Thinned Nonhomogenous Poisson Processes for Presence-Only data

Max Savery

4/6/25

Table of contents

Welcome **3**

1 Introduction **4**

2 Poisson process theory **5**

3 Presence-only data generation **8**

Welcome

This is a discussion about using Nonhomogenous Poisson Point Processes for modelling Presence-Only data. It reviews the Poisson Point Process, Presence-Only data, data generation using the process, and relevant modelling details using the Bayesian programming language Stan.

Please reach out to max.savery@ugent.be with any questions

1 Introduction

In this document we discuss models for presence-only data, using Thinned Nonhomogenous Poisson Processes and Log Gaussian Cox Processes. The initial planned layout is as follows: Introduce presence-only data, Poisson Process models and the data generating process. Then, proceed to fit a Thinned Nonhomogenous Poisson Process in Stan (and in spatstat), on simulated presence-only data. Later blog posts will discuss the relationship between PO and PA data and their intertwined data generation.

The rest of this document follows the above description, including theory where necessary.

2 Poisson process theory

The poisson process describes a set of points S where $\{s_i\} \subseteq D$ with intensity λ. Given an area or quadrat called subregion A of D of size $|A|$ centered around site s_i we can model the number of individuals occurring at that site $N_s(A)$ with a poisson distribution. This distribution will have mean

$$\Lambda(A) = \int \lambda(s)ds$$

For the demonstration here, we assume $|A| = 1$. However, the parameters in the Poisson Point Process are invariant to the scale of the data.

Typically, we model the intensity with a log-linear function, for example with parameters α and β,

$$\lambda(s) = \exp[\alpha + \beta' x(s)]$$

so that the expected number of observations in region A is now

$$\Lambda(A) = \int_A \exp[\alpha + \beta' x(s)]ds$$

We quickly notice that this integral is intractable and we will need a method of numerical integration. Indeed this will be discussed and explored here.

Consider, however, that we are going to be dealing with environmental data with environmental covariates, which are going to be discretized at some resolution inherently. Therefore, we can think of the inhomogenous poisson process as a continuous limit of a discretized set of conditionally independent poisson random variables, where the IPP emerges as the discretization grows smaller, i.e., $\Lambda(ds), ds \to 0$. In fact, using this continuous limit is how the likelihood for the PP can be derived (Banerjee, 2014. p203-4). However, considering that we will be working in a discretized space, the continuous framework ends up being less practically useful.

We now define the likelihood for the NHPP (Banerjee, 2014, p214)

$$f(s_1, s_2 ... s_n | N(D) = n) = \prod_i \frac{\lambda(s_i)}{(\lambda(D))^n}$$

and the joint density will be

$$f(s_1, s_2 ... s_n, N(D) = n) = \prod_i \frac{\lambda(s_i)}{(\lambda(D))^n} \left[(\lambda(D))^n \frac{\exp(-\lambda(D))}{n!} \right]$$

where we can see the second term on the right corresponds to the poisson likelihood for the number of total observations in space D. The likelihood will then be

$$L(\lambda(s)|s_1, s_2, ..., s_n) = \prod_i \lambda(s_i) \frac{\exp(-\lambda(D))}{n!}$$

We can move between the continuous and discrete versions of this likelihood by partitioning D into a grid of c cells and taking the poisson likelihood

$$L(\lambda) = \prod_c (\lambda(A_c))^{N(A_c)} \exp(-\lambda(A_c)) \tag{2.1}$$

Noticing that we can sum all the exponents, we get

$$L(\lambda) = \prod_c (\lambda(A_c))^{N(A_c)} \exp(-[\lambda(A_1) + \lambda(A_2) + \cdots + \lambda(A_c)) = \prod_c (\lambda(A_c))^{N(A_c)} \exp(-\lambda(D)]$$

and $N(A_c) = 1/0, |A_c| \to 0$. The term on the right indeed reduces to 2.1. We can then safely work with the Poisson likelihood given our grid is fine enough. This will be explored later.

We still have, however, one issue. We originally defined the mean of the of region A according to the properties of the PP,

$$\Lambda(A) = \int_A \lambda(s) ds = \int_A \exp[\alpha + \beta' x(s)] ds$$

Suddenly we are saying that we can just work with the intensity for region A without clarifying what we should do about the integral over the functional for intensity. Unfortunately, while we may want the integral over s, we don't have information at the resolution of ds. All we have are covariates $\beta' X$. That is,

$$\int_A \exp[\alpha + \beta' x(s)] ds \approx |A| \exp[\alpha + \beta' x]$$

Notice that we no longer have $x(s)$. We assume the covariate information is relative constant over region A. If this doesn't hold, our approximation will not be correct. But, given that it is all the information we have, the assumption is built into any modelling procedure and is less an issue with the Poisson Process and more just an issue of covariate resolution which is a common modelling issue in spatial-temporal statistics. As is stated in Banerjee, 2014 (p216) "In the absence of finer covariate resolution, we cannot do better with regard to the ecological fallacy."

Also, originally, we mentioned that it is convenient to assume $|A| = 1$. This is now the reason, so we don't have to worry about the scaling factor.

Given this approximation and scaling assumption, we now have

$$\Lambda(A) = \int_A \exp[\alpha + \beta' x(s)] ds \approx |A| \exp[\alpha + \beta' x] = \frac{|D|}{c} \exp[\alpha + \beta' x]$$

where c is the number of cells and $|D|$ is our total area. Then the counts in subregion A will be distributed as

$$N(A_i) \sim Poisson\left(\frac{|D|}{c}\exp[\alpha + \beta' x_i]\right) = Poisson\left(\exp[\alpha + \beta' x_i]\right)$$

if the number of cells is equal to the area ($|A| = 1$). Based on this, we end up with the log-likelihood for our entire presence-only dataset:

$$l(\alpha, \beta) = \sum_{i \in B} N(A_i)(\alpha + \beta' x_i) - \frac{|D|}{c}\sum_{i \in B}\exp[\alpha + \beta' x_i] - \sum_{i \in B}\log(N(A_i))!$$

where $i \in B$ refers to the set of background points $B = 1, 2, ..., c$. We take this likelihood over the entire set of background points in order to appropriately estimate the integral

$$\Lambda(D) = \int_D \lambda(s) ds$$

We need all points in the grid to approximate the integral, as is clear in the original likelihood. Thus, we have defined our model and the likelihood that we can use for PO data.

3 Presence-only data generation

Having introduced some theory for Poisson Point Process models, we next move on to data simulation. Here we simulate some presence-only data, in preparation to fit a nonhomogenous poisson process model. We specify the intensity for each site, the total number of individuals, the area of each site (assumed to be 1 here), and the sampling bias.

Following Fithian 2015, each observation i is associated with a site s_i area D. For each site s_i, there are associated covariates $x_i = x(s_i)$ and $z_i = z(s_i)$. For each survey site, s_i represents the centroid of a quadrat Ai. This is an essential assumption in our model, where we will use the Poisson likelihood to for the number of counts in each quadrat. At s_i we observe counts $N_i = N(A_i)$ or binary presence/absence indicators y_i, with $y_i = 1$ if $N_i > 0$ and $y_i = 0$ otherwise.

One issue in generating the data is that in using a poisson model to simulate the counts for each site, we lose individual point information in aggregating the points to counts per quadrat. However, this issue of aggregation is unavoidable, because we are using covariates that correspond to each cell. This is an important point. The assumptions we make about the grid will actually effect the way we simulate data. For example, when the only covariate data we have is that which is associated with the quadrat we are working with, then to simulate the data, each point realization will be created in aggregate, that is, a draw from a poisson distribution characterized by an functional of intensity dependent only on the covariates.

If our covariates exist on a finer scale than the quadrat, we can simulate the point locations more finely as well, and take a simulation approach outlined on page 2020 of Banerjee 2014 and followed by Koshkina, 2017. In this approach, we simulate first the total number of observations, create a homogenous poisson process, and then thin the process at the level of the grid of covariates we are working with. It doesn't completely make sense to simulate PPP data in this way here, because our covariates only exist at the level of the aggregated quadrat. That is, we assume that the intensity is constant across our region A, because according to our true data generating process, it is. This is also convenient for the approximation we are using.

We must also take into account the bias associated with the observation of species with true intensity $\lambda(s)$. In order to account for the bias caused by collecting presence-only observations, the poisson process is thinned by the sampling bias b. For more information about thinning a poisson process, see https://math.stackexchange.com/questions/580883/thinning-a-poisson-process

$$\lambda(s) = \theta(s)b(s) = \exp[\alpha + \beta' x(s) + \gamma + \delta' z(s)].$$

The question of the data generating process under our modelling assumptions is an interesting one. Referencing our initial theoretical outline, we need to simulate points with likelihood

$$\prod_i \lambda(s_i) \frac{\exp(-\lambda(D))}{n!}.$$

In the homogenous case, we can simulate n observations from

$$N(D) \sim Poisson(\lambda|D|),$$

($\lambda(s) = \lambda$, and is thus constant over D), and then according to the n, distribute these uniformly over D. In the case of the Nonhomogenous Poisson Process, we can do the same, drawing n now from $N(D) \sim Poisson(\lambda(D)) = Poisson(\int_D \lambda(s)ds)$, and then distributing their locations over D. Instead of uniformly distributing them however, they are placed according to the distribution $\frac{\lambda(s)}{\lambda(D)}$. There are two issues here, both relating to the evaluation of the integral for the expected number of observations.

Remember that we are using as intensity

$$\lambda(A_i) = \int_A \lambda(s)ds = \int_A \exp[\alpha + \beta' x(s)]ds = \exp[\alpha + \beta' x_i],$$

because we only have information X_i for quadrat A_i centered around s_i. While we are doing simulations, we could arbitrarily create a finer covariate grid but we won't have access to this during the modelling process. Using the x_i we have, both $\lambda(A)$ and $\lambda(D)$ will be approximations at the covariate resolution, using the assumption that each x_i is constant over A_i. This being the case, then for each observation, we will use the same distribution as the other points that fall into the same A_i. That is, multiple points will have the same distribution $\frac{\exp[\alpha+\beta' x_i]}{\sum_i \exp[\alpha+\beta' x_i]}$. This goes somewhat against the idea of distributing the points in continuous space. Therefore, given the independence of each A_i conditional on the covariates, another approach is to generate counts for each A_i and then uniformly distribute them within each A_i. If x_i is constant across A_i, this is ok. However, if we cannot assume this, using the continuous probability distribution for each point ($\lambda(s)/\lambda(D)$) is necessary to avoid the integral of $\lambda(A)$. We can do so in a way that avoids computing $\lambda(D)$ as well (Banerjee, 2014. p220), by computing λ_{max}, simulating n from $N(D) \sim Poisson(\lambda_{max}|D|)$, and the rejecting some of these points by $\lambda(s_i)/\lambda_{max}$. In this way, the only thing we need is information at the level of $\lambda(s_i)$ and avoid $\lambda(D)$ and $\lambda(A)$.

So note that we take the route of simulation depending on our available information. If we only have x_i for each quadrat, directly generating the points as a poisson random variable is a good option because it avoids any complicated integrals. But if we have finer covariates than at the quadrat level we are working with, then thinning the process down with the approach of Banerjee and others is better because it also avoids the integrals and has finer resolution.

There are a few things to during the data gerating process with the NHPP. The area of the total space, $|D|$, the area of each site (quadrat) $|A_i|$, and the number of cells c that D is discretized

into. The area of A depends on the number of discretizations (or cells): $|A_i| = |D|/c$. The intensity will be scaled by this factor $|D|/c$, so that as the area of $|A|$ becomes smaller, the intensity converges to that of a "true" continuous poisson process with its center at A_i. We also assume that the intensity is constant over A, so that we can indeed scale the intensity by the area factor, instead of needing to integrate over A. This assumption is in contrast to assuming that we model the instensity directly as $\Lambda(A_i) = \exp(\beta' X)$ (versus $\Lambda(A_i) = \frac{|D|}{c} \exp(\beta' X)$).