# Study the galaxy distribution characterisation via Bayesian statistical learning of spatial marked point processes

N Gillot, Radu S. Stoica, Didier Gemmerlé

# Study the galaxy distribution characterisation via Bayesian statistical learning of spatial marked point processes.

N. Gillot[1], R. S. Stoica[1], and D. Gemmerlé[2]

[1] *Université de Lorraine, CNRS, Inria, IECL, F-54500 Nancy France*
[2] *Université de Lorraine, CNRS, IECL, F-54500 Nancy France*

September 2023

**Abstract**

Marked point process and Bayesian inference are powerful tools for analysing spatial data. Here the work done by Hurtado Gil et al. (2021) is analysed and a new in-homogeneous with superposed interaction is proposed. The results indicate a correct fit of the model and allow the study of the significance of the parameter at the corresponding pre-fixed interaction ranges. To this work in progress, immediate conclusions and perspectives are outlined.

## Introduction

Galaxies are not uniformly distributed in the observable Universe. Their positions induce structures such as filaments, void zones or even clusters of galaxies. The complexity of these structures and the amount of data available on the subject led to the idea of a probabilistic approach to explain the characteristics of these structures, based on point process models ((Møller & Waagepetersen, 2004) , (van Lieshout, 2019)). An important part of this probabilistic framework is to use algorithms able to estimate the parameters of the models proposed to fit the observed data such as Approximate Bayesian Computation (ABC) algorithms. ( Stoica et al. (2017)).

The paper presents and tries to extend the modelling, simulation and inference approach for point process models given by ((Hurtado Gil et al., 2021)) while introducing a model based on distance to galactic filaments.

## 1 Materials and methods

Let's assume that a pattern of points $\mathbf{x} = \{x_1, ..., x_n\}$ is observed in a compact window $W \subset \mathbb{R}^d$. From now on, we assume that the data we observe have the following properties :

• The Universe can be seen as the representation of a stochastic process where galaxies are randomly located points in space.

• Two such points cannot share the same position: for a given point $\xi \in W$, no other point has the same coordinates in $W$.

This means that we can see the galaxies distribution in our Universe as a realisation of a point process and that we suppose existing an underlying probability density that characterise this point process. The probability density we'll consider can be written in the exponential form :

$$f(\mathbf{x}|\theta) = \frac{\exp\left(\langle t(\mathbf{x}), \theta\rangle\right)}{c(\theta)} \tag{1}$$

with $\mathbf{x} = \{x_1, ..., x_n\}$ the point pattern, $t : \Omega \to \mathbb{R}^d$ the vector of sufficient statistics, $\theta \in \Theta$ the model parameters and $c(\theta)$ the partition function.

In this section, we first show some examples of point processes and how we can easily create new models characterised by some unnormalised densities. We then focus on the simulation of these models with the Metropolis-Hastings algorithm. Finally, we'll discuss two methods for parameters estimation and the asymptotic results.

## 1.1 Some examples of points processes

### 1.1.1 Poisson point process

This point process exhibit no interactions among points. It's used in practice as a reference point process to build probability densities with respect to the standard (homogeneous with unit intensity) Poisson point process ( Møller and Waagepetersen (2004), Stoica (2014)). For an intensity function $\rho : W \to [0, +\infty[$, the Poisson point process density can be written as :

$$f(\mathbf{x}|\rho) \propto \exp\left(\sum_{i=1}^{n(\mathbf{x})} \log(\rho(x_i))\right) \tag{2}$$

where $n(\mathbf{x})$ is the number of points in $\mathbf{x}$. If $\rho$ is a constant, the point process will be called homogeneous.

### 1.1.2 Strauss point process

The Strauss point process is a model with interaction that penalise the probability of having two points at a distance closer to a fixed radius, $r$. With respect to the standard Poisson point process, its probability density is given by

$$f(\mathbf{x}|\rho, \gamma_s) \propto \exp(n(\mathbf{x}) \log(\rho) + s_r(\mathbf{x}) \log(\gamma_s)) \tag{3}$$

where $s_r(\mathbf{x})$ represent the number of pairs of points closer than the distance $r$, $\gamma_s \in ]0, 1]$ the model parameter. In this model, $n(\mathbf{x})$ and $s_r(\mathbf{x})$ are the sufficient statistics. Note that if $\gamma_s = 1$, the model boils down to the Poisson process of intensity $\rho$.

### 1.1.3 Area Interaction process

The Area Interaction point process is a model with interaction that takes into account the area of balls of a fixed radius $R$ around the points. This is also a good example to show how to create new probability densities with respect to the Poisson point process of intensity 1 by introducing a new sufficient statistic of interest. In the homogeneous case, its density is given by

$$f(\mathbf{x}|\rho, \gamma_a) \propto \exp(n(\mathbf{x}) \log(\rho) + a_R(\mathbf{x}) \log(\gamma_a)) \tag{4}$$

where $a_R(\mathbf{x}) = -|\cup_{\xi \in \mathbf{x}} b(\xi, R)|$ represent the $d$-volume of the union of balls of radius $R$ attached to the points, $\gamma_a \geq 0$ is the model parameter. In this model, $n(\mathbf{x})$ and $a_R(\mathbf{x})$ are the sufficient statistic. Once more, if $\gamma_a = 1$, the model becomes the Poisson process of intensity $\rho$.

### 1.1.4 Superposition of two models : Area Interaction and Strauss point process

Another way to create new probability densities is to combine two existing point processes. Here, we combine the Area Interaction and the Strauss point process : this makes a model that takes into account the area of the balls, the amount of pairs of points and the number of points. Its density can be written as

$$f(\mathbf{x}|\rho, \gamma_s, \gamma_a) \propto \exp(n(\mathbf{x}) \log(\rho) + s_r(\mathbf{x}) \log(\gamma_s) + a_R(\mathbf{x}) \log(\gamma_a)) \tag{5}$$

with the same parameters as in the previous examples.

### 1.2 Simulation : the Metropolis-Hastings algorithm

The simulation algorithm has the following pseudo-code :

1) Set $\mathbf{x^{(0)}}$, the configuration of points at the beginning, $N \in \mathbb{N}$ the number of steps and $\theta$ the model's parameter.

2) For $k = 1, ..., N$, generate $\mathbf{x^{(k)}}$ distributed with density $f(\mathbf{x}^{k-1}|\theta)$

3) Set $\mathbf{x} = \mathbf{x}^{(k)}$

For the sampling of $p(\mathbf{x}|\theta)$, the following Metropolis-Hastings procedure is described bellow :

1) Set $p_b, p_d$ with $p_b + p_d = 1$

2) With probability $p_b$ choose to add a point (birth) and with probability $p_d$ choose to delete a point (death).

- **birth**
  a) generate a random point $\xi$ on $W$ and set $\mathbf{x}' = \mathbf{x} \cup \xi$
  b) compute $r_b = \min\{1, \frac{p_d}{p_b} \frac{f(\mathbf{x} \cup \xi|\theta)}{f(\mathbf{x}|\theta)} \frac{|W|}{n(\mathbf{x})+1}\}$
- **death**
  a) choose a random point $\xi$ of $\mathbf{x}$ and set $\mathbf{x}' = \mathbf{x} \setminus \xi$
  b) compute $r_d = \min\{1, \frac{p_b}{p_d} \frac{f(\mathbf{x} \setminus \xi|\theta)}{f(\mathbf{x}|\theta)} \frac{n(\mathbf{x})}{|W|}\}$

3) Accept the new configuration $\mathbf{x}'$ with probability $r_b$ or $r_d$ (depending on the choice of birth or death). Otherwise, remain in the same state $\mathbf{x}$.

This algorithm generate a Markov Chain that is $\Phi-$irreducible, Harris recurrent and geometric ergodic. Thus, the algorithm converges toward the distribution of interest given by the density $f(\mathbf{x}|\theta)$ ( Møller and Waagepetersen (2004) ; Stoica (2014) ; van Lieshout (2019)).

### 1.3 Statistical inference : the ABC Shadow algorithm

We now turn to the parameters estimation. To do so, we'll use an Approximate Bayesian Computing (ABC) algorithm. In the Bayesian framework, this will consists in sampling the following posterior law :

$$f(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)p(\theta) \tag{6}$$

where $p(\cdot)$ is the prior distribution of the parameters.

Performing such inference from the posterior distribution is a challenging problem in mathematics. Indeed, the partition function $c(\theta)$ isn't available in analytic closed form for the model class we're considering in this article. To bypass this problem, we'll use the ABC Shadow algorithm (Stoica et al. (2017)) : the outputs are approximate samples from the posterior distribution of interest.

The algorithm's pseudo-code is the following :

1) Set $\delta$ a perturbation parameter, $\theta_0$ an initial condition and $m$ number of iterations. Assume that a pattern $\mathbf{x}$ is observed.

2) With the Metropolis Hastings algorithm, generate $\mathbf{y}$ according to $f(\mathbf{y}|\theta_\mathbf{0})$

3) For $k = 1$ to $m$ :

a) Generate a new parameter $\psi$ according to the density $U_\delta(\theta_{k-1} \to \psi)$ defined by $U_\delta(\theta \to \psi) = \frac{1}{|b(\theta,\delta/2)|}\mathbf{1}_{b(\theta,\delta/2}\{\psi\}$.

b) The new state $\theta_k = \psi$ is accepted with probability $\alpha_s(\theta_{k-1} \to \psi) = \min\{1, \frac{f(\mathbf{x}|\theta_k)p(\theta_k)}{f(\mathbf{x}|\theta_{k-1})p(\theta_{k-1})} \times \frac{f(\mathbf{y}|\theta_{k-1})}{f(\mathbf{y}|\theta_k)}\}$

4) Return $\theta_m$.

5) If more samples are needed, go to step 1 and set $\theta_0 = \theta_m$

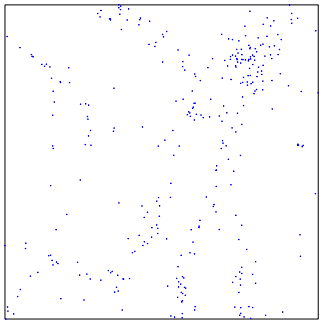
### 1.4 Asymptotic errors

The asymptotic normality of the maximum likelihood estimation allows to compute two types of estimation error. The first one is an approximation of the difference between the unknown exact maximum likelihood estimator (MLE) and the true parameter value: $\hat{\theta} - \theta_0$. The other one is the difference between the Monte Carlo Maximum Likelihood Estimation and the unknown exact MLE: $\hat{\theta_n} - \hat{\theta}$. We can compute an estimation of these errors in order to control the parameter estimation as done as in Geyer (1994, 1999); van Lieshout and Stoica (2003).

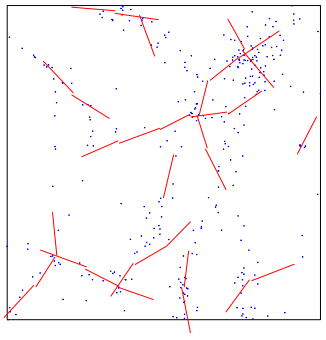
## 2 Application

### 2.1 Data presentation

The data set presented here is the cosmological simulation that was used to set up the first filaments pattern detector based on marked point process Stoica et al. (2005). Here a region from this data field and the corresponding detected filaments are selected. The aim is to fit a model to the galaxy distribution conditionally to the observed point field and the attached filaments. The selected sample and the filaments are shown in Figure 1b



(a) *Galaxies pattern (cosmological simulation)*



(b) *Corresponding detected filaments*

### 2.2 Studied model

Here, the ABC Shadow algorithm was used to fit a superposition of models like (5) with the following components

- Poisson component : in-homogeneity that takes into account $d(\xi, F)$, the shortest distance from a point $\xi \in W$ to the the given filament network. This distance is presented in the Figure 2 below. The sufficient statistic attached to this component is: $\sum_{i=1}^{n(\mathbf{x})} \mathbf{1}_{d(\xi_i,F)\leq 0.05}(\xi) \times \frac{1}{1+d(\xi_i,F)}$.

- Strauss component : the same as the interaction part in (3)
- Area-Interaction component : the same as the interaction part in (4)

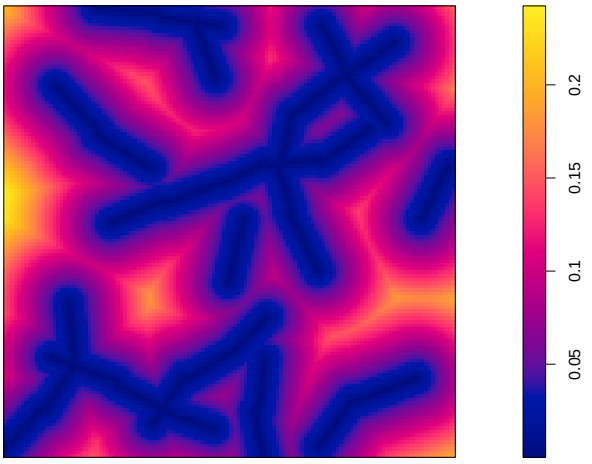



Figure 2: *The shortest distance between any point in the domain to the given filament network.*

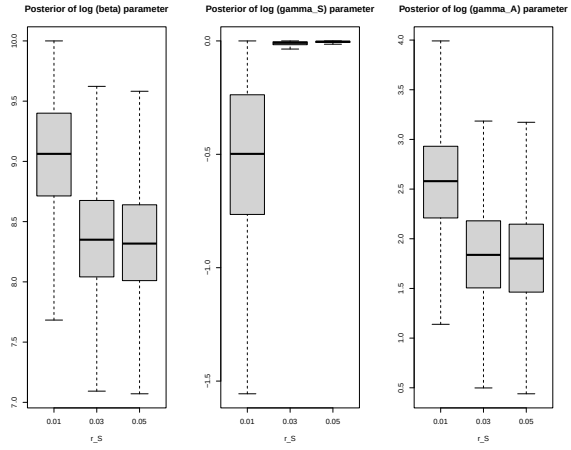
The observed statistics are given in the table below: this will lead to 6 parameters estimation for different fixed radius for both Strauss ($r_S$) and Area Interaction ($r_A$) components.

| $r_S, r_A$ | 0.01 | 0.03 | 0.05 |
|---|---|---|---|
| $n(\mathbf{x})$ | 334 | 334 | 334 |
| $s_{r_S}(\mathbf{x})$ | 71 | 539 | 1268 |
| $-a_{r_A}(\mathbf{x})$ | 272 | 143 | 83 |

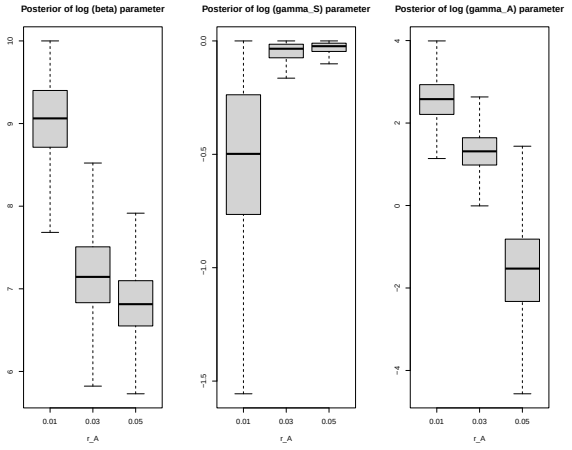
### 2.3 Results

For each radius tuple $(r_S, r_A)$ among $(0.01, 0.01)$ ; $(0.01, 0.03)$ ; $(0.01, 0.05)$ ; $(0.03, 0.01)$ ; $(0.05, 0.01)$, the ABC Shadow algorithm was initialised with the observed pattern's sufficient statistics. The prior density $p(\theta)$ is the uniform distribution on the interval $[0, 10] \times [-10, 0] \times [-10, 10]$. At every step, the auxiliary variable was sampled with 250 iterations of the Metropolis-Hastings algorithm. $\delta$ was set to $(0.01, 0.01, 0.01)$, $m$ to 100 and $\theta_0$ was set randomly inside the prior density interval. This procedure was run $10^4$ times, giving us a sample of size $10^4$ of the estimated parameters.

Figure 3a shows the box-plot of the posterior distribution for the model's parameters for different values of $r_S$ and $r_A = 0.01$, figure 3b shows the box-plot of the posterior distribution for the model's parameters for different values of $r_A S$ and $r_S = 0.01$.




(a) *Box-plot for parameter estimation with respect to* $r_A = 0.01$ *and different* $r_S$




(b) *Box-plot for parameter estimation with respect to* $r_S = 0.01$ *and different* $r_A$

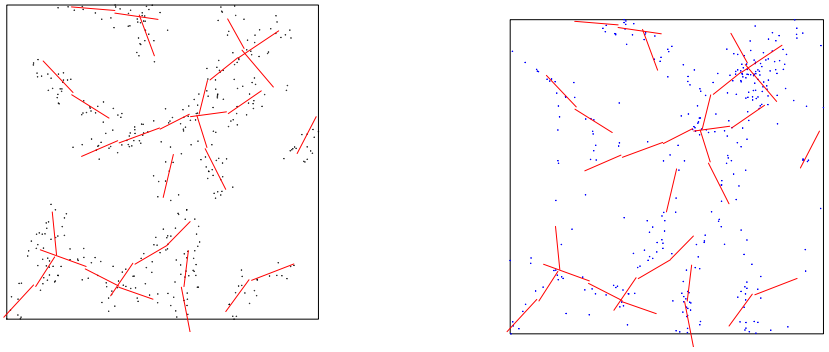
*Apart for* $(r_S, r_A) = (0.01, 0.01)$*, the parameter estimation for the Strauss component is very close to zero. This leads to consider that, for the considered ranges, there is no repulsion between the galaxies. For the estimation of the Area Interaction parameter, however, the values are far from* 0*, this may be interpreted as a clustering tendency between galaxies.*

In the table below, we summarise the parameters inference for the different radius with their standard error:

| Radius | Estimates of $\log(\rho)$, $\log(\gamma_S)$ and $\log(\gamma_A)$ | | |
|---|---|---|---|
| $(r_S, r_A)$ | $\log(\rho)$ | $\log(\gamma_S)$ | $\log(\gamma_A)$ |
| $(0.01, 0.01)$ | $9.04 \pm 0.24$ | $-0.52 \pm 0.16$ | $2.55 \pm 0.28$ |
| $(0.01, 0.03)$ | $7.19 \pm 0.08$ | $-0.05 \pm 0.12$ | $1.31 \pm 0.32$ |
| $(0.01, 0.05)$ | $6.83 \pm 0.09$ | $-0.03 \pm 0.17$ | $-1.57 \pm 0.93$ |
| $(0.03, 0.01)$ | $8.36 \pm 0.20$ | $-0.02 \pm 0.03$ | $1.84 \pm 0.21$ |
| $(0.05, 0.01)$ | $8.33 \pm 0.21$ | $-0.009 \pm 0.02$ | $1.8 \pm 0.20$ |

*We can see that the standard errors are rather smalls except in the* $(0.01, 0.05)$ *case for* $\log(\gamma_A)$ *which can be explained by the rather high value of the statistic.*

Figure 4a shows the simulated pattern with the corresponding estimation for $(r_S, r_A) = (0.01, 0.03)$. Figure 5 shows the histogram of the posterior approximation used for parameter estimation for this specific case. The posterior maximum value are indicated in red.



(a) *Simulated galaxies distribution using the estimated parameters*

(b) *Observed galaxies distribution*

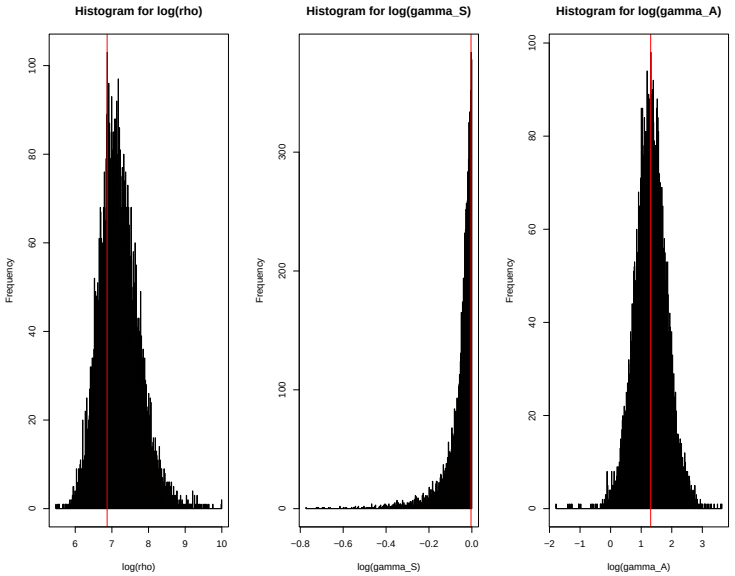



Figure 5: *Histogram of the posterior approximation used for parameter estimation with* $(r_S, r_A) = (0.01, 0.03)$.

## Conclusions

The presented inference framework allows to study and assess significance of the chosen modelling components to be fitted to the data. This is work in progress. As perspectives we mention: chose a random interaction radius for each model component, model validation and 3d data analysis.

## References


Geyer CJ. (1994). On the convergence of monte carlo maximum likelihood calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*(1), 261–274.

Geyer CJ. (1999). *Chapter 1 Likelihood Inference for Spatial Point Processes.*

Hurtado Gil L, Stoica RS, Martínez VJ, & Arnalte Mur P. (2021). Morphostatistical characterization of the spatial galaxy distribution through Gibbs point processes. *Monthly Notices of the Royal Astronomical Society*, *507*(2), 1710-1722.

van Lieshout MNM. (2019). *Theory of Spatial Statistics : A concise Introduction.* Chapman & Hall.

van Lieshout MNM, & Stoica RS. (2003). The Candy model : properties and inference. *Statistica Neerlandica*, *57*(2), 177-206.

Møller J, & Waagepetersen RP. (2004). *Statistical Inference and Simulation for Spatial Point Processes.* Chapman & Hall/CRC.

Stoica RS. (2014). *Modélisation probabiliste et inférence statistique pour l'analyse des données spatialisées.* Habilitation à Diriger des Recherches thesis - Université de Lille.

Stoica RS, Martinez VJ, Mateu J, & Saar E. (2005). Detection of cosmic filaments using the Candy model. *Astronomy and Astrophysics - A&A*, *434*(2), 423-432. doi: 10.1051/0004-6361:20042409

Stoica RS, Philippe A, Gregori P, & Mateu J. (2017). ABC Shadow algorithm : a tool for statistical analysis of spatial patterns. *Statistics and Computing*, *27(5)*, 1225-1238.