Chapter 2

Theory for Spatial Point Patterns

Of the three types of spatial data described in Chapter 1, spatial point patterns have received the least attention with regard to formal inference. Historically, they have been developed within a formal probabilistic framework with modest attention to application though this has changed substantially during the past twenty years. The Bayesian literature is more scant which has motivated the development of this monograph. We utilize this chapter to present a formal development of the basics of spatial point process theory as well as the development of a selection of spatial point pattern models. In the process, we also discuss tools for exploring these point patterns. These tools historically might have been viewed as *inference* but in the 21st century landscape described in Chapter 1, they have to be viewed solely as exploratory data tools. In Chapter 3, we do turn to formal inference, building from a Bayesian perspective but also mentioning some of the more recent classical methods. We also take up model assessment and model comparison. Within the limitations of this monograph-length presentation, it is not possible to do a full elaboration; indeed, there are sources for this. So, instead, we cite connections to potentially helpful literature as we go through the chapter.

2.1. Theory for spatial point patterns

What is a point pattern? For a specified, bounded region D, it is a set of locations, $\mathbf{s}_i, i = 1, 2..., n$. Most importantly, the locations are viewed as "random." We need not have variables observed at the locations, rather we just have the pattern of points. To start, we may look for crude features of the patterns, e.g., behavior that is complete spatial randomness, clustering/attraction, inhibition/repulsion, regular/systematic (we define all of these notions below). Of course, we can add variables at locations; these variables are referred to as "marks." Marks may be discrete, e.g., a label indicating which species was at that location or continuous, e.g., the strength of an earthquake with its epicenter being the location. This leads to modeling challenges taken up in Section 2.6. With discrete marks and seeking a joint model for the locations and the marks, do we model the marks and then a point pattern of locations for each of the marks? Or, do we model the point pattern of locations and then assign a mark to each location? We shall see that these two conditioning choices are not compatible; the joint distributions for the two cases do not reside on the same space. With continuous marks, we may be more concerned with directly specifying the joint distribution.

Let's look at some spatial point patterns. Figure 2.1 shows an example of a point pattern of tree sapling data which we will consider further below. Informally, it is clear that, while we don't know why, the points clearly depart from what we would expect a set dropped *uniformly* over the domain to look like. Figure 2.2 shows examples of point patterns driven by three models discussed below: a homogeneous Poisson process (HPP), a nonhomogeneous Poisson process (NHPP), and

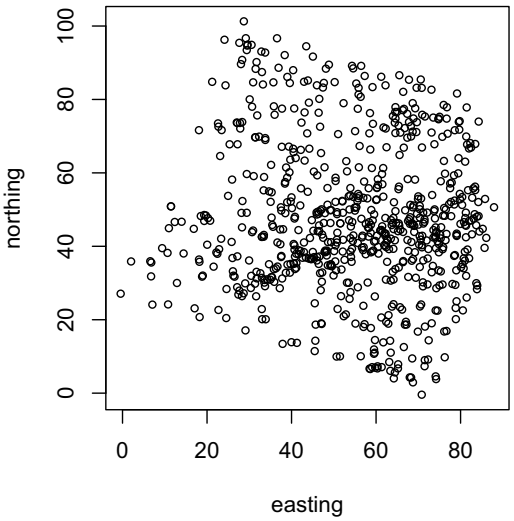

FIG 2.1. *Locations of tree saplings (dbh < 12.7cm) in a forest stand at the Coweeta Hydrologic Laboratory in the southern Appalachians [42, 181].*

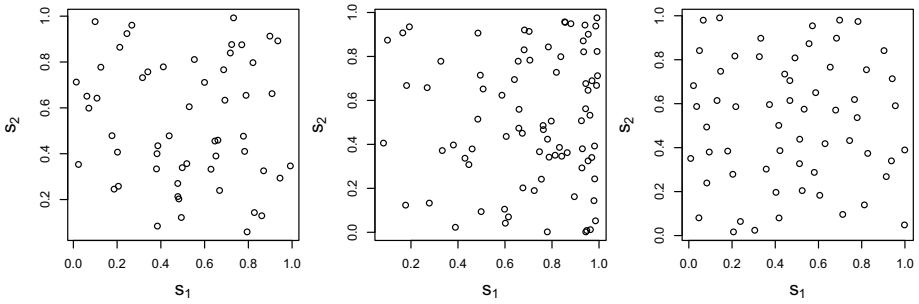

FIG 2.2. *Simulated point patterns from (left) a homogeneous Poisson process (HPP), (middle) a nonhomogeneous Poisson process (NHPP), and (right), a Strauss process over the unit square. The HPP was simulated with intensity* $\lambda = 60$. *The NHPP was simulated with* $\lambda = 8e^{3.6s_1}$. *The Strauss process was simulated with intensity* $\lambda = 170$, *interaction parameter* 0.2, *and interaction radius* 0.075. *Details regarding these processes are supplied in subsequent subsections.*

a Strauss process. These stochastic models provide an explanation of the disparate distributions of points we see. Figure 2.3 shows several realizations under spatial homogeneity, what we refer to as complete spatial randomness. These are realizations from an HPP. Despite the substantial variation in the patterns, all six panels arise from dropping a random number of points uniformly over the unit square. Figure 2.4 shows several realizations under departure from spatial homogeneity. The top three panels arise from a clustering process, the bottom three panels from an inhibition process. Figure 2.5 shows the intensity surface of an NHPP as both a 3-dimensional surface and a contour plot; spatial heterogeneity is evident. Also shown is one realization of the NHPP overlaid on the contours of the intensity. It is evident that areas where the intensity is high tend to produce more points, and areas where the intensity is low tend to produce fewer points. Figure 2.6 shows multiple realizations from the same NHPP, i.e., the same intensity. While there is

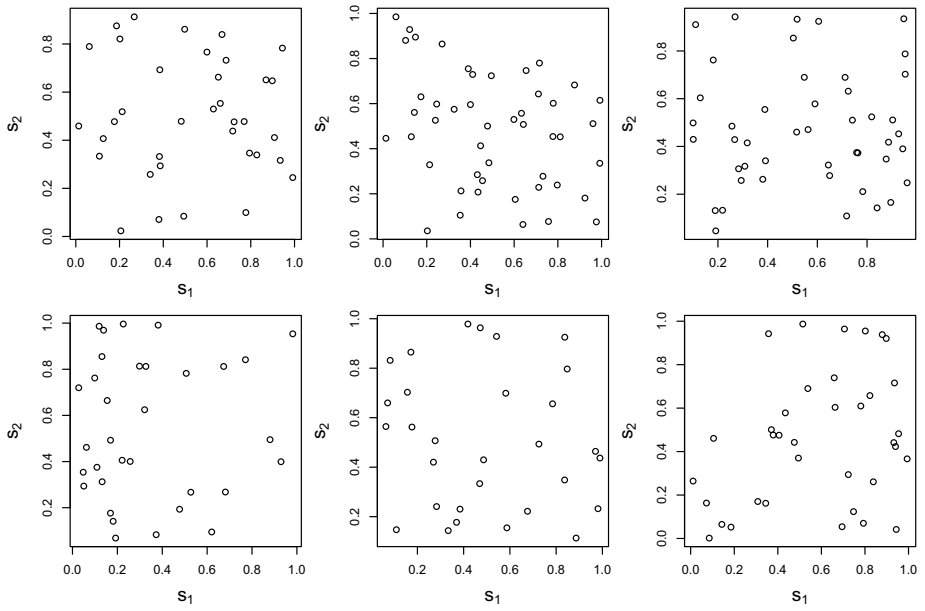

FIG 2.3. *Six random realization from an HPP($\lambda = 40$).*

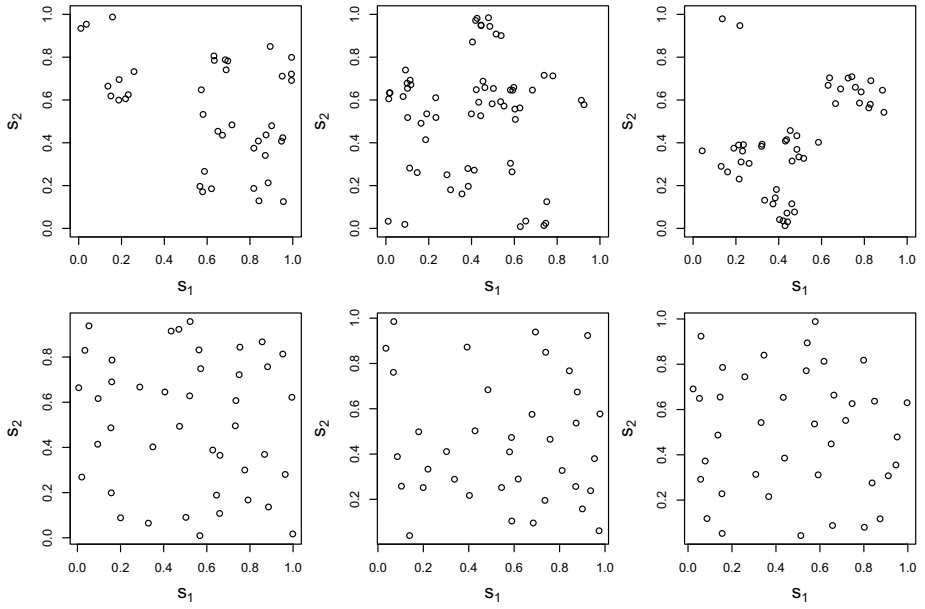

FIG 2.4. *Three random realizations from (top) a Neyman Scott process and (bottom) a Strauss process. The parent process of the Neyman Scott process had intensity $\lambda_p = 10$. The offspring process had intensity $\lambda_o = 5$ and the random number of offspring were located within a 0.1 radius circle of each parent. The Strauss process was simulated with intensity $\lambda = 100$, interaction parameter* 0.1*, and interaction radius* 0.08*. Details regarding these processes are supplied in subsequent subsections.*

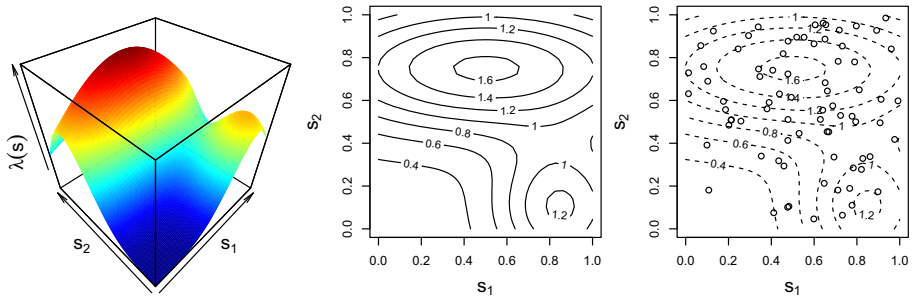

FIG 2.5. *The intensity surface of an NHPP shown as (left) a 3D surface and (middle) a contour plot. (right) One realization from the NHPP.*

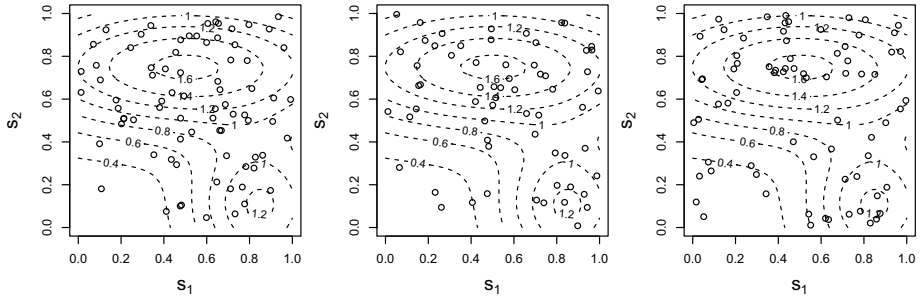

FIG 2.6. *Three realizations from the NHPP specified by the intensity in Figure 2.5.*

substantial variation in the realizations, the effect of the driving intensity is seen in each one. Figure 2.7 returns to the tree sapling data of Figure 2.1 but now provides an estimate of the intensity (a kernel intensity estimate, Section 2.2) which drives the point pattern.

We mention some further examples. In looking at ecological processes, there is interest in the pattern of occurrences of species, e.g., the pattern of trees in a forest, say junipers and pinions. The trees can receive a mark, juniper or pinion. In spatial epidemiology, we seek to find pattern in disease cases, perhaps different patterns for cases vs. controls. An illustration would be the pattern of breast cancer cases. Marks might be the treatment option elected by the woman - mastectomy or radiation. In *syndromic surveillance* we seek to identify disease outbreaks, looking for clustering of cases over time. In studying the evolution or growth of a city over time, i.e., urban development, we could study the pattern of development of single family homes or of commercial property.

It may be useful to note several very active, key players in the spatial point patterns community (hoping not to offend those not noted!). First, we mention Adrian Baddeley whose resume includes impressive theoretical contributions but who has recently turned to more applied effort with, e.g., likelihood-based inference methods, exploratory tools, and residual analysis. He is the driving force behind the `spatstat` package [12], which is an exceptionally useful data analysis R package. Peter Diggle is a researcher whose vision has been consistently ahead of his time. He developed much valuable, lovely early theory. However, he has always had broad spatial interests, always a strong practical bent. He has written an accessible, arguably, classic book [54] in the area and maintains a useful website. Jesper Møller is

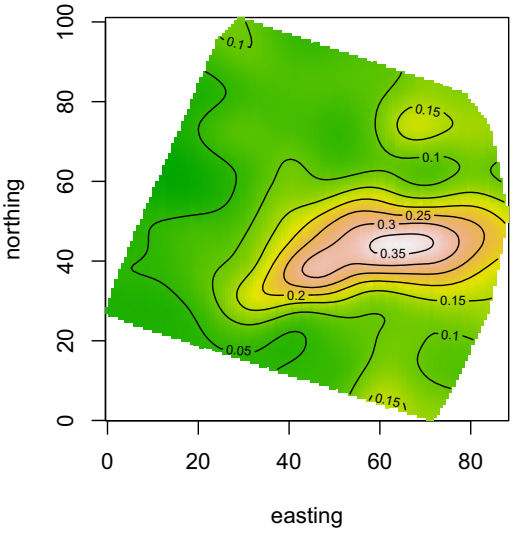

FIG 2.7. *Kernel intensity estimate of the tree sapling data of Figure 2.1.*

an outstanding theoretician who is responsible for developing rich classes of spatial point pattern models. Moreover, to ensure their utility, he has also provided simulation and model fitting approaches for these classes, e.g., algorithms for Markov and Cox processes. Many of his ideas have been gathered into a book with Rasmus Waagepetersen [142]. Lastly here, we mention Janine Illian, a passionate promoter of the role of spatial point pattern modeling in studying ecological processes. The recent book for which she is the lead author, [101] is a broad, richly exemplified, accessible volume.

The contribution of this monograph is modeling and distribution theory for spatial point patterns. Given the goal of model fitting, we focus on inference within a Bayesian framework. As noted earlier, from an inferential perspective, spatial point pattern work is the least developed spatial data area and even more so within the Bayesian framework. Our primary approach for implementing Bayesian inference is the use of simulation to enable full inference with uncertainty, as well as motivating ideas for residual analysis, model adequacy, and model comparison. In a sense, we are merely paralleling what has been the driver of the rapid growth in the use of Bayesian modeling and hierarchical modeling across many data settings and scientific fields. We are merely providing details in the context of point pattern analysis. However, first, we need to develop some theory and detail for spatial point patterns which is the intent of the remainder of this chapter.

2.1.1. The basics

Spatial point processes attempt to describe the randomness associated with a set of spatial locations of the points. For a bounded region D, denote the realization as $\mathbf{s}_i, i = 1, 2..., n$ with **both** n and the $\mathbf{s}_i$ random. A first question we might ask is, "are we seeing a finite realization of an infinite point pattern as a result of imposing D (edge effects and the shape of D might matter), or, are we seeing a finite point pattern associated with a specified D (e.g., an island, a forest, a city)?" Appropriate modeling depends upon the setting. The second case is better suited to application. It is more naturally the way data are collected and enables more flexible modeling.

In particular, it enables inclusion of covariate information across the domain. The first case is usually appropriate for physical systems to capture desired mechanics of the behavior of points in the system. As such, the region is generic and covariates are not part of the story.

Returning to a model for the point pattern, complete spatial randomness (CSR) is a place to start, and, evidently, a model we hope to criticize. Why? Because in applications, it would almost certainly not be operating. In fact, a goal would be to shed light on where there is departure from randomness and what its nature might be. In terms of such departure, it can result from environmental features whence we could use regression models to explain the pattern we have observed. Instead, a model incorporating clustering or attraction, possibly inhibition or repulsion, perhaps regular or systematic behavior, would provide the desired explanation. So it seems we need to clarify what the words "no spatial pattern" mean. What the words, "a *uniform* distribution of points" mean. What is the definition of CSR?

Again, we focus on point patterns over $D \subset \mathbb{R}^2$. We consider a bounded, connected subset D. We denote a random realization of a point pattern by $\mathcal{S}$ with elements $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$. $\mathcal{S}$ is random and so are any features calculated from it. A probabilistic model for $\mathcal{S} \in D$ must place a distribution over all possible realizations in D. This is, evidently, a daunting task. In practice, it is often easier to examine features/functionals of this distribution than to specify the distribution. However, we may consider a generative specification: (i) a distribution over $\{0, 1, 2, ...\}$ to provide the number of points then, (ii) a distribution to *jointly* locate these points over D. It should be clear that only a limited class of models can be generated in this fashion, as we shall elaborate in the sequel.

Following this generative path, we need two ingredients to specify a generative probabilistic model for $\mathcal{S}$. We need a distribution for $N(D)$, the number of points in D which is a distribution over the set $n \in \{0, 1, .., \infty\}$. Then, we need a multivariate *location density* over D^n, for any n, say $f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n)$. Since points are unordered/unlabeled, f must be symmetric in its arguments. With $\partial\mathbf{s}$ denoting a small circular neighborhood around $\mathbf{s}$, $P(N(\partial\mathbf{s}_1) = 1, N(\partial\mathbf{s}_2) = 1, \ldots, N(\partial\mathbf{s}_n) = 1) \approx f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n)\Pi_i|\partial\mathbf{s}_i|$, with $|\partial\mathbf{s}|$ denoting the area of $\partial\mathbf{s}$.

Formally, we need to specify f consistently over all $\mathcal{S}$. Note that the joint distribution has marginal-conditional form $P(N(D) = n)n!f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n)$. The $n!$ appears because the unordered points can be assigned to the n locations in $n!$ ways.

We defined stationarity for a Gaussian process in Section 1.4. Similarly, we can define a stationary point pattern model. We require $f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n) = f(\mathbf{s}_1+\mathbf{h}, \mathbf{s}_2+\mathbf{h}, \ldots, \mathbf{s}_n+\mathbf{h})$ for all n, $\mathbf{s}_i$, and $\mathbf{h}$. This condition would naturally be proposed over $\mathbb{R}^2$ and applied, suitably, over D. Clearly, stationarity is a model property, not a model specification. We will return to stationarity below.

An attractive way to look at a realization of a spatial point process is through a counting measure. Analogous to $N(D)$, suppose we introduce count variables, $N(B)$, i.e., $N(B) = \sum_{\mathbf{s}_i \in \mathcal{S}} 1(\mathbf{s}_i \in B)$. $N(B)$ is computed by looking at the points in $\mathcal{S}$ individually, referred to as a first order property. The observed point pattern provides a realization of a random counting measure over a σ-algebra of sets, say the Borel sets, $\mathcal{B}$. An important point to understand is that a realization of a point pattern is equivalent to a realization of a counting measure (including *void sets*). We won't attempt a rigorous proof of this here but clearly the point pattern provides counts for $B \in \mathcal{B}$. Conversely, through arbitrary unions and intersections of sets in $\mathcal{B}$, we can isolate all of the points in the point pattern, i.e., we can determine the point pattern.

The key idea here is that if we can provide a distributional specification for a random counting measure, we can provide a specification for a random point pattern. That is, we need to provide a recipe for the joint distribution of a finite collection of sets in $\mathcal{B}$. We now turn to this.

The literature for processes over a subset of $\mathbb{R}^1$ is enormous [see, e.g., 117]. This is due to the fact that point processes on the real line are particularly amenable to study because the points are ordered in a natural way, and the whole point process can be described completely by the (random) intervals between the points. Processes over a subset of $\mathbb{R}^2$ sacrifice order, hence history, arguably making them more difficult and more interesting to study.

Historically, the Poisson process is the most well-known and widely-used model for point patterns. With focus on $\mathbb{R}^2$, we recall the definition of a Poisson process over a set D with *intensity* $\lambda(\mathbf{s})$. For any B in $\mathcal{B}$, $N(B) \sim \text{Po}(\lambda(B))$ where $\lambda(B) = \int_B \lambda(\mathbf{s})d\mathbf{s}$. In addition, if B_1 and B_2 are disjoint, then $N(B_1)$ and $N(B_2)$ are independent. The random Poisson measure induced by $\lambda(\mathbf{s})$ can be written as $\lim_{\partial\mathbf{s}\to 0}\frac{N(\partial\mathbf{s})}{|\partial\mathbf{s}|} = N(\mathbf{s})$ or equivalently, $N(B) = \int_B N(\mathbf{s})d\mathbf{s}$.

The independence of disjoint sets implies $f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n) = \Pi_i f(\mathbf{s}_i) = \Pi_i \lambda(\mathbf{s}_i)/\lambda(D)$ where $\lambda(D) = \int_D \lambda(\mathbf{s})d\mathbf{s}$. That is, under a Poisson process, we have a conditionally independent location distribution. Points are assigned to locations independently with the distribution $f(\mathbf{s}) = \lambda(\mathbf{s})/\lambda(D)$. Furthermore, $P(N(\partial\mathbf{s}) = 1) \approx E(N(\partial\mathbf{s})) = \lambda(\partial\mathbf{s}) \approx \lambda(\mathbf{s})|\partial\mathbf{s}|$.

The homogeneous Poisson Process (HPP) specifies $\lambda(\mathbf{s}) = \lambda$ so that $\lambda(B) = \lambda|B|$, i.e., the intensity is proportional to the area. In particular, the location density for any point under an HPP is $1/|D|$. We finally have a definition of complete spatial randomness; CSR is a realization from an HPP(λ). That is, a realization of an HPP arises from the joint location density, $f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n) = 1/|D|^n$. Recalling the definition of stationarity above, it implies that $\lambda(\mathbf{s}) = \lambda$ for all $\mathbf{s}$ and thus, $\lambda(B) = \lambda|B|$ for all $B \subseteq D$. Thus, the HPP is a stationary spatial point process model. However, it is only one stationary spatial process specification. It specifies a constant intensity with conditionally independent locations. More general models include interactions between points, e.g., the stationary Gibbs processes discussed in Section 2.5. Practically speaking, the HPP can be a *null model* for certain types of data such as physical processes in a homogeneous environment, for example, interacting particle models.

In applications, the environment will usually not be homogeneous; the intensity of the process will not be constant, suggesting a nonhomogeneous Poisson process (NHPP) model, i.e., a more general form for $\lambda(\mathbf{s})$. Discussion regarding specification for $\lambda(\mathbf{s})$ is offered below. In any event, as a Poisson process, the NHPP has conditionally independent locations with location density, $f(\mathbf{s}) = \lambda(\mathbf{s})/\lambda(D)$.

2.1.2. A bit more theory

To begin, let us take a look at moment measures. We start with first order properties, i.e., properties of the model for the point pattern which consider the points individually. In particular, the first moment measure is $\{E(N(B)) : B \in \mathcal{B}\}$. Given $\lambda(\mathbf{s})$, we can compute $E(N(B)) = \int_B \lambda(\mathbf{s})d\mathbf{s}$. However, given that the collection, $\{E(N(B)) : B \in \mathcal{B}\}$ is a measure, we can extract the *first-order* intensity: $\lambda(\mathbf{s}) = \lim_{|\partial\mathbf{s}|\to 0}\frac{E(N(\partial\mathbf{s}))}{|\partial\mathbf{s}|}$.

Second order properties refer to properties of the model that consider the points in pairs. For second-order properties, consider $\gamma(B_1 \times B_2) \equiv E_{\mathcal{S}}\sum_{\mathbf{s},\mathbf{s}'\in\mathcal{S}} 1(\mathbf{s} \in$

$B_1, \mathbf{s}' \in B_2$). Define $\gamma(\mathbf{s}, \mathbf{s}')$, the *second order intensity* through $\gamma(B_1 \times B_2) = \int_{B_1}\int_{B_2}\gamma(\mathbf{s},\mathbf{s}')d\mathbf{s}'d\mathbf{s}$. As a result, if B_1, B_2 disjoint, $E_{\mathcal{S}}(N(B_1)N(B_2)) = \int_{B_1}\int_{B_2}\gamma(\mathbf{s},\mathbf{s}')d\mathbf{s}'d\mathbf{s}$. Hence, with sufficiently small sets, $\gamma(\mathbf{s},\mathbf{s}') = \lim_{|\partial\mathbf{s}|\to 0, |\partial\mathbf{s}'|\to 0} \frac{E(N(\partial\mathbf{s})N(\partial\mathbf{s}'))}{|\partial\mathbf{s}||\partial\mathbf{s}'|}$.

The *pair correlation function*, $g(\mathbf{s}, \mathbf{s}')$, is defined as $\gamma(\mathbf{s},\mathbf{s}')/\lambda(\mathbf{s})\lambda(\mathbf{s}')$. When $\lambda(\mathbf{s}) = \lambda$, $g(\mathbf{s},\mathbf{s}')$ simplifies to $\gamma(\mathbf{s},\mathbf{s}')/\lambda^2$. In fact, $g(\mathbf{s},\mathbf{s}') = 1$ under CSR. Furthermore, when $g(\mathbf{s},\mathbf{s}') > 1$, attraction is implied, while $g(\mathbf{s},\mathbf{s}') < 1$ implies repulsion. Under stationarity, $\gamma(\mathbf{s},\mathbf{s}') = \gamma(\mathbf{s}-\mathbf{s}')$. In this context, isotropy means $\gamma(\mathbf{s},\mathbf{s}') = \gamma(||\mathbf{s}-\mathbf{s}'||)$.

Further insight can be obtained by noticing that, if $\partial\mathbf{s}$ is sufficiently small, $P(N(\partial\mathbf{s}) > 1)$ will be negligible so $E(N(\partial\mathbf{s})) \approx P(N(\partial\mathbf{s}) = 1) \approx P(N(\partial\mathbf{s}) > 0)$. Similarly, $E(N(\partial\mathbf{s})N(\partial\mathbf{s}')) \approx P(N(\partial\mathbf{s}) > 0, N(\partial\mathbf{s}') > 0)$. But, since $E(N(\partial\mathbf{s})N(\partial\mathbf{s}')) \approx \gamma(\mathbf{s},\mathbf{s}')|\partial\mathbf{s}||\partial\mathbf{s}'|$, we find that $\gamma(\mathbf{s},\mathbf{s}') \approx P(N(\partial\mathbf{s}) > 0, N(\partial\mathbf{s}') > 0)/|\partial\mathbf{s}||\partial\mathbf{s}'|$. That is, $\gamma(\mathbf{s},\mathbf{s}')|\partial\mathbf{s}||\partial\mathbf{s}'|$ is approximately the probability of a point of $\mathcal{S}$ in $\partial\mathbf{s}$ and a point of $\mathcal{S}$ in $\partial\mathbf{s}'$, with an evident limiting interpretation as an intensity. For instance, if γ is isotropic, $\gamma(||\mathbf{s}-\mathbf{s}'||)$ can be interpreted loosely as the density for inter-point distances.

One further feature we need to define is the Papangelou conditional intensity. We do not attempt a rigorous development. Rather, informally, let's define $\lambda(\mathbf{s}|\mathcal{S})$ for a given location $\mathbf{s}$ and a given realization $\mathcal{S}$. $\lambda(\partial\mathbf{s})|\mathcal{S}) \approx \lambda(\mathbf{s}|\mathcal{S})d\mathbf{s}$ is interpreted as the conditional probability that there is a point of the process in $\partial\mathbf{s}$ and the rest of the process coincides with $\mathcal{S}$. Roughly, $\lambda(\partial\mathbf{s}|\mathcal{S})$ is the probability that there is a point of $\mathcal{S}$ in $\partial\mathbf{s}$ and the rest of $\mathcal{S}$ lies outside of $\partial\mathbf{s}$. In other words, $\lambda(\mathbf{s}|\mathcal{S}) = \lambda(\mathbf{s}|\mathcal{S}/\mathbf{s}), \mathbf{s} \in \mathcal{S}$; $= \lambda(\mathbf{s}|\mathcal{S}), \mathbf{s}$ not $\in \mathcal{S}$. Clearly, $\lambda(\mathbf{s}|\mathcal{S})$ is random since $\mathcal{S}$ is and $E_{\mathcal{S}}(\lambda(\mathbf{s}|\mathcal{S})) = \lambda(\mathbf{s})$.

Formally, this suggests that, with n random, we view

$$\lambda(\partial\mathbf{s}|\mathcal{S}) = \int_{\partial\mathbf{s}} \frac{f(\mathbf{u},\mathbf{s}_1,\mathbf{s}_2,\ldots,\mathbf{s}_n)}{f(\mathbf{s}_1,\mathbf{s}_2,\ldots,\mathbf{s}_n)}d\mathbf{u} \approx \frac{f(\mathbf{s},\mathbf{s}_1,\mathbf{s}_2,\ldots,\mathbf{s}_n)}{f(\mathbf{s}_1,\mathbf{s}_2,\ldots,\mathbf{s}_n)}|\partial\mathbf{s}|. \tag{2.1}$$

Passing to the limit, $\lambda(\mathbf{s}|\mathcal{S}) = \frac{f(\mathbf{s},\mathcal{S})}{f(\mathcal{S})}$ where $f(\mathcal{S})$ is the *density* of the spatial point process (with respect to an HPP(1)). To attempt to clarify why we have λ on the left side with location distributions on the right side, we note that, in the integration, $f(\mathcal{S})$ is a random function that is not of fixed dimension; it is specified through the intensity up to normalizing constant which cancels from the ratio in the integral. Lastly, for conditionally independent locations $\lambda(\mathbf{s}|\mathcal{S}) = \lambda(\mathbf{s})$.

2.2. Exploratory tools

Again, complete spatial randomness (CSR) $\equiv$ HPP(λ) provides a baseline for point patterns which we do not expect to be the truth. So, here, the goal is simply to offer exploratory tools which can criticize CSR. The approaches are distance-based; they look at pairwise distances between points to see if their behavior is in disagreement with what we expect under CSR. These distance-based approaches are the G, F, and K functions.

We start with $G(d)$, the "nearest neighbor" distribution, i.e., the c.d.f. of the nearest neighbor distance, which is event to event. By definition, $G(d) = Pr(\texttt{nearest event} \leq d)$. By analogy, $F(d)$ is the "empty space" distribution, i.e., for an arbitrary location, the c.d.f. of the nearest neighbor distance, now point to event. So, $F(d) = Pr(\texttt{nearest event} \leq d)$. Under CSR, it is clear that

$G(d) = F(d) = 1 - \exp(-\lambda\pi d^2)$. Evidently, the distribution of πd^2 is an exponential so that G places a lot of mass on small distances. We expect to see some clustering under CSR. It is important to appreciate that these conceptual quantities are only sensible for stationary processes. For nonstationary processes, we will be focused on estimating the associated first and, possibly, second order characteristics.

Let us be a bit more formal and view the process over all of $\mathbb{R}^2$, i.e., an infinite point pattern which becomes finite under restriction to D. Consider the random variable $N(\mathbf{s}, d; \mathcal{S})$, where $\mathbf{s} \in \mathcal{S}$, $\partial_d\mathbf{s}$ is a circle of radius d centered at $\mathbf{s}$, and N counts the number of points in the circle from $\mathcal{S}$, excluding $\mathbf{s}$. By stationarity, $N(\mathbf{s}, d; \mathcal{S}) \sim N(\mathbf{0}, d; \mathcal{S} - \mathbf{s})$, where $\mathcal{S} - \mathbf{s}$ is the translation of $\mathcal{S}$ by $\mathbf{s}$. This distributional result is equivalent to saying that every point in $\mathcal{S}$ is a *typical* point, in the sense that each one can be viewed as equivalent to $\mathbf{0}$ under translation.

Under restriction to a bounded set D, consider

$$E_{\mathcal{S}}\left(\sum_{\mathbf{s}_i \in \mathcal{S}, \mathcal{S} \in D} 1(N(\mathbf{s}_i, d; \mathcal{S}) > 0)\right) = \lambda|D|P(N_D(\mathbf{s}, d; \mathcal{S}) > 0). \tag{2.2}$$

The right side of (2.2) follows by noting that the expectation over $\mathcal{S}$ can be calculated in two stages, over $\mathcal{S}$ given $N(D)$ and then over $N(D)$. Here, $N_D(\mathbf{s}, d; \mathcal{S})$ is the count under restriction of the random $\mathcal{S}$ to D. Using the left side of (2.2), we see an obvious Monte Carlo integration for it. Moreover, it is clear that this integration arises in two stages. First, $\mathcal{S}$ is sampled, then N is calculated, given $\mathcal{S}$. Again, note that to obtain (2.2) requires restriction to a bounded set D and enables a Monte Carlo integration for $P(N_D(\mathbf{s}, d; \mathcal{S}) > 0)$, not for $P(N(\mathbf{s}, d; \mathcal{S}) > 0)$ $(\geq P(N_D(\mathbf{s}, d; \mathcal{S}) > 0))$. Empirical estimation of the latter requires an *edge correction* (see below). From the definition above, $G(d) \equiv P(N(\mathbf{s}, d; \mathcal{S}) > 0)$. That is, this probability does not depend upon $\mathbf{s}$, consistent with the notion of a typical point. We see that $G(d)$ increases in d and, in fact, can be viewed as a cdf in distance d.

With regard to estimation of $G(d)$, the empirical c.d.f., $\hat{G}(d)$, arises from the n nearest neighbor distances (for $\mathbf{s}_1$, for $\mathbf{s}_2$, etc.). Denote this set by $\{d_1, d_2, \ldots, d_n\}$. Again, with bounded D, we will need an edge correction. For example, if, for $\mathbf{s}_i$, $d > b_i$, where b_i is the *distance* from $\mathbf{s}_i$ to edge of D, then the event $\{d_i < d\}$ is not observed. Therefore, we set $\hat{G}(d) = \frac{\sum_i I(d_i \leq d < b_i)}{\sum_i I(d < b_i)}$.

Comparison of $\hat{G}$ with G under CSR is usually through a theoretical Q-Q plot. Shorter tails suggest clustering/attraction, i.e., nearest neighbor distances are shorter than expected. Longer tails suggest inhibition/repulsion, i.e., nearest neighbor distances are longer than expected. It is important to note that, technically, $\hat{G}(d)$ is not exactly an empirical c.d.f. since the d_i's are not independent. However, with EDA intentions, perhaps this issue can be ignored.

An alternative to $G(d)$ in the literature is $F(d)$ where now $N(\mathbf{s}, d; \mathcal{S})$ would assume $\mathbf{s}$ is not in $\mathcal{S}$. We might distinguish these two definitions of event N by subscript, say N_G and N_F. More importantly, $G(d)$ need not equal $F(d)$. For instance, inhibition might preclude two points in $\mathcal{S}$ from being within distance d of each other. Also, the empirical c.d.f. for F is different from that for G since the number of "points" is arbitrary. That is, $\hat{F}(d)$ is the empirical c.d.f. arising from the m nearest neighbor distances associated with a randomly selected set of m points in D. Evidently, m and the associated locations are arbitrary, diminishing the value of this diagnostic. In any event, $\hat{G} \neq \hat{F}$ and we may be interested in looking at the differences for various choices of the m points. A potentially useful function is

the J function, $J(d) = \frac{1-G(d)}{1-F(d)}$. This function avoids comparison with CSR (though it equals 1 in that case), rather bringing the interpretation of clustering for $J(d) < 1$ and inhibition for $J(d) > 1$. $\hat{J}(d) = \frac{1-\hat{G}(d)}{1-\hat{F}(d)}$ is the customary estimate of $J(d)$.

Next, the K function considers the *expected number* of points within distance d of an arbitrary point. Under stationarity, this expectation is the same for any point. Explicitly, $K(d) = (\lambda)^{-1}\mathrm{E}(\#$ of points within d of an arbitrary point). The scaling by $1/\lambda$, along with stationarity, scales $K(d)$ to be free of λ. For example, under CSR, $K(d) = \lambda\pi d^2/\lambda = \pi d^2$, i.e., the area of a circle of radius d.

Again, for a stationary process, being more formal, now consider $E(N(\mathbf{s}, d; \mathcal{S}))$ which is the expected number of points in $\partial_d\mathbf{s}$, a circle of radius d, centered at $\mathbf{s}$, when $\mathbf{s} \in \mathcal{S}$ but not including $\mathbf{s}$. Using the foregoing notation and a similar calculation, we have

$$E_{\mathcal{S}}\left(\sum_{\mathbf{s}_i \in \mathcal{S}, \mathcal{S} \in D} N(\mathbf{s}_i, d; \mathcal{S})\right) = \lambda|D|E(N_D(\mathbf{s}, d; \mathcal{S})). \tag{2.3}$$

We have formalized an expectation of interest. That is, with respect to the set D, the right side is the expected number of points from a random $\mathcal{S}$ within distance d of a point in $\mathcal{S}$. The left side of the equality motivates a natural Monte Carlo integration and empirical estimation of $E(N(\mathbf{s}, d; \mathcal{S})) \geq E(N_D(\mathbf{s}, d; \mathcal{S}))$ requires edge correction (see below). Again, $E(N(\mathbf{s}, d; \mathcal{S})$ does not depend on $\mathbf{s}$. And from above, we have $\lambda K(d) \equiv E(N(\mathbf{s}, d; \mathcal{S})$.

An empirical estimator of $K(d)$ becomes $\hat{K}(d) = (\hat{\lambda})^{-1}\sum_i \sum_{j\neq i} 1(d_{ij} \equiv ||\mathbf{s}_i - \mathbf{s}_j|| \leq d)/n = (n\hat{\lambda})^{-1}\sum_i r_i$ where $\hat{\lambda} = n/|D|$ and r_i is number of $\mathbf{s}_j$ within d of $\mathbf{s}_i$. Again, we need edge correction, w_{ij}, for $\mathbf{s}_i$ too near the boundary of D. Explicitly, w_{ij} is the conditional probability that an event is in D given that it is exactly distance d_{ij} from $\mathbf{s}_i$. It is *approximated* as the proportion of the circumference of a circle centered at $\mathbf{s}_i$ with radius $\|\mathbf{s}_i - \mathbf{s}_j\|$ that lies within D.

As with G and F, we compare $\hat{K}(d)$ with $K(d) = \pi d^2$. Regularity/inhibition implies $\hat{K}(d) < \pi d^2$; clustering implies $\hat{K}(d) > \pi d^2$. A plot which has been proposed in the literature [see, e.g., 47] is $L(d)$ vs. d where $L(d) = \sqrt{\left(\frac{\hat{K}(d)}{\pi} - d\right)}$. Evidently, $L(d) = 0$ under CSR, suggesting that we look for peaks and valleys in the plot. For instance, a peak at distance d would suggest clustering at that distance.

We can connect $K(d)$ to the pair correlation function from the previous section. In fact, for a stationary process, we can define $K(d) = \int_{||\mathbf{u}||\leq d} g(\mathbf{u})d\mathbf{u}$, with g the pair correlation function. As a result, the second moment measure $\gamma(d) = \frac{\lambda^2 K'(d)}{2\pi d}$. This suggests the possibility of creating a $\hat{\gamma}(d)$ through a smoothed version of $\hat{K}(d)$.

Finally, for nonstationary processes, moving away from CSR leads to interest in estimating the first order intensity, $\lambda(\mathbf{s})$. In the absence of a model for $\lambda(\mathbf{s})$, we consider empirical estimates. The first is the analogue of a histogram, the second of a kernel density estimate.

Imagine a refined grid over D. Then, as above, $\lambda(\partial\mathbf{s}) = \int_{\partial\mathbf{s}} \lambda(\mathbf{s})d\mathbf{s} \approx \lambda(\mathbf{s})|\partial\mathbf{s}|$. So, for grid cell A_l, assume $\lambda(\mathbf{s})$ is constant over A_l. Then, the natural estimate is $N(A_l)/|A_l|$. Evidently, a picture of this estimate will reveal a two dimensional step surface which we might call a tile surface. It's appearance will resemble a two-dimensional histogram but the area under the surface will be the number of points in the pattern.

Kernel density estimates are widely used, providing a smoothing of a histogram [185]. In the same spirit, a kernel intensity estimate takes the form

$$\hat{\lambda}_\tau(\mathbf{s}) = \sum_i h(||\mathbf{s} - \mathbf{s}_i||/\tau)/\tau^2, \mathbf{s} \in D. \tag{2.4}$$

In (2.4), h is a radially symmetric bivariate pdf (usually a bivariate normal) while τ is a "bandwidth" which controls the smoothness of $\hat{\lambda}_\tau(\mathbf{s})$. The power τ^2 reflects the fact that the scaling is done in $\mathbb{R}^2$. Finally, note that we don't divide by n, as with kernel density estimates. The reason is that we *cumulate* intensity whereas we *normalize* a density.

2.3. Modeling $\boldsymbol{\lambda(s)}$

Next, we turn to modeling the intensity, $\lambda(\mathbf{s})$. We can think of attaching such specification to a nonhomogeneous Poisson process (NHPP), again with conditionally independent locations assigned using location density, $f(\mathbf{s}) = \lambda(\mathbf{s})/\lambda(D)$. Possibilities include:

(i) a scaling form: $\lambda(\mathbf{s};\theta) = \lambda f(\mathbf{s};\theta)$ where f a bivariate density function truncated to D. This specification provides the nice interpretation of λ as the expected number of points in the pattern. Sufficiently rich choices for f would require mixture models, e.g., $f(\mathbf{s}) = \sum_{k=1}^{K} p_k f_k(\mathbf{s})$. However, these mixtures are difficult to identify (recall Section 1.3). Furthermore, it is awkward to specify bivariate densities rich enough to provide rich enough $\lambda(\mathbf{s})$ surfaces. Additionally, the normalization of f to an irregularly shaped D can be computationally expensive.

(ii) One might imagine a trend surface to capture $\lambda(\mathbf{s})$, i.e., a function of location. To ensure positivity, such surfaces must be specified on the log scale. Simple polynomial surfaces will not be flexible enough. We have the familiar "tail wagging the dog" phenomenon, that is, with a polynomial, telling you what it looks like at a few places in D tells you what is must look like elsewhere in D. This suggests the possibility of a spline surface specification [51] as available in the `spatstat` package [12]. Here, we avoid spline specifications in favor of a Gaussian process choice presented below.

(iii) The most common form is $\log \lambda(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\gamma}$; that is, spatially-referenced covariates are assumed to drive the point pattern. This specification is attractive from an explanatory perspective and may be adequate in certain applications. However, often the covariates are not able to provide an adequate story.

We note that, in order to calculate the likelihood, we need to calculate $\int_D e^{\mathbf{x}^T(\mathbf{s})\boldsymbol{\gamma}} d\mathbf{s}$ (see below). This raises the question of how the covariate surfaces are supplied, i.e., what does $\mathbf{x}(\mathbf{s})$ look like? In practice, it is usually a tiled surface. This opens the door to the ecological fallacy [209, 210] with regard to inference. This fallacy notes that covariate levels associated with areal units need not match well point level values for the covariates. In turn, this will affect the performance of the estimated intensity. However, without finer covariate resolution, we can not do better.

To fully enrich the $\lambda(\mathbf{s})$ surface, we turn to the Gaussian process developed in Section 1.4. This leads to what is referred to as the log Gaussian Cox process (LGCP). We write $\lambda(\mathbf{s}) = g(\mathbf{x}(\mathbf{s})^T\boldsymbol{\gamma})\lambda_0(\mathbf{s})$. We require $g(\cdot) \geq 0$ and think of $\lambda_0(\mathbf{s})$ as the local adjustment process, a realization of a positive stochastic process which pushes up or pulls down the intensity appropriately at each location in D. In particular, if $z(\mathbf{s}) \equiv \log\lambda_0(\mathbf{s})$ is a Gaussian process with covariance function $\sigma^2\rho(\cdot)$,

then $\lambda_0(\mathbf{s})$ is a log Gaussian process and $\lambda(\mathbf{s})$ provides a LGCP. The natural center for $\lambda_0(\mathbf{s})$ is 1 suggesting that we center $z(\mathbf{s})$ at $-\sigma^2/2$. Note that, conditional on $\{\lambda_0(\mathbf{s}), \mathbf{s} \in D\}$ (and $\boldsymbol{\gamma}$), we have an NHPP. In this regard, we can consider the LGCP as a two stage process, $[\mathcal{S}|\lambda(\mathbf{s})][\lambda(\mathbf{s})]$.

The likelihood

For an NHPP or a LGCP, what is the likelihood? As a function of $\lambda(\mathbf{s})$, we now argue that

$$L(\{\lambda(\mathbf{s}), \mathbf{s} \in D\}; \mathcal{S}_{obs}) = e^{-\lambda(D)} \Pi_i \lambda(\mathbf{s}_i). \tag{2.5}$$

Note that the likelihood is a function of the entire intensity surface. For the NHPP model it will be a parametric function, for the LGCP, it will be a process realization. Therefore, $\lambda(D) = \int_D \lambda(\mathbf{s}) d\mathbf{s}$ is either a regular or a stochastic integral. We will not develop stochastic integrals here. References include [116, 135]. We only emphasize the fact that, rather than a customary integral of a function, we have an integral over a realization of a stochastic process, hence a random variable. Such integrals can be challenging to define and, in any event are never available explicitly. Integral approximation is required (see Section 4.2 below) and care must be taken to ensure that the approximation, as a random variable, converges to the actual integral, as a random variable.

Returning to the likelihood, we can develop it in two ways. First, given $N(D) = n$, the location density is

$$f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n | N(D) = n) = \prod_i \frac{\lambda(\mathbf{s}_i)}{(\lambda(D))^n} ,$$

where, again $\lambda(D) = \int_D \lambda(\mathbf{s}) d\mathbf{s}$. So, the "joint density','

$$f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n, N(D) = n) = \prod_i \frac{\lambda(s_i)}{(\lambda(D))^n} \times (\lambda(D))^n \frac{\exp(-\lambda(D))}{n!} ,$$

providing the likelihood.

Alternatively, if we partition D into a fine grid, the Poisson assumption implies that the likelihood will be a product over the grid cells. That is, $\prod_l \exp(-\lambda(A_l))(\lambda(A_l))^{N(A_l)}$. Regardless of the grid, the product of the exponential terms is $\exp(-\lambda(D))$. Moreover, as the grid becomes finer, $N(A_l) = 1$ or 0 according to whether there is a $\mathbf{s}_i$ in A_l or not. In the limit, we obtain (2.5).

It is important to note that the likelihood is a function of the entire intensity surface. We will have an uncountable dimensional model unless we provide a parametric form for $\lambda(\mathbf{s})$.

2.4. General Cox processes

We begin with the Neyman Scott process. Suppose we generate K *parent* events from an NHPP with $\lambda(\mathbf{s})$ with locations, say, $\boldsymbol{\mu}_k, k = 1, 2, \ldots, K$. Next, suppose each parent produces a random number of offspring, N_k, where the N_k are i.i.d. according say, $g = \texttt{Po}(\delta)$. We locate the offspring relative to the parent. For the kth parent, we locate the offspring according to i.i.d. draws from a bivariate density, $f(\mathbf{s}; \boldsymbol{\mu}_k)$, centered at $\boldsymbol{\mu}_k$. Finally, only the offspring are retained to yield the

point pattern. Evidently, such a model will yield clustering in the point pattern realizations, i.e., clusters associated with each deleted parent. If the bivariate density is $N(\boldsymbol{\mu}_k, \sigma^2 I)$, we have a (modified) Thomas process. If the offspring at $\boldsymbol{\mu}_k$ are distributed uniformly in a circle of radius R (a parameter) around $\boldsymbol{\mu}_k$, we have a Matérn process. If we have a degenerate offspring density at $\boldsymbol{\mu}_k$, then the offspring count at $\boldsymbol{\mu}_k$ is viewed as a 'mark' at that location. More generally, suppose we combine the steps of generating the number of children and their locations. That is, generate $N \sim g_K$ and generate $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N$ i.i.d $\sim \sum_{k=1}^{K} \frac{1}{K} f(\mathbf{s}; \boldsymbol{\mu}_k, \Sigma)$. For example, with the above, we might take $g_K = Po(K\lambda)$.

An alternative class of spatial clustering models are shot noise processes. These are also Cox processes and, conditionally, NHPPs. Shot noise processes provide a random realization of an intensity through a model which offers an alternative to a LGCP. Again, $\lambda(\mathbf{s}) = e^{X^T(\mathbf{s})\beta}\lambda_0(\mathbf{s})$. However, now $\lambda_0(\mathbf{s})$ is a mean 1 shot noise process where $\lambda(\mathbf{s})$ is *centered* around the deterministic component.

A simple form for a realization of a shot noise process is $\lambda_0(\mathbf{s}) = \sum_{\mathbf{s}_i \in \mathcal{S}} f(\mathbf{s} - \mathbf{s}_i) m(\mathbf{s}_i)$. Here, $\mathcal{S}$, is drawn from an HPP(λ), f is a unimodal density over D centered at 0, and $m(\mathbf{s}_i) \geq 0$. $m(\mathbf{s}_i)$ is referred to as the "shot" at $\mathbf{s}_i$ and is often taken to be a fixed constant, m. The density, f, spreads the influence of the shot at $\mathbf{s}_i$ on $\lambda(\mathbf{s})$ according to the distance that $\mathbf{s}$ is from $\mathbf{s}_i$. As a result, $\lambda_0(\mathbf{s})$ accumulates the shots arising from $\mathcal{S}$. When $m(\mathbf{s}) = m$, a constant, we can calculate $E(\lambda_0(\mathbf{s})) = m\lambda = 1$. So, to center at 1, $m = 1/\lambda$ (and, then, of course, $E(\lambda_0(D)) = |D|$).

Again, $\lambda_0(\mathbf{s})$ is random because the realization, $\mathcal{S}$ from the HPP is random. Moreover, the shot noise process intensity will tend to encourage clustering because its intensity will tend to exhibit peaks at the $\mathbf{s}_i$. We may view $\lambda_0(\mathbf{s})$ as the intensity associated with a marked point process (see Section 2.6 below) in the sense that we can view $f(\mathbf{s} - \mathbf{s}_i)m(\mathbf{s}_i)$ as an intensity associated with the i-th mark. The shot noise intensity sums these individual mark intensities to give the overall intensity.

One further intensity which provides clustering is the Poisson-Gamma process [217]. In fact, the Poisson-Gamma process is an example of a shot noise process. It allows both over and under-dispersion relative to an HPP. In general, a gamma process provides a random positive spatial surface, i.e., the generator is $\Gamma(d\mathbf{u}) \sim Ga(\alpha(d\mathbf{u}), \beta^{-1})$ so that $\int_A \Gamma(d\mathbf{u}) = \Gamma(A)$. We can use kernel mixing to obtain the random intensity $\lambda(\partial\mathbf{s}) \approx \lambda(\mathbf{s})|\partial\mathbf{s}| = \int_D f(\mathbf{s} - \mathbf{u})\Gamma(d\mathbf{u})|\partial\mathbf{s}|$. Again, we draw a realization of an HPP over D to obtain $\mathcal{S}^* = \{s_j^*, j = 1, 2, \ldots, m\}$. Finally, we simplify the intensity by discretizing D to $\mathcal{S}^*$ yielding $\lambda(\mathbf{s}) = \sum_{s_j^* \in \mathcal{S}^*} f(\mathbf{s} - \mathbf{s}_j^*) w(\mathbf{s}_j^*)$, $w(\mathbf{s}_j^*)$, a Gamma variable. The shot noise form is clear.

2.5. Inhibition processes

Next we describe some so-called Markov or Gibbs processes which, for us, play the role of inhibition processes. A finite Gibbs process is defined by location density of the form $f(\mathcal{S}) = \exp(-Q(\mathcal{S}))$ with regard to an HPP with unit intensity. That is, $f(\mathcal{S})$ is essentially the *Radon-Nikodym* derivative of a Gibbs process with respect to an HPP(1). Here, $Q(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n) = c_0 + \sum_{i=1}^{n} h_1(\mathbf{s}_i) + \sum_{i \neq j} h_2(\mathbf{s}_i, \mathbf{s}_j) + \ldots + h_n(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n)$. The h's are parametric functions, symmetric in their arguments, and c_0 is a *normalizing* constant over D^n, a function of the parameters in the h's. c_0 is almost always intractable, making $E(N(D))$ intractable. The h's are potentials of order 1, 2,...n, respectively, again, each symmetric in its arguments. Connections to Markov random field models [16] are immediate but are not discussed further here.

With potentials only of order 1, we obtain an NHPP with $\lambda(\mathbf{s}) = e^{-h_1(\mathbf{s})}$. Evidently, we want higher order potentials in order to capture/control interaction. In practice, there seems no reason to go beyond pairwise interactions so we only include h_1 and h_2 in Q. To guarantee integrability, we must take $h_2 \geq 0$. This implies we can only capture inhibition. To see this, if we require $Q(\mathbf{s}_1, \mathbf{s}_2) \geq c_0 + h_1(\mathbf{s}_1) + h_1(\mathbf{s}_2)$, this means for pairs of points at a given distance, $f(\mathbf{s}_1, \mathbf{s}_2)$ puts less mass under the Gibbs specification than with the corresponding NHPP; we encourage inhibition. If $h_1(\mathbf{s})$ is constant, we have a *homogeneous* Gibbs process.

The most common form for h_2 is $\phi(||\mathbf{s} - \mathbf{s}'||)$, e.g., $||\mathbf{s} - \mathbf{s}'||^2/\tau^2$. Furthermore, the Papangelou conditional intensity has a simple form in this case, $\lambda(\mathbf{s}|\mathcal{S}) = \exp(-(h_1(\mathbf{s}) + \sum_{i=1}^{n} \phi(\|\mathbf{s} - \mathbf{s}_i\|)))$. Conveniently, the unknown normalizing constant cancels from the conditional intensity.

Two examples for $\phi(d)$, where d is an inter-point distance, are the Strauss process and the hardcore process. The Strauss process sets $\phi(d) = \beta, d \leq d_0, = 0, d > d_0$. With $\beta > 0$, we have $e^{-\phi(d)} \leq 1$ for all d. Therefore, the interaction term down-weights patterns with more points close to each other. The hardcore process, as its name suggests, is much stronger. It sets $\phi(d) = \infty, d \leq d_0, = 0, d > d_0$. Now, the density is 0 for all $\mathcal{S}$ with a pair of points less than d_0 apart. Lastly, a recent richer class of Gibbs processes is discussed in [86].

We conclude with a very different class of inhibition processes, the determinantal point processes [119, 184]. These processes are elegant in specification but are very demanding to simulate as well as for model fitting. We offer a brief description.

Suppose, for a finite spatial point process $\mathcal{S}$ on $D \subset \mathbb{R}^2$,

$$f^{(n)}(\mathbf{s}_1, \ldots, \mathbf{s}_n) = \det\{[C](\mathbf{s}_1, \ldots, \mathbf{s}_n)\}, \quad (\mathbf{s}_1, \ldots, \mathbf{s}_n), \quad n = 1, 2, \ldots \tag{2.6}$$

is the n-th order product density function, i.e., the joint location density for a realization of n points. Here, $C(\mathbf{s}_i, \mathbf{s}_j)$ is a covariance function for locations $\mathbf{s}_i$ and $\mathbf{s}_j$ and $\det\{[C](\mathbf{s}_1, \ldots, \mathbf{s}_n)\}$ denotes the determinant with (i, j)-th entry, $C(\mathbf{s}_i, \mathbf{s}_j)$. Then $\mathcal{S}$ is called a realization of a determinantal point pattern with covariance kernel C restricted to D; we write $\mathcal{S} \sim DPP_D(C)$. Hence, the first order density function (Section 2.1) is $f(\mathbf{s}) = C(\mathbf{s}, \mathbf{s})$ and the pair correlation function (Section 2.1) is

$$g(\mathbf{s}_1, \mathbf{s}_2) = 1 - \frac{C(\mathbf{s}_1, \mathbf{s}_2)C(\mathbf{s}_2, \mathbf{s}_1)}{C(\mathbf{s}_1, \mathbf{s}_1)C(\mathbf{s}_2, \mathbf{s}_2)}, \quad \text{if} \quad C(\mathbf{s}_1, \mathbf{s}_1) > 0 \quad \text{and} \quad C(\mathbf{s}_2, \mathbf{s}_2) > 0 \tag{2.7}$$

whereas it is 0 otherwise. Since C is a covariance function, then $f^{(n)}(\mathbf{s}_1, \ldots, \mathbf{s}_n) \leq f(\mathbf{s}_1) \ldots f(\mathbf{s}_n)$ for any $n > 1$, implying repulsion, and $g \leq 1$ [119].

2.6. Marked point processes

Marked point process models add considerable opportunity to understand point pattern behavior since each point carries the extra information of a mark which captures a feature of whatever object was observed at that point. Marks are often discrete such as providing labels for different types of cancers over a point pattern of cancer cases or labels for different species of trees in a forest. We would be interested in seeing differences between the point patterns; aggregating them would lose this opportunity.

Continuous marks may also be of interest. For trees, we may record a height or a basal area. For an earthquake, the mark may be its strength, say on the Richter

scale. We view both the location and the mark as random, contrasting with the usual geostatistical analysis [16] where only the feature at a location is viewed as random.

From a mathematical perspective, a mark is merely viewed as adding an extra coordinate to the observation, i.e., we observe $(\mathbf{s}, m)$ as a point over $D \times M$ where M is the support set for the marks. If the marks are continuous, M will be some subset of $\mathbb{R}^1$ and the marked point pattern is equivalent to a point pattern over $D \times M$. If the marks are discrete, M will be a set of labels, say $l, l = 1, 2, \ldots, L$ and the overall point pattern can be viewed as a set of L point patterns, each over D. The notation $(\mathbf{s}, m)$ is often modified according to interpretation of the marked point process. The notation $m(\mathbf{s}), \mathbf{s} \in \mathcal{S}$ suggests drawing locations and then assigning labels. The notation $\mathcal{S}_m$ suggests selecting labels and then drawing locations.

If we follow a product space representation for a marked point process, then a marked point process is really just a point process over this product space. So, we can adopt much of the earlier theory. For instance, $N(B \times A)$ is the number of points with locations in $B \subseteq D$ and marks in $A \subseteq M$. Defining a random counting measure leads us defining count random variables on a σ-algebra of sets over $D \times M$. In turn, this suggests a Poisson marked point process model where $N(B \times A) \sim Po(\lambda(B \times A))$ for a suitable intensity measure $\lambda(B \times A)$, with independence of the count variables over disjoint product sets.

The process may have an intensity function, i.e., $\lambda(\mathbf{s}, m)$ such that $E(N(B \times A)) = \lambda(B \times A) = \int_B \int_A \lambda(\mathbf{s}, m) d\mu(m) d\mathbf{s}$. If the marks are continuous, we usually take $\mu(m)$ to be Lebesgue measure. If the marks are discrete/categorical, we take $\mu(m)$ to be counting measure and write $E(N(B \times A)) = \lambda(B \times A) = \int_B \sum_{l \in A} \lambda_l(\mathbf{s}) d\mathbf{s}$.

With continuous marks, integrating over m yields $\lambda(\mathbf{s}) = \int_M \lambda(\mathbf{s}, m) dm$, the intensity for the point process of locations. In fact, $f(m|\mathbf{s}) = \frac{\lambda(\mathbf{s}, m)}{\lambda(\mathbf{s})}$ is the conditional density for the mark at location $\mathbf{s}$. For categorical marks, the marginal intensity is $\lambda(\mathbf{s}) = \sum_{l=1}^{L} \lambda_l(\mathbf{s})$. Now, the conditional probability for mark l at location $\mathbf{s}$ is $\frac{\lambda_l(\mathbf{s})}{\lambda(\mathbf{s})}$.

We can also envision random field mark processes. Such processes assume that the realization of the point pattern $\mathcal{S}$ over D is independent of the realization of the stochastic process $m_D = \{m(\mathbf{s}) : \mathbf{s} \in D\}$. This model would be appropriate if we wanted to assign marks that exhibited spatial structure, e.g., marks are more similar at locations closer to each other than at locations more distant from each other.

Another option is to extend Gibbs process models with discrete marks. Suppose we introduce a Gibbs process for each l. Then, we can obtain a Papangelou conditional intensity for each l. Now, we might investigate relative conditional intensities.

Modeling clarification for marked point processes

Clarification of some modeling issues may prove useful here. Suppose we obtain data in the form $(\mathbf{s}_i, m_i), i = 1, 2, \ldots, n$, where the $\mathbf{s}_i$'s are observed locations and we think of $m_i = L(\mathbf{s}_i)$ as a discrete label, say from $l = 1, 2, \ldots, L$. So, we think of the $L(\cdot)$'s as marks, indicating which mark was assigned to each of the observed points. If we ignored the marks, under an NHPP model, we know the joint distribution of $(n, \{s_1, s_2, \ldots, s_n\}|\lambda_D)$ where $\lambda_D = \{\lambda(\mathbf{s}) : \mathbf{s} \in D\}$. From a Bayesian perspective, we only need to model λ_D to complete the specification. Adopting this perspective, with marks as above, we imagine a point pattern for

each label/mark value and would extend to $\lambda_{l,D}$, the intensity associated with label value l. Then, assuming marks are also random, we would assign a prior on labels say $p_l, l = 1, 2, ...L$. In this fashion, we model the joint distribution of location and label as $[location|label][label]$ and we would assume the pairs $(\mathbf{s}_i, L(\mathbf{s}_i))$ are conditionally independent given the $\lambda_l(\mathbf{s})$'s. Under this modeling, we have specified $[\mathcal{S}|\ L = l; \{\lambda_l(\mathbf{s}) : \mathbf{s} \in D\}]$.

The cumulative intensity is $\lambda(\mathbf{s}) = \sum_l \lambda_l(\mathbf{s})$, which has nothing to do with the p_l's, and

$$f(\mathbf{s}) = \frac{\lambda(\mathbf{s})}{\lambda(D)} = \sum_l \frac{\lambda_l(\mathbf{s})}{\sum_l \lambda_l(D)}$$

is the marginal location density. Turning to the joint distribution, we have $f_l(\mathbf{s})p_l$ where $f_l(\mathbf{s}) = \lambda_l(\mathbf{s})/\lambda_l(D)$, the location density associated with mark l. We interpret $f_l(\mathbf{s})p_l$ as drawing a label $L = l$ and then locating the label at $\mathbf{s}$ given $L = l$. That is, the draw $(\mathbf{s}, L = l)$ creates the event $(\mathbf{s}, L(\mathbf{s}) = l)$. Note that this has nothing to do with the joint intensity $\lambda(\mathbf{s}, m)$ which adopts counting measure for m when m is discrete. This joint intensity yields $\lambda(\partial\mathbf{s}, \{l\}) \approx \lambda(\mathbf{s}, l)|\partial\mathbf{s}|$ and thus $\lambda(\mathbf{s}, l) = \lambda_l(\mathbf{s})$, again, free of the p_l's.

Under the above modeling, suppose we imagine drawing a location and then assigning a label to the location. The draw $(\mathbf{s}, L = l)$ creates the event $(\mathbf{s}, L(\mathbf{s}) = l)$. Using Bayes' Theorem, we can calculate $P(L = l|\mathbf{s}) = f_l(\mathbf{s})p_l/\tilde{f}(\mathbf{s})$. We have the familiar rescaling of the prior weights with $\tilde{f}(\mathbf{s}) = \sum_l p_l f_l(\mathbf{s})$, a mixture density having nothing to do with the foregoing marginal location density $f(\mathbf{s})$.

We can imagine that the modeling situation is reversed. The label is viewed as the response at a location; now we would be modeling the joint distribution as $[label|location][location]$. The model for location would now have a single λ_D and the distribution for label given location would be a multinomial trial with location-specific probabilities. In general, the joint distribution becomes $P(L = l|\mathbf{s})f(\mathbf{s})$ where, as usual, $f(\mathbf{s}) = \lambda(\mathbf{s})/\lambda(D)$. This means that we draw a location $\mathbf{s}$ and then assign a label $L = l$ to the location, creating the event $(\mathbf{s}, L(\mathbf{s}) = l)$. Turning to Bayes' Theorem, $f(\mathbf{s}|L = l) = P(L = l|\mathbf{s})f(\mathbf{s})/\int_D P(L = l|\mathbf{s})f(\mathbf{s})d\mathbf{s}$ and, in fact, $f(\mathbf{s}|L = l) = f_l(\mathbf{s})$, the location density associated with mark l. Thus, $\lambda_l(\mathbf{s}) = c_l P(L = l|\mathbf{s})f(\mathbf{s})$ where the constant c_l can not be identified; we can only learn about the location density for mark l but not the intensity for this mark. A last calculation shows that $\lambda(\mathbf{s}) = \sum_l \lambda_l(\mathbf{s}) = \frac{\sum_l c_l P(L=l|\mathbf{s})}{\lambda(D)}\lambda(\mathbf{s})$. Hence, $\sum_l c_l P(L = l|\mathbf{s}) = \lambda(D)$ but, again, the c_l are not determined.

In summary, note the fundamental difference between the two joint modeling scenarios. In the first case, it is most natural that the L's have nothing to do with locations. There is a single distribution for them and given a realization (label), we have an associated intensity which provides the joint distribution for the points having that label. In the second case, we formalize an uncountable collection of $L(\mathbf{s})$'s with a single intensity for the observed points. In other words, conceptually, the joint distributions for the first case live in a different space from the joint distributions for the second case. See, also, [8] in this regard.