Stat 8501 Lecture Notes
Spatial Point Processes
Charles J. Geyer
February 23, 2020

1 Introduction

A *spatial point process* is a random pattern of points, both the number of points and their locations being random. This is unlike the stochastic processes we have been studying, in that it is not thought of as a collection of random variables $X(t)$ for $t \in T$ for some set T. We are only interested in processes having the following properties

(i) At most one point of the process is observed at any location.

(ii) The probability of observing a point at any prespecified location is zero.

(iii) The probability of observing an infinite number of points is zero.

These properties are called (i) *simple*, (ii) *no fixed atoms*, and (iii) *finite.* In some contexts (iii) is replaced by the following.

(iv) The probability of observing an infinite number of points in a bounded region is zero.

Property (iv) is called *boundedly finite.* So, when we use (iv) instead of (iii), we are interested in simple, boundedly finite spatial point processes with no fixed atoms. When we use (iii), the domain A can be an arbitrary set. When we use (iv), the domain A must have some notion of boundedness. This is no problem when A is a subset of $\mathbb{R}^d$ for some d. We just use the usual notion of boundedness.

If we were to try to characterize the process by a Bernoulli random variable $X(t)$ at each location t, then property (ii) would just make all the Bernoulli random variables almost surely zero. So that won't work.

Another way to think about the process is that the total number of points N is a nonnegative-integer-valued random variable and conditional on N the locations of the N points are a random vector taking values in A^n.

Yet another way is to think the process as given by random variables $N(B)$ for $B \in \mathcal{B}$, where $\mathcal{B}$ is a sigma-algebra in A. $N(B)$ is the number of points of the process in the region B. So this is somewhat like our previous notion of thinking of a stochastic process as a collection of random variables, but $B \in \mathcal{B}$ is quite different from anything we have seen before. Just from the definition, if B_1, ..., B_n are disjoint regions, then

$$N\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} N(B_i)$$

so there is strong dependence among these "counts" variables.

2 Poisson Processes

Theorem 1. *For a simple, boundedly finite spatial point process with no fixed atoms in $\mathbb{R}^d$ suppose the following condition holds ($\mathcal{B}$ is the Borel sigma-algebra for $\mathbb{R}^d$).*

(v) *For any choice B_1, B_2, ..., B_k of disjoint elements of $\mathcal{B}$, the random variables $N(B_1)$, $N(B_2)$, ..., $N(B_k)$ are independent.*

Then the random variables $N(B)$, $B \in \mathcal{B}$ all have Poisson distributions, and

$$\Lambda(B) = E\{N(B)\}, \qquad B \in \mathcal{B}, \tag{1}$$

defines a countably additive measure Λ on $(A, \mathcal{B})$ that has no atoms.

In the conclusion of the theorem we allow degenerate Poisson distributions concentrated at zero. Indeed condition (ii) says that $N(\{x\})$ is zero almost surely for any location x.

We also allow degenerate Poisson distributions concentrated at infinity. This is the limit as $\mu \to \infty$ of Poisson distributions with mean μ. Because

$$\frac{\Pr(X \le n)}{\Pr(X > n)} = \frac{\sum_{x=0}^{n} \frac{\mu^x}{x!} e^{-\mu}}{\sum_{x=n+1}^{\infty} \frac{\mu^x}{x!} e^{-\mu}} = \frac{\sum_{x=0}^{n} \frac{\mu^{x-n}}{x!}}{\sum_{x=n+1}^{\infty} \frac{\mu^{x-n}}{x!}}$$

and all the terms in the numerator on the right-hand side go to zero or are constant when $\mu \to \infty$ while and all the terms in the denominator on the right-hand side go to infinity, the limit distribution gives probability zero to $\{0, \ldots, n\}$ for any n. Hence the limit distribution is concentrated at infinity (if we allow ∞ as a value for the random variable). If $N(B) = \infty$ almost surely, then $\Lambda(B) = \infty$ too. By assumption (iv) a necessary condition for $\Lambda(B) = \infty$ is that B is an unbounded region, but this is not a sufficient condition (an unbounded region can have $\Lambda(B)$ finite).

The measure Λ defined by (1) is called the *parameter measure* of the process. Of course, the mean is the usual parameter of the Poisson distribution, so $N(B)$ is Poisson with mean $\Lambda(B)$.

The Poisson process is said to be *homogeneous* if Λ is proportional to Lebesgue measure (length in one dimension, area in two dimensions, volume in three dimensions, and so forth). Otherwise it is said to be *inhomogeneous*.

If one wants a Poisson process in a region A that is a measurable subset of $\mathbb{R}^d$, the theorem applies if we extend the process on A to all of $\mathbb{R}^d$, by defining $N(A^c) = 0$ almost surely.

Daley and Vere-Jones (2003, Theorems 2.4.II, 2.4.III and 2.4.VII) prove this theorem with the domain of the point process being a complete separable metric space.

Proof. Define

$$P_0(B) = P\{N(B) = 0\}, \qquad B \in \mathcal{B}.$$

If the conclusion of the theorem held (which we are not assuming), then we would have

$$P_0(B) = e^{-\Lambda(B)}$$

so we use this as motivation to define Λ by

$$\Lambda(B) = -\log P_0(B), \qquad B \in \mathcal{B} \tag{2}$$

(this is now our definition, not (1), which we now have to prove).

By definition of point process, if B_1 and B_2 are disjoint regions, the event $N(B_1 \cup B_2) = 0$ is the same as the event $N(B_1) = 0$ and $N(B_2) = 0$. Hence the multiplication rule gives

$$\begin{aligned}
\Lambda(B_1 \cup B_2) &= -\log[P_0(B_1 \cup B_2)] \\
&= -\log[P_0(B_1)P_0(B_2)] \\
&= -\log P_0(B_1) - \log P_0(B_2) \\
&= \Lambda(B_1) + \Lambda(B_2)
\end{aligned}$$

Hence Λ is finitely additive by mathematical induction.

To show that Λ is countably additive, let B_1, B_2, $\ldots$ be a sequence of disjoint measurable sets, and define

$$C_n = \bigcup_{i=1}^{n} B_i$$

and

$$C_\infty = \bigcup_{i=1}^\infty B_i$$

To show Λ is countably additive, we need to show $\Lambda(C_n) \uparrow \Lambda(C_\infty)$. Let D_n denote the event $N(C_n) = 0$ and D_∞ the event $N(C_\infty) = 0$. It follows from $C_1 \uparrow C_\infty$ that $N(C_n) \uparrow N(C_\infty)$ and $D_n \downarrow D_\infty$. Hence $P_0(D_n) \downarrow P_0(D_\infty)$ by continuity of probability, and that implies $\Lambda(C_n) \uparrow \Lambda(C_\infty)$.

Now Λ has an atom at x if and only if $\Lambda(\{x\}) > 0$, which happens if and only if $P_0(\{x\}) < 1$, which happens if and only if $P(N(\{x\})) > 0$, which is what it means for x to be a fixed atom of the point process. Hence assumption (ii) implies Λ has no atoms.

Fix a bounded set B, and choose a hyperrectangle

$$R = \{\, x \in \mathbb{R}^d : l_i \le x_i < u_i \text{ for all } i \,\}$$

that contains B. For each positive integer n define

$$J_n = \{\, j \in \mathbb{N}^d : 1 \le j_i \le 2^n \text{ for all } i \,\}$$

(so the cardinality of J_n is 2^{nd}) and define

$$A_{nj} = \left\{\, x \in \mathbb{R}^d : 2^{-n}(j_i - 1) \le \frac{x_i - l_i}{u_i - l_i} < 2^{-n} j_i \text{ for all } i \,\right\}, \qquad j \in J_n$$

(note that in A_{nj} the index n is a scalar integer but the index j is a vector with integer components) so for each n the sets A_{nj} are 2^{nd} hyperrectangles that partition R. Moreover, these partitions are nested, meaning every $A_{n+1,j}$ is contained in some $A_{nj'}$.

Then define Bernoulli random variables

$$X_{nj} = \begin{cases} 1, & N(B \cap A_{nj}) > 0 \\ 0, & \text{otherwise} \end{cases}$$

(again the subscript n is scalar but the subscript j is a vector, and for each n there are 2^{nd} of these random variables), and define for each n

$$N_n(B) = \sum_{j \in J_n} X_{nj}. \tag{3}$$

Because of the independence assumption and because the A_{nj} for fixed n and varying j are disjoint, the random variables summed in (3) are independent and the generating function of $N_n(B)$ is

$$\begin{aligned}\varphi_n(s) &= E\{s^{N_n(B)}\} \\ &= \prod_{j \in J_n} E\{s^{X_{nj}}\} \\ &= \prod_{j \in J_n} \left(e^{-\Lambda(B \cap A_{nj})} + s \left[1 - e^{-\Lambda(B \cap A_{nj})} \right] \right) \\ &= \prod_{j \in J_n} \left(1 + (s - 1) \left[1 - e^{-\Lambda(B \cap A_{nj})} \right] \right)\end{aligned}$$

Because of assumption (iii) $N(B)$ is almost surely finite, that is, $N(B)(\omega)$ is finite for almost all ω in the sample space. Because of assumption (i) the points counted by $N(B)(\omega)$ are at distinct locations. Hence for sufficiently large n, each point is in a different A_{nj} and $N_n(B)(\omega) = N(B)(\omega)$ for that n and all larger n. Also $N_n(B)(\omega)$ is nondecreasing in n for all ω. Hence $N_n(B) \uparrow N(B)$ and by monotone convergence the generating function of $N(B)$ is

$$\begin{aligned}\varphi(s) &= E\{s^{N(B)}\} \\ &= \lim_{n \to \infty} \varphi_n(s) \\ &= \lim_{n \to \infty} \prod_{j \in J_n} \left(1 + (s - 1) \left[1 - e^{-\Lambda(B \cap A_{nj})} \right] \right)\end{aligned}$$

We now claim

$$\Lambda(A_{nj}) \to 0, \qquad \text{as } n \to \infty, \tag{4}$$

which we prove by contradiction. Write

$$\varepsilon_n = \max_{j \in J_n} \Lambda(A_{nj})$$

so (4) is the same as $\varepsilon_n \to 0$. The sequence ε_n is nonincreasing because of the nesting of the partitions. Hence, if the claim is false, then there is $\delta > 0$ such that $\varepsilon_n \geq \delta$ for all n. Hence there exists a sequence j_n such that $\Lambda(A_{n,j_n}) \geq \delta$ for all n. Choose $x_n \in A_{n,j_n}$. Then because the closure of R is compact, there exists a convergent subsequence $x_{n_k} \to x$. Let

$$S_\eta = \{\, y \in \mathbb{R}^d : |y_i - x_i| < \eta \text{ for all } i \,\}$$

So there exists an integer k such that $A_{n_k,j_{n_k}} \subset S_\eta$, which implies $\Lambda(S_\eta) > \delta$. But $S_\eta \downarrow \{x\}$ as $\eta \downarrow 0$, and by continuity this implies

$$\Lambda(\{x\}) = \lim_{\eta \downarrow 0} \Lambda(S_\eta) \geq \delta$$

which contradicts Λ having no atoms, and this contradiction proves (4).

We return to our generating function calculation. The Taylor series with remainder $1 - e^{-\delta} = \delta + O(\delta^2)$ implies

$$1 - e^{-\Lambda(B \cap A_{nj})} = \Lambda(B \cap A_{nj}) + O(\Lambda(B \cap A_{nj})^2)$$

and

$$\begin{aligned} \log \varphi(s) &= \lim_{n \to \infty} \sum_{j \in J_n} \log \left(1 + (s-1) \left[1 - e^{-\Lambda(B \cap A_{nj})} \right] \right) \\ &= \lim_{n \to \infty} \sum_{j \in J_n} \log \left(1 + (s-1) \left[\Lambda(B \cap A_{nj}) + O(\Lambda(B \cap A_{nj})^2) \right] \right) \end{aligned}$$

Now use the Taylor series with remainder $\log(1 + \delta) = \delta + O(\delta^2)$

$$\begin{aligned} \log \varphi(s) &= \lim_{n \to \infty} \sum_{j \in J_n} \log \left(1 + (s-1) \left[\Lambda(B \cap A_{nj}) + O(\Lambda(B \cap A_{nj})^2) \right] \right) \\ &= \lim_{n \to \infty} \sum_{j \in J_n} (s-1) \left[\Lambda(B \cap A_{nj}) + O(\Lambda(B \cap A_{nj})^2) \right] \\ &= \lim_{n \to \infty} (s-1) \left[\Lambda(B) + \sum_{j \in J_n} O(\Lambda(B \cap A_{ni})^2) \right] \\ &= (s-1)\Lambda(B) \end{aligned}$$

the last step being because

$$\sum_{j \in J_n} \Lambda(B \cap A_{ni})^2 \le \varepsilon_n \sum_{j \in J_n} \Lambda(B \cap A_{ni}) = \varepsilon_n \Lambda(B)$$

which goes to zero. Hence the generating function of $N(B)$ is

$$\varphi(s) = e^{(s-1)\Lambda(B)}$$

which is easily checked to be the generating function of the Poisson distribution with mean $\Lambda(B)$. Hence for bounded sets B the assertion of the theorem follows from generating functions corresponding to unique distributions (Appendix A).

For unbounded sets B the assertion of the theorem follows from what we have already proved about bounded B. Define

$$R_n = \{ \, x \in \mathbb{R}^d : |x_i| \le n \, \}$$

Then $N(B \cap R_n)$ is Poisson with mean $\Lambda(B \cap R_n)$ and

$$N(B \cap R_n) \uparrow N(B).$$

It follows that $N(B)$ is Poisson with mean $\Lambda(B)$, regardless of whether $\Lambda(B)$ is finite or infinite. □

3 Cox Processes

A *Cox process* (Cressie, 1993, Section 8.5.2) is a hierarchical model in which the lowest level is an inhomogeneous Poisson process with parameter measure Λ, which at the next level of the hierarchy we take to be a random measure.

Although Cox processes can have quite complicated dependence structure, none of that structure can be seen in a single realization, which is what many, perhaps most, data sets which can be modeled by a spatial point process comprise. If multiple independent realizations of the process were in the data, then something could, at least in principle be inferred about the distribution of the random measure Λ. But for one realization of the process, you just have one realization of Λ, and you cannot infer anything about a distribution (any distribution) from one data point.

We will not consider them further.

4 Neyman-Scott Processes

A *Neyman-Scott process* (Cressie, 1993, Section 8.5.3) is a hierarchical model with three levels. The upper level is an inhomogeneous Poisson process, the points of which are called "centers" and are not observable. In the middle level independent and identically distributed (II) nonnegative-integer-valued random variables are generated, which are called "counts" and are also not observable. In the lower level, for each center x and corresponding count n_x, this many IID points are generated from a probability distribution centered at x (the same distribution for each center). The observed data are only the points generated in the lower layer of the hierarchy; these points are not labeled as to which center they correspond to.

If the random number of points for each center has a Poisson distribution, then the Neyman-Scott process is a Cox process, but not otherwise.

Cressie (1993, Section 8.4.3) shows that for a Neyman-Scott process whose upper level is a homogeneous Poisson process, whose middle level distribution has finite expectation, and whose lower level distribution is spherically symmetric, method of moments estimators of the parameters of the three levels are available.

5 More on General Point Processes

5.1 Sets or Vectors

For any simple, finite point process taking place in a region A, the realizations of the process are point patterns, each of which has some number of points. Since the points are indistinguishable, the points in a pattern have no order. Since the process is simple, there are no duplicates. Thus the point patterns have the properties of mathematical sets. Hence we can think of the point patterns as finite sets

$$x = \{x_1, \ldots, x_n\}$$

and use the notation of set theory for them. The point pattern having no points we denote by $\varnothing$. If x and y are point patterns, then $y \subset x$ means every point in y is also in x. If ξ is a point and x is a point pattern, then $\xi \in x$ means ξ is one of the points in x, and so forth. In one respect we will be a bit sloppy using this notation: if ξ is a point and x is a point pattern, then we will write $x \cup \xi$ to mean the point pattern whose points are ξ and all of the points in x, when to be fussy we should write $x \cup \{\xi\}$.

So far, so good, but we run into trouble when we want to do integrals. When we do multiple integrals the variables are vectors (ordered tuples in which duplicates are allowed) rather than sets (unordered tuples in which duplicates are not allowed). So in order to use the notation of calculus we need to also think of point patterns as vectors (of any length, including length zero, which is the empty point pattern $\varnothing$). When we think of the realizations of the point process as variable length vectors rather than sets, we can write the state space as

$$\Omega = \bigcup_{n=0}^{\infty} A^n \tag{5}$$

where $A^0 = \{\varnothing\}$, $A^1 = A$, $A^2 = A \times A$ and so forth. Note that A^0 is a set with one element (the empty point pattern).

When we say $x \in \Omega$ is a point pattern, that means $x \in A^n$ for some n, which means x can be though of as a vector of length n all of whose components are points of A. If A is a subregion of $\mathbb{R}^d$, then each point of A requires d numbers to specify it, so thought of as a vector of numbers rather than as a vector of points x has length nd rather than n, but this complication only arises writing computer programs for point processes; it needn't complicate our notation.

The vector notation is wrong in the sense that it seems to indicate that order matters when it does not. But as long as we don't make the mistake of thinking that order matters, it is o. k.

5.2 Integration with respect to the Poisson Process

Let P denote the probability measure of a Poisson process with parameter measure Λ on a region A. We assume $\Lambda(A)$ is finite. The probability of observing n points is

$$\frac{\Lambda(A)^n}{n!} e^{-\Lambda(A)}$$

and the conditional distribution of x given that x has n points is

$$\frac{\Lambda^n(dx)}{\Lambda(A)^n}$$

where Λ^n is n-fold product measure on A^n. Then for any measurable function $h : \Omega \to \mathbb{R}$ we have

$$\int_\Omega h(x) P(dx) = e^{-\Lambda(A)} \sum_{n=0}^\infty \frac{1}{n!} \int_{A^n} h(x) \Lambda^n(dx) \tag{6}$$

provided the integral exists. We will actually only use the special case where P is the probability measure of a homogeneous Poisson process, in which case we have

$$\Lambda(A) = \lambda v(A),$$

where

$$v(A) = \int_A dx$$

is Lebesgue measure of A (length in one dimension, area in two dimensions, volume in three dimensions, and so forth), and

$$\Lambda^n(dx) = \lambda^n \, d\xi_1 \cdots d\xi_n.$$

Thus

$$\int_\Omega h(x) P(dx) = e^{-\lambda v(A)} \left[h(\varnothing) + \sum_{n=1}^\infty \frac{\lambda^n}{n!} \int_A \cdots \int_A h(\xi_1, \ldots, \xi_n) \, d\xi_1 \cdots d\xi_n \right]$$

5.3 Densities with respect to the Poisson Process

We say h is an unnormalized density with respect to a Poisson process measure P if it is nonnegative and integrates to something (called the *normalizing constant*) that is nonzero and finite. Then h divided by the normalizing constant is a proper (normalized) probability density.

Thus to specify a point process model having an unnormalized density with respect to a Poisson process, we only need to specify a nonnegative-real-valued function h on Ω and then check, first, that h is not zero almost everywhere $[P]$ and, second, that $\int h(x)\, P(dx)$ is finite.

5.4 Conditional Intensity Functions

An unnormalized density h with respect to a Poisson process *has the hereditary property* if

$$h(x) > 0 \text{ and } y \subset x \text{ implies } h(y) > 0. \tag{7}$$

Let h be an unnormalized density with respect to the probability measure P of a homogeneous Poisson process on a bounded region A of $\mathbb{R}^d$, and suppose h has the hereditary property. The *Papangelou conditional intensity function* of the process having density h with respect to P is

$$\lambda(\xi \mid x) = \frac{h(x \cup \xi)}{h(x)}, \qquad x \in \Omega \text{ and } \xi \in A. \tag{8}$$

The hereditary property implies that we can only have divide by zero in (8) when both the numerator and denominator are zero, and in this case we set $\lambda(\xi \mid x) = 1$ (a convention).

Since we are not interested in other kinds of conditional intensity functions (Daley and Vere-Jones, 2003, Section 7.2, discuss the different kinds of conditional intensity functions), we will just call (8) the "conditional intensity function" (omitting the eponym).

6 Markov Point Processes

A *Markov point process* (Ripley and Kelly, 1977; Cressie, 1993, Section 8.5.5; Daley and Vere-Jones, 2003, Section 7.1) is a spatial point process having a characteristic property defined in terms of the conditional intensity function.

First Ripley and Kelly define a neighbor relation $\sim$ among points, which is symmetric ($\xi \sim \zeta$ implies $\zeta \sim \xi$) and reflexive $\xi \sim \xi$ for all ξ). An example of a neighbor relation is $\xi \sim \zeta$ if and only if $d(\xi, \zeta) \leq \rho$, where d is Euclidean distance and ρ is a known constant called the *range*.

A spatial point process having unnormalized density h with respect to a finite Poisson process has the *spatial Markov property* if h has the hereditary property and the conditional intensity (8) actually depends on x only through points in x that are neighbors of ξ.

If x is a point pattern and $y \subset x$, then we say y is a *clique* if every pair of points in y are neighbors. Let $\operatorname{clq}(x)$ denote the set of all cliques in x. Note that $\operatorname{clq}(x)$ is never empty, because the empty point pattern is a clique.

The so-called Hammersley-Clifford theorem for spatial point processes (proved by Ripley and Kelly, the original Hammersley-Clifford theorem, proved by Hammersley and Clifford, was for spatial lattice processes) is the following.

Theorem 2. *A spatial point process having unnormalized density h with respect to a finite Poisson process has the spatial Markov property if and only if h has the form*

$$h(x) = \prod_{y \in \operatorname{clq}(x)} \varphi(y) \tag{9}$$

for some nonnegative-valued function φ on Ω.

Proof. If h has the form (9) and $h(x) > 0$ then

$$\lambda(\xi \mid x) = \prod_{\substack{y \subset \operatorname{clq}(x \cup \xi) \\ y \not\subset \operatorname{clq}(x)}} \varphi(y)$$

and every element of every y on the right-hand side is a neighbor of ξ. If $h(x) = 0$ then $\lambda(\xi \mid x) = 1$ by convention and does not depend on x. That proves one direction.

Now suppose h has the spatial Markov property. We must have $h(\varnothing) > 0$ because $h(x)$ cannot be zero for all x and $h(x) > 0$ for any x implies $h(\varnothing) > 0$ by the hereditary property. In order to have (9) we must define

$$\varphi(\varnothing) = h(\varnothing). \tag{10}$$

Now define $\varphi(x)$ for nonempty cliques x inductively by

$$\varphi(x) = 0, \qquad \text{if } h(x) = 0, \tag{11a}$$

and

$$\varphi(x) = \frac{h(x)}{\prod_{y \subsetneq x} \varphi(y)}, \qquad \text{if } h(x) > 0. \tag{11b}$$

In order for this definition to make sense, we must show that the denominator in (11b) is always nonzero. This follows from mathematical induction on the number of points in x. The induction hypothesis is x is a clique and $h(x) > 0$ imply $\varphi(y) > 0$ for all $y \subsetneq x$. Call that $\mathrm{IH}(x)$. The base of the induction is when x contains a single point, in which case the only proper subset of x is the empty pattern, and we already know $\varphi(\varnothing) > 0$. The induction step must prove that $\mathrm{IH}(x)$ from the assumption that $\mathrm{IH}(y)$ holds for all $y \subsetneq x$. If x is not a clique or $h(x) = 0$, then $\mathrm{IH}(x)$ holds vacuously. So assume x is a clique and $h(x) > 0$. Then for $y \subsetneq x$ we know that $h(y) > 0$ by the hereditary property and the denominator in

$$\varphi(y) = \frac{h(y)}{\prod_{z \subsetneq y} \varphi(z)}$$

is nonzero by $\mathrm{IH}(y)$. Hence $\varphi(y) > 0$.

Now we prove that (9) holds for our definition of φ. This too we prove by mathematical induction on the number of points in x. It is clear that (9) holds whenever x has zero or one points by (10) and

$$\varphi(x) = \frac{h(x)}{\varphi(\varnothing)}$$

when x has one point. When x has two or more points, there are two cases. If x is a clique, then (11a) and (11b) obviously imply (9). If x is not a clique, then there are ξ and ζ in x such that $\xi \not\sim \zeta$. Write $z = x \setminus \{\xi, \zeta\}$ so $x = z \cup \xi \cup \zeta$. If $h(x) = 0$, then (11a) implies (9). So we are left with the case $h(x) > 0$, which implies $h(z) > 0$ by the hereditary property. And

$$\begin{aligned} h(x) &= h(z \cup \xi)\lambda(\zeta \mid z \cup \xi) \\ &= h(z \cup \xi)\lambda(\zeta \mid z) \\ &= \frac{h(z \cup \xi)h(z \cup \zeta)}{h(z)} \end{aligned}$$

because $\lambda(\zeta \mid x \cup \xi)$ only depends on neighbors of ζ and ξ is not a neighbor of ζ. The induction hypothesis assumes that (9) holds for all sets of cardinality less than x, hence

$$\begin{aligned} h(x) &= \left[\prod_{y \in \mathrm{clq}(z \cup \xi)} \varphi(y)\right] \left[\prod_{y \in \mathrm{clq}(z \cup \zeta)} \varphi(y)\right] \Bigg/ \left[\prod_{y \in \mathrm{clq}(z)} \varphi(y)\right] \\ &= \prod_{y \in \mathrm{clq}(x)} \varphi(y) \end{aligned}$$

and we are done. □

7 Strauss Process

The *Strauss process* (Strauss, 1975) is a spatial point process that uses the neighbor relation $\xi \sim \zeta$ if and only if $d(\xi, \zeta) \leq \rho$, where d is Euclidean distance and $\rho > 0$ is a fixed constant, and uses only cliques of size 2 or less in the Hammersley-Clifford expansion. For any $x \in \Omega$, let

- $t_1(x)$ denote the number of points in x and
- $t_2(x)$ denote the number of unordered pairs of distinct points in x that are neighbors.

Then the Strauss process is the exponential family of distributions having $t_1(x)$ and $t_2(x)$ as its canonical statistics, which means the unnormalized density is

$$h_\theta(x) = e^{t_1(x)\theta_1 + t_2(x)\theta_2} = e^{\langle t(x), \theta \rangle}$$

where in the last expression $t(x)$ and θ are vectors of length 2.

We are specifying a family of probability distributions (a statistical model) with parameter vector θ (the canonical parameter vector of this exponential family). Since we have a family of densities, the normalizing "constant" is a function of θ (but not a function of x)

$$\begin{aligned} c(\theta) &= \int_\Omega h_\theta(x) P(dx) \\ &= e^{-\lambda v(A)} \sum_{n=0}^\infty \frac{\lambda^n}{n!} \int_A \cdots \int_A e^{n\theta_1 + t_2(x_1, \ldots, x_n)\theta_2} \, dx_1 \cdots dx_n \qquad (12) \\ &= e^{-\lambda v(A)} \sum_{n=0}^\infty \frac{\lambda^n e^{n\theta_1}}{n!} \int_A \cdots \int_A e^{t_2(x_1, \ldots, x_n)\theta_2} \, dx_1 \cdots dx_n \end{aligned}$$

It was a slight embarrassment for Strauss that the title of Strauss (1975) is "a model for clustering" but, as pointed out by Kelly and Ripley (1976), the Strauss process is *not* "a model for clustering" but only a model for anti-clustering because of the following theorem. (Not a major embarrassment because Strauss did propose the first Markov spatial point process widely used for real data.)

Theorem 3. *The integral* (12) *exists if and only if* $\theta_2 \leq 0$.

Proof. First, suppose $\theta_2 \leq 0$. Then, since $t_2(x) \geq 0$, we have

$$\int_A \cdots \int_A e^{t_2(x_1, \ldots, x_n)\theta_2} \, dx_1 \cdots dx_n \leq \int_A \cdots \int_A dx_1 \cdots dx_n = v(A)^n$$

and

$$\sum_{n=0}^{\infty} \frac{\lambda^n e^{n\theta_1} v(A)^n}{n!} = e^{\log(\lambda) + \theta_1 + \log v(A)}$$

Hence $c(\theta) < \infty$.

Second, suppose $\theta_2 > 0$, and consider a subset B of A having nonzero Lebesgue measure and diameter less than or equal to ρ. Then $x \in B^n$ implies $t_2(x) = n(n-1)/2$ (every pair of points of x is a neighbor pair), and

$$\begin{aligned} \int_A \cdots \int_A e^{t_2(x_1, \ldots, x_n)\theta_2} \, dx_1 \cdots dx_n &\ge e^{n(n-1)\theta_2/2} \int_B \cdots \int_B dx_1 \cdots dx_n \\ &= e^{n(n-1)\theta_2/2} v(B)^n \end{aligned}$$

and

$$c(\theta) \ge e^{-\lambda v(A)} \sum_{n=0}^{\infty} \frac{\lambda^n e^{n(n-1)\theta_2/2} v(B)^n}{n!}$$

and this infinite sum is not finite because Stirling's approximation says

$$\log(n!) = n \log(n) - n + O\bigl(\log(n)\bigr)$$

so

$$\begin{aligned} &\log\left(\frac{\lambda^n e^{n(n-1)\theta_2/2} v(B)^n}{n!}\right) \\ &\qquad \ge n \log \lambda + \frac{n(n-1)\theta_2}{2} + n \log v(B) - n \log n + n + O\bigl(\log(n)\bigr) \end{aligned}$$

and this goes to infinity as $n \to \infty$, hence so do the terms of the infinite sum, which consequently cannot converge. □

Thus the canonical parameter space of the full exponential family for the Strauss process is

$$\Theta = \{\, \theta \in \mathbb{R}^2 : \theta_2 \le 0 \,\} \tag{13}$$

This and related models (Section 10 below) are the only examples of non-regular exponential families that I know of that arise in actual applications.

An exponential family is *regular* (Barndorff-Nielsen, 1978, p. 116) if its full canonical parameter space Θ is an open set, which (13) is not. The point of regularity, is that it implies that the maximum likelihood estimate (MLE), if it exists, is any point where the first derivative of the log likelihood is zero

(Barndorff-Nielsen, 1978, Theorems 9.13 and 9.14). When the boundary of Θ is nonempty, as in (13), then if a boundary point is the MLE, the first derivative is not even defined there, and one has to use the methods of constrained optimization (Geyer and Møller, 1994).

For the Strauss process, the boundary points (where $\theta_2 = 0$) are homogeneous Poisson processes because for such θ we have

$$\begin{aligned} c(\theta) &= e^{-\lambda v(A)} \sum_{n=0}^{\infty} \frac{\lambda^n e^{n\theta_1}}{n!} \int_A \cdots \int_A dx_1 \cdots dx_n \\ &= e^{-\lambda v(A)} \sum_{n=0}^{\infty} \frac{\lambda^n e^{n\theta_1} v(A)^n}{n!} \\ &= e^{-\lambda v(A) + \lambda v(A) e^{\theta_1}} \end{aligned}$$

and for any measurable function f we have

$$\begin{aligned} E_\theta\{f(x)\} &= \frac{e^{-\lambda v(A)}}{c(\theta)} \sum_{n=0}^{\infty} \frac{\lambda^n e^{n\theta_1}}{n!} \int_A \cdots \int_A f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n \\ &= e^{-\lambda v(A) e^{\theta_1}} \sum_{n=0}^{\infty} \frac{\lambda^n e^{n\theta_1}}{n!} \int_A \cdots \int_A f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n \end{aligned}$$

and this is the same formula as for a Poisson process, the only difference is that the rate parameter λ has been replaced by λe^{θ_1}.

From the theory of exponential families (Barndorff-Nielsen, 1978, Theorem 8.2)

$$\begin{aligned} E_\theta\{t(x)\} &= \nabla \log c(\theta) \\ \operatorname{var}_\theta\{t(x)\} &= \nabla^2 \log c(\theta) \end{aligned}$$

Consequently,

$$\frac{\partial E_\theta\{t_2(x)\}}{\partial \theta_2} = \frac{\partial^2 c(\theta)}{\partial \theta_2^2} = \operatorname{var}_\theta\{t_2(x)\} > 0$$

Thus decreasing θ_2 decreases the expectation of $t_2(x)$, and a Strauss process with $\theta_2 < 0$ has fewer neighbor pairs on average than expected for a Poisson processes (the $\theta_2 = 0$ case).

8 The Hard Core Process

Geyer (2009), generalizing the theory in Barndorff-Nielsen (1978) and Brown (1986) constructs the completion of an exponential family by taking limits as parameters go to infinity. The limit in a direction δ exists (Geyer, 2009, Theorem 6) if and only if δ is a direction of recession of the family, which is characterized by Theorem 3 in Geyer (2009). A vector δ is a direction of recession if and only if there exists a constant M such that $\langle t(x), \delta \rangle \leq M$ for all $x \in \Omega$, in which case the limiting distributions (the limiting conditional model (LCM)) are the distributions in the original family conditioned on the event

$$\langle t(x), \delta \rangle = \max_{y \in \Omega} \langle t(y), \delta \rangle$$

For the Strauss process $t_2(x)$ is bounded below by zero. The LCM in the direction $(0, -1)$, which conditions on $t_2(x)$ having its minimum value, is called the *hard core* process. It is the Strauss process conditioned on the event $t_2(x) = 0$, that is, there are no neighbor pairs, every point is separated by a distance of at least ρ from every other point.

Unlike the Poisson process and the Strauss process, the hard core process cannot have an arbitrarily large number of points. When the region A in which the process takes place is bounded, then there is a maximum number of points that can be crammed into A while maintaining a separation of at least ρ (this maximum number may be difficult to calculate, but it does exist).

In taking the limit to form the LCM one loses a parameter. For the hard core process the value of θ_2 is irrelevant. It controls the mean value of $t_2(x)$, but we are fixing $t_2(x)$ at zero, leaving nothing for θ_2 to do. (More formally, if we calculate the conditional distributions, we see that θ_2 drops out of the formulas.)

Thus the hard core process is a one-parameter exponential family. Unlike the Strauss process, it is a regular exponential family. The full canonical parameter space is the whole real line.

We can continue the process of taking limits. If we take the limit as $\theta_1 \to -\infty$ we get the empty process, that has no points with probability one (and no parameters). If we take the limit as $\theta_1 \to +\infty$ we get the Poisson process conditioned on $t_2(x) = 0$ and $t_1(x)$ conditioned on having its maximal value (which we said was difficult to calculate). Again this has no parameters, if n is the maximal number of points that can be crammed into A while maintaining a separation of at least ρ, then this process is the distribution of n points uniformly distributed in A conditioned on the event $t_2(x) = 0$.

9 MCMC Simulation

Geyer and Møller (1994) proposed an MCMC simulation method for spatial point processes. Later on, it turned out to be a special case of the Metropolis-Hastings-Green (MHG) algorithm (Green, 1995), but rather than present all of the theory of the general MHG algorithm, we will just prove what we want to prove about the algorithm of Geyer and Møller.

Here is the method. Suppose we want to simulate a spatial point process that has an unnormalized density h with respect to the Poisson process on a region A that has parameter measure Λ. We describe one iteration of the Markov chain starting at x (a point pattern). This is an equal mixture of two Metropolis-like updates, which we call up moves and down moves. These work as follows. The up move.

- Generate a point ξ having distribution $\Lambda(\,\cdot\,)/\Lambda(A)$. Set $y = x \cup \xi$.
- Calculate
  $$R_{\text{up}} = \frac{h(y)\Lambda(A)}{h(x)(n+1)} \tag{14}$$
  where n is the number of points in x.
- Generate U uniform on $(0,1)$. If $U < R_{\text{up}}$ the state of the Markov chain at the next time is y. Otherwise, the state at the next time is x.

And the down move.

- If $x = \varnothing$, do nothing. The chain stays at x.
- If $x \neq \varnothing$ do the following.
  - Choose a point $\xi \in x$ uniformly at random. Set $y = x \setminus \xi$.
  - Calculate
    $$R_{\text{down}} = \frac{h(y)(n+1)}{h(x)\Lambda(A)} \tag{15}$$
    where n is the number of points in y.
  - Generate U uniform on $(0,1)$. If $U < R_{\text{down}}$ the state of the Markov chain at the next time is y. Otherwise, the state at the next time is x.

Those familiar with the Metropolis algorithm will notice that this algorithm is quite similar. There is a proposal (y), a ratio R_{up} or R_{down} is calculated, and "Metropolis rejection" is done based on the ratio. The only differences from the Metropolis algorithm are that x and y have different dimensions and (14) and (15) are not the Metropolis ratio, although they are Green ratios that occur in the general MHG algorithm.

If the current state x is possible under the Poisson process and satisfies $h(x) > 0$, then so will the next state. In (14) we have $h(x) > 0$ by assumption and, if $h(y) = 0$, then the proposal y is accepted with probability zero and cannot be the next state. Hence the next state must (with probability one) satisfy $h(y) > 0$. In (15) we have $h(x) > 0$ by assumption and $h(y) > 0$ by the hereditary property. Thus, if started in a state x that is possible under the Poisson process and satisfies $h(x) > 0$, the chain satisfies this condition at all times, and there can never be divide by zero in (14) or (15).

Theorem 4. *An invariant distribution of the sampler described above is the distribution that has an unnormalized density h with respect to the Poisson process on A that has parameter measure Λ.*

Proof. The transition probability kernel of the sampler is

$$P(x, B) = \frac{1}{2} I(\varnothing, B) + \frac{1}{2} \sum_{n=0}^{\infty} P_n(x, B)$$

where P_n describes the up move from n points to $n+1$ points and the down move from $n+1$ points to n points.

We will show that every P_n is reversible with respect to the desired invariant distribution. Since the identity kernel is reversible with respect to every distribution, and the sum of reversible is reversible, this implies P is reversible with respect to the desired invariant distribution. Since reversible with respect to a distribution implies that distribution is invariant, that proves the assertion of the theorem. Thus it only remains to show that each P_n is reversible.

Now

$$P_n(x, B) = r_n(x) I(x, B) + \int_B Q_n(x, dy) a_n(x, y)$$

where Q_n is the conditional distribution of the proposal y given the current state x for moves between n and $n+1$ points, where $a_n(x, y)$ is the acceptance probability in the Metropolis rejection, either $\min(1, R_{\text{up}})$ for up moves or $\min(1, R_{\text{down}})$ for down moves, and where

$$r_n(x) = 1 - \int Q_n(x, dy) a_n(x, y)$$

Let μ denote the measure of the Poisson process. Then reversibility is for any bounded measurable function f

$$\iint f(x,y)h(x)\mu(dx)P_n(x,dy)$$

is unchanged if $f(x,y)$ is replaced by $f(y,x)$. Using (6) and taking account of the fact that $P_n(x,B)$ is zero unless x has n or $n+1$ points, we get

$$\iint f(x,y)h(x)\mu(dx)P_n(x,dy) \\ = e^{-\Lambda(A)} \sum_{m=n}^{n+1} \frac{1}{m!} \int_{A^m} \int_\Omega f(x,y)h(x)\Lambda^m(dx)P_n(x,dy)$$

we can divide out the constants $e^{-\Lambda(A)}$ and $n!$ without affecting the reversibility verification. Hence it remains to show that

$$\int_{A^n} \int_\Omega f(x,y)h(x)\Lambda^n(dx)P_n(x,dy) \\ + \frac{1}{n+1} \int_{A^{n+1}} \int_\Omega f(x,y)h(x)\Lambda^{n+1}(dx)P_n(x,dy)$$

is unchanged if $f(x,y)$ is replaced by $f(y,x)$. Now

$$Q_n(x,B) = \Lambda(B)/\Lambda(A), \qquad x \in A^n$$

and

$$Q_n(y \cup \xi, B) = I(y,B), \qquad y \in A^n \text{ and } \xi \in A$$

(Why is there not a factor of $1/(n+1)$ because we are deleting a point ξ chosen "at random" from among the points of $y \cup \xi$? Because the points are indistinguishable. Another way to think of this is that the operation of permuting the points preserves every distribution, so could do a random permutation of the points at the beginning of each iteration of the Markov chain without changing any invariant distributions the chain has. Then when we delete a particular point, say the last one, it is already random.)

Hence

$$
\begin{aligned}
&\int_{A^n} \int_{\Omega} f(x, y) h(x) \Lambda^n(dx) P_n(x, dy) \\
&\qquad = \int_{A^n} \int_{\Omega} f(x, y) h(x) \Lambda^n(dx) \left[r_n(x) I(x, dy) + a_n(x, y) Q_n(x, dy) \right] \\
&\qquad\qquad = \int_{A^n} f(x, x) h(x) r_n(x) \Lambda^n(dx) \\
&\qquad\qquad\quad + \frac{1}{\Lambda(A)} \int_{A^n} \int_{A} f(x, x \cup \xi) h(x) a_n(x, x \cup \xi) \Lambda^n(dx) \Lambda(d\xi)
\end{aligned}
$$

and the first term on the right-hand side is unchanged when the arguments of f are swapped, so we don't need to worry about it further. And hence

$$
\begin{aligned}
&\int_{A^{n+1}} \int_{\Omega} f(x, y) h(x) \Lambda^{n+1}(dx) P_n(x, dy) \\
&\qquad = \int_{A^{n+1}} \int_{\Omega} f(x, y) h(x) \Lambda^{n+1}(dx) \left[r_n(x) I(x, dy) + a_n(x, y) Q_n(x, dy) \right] \\
&\qquad\qquad = \int_{A^{n+1}} f(x, x) h(x) r_n(x) \Lambda^{n+1}(dx) \\
&\qquad\qquad\quad + \int_{A^n} \int_{A} f(y \cup \xi, y) h(y \cup \xi) a_n(y \cup \xi, y) \Lambda^n(dy) \Lambda(d\xi)
\end{aligned}
$$

and, again, the first term on the right-hand side is unchanged when the arguments of f are swapped, so we don't need to worry about it further. Now what remains to be shown is that

$$
\begin{aligned}
&\frac{1}{\Lambda(A)} \int_{A^n} \int_{A} f(x, x \cup \xi) h(x) a_n(x, x \cup \xi) \Lambda^n(dx) \Lambda(d\xi) \\
&\qquad + \frac{1}{n+1} \int_{A^n} \int_{A} f(x \cup \xi, x) h(x \cup \xi) a_n(x \cup \xi, x) \Lambda^n(dx) \Lambda(d\xi)
\end{aligned}
$$

is unchanged when the arguments of f are swapped, and this will be the case if it is true of the integrands, that is, if

$$
\frac{1}{\Lambda(A)} f(x, x \cup \xi) h(x) a_n(x, x \cup \xi) + \frac{1}{n+1} f(x \cup \xi, x) h(x \cup \xi) a_n(x \cup \xi, x) \tag{16}
$$

is unchanged when the arguments of f are swapped for all $x \in A^n$ and $\xi \in A$ such that there is no divide by zero in the calculation of a_n (since we already know that happens with probability zero). Now we note that (14) and (15) are reciprocals so long as both $h(x) > 0$ and $h(y) > 0$. Thus we may assume

$h(x) > 0$ and $h(x \cup \xi) > 0$ when verifying that (16) is unchanged when the arguments of f are swapped. Now we have two cases. First, suppose $a_n(x, x \cup \xi) = 1$. Then (16) is

$$\frac{1}{\Lambda(A)} f(x, x \cup \xi) h(x) + \frac{1}{n+1} f(x \cup \xi, x) h(x \cup \xi) \cdot \frac{h(x)(n+1)}{h(x \cup \xi)\Lambda(A)}$$

and this is unchanged when the arguments of f are swapped. Second, suppose $a_n(x \cup \xi, x) = 1$. Then (16) is

$$\frac{1}{\Lambda(A)} f(x, x \cup \xi) h(x) \cdot \frac{h(x \cup \xi)\Lambda(A)}{h(x)(n+1)} + \frac{1}{n+1} f(x \cup \xi, x) h(x \cup \xi)$$

and this is unchanged when the arguments of f are swapped. □

Theorem 5. *The sampler described above is irreducible if h has the hereditary property.*

Recall that the hereditary property is (7).

Proof. We have to show the sampler is φ-irreducible for some positive measure φ that is not the zero measure. We choose φ to be concentrated at the empty point pattern. Thus we need to show that the sampler can get from any point pattern x to the empty point pattern in a finite number of steps. We claim if x has n points that it can do so in n steps. The sampler can do a down move in each of the n steps with positive probability, and each of those proposals is accepted with positive probability because of the hereditary property: if $h(x) > 0$ then $h(y) > 0$ for all $y \subset x$ so (15) is never zero. □

So now we know that h is the unnormalized density of the unique invariant distribution.

Theorem 6. *The sampler described above is geometrically ergodic if the distribution having unnormalized density h has bounded conditional intensity.*

Recall that the conditional intensity is (8), the assumption of the theorem is that there exists a constant M such that

$$h(x \cup \xi) \le M h(x), \qquad x \in \Omega \text{ and } \xi \in A. \tag{17}$$

Proof. First note that (17) implies that h has the hereditary property.

Next we claim that every set consisting of point patterns having at most n points is a small set in the terminology of Markov chain theory (Meyn and Tweedie, 2009, Section 5.2): a set C is *small* if there exists a positive measure ν that is not the zero measure and an integer m such that

$$P^m(x, B) \geq \nu(B), \qquad x \in C,\ B \in \mathcal{B},$$

where P is the transition probability kernel of the Markov chain and $\mathcal{B}$ is the sigma-algebra for the state space. Here we take ν to be concentrated at the empty point pattern and $m = n$. The proof is just like the proof of Theorem 5 except for using (17). Again we know that every down step is possible, so the chain can go from having n points to having zero points in n steps. Now we calculate the probability of a down move is

$$\frac{1}{2} \cdot \frac{h(y)(m+1)}{h(x)\Lambda(A)} \geq \frac{1}{2M\Lambda(A)}$$

where m is the number of points in y, so long as the right-hand side is less than one, which we can assure by increasing M if necessary (since, if (17) holds, then it also holds if M is increased). Hence the probability of going from n points to zero points in n steps is at least $1/[2M\Lambda(A)]^n$. If we take ν to be the measure having that mass concentrated at zero, then we satisfy the small set condition.

Now we verify the geometric drift condition

$$PV(x) \leq \lambda V(x) + L$$

for some $\lambda < 1$, some constant L, and some function V on the state space having the property that every sublevel set $\{\, x \in \Omega : V(x) \leq \alpha \,\}$ is small, where P is the transition probability kernel of the Markov chain. This proves geometric ergodicity Meyn and Tweedie (2009, Proposition 5.5.3, definition of unbounded off petite sets on p. 189, Theorem 15.0.1, and Lemma 15.2.8).

We choose $V(x) = r^{n(x)}$, where $n(x)$ is the number of points in x and $r > 1$ is a constant to be named later. Clearly V has small level sets. Then if x has n points

$$PV(x) = p_{\text{down}}(x) r^{n-1} + p_{\text{same}}(x) r^n + p_{\text{up}}(x) r^{n+1}$$

where the $p_{\text{down}}(x)$ is the probability the sampler accepts a down step when at x, where the $p_{\text{up}}(x)$ is the probability the sampler accepts an up step when at x, and $p_{\text{same}} = 1 - p_{\text{down}}(x) - p_{\text{up}}(x)$. Choose $r > 1$ and $r > M\Lambda(A)$. Then

$$p_{\text{up}}(x) = \min\left(1, \frac{h(x \cup \xi)\lambda(A)}{h(x)(n+1)}\right) \leq \frac{M\lambda(A)}{n+1} \tag{18}$$

For $0 < \varepsilon < 1$ we have (18) less than ε when $n(x) \geq K_\varepsilon = M\Lambda(A)/\varepsilon$. And

$$p_{\text{down}}(x) = \min\left(1, \frac{nh(x \setminus \xi)}{\lambda(A)h(x)}\right) \geq 1 \tag{19}$$

when $n(x) \geq K_\varepsilon$. Hence being a probability ≥ 1 it must be equal to one. Hence

$$PV(x) \leq \left[\frac{1}{2r} + \frac{1-\varepsilon}{2} + \frac{\varepsilon r}{2}\right] V(x).$$

Since the term in square brackets converges to $(1 + 1/r)/2 < 1$ as $\varepsilon \to 0$ we can choose ε so that the term in square brackets is strictly less than 1; call it λ_ε. Then we have

$$PV(x) \leq \lambda_\varepsilon V(x), \qquad n(x) \geq K_\varepsilon.$$

and we have

$$PV(x) \leq r^{n(x)} \leq r^{K_\varepsilon}, \qquad n(x) < K_\varepsilon;$$

call the right hand side L_ε. Then we have

$$PV(x) \leq \lambda_\varepsilon V(x) + L_\varepsilon, \qquad \text{for all } x,$$

and we are done. □

Now, as an application, we want to show that the Strauss process has bounded conditional intensity. We have

$$\frac{h(x \cup \xi)}{h(x)} = e^{\theta_1 + [t_2(x \cup \xi) - t_2(x)]\theta_2} \leq e^{\theta_1}$$

when $\theta_2 \leq 0$. So the sampler for the Strauss process is geometrically ergodic.

Theorem 7. *If a putative unnormalized density h with respect to a Poisson process having probability measure P has bounded conditional intensity, then its integral with respect to P (6) is finite, so h is in fact an unnormalized density of a spatial point process unless it is zero almost everywhere $[P]$.*

Proof. Suppose (17) holds. If $n(x)$ is the number of points in x, then (17) implies

$$h(x) \leq M^{n(x)} h(\varnothing), \qquad x \in \Omega,$$

and (6) is

$$
\begin{aligned}
\int_\Omega h(x) P(dx) &= e^{-\Lambda(A)} \sum_{n=0}^\infty \frac{1}{n!} \int_{A^n} h(x) \Lambda^n(dx) \\
&\le e^{-\Lambda(A)} \sum_{n=0}^\infty \frac{M^n h(\varnothing) \Lambda(A)^n}{n!} \\
&= h(\varnothing) e^{(M-1)\Lambda(A)}
\end{aligned}
$$

□

Thus in the future, we just want to show that a process has bounded conditional intensity. We do not need a separate proof that the process exists, since that follows from bounded conditional intensity by Theorem 7. Hence if we had proved bounded conditional intensity first, we would not have needed one direction of the proof of Theorem 3. But we still need the other direction, proving when the process does not exist.

10 The Triplets Process

The next step after the Strauss process, going from cliques of size 2 to cliques of size 3, is what Geyer (1999) called the *triplets process.* This is a spatial point process that uses the same neighbor relation as the Strauss process, $\xi \sim \zeta$ if and only if $d(\xi, \zeta) \le \rho$, where d is Euclidean distance and $\rho > 0$ is a fixed constant. For any $x \in \Omega$, let

- $t_1(x)$ denote the number of points in x, which is the number of cliques of size one in x.
- $t_2(x)$ denote the number of unordered pairs of distinct points in x that are neighbors, which is the number of cliques of size two in x
- $t_3(x)$ denote the number of unordered triplets of distinct points in x that are all neighbors, which is the number of cliques of size three in x.

Then the triplets process is the exponential family of distributions having $t_1(x)$, $t_2(x)$, $t_3(x)$ as its canonical statistics, which means the unnormalized density is

$$h_\theta(x) = e^{t_1(x)\theta_1 + t_2(x)\theta_2 + t_3(x)\theta_3} = e^{\langle t(x), \theta \rangle}$$

where in the last expression $t(x)$ and θ are vectors of length 3.

Theorem 8. *The full canonical parameter space for the triplets process is*

$$\Theta = \{\, \theta \in \mathbb{R}^3 : \theta_3 < 0 \text{ or } (\theta_3 = 0 \text{ and } \theta_2 \leq 0) \,\}. \tag{20}$$

The process has bounded conditional intensity for $\theta \in \Theta$.

Since (20) is not an open set, the triplets process (like the Strauss process) is not a regular exponential family. Since (20) is not a closed set either, maximum likelihood for the triplets process is even trickier than for the Strauss process (Geyer, 1999).

Proof. First, we do bounded conditional intensity.

In case $\theta_3 = 0$ the process becomes a Strauss process, and we know from Theorem 3 that the process then exists if and only if $\theta_2 \leq 0$.

Now for the case $\theta_3 < 0$. Fix a point ξ and let S_ξ denote the closed ball centered at ξ having radius ρ. For $\zeta \in S_\xi$, let W_ζ denote the open ball having radius $\rho/2$ centered at ζ. The sets W_ζ, $\zeta \in S_\xi$ are an open cover of S_ξ. Hence, since S_ξ is a compact set, there is a finite subcover W_{ζ_1}, W_{ζ_2}, $\ldots$, W_{ζ_k}. Now define recursively

$$B_1 = S_\xi \cap W_{\zeta_1}$$

and

$$B_j = (S_\xi \cap W_{\zeta_j}) \setminus (B_1 \cup \cdots \cup B_{j-1}), \qquad \text{for } j \geq 2$$

Then B_1, $\ldots$, B_k partition S_ξ, and every pair of points in one of these B_i are neighbors (their separation is less than ρ, because they are contained of some ball of radius $\rho/2$) and every point in one of these B_i is a neighbor of ξ (because it is contained in S_ξ). Let $n_i(x)$ denote the number of points in x that are in B_i.

For any point pattern x not containing ξ

$$t_2(x \cup \xi) - t_2(x) = n_1(x) + \cdots + n_k(x)$$

and

$$t_3(x \cup \xi) - t_3(x) \geq \sum_{i=1}^{k} \frac{n_i(x)[n_i(x) - 1]}{2}$$

hence

$$\log \frac{h_\theta(x \cup \xi)}{h_\theta(x)} \leq \theta_1 + \sum_{i=1}^{k} \left[n_i(x)\theta_2 + \frac{n_i(x)[n_i(x) - 1]\theta_3}{2} \right]$$

The function

$$n \mapsto n\theta_2 + \frac{n(n-1)\theta_3}{2}$$

is quadratic with negative leading coefficient, hence it achieves its maximum and thus is bounded.

Now for the non-existence part. For the case $\theta_3 = 0$ and $\theta_2 > 0$ we know the process does not exist by Theorem 3. For the case $\theta_3 > 0$, we do a proof very similar to that in Theorem 3.

Consider a subset B of A having nonzero Lebesgue measure and diameter less than or equal to ρ. Then $x \in B^n$ implies

$$\begin{aligned} t_2(x) &= n(n-1)/2 \\ t_3(x) &= n(n-1)(n-2)/6 \end{aligned}$$

and

$$\begin{aligned} \int_\Omega h_\theta(x) P(dx) &= e^{-\Lambda(A)} \sum_{n=0}^{\infty} \frac{1}{n!} \int_{A^n} h(x) \Lambda^n(dx) \\ &\ge e^{-\Lambda(A)} \sum_{n=0}^{\infty} \frac{1}{n!} \int_{B^n} h(x) \Lambda^n(dx) \\ &\ge e^{-\Lambda(A)} \Bigg[1 + e^{\theta}_{1} \Lambda(B) \\ &\qquad + \sum_{n=2}^{\infty} \frac{e^{n\theta_1 + n(n-1)\theta_2/2 + n(n-1)(n-2)\theta_3/6} \Lambda(B)^n}{n!} \Bigg] \end{aligned}$$

and, as in the proof of Theorem 3, the terms of the infinite sum go to infinity as $n \to \infty$, hence the sum does not converge. □

Unlike the Strauss process, the triplets process really is a "model for clustering." With slightly negative θ_3 and highly positive θ_2 the process will have much larger expectation of $t_2(x)$ than a Poisson process with the same expectation of $t_1(x)$. Geyer (1999) shows specific examples.

11 The Saturation Process

Inventing non-Poisson Markov spatial point processes having bounded conditional intensity is (IMHO) fairly easy. Here is another one (Geyer, 1999). The motivation for it does not come from Hammersley-Clifford representation. The unnormalized density isn't a function of cliques of certain sizes.

As with the Strauss process and the triplets process, it is an exponential family. As with the Strauss process and the triplets process, the first canonical statistic $t_1(x)$ is the number of points in x. For each point pattern x and each $\xi \in x$, let $m_\xi(x)$ denote the number of neighbors of ξ in x. Then $\sum_{\xi \in x} m_\xi(x)$ would be $2t_2(x)$ for the $t_2(x)$ used for the Strauss and triplets processes. But we don't go there. Let σ be a known constant called the *saturation parameter*, and define

$$t_2(x) = \sum_{\xi \in x} \min\bigl(\sigma, m_\xi(x)\bigr)$$

The story is that $t_2(x)$ only counts neighbors of ξ up to a certain level σ after which ξ has all the neighbors it can handle and any more are irrelevant. Clearly, $t_2(x) \leq \sigma t_1(x)$, and

$$\log \frac{h_\theta(x \cup \xi)}{h_\theta(x)} \leq \theta_1 + \sigma\theta_2$$

so the process has bounded conditional intensities for all values of the parameter, the full canonical parameter space is $\Theta = \mathbb{R}^2$, and this is a regular exponential family.

A Generating Functions

The generating function of the nonnegative-integer-valued random variable X having probability mass function f is

$$\varphi(s) = E\{s^X\} = \sum_{x=0}^{\infty} f(x)s^x = f(0) + \sum_{x=1}^{\infty} f(x)s^x$$

where we write $\varphi(s) = \infty$ if the infinite sum diverges.

Since the terms of the infinite sum are all nonnegative and $s \mapsto s^x$ is strictly increasing for $x > 0$, the generating function is strictly increasing on the interval $\{\, s > 0 : \varphi(s) < \infty \,\}$. Since $\varphi(1) = 1$, we know the radius of convergence of the power series defining the moment generating function is at least one. Since every function defined by a power series with a positive radius of convergence uniquely determines the power series ($f(x) = \varphi^{(x)}(0)/x!$) (Browder, 1996, Corollary 4.37). Hence each moment generating function corresponds to a unique distribution.

References

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, New York.

Browder, A. (1996). *Mathematical Analysis: An Introduction.* New York: Springer-Verlag.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory.* Institute of Mathematical Statistics, Hayward, CA.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*, revised ed. New York: John Wiley.

Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*, 2nd ed., vol I. New York: Springer-Verlag.

Geyer, C. J. (1999). Likelihood inference for spatial point processes. In *Stochastic Geometry: Likelihood and Computation*, W. Kendall, O. Barndorff-Nielsen and M. N. M. van Lieshout, eds. London: Chapman and Hall/CRC, 141–172.

Geyer, Charles J. (2009). Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, **3**, 259–289.

Geyer, C. J. and Møller, J. (1994). Simulation and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, **21**, 359–373.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Kelly, F. P. and Ripley, B. D. (1976). A note on Strauss's model for clustering. *Biometrika*, **63**, 357–360.

Meyn, S. P. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*, second edition. Cambridge: Cambridge University Press.

Ripley, B. D. and Kelly, F. P. (1977). Markov point processes. *Journal of the London Mathematical Society*, **15**, 188–192.

Strauss, D. J. (1975). A model for clustering. *Biometrika*, **62**, 467–475.