since Sibuya processes are finite with probability 1, then the cluster process is finite or infinite depending on the finiteness of the Poisson centre process.
In the next chapter we will need also the following result about ergodicity

Theorem 10. *A stationary DαS point process N in $\mathbb{R}^d$ is ergodic if and only if*

$$n^{-d} \int_{[-n/2,n/2]^d} (1 - e^{-\mu(B+x)})dx \to 0 \; as \; n \to +\infty.$$

for all relatively compact Borel set B and all μ from the support of σ.

We note that, for example, the condition given by the theorem is satisfied when σ is concentrated on a set of measures having compact support or when it is concentrated on measures that decay rapidly at infinity.

Chapter 3

Parameter inference

In the previous chapter we defined DαS point processes. We stated that, if the Levy measure is supported only by finite measures, we can represent the DαS process as a cluster process with Poisson centre process on the space of probability measures and Sibuya distributed clusters (see Th.9, Chapter 2). The intensity of the Poisson centre process is given by the spectral measure σ.
In principle we could have clusters distributed according to any probability measure but for our statistical purpose it is more interesting to concentrate on the class of the stationary and ergodic models. To obtain stationarity we have to require distributions of the different clusters to differ only by a translation and the distribution of the clusters' positions to be uniform on a certain region. Ergodicity can be guaranteed under certain hypotheses on the distribution of the points within the single cluster (see below for details).

3.1 Description of the model

From now on, we consider DαS point processes N on $\mathbb{R}^d$ satisfying the following assumptions:

1. the spectral measure σ is concentrated on a set $\{\mu_x\}_{x\in A}$, where $A \subseteq \mathbb{R}^d$ and $\mu_x = \mu(\cdot - x)$ is obtained by translation of a probability distribution μ;

2. σ is uniform on the set $\{\mu_x\}_{x\in A}$, that is σ is the image under the map $x \to \mu(\cdot - x)$ of a multiple of the Lebesgue measure ν. In formulas, for

any Borel set of $B \in \mathcal{B}(A)$,

$$\sigma(\{\mu_x | x \in B\}) = C \cdot \nu(B),$$

where C is a positive constant and in the following we will set $C = \lambda$ and we will say that λ is the intensity of the Poisson centre process.

3. σ is supported by diffuse measures only.

This last hypothesis implies that we concentrate only on simple processes. In Chapter 1 we gave the definition of ergodic point process and in Chapter 2 we gave a necessary and sufficient condition on the spectral measure σ for a DαS process Nin $\mathbb{R}^d$ to be ergodic. This condition was the following:

$$n^{-d} \int_{[-n/2,n/2]^d} (1 - e^{-\mu(B+x)}) dx \to 0 \text{ as } n \to +\infty.$$

for all relatively compact Borel set B and all μ from the support of σ.
In the case of uniformly distributed clusters σ is concentrated on a set of measures that have finite support. Hence for each μ in supp(σ) and each Borel set B the integrand function is zero outside a compact set and, obviously, it is bounded. This is enough to guarantee that the condition holds.
In the case of Gaussian distributed clusters the rapid decay of the Gaussian density at infinity is enough to guarantee that the condition holds in the same way. More easily we could also make use of a result in [3] according to which a stationary cluster process is ergodic whenever the cluster centre process has the same property. In our case the centre process is a Poisson homogeneous process which is trivially ergodic.
Hence all the point processes we consider are ergodic. We will work with two parametric families of measures: uniform probability measures on balls of variable radius and Gaussian distributions with arbitrary covariance matrix.

3.2 Estimation of the Sibuya parameter measure

The crucial thing for our statistical inference is the knowledge of the measure μ according to which all the clusters are distributed. In fact, as it will be explained in the next chapter, if we can assume that we know μ, then we have robust methods to estimate λ and α.
The bursty nature of our processes provides a very intuitive way of estimating

μ: if it is possible to identify a big cluster and to isolate it from the rest of the data we can be confident than it would give an accurate estimation for the measure μ.
In principle, when we are given a data set that we want to fit with our model, we could have three possible situations:

- μ is already known: in this case we can use this information to estimate λ and α and we are done;
- μ is unknown but for some reason we know that it belongs to a parametric class (for example it could be Gaussian with unknown covariance matrix or uniform on a ball with unknown radius);
- μ is totally unknown.

This last case is not analyzed in detail in this report: we limit ourselves to notice that if we don't have any previous information about μ, two possible ways are open. The first one is to use standard procedures of kernel density estimation to estimate the measure or, at least, to understand which parametric class it belongs to. The second one is to make direct use of the sample measure that can be obtained by the big cluster.
In our work we concentrate on the case in which μ is unknown but we know that it belongs to a certain parametric class. The synthetic samples we were working with were generated with μ being the uniform measure on a ball of variable radius r or with Gaussian distributed clusters for which we assume the covariance matrix to be multiple of the identity matrix. Hence our goal is simply that of isolating a big cluster and using it to estimate the radius r of the ball (in the first case) or the positive real number σ such that the covariance matrix is $\sigma \mathrm{I}$ in the second case.
We start facing the problem of identifying a big cluster in the data set.

3.2.1 How to identify a big cluster

In order to identify a big cluster in our data set we need to run some clustering procedure. In literature several methods are available but the majority of them are not very suitable for us because of the very special features of our data. In particular our data sets have the following characteristics that should be taken into account:

- the total number of clusters is unknown and we can't even make the hypothesis of having an *a priori* estimation for it since it is related to λ (and we can't provide any estimation of λ before knowing μ.);
- clusters are very different in size;
- it is possible to have some overlapping clusters.

Because of all these features, many clustering methods fail on our data. We found that a good method is given by the combination of an Expectation-Maximization algorithm (often referred to as EM algorithm) and the use of the Bayesian Information Criterion (BIC). In particular we think at our data as coming from a mixture model with an unknown number of components G, we estimate the parameter for this model with the EM algorithm and then we decide which model is the best one (i.e. which is the most likely number of components) comparing values of the BIC. We go a little bit more into details of these ideas, following the article from Fraley and Raftery [5].

3.2.2 The Expectation-Maximization algorithm for mixture models

Let's consider our data set to be $\mathbf{y} = (y_1, ..., y_n)$. When we say that the data are distributed according to a mixture model with G components we mean that the likelihood function is

$$\mathcal{L}_{\text{MIX}}(\theta_1, .., \theta_G; \tau_1, .., \tau_G | \mathbf{y}) = \prod_{i=1}^{n} \sum_{i=1}^{G} \tau_k f_k(y_i|\theta_k),$$

where, for $k \in \{1, .., G\}$, f_k (with parameters θ_k) is the density distribution of the k-th component and we denote by τ_k the probability that an observation belongs to the k-th component ($\sum_{k=1}^{G} \tau_k = 1$). In order to use the EM algorithm we change our data in the following way. We associate to each y_i the couple $x_i = (y_i, \mathbf{z}_i)$ where $\mathbf{z}_i = (z_{i1}, .., z_{iG})$ and

$$z_{ik} = \begin{cases} 1 & \text{if } x_i \text{ belongs to component } k \\ 0 & \text{otherwise} \end{cases}.$$

The vector $\mathbf{x}$ whose components are the couples x_i is called the vector of the *complete data* while $\mathbf{y}$ is the vector of the *observed data*. $\mathbf{x}$ is said to be complete because it contains also information about the division of data into clusters.

The EM algorithm alternates two steps:

- E step: in this step conditional expectation of the log-likelihood given the observed data and the current parameter estimates is computed. In our case expectation for z_{ik} is taken, using current estimations for θ_k and τ_k. In fact the log-likelihood is given by

$$l(\theta_k, \tau_k, z_{ik}|\mathbf{x}) = \sum_{i=1}^{n} \sum_{k=1}^{G} z_{ik} \log[\tau_k f_k(y_i|\theta_k)]$$

  and taking expectation of this expression given the observed data and the current estimations for the parameters requires taking expectation of z_{ik} only. Hence this step can be expressed as

$$\widehat{z}_{ik} \leftarrow \frac{\widehat{\tau}_k f_k(y_i|\widehat{\theta}_k)}{\sum_{j=1}^{G} \widehat{\tau}_j f_j(y_j|\widehat{\theta}_k)}.$$

  Here the expression on the right is exactly the expectation for z_{ik}.

- M step: this step consists of the determination of parameters that maximize the expected log-likelihood coming from the "E step". In our model the log-likelihood should be maximized with respect to θ_k and τ_k.

3.2.3 Model selection via Bayesian Information Criterion

The basic idea of model selection through Bayesian Information Criterion is maximization of the *a posteriori* likelihood.
Let $M_1, .., M_G$ the set of model from which we want to select one. The quantity that we want to maximize is $\mathbf{P}(\mathbf{y}|M_k)$ where k is in the range $\{1, .., G\}$. By the law of total probabilities the latter is obtained by integrating over the parameters:

$$\mathbf{P}(\mathbf{y}|M_k) = \int \mathbf{P}(\mathbf{y}|M_k, \theta_k)\mathbf{P}(\theta_k|M_k)\mathrm{d}\theta_k,$$

where $\mathbf{P}(\theta_k|M_k)$ is the prior distribution for the parameter θ_k of model M_k. An approximation for the logarithm of this integral was given by Schwartz (1978):

$$2 \log \mathbf{P}(\mathbf{y}|M_k) \approx 2 \log \mathbf{P}(\mathbf{y}|\widehat{\theta}_k, M_k) - \nu_k \log(n) = \mathrm{BIC}_k.$$

Here n is the size of the data set while ν_k is the number of parameters to estimate in model M_k. Hence what we do when using the so called BIC criterion, is to select the model M_k which corresponds to the highest value for BIC_k.

3.2.4 Adding a Poisson homogeneous noise to the model

In many applications data present a noise component: there are points which do not belong to any of the clusters and disturb the clustering procedure. To handle these cases it is possible to change the mixture model in such a way that it includes also a Poisson homogeneous noise. Then it is still possible to use EM algorithm and BIC as described above. More specifically what we do is to change our likelihood for the model to be

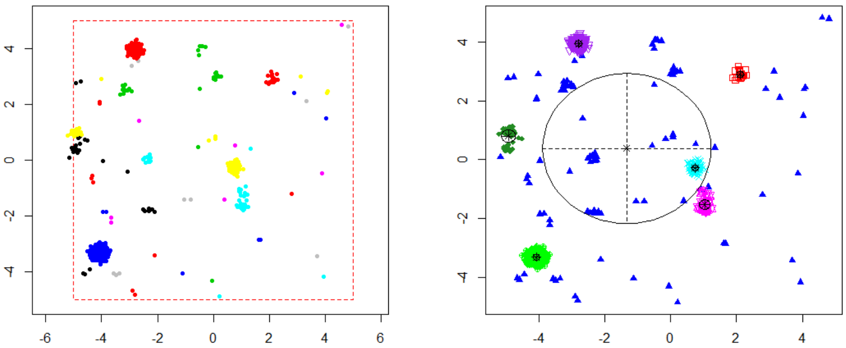
$$\widetilde{\mathcal{L}}_{\text{MIX}}(\theta_1,..,\theta_G;\tau_1,..,\tau_G|\mathbf{y}) = \prod_{i=1}^{n}\Big[\frac{\tau_0}{V} + \sum_{i=1}^{G}\tau_k f_k(y_i|\theta_k)\Big]$$

where V is the area of the data region and τ_0 is the probability that an observation comes from the noise component and the condition $\sum_{k=0}^{G}\tau_k = 1$ holds. This approach will be very useful for our analysis since in our data set we have few very big clusters and many very small clusters (most of them constituted of 1 point only). Hence the latter can actually be modeled as homogeneous Poisson noise and in this way big clusters will be identified more easily.

3.2.5 Discussion of some examples

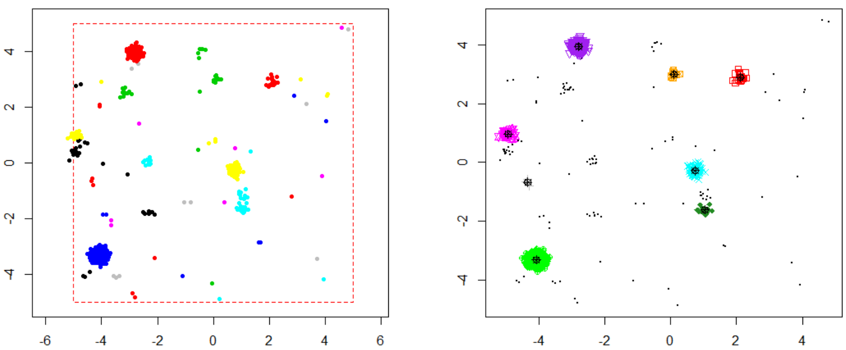
Gaussian distributed clusters

For Gaussian distributed clusters this procedure of model selection through EM algorithm and BIC is implemented in **R** in a library called **mclust** and particularly in a function called **mclustBIC**. We tried it on our synthetic samples to see if we can use it to obtain good estimation for the parameters that characterize the measure μ. Figure 1 shows the comparison between the real division in clusters of a synthetic sample with Gaussian clusters (in 2 dimensional case) and the clustering provided by the algorithm and the BIC selection. We see that the algorithm tends to identify a big and very spread cluster which is instead formed by several very small clusters present in the configuration. This shows why it is a good idea to add a component of Poisson homogeneous noise to the model. Figure 2 shows the output obtained in this case. Not only the big clusters are well separated and the small one are removed, we also have that the algorithm is considerably less expensive in terms of computation required. An intuitive reason for this is that the points which are quite isolated are simply excluded and the algorithm does



(a) Original process (b) Clustered process

Figure 1: Clustering according to mixture model without noise correction for DαS process with Gaussian clusters: $\lambda = 0.5$, $\alpha = 0.6$, covariance matrix $(0.1)^2$I



(a) Original process (b) Clustered process

Figure 2: Clustering according to mixture model with noise correction for DαS process with Gaussian clusters: $\lambda = 0.5$, $\alpha = 0.6$, covariance matrix $(0.1)^2$I

not have to try to fit them into a particular cluster.
When we have a reliable division in clusters we can isolate the biggest one (or more than one) and use the selected data to provide estimation for the covariance matrix. Some estimations for the covariance matrix of the different clusters are included in the output of the function **mclustBIC**. Otherwise the more general function **fitdistr** in the **R** library **MASS** could be used. The whole procedure can be repeated in 1-dimensional case. Moreover it can be easily extended to the more general case when no assumptions on the covariance matrix are made.

Uniformly distributed clusters

For uniformly distributed clusters the idea is basically the same: to use EM algorithm for mixture models and BIC model selection. Also in this case it is very useful to correct the model adding a homogeneous Poisson noise because in this way we can get rid of very small clusters present in the configuration.
After selecting the biggest clusters we can use them to estimate our parameter. In this case the parameter is the radius of the ball where the points are distributed. A simple way to do this is the following. First an estimation of the centre of the ball is obtained averaging out the positions of the different points in the cluster: if the number of points in the cluster is big enough, this estimation is reliable. Then we estimate the radius of the ball using the proposition below.

Proposition 8. *Let $x \in \mathbb{R}^2$ and let $\{X_i\}_{i=1}^n$ be i.i.d. observations of a random variable $X \sim U(B_R(x))$. If $R_{\text{MAX}} = \max_{i \in \{1,..,n\}} \|X_i - x\|_2$, then*

$$\frac{2n+1}{2n} R_{\text{MAX}}$$

is an unbiased estimator for R.

Proof. The result follows easily from

$$\mathbf{E}[R_{\text{MAX}}] = \int_0^R r \mathrm{d}\mathbf{P}(\{R_{\text{MAX}} \leq r\})$$

and

$$\mathbf{P}((\{R_{\text{MAX}} \leq r\})) = (P(|X_1 - x| \leq r))^n = \Big(\frac{\pi r^2}{\pi R^2}\Big).$$

□

In the same way, in the case of 1-dimensional clusters the radius of the ball can be estimated using the following:

Proposition 9. *Let* $x \in \mathbb{R}$ *and let* $\{X_i\}_{i=1}^n$ *be i.i.d. observations of a random variable* $X \sim U(B_R(x))$*. If* $X_{\text{MAX}} = \max_{i\in\{1,..,n\}} X_i$ *and* $X_{\text{MIN}} = \min_{i\in\{1,..,n\}} X_i$*, then*

$$\frac{n+1}{n}\Big(\frac{X_{\text{MAX}} - X_{\text{MIN}}}{2}\Big)$$

is an unbiased estimator for R*.*

This last result can be proved exactly in the same way as the previous one.

Chapter 4

Estimation of the stability parameter α and of the intensity λ of the centre process

This chapter is again in the framework of the previous one and all the hypotheses of Section 3.1 are assumed.
Here we will suppose that we already know μ and we will concentrate on the task of estimating the Sibuya exponent α and the Poisson parameter λ.

4.1 Estimation of void probabilities

4.1.1 Spherical contact distribution function

Since the process we are studying is a simple point process, its distribution is uniquely determined by the void probabilities (see Prop. 4, Chapter 1). Our strategy will be exactly that of estimating void probabilities and using them to make inferences on the parameters. Let N be a point process on $\mathbb{R}^d$ satisfying all the hypotheses listed in the beginning of the chapter. We will make use of the following:

Definition 27 (Spherical contact distribution function)**.** *Give a point process N on $\mathbb{R}^d$, for $r > 0$ we define the spherical contact distribution function to be*

$$H_s(r) = 1 - \mathbf{P}\{N(S_r(\mathbf{0})) = 0\}.$$

This means that $H_s(r)$ is the probability of having in the configuration a point whose distance from the origin is less than r. Obviously the definition is of particular interest when the point process N is stationary and the origin can actually be replaced by any point in $\mathbb{R}^d$. The quantity $\mathbf{P}\{N(S_r(\mathbf{0})) = 0\}$ which appears in the definition is the void probability we are interested in. We can rewrite $H_s(r)$ as follows:

$$H_s(r) = \mathbf{P}\{\text{supp}(N) \cap S_r(\mathbf{0}) \neq 0\} = Pr\{0 \in \text{supp}(N) \oplus S_r(\mathbf{0})\},$$

where $\oplus$ denotes the Minkowski sum i.e. the operation defined on any couple of sets A, B as $A \oplus B = \{x + y | \ x \in A, \ y \in B\}$.
Now consider a finite sequence of points $\{x_i\}_{i=1}^n$, which we will refer to as *test points*, in a set where we can assume stationarity for the process (for our purposes we need $\mathbf{P}\{x_i \notin \text{supp}(N) \oplus S_r(\mathbf{0})\}$ to be independent of the position of x_i at least for small values of r). We define an unbiased estimator for the void probability function:

Proposition 10. *Let N be a point process on $\mathbb{R}^d$ and $\{x_i\}_{i=1}^n$ a sequence of test points included in a set where we can assume stationarity for the process. For each $i \in \{1, .., n\}$ let $r_i = \text{dist}(x_i, \text{supp}(N))$. The estimator $\widehat{G}(r)$ defined as*

$$\widehat{G}(r) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{r_i > r\}} .$$

is an unbiased estimator for the void probability $\mathbf{P}\{N(S_r(\mathbf{0}) = 0)\}$.

Proof.

$$\begin{aligned}
\mathbf{E}[\widehat{G}(r)] &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}\left[\ \mathbb{I}_{\{r_i > r\}} \right] = \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbf{P}\{x_i \notin \text{supp}(N) \oplus S_r(\mathbf{0})\} = \\
&= 1 - H_s(r) = \mathbf{P}\{N(S_r(\mathbf{0}) = 0)\}.
\end{aligned}$$

We notice that in the second last equality, we used the stationarity property.
$\square$

In the previous chapter we recalled that the void probabilities of a $D\alpha S$ point process are given, for any Borel set B, by

$$\mathbf{P}\{N(B) = 0\} = \exp\Big(- \int_{\mathbb{S}} \mu(B)^\alpha \sigma(d\mu)\Big).$$

Under the hypotheses given in the beginning of the chapter we have

$$\mathbf{P}\{N(B) = 0\} = \exp\Big\{ - \int_A \lambda\big(\mu_x(B)\big)^\alpha dx\Big\}. \tag{4.1}$$

If we can get explicit version of this last formula dependently on α and λ then we are done. In fact we could run an optimization procedure and look for the values of α and λ which minimize the "distance" between the real void probabilities $p_{\alpha,\lambda}(r)$ and the estimated $\widehat{G}(r)$. In the following we will discuss which notion of distance we would better adopt.

4.1.2 Example: uniformly distributed clusters in 1-dimensional case

We now consider the specific case where μ is the uniform measure on the ball $B_1(0)$ and we assume the dimension of the space to be $d = 1$. Under these hypotheses $\mu_x(\cdot)$ denotes the uniform probability measure on $S_1(x)$. We take A as the symmetric window $[-a/2, a/2]$ for some large enough a and explicit calculation of the integral above leads to

$$p_{\alpha,\lambda}(r) \equiv \mathbf{P}\{N(B_r(0)) = 0\} = \begin{cases} \exp(-2\lambda r^\alpha(1 - r + \frac{2r}{\alpha+1})) & \text{if } r < 1 \\ \exp(-2\lambda(r - 1 + \frac{2}{\alpha+1})) & \text{if } r \geq 1 \end{cases}.$$

As a consequence we can easily implement the method described for this kind of processes.

Here there are some questions that arise naturally and that we will try to answer. How many test points should be used? How should they be laid? How big our data set should be in order to get an accurate estimation? We will answer directly to those questions in the next section. Here we start showing some results obtained trying this first method on some synthetic samples we generated.

In our simulations, when the width of the window considered is a, the center process is generated in the window $[-a/2, a/2]$. Since each cluster is uniformly distributed on a ball of radius 1, we could assume that the process was stationary in $[-a/2 + 1, a/2 - 1]$. We also observed that the void probabilities are usually very close to zero for $r > 2$ and so we generate the sequence $\{x_i\}_{i=1}^n$ in the window $[-a/2 + 3, a/2 - 3]$. Alternatively we should take edge-effects into account (see subsection below). Figure 1 shows approximations for void probabilities obtained with points randomly chosen and increasing density. As we could expect the approximations get better

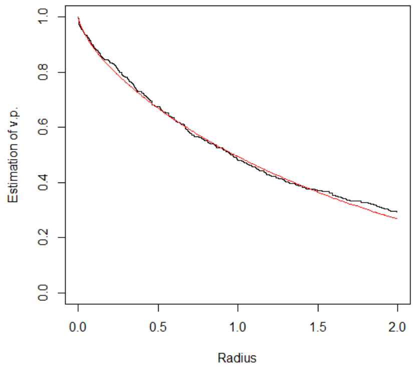
when the density of test points increases. Figure 2 repeat the same test but with regularly distributed test points (i.e. they are on a grid with decreasing mesh) and we can observe the same improving effect. Finally Figure 3 shows approximations for the void probabilities obtained considering an increasing quantity of data: from the same configuration we extract of size 200 in the first case and we enlarge it to be of size 500 and 1000 in the second and in the third case. We consider these windows in such a way that each one is contained in the next one. Since the behavior of our processes is highly irregular we could expect the strong improving in the estimations that is shown in our plots.
More comments about these first results will be given in the next section.

4.1.3 Consistency of the estimator

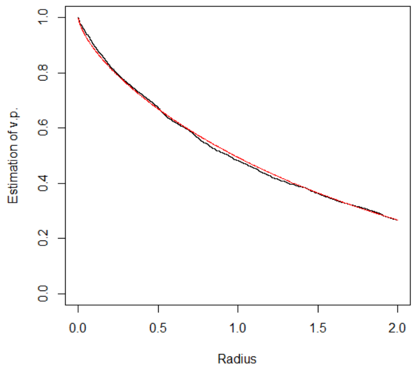
Here we begin to express in a more formal way the observations above by studying the consistency of our estimator. Namely we want to verify if the variance of our estimator tends to zero when the amount of data used for the estimation increases. In the following let l be the length of the window (in 1-dimensional case) or the area (in 2-dimensional case) and k the number of test points per unity (i.e. $n = k \cdot l$).
What we need is to calculate the variance of

$$\widehat{G}(r) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{r_i > r\}} \,.$$

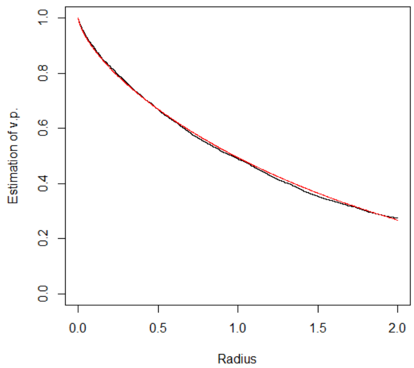
Each term of the sum is a Bernoulli random variable with parameter $p_{\alpha,\lambda}(r)$, i.e. the probability we are interested in. This makes it very simple to calculate the variance of each term but, without any further hypotheses we can't assume independence between the terms of this sum and we have to take



(a) 1 test point per unity

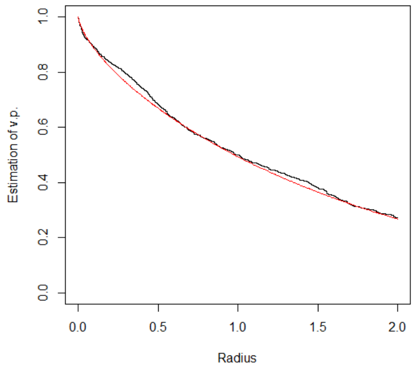


(b) 5 test points per unity

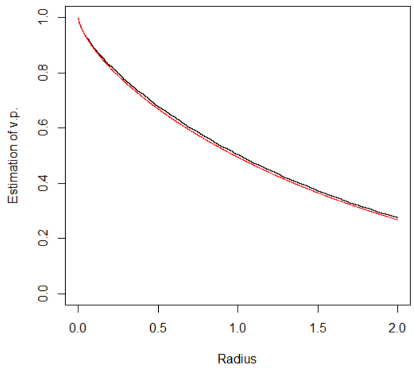


(c) 10 test points per unity

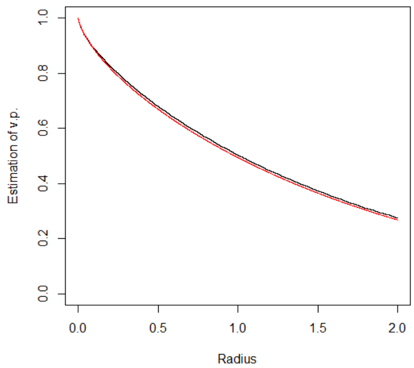
Figure 1: Estimations for the void probabilities: $\alpha = 0.7$, $\lambda = 0.3$, width of the window=500, test points randomly chosen



(a) 1 test point per unity

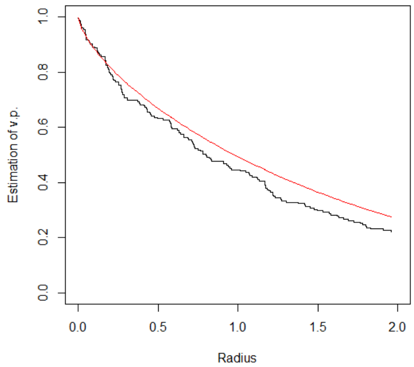


(b) 5 test points per unity

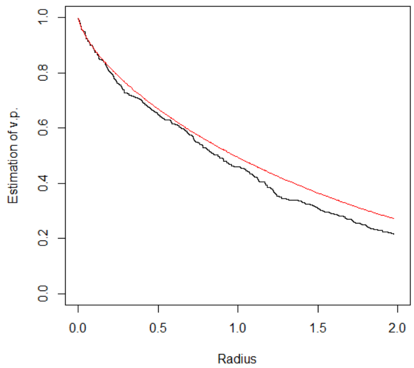


(c) 10 test points per unity

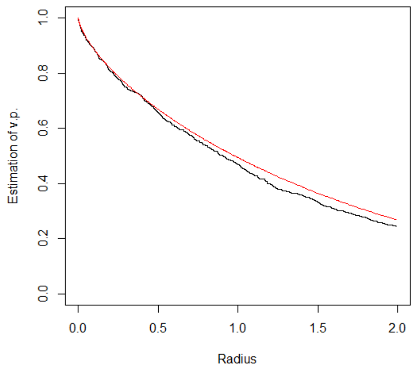
Figure 2: Estimations for the void probabilities: $\alpha = 0.7$, $\lambda = 0.3$, width of the window=500, test points on a grid



(a) width of the window=200



(b) width of the window=500



(c) width of the window=1000

Figure 3: Estimations for the void probabilities: $\alpha = 0.7$, $\lambda = 0.3$, 1 test point per unity, test points on a grid

into account also the covariance terms:

$$
\begin{aligned}
\mathbf{var}(\widehat{G}(r)) &= \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbf{cov}(\mathbb{1}_{\{r_i>r\}},\mathbb{1}_{\{r_j>r\}}) = \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{var}(\mathbb{1}_{\{r_i>r\}}) + \sum_{i\neq j}\mathbf{cov}(\mathbb{1}_{\{r_i>r\}},\mathbb{1}_{\{r_j>r\}}) = \\
&= \frac{1}{n}\,\mathbf{var}(\mathbb{1}_{\{r_1>r\}}) + \\
&\qquad + \frac{1}{n^2}\sum_{i\neq j}\Big[\mathbf{E}[\mathbb{1}_{\{r_i>r\}}\cdot\mathbb{1}_{\{r_j>r\}}] - \mathbf{E}[\mathbb{1}_{\{r_i>r\}}]\cdot\mathbf{E}[\mathbb{1}_{\{r_j>r\}}]\Big] = \\
&= \frac{1}{n}p_{\alpha,\lambda}(r)(1-p_{\alpha,\lambda}(r)) + \\
&\qquad + \frac{1}{n^2}\sum_{i\neq j}\Big[\mathbf{P}\{N(S_r(x_i))=0, N(S_r(x_j))=0\} - p_{\alpha,\lambda}(r)^2\Big]
\end{aligned}
$$

To complete our calculation we need to compute or to estimate the probability $\mathbf{P}\{N(S_r(x_i)) = 0, N(S_r(x_j)) = 0\}$. Let's suppose, for example, that the maximum radius we are to consider is $r = 1$. Then for each x_i we have a maximum of $4k$ different points $x_j \geq x_i$ for which this probability does not factorize and with a very rough estimate we can say

$$
\begin{aligned}
\mathbf{var}(\widehat{G}(r)) &\leq \frac{1}{k\cdot l}p_{\alpha,\lambda}(r)(1-p_{\alpha,\lambda}(r)) + \frac{8}{l}p_{\alpha,\lambda}(r)(1-p_{\alpha,\lambda}(r)) = \\
&= \left(\frac{1}{k\cdot l} + \frac{8}{l}\right)p_{\alpha,\lambda}(r)(1-p_{\alpha,\lambda}(r))
\end{aligned}
$$

The last formula shows that increasing the density k of test points brings to a smaller variance but, for the variance to tend to zero what is really needed is l going to infinity. In fact we can rewrite the last term in the series of inequalities as

$$
\frac{8+1/k}{l}p_{\alpha,\lambda}(r)(1-p_{\alpha,\lambda}(r))
$$

and in this formulation we note that the upper limit for the variance decreases much more quickly when we increase l than when we increase k. If we look back to Figures 1 and 3 we see that there is even no big difference, for example, between approximation obtained with 5 or 10 points for unity. Hence we conclude that it is not really worth considering more than 5 test points per unity.

Looking again at Figures 1 and 2 we note that there is no big difference between the goodness of the approximations obtained with points laid on a grid or chosen randomly. Anyway the better choice is to lay our test points on a grid. In this way we avoid to introduce in our estimations a randomness that is not really needed and that could also influence our results if, for example, we happen to have areas with a too high density of test points. This can be easily understood from the calculations we have just done for the variance of our estimator. Randomness in the choice of the points would result in additional terms in the expression.

4.1.4 Estimation of α and λ

We estimate α and λ optimizing the "distance" between the estimated void probabilities and the function $p_{\alpha,\lambda}(r)$. As above let $r_i = \text{dist}(x_i, \text{supp}(N))$ and $r_{(1)}, ..., r_{(n)}$ the set of the ordered $r_1, ..., r_n$. We consider the two following notions of distance:

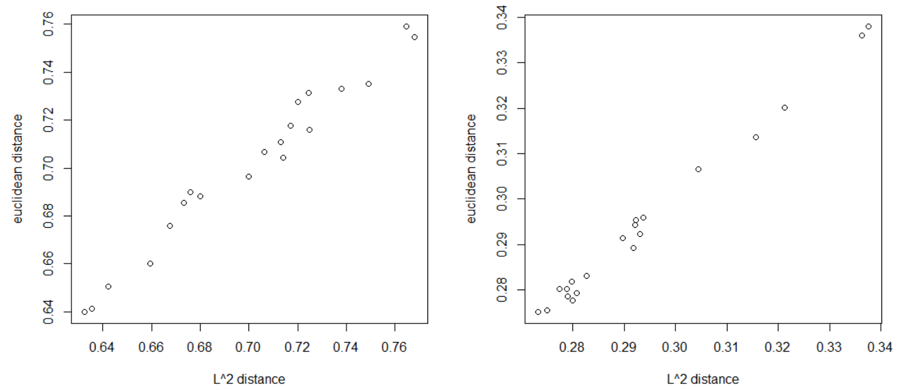
1. *L^2-distance*
   We minimize L^2 distance between the exact void probabilities and the step function $\widehat{G}(r)$:

$$(\widehat{\alpha}_1, \widehat{\lambda}_1) \doteq \arg \min_{\alpha\in(0,1),\lambda\in\mathbb{R}_+} \int_{r_{(1)}}^{r_{(n)}} |\widehat{G}(r) - p_{\alpha,\lambda}(r)|^2 dr =$$
$$= \arg \min_{\alpha\in(0,1),\lambda\in\mathbb{R}_+} \sum_{i=1}^{n-1} \int_{r_{(i)}}^{r_{(i+1)}} \left|\widehat{G}(r_{(i)}) - p_{\alpha,\lambda}(r)\right|^2 dr =$$
$$= \arg \min_{\alpha\in(0,1),\lambda\in\mathbb{R}_+} \sum_{i=1}^{n-1} \int_{r_{(i)}}^{r_{(i+1)}} \left|\frac{n-i}{n} - p_{\alpha,\lambda}(r)\right|^2 dr.$$

2. *Euclidean distance*
   We minimize the euclidean distance between the vector of the exact values of $p_{\alpha,\lambda}$ and the approximated values in the same points:

$$(\widehat{\alpha}_2, \widehat{\lambda}_2) \doteq \arg \min_{\alpha\in(0,1),\lambda\in\mathbb{R}_+} \sum_{i=1}^{n-1} \left|\widehat{G}(r_{(i)}) - p_{\alpha,\lambda}(r_{(i)})\right|^2 =$$
$$= \arg \min_{\alpha\in(0,1),\lambda\in\mathbb{R}_+} \sum_{i=1}^{n-1} \left|\frac{n-i}{n} - p_{\alpha,\lambda}(r_{(i)})\right|^2.$$

Figure 4 shows the relation between the estimations obtained with the two different definitions of distance in the case of uniformly distributed clusters in 1-dimensional case ($\alpha = 0.7$, $\lambda = 0.3$, width of the window=1000). It seems that both for α and λ the points follow the diagonal of the plot and



(a) $\widehat{\alpha}$

(b) $\widehat{\lambda}$

Figure 4: Estimation obtained with different distances: L^2 on the x axis and euclidean distance on the y axis

this means that the two distances give similar estimations. Anyway we think that, even if the first notion is more natural when we want to fit a function, the second one, i.e. the euclidean distance, should be preferred. In fact, since the radii $\{r_{(i)}\}$ are not placed on an homogeneous grid, we think that the first method tends to weight too much errors related to values of r which lie in an area where we have low density of the sequence $\{r_{(i)}\}$.

4.1.5 Edge-correction

If, as it is likely to be, we can observe the process only in a limited window W, while computing our statistics we should take into account edge-effects. One way of doing it could be the following. If we distribute randomly the test points $x_1, ..., x_n$ in the window W it may happen that some of them, let's say x_i, fall close to the border of the window in such a way that $B_r(x_i)$ is not entirely contained in the window. In this case we should change our statistic in the following way:

$$\widehat{G}(r) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{r_i > r\}}\, \mathbf{P}\{N(B_r(x_i) \cap W^c) = 0 | N(B_r(x_i) \cap W) = 0\} =$$
$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{r_i > r\}}\, \frac{\mathbf{P}\{N(B_r(x_i)) = 0\}}{\mathbf{P}\{N(B_r(x_i) \cap W) = 0\}}.$$

However, since the weights

$$\frac{\mathbf{P}\{N(B_r(x_i)) = 0\}}{\mathbf{P}\{N(B_r(x_i) \cap W) = 0\}}$$

still depend on the parameters we have to estimate we can't compute them explicitly. It is natural to substitute them with the percentage of $B_r(x_i)$ that is contained in the window. Our statistic will be:

$$\widehat{G}(r) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{r_i > r\}} \frac{|B_r(x_i) \cap W|}{|B_r(0)|}.$$

In 1-dimensional case with the window $W = [-a/2, a/2]$ we have $\frac{|B_r(x_i) \cap W|}{|B_r(0)|} = r - |x_i - a/2|$.

4.2 Void probabilities of the p-thinned process

In the previous section we described a strategy that, under the hypothesis that we know μ, could provide good approximations for α and λ both in 1-dimensional and 2-dimensional case whatever measure μ is. However we saw that in order to get good approximations for our parameters we need to have a very big data set. So the question arises if it is possible to improve our estimation with a statistic that takes into account more information coming from our data set. In fact the previous statistic was using only a little part of the information available, that is only the distances from the test points to the closest point into the configuration. For example we can try to look for a statistic that takes into account also distances from test points to the second closest point in the configuration, to the third, and so on.

Furthermore we saw that the method described in the previous section required explicit knowledge of the void probability function $p_{\alpha,\lambda}(r)$. This is something that in principle we can have without any difficulty, provided that we know the measure μ according to which points within any single clusters are distributed. On the other hand, apart from very simple cases such as the case of uniformly distributed clusters in 1-dimensional case, a computational problem arise because of the large number of numeric integrations required. In this section we describe a strategy to estimate both λ and α which is a good answer to both of these problems. It is based on the knowledge of the expression for the void probabilities of a p-thinned point process and on the use of a natural statistic to estimate them. This statistic involves distances from test points also to the second closest point in the configuration, the third, and so on.

4.2.1 Properties of the p-thinned process

Let's start by recalling that the p.g.fl. of the point process we are studying has the form

$$G_N[h] = \exp\Big\{ - \int_{\mathbb{S}} \langle 1-h, \mu\rangle^{\alpha} \sigma(d\mu)\Big\}, \quad 1-h \in \mathrm{BM}(\mathcal{X}),$$

where σ, as in the beginning of this chapter, is such that, for any Borel set $B \in \mathcal{B}(A)$,

$$\sigma(\{\mu_x | x \in B\}) = \lambda \cdot \nu(B).$$

As we have seen in Chapter 2, the p.g.fl. of a p-thinned point process is

$$G_{p\circ N}[h] = G_N[p \cdot h + 1 - p], \quad 1-h \in \mathrm{BM}(\mathcal{X}).$$

In our case this formula gives

$$G_{p\circ N}[h] = \exp\Big\{ - p^{\alpha} \int_{\mathbb{S}} \langle 1-h, \mu\rangle^{\alpha} \sigma(d\mu)\Big\}, \quad 1-h \in \mathrm{BM}(\mathcal{X}).$$

This means that, after applying p-thinning, we have a process that is distributed as the original one, apart from the new Poisson parameter

$$\lambda_{new} = \lambda \cdot p^{\alpha}.$$

Hence if we had, dependently on the parameters, an explicit expression for the void probabilities of the original process at a given radius r we could have also the expression for the new process and the same radius without any additional effort than the change of an exponent in formula 4.1.

4.2.2 Definition of an estimator

What is interesting about the strategy we are to present is the fact that, in order to estimate void probabilities for the p-thinned process we don't need to simulate p-thinning on the original configuration, all the information we need is already in our data.
Here we describe what our strategy will be. Let's consider some test points $\{x_i\}_{i=1}^n$. According to what we discussed in the previous section it is more convenient to lay test points on a grid, obviously in a set where we can

assume the stationarity of the process. For each point x_i we now consider the sequence $\{r_i^{(k)}\}$ of the distances from x_i to the k-th closest point in the configuration. In this setting $r_i^{(1)}$ is exactly what was denoted with r_i in the previous section. The following proposition shows how we can use these new sequences we have defined to identify an unbiased estimator for the void probabilities of the p-thinned process.

Proposition 11. *Let N a DαS point process satisfying all the hypotheses listed in the beginning of the chapter. Suppose to have a sample configuration constituted by m points in a window W such that $\mathbf{0} \in W$ containing m points . Let $\{x_i\}_{i=1}^n$ a sequence of test points as above and the sequences $\{r_i^{(k)}\}$ as we have just defined them. Then*

$$\widehat{H}(r) = \frac{1}{n} \sum_{i=1}^{n} \Big(\sum_{k=1}^{m} p(1-p)^{k-1} \, \mathbb{I}_{\{r_i^{(k)} > r\}} \Big)$$

is an unbiased estimator for $\mathbf{P}\{p \circ N(S_r(0)) = 0\}$.

Proof. We can express the void probabilities for the p-thinned process in the following way:

$$\begin{aligned} &\mathbf{P}\{p \circ N(S_r(0)) = 0\} = \\ &= \sum_{k=1}^{m} \mathbf{P}\{\text{“the closest survived point is the k-th”}\} \mathbf{P}\{r_i^{(k)} > r\}, \end{aligned}$$

where we used the fact that the probability of $S_r(0)$ being empty conditional on the fact that the closest survived point is the k-th is the probability of $r_i^{(k)}$ being greater than r. Furthermore we have

$$\mathbf{P}\{\text{“the closest survived point is the k-th”}\} = p(1-p)^{k-1}$$

and

$$\mathbf{P}\{p \circ N(S_r(x_i)) = 0\} = \sum_{k=1}^{N} p(1-p)^{k-1} \mathbf{P}\{r_i^{(k)} > r\}.$$

Hence

$$\mathbf{E}\Big[\sum_{k=1}^{N} p(1-p)^{k-1} \, \mathbb{I}_{\{r_i^{(k)} > r\}} \Big] = \mathbf{P}\{p \circ N(S_r(0)) = 0\}$$

and our thesis follows. □

4.2.3 Implementation

We observe that when k becomes bigger the quantity $p(1-p)^{k-1}$ tends to zero. So we can choose to make an approximation in our estimator and to consider only values of k such that $p(1-p)^{k-1}$ is greater than a fixed value that we can choose to be arbitrary small. We tried the method on synthetic samples both in the case of Gaussian distributed clusters and uniformly distributed clusters and in 1 and 2 dimensional case. What we did was simply to compute estimations for the void probabilities of $p \circ N$ in $r = 1$ for a sequence of different values of retention probability p, e.g. $p \in \{1/30, 2/30, .., 29/30\}$. The choice of r is arbitrary but obviously it is good idea to choose a value that is approximately in the middle of the range of distances which is "typical" for the original process. Finally the **R** procedure **optim** can be used again to optimize the euclidean distance between the estimated values and the real ones.

In Figure 5 we show an approximation of void probabilities (in black) and the real values (in red) for the same case of uniform distributed clusters in 1-dimensional case and the same parameters for which we showed approximation in the previous section. In this case the **R** procedure **optim** gave $\widehat{\alpha} = 0.72$ and $\widehat{\lambda} = 0.29$ while the real values we used to generate the process where $\alpha = 0.7$ and $\lambda = 0.3$. We note that in this second method the amount of data needed to get a good approximation of the real void probabilities function is smaller. For example, if we consider the same synthetic sample that we used in the last approximation of Figure 3 we obtain the approximation of the void probabilities shown in black in Figure 5. The next table shows the estimated values for the parameters obtained with the two different methods.

	$\widehat{\alpha}$	$\widehat{\lambda}$
Real Values	0.7	0.3
Estimated with 1^{st} method	0.744	0.333
Estimated with 2^{nd} method	0.699	0.290

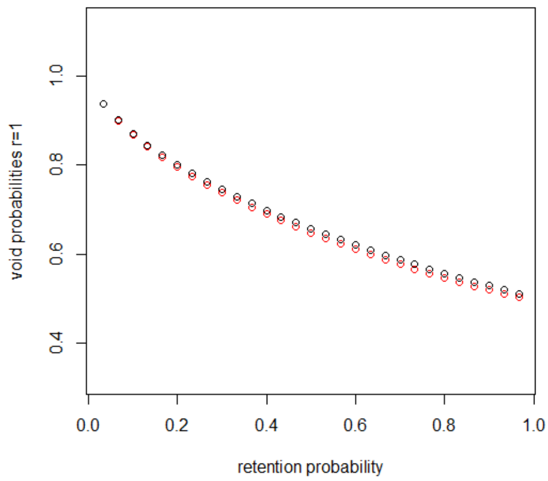


Figure 5: Uniformly distributed clusters in 1 dimensional case. $\lambda = 0.3$, $\alpha = 0.7$, width=1000, density of test points=1

Appendix: R codes

Here we present some of the **R** codes that we implemented for our simulations.
In the first section we have programs dealing with the generation of synthetic samples. As an example we show the codes we wrote for uniformly distributed clusters in 1-dimensional case. The case of Gaussian clusters and the 2-dimensional case were handled exactly in the same way.
In the second section codes for different parameter estimations are presented again only in the case of 1-dimensional uniformly distributed clusters since the other cases required just few modifications.

Generation of synthetic samples

First of all we needed a function to simulate Sibuya random variables. Our variable is simulated as a sequence of Bernoulli variables with decreasing probability of success $\alpha, \alpha/2, \alpha/3, ...$ The value of the simulated variable is the position of the first success that is the first 1 in the series of the Bernoulli trials.

```
rsib=function(n,alpha){

#n=number of sibuya distributed numbers generated
#alpha=sibuya exponent

x=rep(0, n) #initialization

for(j in 1:n){
        i=rbinom(1,1,alpha) #first trial
        k=1
        while(i!=1){
        k=k+1
```

```
            i=rbinom(1,1,alpha/k)
            }
            x[j]=k;
}
return(x)
}
```

The following code was needed to generate a 1-dimensional synthetic sample for a DαS process satisfying the hypotheses stated in the beginning of Chapter 3 with spectral measure concentrated on the set of uniform measures on balls $B_1(x)$, $x \in A \subseteq \mathbb{R}$. In this code we also define a new *object class* and we call it *clpr*. This is an easy and efficient way to work with our synthetic samples of processes. In fact it allows us to access very quickly all the information we have on the configuration. In the last part of the code some basic functions for these new objects are defined. They are the following:

- **is.clpr**: checks if an object is in the *clpr* class
- **print.clpr**: prints just basic information about the process (number of points in the configuration and number of clusters)
- **summary.clpr**: prints more detailed information about the configuration (number of points in each cluster, intensity of the Poisson centre process, width of the window in which the process was generated)
- **plot.clpr**: plots the points in the configuration on a $2-d$ plot: points in different clusters are given different values of the y coordinate and different colors.

The notation *FunctionName.clpr* is standard notation in **R** to name functions that work on a particular class of objects.

```
#NOTATIONS:
#a=width of the window
#lambda=intensity of the centre process
#alpha=exponent ot the Sibuya distribution


clpr.un=function(lambda, alpha, a=100){ #generates a process
```

```
  n=rpois(1,lambda*a) #number of centres
  xp=runif(n,-a/2,a/2) # parents' coordinate
  ncl=rsib(n,alpha) # no.s of points in each cluster
  m=rep(1:n,ncl) # marks (cluster number) of the points

  # coordinates of points with the corresp mark m of the clusters:
  pts=cbind(runif(sum(ncl),rep(xp,ncl)-1, rep(xp,ncl)+1), m)
  #xm=matrix(pts[-a/2< pts[,1] & pts[,1]<a/2],ncol=2) # select points in the window
  #nclv=as.vector(table(xm[,2])) #number of visible points for each cluster

  # definition of an object of clpr class
  clprl=list(xm=pts,xp=xp,ncl=ncl,lambda=lambda,alpha=alpha,a=a)
  class(clprl)="clpr"

  return(clprl)
}


is.clpr=function(pr) #function to test if an object as clpr class
  return(class(pr)=="clpr")


print.clpr=function(pr)
  {
    if(!is.clpr(pr))
      stop("The argument is not clpr-class. Aborting.")
    cat("Total number of points:",length(pr$xm[,1]),"(see entry$xm for details)\n")
    cat("Number of clusters:",length(pr$xp), "\n")
  }


summary.clpr=function(pr)
        {
        if(!is.clpr(pr))
      stop("The argument is not clpr-class. Aborting.")
```

```
        cat("Width of the window in which the centre process
             is simulated:", pr$a, "\n")
        cat("Intensity of the Poisson process:", pr$lambda,
            "\n")
        cat("Total number of points:",length(pr$xm[,1]),"\n"
            )
        cat("Number of clusters:",length(pr$xp), "\n")
        cat("Number of points in each clusters:", pr$ncl, "\
            n")
        cat("Number of visible points in each clusters:", pr
            $nclv, "\n")
        }


plot.clpr=function(pr){
        if(!is.clpr(pr))
        stop("The argument is not clpr-class. Aborting.")
        plot(c(min(pr$xm[,1],-pr$a/2)-0.1,max(pr$xm[,1],pr$a
            /2)+0.1),c(-0.1,1.1),ty="n",asp=1,xlab="",ylab=
            "")
        lines(c(-pr$a/2,-pr$a/2),c(-0.1,1.1), col="red", lty
            =2)
        lines(c(pr$a/2,pr$a/2),c(-0.1,1.1), col="red", lty
            =2)
        points(pr$xm[,1],pr$xm[,2]/length(pr$ncl),pch=20,col
            =pr$xm[,2])
}
```

All the previous functions needed to be just slightly modified for 2-dimensional case and for the case of Gaussian clusters.

Parameter estimation

The first code we present is the one we used to estimate void probabilities through the first method we described in Chapter 3. First we have a function that receives as impute a synthetic configuration and the real values of the parameters according to which was generated. It returns two vectors: one contains the radii r where the estimation of the void probabilities was computed, the other is the vector of the corresponding estimations. This functions have options to decide how the test points are laid. It is also possible to obtain a plot which compares estimated values and real values (see Figures 1 - 3). Then two functions to compute estimations of the parameters are given: they both need the 2 vectors returned by the previous function as input and return respectively, estimation in L^2 distance and in the euclidean vectorial distance.

```
#NOTATIONS:
#pr=process
#k=number of test points per unity
#lambda=intensity of the centre process
#alpha=sibuya exponent
#a=width of the window

vp.est.unif=function(lambda,alpha,pr,k=1,grid=TRUE,plot=TRUE
    ){#estimates vp
        #k=number of test points per unity
        #grid: if TRUE points are laid on a uniform grid, if
              FALSE they are chosen randomly
        #plot: if TRUE will give the plot with the estimated
              v.p. in black and the real v.p. in red

        rmax=2 #maximum radius we are to consider

        #test points
        if(grid==TRUE){
                pt=seq(-pr$a/2+1+rmax,pr$a/2-1-rmax,by=1/k)
                }
        else{
                pt=cbind(runif(k*(pr$a-2-2*rmax),-pr$a/2+1+
                      rmax,pr$a/2-1-rmax))
        }

        n=length(pt)

        d1=matrix(rep(pr$xm[,1],times=n),ncol=n)
        d2=matrix(rep(pt,times=nrow(pr$xm)),nrow=nrow(pr$xm)
            ,byrow=TRUE)
```

```
        d=apply(abs(d1-d2),2,min) #calculates the distance between test points and the process
        e=sort(d[d<rmax]) #ordered distances
        y=(n-seq(1,length(e),by=1))/length(pt) #estimation of vp for each value in e

        if(plot==TRUE){
        plot(c(0,rmax), c(0,1),ty="n",xlab="Radius",ylab="Estimation of v.p.")
        points(e,y,ty="s")
        points(e,vp(e,alpha,lambda), col="red",ty="l")
        }

        l=list(radii=e,est=y)
        return(l)
}


alphalambda.est.1=function(e,y){ #estimates alpha using L^2 distance between v.p.
g=function(a){ #calculates the distance
        q=numeric()
        vp1=function(ro) evp1(ro,a[1],a[2])
        vp2=function(ro) evp2(ro,a[1],a[2])
        i=1
        while(e[i+1]<1 & i+1<=length(e)){ #integrate the exact v.p.
                h=function(x) (abs(vp1(x)-y[i]))^2
                q[i]=as.numeric(integrate(h,e[i],e[i+1])[1])
                i=i+1
        }
        if(i+1 != length(e)){
                while(i+1<=length(e)){
                        h=function(x) (abs(vp2(x)-y[i]))^2
                        q[i]=as.numeric(integrate(h,e[i],e[i+1])[1])
                        i=i+1
                }
        }
        return(sum(q))
}
```

```
return(optim(c(1,1),g))#returns the values of alpha and lambda which minimizes the distance

}


alphalambda.est.2=function(e,y){ #estimates alpha using euclidean distance between v.p.
g=function(a){
      sum((c(evp1(e[e<=1],a[1],a[2]),evp2(e[e>1],a[1],a[2]))-y)^2) #calculates the squared euclidean distance
}

return(optim(c(1,1),g)) #returns the values of alpha and lambda which minimizes the distance

}


evp1=function(r,alpha,lambda){ #exact void probabilities for r<1
      b=exp(-2*lambda*(r^(alpha))*(1-r+2*r/(alpha+1)))
      return(b)
      }


evp2=function(r,alpha,lambda){ #exact void probabilities for r>1
      b=exp(-2*lambda*(r-1+2/(alpha+1)))
      return(b)
      }


vp=function(r,alpha,lambda){ #exact void probabilities
    if(r<1){
      b=evp1(r,alpha,lambda)
      }
      else{b=evp2(r,alpha,lambda)}
      return(b)
```

}

Finally the following code was used to estimate void probabilities of the thinned process. The main function receives a process as input and returns estimations for the parameters, estimated values for the void probabilities and exact values. Then we have an auxiliary function which computes the estimator for different values of retention probability p. Since the number of points in the configuration can be huge, one of the sum which is present in the definition of our estimator can have many terms. We decide how many terms we take into account by neglecting all the terms that are smaller than a fixed *tolerance* (see Section 3.2 for details).

```
#NOTATIONS:
#pr=process
#lambda=intensity of the centre process
#alpha=sibuya exponent
#k=number of test points per unity
#a=width of the window


vp.est.un=function(lambda, alpha, pr, a=1000, plot=TRUE, grid=
      TRUE, k=1){  #estimates vp for uniformly distributed
       values of p and plots the estimation
          #plot=TRUE will give no plot
          #plot=FALSE will give the plot of the estimated v.p.
                 (black) compared with the exact ones (red)

          rmax=2 #maximum radius we are to consider

          #test points
          if(grid==TRUE){
                    pt=seq(-pr$a/2+1+rmax, pr$a/2-1-rmax, by=1/k)
                    }
          else{
                    pt=cbind(runif(k*(pr$a-2-2*rmax),-pr$a/2+1+
                          rmax, pr$a/2-1-rmax))
          }
          n=length(pt)
```

```
        p=(1:29)/30 #range of retention probabilities

        d1=matrix(rep(pr$xm[,1],times=n),ncol=n)
        d2=matrix(rep(pt,times=sum(pr$ncl)),nrow=sum(pr$ncl),byrow=TRUE)
        d=apply(abs(d1-d2),2,sort) #calculates the distances between test points and points in the process and order them

        est.dr=function(pp){return(estimation(pp,d,r))}
        vpr=sapply(p,est.dr) #estimated void probabilities

        q=function(x){ #exact void probabilities
        v=exp(-4*x[1]*p^x[2]/(x[2]+1))
        return(sum((v-vpr)^2))
        }

        if(plot==TRUE){
        #plot of the estimated vp (black) and of the exact vp (red)

        plot(p, exp(-4*lambda*p^alpha/(alpha+1)),ty="p",xlab="retention probability",ylab="void probabilities r=1",asp=1,col="red")
        points(p, vpr, col="black",ty="p")
        }

        par=return(optim(c(1,1),q,lower=c(0,0),upper=c(1,1))$par)

        return(list(par,vpr,real=exp(-4*lambda*p^alpha/(alpha+1))))
}

estimation=function(p,d,r=1,t=0.00001){
        #p=retention probabilities
        #d=matrix of distances between test points and the configuration
        #t=tolerance
        v=numeric()
```

```
    z=numeric()
    j=ceiling(log(t/p)/log(1-p)) #last index to be considered in the sum
    for(i in 1:ncol(d)){
        v[i]=nrow(d)-length(d[,i][d[,i]>r])+1 #position of the first point in the configuration which is further than r
        k=v[i]:j
        z[i]=sum(((1-p)^(k-1))*p) #computation of the sum for each test points
        }
    stat=(1/ncol(d))*sum(z) #value of the estimator
}
```

Conclusions

In the present work we suggested efficient methods to estimate the three parameters that characterize our model. We also obtained some good results while testing them on synthetic samples. Future development for our work are twofold.

Firstly it would be interesting to complete the implementation side with the estimation of the Sibuya parameter measure μ in the cases when we cannot assume that it belongs to any parametric class. More particularly a new code could be implemented to extract the sample measure from a big cluster and this extracted sample measure could be used as an approximation of our measure when estimating λ and α.

Secondly we plan to try to use our model to fit data regarding calling activity during events involving big groups of people (see more details and a picture in the introduction). In fact the present thesis confirmed that our work have all the potentials to be well employed in this task.

Finally we are sure that trying our hands on real data will arise the more interesting tasks apart from those that we can foresee and will give new inspiration both for the theoretical and the computational side.

Bibliography

[1] R. S. Bivand, E. J. Pebesma, and V. Gomez-Rubio. *Applied Spatial Data Analysis with R.* Springer, 2008.

[2] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods.* Springer, 2002.

[3] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure.* Springer, 2002.

[4] Y. Davydov, I. Molchanov, and S. Zuyev. Stability for random measures, point processes and discrete semigroups. *Bernoulli*, 17, 2011.

[5] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Society*, 97, 2002.

[6] C. Fraley and A. E. Raftery. Mclust version 3 for r: Normal mixture modeling and model-based clustering. *Technical Report No. 504*, 2006.

[7] D. Stoyan, W. S. Kendall, and J. Mecke. *Stochastic Geometry and its Applications.* Wiley, 1995.