STAT 8025

Lecture 13: Point Process

Dr. Emily Lei Kang

Division of Statistics & Data Science
Department of Mathematical Sciences
University of Cincinnati

Introduction

- Point processes: Locations are random
- Example: Tree locations
  - Ambrosia dumosa is a drought deciduous shrub with 20-60cm in height which is abundant on well drained soils below one thousand meter elevation. The data were collected within a hectare ($100 \times 100\text{m}^2$) area in the Colorado Desert in 1984. The data contains
    - locations 4358 Ambrosia dumosa trees
    - the height of the plant canopy
    - the length of the major axis of the plant canopy
    - the length of the minor axis of the plant canopy
    - the volume of the plant canopy.
- Examples: occurrence of earthquake, crime, forest fire
- We observe $\text{s}_1, \ldots, \text{s}_n$, and model n and $\text{s}_1, \ldots, \text{s}_n$ as random.

Types of Point Patterns

- Completely random: $\mathrm{s}_i | n \overset{iid}{\sim} Uniform(\mathcal{S})$
- Clustered: s_i's form clusters
- Inhomogeneous: Events more likely in some regions than others (sometimes it can be hard to distinguish this from clustered pattern)
- Regular: s_i's repulse each other

Ripley's K Function

- Ripley's K function is analogous to the variogram in geostatistics
- Let $d_{ij} = |\mathrm{s}_i - \mathrm{s}_j|$ be the distance between s_i and s_j
- Ripley's empirical K function is

$$\hat{K}(t) = |\mathcal{S}| \frac{1}{n^2} \sum_{i \neq j} I(d_{ij} \leq t)$$

  - estimating proportion that a pair of points are within t of each other
  - this is the empirical function; as variogram, different modles have different true K-functions and plotting $\hat{K}$ will help choose a true model

Point Processes

- Let $\mathcal{S} \subset \mathbb{R}^d$ be the domain. Let $N(A)$ be the number of points in A for *any* $A \subset \mathcal{S}$. Then, the distribution of N is given by

$$P_k[N(A_1) = n_1, \cdots, N(A_k) = n_k]$$

  for any $A_1, \ldots, A_k \subset \mathcal{S}$ and $k \in \mathcal{N}^+$.

- The k-th order intensity function of N (if exists) is

$$\lambda_k(\mathrm{s}_1, \ldots, \mathrm{s}_k) = \lim_{|d\mathrm{s}_i| \to 0, i=1,\ldots,k} \left\{ \frac{E[N(d\mathrm{s}_1) \cdots N(d\mathrm{s}_k)]}{|d\mathrm{s}_1| \cdots |d\mathrm{s}_k|} \right\}$$

  where s_i are distinct points in $\mathcal{S}$.

- Both P_k and λ_k can be used for the distribution of N, but people focus on λ_k more.

- The pair correlation function:

$$g(s_1, s_2) = \frac{\lambda_2(s_1, s_2)}{\lambda(s_1)\lambda(s_2)}$$

- The mean function $\mu(A) = \int_A \lambda(s)ds$
- The covariance structure N is

$$\begin{aligned} & Cov[N(A_1), N(A_2)] \\ = \ & \int_{A_1} \int_{A_2} [\lambda_2(s_1, s_2) - \lambda(s_1)\lambda(s_2)] ds_2 ds_1 + \int_{A_1 \cup A_2} \lambda(s) ds \end{aligned}$$

K-function

- Suppose N is stationary. Let λ be the first-order intensity function.
- The K-function is defined as

$$K(t) = \frac{1}{\lambda}E[\text{number of extra events within distance of } t$$

$$\text{of a randomly chosen event}]$$

- The L-function is

$$L(t) = \sqrt{\frac{K(t)}{\pi}}$$

- $K(t)$ is more often used.

Stationarity

- A spatial point process N is said to be strongly stationary if for any $A_1, \ldots, A_k$, the joint distribution of

$$N(A_1 + \mathsf{s}), \ldots, N(A_k + \mathsf{s})$$

does not depend on s, where

$$A_i + \mathsf{s} = \left\{\mathsf{s}' + \mathsf{s} : \mathsf{s}' \in A_i\right\}$$

- If $\lambda(\mathsf{s})$ is constant and

$$\lambda_2(\mathsf{s}_1, \mathsf{s}_2) = \lambda_2(\mathsf{s}_1 - \mathsf{s}_2)$$

N is called second-order stationary. In addition, if

$$\lambda_2(\mathsf{s}_1, \mathsf{s}_2) = \lambda_2(|\mathsf{s}_1 - \mathsf{s}_2|)$$

N is called isotropic.

Poisson Point Process

A poisson point process is derived if $N(A_1), \ldots, N(A_k)$ are independent Poisson random variable with mean $\mu(|A_1|), \ldots, \mu(|A_k|)$ if $A_1, \ldots, A_k$ are disjoint subsets.

- The simplest case with $\lambda(\mathsf{s}) = \lambda$ and this implies: $n \sim Poisson(\lambda|\mathcal{S}|)$ and $\mathsf{s}_1, \ldots, \mathsf{s}_n | n \sim Uniform(\mathcal{S})$
  - Called the homogeneous Poisson process, $n \sim Poisson(\lambda|\mathcal{S}|)$ and is stationary and sampled completely at random.
  - The sufficient statistics for the single parameter is n and $\hat{\lambda} = \frac{n}{|\mathcal{S}|}$ is the MLE

- Inhomogeneous Poisson point process: $\lambda(\mathsf{s})$ spatially-varying
  - $\mu(A) = \int_A \lambda(\mathsf{s})d\mathsf{s}$ so the total number of event $n \sim Poisson(\mu(\mathcal{S}))$
  - Given n, $\mathsf{s}_1, \ldots, \mathsf{s}_n \overset{iid}{\sim} f(\mathsf{s})$ with

$$f(\mathsf{s}) = \frac{\lambda(\mathsf{s})}{\int_{\mathcal{S}} \lambda(t)dt}$$

UNIVERSITY OF Cincinnati

Simulate an inhomogeneous Poisson point process?

- Let f be a pdf on s.
- We generate $n \sim Poisson(B)$
- Generate n observations idependently from f, $s_1, \ldots, s_n$
- Then $s_1, \ldots, s_n$ is a sample from Point point process with intensity

$$\lambda(s) = Bf(s)$$

Thinned Poisson Point Process

- We have $\lambda(\mathsf{s}) = Z(\mathsf{s})\lambda$ where $Z(\mathsf{s}) \in [0, 1]$
- Simulating a thinned Poisson process
  1. Generate $n \sim \mathit{Poisson}(\lambda|\mathcal{S}|)$
  2. Generate $\mathsf{s}_1, \ldots, \mathsf{s}_n \sim \mathit{Uniform}(\mathcal{S})$
  3. Keep s_i wit probability $Z(\mathsf{s}_i)$

Modeling Inhomogeneous Poisson Process

- We can model the log intensity function:

$$log(\lambda(\mathrm{s})) = \beta_0 + \sum_{i=1}^{p} X_i(\mathrm{s})\beta_i$$

- The likelihood is:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\lambda(\mathrm{s}_i)}{\int_{\mathcal{S}} \lambda(\mathrm{s})ds} = \frac{exp[\sum_{i=1}^{n} X(\mathrm{s}_i)'\boldsymbol{\beta}]}{[\int_{\mathcal{S}} \lambda(\mathrm{s})ds]^n}$$

  - The integration is difficult to compute.
  - In practice, this integration is often approximated with a sum: Partition $\mathcal{S}$ into $B_1, \ldots, B_m$ and the integration is approximated with $\sum_{i=1}^{m} exp(\mathrm{X}_i'\boldsymbol{\beta})$
- Reading assignment: Hessellund, Xu, Guan & Waagepetersen (2021) Semiparametric Multinomial Logistic Regression for Multivariate Point Pattern Data, Journal of the American Statistical Association, DOI: 10.1080/01621459.2020.1863812.

UNIVERSITY OF Cincinnati

Clustering

- An intuitive clustering model is to build it based on a parent Poisson point process:
  - The parent process is a Poisson point process, and we get $v_1, \dots, v_m$ follow a homogeneous Poisson point process. These m locations are centers of clusters.
  - For cluster j, $j = 1, \dots, m$, we assume:

$$n_j \sim Poisson(\lambda_0)$$

$$s_{ji} \sim N(v_j, \sigma^2 I), \text{ for } i = 1, \dots, n_j$$

- It can be fit using EM algorithm or MCMC.

Cox Process

Log-Gaussian Cox Process:

$$log(\lambda(\mathsf{s})) = \mathsf{X}(\mathsf{s})'\boldsymbol{\beta} + \theta(\mathsf{s})$$

$$\theta(\mathsf{s}) \sim GP(\mu(\cdot), C(\cdot, \cdot))$$

- Likelihood is difficult to deal with due to the integration... approximation is needed.
- Reading assignment:
  - Chakraborty, Avishek, et al. Point Pattern Modelling for Degraded Presence-Only Data over Large Regions. Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 60, no. 5, 2011, pp. 757-776.
  - D. Simpson, J. B. Illian, F. Lindgren, S. H. Sorbye, H. Rue, Going off grid: computationally efficient inference for log-Gaussian Cox processes, Biometrika, 103, 2016, 49-70.

Marked Point Process

- A marked point process is composed of a point process and associate marks:

$$\{(\mathrm{s}_i, m_i) : i = 1, \ldots, n\}$$

  where $\mathrm{s}_1, \ldots, \mathrm{s}_n$ are locations and $m_1, \ldots, m_n$ are associated marks.
- Example:
  - Simulate from a homogeneous Poisson point process with $\lambda(\mathrm{s}) = 10$ on $[0, 1]^2$
  - Generate $m_i \overset{iid}{\sim} Exp(1)$ at each s_i.
  - The marked point process $\{(\mathrm{s}_i, m_i)\}$ is obtained.

A Parametric Model

Ho and Stoyan (2008; Statistics and Probability Letters, 1194-1199):

$$\lambda(\mathrm{s}) = exp(\alpha + \beta S(\mathrm{s}))$$

$$m(\mathrm{s}) = S(\mathrm{s}) + \epsilon(\mathrm{s})$$

$$S(\mathrm{s}) \sim \mathcal{GP}(\mu(\cdot), C(\cdot,\cdot))$$

That is, using a GP to link m and λ.

- There are a lot of interesting research problems related to point process:
    - Nonstationary proces: estimation of intensity functions
    - Spatio-tempolra point process
    - Relationship between points and marks
- Requiring more work

Summary

- Point process