# A Guide to NScluster: R Package for Maximum Palm Likelihood Estimation for Cluster Point Process Models using OpenMP

Ushio Tanaka
Osaka Prefecture University

Masami Saga
Indigo Corporation

Junji Nakano
The Institute of Statistical Mathematics

December 8, 2019

## 1 Preliminaries

A *point process* is a stochastic model governing the location of events in a given set. We consider the point process in a subset of Euclidean space. A *point pattern* is considered a realization of the point process. To analyze the point pattern, we first plot it as observed in the subset, which is considered an *observation window* denoted $W$. Following the preceding study, for simplicity, we restrict our discussion to $W$ of a two-dimensional Euclidean space $\mathbb{R}^2$ to be standardized, i.e., a unit square ($W = [0,1] \times [0,1]$). Thus, throughout NScluster, we employ a unit square as the observation window. If the window is a rectangular domain or is irregularly shaped, we select the largest possible square from the window, and consider it as the unit. We assume that $W$ satisfies a periodic boundary condition to consider it as a torus.

Throughout NScluster, we refer readers to Tanaka et al. [1, 2] for details.

## 2 Overview of models

We assume point processes on $W$ satisfy conditions of local finiteness, simplicity, uniformity and isotropy. Note that by virtue of uniformity, point processes are homogeneous, i.e., they are of constant intensity.

First, we generate a homogeneous Poisson point process with intensity $\mu$. The generated points are referred to as parent points. Each parent point generates a random number $M$ of descendant points, which are realized independently and identically. Let $\nu$ be the expectation of $M$. The descendent points are distributed isotropically around each parent point, and the distances between each parent point and its descendent points are distributed independently and identically according to a probability density function (PDF) relative to the distance from a parent point to its descendent point. We call the PDF a dispersal kernel and denote it by $q_\tau$, where $\tau$ indicates the parameter set of the dispersal kernel. The *Neyman-Scott cluster point process* is a union of all descendant points,

with the exception of all parent points. In other words, the cluster process is unobservable for each cluster center. The Neyman-Scott cluster point process is also homogeneous, and its intensity $\lambda$ equals $\mu\nu$.

We describe five cluster point process models, i.e., the Thomas and Inverse-power type models, and the extended Thomas models of type A, B, and C.

## 2.1 Neyman-Scott cluster point process model

### 2.1.1 Thomas model

The *Thomas model* is the most utilized Neyman-Scott cluster point process model. In this model, descendant points are scattered according to bivariate Gaussian distribution with zero mean and covariance matrix $\sigma^2 I$, $\sigma > 0$, where $I$ is a $2 \times 2$ identity matrix. The corresponding dispersal kernel with $\tau = \sigma$ is given by

$$q_\sigma(r) := \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right), \quad r \geq 0.$$

In previous studies that analyzed clustering point pattern data, the Thomas model has been representatively situated to be fitted to such data because one can explicitly derive classical summary statistics, e.g., Ripley's $K$-function of the Thomas model, which is closely related to the Palm intensity (Section 3.2.1).

### 2.1.2 Inverse-power type model

The *Inverse-power type model* originated from the frequency of aftershocks per unit time interval (one day, one month, etc.), which has been referred to as the "modified Omori formula". The corresponding dispersal kernel with $\tau = (p, c)$ is given by

$$q_{(p,c)}(r) := \frac{c^{p-1}(p-1)}{(r+c)^p}, \quad r \geq 0,$$

where $p > 1$ and $c > 0$ imply the decay order and scaling with respect to the distance between each parent point and its descendant points, respectively.

### 2.1.3 Type A model

The *extended Thomas model of type A* (*Type A model* for short) is a Neyman-Scott cluster point process model where the dispersal kernel is mixed by that of the two Thomas models with variable cluster sizes as follows:

$$q_{(a,\sigma_1,\sigma_2)}(r) := a q_{\sigma_1}(r) + (1-a) q_{\sigma_2}(r), \quad r \geq 0, \tag{1}$$

where $a$ is a mixture ratio parameter with $0 < a < 1$. From Equation (1), it can be inferred that the Type A model is suitable for densely and vaguely clustering point pattern data to be fitted by mixing the Thomas model with the mixture ratio $a$.

## 2.2 Superposed Neyman-Scott cluster point process model

We extend the Neyman-Scott cluster point processes to superposed ones. The superposition is one of extension manners.

Here, we focus on the superposed Thomas model. The parameters to be estimated are given by those of two Thomas models: $(\mu_i, \nu_i, \sigma_i)$, where $i = 1, 2$. Note that the intensity $\lambda$ of superposed uniform point processes with intensity $\lambda_i$ $(= \mu_i \nu_i)$, $i = 1, 2$, is given by

$$\lambda = \lambda_1 + \lambda_2.$$

### 2.2.1 Type B and C models

We handle two types of the superposed Thomas model, which are referred to as the *etended Thomas model of type B* (*Type B model* for short) if $\nu_1 = \nu_2$ and the *extended Thomas model of type C* (*Type C model* for short) if $\nu_1 \neq \nu_2$.

# 3 Overview of functions

The package NScluster comprises four tasks, i.e., simulation, MPLE, confidence interval estimation, and non-parametric and parametric Palm intensity comparison.

## 3.1 Simulation

The first and most intuitive step to understand the model characteristics is to observe the data generated by the model. This can be realized using the `sim.cppm` function.

## 3.2 MPLE

### 3.2.1 Palm intensity

We begin with a brief overview of the Palm intensity of the point processes. Translating each point of the given point process into the origin $\boldsymbol{o} \in \mathbb{R}^2$, we obtain a superposed point process at $\boldsymbol{o}$. We call it the *difference process*. The difference process is symmetric with respect to $\boldsymbol{o}$. The Palm intensity focuses on the difference process induced from pairwise coordinates of the original process rather than the original given point process.

Let us define the Palm intensity. We denote by $N$ a counting measure, i.e., the total mass of random geometrical objects such as the number of points, lengths of fibers, areas of surfaces, and volume of grains within Borel sets. The *Palm intensity* $\lambda_{\boldsymbol{o}}$ is defined as follows:

$$\lambda_{\boldsymbol{o}}(\boldsymbol{x}) := \frac{\Pr(\{\, N(\boldsymbol{dx}) \geq 1 \mid N(\{\boldsymbol{o}\}) = 1 \,\})}{\mathrm{Vol}(\boldsymbol{dx})}, \tag{2}$$

where $\boldsymbol{dx}$ represents an infinitesimal set containing an arbitrary given point $\boldsymbol{x} \in W$. Here, we examine Equation (2). $\lambda_{\boldsymbol{o}}$ implies the occurrence rate at an arbitrary given point $\boldsymbol{x}$ provided that a point is at $\boldsymbol{o}$. Let $r$ be the distance from $\boldsymbol{o}$ to $\boldsymbol{x}$. We see that $\lambda_{\boldsymbol{o}}$ depends only on $r$. Thus, we obtain its polar coordinate representation with respect to distance $r$ as follows:

$$\lambda_{\boldsymbol{o}}(\boldsymbol{x}) = \lambda_{\boldsymbol{o}}(r, \theta) = \lambda_{\boldsymbol{o}}(r), \quad r \geq 0, \quad 0 \leq \theta < 2\pi.$$

Generally, the Palm intensity of cluster point processes cannot be derived analytically, say the aforementioned Inverse-power type and the Type A models.

Here, we further assume the point processes to be orderly, i.e., $\Pr(\{ N(\boldsymbol{dx}) \geq 2\})$ is of a smaller order of magnitude than $\mathrm{Vol}(\boldsymbol{dx})$. The orderliness allows us to represent the Palm intensity in terms of Ripley's $K$-function, which is defined as the average number of other points that have appeared within the distance from the typical point.

### 3.2.2 Palm likelihood function

The maximum Palm likelihood estimation procedure is based on the assumption that the difference process is well approximated by an isotropic and inhomogeneous Poisson point process with intensity function $N(W)\,\lambda_{\boldsymbol{o}}(r)$, which is centered at $\boldsymbol{o}$.

We are positioned to state the log-*Palm likelihood function*. Let $\boldsymbol{\theta}$ denote the parameter set of the cluster point process models. The log-Palm likelihood function, denoted $\ln L$ based on the Palm intensity $\lambda_{\boldsymbol{o}}$ (including $\boldsymbol{\theta}$) is given as follows:

$$\ln L(\boldsymbol{\theta}) = \sum_{i,j;i<j,0<r_{ij}\leq 1/2} \ln\left(N(W)\,\lambda_{\boldsymbol{o}}(r_{ij})\right) - 2\pi\, N(W) \int_0^{1/2} \lambda_{\boldsymbol{o}}(r)\, r\, dr, \tag{3}$$

where the summation is taken over each pair $(i, j)$ with $i < j$ such that the distance $r_{ij}$ between distinct points $x_i$ and $x_j$ of the cluster point processes satisfies $0 < r_{ij} \leq 1/2$. Note that in Equation (3) "$i < j$" and "$1/2$" are due to the symmetry of difference processes and the periodic boundary condition for $W = [0, 1] \times [0, 1]$, respectively.

The *maximum Palm likelihood estimate*s (*MPLE*s for short) are those that maximize Equation (3). Note that maximizing $\ln L(\boldsymbol{\theta})$ in Equation (3) to obtain MPLEs, $N(W)$ assigning the non-parametric part of Equation (3) is removable.

The `mple.cppm` function improves the given initial parameters using the simplex method to maximize $\ln L(\boldsymbol{\theta})$ in Equation (3).

## 3.3 Confidence interval of parameter estimates

We develop a confidence interval of parameters using bootstrap method. When we estimate one model, we generate simulated data several times for the estimated model, then, we estimate the parameters and repeatedly. The empirical distribution of given parameters can be used to decide the interval estimation of the parameter.

## 3.4 Display of normalized Palm intensity

To determine the adequacy of MPLEs, NScluster provides users with a non-parametric estimation of the Palm intensity. NScluster can depict the Palm intensity of the five cluster point process models using the `palm.cppm` function.

# References

[1] Tanaka U, Ogata Y, Katsura K (2008). "Simulation and Estimation of the Neyman-Scott Type Spatial Cluster Models." *Computer Science Monographs*, **34**, 1–44: URL https://www.ism.ac.jp/editsec/csm/

[2] Tanaka U, Ogata Y, Stoyan D (2008). "Parameter estimation and model selection for Neyman-Scott point processes." *Biometrical Journal*, **50**, 43–57.